

Editor's Special Invited Paper

Two-Stage Procedures for High-Dimensional Data

Makoto Aoshima¹ and Kazuyoshi Yata²

^{1,2}Institute of Mathematics, University of Tsukuba, Ibaraki, Japan

Abstract: In this paper, we consider a variety of inference problems for high-dimensional data. The purpose of this paper is to suggest directions for future research and possible solutions about $p \gg n$ problems by using new types of two-stage estimation methodologies. This is the first attempt to apply sequential analysis to high-dimensional statistical inference ensuring prespecified accuracy. We offer the sample size determination for inference problems by creating new types of multivariate two-stage procedures. To develop theory and methodologies, most important and basic idea is the asymptotic normality when $p \rightarrow \infty$. By developing asymptotic normality when $p \rightarrow \infty$, we first give (1) a *given-bandwidth confidence region* for the square loss. In addition, we give (2) a *two-sample test to assure prespecified size and power simultaneously* together with (3) an *equality-test procedure for two covariance matrices*. We also give (4) a *two-stage discriminant procedure that controls misclassification rates* being no more than a prespecified value. Moreover, we propose (5) a *two-stage variable selection procedure* that provides screening of variables in the first stage and selects a significant set of associated variables from among a set of candidate variables in the second stage. Following the variable selection procedure, we consider (6) *variable selection for high-dimensional regression* to compare favorably with the Lasso in terms of the assurance of accuracy and the computational cost. Further, we consider variable selection for classification and propose (7) a *two-stage discriminant procedure after screening some variables*. Finally, we consider (8) *pathway analysis for high-dimensional data by constructing a multiple test of correlation coefficients*.

Keywords: Asymptotic normality; Classification; Confidence region; HDLSS; Lasso; Pathway analysis; Regression; Sample size determination; Testing equality of covariance matrices; Two-sample test; Variable selection.

Subject Classifications: 62L10, 62H10; 60F05.

1. INTRODUCTION

High-dimensional data situation occurs in many areas of modern science such as genetic microarrays, medical imaging, text recognition, finance, chemometrics, and so on. A common feature of high-dimensional data is that, while the data dimension is high, the sample size is relatively small. This is the so-called “HDLSS” or “large p , small n ” situation where $p/n \rightarrow \infty$; here p is the data dimension and n is the sample size.

The asymptotic studies of this type of data are becoming increasingly relevant. In recent years, substantial work has been done on the asymptotic behavior of eigenvalues of

Address correspondence to Makoto Aoshima, Institute of Mathematics, University of Tsukuba, Ibaraki 305-8571, Japan; Fax: +81-298-53-6501; E-mail: aoshima@math.tsukuba.ac.jp

the sample covariance matrix in the limit as $p \rightarrow \infty$, see Johnstone (2001), Baik et al. (2005) and Paul (2007) for Gaussian assumptions and Baik and Silverman (2006) for non-Gaussian but i.i.d. assumptions when p and n increase at the same rate, i.e. $n/p \rightarrow c > 0$. The HDLSS asymptotics, where only $p \rightarrow \infty$ while n is fixed, were studied by Hall et al. (2005), Ahn et al. (2007) and Yata and Aoshima (2010c). They explored conditions to give a geometric representation of HDLSS data. Jung and Marron (2009) investigated consistency properties of both eigenvalues and eigenvectors of the sample covariance matrix in the HDLSS data situations. Many of these focus on the spiked covariance model introduced by Johnstone (2001). The HDLSS asymptotics usually regulate either the population distribution by the normality or the dependency of the random variables in the sphered data matrix by a ρ -mixing condition. However, Yata and Aoshima (2010b) have developed the HDLSS asymptotics without assuming either the normality or a ρ -mixing condition. In addition, Yata and Aoshima (2009b) have succeeded in investigating the consistency properties of both eigenvalues and eigenvectors of the sample covariance matrix in more general settings that include the case when all eigenvalues are in the range of sphericity. Furthermore, Yata and Aoshima (2010a) have recently developed the *cross-data-matrix methodology* that provides effective inference on PCA and clustering for HDLSS data.

Suppose we have independent and p -variate populations, π_i , $i = 1, \dots, k$, having *unknown* mean vector $\boldsymbol{\mu}_i = (\mu_{i1}, \dots, \mu_{ip})^T$ and *unknown* covariance matrix $\boldsymbol{\Sigma}_i (> \mathbf{O})$ for each i . We do not assume that $\boldsymbol{\Sigma}_1 = \dots = \boldsymbol{\Sigma}_k$. The eigen-decomposition of $\boldsymbol{\Sigma}_i$ ($i = 1, \dots, k$) is $\boldsymbol{\Sigma}_i = \mathbf{H}_i \boldsymbol{\Lambda}_i \mathbf{H}_i^T$, where $\boldsymbol{\Lambda}_i$ is a diagonal matrix of eigenvalues $\lambda_{i1} \geq \dots \geq \lambda_{ip} > 0$ and $\mathbf{H}_i = [\mathbf{h}_{i1}, \dots, \mathbf{h}_{ip}]$ is an orthogonal matrix of corresponding eigenvectors. Having recorded i.i.d. samples, $\mathbf{x}_{i1}, \dots, \mathbf{x}_{in_i}$, from each π_i , we have a $p \times n_i$ ($p > n_i$) data matrix $\mathbf{X}_i = [\mathbf{x}_{i1}, \dots, \mathbf{x}_{in_i}]$, where $\mathbf{x}_{ij} = (x_{i1j}, \dots, x_{ipj})^T$, $j = 1, \dots, n_i$. We assume $n_i \geq 4$, $i = 1, \dots, k$. Then, $\mathbf{Z}_i = \boldsymbol{\Lambda}_i^{-1/2} \mathbf{H}_i^T (\mathbf{X}_i - [\boldsymbol{\mu}_i, \dots, \boldsymbol{\mu}_i])$ is a $p \times n_i$ sphered data matrix from a distribution with the identity covariance matrix. Here, we write $\mathbf{Z}_i = [\mathbf{z}_{i1}, \dots, \mathbf{z}_{in_i}]$ and $\mathbf{z}_{ij} = (z_{i1j}, \dots, z_{ipj})^T$, $j = 1, \dots, n_i$. Note that $E(z_{ijl}^2) = 1$ and $E(z_{ijl} z_{ijl'}) = 0$ for $i = 1, \dots, k$; $j (\neq j') = 1, \dots, p$; $l = 1, \dots, n_i$. We assume that $\lambda_{ip} > 0$ ($i = 1, \dots, k$) as $p \rightarrow \infty$ and the fourth moments of each variable in \mathbf{Z}_i are uniformly bounded. Let $\sigma_{(i)j} (> 0)$, $j = 1, \dots, p$, be diagonal elements of $\boldsymbol{\Sigma}_i$. In this paper, we assume one of the following three assumptions for π_i 's as necessary:

$$\text{(A-i)} \quad \pi_i : N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \text{ for } i = 1, \dots, k;$$

$$\text{(A-ii)} \quad z_{ijl}, j = 1, \dots, p \text{ are independent for } i = 1, \dots, k;$$

$$\text{(A-iii)} \quad E(z_{ijl}^2 z_{isl}^2) = 1 \text{ and } E(z_{ijl} z_{isl} z_{itl} z_{iul}) = 0, j \neq s, t, u, \text{ and } \{x_{ijl} - \mu_{ij}\}_{j \in N} \text{ is a strictly stationary sequence and } \rho\text{-mixing for } i = 1, \dots, k.$$

Note that (A-i) implies (A-ii). We also assume the following condition for $\boldsymbol{\Sigma}_i$'s as necessary:

$$\text{(A-iv)} \quad \frac{\text{tr}(\boldsymbol{\Sigma}_i^t)}{p} < \infty \text{ (} t = 1, 2 \text{) and } \frac{\text{tr}(\boldsymbol{\Sigma}_i^4)}{p^2} \rightarrow 0 \text{ as } p \rightarrow \infty \text{ for } i = 1, \dots, k.$$

We assume the following extra condition when applying (A-iii):

$$\text{(A-v)} \quad \frac{\text{tr}(\boldsymbol{\Sigma}_i \boldsymbol{\Sigma}_j)}{p} \rightarrow c_{ij} \text{ as } p \rightarrow \infty \text{ for all } i, j = 1, \dots, k, \text{ where } c_{ij}\text{'s are positive constants.}$$

Remark 1.1. If all λ_{ij} 's are bounded, (A-iv) trivially holds. For a spiked model such as $\lambda_{ij} = a_{ij} p^{\alpha_{ij}}$ ($j = 1, \dots, m_i$) and $\lambda_{ij} = c_{ij}$ ($j = m_i + 1, \dots, p$) with positive constants

a_{ij} 's, c_{ij} 's and α_{ij} 's, (A-iv) holds under the condition that $\alpha_{ij} < 1/2$, $j = 1, \dots, m_i (< \infty)$, $i = 1, \dots, k$. See Yata and Aoshima (2009b, 2010a) for the details of a spiked model. As an interesting example, both (A-iv) and (A-v) hold for $\Sigma_{i'} = c_{i'}(\rho_{i'}^{|i-j|^{q_{i'}}})$, $i' = 1, \dots, k$, where $c_{i'}$'s, $q_{i'}$'s and $\rho_{i'}$'s (< 1) are positive constants.

The concept of ρ -mixing was first developed by Kolmogorov and Rozanov (1960). See Bradley (2005) for a clear and insightful discussion. See also Jung and Marron (2009). For $-\infty \leq J \leq K \leq \infty$, let \mathcal{F}_J^K denote the σ -field of events generated by the random variables $(Y_j, J \leq j \leq K)$. For any σ -field \mathcal{A} , let $L_2(\mathcal{A})$ denote the space of square-integrable, \mathcal{A} measurable (real-valued) random variables. For each $r \geq 1$, define the maximal correlation coefficient

$$\rho(r) = \sup |Corr(f, g)|, \quad f \in L_2(\mathcal{F}_{-\infty}^i), g \in L_2(\mathcal{F}_{i+r}^\infty),$$

where sup is over all f, g , and i is a positive integer. The sequence $\{Y_j\}$ is said to be ρ -mixing if $\rho(r) \rightarrow 0$ as $r \rightarrow \infty$. Note that when $(x_{i1l}, x_{i2l}, \dots)$ is ρ -mixing and $Var(x_{ijl}^2) < \infty$, $j = 1, 2, \dots$, it holds for $j, j' = 1, 2, \dots$ with $|j - j'| = r$ that

$$|Corr(x_{ijl}^2, x_{ij'l}^2)| \leq \rho(r) \rightarrow 0 \quad \text{as } r \rightarrow \infty.$$

Let $\boldsymbol{\mu} = \sum_{i=1}^k b_i \boldsymbol{\mu}_i$ with b_i 's known and nonzero scalars. Let $\mathbf{T}_n = \sum_{i=1}^k b_i \bar{\mathbf{x}}_{in_i}$, where $\mathbf{n} = (n_1, \dots, n_k)$ and $\bar{\mathbf{x}}_{in_i} = \sum_{j=1}^{n_i} \mathbf{x}_{ij}/n_i$. One choice of making inference on $\boldsymbol{\mu}$ is to construct a confidence region by $\mathbf{R}_n = \{\boldsymbol{\mu} \in R^p : \|\mathbf{T}_n - \boldsymbol{\mu}\| \leq d\}$, where $\|\cdot\|$ denotes the Euclidean norm. Let $\boldsymbol{\theta} = (\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_k)$. For given and fixed $d (> 0)$ and $\alpha \in (0, 1)$, the requirement is established by

$$P_{\boldsymbol{\theta}}(\boldsymbol{\mu} \in \mathbf{R}_n) \geq 1 - \alpha. \quad (1.1)$$

There is a huge literature out there addressing scenarios related to this problem when p is fixed less than n_i . One may refer to Ghosh et al. (1997), Aoshima and Mukhopadhyay (1998), Aoshima et al. (2002), Aoshima et al. (2003), Aoshima and Takada (2004), Aoshima (2005) and Yata and Aoshima (2009a) among others in which Stein (1945)-type two-stage procedures were proposed in a typical multivariate context. Recently, Aoshima and Yata (2010) provided a general methodology to make a Stein-type two-stage procedure asymptotically second-order consistent for a variety of multivariate inference problems such as multiple comparisons and bioequivalence tests. For the concept of second-order efficiency, refer to Ghosh et al. (1997). In a high-dimensional case, those methodologies tend to satisfy the probability requirement such as (1.1) excessively by taking overly samples. To overcome this inconvenience, Yata (2010) gave a two-stage procedure that meets the equality in (1.1) approximately with a moderate sample size when p is large. However, the high-dimensional cases discussed by Yata (2010) were restricted to a high dimension, large sample size context such as $p/n_i < \infty$.

In this paper, we consider a variety of inference problems for high-dimensional data such as $p/n \rightarrow \infty$ in the context of sequential analysis. *The most challenging issue is to develop the new asymptotic theory when $p \rightarrow \infty$ instead of the large sample asymptotic theory in which $n \rightarrow \infty$ while p is fixed.* We emphasize that high-dimensional statistical inference can be ensured prespecified accuracy with the help of the new asymptotic theory when $p \rightarrow \infty$. *We do not assume $\boldsymbol{\Sigma}_1 = \dots = \boldsymbol{\Sigma}_k$* because it is a rather strong assumption and most importantly such an assumption is difficult to verify specially for non-Gaussian and high-dimensional data.

The purpose of this paper is to suggest directions for future research and possible solutions about $p \gg n$ problems by using new types of two-stage estimation methodologies. This is the first attempt to apply sequential analysis to high-dimensional statistical inference ensuring prespecified accuracy. We offer the sample size determination for each inference problem by creating new types of multivariate two-stage procedures. *To develop theory and methodologies, most important and basic idea is the asymptotic normality when $p \rightarrow \infty$.* We develop the asymptotic normality when $p \rightarrow \infty$ for the high-dimensional statistics given in this paper. We emphasize that one cannot apply the existing multivariate two-stage procedures, that are based on the large sample asymptotic theory, to $p \gg n$ problems because of the curse of dimensionality. One may also refer to Sen et al. (2007).

In Section 2, we consider a new type of confidence region that has a given-bandwidth for the square loss. Here, $\boldsymbol{\mu}$ is included in a region sandwiched by two p -dimensional spheres with a certain radius from centre \mathbf{T}_n . We give a two-stage estimation procedure to assure a prespecified coverage probability. In Section 3, we give a two-stage test procedure that provides a two-sample test having prespecified size and power together with a equality-test procedure for two covariance matrices. In Section 4, we give a two-stage discriminant procedure that controls misclassification rates being no more than a prespecified value. In Section 5, we propose a two-stage variable selection procedure that provides screening of variables in the first stage. We select a significant set of associated variables from among a set of candidate variables in the second stage. In Section 6, we consider variable selection for high-dimensional regression to compare favorably with the Lasso in terms of the assurance of accuracy and the computational cost. In Section 7, we consider variable selection for classification and propose a two-stage discriminant procedure after screening some variables. Finally, in Section 8, we consider pathway analysis for high-dimensional data by constructing a multiple test of correlation coefficients.

Throughout this paper, let $n_{i1} = \lfloor n_i/2 \rfloor + 1$ and $n_{i2} = n_i - n_{i1}$, where $\lfloor x \rfloor$ denotes the largest integer less than x . We define for each π_i

$$\begin{aligned} \mathbf{S}_{in_i} &= \frac{\sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_{in_i})(\mathbf{x}_{ij} - \bar{\mathbf{x}}_{in_i})^T}{n_i - 1}, & \mathbf{S}_{in_i(1)} &= \frac{\sum_{j=1}^{n_{i1}} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_{in_{i1}})(\mathbf{x}_{ij} - \bar{\mathbf{x}}_{in_{i1}})^T}{n_{i1} - 1}, \\ \text{and } \mathbf{S}_{in_i(2)} &= \frac{\sum_{j=n_{i1}+1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_{in_{i2}})(\mathbf{x}_{ij} - \bar{\mathbf{x}}_{in_{i2}})^T}{n_{i2} - 1}, \end{aligned} \quad (1.2)$$

where $\bar{\mathbf{x}}_{in_{i1}} = \sum_{j=1}^{n_{i1}} \mathbf{x}_{ij}/n_{i1}$ and $\bar{\mathbf{x}}_{in_{i2}} = \sum_{j=n_{i1}+1}^{n_i} \mathbf{x}_{ij}/n_{i2}$.

2. CONFIDENCE REGION FOR HIGH-DIMENSIONAL DATA

First, we note that a HDLSS data set has a geometric representation given by Hall et al. (2005). A confidence region defined by (1.1) is not available for a given and fixed $d (> 0)$ in the HDLSS context. Let $\Sigma_n = \sum_{i=1}^k b_i^2 \text{tr}(\Sigma_i)/n_i$. Note that $E_{\boldsymbol{\theta}}(\|\mathbf{T}_n - \boldsymbol{\mu}\|^2) = \Sigma_n$. Under (A-i) and (A-iv), we have as $p \rightarrow \infty$ that $\|\mathbf{T}_n - \boldsymbol{\mu}\|^2/\Sigma_n = 1 + o_p(1)$. Thus it holds that $\|\mathbf{T}_n - \boldsymbol{\mu}\|^2$ behaves around Σ_n and $P_{\boldsymbol{\theta}}(\boldsymbol{\mu} \in \mathbf{R}_n) = P_{\boldsymbol{\theta}}(\|\mathbf{T}_n - \boldsymbol{\mu}\| \leq d) \rightarrow 0$ under $n_i/p \rightarrow 0$, $i = 1, \dots, k$. In this section, we consider constructing a given-bandwidth confidence region for the square loss defined by

$$\mathbf{R}_{\Sigma_n} = \{\boldsymbol{\mu} \in R^p : \max\{-\delta + \Sigma_n, 0\} \leq \|\mathbf{T}_n - \boldsymbol{\mu}\|^2 \leq \delta + \Sigma_n\} \quad (2.1)$$

for given $\delta (> 0)$. We assume $\delta = o(p^{1/2})$. For given $\delta (> 0)$ and $\alpha \in (0, 1)$, we are interested

in constructing a confidence region $\mathbf{R}_{\Sigma_{\mathbf{n}}}$ such that

$$P_{\theta}(\boldsymbol{\mu} \in \mathbf{R}_{\Sigma_{\mathbf{n}}}) \geq 1 - \alpha. \quad (2.2)$$

2.1. ASYMPTOTIC NORMALITY AND SAMPLE SIZE DETERMINATION

Let us consider the region $\mathbf{R}_{\Sigma_{\mathbf{n}}}$. It indicates for $\Sigma_{\mathbf{n}} > \delta$ that $\boldsymbol{\mu}$ is included in the region sandwiched by the two p -dimensional spheres with radius of $\sqrt{\Sigma_{\mathbf{n}} + \delta}$ and $\sqrt{\Sigma_{\mathbf{n}} - \delta}$ from centre $\mathbf{T}_{\mathbf{n}}$. In Fig.1, the gray zone represents the sandwiched region when $p = 2$. Note that one can control the loss function $\|\mathbf{T}_{\mathbf{n}} - \boldsymbol{\mu}\|^2$ by using $\mathbf{R}_{\Sigma_{\mathbf{n}}}$.

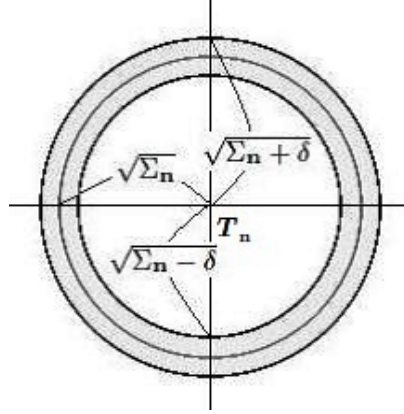


Figure 1. The gray zone represents a confidence region $\mathbf{R}_{\Sigma_{\mathbf{n}}}$ when $p = 2$.

We have the following theorems.

Theorem 2.1. Assume (A-i) and (A-iv). Then, we have that

$$\frac{\|\mathbf{T}_{\mathbf{n}} - \boldsymbol{\mu}\|^2 - \Sigma_{\mathbf{n}}}{\sqrt{2 \sum_{i,j} b_i^2 b_j^2 \text{tr}(\boldsymbol{\Sigma}_i \boldsymbol{\Sigma}_j) / (n_i n_j)}} \Rightarrow N(0, 1) \quad (2.3)$$

when $p \rightarrow \infty$ and either $n_i \rightarrow \infty$ or n_i is fixed for $i = 1, \dots, k$, where “ \Rightarrow ” denotes the convergence in distribution and $N(0, 1)$ denotes a random variable distributed as the standard normal distribution.

Theorem 2.2. Assume (A-iv) and either (A-ii) or (A-iii) with (A-v). Then, we have (2.3) as $p \rightarrow \infty$ and $n_i \rightarrow \infty$, $i = 1, \dots, k$.

It should be noted that the result in Theorem 2.1 can be claimed even when n_i is fixed for $i = 1, \dots, k$. The condition that $p \rightarrow \infty$ and $n_i \rightarrow \infty$ does not restrict n_i to either $p/n_i \rightarrow \infty$ or $p/n_i < \infty$.

Corollary 2.1. Let $\hat{\Sigma}_{\mathbf{n}} = \sum_{i=1}^k b_i^2 \text{tr}(\mathbf{S}_{in_i}) / n_i$. Assume (A-iv) and either (A-ii) or (A-iii) with (A-v). Then, we have as $p \rightarrow \infty$ and $n_i \rightarrow \infty$, $i = 1, \dots, k$, that

$$\frac{\|\mathbf{T}_{\mathbf{n}} - \boldsymbol{\mu}\|^2 - \hat{\Sigma}_{\mathbf{n}}}{\sqrt{2 \sum_{i,j} b_i^2 b_j^2 \text{tr}(\boldsymbol{\Sigma}_i \boldsymbol{\Sigma}_j) / (n_i n_j)}} \Rightarrow N(0, 1).$$

From the fact that $\text{tr}(\mathbf{\Sigma}_i \mathbf{\Sigma}_j) \leq \sqrt{\text{tr}(\mathbf{\Sigma}_i^2) \text{tr}(\mathbf{\Sigma}_j^2)}$, it holds that $\sqrt{\sum_{i,j} b_i^2 b_j^2 \text{tr}(\mathbf{\Sigma}_i \mathbf{\Sigma}_j) / (n_i n_j)} \leq \sum_{i=1}^k b_i^2 \sqrt{\text{tr}(\mathbf{\Sigma}_i^2) / n_i}$. Let $z_{\alpha/2}$ be the upper $\alpha/2$ point of $N(0, 1)$. Here, we consider n_i 's such that

$$\min \sum_{i=1}^k n_i \quad \text{subject to} \quad \sqrt{2} \sum_{i=1}^k b_i^2 \sqrt{\text{tr}(\mathbf{\Sigma}_i^2) / n_i} \leq \delta / z_{\alpha/2}.$$

Then, we find the sample size for each π_i as

$$n_i \geq \frac{z_{\alpha/2} \sqrt{2}}{\delta} |b_i| \text{tr}(\mathbf{\Sigma}_i^2)^{1/4} \sum_{j=1}^k |b_j| \text{tr}(\mathbf{\Sigma}_j^2)^{1/4} \quad (= C_i, \text{ say}). \quad (2.4)$$

Note that $C_i = o(p/\delta)$ for $i = 1, \dots, k$, under (A-iv). Thus it holds that $C_i/p \rightarrow 0$ as $p \rightarrow \infty$. We also notice that $n_i \rightarrow \infty$, $i = 1, \dots, k$, as $p \rightarrow \infty$. Let

$$\mathbf{R}_{\widehat{\Sigma}_{\mathbf{n}}} = \{\boldsymbol{\mu} \in R^p : \max\{-\delta + \widehat{\Sigma}_{\mathbf{n}}, 0\} \leq \|\mathbf{T}_{\mathbf{n}} - \boldsymbol{\mu}\|^2 \leq \delta + \widehat{\Sigma}_{\mathbf{n}}\}, \quad (2.5)$$

where $\widehat{\Sigma}_{\mathbf{n}} = \sum_{i=1}^k b_i^2 \text{tr}(\mathbf{S}_{in_i}) / n_i$. Then, we have the following theorem.

Theorem 2.3. *Assume (A-iv) and either (A-ii) or (A-iii) with (A-v). Then, for n_i satisfying (2.4), we have as $p \rightarrow \infty$ that*

$$\liminf P_{\boldsymbol{\theta}}(\boldsymbol{\mu} \in \mathbf{R}_{\widehat{\Sigma}_{\mathbf{n}}}) \geq 1 - \alpha.$$

Remark 2.1. The same assertion as in Theorem 2.3 holds for $\mathbf{R}_{\Sigma_{\mathbf{n}}}$.

2.2. TWO-STAGE PROCEDURE FOR CONFIDENCE REGION

Since $\mathbf{\Sigma}_i$'s are unknown, it is necessary to estimate C_i 's in (2.4) with some pilot samples. We consider a two-stage procedure to construct a confidence region $\mathbf{R}_{\widehat{\Sigma}_{\mathbf{n}}}$. Along the line of Mukhopadhyay and Duggan (1997, 1999), we suppose the following assumption: There exists a known and positive lower bound $\sigma_{i\star}$ for $\sqrt{\text{tr}(\mathbf{\Sigma}_i^2)}$ such that $\sigma_{i\star} / \sqrt{\text{tr}(\mathbf{\Sigma}_i^2)} \in (0, 1)$, $i = 1, \dots, k$, as $p \rightarrow \infty$. We proceed the following two steps:

1. Let $\tau_{\star} = \min_{1 \leq i \leq k} |b_i| \sqrt{\sigma_{i\star}} \sum_{j=1}^k |b_j| \sqrt{\sigma_{j\star}}$. Having a fixed integer m_0 (≥ 4), define

$$m = \max \left\{ m_0, \left\lceil \frac{z_{\alpha/2} \sqrt{2}}{\delta} \tau_{\star} \right\rceil + 1 \right\}. \quad (2.6)$$

According to (2.6), take pilot samples \mathbf{x}_{ij} , $j = 1, \dots, m$, of size m from each π_i . Then, calculate \mathbf{S}_{im} , $\mathbf{S}_{im(1)}$ and $\mathbf{S}_{im(2)}$ according to (1.2) for each π_i . Define the total sample size for each π_i by

$$N_i = \max \left\{ m, \left\lceil \frac{z_{\alpha/2} \sqrt{2}}{\delta} |b_i| \text{tr}(\mathbf{S}_{im(1)} \mathbf{S}_{im(2)})^{1/4} \sum_{j=1}^k |b_j| \text{tr}(\mathbf{S}_{jm(1)} \mathbf{S}_{jm(2)})^{1/4} \right\rceil + 1 \right\}. \quad (2.7)$$

Note that $\text{tr}(\mathbf{S}_{im(1)} \mathbf{S}_{im(2)}) \geq 0$ w.p.1. Let $\mathbf{N} = (N_1, \dots, N_k)$.

2. Take additional samples \mathbf{x}_{ij} , $j = m + 1, \dots, N_i$, of size $N_i - m$ from each π_i . By combining the initial samples and the additional samples, calculate $\bar{\mathbf{x}}_{iN_i} = \sum_{j=1}^{N_i} \mathbf{x}_{ij}/N_i$ and $\mathbf{S}_{iN_i} = \sum_{j=1}^{N_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_{iN_i})(\mathbf{x}_{ij} - \bar{\mathbf{x}}_{iN_i})^T / (N_i - 1)$ for each π_i . Then, define $\hat{\Sigma}_{\mathbf{N}} = \sum_{i=1}^k b_i^2 \text{tr}(\mathbf{S}_{iN_i})/N_i$ and the region $\mathbf{R}_{\hat{\Sigma}_{\mathbf{N}}}$ according to (2.5) with $\mathbf{T}_{\mathbf{N}} = \sum_{i=1}^k b_i \bar{\mathbf{x}}_{iN_i}$.

We have the following theorem.

Theorem 2.4. Assume (A-iv) and either (A-ii) or (A-iii) with (A-v). For the two-stage procedure given by (2.6)-(2.7), we have as $p \rightarrow \infty$ that

$$\liminf P_{\boldsymbol{\theta}}(\boldsymbol{\mu} \in \mathbf{R}_{\hat{\Sigma}_{\mathbf{N}}}) \geq 1 - \alpha. \quad (2.8)$$

Remark 2.2. Assume (A-iv) and either (A-ii) or (A-iii). It holds as $p \rightarrow \infty$ that $N_i/C_i = 1 + o_p(1)$, $i = 1, \dots, k$, that are in the HDLSS situation in the sense that $N_i/p = o_p(1)$, $i = 1, \dots, k$. Further, under (A-i), we have higher-order results as in the following theorem.

Theorem 2.5. Assume (A-i) and (A-iv). For the two-stage procedure given by (2.6)-(2.7), it holds as $p \rightarrow \infty$ that

$$\limsup |E_{\boldsymbol{\theta}}(N_i - C_i)| \leq 1 \quad \text{and} \quad \text{Var}_{\boldsymbol{\theta}}(N_i) = o(p^{1/2}/\delta) \quad \text{for } i = 1, \dots, k.$$

Remark 2.3. One of the choices of $\sigma_{i\star}$ is, for example, a positive lower bound, σ_{i0} , for $\text{tr}(\Sigma_i)/\sqrt{p}$ such that $\sigma_{i0}\sqrt{p}/\text{tr}(\Sigma_i) \in (0, 1)$ as $p \rightarrow \infty$. Then, it holds from Schwartz's inequality and (A-iv) that $0 < \sigma_{i0}/\sqrt{\text{tr}(\Sigma_i^2)} = (\sigma_{i0}\sqrt{p}/\text{tr}(\Sigma_i))(\text{tr}(\Sigma_i)/\sqrt{p\text{tr}(\Sigma_i^2)}) < 1$ as $p \rightarrow \infty$. We emphasize that the two-stage procedure still holds (2.8) as long as $\sigma_{i\star}/p^{1/2} > 0$ as $p \rightarrow \infty$ for $i = 1, \dots, k$. In that sense, the two-stage procedure is quite robust for the misidentification of $\sigma_{i\star}$.

Remark 2.4. Under (A-i), it holds that $E_{\boldsymbol{\theta}}\{\text{tr}(\mathbf{S}_{im}^2)\} = (1 + (m - 1)^{-1})\text{tr}(\Sigma_i^2) + \text{tr}(\Sigma_i)^2/(m - 1)$. Hence, the naive estimator of $\text{tr}(\Sigma_i^2)$ is overly biased when $p \rightarrow \infty$. Yata (2010) considered an unbiased estimator of $\text{tr}(\Sigma_i^2)$ by $\text{tr}(\mathbf{S}_{im(1)}\mathbf{S}_{im(2)})$. Note that $E_{\boldsymbol{\theta}}(\mathbf{S}_{im(1)}\mathbf{S}_{im(2)}) = \text{tr}(\Sigma_i^2)$ and $\text{tr}(\mathbf{S}_{im(1)}\mathbf{S}_{im(2)}) \geq 0$ w.p.1. Under either (A-ii) or (A-iii), it holds as $p \rightarrow \infty$ and $m \rightarrow \infty$ that

$$\text{Var}_{\boldsymbol{\theta}}\left(\frac{\text{tr}(\mathbf{S}_{im(1)}\mathbf{S}_{im(2)})}{\text{tr}(\Sigma_i^2)}\right) = \frac{8}{m^2}(1 + o(1)) + O\left(\frac{\text{tr}(\Sigma_i^4)}{\text{tr}(\Sigma_i^2)^2 m}\right).$$

On the other hand, Bai and Saranadasa (1996) and Srivastava (2005) considered an estimator of $\text{tr}(\Sigma_i^2)$ by $\widehat{\text{tr}}(\Sigma_i^2) = c_m^{-1}\{\text{tr}(\mathbf{S}_{im}^2) - \text{tr}(\mathbf{S}_{im})^2/(m - 1)\}$ with $c_m = (m - 2)(m + 1)/(m - 1)^2$. Then, it holds under (A-i) that $E_{\boldsymbol{\theta}}(\widehat{\text{tr}}(\Sigma_i^2)) = \text{tr}(\Sigma_i^2)$ and

$$\text{Var}_{\boldsymbol{\theta}}\left(\frac{\widehat{\text{tr}}(\Sigma_i^2)}{\text{tr}(\Sigma_i^2)}\right) = \frac{4}{m^2}(1 + o(1)) + \frac{8\text{tr}(\Sigma_i^4)}{\text{tr}(\Sigma_i^2)^2 m}(1 + o(1))$$

as $p \rightarrow \infty$ and $m \rightarrow \infty$. One might consider $\widehat{\text{tr}}(\Sigma_i^2)$ for $\text{tr}(\mathbf{S}_{im(1)}\mathbf{S}_{im(2)})$ in (2.7). However, it should be noted that $\widehat{\text{tr}}(\Sigma_i^2)$ is not unbiased unless (A-i) holds. In addition, it does not

hold $Var_{\theta}(\widehat{tr(\Sigma_i^2)})/tr(\Sigma_i^2) < \infty$ when the eighth moments of each variable in \mathbf{Z}_i are not uniformly bounded.

2.3. SIMULATION

In order to study the performance of the two-stage procedure given by (2.6)-(2.7), we took resort to computer simulations. We set $k = 2$, $p = 1600$, $b_1 = b_2 = 1$ and $\delta = 5$. Our goal was to construct a 95% given-bandwidth confidence region $\mathbf{R}_{\widehat{\Sigma}_{\mathbf{N}}}$. In other words, we set $\alpha = 0.05$. We set $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = (0, \dots, 0)^T$. Independent pseudorandom normal observations were generated for $\pi_i : N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, $i = 1, 2$. We considered the covariance matrix such as $\boldsymbol{\Sigma}_1 = c_1 \mathbf{B}(\rho_1^{|i-j|^{1/3}}) \mathbf{B}$ and $\boldsymbol{\Sigma}_2 = c_2 \mathbf{B}(\rho_2^{|i-j|^{1/3}}) \mathbf{B}$, where $\rho_i \in (0, 1)$, $i = 1, 2$, and

$$\mathbf{B} = \text{diag} \left(\sqrt{0.5 + 1/(p+1)}, \sqrt{0.5 + 2/(p+1)}, \dots, \sqrt{0.5 + p/(p+1)} \right). \quad (2.9)$$

Note that $tr(\boldsymbol{\Sigma}_i) = c_i p$ ($i = 1, 2$). We considered the following three cases: (i) $(c_1, c_2) = (1, 1)$ and $(\rho_1, \rho_2) = (0.3, 0.3)$, i.e., $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$; (ii) $(c_1, c_2) = (1, 1)$ and $(\rho_1, \rho_2) = (0.3, 0.4)$, i.e., $tr(\boldsymbol{\Sigma}_1) = tr(\boldsymbol{\Sigma}_2)$ and $tr(\boldsymbol{\Sigma}_1^2) \neq tr(\boldsymbol{\Sigma}_2^2)$; (iii) $(c_1, c_2) = (1, 1.5)$ and $(\rho_1, \rho_2) = (0.3, 0.3)$, i.e., $\boldsymbol{\Sigma}_2 = 1.5 \boldsymbol{\Sigma}_1$.

For the two-stage procedure (2.6)-(2.7), Table 1 gives the findings obtained by averaging the outcomes from 2000 ($= R$, say) replications. We set $\sigma_{i*} = tr(\boldsymbol{\Sigma}_i^2)^{1/2}/3$, $i = 1, 2$, so that $m = 20$. The findings for case (i) were given in the first block and the ones for cases (ii) and (iii) followed after the block. Under a fixed scenario, suppose that the r th replication ends with $N_i = n_{ir}$ ($i = 1, 2$) observations and the corresponding confidence region with $\mathbf{n}_r = (n_{1r}, n_{2r})$ for $r = 1, \dots, R$. Let $\bar{n}_i = R^{-1} \sum_{r=1}^R n_{ir}$ and $Var(n_i) = (R-1)^{-1} \sum_{r=1}^R (n_{ir} - \bar{n}_i)^2$. Then, \bar{n} ($= \bar{n}_1 + \bar{n}_2$) estimates $C = C_1 + C_2$ defined by (2.4) with its estimated variance $Var(n)$, computed analogously. In the end of the r th replication, we checked whether $\boldsymbol{\mu}$ does (or does not) belong to the corresponding confidence region and defined $P_r = 1$ (or 0) accordingly. Let $\bar{P} = R^{-1} \sum_{r=1}^R P_r$, which estimates the target coverage probability, having its estimated standard error $s(\bar{P})$ where $s^2(\bar{P}) = R^{-1} \bar{P}(1 - \bar{P})$.

Let us explain, for example, the entries from the second block for case (ii) in Table 1. We had $C_1 = 61.87$, $C_2 = 69.79$ and $C = 131.66$ from (2.4). From 2000 independent replications, we observed $\bar{n}_1 = 62.17$ ($\bar{n}_1 - C_1 = 0.30$), $\bar{n}_2 = 70.07$ ($\bar{n}_2 - C_2 = 0.28$), $\bar{n} = 132.24$ ($\bar{n} - C = 0.58$) and $\bar{p} = 0.950$ together with $Var(n_1) = 16.60$, $Var(n_2) = 27.08$, $Var(n) = 69.54$ and $s(\bar{p}) = 0.00487$. Throughout, the two-stage procedure constructed required confidence region successfully.

3. TWO-SAMPLE TEST FOR HIGH-DIMENSIONAL DATA

Suppose we have two independently distributed populations, π_i , $i = 1, 2$. We do not assume $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$. A well-pursued interest in high-dimensional data analysis is to test if the two high-dimensional populations have the same mean or not, namely,

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \quad \text{vs.} \quad H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2. \quad (3.1)$$

The hypothesis H_0 consists of p marginal hypotheses regarding the means of each data dimension. Bai and Saranadasa (1996), Srivastava (2007) and Chen and Qin (2010) considered testing hypothesis (3.1). We should note that Hotelling's classical T^2 test does not work for HDLSS situations. Let $\Delta = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2$. We are interested in designing a test of

Table 1. Required sample size and the coverage probability by (2.6)-(2.7) with $\delta = 5$.

C	\bar{n}	$\bar{n} - C$	$Var(n)$	\bar{P}	$s(\bar{P})$
Case (i): $(c_1, c_2) = (1, 1)$ and $(\rho_1, \rho_2) = (0.3, 0.3)$					
116.29	117.00	0.72	47.81	0.943	0.00518
58.14	58.50	0.36	15.13		
58.14	58.50	0.36	14.83		
Case (ii): $(c_1, c_2) = (1, 1)$ and $(\rho_1, \rho_2) = (0.3, 0.4)$					
131.66	132.24	0.58	69.54	0.950	0.00487
61.87	62.17	0.30	16.60		
69.79	70.07	0.28	27.08		
Case (iii): $(c_1, c_2) = (1, 1.5)$ and $(\rho_1, \rho_2) = (0.3, 0.3)$					
143.89	144.21	0.32	74.89	0.946	0.00505
64.68	64.88	0.20	17.53		
79.21	79.33	0.12	29.48		

(3.1) with size α and power no less than $1 - \beta$ when $\Delta \geq \Delta_L$, where $\alpha, \beta \in (0, 1/2)$ and $\Delta_L (> 0)$ are prespecified constants. We assume $\Delta_L = o(p^{1/2})$.

3.1. ASYMPTOTIC NORMALITY AND SAMPLE SIZE DETERMINATION

Having recorded $\mathbf{x}_{i1}, \dots, \mathbf{x}_{in_i}$ from each π_i , Chen and Qin (2010) gave an estimator of Δ by

$$\tilde{T}_{\mathbf{n}} = \sum_{i=1}^2 \frac{\sum_{j \neq j'}^{n_i} \mathbf{x}_{ij}^T \mathbf{x}_{ij'}}{n_i(n_i - 1)} - 2 \frac{\sum_{j=1}^{n_1} \sum_{j'=1}^{n_2} \mathbf{x}_{1j}^T \mathbf{x}_{2j'}}{n_1 n_2}.$$

We note that the above description is equivalent to

$$\tilde{T}_{\mathbf{n}} = \|\bar{\mathbf{x}}_{1n_1} - \bar{\mathbf{x}}_{2n_2}\|^2 - \sum_{i=1}^2 \frac{\text{tr}(\mathbf{S}_{in_i})}{n_i}. \quad (3.2)$$

They showed that $E_{\boldsymbol{\theta}}(\tilde{T}_{\mathbf{n}}) = \Delta$ and

$$Var_{\boldsymbol{\theta}}(\tilde{T}_{\mathbf{n}}) = \sum_{i=1}^2 \frac{2}{n_i(n_i - 1)} \text{tr}(\boldsymbol{\Sigma}_i^2) + \frac{4}{n_1 n_2} \text{tr}(\boldsymbol{\Sigma}_1 \boldsymbol{\Sigma}_2) + \sum_{i=1}^2 \frac{4}{n_i} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_i (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2). \quad (3.3)$$

We consider estimating $Var_{\boldsymbol{\theta}}(\tilde{T}_{\mathbf{n}})$ by

$$\hat{Var}_{\boldsymbol{\theta}}(\tilde{T}_{\mathbf{n}}) = \sum_{i=1}^2 \frac{2}{n_i(n_i - 1)} \text{tr}(\mathbf{S}_{in_i(1)} \mathbf{S}_{in_i(2)}) + \frac{4}{n_1 n_2} \text{tr}(\mathbf{S}_{1n_1} \mathbf{S}_{2n_2}).$$

Then, we have the following theorem.

Theorem 3.1. Assume that $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_i (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = o(\text{tr}(\boldsymbol{\Sigma}_i^2)/n_i)$, $i = 1, 2$. Assume also (A-iv) and either (A-ii) or (A-iii) with (A-v). Then, it holds as $p \rightarrow \infty$ and $n_1, n_2 \rightarrow \infty$ that

$$\frac{\tilde{T}_{\mathbf{n}} - \Delta}{\sqrt{\hat{Var}_{\boldsymbol{\theta}}(\tilde{T}_{\mathbf{n}})}} \Rightarrow N(0, 1).$$

Remark 3.1. Assume that $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_i (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = o(\text{tr}(\boldsymbol{\Sigma}_i^2)/n_i)$, $i = 1, 2$. Assume also (A-iv) and either (A-ii) or (A-iii) with (A-v). Then, it holds that

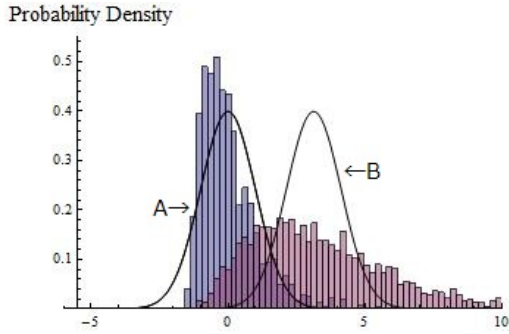
$$\frac{\tilde{T}_{\mathbf{n}} - \Delta}{\sqrt{\text{Var}_{\boldsymbol{\theta}}(\tilde{T}_{\mathbf{n}})}} \Rightarrow N(0, 1)$$

when $p \rightarrow \infty$ and $n_i \rightarrow \infty$ for $i = 1, 2$. Chen and Qin (2010) gave the asymptotic normality under different conditions.

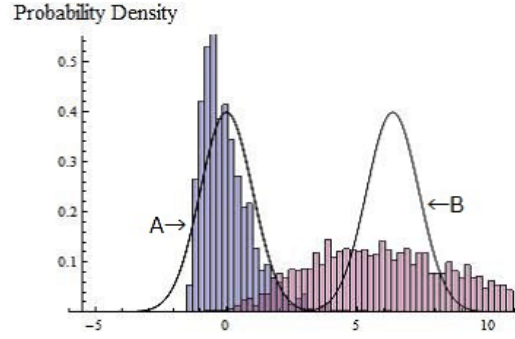
Remark 3.2. Chen and Qin (2010) gave a different estimator of $\text{Var}_{\boldsymbol{\theta}}(\tilde{T}_{\mathbf{n}})$ by $2 \sum_{i=1}^2 \text{tr}(\widehat{\boldsymbol{\Sigma}_i^2}) / (n_i(n_i - 1)) + 4 \text{tr}(\widehat{\boldsymbol{\Sigma}_1 \boldsymbol{\Sigma}_2}) / (n_1 n_2)$ with

$$\begin{aligned} \text{tr}(\widehat{\boldsymbol{\Sigma}_i^2}) &= (n_i(n_i - 1))^{-1} \text{tr} \left\{ \sum_{j \neq k}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_{in_i(j,k)}) \mathbf{x}_{ij}^T (\mathbf{x}_{ik} - \bar{\mathbf{x}}_{in_i(j,k)}) \mathbf{x}_{ik}^T \right\}, \\ \text{tr}(\widehat{\boldsymbol{\Sigma}_1 \boldsymbol{\Sigma}_2}) &= (n_1 n_2)^{-1} \text{tr} \left\{ \sum_{j=1}^{n_1} \sum_{k=1}^{n_2} (\mathbf{x}_{1j} - \bar{\mathbf{x}}_{1n_1(j)}) \mathbf{x}_{1j}^T (\mathbf{x}_{2k} - \bar{\mathbf{x}}_{2n_2(k)}) \mathbf{x}_{2k}^T \right\}. \end{aligned}$$

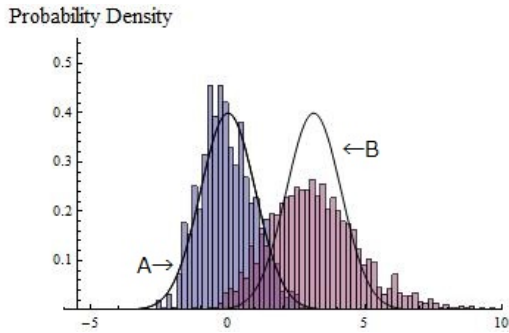
Here, $\bar{\mathbf{x}}_{in_i(j,k)}$ is the i -th sample mean after excluding \mathbf{x}_{ij} and \mathbf{x}_{ik} , and $\bar{\mathbf{x}}_{in_i(j)}$ is the i -th sample mean without \mathbf{x}_{ij} . Then, they claimed the asymptotic normality under several assumptions similar to Theorem 3.1. However, from the proof of Theorem 2 given in Chen and Qin (2010), it should be noted that $E_{\boldsymbol{\theta}}(\text{tr}(\widehat{\boldsymbol{\Sigma}_i^2})) = \text{tr}(\boldsymbol{\Sigma}_i^2) + \boldsymbol{\mu}_i^T \boldsymbol{\Sigma}_i \boldsymbol{\mu}_i / (n_i - 2)$. When $\|\boldsymbol{\mu}_i\|^2$ is large such as $\|\boldsymbol{\mu}_i\|^2 = O(p)$, the bias of $\text{tr}(\widehat{\boldsymbol{\Sigma}_i^2})$ becomes formidably large.



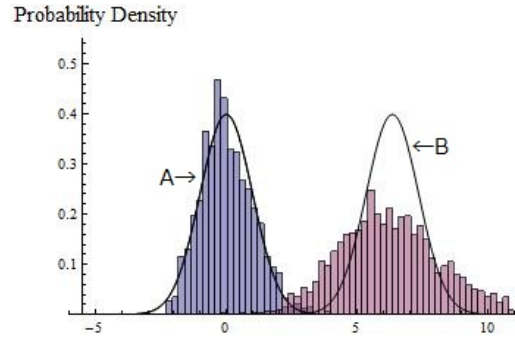
(a-i) $p = 4 : n_1 = n_2 = 10$



(a-ii) $p = 4 : n_1 = n_2 = 20$



(b-i) $p = 32 : n_1 = n_2 = 10$



(b-ii) $p = 32 : n_1 = n_2 = 20$

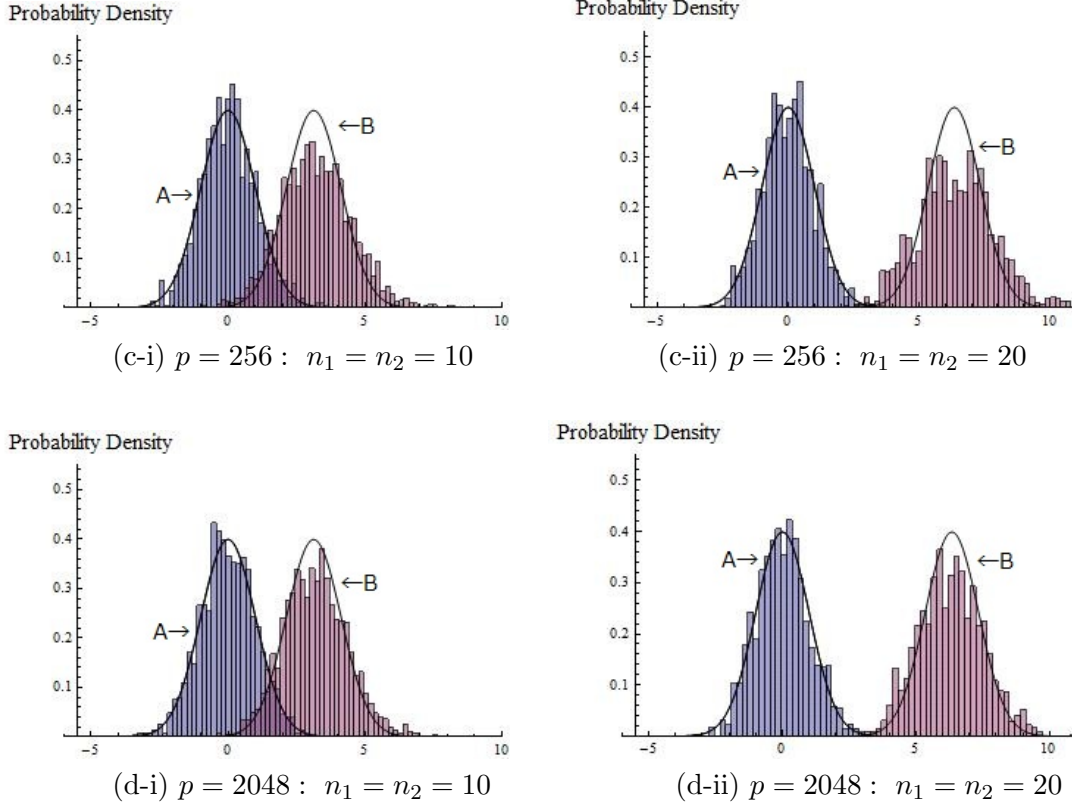


Figure 2. The solid lines are probability densities of A: $N(0, 1)$ and B: $N(\delta, 1)$. The histogram of $\tilde{T}_{\mathbf{n}}/\sqrt{\widehat{Var}_{\boldsymbol{\theta}}(\tilde{T}_{\mathbf{n}})}$ with $\|\boldsymbol{\mu}_2\|^2 = 0$ or $\|\boldsymbol{\mu}_2\|^2 = p^{1/2}$ fits well the solid lines with increasing dimension: $p = 4$ for (a-i) and (a-ii), $p = 32$ for (b-i) and (b-ii), $p = 256$ for (c-i) and (c-ii), and $p = 2048$ for (d-i) and (d-ii).

Let us observe Theorem 3.1. We considered an easy example such as $\|\boldsymbol{\mu}_1\|^2 = 0$, $\boldsymbol{\Sigma}_1 = \mathbf{I}_p$, $\|\boldsymbol{\mu}_2\|^2 = 0$ or $\|\boldsymbol{\mu}_2\|^2 = p^{1/2}$, and $\boldsymbol{\Sigma}_2 = 1.2\mathbf{I}_p$. We considered two cases for each dimension: $n_1 = n_2 = 10$ and $n_1 = n_2 = 20$. Here, \mathbf{x}_{ij} ($i = 1, 2$; $j = 1, \dots, n_i$) were generated from independent pseudorandom normal distribution with mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$ for $p = 4, 32, 256$ and 2048 . Fig. 2 gives the histograms of 2000 independent outcomes of $\tilde{T}_{\mathbf{n}}/\sqrt{\widehat{Var}_{\boldsymbol{\theta}}(\tilde{T}_{\mathbf{n}})}$ when $\|\boldsymbol{\mu}_2\|^2 = 0$ or $\|\boldsymbol{\mu}_2\|^2 = p^{1/2}$. Let $\delta = p^{1/2}/(\sum_{i=1}^2 2\text{tr}(\boldsymbol{\Sigma}_i^2)/(n_i(n_i-1)) + 4\text{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)/(n_1n_2))^{1/2}$. From Theorem 3.1, we expect that $\tilde{T}_{\mathbf{n}}/\sqrt{\widehat{Var}_{\boldsymbol{\theta}}(\tilde{T}_{\mathbf{n}})}$ is close to $N(0, 1)$ when $\|\boldsymbol{\mu}_2\|^2 = 0$, and $\tilde{T}_{\mathbf{n}}/\sqrt{\widehat{Var}_{\boldsymbol{\theta}}(\tilde{T}_{\mathbf{n}})}$ is close to $N(\delta, 1)$ when $\|\boldsymbol{\mu}_2\|^2 = p^{1/2}$. When $p = 4$ and $p = 32$, the histograms appear quite different from the probability densities specially when $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\| \neq 0$. However, as expected, the histograms fit well the probability densities as p increases. We can observe for each dimension that taking more samples makes more difference of those two hypotheses.

For testing the hypothesis (3.1), we find the sample size for each π_i as

$$n_i \geq \frac{(z_{\alpha} + z_{\beta})\sqrt{2}}{\Delta_L} \text{tr}(\boldsymbol{\Sigma}_i^2)^{1/4} \sum_{j=1}^2 \text{tr}(\boldsymbol{\Sigma}_j^2)^{1/4} \quad (= C_i, \text{ say}) \quad (3.4)$$

and test the hypothesis by

$$\text{rejecting } H_0 \iff \tilde{T}_{\mathbf{n}} > \frac{\Delta_L z_\alpha}{z_\alpha + z_\beta}, \quad (3.5)$$

where z_α is the upper α point of $N(0, 1)$. Then, we have the following theorem.

Theorem 3.2. *Assume (A-iv) and either (A-ii) or (A-iii) with (A-v). The test given by (3.4)-(3.5) has as $p \rightarrow \infty$ that*

$$\limsup \text{size} \leq \alpha \quad \text{and} \quad \liminf \text{power}(\Delta_L) \geq 1 - \beta, \quad (3.6)$$

where $\text{power}(\Delta_L)$ is the power when $\Delta = \Delta_L$.

Remark 3.3. Assume (A-iv) and either (A-ii) or (A-iii). If it holds that $\Delta_L/p^{1/2} > 0$ as $p \rightarrow \infty$, the test given by (3.5) has as $p \rightarrow \infty$ and $n_1, n_2 \rightarrow \infty$ that the size $\rightarrow 0$ and the power $\rightarrow 1$ when $\Delta \geq \Delta_L$.

3.2. TWO-STAGE PROCEDURE FOR TWO-SAMPLE TEST

Since Σ_i 's are unknown, it is necessary to estimate C_i 's in (3.4) with some pilot samples. We propose a two-stage test procedure to determine the sample sizes \mathbf{n} . We suppose the following assumption: There exists a known and positive lower bound $\sigma_{i\star}$ for $\sqrt{\text{tr}(\Sigma_i^2)}$ such that $\sigma_{i\star}/\sqrt{\text{tr}(\Sigma_i^2)} \in (0, 1)$, $i = 1, 2$, as $p \rightarrow \infty$. We proceed the following two steps:

1. Let $\tau_\star = \min_{i=1,2} \sqrt{\sigma_{i\star}} \sum_{j=1}^2 \sqrt{\sigma_{j\star}}$. Having a fixed integer m_0 (≥ 4), define

$$m = \max \left\{ m_0, \left\lceil \frac{(z_\alpha + z_\beta)\sqrt{2}}{\Delta_L} \tau_\star \right\rceil + 1 \right\}. \quad (3.7)$$

According to (3.7), take pilot samples \mathbf{x}_{ij} , $j = 1, \dots, m$, of size m from each π_i . Then, calculate $\mathbf{S}_{im(1)}$ and $\mathbf{S}_{im(2)}$ for each π_i according to (1.2). Define the total sample size for each π_i by

$$N_i = \max \left\{ m, \left\lceil \frac{(z_\alpha + z_\beta)\sqrt{2}}{\Delta_L} \text{tr}(\mathbf{S}_{im(1)} \mathbf{S}_{im(2)})^{1/4} \sum_{j=1}^2 \text{tr}(\mathbf{S}_{jm(1)} \mathbf{S}_{jm(2)})^{1/4} \right\rceil + 1 \right\}. \quad (3.8)$$

Let $\mathbf{N} = (N_1, N_2)$.

2. Take additional samples \mathbf{x}_{ij} , $j = m + 1, \dots, N_i$, of size $N_i - m$ from each π_i . By combining the initial samples and the additional samples, calculate $\tilde{T}_{\mathbf{N}}$ according to (3.2). Then, test the hypothesis (3.1) by

$$\text{rejecting } H_0 \iff \tilde{T}_{\mathbf{N}} > \frac{\Delta_L z_\alpha}{z_\alpha + z_\beta}. \quad (3.9)$$

We have the following theorems.

Theorem 3.3. *Assume (A-iv) and either (A-ii) or (A-iii) with (A-v). The test given by (3.9) with (3.7)-(3.8) has (3.6) as $p \rightarrow \infty$.*

Table 2. Required sample size and the size and power by (3.9) with (3.7)-(3.8).

C	\bar{n}	$\bar{n} - C$	$Var(n)$	$\bar{\alpha}$	$s(\bar{\alpha})$	$1 - \bar{\beta}$	$s(\bar{\beta})$
When $p = 400$: $m=14$							
53.57	53.88	0.31	27.30	0.062	0.00541	0.852	0.00794
24.08	24.26	0.18	6.45				
29.49	29.62	0.13	10.69				
When $p = 800$: $m=19$							
75.89	76.63	0.73	26.04	0.063	0.00545	0.875	0.00741
34.11	34.50	0.38	6.21				
41.78	42.13	0.35	10.63				
When $p = 1200$: $m=23$							
93.00	93.67	0.66	24.09	0.055	0.00510	0.865	0.00765
41.80	42.16	0.36	5.74				
51.20	51.50	0.30	9.54				
When $p = 1600$: $m=27$							
107.42	108.19	0.77	21.81	0.052	0.00499	0.882	0.00723
48.28	48.67	0.38	5.41				
59.14	59.52	0.38	8.40				
When $p = 2000$: $m=30$							
120.12	120.98	0.86	21.71	0.052	0.00499	0.888	0.00707
53.99	54.44	0.45	5.28				
66.13	66.54	0.41	8.51				

Theorem 3.4. Assume (A-i) and (A-iv). For the two-stage procedure given by (3.7)-(3.8), it holds as $p \rightarrow \infty$ that

$$\limsup |E_{\theta}(N_i - C_i)| \leq 1 \quad \text{and} \quad Var_{\theta}(N_i) = o(p^{1/2}/\Delta_L) \quad \text{for } i = 1, 2.$$

Remark 3.4. Assume (A-iv) and either (A-ii) or (A-iii). Then, it holds as $p \rightarrow \infty$ that $N_i/C_i = 1 + o_p(1)$ for $i = 1, 2$, that are in the HDLSS situation in the sense that $N_i/p = o_p(1)$, $i = 1, 2$.

3.3. SIMULATION FOR TWO-STAGE TEST

In order to study the performance of the two-stage test procedure given by (3.9) with (3.7)-(3.8), we took resort to computer simulations. We fixed $\Delta_L = 10$. Our goal was to construct a test with size $\alpha = 0.05$ and power no less than $1 - \beta = 0.9$ when $\Delta \geq \Delta_L$. Independent pseudorandom normal observations were generated from $\pi_i : N_p(\mu_i, \Sigma_i)$, $i = 1, 2$. We considered $\Sigma_1 = B(0.3^{|i-j|^{1/3}})B$ and $\Sigma_2 = 1.5 B(0.3^{|i-j|^{1/3}})B$, where B is defined by (2.9). From Remark 2.3, we set $\sigma_{i\star} = 0.8 \times \text{tr}(\Sigma_1)/p^{1/2}$, $i = 1, 2$. Then, we obtained $m=14, 19, 23, 27$ and 30 from (3.7) for $p = 400(400)2000$, respectively.

In Table 2, each block gives the findings when $p = 400(400)2000$. The findings were obtained by averaging the outcomes from 4000 ($= R$, say) replications, where the first 2000 replications were generated by setting as $\Delta = 0$ ($\mu_1 = \mu_2 = (0, \dots, 0)^T$) and the

last 2000 replications were generated by setting as $\Delta = 10$ ($\boldsymbol{\mu}_1 = (1, \dots, 1, 0, \dots, 0)^T$ whose first 10 elements are 1 and $\boldsymbol{\mu}_2 = (0, \dots, 0)^T$). Under a fixed scenario, suppose that the r th replication ends with $N_i = n_{ir}$ ($i = 1, 2$) observations given by (3.8) and the test result given by (3.9). Let $\bar{n}_i = R^{-1} \sum_{r=1}^R n_{ir}$ and $\text{Var}(n_i) = (R-1)^{-1} \sum_{r=1}^R (n_{ir} - \bar{n}_i)^2$. Then, \bar{n} ($= \bar{n}_1 + \bar{n}_2$) estimates $C = C_1 + C_2$, defined by (3.4), with its estimated variance $\text{Var}(\bar{n})$, computed analogously. In the end of the r th replication, we defined $P_r = 1$ (or 0) accordingly as $\Delta = 0$ was falsely rejected (or not) and $\Delta = 10$ was rightly rejected (or not). We defined $\bar{\alpha} = (R/2)^{-1} \sum_{r=1}^{(R/2)} P_r$ to estimate the size and $1 - \bar{\beta} = (R/2)^{-1} \sum_{r=R/2+1}^R P_r$ to estimate $\text{power}(\Delta_L)$, having their estimated standard errors $s(\bar{\alpha})$ and $s(\bar{\beta})$, where $s^2(\bar{\alpha}) = (R/2)^{-1} \bar{\alpha}(1 - \bar{\alpha})$ and $s^2(\bar{\beta}) = (R/2)^{-1} \bar{\beta}(1 - \bar{\beta})$. Throughout, we observed that the test given by (3.9) with (3.7)-(3.8) gave good performances especially in a very high-dimensional case.

3.4. TESTING THE EQUALITY OF TWO COVARIANCE MATRICES

We consider testing the equality of two covariance matrices as follows:

$$H_0 : \text{tr}(\boldsymbol{\Sigma}_1) = \text{tr}(\boldsymbol{\Sigma}_2) \quad \text{vs.} \quad H_1 : \text{tr}(\boldsymbol{\Sigma}_1) \neq \text{tr}(\boldsymbol{\Sigma}_2). \quad (3.10)$$

This type of equality test is essential for high-dimensional data. See Section 4 for example. We are interested in designing a test of (3.10) with size α and power no less than $1 - \beta$ when $|\text{tr}(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2)| \geq \Delta_{\boldsymbol{\Sigma}}$, where $\alpha, \beta \in (0, 1/2)$ and $\Delta_{\boldsymbol{\Sigma}} (> 0)$ are prespecified constants. We assume $\Delta_{\boldsymbol{\Sigma}} = o(p^{1/2})$. Under (A-i) and (A-iv), it holds that

$$\frac{\text{tr}(\mathbf{S}_{1n_1} - \mathbf{S}_{2n_2}) - \text{tr}(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2)}{\sqrt{2\text{tr}(\boldsymbol{\Sigma}_1^2)/(n_1 - 1) + 2\text{tr}(\boldsymbol{\Sigma}_2^2)/(n_2 - 1)}} \Rightarrow N(0, 1) \quad (3.11)$$

when $p \rightarrow \infty$ and either $n_i \rightarrow \infty$ or n_i is fixed for $i = 1, 2$. For testing the hypothesis (3.10), we find the sample size for each π_i as

$$n_i \geq \frac{2u(\alpha, \beta)^2}{\Delta_{\boldsymbol{\Sigma}}^2} \text{tr}(\boldsymbol{\Sigma}_i^2)^{1/2} \sum_{j=1}^2 \text{tr}(\boldsymbol{\Sigma}_j^2)^{1/2} \quad (= C_i, \text{ say}) \quad (3.12)$$

and test the hypothesis by

$$\text{rejecting } H_0 \iff |\text{tr}(\mathbf{S}_{1n_1} - \mathbf{S}_{2n_2})| > \frac{z_{\alpha/2} \Delta_{\boldsymbol{\Sigma}}}{u(\alpha, \beta)}, \quad (3.13)$$

where $u(\alpha, \beta) (> 0)$ is a solution of the equation $P(|N(0, 1) + u(\alpha, \beta)| > z_{\alpha/2}) = 1 - \beta$. Then, we have the following theorem.

Theorem 3.5. *Assume (A-i) and (A-iv). The test given by (3.12)-(3.13) has as $p \rightarrow \infty$ that*

$$\limsup \text{size} \leq \alpha \quad \text{and} \quad \liminf \text{power}(\Delta_{\boldsymbol{\Sigma}}) \geq 1 - \beta, \quad (3.14)$$

where $\text{power}(\Delta_{\boldsymbol{\Sigma}})$ is the power when $|\text{tr}(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2)| = \Delta_{\boldsymbol{\Sigma}}$.

Remark 3.5. Assume (A-iv) and either (A-ii) or (A-iii). Suppose that $z_{\alpha/2}/u(\alpha, \beta) \in$

$(0, 1)$. If it holds that $\Delta_{\Sigma}/p^{1/2} > 0$ as $p \rightarrow \infty$, the test given by (3.13) has as $p \rightarrow \infty$ and $n_1, n_2 \rightarrow \infty$ that the size $\rightarrow 0$ and the power $\rightarrow 1$ when $|\text{tr}(\Sigma_1 - \Sigma_2)| \geq \Delta_{\Sigma}$.

Since C_i depends on unknown Σ_i 's, we proceed the following two steps:

1. Let $\tau_{\star} = \min_{i=1,2} \sigma_{i\star} \sum_{j=1}^2 \sigma_{j\star}$, where $\sigma_{i\star}$'s are given in Section 3.2. Having a fixed integer m_0 (≥ 4), define

$$m = \max \left\{ m_0, \left\lceil \frac{2u(\alpha, \beta)^2}{\Delta_{\Sigma}^2} \tau_{\star} \right\rceil + 1 \right\}. \quad (3.15)$$

According to (3.15), take pilot samples \mathbf{x}_{ij} , $j = 1, \dots, m$, of size m from each π_i . Then, calculate $\mathbf{S}_{im(1)}$ and $\mathbf{S}_{im(2)}$ for each π_i according to (1.2). Define the total sample size for each π_i by

$$N_i = \max \left\{ m, \left\lceil \frac{2u(\alpha, \beta)^2}{\Delta_{\Sigma}^2} \text{tr}(\mathbf{S}_{im(1)} \mathbf{S}_{im(2)})^{1/2} \sum_{j=1}^2 \text{tr}(\mathbf{S}_{jm(1)} \mathbf{S}_{jm(2)})^{1/2} \right\rceil + 1 \right\}. \quad (3.16)$$

2. Take additional samples \mathbf{x}_{ij} , $j = m + 1, \dots, N_i$, of size $N_i - m$ from each π_i . By combining the initial samples and the additional samples, calculate \mathbf{S}_{iN_i} , $i = 1, 2$. Then, test the hypothesis (3.10) by

$$\text{rejecting } H_0 \iff |\text{tr}(\mathbf{S}_{1N_1} - \mathbf{S}_{2N_2})| > \frac{z_{\alpha/2} \Delta_{\Sigma}}{u(\alpha, \beta)}. \quad (3.17)$$

We have the following theorems.

Theorem 3.6. Assume (A-i) and (A-iv). The test given by (3.17) with (3.15)-(3.16) has (3.14) as $p \rightarrow \infty$.

Theorem 3.7. Assume (A-i) and (A-iv). For the two-stage procedure given by (3.15)-(3.16), it holds as $p \rightarrow \infty$ that

$$\limsup |E_{\theta}(N_i - C_i)| \leq 1 \quad \text{and} \quad \text{Var}_{\theta}(N_i) = o(p/\Delta_{\Sigma}^2) \quad \text{for } i = 1, 2.$$

Remark 3.6. Assume (A-iv) and either (A-ii) or (A-iii). Then, it holds as $p \rightarrow \infty$ that $N_i/C_i = 1 + o_p(1)$ for $i = 1, 2$, that are in the HDLSS situation in the sense that $N_i/p = o_p(1)$, $i = 1, 2$, under $\Delta_{\Sigma} \rightarrow \infty$ as $p \rightarrow \infty$.

4. HIGH-DIMENSIONAL CLASSIFICATION

Suppose we have two independently distributed populations, π_i , $i = 1, 2$. We do not assume $\Sigma_1 = \Sigma_2$. Let \mathbf{x}_0 be an observation vector on an individual belonging to π_1 or to π_2 . Having recorded $\mathbf{x}_{i1}, \dots, \mathbf{x}_{in_i}$ from each π_i , we estimate $\boldsymbol{\mu}_i$ and Σ_i by $\bar{\mathbf{x}}_{in_i}$ and \mathbf{S}_{in_i} . A typical discriminant rule is that one classifies \mathbf{x}_0 into π_1 if

$$(\mathbf{x}_0 - \bar{\mathbf{x}}_{1n_1})^T \mathbf{S}_{1n_1}^{-1} (\mathbf{x}_0 - \bar{\mathbf{x}}_{1n_1}) - \log \left\{ \frac{\det(\mathbf{S}_{2n_2})}{\det(\mathbf{S}_{1n_1})} \right\} < (\mathbf{x}_0 - \bar{\mathbf{x}}_{2n_2})^T \mathbf{S}_{2n_2}^{-1} (\mathbf{x}_0 - \bar{\mathbf{x}}_{2n_2}), \quad (4.1)$$

and into π_2 otherwise. However, the inverse matrix of \mathbf{S}_{in_i} does not exist in the HDLSS context ($p > n_i$). When $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$, Saranadasa (1993) considered using \mathbf{I}_p . Srivastava and Kubokawa (2007) considered using three types of the inverse covariance matrix: the Moore-Penrose inverse matrix; the inverse matrix defined by only diagonal elements of \mathbf{S}_{in_i} ; and the empirical Bayes inverse matrix estimator. On the other hand, Yata and Aoshima (2010c) considered using a ridge-type inverse covariance matrix derived by the *noise reduction methodology*. When $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$, Dudoit et al. (2002) considered the quadratic discriminant rule using the inverse matrix defined by only diagonal elements of \mathbf{S}_{in_i} . On the other hand, Hall et al. (2008) considered the distance-based classifiers and showed an asymptotic normality about $\|\mathbf{x}_0 - \mathbf{x}_{1j}\|^2 - \|\mathbf{x}_0 - \mathbf{x}_{2j'}\|^2$ ($j = 1, \dots, n_1$; $j' = 1, \dots, n_2$) by ignoring the covariance matrices.

We consider a discriminant rule given by replacing $\mathbf{S}_{in_i}^{-1}$ with $(\text{tr}(\mathbf{S}_{in_i})/p)^{-1}\mathbf{I}_p$ such that one classifies \mathbf{x}_0 into π_1 if

$$\frac{p\|\mathbf{x}_0 - \bar{\mathbf{x}}_{1n_1}\|^2}{\text{tr}(\mathbf{S}_{1n_1})} - \frac{p\|\mathbf{x}_0 - \bar{\mathbf{x}}_{2n_2}\|^2}{\text{tr}(\mathbf{S}_{2n_2})} - p \log \left\{ \frac{\text{tr}(\mathbf{S}_{2n_2})}{\text{tr}(\mathbf{S}_{1n_1})} \right\} - \frac{p}{n_1} + \frac{p}{n_2} + \gamma < 0 \quad (4.2)$$

and into π_2 otherwise. Here, $-p/n_1 + p/n_2$ is a bias-correction and γ is a tuning parameter. We denote the error rate of misclassifying an individual from π_1 (into π_2) or from π_2 (into π_1) by $e(2|1)$ or $e(1|2)$. Let $\Delta = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2$ and $\Delta_{\mathbf{\Sigma}_i} = (\text{tr}(\mathbf{\Sigma}_1) - \text{tr}(\mathbf{\Sigma}_2))^2 / \text{tr}(\mathbf{\Sigma}_i)$, $i = 1, 2$. Then, let us write that $\Delta_i = \Delta + \Delta_{\mathbf{\Sigma}_i}/2$, $i = 1, 2$, and $\Delta_\star = \min_{i=1,2} \Delta_i$. We are interested in designing the discriminant rule (4.2) having both $e(2|1) \leq \alpha$ and $e(1|2) \leq \beta$ when $\Delta_\star \geq \Delta_L$, where $\alpha, \beta \in (0, 1/2)$ and $\Delta_L (> 0)$ are prespecified constants. We assume $\Delta_L = o(p^{1/2})$.

4.1. ASYMPTOTIC NORMALITY AND SAMPLE SIZE DETERMINATION

We assume the followings:

$$\begin{aligned} \text{(A-vi)} \quad & \frac{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{\Sigma}_i (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{\Delta_\star^2} \rightarrow 0 \text{ as } p \rightarrow \infty \text{ for } i = 1, 2; \\ \text{(A-vii)} \quad & \frac{\text{tr}(\mathbf{\Sigma}_i^2)}{n_i^2 \Delta_\star^2} \rightarrow 0 \text{ and } \frac{\text{tr}(\mathbf{\Sigma}_i^2)}{n_i \Delta_\star^2} > 0 \text{ as } p \rightarrow \infty \text{ and } n_i \rightarrow \infty \text{ for } i = 1, 2. \end{aligned}$$

Then, we have the following theorem.

Theorem 4.1. *Assume that $\text{tr}(\mathbf{\Sigma}_1)/\text{tr}(\mathbf{\Sigma}_2) \rightarrow 1$ as $p \rightarrow \infty$. Assume also (A-iv), (A-vi), (A-vii) and either (A-ii) or (A-iii) with (A-v). Let*

$$\omega(\mathbf{x}_0) = \frac{p\|\mathbf{x}_0 - \bar{\mathbf{x}}_{1n_1}\|^2}{\text{tr}(\mathbf{S}_{1n_1})} - \frac{p\|\mathbf{x}_0 - \bar{\mathbf{x}}_{2n_2}\|^2}{\text{tr}(\mathbf{S}_{2n_2})} - p \log \left\{ \frac{\text{tr}(\mathbf{S}_{2n_2})}{\text{tr}(\mathbf{S}_{1n_1})} \right\} - \frac{p}{n_1} + \frac{p}{n_2}.$$

Then, we have as $p \rightarrow \infty$ and $n_1, n_2 \rightarrow \infty$ that

$$\begin{aligned} \frac{\omega(\mathbf{x}_0) + \Delta_2(\text{tr}(\mathbf{\Sigma}_2)/p)^{-1}}{2\sqrt{(\text{tr}(\mathbf{\Sigma}_1)/p)^{-2}\text{tr}(\mathbf{\Sigma}_1^2)/n_1 + (\text{tr}(\mathbf{\Sigma}_2)/p)^{-2}\text{tr}(\mathbf{\Sigma}_1\mathbf{\Sigma}_2)/n_2}} &\Rightarrow N(0, 1) \quad \text{when } \mathbf{x}_0 \in \pi_1; \\ \frac{\omega(\mathbf{x}_0) - \Delta_1(\text{tr}(\mathbf{\Sigma}_1)/p)^{-1}}{2\sqrt{(\text{tr}(\mathbf{\Sigma}_2)/p)^{-2}\text{tr}(\mathbf{\Sigma}_2^2)/n_2 + (\text{tr}(\mathbf{\Sigma}_1)/p)^{-2}\text{tr}(\mathbf{\Sigma}_1\mathbf{\Sigma}_2)/n_1}} &\Rightarrow N(0, 1) \quad \text{when } \mathbf{x}_0 \in \pi_2. \end{aligned}$$

Remark 4.1. Assume (A-iv) and either (A-ii) or (A-iii). If it holds $p/\Delta_\star^2 \rightarrow 0$ as $p \rightarrow \infty$, there exist positive constants c_1 and c_2 such that

$$\frac{\omega(\mathbf{x}_0)}{\Delta_\star} < -c_1 \quad \text{when } \mathbf{x}_0 \in \pi_1; \quad \frac{\omega(\mathbf{x}_0)}{\Delta_\star} > c_2 \quad \text{when } \mathbf{x}_0 \in \pi_2.$$

Then, for the discriminant rule given by (4.2) with $\gamma = 0$, we have as $p \rightarrow \infty$ that

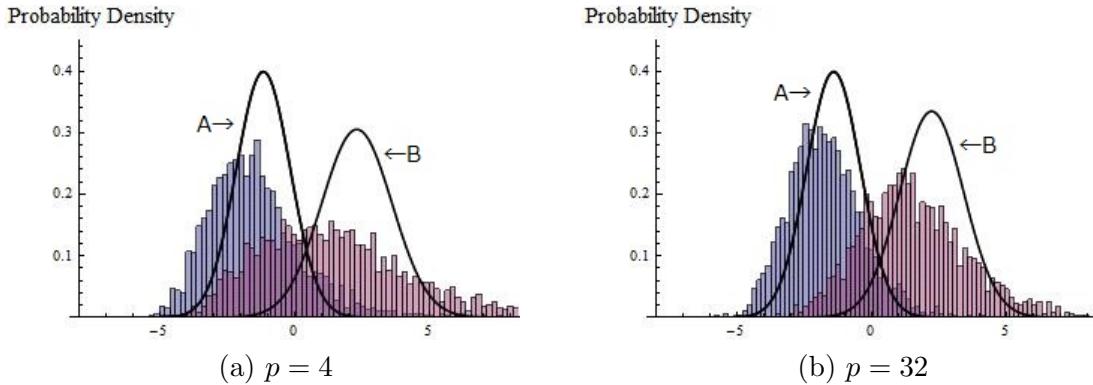
$$e(2|1) \rightarrow 0 \quad \text{and} \quad e(1|2) \rightarrow 0. \quad (4.3)$$

For instance, let us consider a case that $\text{tr}(\Sigma_1)/\text{tr}(\Sigma_2) \neq 1$ as $p \rightarrow \infty$. (For testing the equality, see Section 3.4). Then, it follows that $\min_{i=1,2} \Delta_{\Sigma_i}/p > 0$ as $p \rightarrow \infty$. Since it holds $p/\Delta_\star^2 \rightarrow 0$ as $p \rightarrow \infty$, we can claim (4.3) in the case.

Let us observe Theorem 4.1. Now, we consider a non-Gaussian example such as $z_{ijl} = (13/15)^{1/2} w_{ijl}$, where w_{ijl} , $i = 1, 2$; $j = 1, \dots, p$ ($l = 1, \dots, n_i$) were independently generated by t -distribution with 15 degrees of freedom. Then, note that $E(z_{ijl}) = 0$, $E(z_{ijl}^2) = 1$ and z_{ijl} , $j = 1, \dots, p$ ($i = 1, 2$; $l = 1, \dots, n_i$) are independent. Let $\mathbf{x}_{il} = c_i^{1/2}(z_{i1l}, \dots, z_{ipl})^T + \boldsymbol{\mu}_i$ ($i = 1, 2$; $l = 1, \dots, n_i$) with $(c_1, c_2) = (1, 1 + p^{-1/4})$, $\boldsymbol{\mu}_1 = \mathbf{0}$ and $\|\boldsymbol{\mu}_2\|^2 = p^{1/2}$, so that $\Sigma_i = c_i \mathbf{I}_p$, $i = 1, 2$. Then, the population distributions of \mathbf{x}_{il} , $i = 1, 2$, satisfy (A-ii) and (A-iv). Since it holds that $\Delta = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2 = p^{1/2}$, $\Delta_{\Sigma_1} = \text{tr}(\Sigma_1 - \Sigma_2)^2/\text{tr}(\Sigma_1) = p^{1/2}$ and $\Delta_{\Sigma_2} = p^{1/2}/(1 + p^{-1/4})$, we have that $\Delta_1 = 3p^{1/2}/2$ and $\Delta_2 = p^{1/2}(1 + 1/(2 + 2/p^{1/4}))$. Let

$$\begin{aligned} \delta_1 &= 2\sqrt{(\text{tr}(\Sigma_1)/p)^{-2}\text{tr}(\Sigma_1^2)/n_1 + (\text{tr}(\Sigma_2)/p)^{-2}\text{tr}(\Sigma_1\Sigma_2)/n_2}, \\ \delta_2 &= 2\sqrt{(\text{tr}(\Sigma_2)/p)^{-2}\text{tr}(\Sigma_2^2)/n_2 + (\text{tr}(\Sigma_1)/p)^{-2}\text{tr}(\Sigma_1\Sigma_2)/n_1}. \end{aligned}$$

Figs. 3(a), (b), (c) and (d) give two histograms of 2000 independent outcomes of $\omega(\mathbf{x}_0)/\delta_1$ when $\mathbf{x}_0 \in \pi_1$ or $\mathbf{x}_0 \in \pi_2$ for $p = 4, 32, 256$ and 2048 , respectively. Here, $\omega(\mathbf{x}_0)$ was calculated from $n_1 = n_2 = 15$ samples for each π_i . Fig. 3 also displays the probability density of $\omega(\mathbf{x}_0)/\delta_1$ claimed by Theorem 4.1. We expect that $\omega(\mathbf{x}_0)/\delta_1$ is close to $N(-\Delta_2/(\delta_1\text{tr}(\Sigma_2)/p), 1)$ when $\mathbf{x}_0 \in \pi_1$ and $\omega(\mathbf{x}_0)/\delta_1$ is close to $N(\Delta_1/(\delta_1\text{tr}(\Sigma_1)/p), \delta_2^2/\delta_1^2)$ when $\mathbf{x}_0 \in \pi_2$. When $p = 4$ and $p = 32$, the histograms appear quite different from the probability densities. However, as expected, the histograms become similar to the probability densities as p increases.



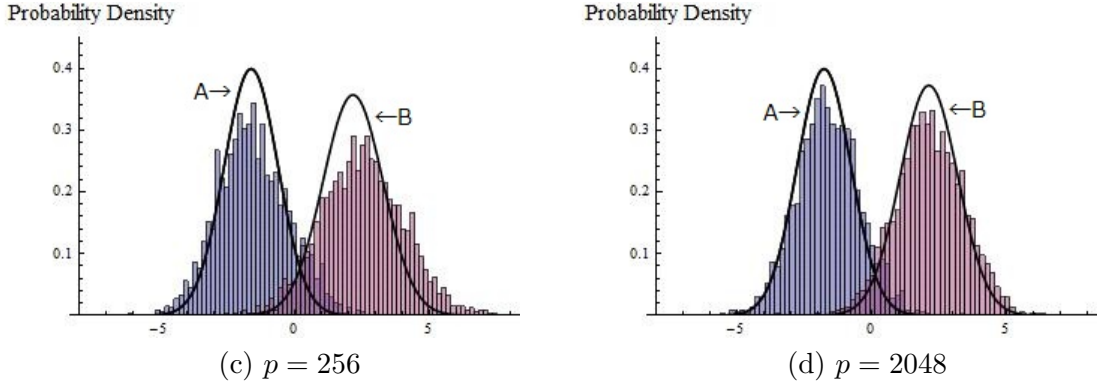


Figure 3. The solid lines are probability densities of A: $N(-\Delta_2(\delta_1 \text{tr}(\Sigma_2)/p)^{-1}, 1)$ and B: $N(\Delta_1(\delta_1 \text{tr}(\Sigma_1)/p)^{-1}, \delta_2^2/\delta_1^2)$. The histograms of $\omega(\mathbf{x}_0)/\delta_1$ for $\mathbf{x}_0 \in \pi_i$, $i = 1, 2$, fit well the solid lines with increasing dimension: (a) $p = 4$, (b) $p = 32$, (c) $p = 256$, and (d) $p = 2048$.

Let $\sigma = \max\{\text{tr}(\Sigma_1^2)^{1/2}, \text{tr}(\Sigma_2^2)^{1/2}\}$. We find the sample size for each π_i as

$$n_i \geq \frac{(z_\alpha + z_\beta)^2 \sigma}{\Delta_L^2} \text{tr}(\Sigma_i^2)^{1/4} \sum_{j=1}^2 \text{tr}(\Sigma_j^2)^{1/4} \quad (= C_i, \text{ say}). \quad (4.4)$$

Note that $C_i = O(p/\Delta_L^2)$ for $i = 1, 2$, under (A-iv). Thus under $\Delta_L \rightarrow \infty$ as $p \rightarrow \infty$, it holds that $C_i/p \rightarrow 0$ as $p \rightarrow \infty$. Then, we have the following theorem.

Theorem 4.2. Assume (A-iv), (A-vi) and either (A-ii) or (A-iii) with (A-v). Let $\gamma = (\text{tr}(\mathbf{S}_{1n_1} + \mathbf{S}_{2n_2})/(2p))^{-1} \Delta_L (z_\beta - z_\alpha)/(z_\alpha + z_\beta)$ in (4.2). Then, for the discriminant rule given by (4.2) with (4.4), it holds as $p \rightarrow \infty$ that

$$\limsup e(2|1) \leq \alpha \quad \text{and} \quad \limsup e(1|2) \leq \beta$$

when $\Delta_\star \geq \Delta_L$.

4.2. TWO-STAGE PROCEDURE FOR CLASSIFICATION

Since Σ_i 's are unknown, it is necessary to estimate C_i 's in (4.4) with some pilot samples. We suppose the following assumption: There exists a known and positive lower bound $\sigma_{i\star}$ for $\sqrt{\text{tr}(\Sigma_i^2)}$ such that $\sigma_{i\star}/\sqrt{\text{tr}(\Sigma_i^2)} \in (0, 1)$, $i = 1, 2$, as $p \rightarrow \infty$. We proceed the following two steps:

1. Let $\tau_\star = \min_{i=1,2} \sigma_{i\star} \sum_{j=1}^2 \sigma_{j\star}$. Having a fixed integer $m_0 (\geq 4)$, define

$$m = \max \left\{ m_0, \left\lceil \frac{(z_\alpha + z_\beta)^2}{\Delta_L^2} \tau_\star \right\rceil + 1 \right\}. \quad (4.5)$$

According to (4.5), take pilot samples \mathbf{x}_{ij} , $j = 1, \dots, m$, of size m from each π_i . Then, calculate \mathbf{S}_{im} , $\mathbf{S}_{im(1)}$ and $\mathbf{S}_{im(2)}$ for each π_i according to (1.2). Let $\hat{\sigma} = \max\{\text{tr}(\mathbf{S}_{1m(1)}\mathbf{S}_{1m(2)})^{1/2}, \text{tr}(\mathbf{S}_{2m(1)}\mathbf{S}_{2m(2)})^{1/2}\}$. Define the total sample size for each π_i by

$$N_i = \max \left\{ m, \left\lceil \frac{(z_\alpha + z_\beta)^2 \hat{\sigma}}{\Delta_L^2} \text{tr}(\mathbf{S}_{im(1)}\mathbf{S}_{im(2)})^{1/4} \sum_{j=1}^2 \text{tr}(\mathbf{S}_{jm(1)}\mathbf{S}_{jm(2)})^{1/4} \right\rceil + 1 \right\}. \quad (4.6)$$

2. Take additional samples \mathbf{x}_{ij} , $j = m + 1, \dots, N_i$, of size $N_i - m$ from each π_i . By combining the initial samples and the additional samples, calculate $\bar{\mathbf{x}}_{iN_i}$, \mathbf{S}_{iN_i} , $i = 1, 2$. Then, we classify \mathbf{x}_0 into π_1 if

$$\frac{p\|\mathbf{x}_0 - \bar{\mathbf{x}}_{1N_1}\|^2}{\text{tr}(\mathbf{S}_{1N_1})} - \frac{p\|\mathbf{x}_0 - \bar{\mathbf{x}}_{2N_2}\|^2}{\text{tr}(\mathbf{S}_{2N_2})} - p \log \left\{ \frac{\text{tr}(\mathbf{S}_{2N_2})}{\text{tr}(\mathbf{S}_{1N_1})} \right\} - \frac{p}{N_1} + \frac{p}{N_2} + \hat{\gamma} < 0 \quad (4.7)$$

and into π_2 otherwise, where $\hat{\gamma} = (\text{tr}(\mathbf{S}_{1N_1} + \mathbf{S}_{2N_2})/(2p))^{-1} \Delta_L (z_\beta - z_\alpha)/(z_\alpha + z_\beta)$.

Then, we have the following theorems.

Theorem 4.3. Assume (A-iv), (A-vi) and either (A-ii) or (A-iii) with (A-v). Then, for the discriminant rule given by (4.7) with (4.5)-(4.6), it holds as $p \rightarrow \infty$ that

$$\limsup e(2|1) \leq \alpha \quad \text{and} \quad \limsup e(1|2) \leq \beta$$

when $\Delta_\star \geq \Delta_L$.

Theorem 4.4. Assume (A-i), (A-iv) and that $\text{tr}(\Sigma_1^2)/\text{tr}(\Sigma_2^2) \neq 1$ as $p \rightarrow \infty$. For the two-stage procedure given by (4.5)-(4.6), it holds as $p \rightarrow \infty$ that

$$\limsup |E_\theta(N_i - C_i)| \leq 1 \quad \text{and} \quad \text{Var}_\theta(N_i) = o(p/\Delta_L^2) \quad \text{for } i = 1, 2.$$

Remark 4.2. Assume (A-iv) and either (A-ii) or (A-iii). Then, it holds as $p \rightarrow \infty$ that $N_i/C_i = 1 + o_p(1)$ for $i = 1, 2$, that are in the HDLSS situation in the sense that $N_i/p = o_p(1)$, $i = 1, 2$, under $\Delta_L \rightarrow \infty$ as $p \rightarrow \infty$.

4.3. SIMULATION

In order to study the performance of the discriminant rule given by (4.7) with (4.5)-(4.6), we took resort to computer simulations. We set $\alpha = 0.1$, $\beta = 0.2$ and $\Delta_L = 20$. Then, we obtained $z_\alpha + z_\beta = 2.12$. Independent pseudorandom observations were generated from $\pi_i : N_p(\boldsymbol{\mu}_i, \Sigma_i)$, $i = 1, 2$. We set $\boldsymbol{\mu}_1 = (1, \dots, 1, 0, \dots, 0)^T$ whose first 25 elements are 1, and $\boldsymbol{\mu}_2 = (0, \dots, 0)^T$. Then, we obtained $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2 = \Delta = 25$. We considered $\Sigma_1 = c_1 \mathbf{B}(0.3^{|i-j|^{1/3}}) \mathbf{B}$ and $\Sigma_2 = c_2 \mathbf{B}(0.3^{|i-j|^{1/3}}) \mathbf{B}$, where \mathbf{B} is defined by (2.9). We considered the following four cases: (i) $(c_1, c_2) = (1, 1)$ when $p = 800$; (ii) $(c_1, c_2) = (0.95, 1.05)$ when $p = 800$; (iii) $(c_1, c_2) = (1, 1)$ when $p = 1600$; (iv) $(c_1, c_2) = (0.95, 1.05)$ when $p = 1600$. Then, we obtained $\Delta_\star = \Delta + \min_{i=1,2} \Delta_{\Sigma_i}/2 = 25, 28.81, 25$ and 32.62 for (i), (ii), (iii) and (iv). From Remark 2.3, we set $\sigma_{i\star} = 0.9 \times \text{tr}(\Sigma_i)/p^{1/2}$, $i = 1, 2$. Then, from (4.5), we obtained $m = 15, 14, 30$ and 28 for (i), (ii), (iii) and (iv).

In Table 3, the findings obtained by averaging the outcomes from 4000 (= R , say) replications were summarized in each situation, where $\mathbf{x}_0 \in \pi_1$ is taken for the first 2000 replications and $\mathbf{x}_0 \in \pi_2$ is taken for the last 2000 replications. Under a fixed scenario, suppose that the r th replication ends with $N_i = n_{ir}$ ($i = 1, 2$) observations for $r = 1, \dots, R$. Let $\bar{n}_i = R^{-1} \sum_{r=1}^R n_{ir}$ and $\text{Var}(n_i) = (R-1)^{-1} \sum_{r=1}^R (n_{ir} - \bar{n}_i)^2$. Then, \bar{n} ($= \bar{n}_1 + \bar{n}_2$) estimates $C = C_1 + C_2$ with its estimated variance $\text{Var}(n)$, computed analogously. In the end of the r th replication, we checked whether the rule (4.7) does (or does not) classify \mathbf{x}_0 correctly and defined $P_r = 1$ (or 0) accordingly. We calculated $1 - \overline{e(2|1)} = (R/2)^{-1} \sum_{r=1}^{R/2} P_r$ and $1 - \overline{e(1|2)} = (R/2)^{-1} \sum_{r=R/2+1}^R P_r$ for the estimates of $1 - e(2|1)$ and $1 - e(1|2)$.

Table 3. Discriminant rule (4.7) with (4.5)-(4.6).

C	\bar{n}	$\bar{n} - C$	$Var(n)$	$1 - \overline{e(2 1)}$	$s(\overline{e(2 1)})$	$1 - \overline{e(1 2)}$	$s(\overline{e(1 2)})$
When $p = 800$ and $(c_1, c_2) = (1, 1)$: $m = 15$							
61.88	66.23	4.35	137.20	0.886	0.00711	0.809	0.00880
30.94	33.17	2.23	36.51				
30.94	33.06	2.13	36.45				
When $p = 800$ and $(c_1, c_2) = (0.95, 1.05)$: $m = 14$							
64.93	66.88	1.95	164.41	0.927	0.00584	0.852	0.00795
31.65	32.57	0.92	36.87				
33.28	34.31	1.04	50.68				
When $p = 1600$ and $(c_1, c_2) = (1, 1)$: $m = 30$							
123.97	128.65	4.68	105.62	0.901	0.00669	0.837	0.00827
61.99	64.35	2.37	28.34				
61.99	64.30	2.31	27.89				
When $p = 1600$ and $(c_1, c_2) = (0.95, 1.05)$: $m = 28$							
130.09	131.51	1.42	140.00	0.949	0.00490	0.893	0.00693
63.42	64.11	0.69	29.45				
66.67	67.40	0.73	44.77				

Their estimated standard errors were given by $s(\overline{e(2|1)})$ and $s(\overline{e(1|2)})$, where $s^2(\overline{e(2|1)}) = (R/2)^{-1} \overline{e(2|1)}(1 - \overline{e(2|1)})$ and $s^2(\overline{e(1|2)}) = (R/2)^{-1} \overline{e(1|2)}(1 - \overline{e(1|2)})$. Throughout, the discriminant rule given by (4.7) with (4.5)-(4.6) gave adequate performances specially when $\text{tr}(\mathbf{\Sigma}_1) \neq \text{tr}(\mathbf{\Sigma}_2)$. This result is quite natural because Δ_\star for $\text{tr}(\mathbf{\Sigma}_1) \neq \text{tr}(\mathbf{\Sigma}_2)$ is greater than that for $\text{tr}(\mathbf{\Sigma}_1) = \text{tr}(\mathbf{\Sigma}_2)$.

4.4. EXAMPLE

We analyzed gene expression data given by Chiaretti et al. (2004) in which data set consisted of 12625 ($= p$) genes and 128 samples. Note that the expression measures were obtained using the three-step robust multichip average (RMA) preprocessing method. Refer to Pollard et al. (2005) as well for the details. The data set had two tumor cellular subtypes, B-cell (95 samples) and T-cell (33 samples). We set π_1 : B-cell and π_2 : T-cell. We set $\alpha = 0.05$, $\beta = 0.05$ and $\Delta_L = 800$. Here, we emphasize that one can make Δ_L in the two sample test of mean vectors and covariance matrices in Section 3. Our goal was to construct a discriminant rule with $e(2|1) \leq 0.05$ and $e(1|2) \leq 0.05$ when $\Delta_\star \geq \Delta_L$. We assume that $\text{tr}(\mathbf{\Sigma}_1^2)^{1/2} > 700$ for B-cell and $\text{tr}(\mathbf{\Sigma}_2^2)^{1/2} > 550$ for T-cell. We set $\sigma_{1\star} = 700$ and $\sigma_{2\star} = 550$ so that $\tau_\star = \min_{i=1,2} \sigma_{i\star}(\sigma_{1\star} + \sigma_{2\star}) = 6.88 \times 10^5$. We chose the pilot sample size for each π_i as

$$m = \left\{ 4, \left\lceil \frac{(z_\alpha + z_\beta)^2 \tau_\star}{\Delta_L^2} \right\rceil + 1 \right\} = 12$$

according to (4.5), where $z_\alpha = z_\beta = 1.64$. So, we took the first 12 samples from each π_i as a pilot sample. Then, we had $\text{tr}(\mathbf{S}_{1m(1)}\mathbf{S}_{1m(2)})^{1/2} = 718$, $\text{tr}(\mathbf{S}_{2m(1)}\mathbf{S}_{2m(2)})^{1/2} = 571$ and $\hat{\sigma} = \max\{\text{tr}(\mathbf{S}_{1m(1)}\mathbf{S}_{1m(2)})^{1/2}, \text{tr}(\mathbf{S}_{2m(1)}\mathbf{S}_{2m(2)})^{1/2}\} = 718$. According to (4.6), the total

sample size for each π_i was given by

$$N_1 = \max \left\{ 12, \left\lceil \frac{(z_\alpha + z_\beta)^2 \hat{\sigma}}{\Delta_L^2} \text{tr}(\mathbf{S}_{1m(1)} \mathbf{S}_{1m(2)})^{1/4} \sum_{i=1}^2 \text{tr}(\mathbf{S}_{im(1)} \mathbf{S}_{im(2)})^{1/4} \right\rceil + 1 \right\} = 17,$$

$$N_2 = \max \left\{ 12, \left\lceil \frac{(z_\alpha + z_\beta)^2 \hat{\sigma}}{\Delta_L^2} \text{tr}(\mathbf{S}_{2m(1)} \mathbf{S}_{2m(2)})^{1/4} \sum_{i=1}^2 \text{tr}(\mathbf{S}_{im(1)} \mathbf{S}_{im(2)})^{1/4} \right\rceil + 1 \right\} = 15.$$

So, we took the next 5 samples from π_1 and the next 3 samples from π_2 . Note that $\hat{\gamma} = 0$ for $\alpha = \beta$. Then, we constructed the discriminant rule (4.7) with $e(2|1) \leq 0.05$ and $e(1|2) \leq 0.05$ when $\Delta_\star \geq \Delta_L$. Note that an estimator of Δ_\star was given by $\tilde{T}_N + \min_{i=1,2} \text{tr}(\mathbf{S}_{1N_1} - \mathbf{S}_{2N_2})^2 / (2\text{tr}(\mathbf{S}_{iN_i})) = 1564$, where \tilde{T}_N is defined in Section 3.

We compared the constructed discriminant rule with two different discriminant rules, DLDR and DQDR, that were given by Dudoit et al. (2002) as follows: Diagonal linear discriminant rule (DLDR) was given by replacing (4.7) with

$$(\mathbf{x}_0 - (\bar{\mathbf{x}}_{1N_1} + \bar{\mathbf{x}}_{2N_2})/2)^T \mathbf{S}_{diag}^{-1} (\bar{\mathbf{x}}_{2N_2} - \bar{\mathbf{x}}_{1N_1}) < 0$$

with $\mathbf{S}_{diag} = \text{diag}(s_{1N}, \dots, s_{pN})$, where $s_{jN} = \sum_{i=1}^2 \sum_{l=1}^{N_i} (x_{ijl} - \bar{x}_{ijN_i})^2 / (N_1 + N_2 - 2)$ and $\bar{x}_{ijN_i} = \sum_{l=1}^{N_i} x_{ijl} / N_i$. Diagonal quadratic discriminant rule (DQDR) was given by replacing (4.7) with

$$(\mathbf{x}_0 - \bar{\mathbf{x}}_{1N_1})^T \mathbf{S}_{diag(1)}^{-1} (\mathbf{x}_0 - \bar{\mathbf{x}}_{1N_1}) - (\mathbf{x}_0 - \bar{\mathbf{x}}_{2N_2})^T \mathbf{S}_{diag(2)}^{-1} (\mathbf{x}_0 - \bar{\mathbf{x}}_{2N_2}) - \log \left\{ \frac{\det(\mathbf{S}_{diag(2)})}{\det(\mathbf{S}_{diag(1)})} \right\} < 0$$

with $\mathbf{S}_{diag(i)} = \text{diag}(s_{(i)1N_i}, \dots, s_{(i)pN_i})$, where $s_{(i)jN_i} = \sum_{l=1}^{N_i} (x_{ijl} - \bar{x}_{ijN_i})^2 / (N_i - 1)$. In Table 4, we investigated the performance of the three discriminant rules with $(N_1, N_2) = (17, 15)$ by using test data sets of $95 - N_1 = 78$ surplus samples from π_1 and $33 - N_2 = 18$ surplus samples from π_2 . The discriminant rule given by (4.7) showed an adequate performance and was best among the three rules.

Table 4. The correct discrimination rates of (4.7), DLDR and DQDR to test data sets of 78 samples from π_1 and 18 samples from π_2 .

	(4.7)	DLDR	DQDR
$1-e(2 1)$	74/78 (=0.949)	64/78 (=0.821)	67/78 (=0.859)
$1-e(1 2)$	18/18 (=1.0)	18/18 (=1.0)	18/18 (=1.0)

5. HIGH-DIMENSIONAL VARIABLE SELECTION

Suppose we have two independently distributed populations, π_i , $i = 1, 2$. We do not assume $\Sigma_1 = \Sigma_2$. We consider a methodology to select a significant set of associated variables from among high-dimensional data sets. We recall that $\boldsymbol{\mu}_i = (\mu_{i1}, \dots, \mu_{ip})^T$, $i = 1, 2$. Then, we consider testing the following univariate hypotheses:

$$H_{0j} : \mu_{1j} = \mu_{2j} \quad \text{vs.} \quad H_{1j} : \mu_{1j} \neq \mu_{2j} \quad \text{for } j = 1, \dots, p. \quad (5.1)$$

Our interest is to select a set of significant variables such that $\mathbf{D} = \{j : \mu_{1j} \neq \mu_{2j}\}$. Fan and Fan (2008), Meinshausen et al. (2009) and Wasserman and Roeder (2009) considered this

type of problem. Assume that $|\mathbf{D}| = S$ for some $S \geq 1$, where $|\mathbf{D}|$ denotes the number of elements in set \mathbf{D} . A variable selection procedure $\widehat{\mathbf{D}}$ maps the data into subsets of $\{1, \dots, p\}$. In this section, we are interested in designing $\widehat{\mathbf{D}}$ such that the asymptotic *family-wise error rate* (FWER) is 0, i.e.,

$$P_{\theta}(|\mathbf{D}^c \cap \widehat{\mathbf{D}}| \neq 0) \rightarrow 0, \quad (5.2)$$

and the asymptotic *average power* (AP) is 1, i.e.,

$$\frac{|\mathbf{D} \cap \widehat{\mathbf{D}}|}{S} \rightarrow 1 \quad \text{when} \quad \min_{j \in \mathbf{D}} |\mu_{1j} - \mu_{2j}|^2 > \delta, \quad (5.3)$$

where $\delta (> 0)$ is a prespecified constant. We should note that the assertion (5.3) does not consider the case when $\min_{j \in \mathbf{D}} |\mu_{1j} - \mu_{2j}|^2 = \delta$.

5.1. SAMPLE SIZE DETERMINATION

Let $\sigma_i = \max_{1 \leq j \leq p} \sigma_{(i)j}$ ($i = 1, 2$), where we recall that $\sigma_{(i)j}$, $j = 1, \dots, p$, are diagonal elements of Σ_i . We assume that $\sigma_{(i)j} < \infty$ for $i = 1, 2$; $j \in \mathbf{D}$, and $E_{\theta}\{\exp(t|x_{ijl} - \mu_{ij}|/\sigma_{(i)j}^{1/2})\} < \infty$, $i = 1, 2$; $j = 1, \dots, p$, for some $t > 0$. We do not assume $\sigma_i < \infty$ as $p \rightarrow \infty$ for $i = 1, 2$. Let $T_{j(\mathbf{n})} = \bar{x}_{1jn_1} - \bar{x}_{2jn_2}$ with $\bar{x}_{ijn_i} = \sum_{l=1}^{n_i} x_{ijl}/n_i$, $i = 1, 2$. Then, for testing the hypotheses (5.1), we find the sample size for each π_i as

$$n_i \geq \frac{2(\log p)^{1+\zeta}}{\delta} \sqrt{\sigma_i} \sum_{j=1}^2 \sqrt{\sigma_j} \quad (= C_i, \text{ say}) \quad (5.4)$$

with $\zeta \in (0, 1)$ chosen, and test the hypothesis for $j = 1, \dots, p$, by

$$\text{rejecting } H_{0j} \iff |T_{j(\mathbf{n})}| > \sqrt{\delta}. \quad (5.5)$$

Let $\widehat{\mathbf{D}} = \{j \mid \text{rejecting } H_{0j}\}$. Then, we have the following theorem.

Theorem 5.1. *The test given by (5.5) with (5.4) has as $p \rightarrow \infty$ that*

$$\begin{aligned} P_{\theta}(|\mathbf{D}^c \cap \widehat{\mathbf{D}}| \neq 0) &= o(1); \\ \frac{|\mathbf{D} \cap \widehat{\mathbf{D}}|}{S} &= 1 + o_p(1) \quad \text{when} \quad \min_{j \in \mathbf{D}} |\mu_{1j} - \mu_{2j}|^2 > \delta. \end{aligned} \quad (5.6)$$

Remark 5.1. Assume (A-i). We choose $\zeta = 0$ in (5.4). Then, the test given by (5.5) with (5.4) has (5.6) as $p \rightarrow \infty$.

Remark 5.2. We consider a test having the asymptotic power

$$P_{\theta}(\mathbf{D} \subseteq \widehat{\mathbf{D}}) \rightarrow 1 \quad \text{when} \quad \min_{j \in \mathbf{D}} |\mu_{1j} - \mu_{2j}|^2 \geq \delta$$

instead of (5.3). We define the sample size for each π_i as

$$n_i \geq \frac{8(\log p)^{1+\zeta}}{\delta} \sqrt{\sigma_i} \sum_{j=1}^2 \sqrt{\sigma_j} \quad (5.7)$$

with $\zeta \in (0, 1)$ chosen, and test the hypothesis for $j = 1, \dots, p$, by

$$\text{rejecting } H_{0j} \iff |T_{j(\mathbf{n})}| > \sqrt{\delta}/2.$$

Let $\widehat{\mathbf{D}} = \{j \mid \text{rejecting } H_{0j}\}$. Then, it holds as $p \rightarrow \infty$ that

$$\begin{aligned} P_{\boldsymbol{\theta}}(|\mathbf{D}^c \cap \widehat{\mathbf{D}}| \neq 0) &= o(1); \\ P_{\boldsymbol{\theta}}(\mathbf{D} \subseteq \widehat{\mathbf{D}}) &= 1 + o(1) \quad \text{when } \min_{j \in \mathbf{D}} |\mu_{1j} - \mu_{2j}|^2 \geq \delta. \end{aligned}$$

Under (A-i), it holds the above results for $\zeta = 0$ in (5.7).

5.2. TWO-STAGE VARIABLE SELECTION PROCEDURE

In this section, we propose a two-stage variable selection procedure that provides screening of variables in the first stage. We select a significant set of associated variables from among a set of candidate variables in the second stage. We proceed the following two steps:

1. Choose a pilot sample size m such that $m = O(\log p)$ and $m \rightarrow \infty$ as $p \rightarrow \infty$. Take pilot samples \mathbf{x}_{il} , $l = 1, \dots, m$, of size m from each π_i . Calculate $T_{j(m)} = \bar{x}_{1jm} - \bar{x}_{2jm}$, $j = 1, \dots, p$, where $\bar{x}_{ijm} = \sum_{l=1}^m x_{ijl}/m$ for each π_i . Then, provide screening of variables by

$$\tilde{\mathbf{D}} = \{j \mid |T_{j(m)}| > \sqrt{\delta}\} \quad (5.8)$$

for a set of candidate variables. Calculate $|\tilde{\mathbf{D}}|$ ($= \tilde{S}$, say) and $\tilde{s}_i = \max_{j \in \tilde{\mathbf{D}}} (m-1)^{-1} \sum_{l=1}^m (x_{ijl} - \bar{x}_{ijm})^2$, $i = 1, 2$. Define the additional sample size for each π_i by

$$N_i = \left\lceil \frac{2 \max\{(\log \tilde{S})^{1+\xi}, (\log p)^\varepsilon\}}{\delta} \sqrt{\tilde{s}_i} \sum_{j=1}^2 \sqrt{\tilde{s}_j} \right\rceil + 1, \quad (5.9)$$

where $\xi \in (0, 1)$ and $\varepsilon \in (0, 1)$ are chosen constants.

2. Regarding $j \in \tilde{\mathbf{D}}$, take new samples x_{ijl} , $l = m+1, \dots, m+N_i$, of size N_i from each π_i . Calculate $T_{j(\mathbf{N})} = \bar{x}_{1j(N_1)} - \bar{x}_{2j(N_2)}$, where $\bar{x}_{ij(N_i)} = \sum_{l=m+1}^{m+N_i} x_{ijl}/N_i$, $j \in \tilde{\mathbf{D}}$ for each π_i . Then, test the hypothesis by

$$\text{rejecting } H_{0j} \iff |T_{j(\mathbf{N})}| > \sqrt{\delta} \quad (5.10)$$

for $j \in \tilde{\mathbf{D}}$, and define

$$\widehat{\mathbf{D}} = \{j \in \tilde{\mathbf{D}} \mid \text{rejecting } H_{0j}\}. \quad (5.11)$$

Select the variables regarding $\widehat{\mathbf{D}}$.

Then, we have the following theorem.

Theorem 5.2. *The two-stage variable selection procedure (5.8)-(5.11) given by (5.10) has (5.6) as $p \rightarrow \infty$.*

We emphasize that the two-stage variable selection procedure allows the experimenter to reduce the cost of sampling in the second stage by taking samples only from $\tilde{\mathbf{D}}$.

5.3. SIMULATION

In order to study the performance of the two-stage variable selection procedure, we took resort to computer simulations. Our goal was to estimate \mathbf{D} with accuracy regarding asymptotic FWER=0 and asymptotic AP=1 when $\min_{j \in \mathbf{D}} |\mu_{1j} - \mu_{2j}|^2 > \delta$. We fixed $\delta = 1$.

We set $\boldsymbol{\mu}_1 = (0, \dots, 0)^T$, and $\boldsymbol{\mu}_2 = (1.5, \dots, 1.5, 0, \dots, 0)^T$ whose first 20 elements are 1.5. Then, $\mathbf{D} = \{j : \mu_{1j} \neq \mu_{2j}\} = \{1, \dots, 20\}$ and $S = |\mathbf{D}| = 20$. We considered $\boldsymbol{\Sigma}_1 = \mathbf{B}(0.3^{|i-j|^{1/3}})\mathbf{B}$ and $\boldsymbol{\Sigma}_2 = 1.2\mathbf{B}(0.3^{|i-j|^{1/3}})\mathbf{B}$, where \mathbf{B} is given by (2.9). Independent pseudorandom observations \mathbf{x}_{ij} , $i = 1, 2$; $j = 1, 2, \dots$, were generated from $\pi_i : N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$. We considered the following four cases: (1) $p = 1600$; (2) $p = 3200$; (3) $p = 4800$ and (4) $p = 6400$. For all cases, we set $m = 20$. We fixed $(\zeta, \varepsilon, \xi) = (0.2, 0.2, 0.2)$ in (5.4) and (5.9).

In Table 5, we compared the performance of the two-stage variable selection procedure (5.8)-(5.11) given by (5.10) with the fixed-sample procedure given by (5.5) when $(n_1, n_2) = ([C_1] + 1, [C_2] + 1)$. The findings obtained by averaging the outcomes from 2000 ($= R$, say) replications were summarized in each case. Under a fixed scenario, suppose that the r th replication provides $\tilde{\mathbf{D}}_r$ from (5.8) in the first stage. We calculated $\tilde{S} = R^{-1} \sum_{r=1}^R |\tilde{\mathbf{D}}_r|$ to estimate the mean of the number of candidate variables, \tilde{S} . Suppose that the r th replication ends with $N_i = n_{ir}$ ($i = 1, 2$) observations from (5.9) and the corresponding test rule (5.10) together with a set of selected variables, $\hat{\mathbf{D}}_r$ from (5.11). We calculated $pm + \tilde{S}n_i = pm + R^{-1} \sum_{r=1}^R |\tilde{\mathbf{D}}_r|n_{ir}$ to estimate the mean of the number of required observations for each π_i . We calculated $\hat{S} = R^{-1} \sum_{r=1}^R |\hat{\mathbf{D}}_r|$ to estimate the mean of the number of selected variables, say \hat{S} . We checked whether $|\mathbf{D}^c \cap \hat{\mathbf{D}}_r| = |\{21, \dots, p\} \cap \hat{\mathbf{D}}_r| \neq 0$ (or $= 0$) and defined $P_r = 1$ (or 0) accordingly. We calculated $\bar{P} = R^{-1} \sum_{r=1}^R P_r$ to estimate the target FWER, having its estimated standard error $s(\bar{P})$, where $s^2(\bar{P}) = R^{-1}\bar{P}(1 - \bar{P})$. We calculated $|\mathbf{D} \cap \hat{\mathbf{D}}|/S = R^{-1} \sum_{r=1}^R |\{1, \dots, 20\} \cap \hat{\mathbf{D}}_r|/20$ to estimate the target asymptotic AP.

Let us explain, for example, the entries from the first block when $p = 3200$. For the two-stage variable selection procedure (5.8)-(5.11) given by (5.10), we observed $\tilde{S} = 32.89$ in the first stage and $\hat{S} = 19.63$ in the second stage. The numbers of required observations, $pm + \tilde{S}n_i$, $i = 1, 2$, were (65342, 65471) on average. Then, we had $\bar{P} = 0.002$ with $s(\bar{P}) = 0.001$ for FWER and $|\mathbf{D} \cap \hat{\mathbf{D}}|/S = 0.982$ for the asymptotic AP. On the other hand, for the fixed-sample procedure given by (5.5), we observed $\tilde{S} = 19.99$. Then, it should be noted that the number of observations was (249600, 272000). We had $\bar{P} = 0.0$ with $s(\bar{P}) = 0.0$ and $|\mathbf{D} \cap \hat{\mathbf{D}}|/S = 0.999$. Throughout, we observed that the number of candidate variables, \tilde{S} , is extremely small compared to p . The two-stage variable selection procedure allows the experimenter to reduce the cost of sampling in the second stage.

Table 5. Two-stage variable selection procedure (5.8)-(5.11) given by (5.10)
vs. Fixed-sample procedure given by (5.5).

\widetilde{S}	$(m + \bar{n}_1, m + \bar{n}_2)$	$(pm + \overline{\widetilde{S}n_1}, pm + \overline{\widetilde{S}n_2})$	\widetilde{S}	\bar{P}	$s(\bar{P})$	$\frac{ \overline{D \cap \widehat{D}} }{S}$
When $p = 1600$						
Two-stage variable selection procedure: $m = 20$						
26.28	(53.06, 56.11)	(32880, 32961)	19.59	0.005	0.00158	0.979
Fixed-sample procedure: $n_1 = 70$ and $n_2 = 76$						
		(112000, 121600)	19.99	0.0	0.0	0.999
When $p = 3200$						
Two-stage variable selection procedure: $m = 20$						
32.89	(60.41, 64.3)	(65342, 65471)	19.63	0.002	0.001	0.982
Fixed-sample procedure: $n_1 = 78$ and $n_2 = 85$						
		(249600, 272000)	19.99	0.0	0.0	0.999
When $p = 4800$						
Two-stage variable selection procedure: $m = 20$						
39.24	(64.96, 69.27)	(97779, 97950)	19.63	0.001	0.0005	0.981
Fixed-sample procedure: $n_1 = 82$ and $n_2 = 90$						
		(393600, 432000)	20.0	0.001	0.0005	1.0
When $p = 6400$						
Two-stage variable selection procedure: $m = 20$						
45.87	(69.29, 74.14)	(130279, 130503)	19.65	0.001	0.00071	0.983
Fixed-sample procedure: $n_1 = 86$ and $n_2 = 94$						
		(550400, 601600)	19.99	0.0	0.0	0.999

5.4. EXAMPLE

We analyzed gene expression data given by Chiaretti et al. (2004) that was used in Section 4.4 as well. The data set consisted of 12625 ($= p$) genes and two tumor cellular subtypes, π_1 : B-cell and π_2 : T-cell. We set $\delta = 2.5^2$. Our goal was to find variables j 's such that $|\mu_{1j} - \mu_{2j}| > 2.5$. We chose the pilot sample size for each π_i as $m = 15$. Then, we took the first 15 samples from each π_i in their data set as pilot samples, that are given in Table 6.

Table 6. Pilot samples, x_{ijl} ($p = 12625$, $m = 15$)

$j \backslash l$	π_1 : B-cell			π_2 : T-cell		
	1	...	15	1	...	15
1	7.597	...	7.892	7.240	...	7.165
\vdots	\vdots		\vdots	\vdots		\vdots
\vdots	\vdots		\vdots	\vdots		\vdots
12625	3.806	...	3.945	3.599	...	2.798

We considered screening variables by $\tilde{D} = \{j \mid |\bar{x}_{1jm} - \bar{x}_{2jm}| > 2.5\}$. Then, we obtained a set of candidate variables as $\tilde{D} = \{106, 122, 1144, \dots, 11576, 11834\}$ with $\tilde{S} = |\tilde{D}| = 35$. For $j \in \tilde{D}$, we calculated $\tilde{s}_i = \max_{j \in \tilde{D}} (m-1)^{-1} \sum_{l=1}^m (x_{ijl} - \bar{x}_{ijm})^2$ and obtained $(\tilde{s}_1, \tilde{s}_2) = (5.563, 2.474)$. We set $(\xi, \varepsilon) = (0.4, 0.4)$. According to (5.9), the additional sample size for each π_i is given by

$$N_1 = \left\lceil \frac{2 \max\{(\log \tilde{S})^{1+\xi}, (\log p)^\varepsilon\}}{\delta} \sqrt{\tilde{s}_1} \sum_{j=1}^2 \sqrt{\tilde{s}_j} \right\rceil + 1 = 18,$$

$$N_2 = \left\lceil \frac{2 \max\{(\log \tilde{S})^{1+\xi}, (\log p)^\varepsilon\}}{\delta} \sqrt{\tilde{s}_2} \sum_{j=1}^2 \sqrt{\tilde{s}_j} \right\rceil + 1 = 12.$$

Regarding $j \in \tilde{D}$, we took additional samples x_{ijl} , $l = m+1, \dots, m+N_i$, of size N_i from each π_i , which are given in Table 7.

Table 7. Additional samples, x_{ijl} , $j \in \tilde{D}$ ($\tilde{S} = 35$, $(N_1, N_2) = (18, 12)$)

$j \backslash l$	π_1 : B-cell			π_2 : T-cell		
	16	...	33	16	...	27
106	9.414	...	8.575	6.377	...	5.439
122	6.620	...	5.457	9.834	...	10.119
\vdots	\vdots		\vdots	\vdots		\vdots
11834	9.461	...	8.379	7.010	...	4.306

We selected significant variables by $\hat{D} = \{j \in \tilde{D} \mid \text{rejecting } H_{0j}\} = \{j \in \tilde{D} \mid |\bar{x}_{1j(N_1)} - \bar{x}_{2j(N_2)}| > 2.5\}$ and finally obtained

$$\hat{D} = \{106, 122, 1271, 2673, 3268, 3740, 5064, 6702, 7106, 7414, 8172, 8173, \\ 8225, 8321, 8399, 8917, 9002, 9478, 9932, 10299, 10670, 11270, 11271, 11834\}$$

with $\widehat{S} = |\widehat{D}| = 24$. For $j \in \widehat{D}$, we calculated $\bar{x}_{ijm+N_i} = \sum_{l=1}^{m+N_i} x_{ijl}/(m+N_i)$ for each π_i and obtained estimates of $\mu_{1j} - \mu_{2j}$ for $j \in \widehat{D}$ as

$$\begin{aligned} & \{\bar{x}_{1jm+N_1} - \bar{x}_{2jm+N_2} | j \in \widehat{D}\} \\ &= \{2.595, -3.315, 2.982, -3.347, -3.068, 2.890, 3.273, 3.301, 3.361, 2.574, 3.849, 4.443, \\ & \quad -3.202, 2.896, -4.486, 3.036, -2.763, 3.831, 3.571, 2.618, 3.236, 3.290, 2.909, 3.671\}. \end{aligned}$$

6. HIGH-DIMENSIONAL REGRESSION

We consider a usual high-dimensional linear regression setup with a response vector $\mathbf{y} = (y_1, \dots, y_n)^T$ and an $p \times n$ fixed design matrix $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]$ such that

$$E(\mathbf{y}) = \beta_0 \mathbf{1} + \mathbf{X}^T \boldsymbol{\beta}$$

with $\mathbf{1} = (1, \dots, 1)^T$. Here, β_0 is an unknown intercept and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ is an unknown p -vector of regression model parameters. Meinshausen et al. (2009) and Wasserman and Roeder (2009) considered variable selection for the regression model by using hypothesis testing to eliminate some variables such as

$$H_{0j} : \beta_j = 0 \quad \text{vs.} \quad H_{1j} : \beta_j \neq 0 \quad \text{for } j = 1, \dots, p. \quad (6.1)$$

We consider the case that $y_i \in \{1, -1\}$. Let us rewrite that $\mathbf{X} = [\mathbf{x}_{1(1)}, \dots, \mathbf{x}_{n_1(1)}, \mathbf{x}_{1(2)}, \dots, \mathbf{x}_{n_2(2)}]$ and $\mathbf{y} = (1, \dots, 1, -1, \dots, -1)^T$ whose first n_1 elements are 1 and last n_2 elements are -1. Note that $n_1 + n_2 = n$. We assume for $i = 1, 2$ that $n_i/n \rightarrow \eta_i \in (0, 1)$, $n_i^{-1} \sum_{j=1}^{n_i} \mathbf{x}_{j(i)} \rightarrow \boldsymbol{\mu}_i$ and $n_i^{-1} \sum_{j=1}^{n_i} \mathbf{x}_{j(i)} \mathbf{x}_{j(i)}^T - \boldsymbol{\mu}_i \boldsymbol{\mu}_i^T \rightarrow \boldsymbol{\Sigma}_i (> \mathbf{O})$ as $n \rightarrow \infty$.

We consider $\beta_{0\star}$ and $\boldsymbol{\beta}_\star$ such that

$$\min_{\beta_0, \boldsymbol{\beta}} \|\mathbf{y} - \beta_0 \mathbf{1} - \mathbf{X}^T \boldsymbol{\beta}\|^2.$$

Then, we have as $n \rightarrow \infty$ that

$$\beta_{0\star} = \sum_{j=1}^n y_j/n - \boldsymbol{\beta}^T \sum_{i=1}^2 \sum_{j=1}^{n_i} \mathbf{x}_{j(i)}/n \rightarrow \eta_1 - \eta_2 - \boldsymbol{\beta}^T (\eta_1 \boldsymbol{\mu}_1 + \eta_2 \boldsymbol{\mu}_2). \quad (6.2)$$

Let $\mathbf{y}_0 = \mathbf{y} - (\eta_1 - \eta_2) \mathbf{1}$ and $\mathbf{X}_0 = \mathbf{X} - (\eta_1 \boldsymbol{\mu}_1 + \eta_2 \boldsymbol{\mu}_2) \mathbf{1}^T$. Then, from (6.2), note that $\|\mathbf{y} - \beta_{0\star} \mathbf{1} - \mathbf{X}^T \boldsymbol{\beta}\|^2 \rightarrow \|\mathbf{y}_0 - \mathbf{X}_0^T \boldsymbol{\beta}\|^2$ as $n \rightarrow \infty$. Since it holds as $n \rightarrow \infty$ that

$$\begin{aligned} \frac{\mathbf{X}_0 \mathbf{y}_0}{n} &\rightarrow 2\eta_1 \eta_2 (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2); \\ \frac{\mathbf{X}_0 \mathbf{X}_0^T}{n} &\rightarrow \eta_1 \boldsymbol{\Sigma}_1 + \eta_2 \boldsymbol{\Sigma}_2 + \eta_1 \eta_2 (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T (= \boldsymbol{\Sigma}, \text{ say}), \end{aligned}$$

we have as $n \rightarrow \infty$ that

$$\boldsymbol{\beta}_\star \rightarrow 2\eta_1 \eta_2 \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

Let $\Delta = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2$. Let $\lambda_{(1)}$ and $\lambda_{(2)}$ be the largest eigenvalues of $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$. We assume that $\lambda_{(i)}/\Delta \rightarrow 0$, $i = 1, 2$, as $p \rightarrow \infty$. Then, we have that $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)/\Delta^2 \rightarrow \eta_1 \eta_2$ as $p \rightarrow \infty$. By noting that $\boldsymbol{\Sigma}/\Delta \rightarrow \eta_1 \eta_2 (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T/\Delta$ as $p \rightarrow \infty$, we have as

$p \rightarrow \infty$ that $2\eta_1\eta_2\Delta\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \rightarrow 2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$. Let $\boldsymbol{\beta}_\star = (\beta_{1\star}, \dots, \beta_{p\star})^T$. Then, it holds as $p \rightarrow \infty$ after $n \rightarrow \infty$ that

$$\Delta\beta_{j\star} \rightarrow 2(\mu_{1j} - \mu_{2j}) \quad \text{for } j = 1, 2, \dots$$

Therefore, we can claim that testing (6.1) is equivalent to testing (5.1).

Wasserman and Roeder (2009) also considered the lasso estimator $\boldsymbol{\beta}_\lambda$ such as

$$\boldsymbol{\beta}_\lambda = \min_{\boldsymbol{\beta}} \left(\|\mathbf{y}_0 - \mathbf{X}_0^T \boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^p |\beta_j| \right),$$

where λ is a smoothing parameter. Here, note that $\|\boldsymbol{\beta}_\lambda\| = O(\|\boldsymbol{\beta}_\star\|)$. When we follow the same arguments stated above, it holds as $p \rightarrow \infty$ after $n \rightarrow \infty$ that

$$\begin{aligned} \argmin_{\boldsymbol{\beta}} \left(\|\mathbf{y}_0 - \mathbf{X}_0^T \boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^p |\beta_j| \right) &\rightarrow \argmin_{\boldsymbol{\beta}} \left(\|\mathbf{X}_0^T (\boldsymbol{\beta} - \boldsymbol{\beta}_\star)\|^2 + \lambda \sum_{j=1}^p |\beta_j| \right) \\ &\rightarrow \argmin_{\boldsymbol{\beta}} \left(\eta_1\eta_2(\boldsymbol{\beta}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))^2 - 4\eta_1\eta_2\boldsymbol{\beta}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \frac{\lambda}{n} \sum_{j=1}^p |\beta_j| \right) \\ &= \argmin_{\boldsymbol{\beta}} \left((\boldsymbol{\beta}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))^2 - 4 \sum_{j=1}^p |\beta_j| \left(\text{sgn}(\beta_j)(\mu_{1j} - \mu_{2j}) - \lambda/(4n\eta_1\eta_2) \right) \right). \end{aligned}$$

We observe that the lasso estimator concludes $\beta_j = 0$ if one chooses λ as $4n\eta_1\eta_2|\mu_{1j} - \mu_{2j}| < \lambda$. Then, one would note that λ is equivalent to $4n\eta_1\eta_2\sqrt{\delta}$, where δ is the prespecified constant discussed in Section 5. Thus, in the above setting, the variable selection given by Section 5 might be promising to compare favorably with the Lasso in terms of the assurance of accuracy and the computational cost.

7. CLASSIFICATION AFTER VARIABLE SELECTION

In this section, we consider applying a variable selection procedure to classification. Fan and Fan (2008) also considered this problem. Suppose we have $\pi_i : N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, $i = 1, 2$. First, we consider testing the hypotheses given by (5.1). We are interested in designing a $\widehat{\mathbf{D}}$ satisfying (5.2) and

$$P_{\boldsymbol{\theta}}(\mathbf{D} \subseteq \widehat{\mathbf{D}}) \rightarrow 1 \quad \text{when } \min_{j \in \mathbf{D}} \frac{|\mu_{1j} - \mu_{2j}|^2}{\sigma_{(1)j} + \sigma_{(2)j}} \geq \delta, \quad (7.1)$$

where $\delta (> 0)$ is a prespecified constant.

Let us choose the sample size for each π_i as

$$n_i \geq \frac{8(\log p)^{1+\zeta}}{\delta} \quad (7.2)$$

with $\zeta \in (0, 1)$ chosen, and test the hypothesis for $j = 1, \dots, p$, by

$$\text{rejecting } H_{0j} \iff \frac{|T_{j(\mathbf{n})}|}{\sqrt{s_{(1)jn_1} + s_{(2)jn_2}}} > \sqrt{\delta}/2, \quad (7.3)$$

where $s_{(i)jn_i} = (n_i - 1)^{-1} \sum_{l=1}^{n_i} (x_{ijl} - \bar{x}_{ijn_i})^2$, $i = 1, 2$, with $\bar{x}_{ijn_i} = \sum_{l=1}^{n_i} x_{ijl} / n_i$. Let $\hat{\mathbf{D}} = \{j \mid \text{rejecting } H_{0j}\}$. Then, we have the following theorem.

Theorem 7.1. *The test given by (7.3) with (7.2) has as $p \rightarrow \infty$ that*

$$P_{\theta}(|\mathbf{D}^c \cap \hat{\mathbf{D}}| \neq 0) = o(1);$$

$$P_{\theta}(\mathbf{D} \subseteq \hat{\mathbf{D}}) = o(1) \quad \text{when} \quad \min_{j \in \mathbf{D}} \frac{|\mu_{1j} - \mu_{2j}|^2}{\sigma_{(1)j} + \sigma_{(2)j}} \geq \delta.$$

Next, we consider a discriminant rule by using only the variables of $\hat{\mathbf{D}}$. We assume that $\min_{j \in \mathbf{D}} (\sigma_{(1)j} + \sigma_{(2)j})^{-1} (\mu_{1j} - \mu_{2j})^2 \geq \delta$. Let $S = |\mathbf{D}|$ and $\hat{S} = |\hat{\mathbf{D}}|$. Let $f_{\mathbf{D}}(\cdot)$ denote the function such that $f_{\mathbf{D}}(\mathbf{x}) = (x_{(1)}, \dots, x_{(S)})^T$ for any vector $\mathbf{x} = (x_1, \dots, x_p)^T$, where $x_{(j)} = x_{j'}$ with j' the j -th smallest subscript in \mathbf{D} . Similarly, let us write $f_{\hat{\mathbf{D}}}(\mathbf{x}) = (x_{(1)}, \dots, x_{(\hat{S})})^T$ for any vector $\mathbf{x} = (x_1, \dots, x_p)^T$. Let $\boldsymbol{\mu}_{i(S)}$ and $\boldsymbol{\Sigma}_{i(S)}$ be a mean vector and a covariance matrix of $f_{\mathbf{D}}(\mathbf{x}_{ik})$ ($k = 1, \dots, n_i$). We assume that $S \rightarrow \infty$ and

$$\frac{(\boldsymbol{\mu}_{1(S)} - \boldsymbol{\mu}_{2(S)})^T \boldsymbol{\Sigma}_{i(S)} (\boldsymbol{\mu}_{1(S)} - \boldsymbol{\mu}_{2(S)})}{\|\boldsymbol{\mu}_{1(S)} - \boldsymbol{\mu}_{2(S)}\|^4} \rightarrow 0$$

for $i = 1, 2$, as $p \rightarrow \infty$. Let $\hat{\mathbf{x}} = f_{\hat{\mathbf{D}}}(\mathbf{x})$ for any vector $\mathbf{x} = (x_1, \dots, x_p)^T$. Now, we calculate $\bar{\mathbf{x}}_{in_i} = \sum_{l=1}^{n_i} \hat{\mathbf{x}}_{il} / n_i$ and $\hat{\mathbf{S}}_{in_i} = \sum_{l=1}^{n_i} (\hat{\mathbf{x}}_{il} - \bar{\mathbf{x}}_{in_i})(\hat{\mathbf{x}}_{il} - \bar{\mathbf{x}}_{in_i})^T / (n_i - 1)$ for each π_i . Let \mathbf{x}_0 be an observation vector on an individual belonging to π_1 or to π_2 . We propose the discriminant rule that one classifies \mathbf{x}_0 into π_1 if

$$\left(\hat{\mathbf{x}}_0 - \frac{\bar{\mathbf{x}}_{1n_1} + \bar{\mathbf{x}}_{2n_2}}{2} \right)^T (\bar{\mathbf{x}}_{2n_2} - \bar{\mathbf{x}}_{1n_1}) - \frac{\text{tr}(\hat{\mathbf{S}}_{1n_1})}{2n_1} + \frac{\text{tr}(\hat{\mathbf{S}}_{2n_2})}{2n_2} < 0 \quad (7.4)$$

and into π_2 otherwise. As for the error rates of misclassification, $e(2|1)$ and $e(1|2)$, we have the following theorem.

Theorem 7.2. *The discriminant rule given by (7.4) has as $p \rightarrow \infty$ that*

$$e(2|1) \rightarrow 0 \quad \text{and} \quad e(1|2) \rightarrow 0.$$

8. PATHWAY ANALYSIS FOR HIGH-DIMENSIONAL DATA

In this section, we consider pathway analysis for high-dimensional data by constructing a multiple test of correlation coefficients. Suppose we have i.i.d. $p + 1$ -variate data vectors, $\mathbf{x}_{j(*)} = (\mathbf{x}_j^T, x_{j(*)})^T$, $j = 1, \dots, n$, where $\mathbf{x}_j^T = (x_{1j}, \dots, x_{pj})$. We assume $n \geq 4$. Here, \mathbf{x}_j has unknown mean vector $\boldsymbol{\mu}$ and unknown covariance matrix $\boldsymbol{\Sigma} (> \mathbf{O})$, and $x_{j(*)}$ has unknown mean μ_* and unknown variance $\sigma_* (< \infty)$. We denote the correlation coefficient vector between \mathbf{x}_j and $x_{j(*)}$ by $\text{Corr}(\mathbf{x}_j, x_{j(*)}) = \boldsymbol{\rho}$, where $\boldsymbol{\rho} = (\rho_1, \dots, \rho_p)^T$. We consider testing the correlation between \mathbf{x}_j and $x_{j(*)}$ as follows:

$$H_0 : \boldsymbol{\rho} = \mathbf{0} \quad \text{vs.} \quad H_1 : \boldsymbol{\rho} \neq \mathbf{0}.$$

The test of the correlation is a very important tool of pathway analysis or graphical modeling for high-dimensional data. For example, Kraft et al. (2003) and Drton and Perlman (2007)

considered the pathway analysis or graphical modeling of microarray data by testing a correlation coefficient individually.

Let $\bar{\mathbf{x}}_{n_{(1)}} = \sum_{j=1}^{n_{(1)}} \mathbf{x}_j / n_{(1)}$, $\bar{\mathbf{x}}_{n_{(2)}} = \sum_{j=n_{(1)+1}}^n \mathbf{x}_j / n_{(2)}$, $\bar{x}_{n_{(1)}(*)} = \sum_{j=1}^{n_{(1)}} x_{j(*)} / n_{(1)}$ and $\bar{x}_{n_{(2)}(*)} = \sum_{j=n_{(1)+1}}^n x_{j(*)} / n_{(2)}$, where $n_{(1)} = \lfloor n/2 \rfloor + 1$ and $n_{(2)} = n - n_{(1)}$. We denote the covariance vector between \mathbf{x}_j and $x_{j(*)}$ by $Cov(\mathbf{x}_j, x_{j(*)}) = \boldsymbol{\sigma}$, where $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_p)^T$. We propose an estimator of $\|\boldsymbol{\sigma}\|^2$ by

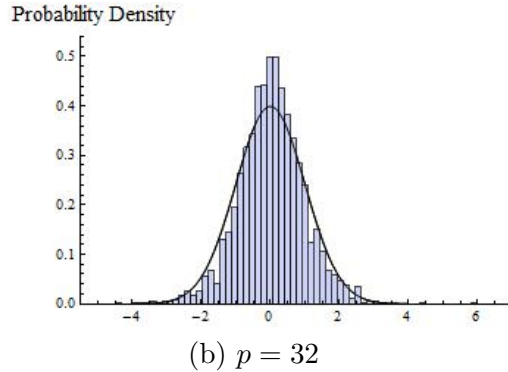
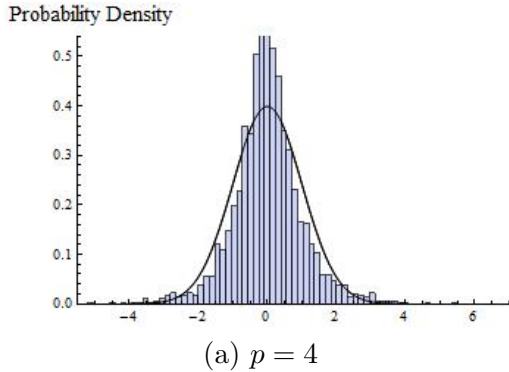
$$\hat{T}_{\boldsymbol{\sigma}} = \left(\sum_{j=1}^{n_{(1)}} \frac{(x_{j(*)} - \bar{x}_{n_{(1)}(*)})(\mathbf{x}_j - \bar{\mathbf{x}}_{n_{(1)}})}{n_{(1)} - 1} \right)^T \left(\sum_{j=n_{(1)+1}}^n \frac{(x_{j(*)} - \bar{x}_{n_{(2)}(*)})(\mathbf{x}_j - \bar{\mathbf{x}}_{n_{(2)}})}{n_{(2)} - 1} \right). \quad (8.1)$$

Note that $E(\hat{T}_{\boldsymbol{\sigma}}) = \|\boldsymbol{\sigma}\|^2$. Let $\boldsymbol{\Sigma} = \mathbf{H}\boldsymbol{\Lambda}\mathbf{H}^T$, where $\boldsymbol{\Lambda}$ is a diagonal matrix of eigenvalues $\lambda_1 \geq \dots \geq \lambda_p > 0$ and \mathbf{H} is an orthogonal matrix of corresponding eigenvectors. Let $\mathbf{z}_j = (z_{1j}, \dots, z_{pj})^T = \boldsymbol{\Lambda}^{-1/2} \mathbf{H}^T (\mathbf{x}_j - \boldsymbol{\mu})$. We assume that z_{ij} , $i = 1, \dots, p$, are independent, the fourth moments of z_{ij} 's and $x_{j(*)}$ are uniformly bounded, $\lambda_p > 0$ as $p \rightarrow \infty$, and \mathbf{x}_j and $x_{j(*)}$ ($j = 1, \dots, n$) are independent when $\boldsymbol{\rho} = \mathbf{0}$. Then, we have the following theorem.

Theorem 8.1. Assume that $tr(\boldsymbol{\Sigma}^4)/tr(\boldsymbol{\Sigma}^2)^2 \rightarrow 0$ as $p \rightarrow \infty$. When $\boldsymbol{\rho} = \mathbf{0}$, we have as $p \rightarrow \infty$ and $n \rightarrow \infty$ that

$$\hat{T}_{\boldsymbol{\sigma}} \sqrt{(n_{(1)} - 1)(n_{(2)} - 1)/(\sigma_*^2 tr(\boldsymbol{\Sigma}^2))} \Rightarrow N(0, 1).$$

Let us observe Theorem 8.1. Now, we consider an easy example such as $\boldsymbol{\mu} = \mathbf{0}$, $\boldsymbol{\Sigma} = (0.3^{|i-j|^{1/3}})$, $\mu_* = 0$, $\sigma_* = 1$ and $n = 40$. Figs. 4(a), (b), (c) and (d) give the histograms of 2000 independent outcomes of $\hat{T}_{\boldsymbol{\sigma}} \sqrt{(n_{(1)} - 1)(n_{(2)} - 1)/(\sigma_*^2 tr(\boldsymbol{\Sigma}^2))}$ when $p = 4, 32, 256$ and 2048. Here, \mathbf{x}_j , $j = 1, \dots, n$, were generated from independent pseudorandom normal distribution with zero mean and covariance matrix $\boldsymbol{\Sigma}$ for $p = 4, 32, 256$ and 2048. Independent of \mathbf{x}_j , $x_{j(*)}$, $j = 1, \dots, n$, were generated from independent pseudorandom normal distribution with zero mean and variance σ_* . Thus it holds $\boldsymbol{\rho} = \mathbf{0}$. When $p = 4$ and $p = 32$, the histograms appear different from $N(0, 1)$. However, when $p = 256$, the histogram becomes quite similar to $N(0, 1)$. When $p = 2048$, the histogram fits $N(0, 1)$ perfectly as expected.



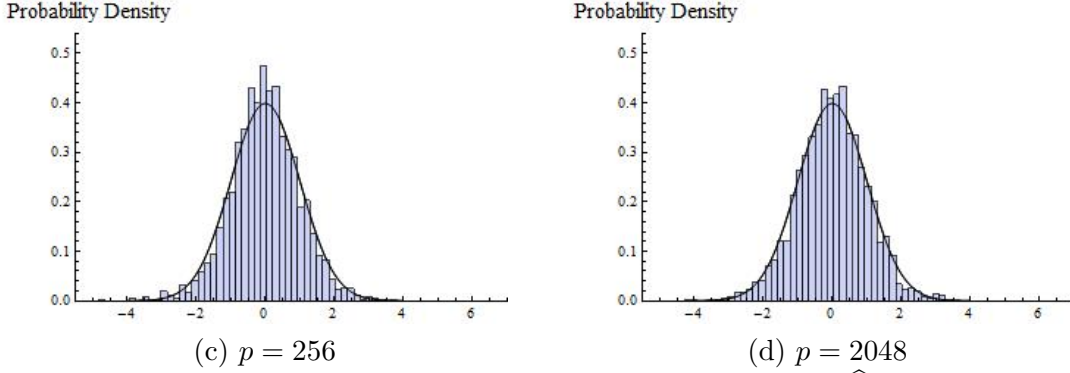


Figure 4. The solid line is probability density of $N(0, 1)$. The histogram of $\hat{T}\boldsymbol{\sigma}\sqrt{(n_{(1)} - 1)(n_{(2)} - 1)}/\sqrt{\sigma_*^2 \text{tr}(\boldsymbol{\Sigma}^2)}$ fits the solid line with increasing dimension: (a) $p=4$, (b) $p=32$, (c) $p=256$, and (d) $p=2048$.

Remark 8.1. Let $\hat{\sigma}_* = (n - 1)^{-1} \sum_{j=1}^n (x_{j(*)} - \bar{x}_{n(*)})^2$, where $\bar{x}_{n(*)} = \sum_{j=1}^n x_{j(*)}/n$. Assume that $\text{tr}(\boldsymbol{\Sigma}^4)/\text{tr}(\boldsymbol{\Sigma}^2)^2 \rightarrow 0$ as $p \rightarrow \infty$. When $\boldsymbol{\rho} = \mathbf{0}$, we have as $p \rightarrow \infty$ and $n \rightarrow \infty$ that

$$\hat{T}\boldsymbol{\sigma}\sqrt{(n_{(1)} - 1)(n_{(2)} - 1)/(\hat{\sigma}_*^2 \text{tr}(\mathbf{S}_{n(1)}\mathbf{S}_{n(2)}))} \Rightarrow N(0, 1),$$

where $\mathbf{S}_{n(1)}$ and $\mathbf{S}_{n(2)}$ are defined similarly to (1.2).

By using Theorem 8.1 (or Remark 8.1), the experimenter can conduct a test whether $\boldsymbol{\rho} = \mathbf{0}$ or $\boldsymbol{\rho} \neq \mathbf{0}$. There are future prospects to develop a two-stage procedure for the correlation test in the pathway analysis for high-dimensional data.

APPENDIX

Proof of Theorem 2.1. Note that $\mathbf{T}_{\mathbf{n}} - \boldsymbol{\mu}$ is distributed as $N_p(\mathbf{0}, \sum_{i=1}^k b_i^2 \boldsymbol{\Sigma}_i/n_i)$. Let $\mathbf{H}_{(\mathbf{n})} = [\mathbf{h}_{1(\mathbf{n})}, \dots, \mathbf{h}_{p(\mathbf{n})}]$ be an orthogonal matrix such that $\mathbf{H}_{(\mathbf{n})}^T (\sum_{i=1}^k b_i^2 \boldsymbol{\Sigma}_i/n_i) \mathbf{H}_{(\mathbf{n})} = \text{diag}(\sum_{i=1}^k b_i^2 \lambda_{i1(\mathbf{n})}/n_i, \dots, \sum_{i=1}^k b_i^2 \lambda_{ip(\mathbf{n})}/n_i)$, where $\mathbf{h}_{j(\mathbf{n})}^T \boldsymbol{\Sigma}_i \mathbf{h}_{j(\mathbf{n})} = \lambda_{ij(\mathbf{n})}$, $i = 1, \dots, k$; $j = 1, \dots, p$. Note that $\sum_{j=1}^p \lambda_{ij(\mathbf{n})} = \text{tr}(\boldsymbol{\Sigma}_i)$. We write that $\|\mathbf{T}_{\mathbf{n}} - \boldsymbol{\mu}\|^2 - \Sigma_{\mathbf{n}} = \sum_{j=1}^p \sum_{i=1}^k b_i^2 \lambda_{ij(\mathbf{n})} (w_j - 1)/n_i$, where w_j , $j = 1, \dots, p$, are independent random variables distributed as a chi-square distribution with 1 degree of freedom. From (A-iv) and the assumption that $\lambda_{ip} > 0$ ($i = 1, \dots, k$) as $p \rightarrow \infty$, it holds as $p \rightarrow \infty$ that $0 < \text{tr}(\boldsymbol{\Sigma}_i \boldsymbol{\Sigma}_{i'})/p \leq \sqrt{\text{tr}(\boldsymbol{\Sigma}_i^2) \text{tr}(\boldsymbol{\Sigma}_{i'}^2)}/p < \infty$. Thus it follows that $\{\text{Var}_{\boldsymbol{\theta}}(\|\mathbf{T}_{\mathbf{n}} - \boldsymbol{\mu}\|^2)\}^{-1} = \{2 \sum_{i,i'} b_i^2 b_{i'}^2 \text{tr}(\boldsymbol{\Sigma}_i \boldsymbol{\Sigma}_{i'})/(n_i n_{i'})\}^{-1} = O(\min_{1 \leq i \leq k} n_i^2/p)$. Let $v_j = \sum_{i=1}^k b_i^2 \lambda_{ij(\mathbf{n})} n_i^{-1} (w_j - 1)/\sqrt{\text{Var}_{\boldsymbol{\theta}}(\|\mathbf{T}_{\mathbf{n}} - \boldsymbol{\mu}\|^2)}$, $j = 1, \dots, p$. Note that $\sum_{j=1}^p E_{\boldsymbol{\theta}}(v_j^2) = 1$. Then, by noting that $\sum_{j=1}^p \lambda_{ij(\mathbf{n})} \lambda_{i'j(\mathbf{n})} \lambda_{lj(\mathbf{n})} \lambda_{l'j(\mathbf{n})} \leq \max_{1 \leq i \leq k} \text{tr}(\boldsymbol{\Sigma}_i^4)$ for $i, i', l, l' = 1, \dots, k$, we have for Lyapunov's condition that

$$\sum_{j=1}^p E_{\boldsymbol{\theta}}(v_j^4) = O\left(\left(\min_{1 \leq i \leq k} n_i\right)^{-4} \max_{1 \leq i \leq k} \text{tr}(\boldsymbol{\Sigma}_i^4)\right) \times O\left(\min_{1 \leq i \leq k} n_i^4/p^2\right) \rightarrow 0$$

under (A-iv). Then, from Lyapunov's central limit theorem, we obtain that

$$\frac{\|\mathbf{T}_{\mathbf{n}} - \boldsymbol{\mu}\|^2 - \Sigma_{\mathbf{n}}}{\sqrt{2 \sum_{i,i'} b_i^2 b_{i'}^2 \text{tr}(\boldsymbol{\Sigma}_i \boldsymbol{\Sigma}_{i'}) / (n_i n_{i'})}} \Rightarrow N(0, 1)$$

when $p \rightarrow \infty$ and either $n_i \rightarrow \infty$ or n_i is fixed for $i = 1, \dots, k$. It concludes the result. \square

Proof of Theorem 2.2 and Corollary 2.1. We first consider (A-ii). We have from (A-iv) that

$$\begin{aligned} & \|\mathbf{T}_{\mathbf{n}} - \boldsymbol{\mu}\|^2 - \widehat{\Sigma}_{\mathbf{n}} \\ &= \sum_{i=1}^k b_i^2 \sum_{l \neq l'}^{n_i} \frac{(\mathbf{x}_{il} - \boldsymbol{\mu}_i)^T (\mathbf{x}_{il'} - \boldsymbol{\mu}_i)}{n_i(n_i - 1)} + 2 \sum_{i < i'} b_i b_{i'} \sum_{l=1}^{n_i} \sum_{l'=1}^{n_{i'}} \frac{(\mathbf{x}_{il} - \boldsymbol{\mu}_i)^T (\mathbf{x}_{i'l'} - \boldsymbol{\mu}_{i'})}{n_i n_{i'}} \\ &= \sum_{i=1}^k b_i^2 \sum_{l \neq l'}^{n_i} \frac{(\mathbf{x}_{il} - \boldsymbol{\mu}_i)^T (\mathbf{x}_{il'} - \boldsymbol{\mu}_i)}{n_i^2} + 2 \sum_{i < i'} b_i b_{i'} \sum_{l=1}^{n_i} \sum_{l'=1}^{n_{i'}} \frac{(\mathbf{x}_{il} - \boldsymbol{\mu}_i)^T (\mathbf{x}_{i'l'} - \boldsymbol{\mu}_{i'})}{n_i n_{i'}} \\ &+ o_p(\sqrt{p} / (\min_{1 \leq i \leq k} n_i)) \end{aligned} \quad (\text{A.1})$$

when $p \rightarrow \infty$ and $n_i \rightarrow \infty$ for $i = 1, \dots, k$. Note that

$$\begin{aligned} \text{Var}_{\boldsymbol{\theta}}(\|\mathbf{T}_{\mathbf{n}} - \boldsymbol{\mu}\|^2 - \widehat{\Sigma}_{\mathbf{n}}) &= 2 \sum_{i=1}^k \frac{b_i^4 \text{tr}(\boldsymbol{\Sigma}_i^2)}{n_i(n_i - 1)} + 2 \sum_{i \neq i'} \frac{b_i^2 b_{i'}^2 \text{tr}(\boldsymbol{\Sigma}_i \boldsymbol{\Sigma}_{i'})}{n_i n_{i'}} \\ &= (1 + o(1)) 2 \sum_{i, i'} \frac{b_i^2 b_{i'}^2 \text{tr}(\boldsymbol{\Sigma}_i \boldsymbol{\Sigma}_{i'})}{n_i n_{i'}}. \end{aligned}$$

Let $n_* = \sum_{i=1}^k n_i$. Let $\mathbf{y}_j = b_1(\mathbf{x}_{1j} - \boldsymbol{\mu}_1)/n_1$ for $j = 1, \dots, n_1$, $\mathbf{y}_{j+\sum_{i'=1}^{j-1} n_{i'}} = b_i(\mathbf{x}_{ij} - \boldsymbol{\mu}_i)/n_i$ for $j = 1, \dots, n_i$ ($i \geq 2$) and $\phi_{ij} = \mathbf{y}_i^T \mathbf{y}_j$. Define $V_{\mathbf{n}j} = \sum_{i=1}^{j-1} \phi_{ij}$ for $j = 2, \dots, n_*$. From (A.1), we write that $\|\mathbf{T}_{\mathbf{n}} - \boldsymbol{\mu}\|^2 - \widehat{\Sigma}_{\mathbf{n}} = 2 \sum_{j=2}^{n_*} V_{\mathbf{n}j} + o_p(\sqrt{p}/(\min_{1 \leq i \leq k} n_i))$. First, we consider the case that $0 < n_i/n_{i'} < \infty$ as $n_i, n_{i'} \rightarrow \infty$ for all $i \neq i' = 1, \dots, k$. By noting the $\text{tr}(\boldsymbol{\Sigma}_i \boldsymbol{\Sigma}_j \boldsymbol{\Sigma}_i \boldsymbol{\Sigma}_{j'}) \leq \max_{1 \leq i \leq k} \text{tr}(\boldsymbol{\Sigma}_i^4)$, in a way similar to the proof of Theorem 1 in Chen and Qin (2010), we can claim that

$$\frac{\|\mathbf{T}_{\mathbf{n}} - \boldsymbol{\mu}\|^2 - \widehat{\Sigma}_{\mathbf{n}}}{\sqrt{2 \sum_{i,i'} b_i^2 b_{i'}^2 \text{tr}(\boldsymbol{\Sigma}_i \boldsymbol{\Sigma}_{i'}) / (n_i n_{i'})}} \Rightarrow N(0, 1) \quad (\text{A.2})$$

when $p \rightarrow \infty$ and $n_i \rightarrow \infty$ for $i = 1, \dots, k$. Next, we consider the case that $n_i/n_{i'} \rightarrow 0$ as $n_i, n_{i'} \rightarrow \infty$ for some $i' (\neq i)$. By noting that $\|\mathbf{T}_{\mathbf{n}} - \boldsymbol{\mu}\|^2 - \widehat{\Sigma}_{\mathbf{n}} = \|\sum_{i \in \setminus i'}^k b_i(\bar{\mathbf{x}}_i - \boldsymbol{\mu}_i)\|^2 - \sum_{i \in \setminus i'}^k b_i^2 \text{tr}(\mathbf{S}_{in_i})/n_i + o_p(\sqrt{p}/(\min_{1 \leq i \leq k} n_i))$, similarly, it holds (A.2). It concludes the result in Corollary 2.1 for (A-ii). As for Theorem 2.2, we note that

$$\frac{\Sigma_{\mathbf{n}} - \widehat{\Sigma}_{\mathbf{n}}}{\sqrt{2 \sum_{i,i'} b_i^2 b_{i'}^2 \text{tr}(\boldsymbol{\Sigma}_i \boldsymbol{\Sigma}_{i'}) / (n_i n_{i'})}} = o_p(1).$$

Thus it follows that

$$\frac{\|\mathbf{T}_n - \boldsymbol{\mu}\|^2 - \Sigma_n}{\sqrt{2 \sum_{i,i'} b_i^2 b_{i'}^2 \text{tr}(\boldsymbol{\Sigma}_i \boldsymbol{\Sigma}_{i'}) / (n_i n_{i'})}} = \frac{\|\mathbf{T}_n - \boldsymbol{\mu}\|^2 - \widehat{\Sigma}_n - \Sigma_n + \widehat{\Sigma}_n}{\sqrt{2 \sum_{i,i'} b_i^2 b_{i'}^2 \text{tr}(\boldsymbol{\Sigma}_i \boldsymbol{\Sigma}_{i'}) / (n_i n_{i'})}} \Rightarrow N(0, 1), \quad (\text{A.3})$$

which concludes the result in Theorem 2.2 for (A-ii).

Next, we consider (A-iii) with (A-v). We write that

$$\begin{aligned} & \|\mathbf{T}_n - \boldsymbol{\mu}\|^2 - \widehat{\Sigma}_n \\ &= \sum_{j=1}^p \left(\sum_{i=1}^k b_i^2 \sum_{l \neq l'}^{n_i} \frac{(x_{ijl} - \mu_{ij})(x_{ijl'} - \mu_{ij})}{n_i(n_i - 1)} + 2 \sum_{i < i'} b_i b_{i'} \sum_{l=1}^{n_i} \sum_{l'=1}^{n_{i'}} \frac{(x_{ijl} - \mu_{ij})(x_{ijl'} - \mu_{i'j})}{n_i n_{i'}} \right) \\ &= \sum_{j=1}^p Y_j \text{ (say)}. \end{aligned}$$

Let $W = \sqrt{2 \sum_{i,i'} b_i^2 b_{i'}^2 c_{ii'} / (n_i n_{i'})}$, where $c_{ii'}$'s are defined in (A-v). Then, by using Theorem 5.2 in Bradley (2005), we note from (A-iii) that $\{Y_j/W\}$ is a strictly stationary sequence and ρ -mixing. From (A-v), it holds that $\text{Var}_{\boldsymbol{\theta}}(\sum_{j=1}^p Y_j/W) = p(1 + o(1))$ as $p \rightarrow \infty$ and $n_i \rightarrow \infty$, $i = 1, \dots, k$. Then, by using Theorem 2.1 in Ibragimov (1975), we claim that $\sum_{j=1}^p Y_j / (p^{1/2} W) \Rightarrow N(0, 1)$. From the fact that $p^{1/2} W / \sqrt{2 \sum_{i,i'} b_i^2 b_{i'}^2 \text{tr}(\boldsymbol{\Sigma}_i \boldsymbol{\Sigma}_{i'}) / (n_i n_{i'})} \rightarrow 1$ as $p \rightarrow \infty$, we can claim (A.2) and (A.3) under (A-iii) with (A-v). Thus it concludes the results. \square

Proof of Theorems 2.3. From (2.4), one can claim that

$$\sqrt{2 \sum_{i,j} b_i^2 b_j^2 \text{tr}(\boldsymbol{\Sigma}_i \boldsymbol{\Sigma}_j) / (n_i n_j)} \leq \sqrt{2} \sum_{i=1}^k b_i^2 \text{tr}(\boldsymbol{\Sigma}_i^2)^{1/2} / n_i \leq \frac{\delta}{z_{\alpha/2}}.$$

Note that $n_i \rightarrow \infty$ as $p \rightarrow \infty$ for $i = 1, \dots, k$. Then, we have from (2.5) and Theorem 2.2 that

$$P_{\boldsymbol{\theta}}(\boldsymbol{\mu} \in \mathbf{R}_{\widehat{\Sigma}_n}) = P_{\boldsymbol{\theta}}(\|\mathbf{T}_n - \boldsymbol{\mu}\|^2 - \widehat{\Sigma}_n \leq \delta) \geq P_{\boldsymbol{\theta}}(|N(0, 1)| \leq z_{\alpha/2}) + o(1) = 1 - \alpha + o(1)$$

as $p \rightarrow \infty$. It concludes the result. \square

Proof of Theorems 2.4 and 2.5. We have under either (A-ii) or (A-iii) that

$$\text{Var}_{\boldsymbol{\theta}} \left(\frac{\text{tr}(\mathbf{S}_{im(1)} \mathbf{S}_{im(2)})}{\text{tr}(\boldsymbol{\Sigma}_i^2)} \right) = O(m^{-2}) + O(\text{tr}(\boldsymbol{\Sigma}_i^4) / (\text{tr}(\boldsymbol{\Sigma}_i^2)^2 m)). \quad (\text{A.4})$$

Here, from (A-i) and (A-iv), one can claim that

$$\begin{aligned} & E_{\boldsymbol{\theta}} \left(\frac{z_{\alpha/2} \sqrt{2}}{\delta} |b_i| \text{tr}(\mathbf{S}_{im(1)} \mathbf{S}_{im(2)})^{1/4} \sum_{j=1}^k |b_j| \text{tr}(\mathbf{S}_{jm(1)} \mathbf{S}_{jm(2)})^{1/4} \right) \\ &= \frac{z_{\alpha/2} \sqrt{2}}{\delta} |b_i| \text{tr}(\boldsymbol{\Sigma}_i^2)^{1/4} \sum_{j=1}^k |b_j| \text{tr}(\boldsymbol{\Sigma}_j^2)^{1/4} + o(1) = C_i + o(1); \\ & E_{\boldsymbol{\theta}} \left\{ \left(\frac{z_{\alpha/2} \sqrt{2}}{\delta} |b_i| \text{tr}(\mathbf{S}_{im(1)} \mathbf{S}_{im(2)})^{1/4} \sum_{j=1}^k |b_j| \text{tr}(\mathbf{S}_{jm(1)} \mathbf{S}_{jm(2)})^{1/4} / C_i - 1 \right)^2 \right\} = o(m^{-1}). \end{aligned} \quad (\text{A.5})$$

Here, in a way similar to the proof of Lemma 2 in Yata and Aoshima (2009a), we have under (A-i) that

$$\begin{aligned} & E_{\boldsymbol{\theta}} \left\{ \left(N_i - \left[\frac{z_{\alpha/2}\sqrt{2}}{\delta} |b_i| \text{tr}(\mathbf{S}_{im(1)} \mathbf{S}_{im(2)})^{1/4} \sum_{j=1}^k |b_j| \text{tr}(\mathbf{S}_{jm(1)} \mathbf{S}_{jm(2)})^{1/4} \right] - 1 \right)^t \right\} \\ &= o(m^{t-1}) \quad (t = 1, 2). \end{aligned} \quad (\text{A.6})$$

Then, from (A.5) and (A.6), it holds that

$$\begin{aligned} & |E_{\boldsymbol{\theta}}(N_i - C_i)| \\ & \leq |E_{\boldsymbol{\theta}} \left(\frac{z_{\alpha/2}\sqrt{2}}{\delta} |b_i| \text{tr}(\mathbf{S}_{im(1)} \mathbf{S}_{im(2)})^{1/4} \sum_{j=1}^k |b_j| \text{tr}(\mathbf{S}_{jm(1)} \mathbf{S}_{jm(2)})^{1/4} \right) - C_i| + 1 + o(1) \\ &= 1 + o(1); \\ & \text{Var}_{\boldsymbol{\theta}}(N_i) = o(m) = o(p^{1/2}/\delta). \end{aligned}$$

It concludes the results in Theorem 2.5.

Next, from (A.4), it holds as $p \rightarrow \infty$ that $|N_i - C_i| = o_p(m^{1/2})$ under (A-iv) and either (A-ii) or (A-iii). Then, we can write that $|N_i - C_i| = O_p(\omega m^{1/2})$, where ω is a variable such that $\omega \rightarrow 0$ as $p \rightarrow \infty$. Let $C_{i\star} = [C_i - (\omega m)^{1/2}]$ ($i = 1, \dots, k$). We claim as $p \rightarrow \infty$ that $m < C_{i\star} < N_i < C_i + (\omega m)^{1/2}$ w.p.1. Here, we write that $\bar{\mathbf{x}}_{iN_i} = \sum_{l=1}^{C_{i\star}} \mathbf{x}_{il}/N_i + \sum_{l=C_{i\star}+1}^{N_i} \mathbf{x}_{il}/N_i$. Then, it holds that

$$\begin{aligned} & \|\bar{\mathbf{x}}_{iN_i} - \boldsymbol{\mu}_i\|^2 - \text{tr}(\mathbf{S}_{iN_i})/N_i \\ &= \sum_{j=1}^p \lambda_{ij} \left(\frac{\sum_{l \neq l'}^{C_{i\star}} z_{ijl} z_{ijl'}}{N_i(N_i - 1)} + 2 \frac{\sum_{l=1}^{C_{i\star}} \sum_{l'=C_{i\star}+1}^{N_i} z_{ijl} z_{ijl'}}{N_i(N_i - 1)} + \frac{\sum_{l \neq l'}^{N_i} z_{ijl} z_{ijl'}}{N_i(N_i - 1)} \right). \end{aligned}$$

By using Markov's inequality and Schwarz's inequality, for any $\tau > 0$, we have from $N_i < [C_i + (\omega m)^{1/2}] + 1$ w.p.1 that

$$\begin{aligned} P_{\boldsymbol{\theta}} \left(\sum_{l \neq l' (\geq C_{i\star}+1)}^{N_i} \left| \sum_{j=1}^p \frac{\lambda_{ij} z_{ijl} z_{ijl'}}{C_{i\star}^2} \right| > \tau \delta \right) &\leq P_{\boldsymbol{\theta}} \left(\sum_{l \neq l' (\geq C_{i\star}+1)}^{[C_i + (\omega m)^{1/2}] + 1} \left| \sum_{j=1}^p \frac{\lambda_{ij} z_{ijl} z_{ijl'}}{C_{i\star}^2} \right| > \tau \delta \right) + o(1) \\ &= O(\omega m \delta^{-1} \sqrt{\text{tr}(\boldsymbol{\Sigma}_i^2)/C_{i\star}^2}) + o(1) \rightarrow 0. \end{aligned} \quad (\text{A.7})$$

Thus from $N_i/C_{i\star} = 1 + o_p(1)$, we claim that

$$\sum_{j=1}^p \frac{\lambda_{ij} \sum_{l \neq l' (\geq C_{i\star}+1)}^{N_i} z_{ijl} z_{ijl'}}{N_i(N_i - 1)} = o_p(\delta).$$

Note that

$$E_{\boldsymbol{\theta}} \left\{ \left(\sum_{j=1}^p \lambda_{ij} \frac{\sum_{l=1}^{C_{i\star}} z_{ijl} z_{ijl'}}{C_{i\star}^2} \right)^2 \right\} = O(\text{tr}(\boldsymbol{\Sigma}_i^2)/C_{i\star}^3) \quad \text{for } l' = C_{i\star} + 1, \dots, N_i.$$

In a way similar to (A.7), we have that

$$P_{\boldsymbol{\theta}} \left(\sum_{l'=C_{i\star}+1}^{N_i} \left| \sum_{j=1}^p \lambda_{ij} \frac{\sum_{l=1}^{C_{i\star}} z_{ijl} z_{ijl'}}{C_{i\star}^2} \right| > \tau \delta \right) \rightarrow 0.$$

Hence, we have that

$$\sum_{j=1}^p \lambda_{ij} \frac{\sum_{l=1}^{C_{i\star}} \sum_{l'=C_{i\star}+1}^{N_i} z_{ijl} z_{ijl'}}{N_i(N_i-1)} = o_p(\delta).$$

Then, it holds as $p \rightarrow \infty$ that

$$\|\bar{\mathbf{x}}_{iN_i} - \boldsymbol{\mu}_i\|^2 - \text{tr}(\mathbf{S}_{iN_i})/N_i = \sum_{j=1}^p \lambda_{ij} \left(\frac{\sum_{l \neq l'}^{C_{i\star}} z_{ijl} z_{ijl'}}{C_{i\star}(C_{i\star}-1)} \right) + o_p(\delta).$$

Similarly, it holds for $i \neq i'$ that

$$(\bar{\mathbf{x}}_{iN_i} - \boldsymbol{\mu}_i)^T (\bar{\mathbf{x}}_{i'N_{i'}} - \boldsymbol{\mu}_{i'}) = (\bar{\mathbf{x}}_{iC_{i\star}} - \boldsymbol{\mu}_i)^T (\bar{\mathbf{x}}_{i'C_{i'\star}} - \boldsymbol{\mu}_{i'}) + o_p(\delta).$$

Then, we have that

$$\|\mathbf{T}_{\mathbf{N}} - \boldsymbol{\mu}\|^2 - \hat{\Sigma}_{\mathbf{N}} = \|\mathbf{T}_{\mathbf{C}_{\star}} - \boldsymbol{\mu}\|^2 - \hat{\Sigma}_{\mathbf{C}_{\star}} + o_p(\delta),$$

where $\mathbf{C}_{\star} = (C_{1\star}, \dots, C_{k\star})$. Hence, similarly to the proof of Theorem 2.3, we have that

$$P_{\boldsymbol{\theta}}(|\|\mathbf{T}_{\mathbf{N}} - \boldsymbol{\mu}\|^2 - \hat{\Sigma}_{\mathbf{N}}| \leq \delta) = P_{\boldsymbol{\theta}}(|\|\mathbf{T}_{\mathbf{C}_{\star}} - \boldsymbol{\mu}\|^2 - \hat{\Sigma}_{\mathbf{C}_{\star}}| \leq \delta) + o(1) \geq 1 - \alpha + o(1).$$

It concludes the result in Theorem 2.4. \square

Proof of Theorem 3.1. From the assumptions, we have that $\hat{Var}_{\boldsymbol{\theta}}(\tilde{T}_{\mathbf{n}})/Var_{\boldsymbol{\theta}}(\tilde{T}_{\mathbf{n}}) \rightarrow 1$ as $p \rightarrow \infty$ and $n_1, n_2 \rightarrow \infty$. The remainder of the proof is the same as in the proof of Theorem 2.2 and Corollary 2.1. \square

Proof of Theorem 3.2. When $\Delta = 0$, we have from (3.3) and (3.4) that $\sqrt{Var_{\boldsymbol{\theta}}(\tilde{T}_{\mathbf{n}})} \leq \Delta_L/(z_{\alpha} + z_{\beta})$. Then, it holds as $p \rightarrow \infty$ that

$$\text{size} = P_{\boldsymbol{\theta}} \left(\tilde{T}_{\mathbf{n}} > \frac{\Delta_L z_{\alpha}}{z_{\alpha} + z_{\beta}} \right) \leq P_{\boldsymbol{\theta}} (N(0, 1) > z_{\alpha}) + o(1) = \alpha + o(1).$$

From (A-iv), it holds $\lambda_{i1}^4 = o(p^2)$. When $\Delta = \Delta_L$, we have as $p \rightarrow \infty$ that

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_i (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) / (n_i \Delta_L^2) \leq \lambda_{i1} / (n_i \Delta_L) = O(\lambda_{i1} / \text{tr}(\boldsymbol{\Sigma}_i^2)^{1/2}) = o(1).$$

Thus we claim as $p \rightarrow \infty$ that $\sqrt{Var_{\boldsymbol{\theta}}(\tilde{T}_{\mathbf{n}})}(z_{\alpha} + z_{\beta})/\Delta_L \leq 1 + o(1)$. Then, it holds as $p \rightarrow \infty$ that

$$\begin{aligned} \text{power}(\Delta_L) &= P_{\boldsymbol{\theta}} \left(\tilde{T}_{\mathbf{n}} > \frac{\Delta_L z_{\alpha}}{z_{\alpha} + z_{\beta}} \right) = P_{\boldsymbol{\theta}} \left((\tilde{T}_{\mathbf{n}} - \Delta_L)(z_{\alpha} + z_{\beta})/\Delta_L > -z_{\beta} \right) \\ &\geq P_{\boldsymbol{\theta}} (N(0, 1) > -z_{\beta}) + o(1) = 1 - \beta + o(1). \end{aligned}$$

The results follow. \square

Proof of Theorems 3.3 and 3.4. The proof is the same as in the proof of Theorems 2.4 and 2.5. We omit the details for brevity. \square

Proof of Theorems 3.5, 3.6 and 3.7. Under (A-i), we claim that $\text{tr}(\mathbf{S}_{in_i}) = \sum_{j=1}^p \lambda_{ij} w_{ijn_i} / (n_i - 1)$, where $w_{ijn_i}, j = 1, \dots, p$, are independently distributed as a chi-square distribution with $n_i - 1$ degrees of freedom. Thus, in a way similar to the proof of Theorem 2.1, we have under (A-iv) that

$$\frac{\text{tr}(\mathbf{S}_{in_i} - \boldsymbol{\Sigma}_i)}{\sqrt{2\text{tr}(\boldsymbol{\Sigma}_i^2)/(n_i - 1)}} \Rightarrow N(0, 1)$$

when $p \rightarrow \infty$ and either $n_i \rightarrow \infty$ or n_i is fixed. Hence, from the fact that \mathbf{S}_{1n_1} and \mathbf{S}_{2n_2} are independent, we obtain (3.11). By using (3.11), similarly to the proof of Theorems 2.4-2.5 and 3.2, we can conclude the results. \square

Proof of Theorem 4.1. From (A-iv) and either (A-ii) or (A-iii), we have as $p \rightarrow \infty$ and $\Delta_\star \rightarrow \infty$ that $\text{tr}(\mathbf{S}_{in_i}) = \text{tr}(\boldsymbol{\Sigma}_i) + o_p(\Delta_\star^{1/2} p^{1/2})$ for $i = 1, 2$. Then, it holds for $\mathbf{x}_0 \in \pi_i$ that

$$\begin{aligned} & \frac{p(\|\mathbf{x}_0 - \boldsymbol{\mu}_i\|^2 - \text{tr}(\boldsymbol{\Sigma}_i))(\text{tr}(\mathbf{S}_{2n_2}) - \text{tr}(\mathbf{S}_{1n_1}))}{\text{tr}(\mathbf{S}_{1n_1})\text{tr}(\mathbf{S}_{2n_2})} \\ &= \frac{p(\|\mathbf{x}_0 - \boldsymbol{\mu}_i\|^2 - \text{tr}(\boldsymbol{\Sigma}_i))(\text{tr}(\boldsymbol{\Sigma}_2) - \text{tr}(\boldsymbol{\Sigma}_1))}{\text{tr}(\mathbf{S}_{1n_1})\text{tr}(\mathbf{S}_{2n_2})} + o_p(\Delta_\star) = o_p(\Delta_\star). \end{aligned} \quad (\text{A.8})$$

We first consider the case when $\mathbf{x}_0 \in \pi_1$. Here, we have that

$$p \log \left(\frac{\text{tr}(\mathbf{S}_{2n_2})}{\text{tr}(\mathbf{S}_{1n_1})} \right) - p \log \left(\frac{\text{tr}(\boldsymbol{\Sigma}_2)}{\text{tr}(\boldsymbol{\Sigma}_1)} \right) = p \frac{\text{tr}(\boldsymbol{\Sigma}_1)}{\text{tr}(\mathbf{S}_{1n_1})} - p - p \frac{\text{tr}(\boldsymbol{\Sigma}_1)}{\text{tr}(\mathbf{S}_{2n_2})} + p \frac{\text{tr}(\boldsymbol{\Sigma}_1)}{\text{tr}(\boldsymbol{\Sigma}_2)} + o_p(\Delta_\star). \quad (\text{A.9})$$

Note that $p \log(\text{tr}(\boldsymbol{\Sigma}_1)/\text{tr}(\boldsymbol{\Sigma}_2)) = p \text{tr}(\boldsymbol{\Sigma}_1)/\text{tr}(\boldsymbol{\Sigma}_2) - p - p \text{tr}(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2)^2 / (2\text{tr}(\boldsymbol{\Sigma}_2)^2) + o(\Delta_\star)$. Then, by combining (A.8) and (A.9), when $\mathbf{x}_0 \in \pi_1$, it holds from (A-vi) and (A-vii) that

$$\begin{aligned} \omega(\mathbf{x}_0) &= \frac{-2(\mathbf{x}_0 - \boldsymbol{\mu}_1)^T(\bar{\mathbf{x}}_{1n_1} - \boldsymbol{\mu}_1)}{\text{tr}(\boldsymbol{\Sigma}_1)/p} + \frac{2(\mathbf{x}_0 - \boldsymbol{\mu}_1)^T(\bar{\mathbf{x}}_{2n_2} - \boldsymbol{\mu}_2) - \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2}{\text{tr}(\boldsymbol{\Sigma}_2)/p} \\ &\quad + p - p \frac{\text{tr}(\boldsymbol{\Sigma}_1)}{\text{tr}(\boldsymbol{\Sigma}_2)} - p \log \left(\frac{\text{tr}(\boldsymbol{\Sigma}_2)}{\text{tr}(\boldsymbol{\Sigma}_1)} \right) + o_p(\Delta_\star) \\ &= \frac{-2(\mathbf{x}_0 - \boldsymbol{\mu}_1)^T(\bar{\mathbf{x}}_{1n_1} - \boldsymbol{\mu}_1)}{\text{tr}(\boldsymbol{\Sigma}_1)/p} + \frac{2(\mathbf{x}_0 - \boldsymbol{\mu}_1)^T(\bar{\mathbf{x}}_{2n_2} - \boldsymbol{\mu}_2)}{\text{tr}(\boldsymbol{\Sigma}_2)/p} - \frac{\Delta_2}{\text{tr}(\boldsymbol{\Sigma}_2)/p} + o_p(\Delta_\star). \end{aligned} \quad (\text{A.10})$$

First, we consider (A-ii). Let us write that $\mathbf{H}_1^T(\mathbf{x}_0 - \boldsymbol{\mu}_1) = (\lambda_{11}^{1/2} z_{01}, \dots, \lambda_{1p}^{1/2} z_{0p})^T$. Then, we have that $(\mathbf{x}_0 - \boldsymbol{\mu}_1)^T((\bar{\mathbf{x}}_{2n_2} - \boldsymbol{\mu}_2)/(\text{tr}(\boldsymbol{\Sigma}_2)/p) - (\bar{\mathbf{x}}_{1n_1} - \boldsymbol{\mu}_1)/(\text{tr}(\boldsymbol{\Sigma}_1)/p)) = \sum_{j=1}^p \sqrt{\lambda_{1j}} z_{0j} (\mathbf{h}_{1j}^T (\bar{\mathbf{x}}_{2n_2} - \boldsymbol{\mu}_2)/(\text{tr}(\boldsymbol{\Sigma}_2)/p) - \sqrt{\lambda_{1j}} \bar{z}_{1jn_1}/(\text{tr}(\boldsymbol{\Sigma}_1)/p))$, where $\bar{z}_{1jn_1} = \sum_{l=1}^{n_1} z_{1jl}/n_1$. Let

$$v_{1j} = \frac{\sqrt{\lambda_{1j}} z_{0j} (\mathbf{h}_{1j}^T (\bar{\mathbf{x}}_{2n_2} - \boldsymbol{\mu}_2)/(\text{tr}(\boldsymbol{\Sigma}_2)/p) - \sqrt{\lambda_{1j}} \bar{z}_{1jn_1}/(\text{tr}(\boldsymbol{\Sigma}_1)/p))}{\sqrt{(\text{tr}(\boldsymbol{\Sigma}_1)/p)^{-2} \text{tr}(\boldsymbol{\Sigma}_1^2)/n_1 + (\text{tr}(\boldsymbol{\Sigma}_2)/p)^{-2} \text{tr}(\boldsymbol{\Sigma}_1 \boldsymbol{\Sigma}_2)/n_2}}, \quad j = 1, \dots, p.$$

Then, it holds for $j = 2, \dots, p$, that $E_{\boldsymbol{\theta}}(v_{1j}|v_{1j-1}, \dots, v_{11}) = 0$. Note that $\sum_{j=1}^p E_{\boldsymbol{\theta}}(v_{1j}^2) = 1$. We consider applying the martingale central limit theorem. Refer to Section 2.6 in Ghosh et al. (1997) for the details of the martingale central limit theorem. Let $I(\cdot)$ be the indicator function. Note that $E_{\boldsymbol{\theta}}\{(\mathbf{h}_{1j}^T(\bar{\mathbf{x}}_{2n_2} - \boldsymbol{\mu}_2))^4\} = O((\mathbf{h}_{1j}^T \boldsymbol{\Sigma}_2 \mathbf{h}_{1j})^2/n_2^2)$. Then, by using Chebyshev's inequality and Schwarz's inequality, from (A-iv), we have for Lindeberg's condition that

$$\begin{aligned} \sum_{j=1}^p E_{\boldsymbol{\theta}}\{v_{1j}^2 I(v_{1j}^2 > \tau)\} &\leq \tau^{-1} \sum_{j=1}^p E_{\boldsymbol{\theta}}(v_{1j}^4) = \sum_{j=1}^p O\left(\frac{\lambda_{1j}^2 (\mathbf{h}_{1j}^T \boldsymbol{\Sigma}_2 \mathbf{h}_{1j}/n_2 + \lambda_{1j}/n_1)^2}{(\text{tr}(\boldsymbol{\Sigma}_1^2)/n_1 + \text{tr}(\boldsymbol{\Sigma}_1 \boldsymbol{\Sigma}_2)/n_2)^2}\right) \\ &= O\left(\frac{\max_{i=1,2} \text{tr}(\boldsymbol{\Sigma}_i^4)/(\min_{i=1,2} n_i)^2}{(\text{tr}(\boldsymbol{\Sigma}_1^2)/n_1 + \text{tr}(\boldsymbol{\Sigma}_1 \boldsymbol{\Sigma}_2)/n_2)^2}\right) \rightarrow 0 \end{aligned}$$

for any $\tau > 0$. Here, we claim that

$$P_{\boldsymbol{\theta}}\left(\left|\sum_{j=1}^p v_{1j}^2 - 1\right| > \tau\right) = O\left(\frac{\max_{i=1,2} \text{tr}(\boldsymbol{\Sigma}_i^4)/(\min_{i=1,2} n_i)^2}{(\text{tr}(\boldsymbol{\Sigma}_1^2)/n_1 + \text{tr}(\boldsymbol{\Sigma}_1 \boldsymbol{\Sigma}_2)/n_2)^2}\right) \rightarrow 0.$$

Hence, by using the martingale central limit theorem, we obtain that

$$\sum_{j=1}^p v_{1j} \Rightarrow N(0, 1). \quad (\text{A.11})$$

Next, we consider (A-iii) with (A-v). In a way similar to the proof of Theorem 2.2 and Corollary 2.1, we can claim (A.11). Hence, when $\mathbf{x}_0 \in \pi_1$, we obtain from (A.10) that

$$\frac{\omega(\mathbf{x}_0) + \Delta_2/(\text{tr}(\boldsymbol{\Sigma}_2)/p)}{2\sqrt{(\text{tr}(\boldsymbol{\Sigma}_1)/p)^{-2} \text{tr}(\boldsymbol{\Sigma}_1^2)/n_1 + (\text{tr}(\boldsymbol{\Sigma}_2)/p)^{-2} \text{tr}(\boldsymbol{\Sigma}_1 \boldsymbol{\Sigma}_2)/n_2}} \Rightarrow N(0, 1)$$

under (A-ii) or (A-iii) with (A-v). When $\mathbf{x}_0 \in \pi_2$, we have the result in a similar way. Thus the proof is completed. \square

Proof of Theorem 4.2. We first consider the case when $\text{tr}(\boldsymbol{\Sigma}_1)/\text{tr}(\boldsymbol{\Sigma}_2) \neq 1$ as $p \rightarrow \infty$. Noting that $\gamma = o_p(\Delta_*)$, it holds (4.3). Next, we consider the case when $\mathbf{x}_0 \in \pi_1$ and $\text{tr}(\boldsymbol{\Sigma}_1)/\text{tr}(\boldsymbol{\Sigma}_2) \rightarrow 1$ as $p \rightarrow \infty$. We have from (4.4) that

$$\text{tr}(\boldsymbol{\Sigma}_1^2)/n_1 + \text{tr}(\boldsymbol{\Sigma}_1 \boldsymbol{\Sigma}_2)/n_2 \leq \frac{\Delta_L^2}{(z_{\alpha} + z_{\beta})^2}. \quad (\text{A.12})$$

Then, it holds from Theorem 4.1 that

$$\frac{\omega(\mathbf{x}_0) + \gamma}{2\Delta_2/(\text{tr}(\boldsymbol{\Sigma}_2)/p)} \Rightarrow N\left(-1/2 + \frac{\gamma}{2\Delta_2/(\text{tr}(\boldsymbol{\Sigma}_2)/p)}, \frac{\text{tr}(\boldsymbol{\Sigma}_1^2)/n_1 + \text{tr}(\boldsymbol{\Sigma}_1 \boldsymbol{\Sigma}_2)/n_2}{\Delta_2^2}\right).$$

Hence, from (A.12) we have as $p \rightarrow \infty$ that

$$P_{\boldsymbol{\theta}}\left(\frac{\omega(\mathbf{x}_0) + \gamma}{2\Delta_2/(\text{tr}(\boldsymbol{\Sigma}_2)/p)} < 0\right) \geq P_{\boldsymbol{\theta}}\left(N\left(\frac{-z_{\alpha} - z_{\beta}}{2} + \frac{z_{\beta} - z_{\alpha}}{2}, 1\right) < 0\right) + o(1) = 1 - \alpha + o(1).$$

For the case when $\mathbf{x}_0 \in \pi_2$, we obtain the result similarly. Thus, the results follow. \square

Proof of Theorem 4.3. By noting that $\hat{\sigma} = \sigma + o_p((p/m)^{1/2})$ and $\text{tr}(\mathbf{S}_{iN_i}) = \text{tr}(\mathbf{\Sigma}_i) + o_p(p^{1/2})$, $i = 1, 2$, similarly to the proof of Theorems 2.4 and 2.5, it concludes the results. \square

Proof of Theorem 4.4. From the assumption that $\text{tr}(\mathbf{\Sigma}_1^2)/\text{tr}(\mathbf{\Sigma}_2^2) \neq 1$ as $p \rightarrow \infty$, the results can be obtained in a way similar to the proof of Lemmas 1 and 5 in Yata and Aoshima (2009a). We omit further details for brevity. \square

Proof of Theorem 5.1. Let $\boldsymbol{\theta}_j = (\mu_{1j}, \mu_{2j}, \sigma_{(1)j}, \sigma_{(2)j})$, $j = 1, \dots, p$. From the assumption that $E_{\boldsymbol{\theta}_j} \{\exp(t|x_{ijl} - \mu_{ij}|/\sigma_{(i)j}^{1/2})\} < \infty$, $i = 1, 2$ ($j = 1, \dots, p$) for some $t > 0$, for any x satisfying $x \rightarrow \infty$ and $x = o(n_i^{1/2})$ as $n_i \rightarrow \infty$, we claim as $n_i \rightarrow \infty$ that

$$\log \left(P_{\boldsymbol{\theta}_j} \left(\left| \left(\frac{n_i}{\sigma_{(i)j}} \right)^{1/2} (\bar{x}_{ijn_i} - \mu_{ij}) \right| > x \right) \right) = -x^2/2 + o(x^2).$$

Refer to Section 4 in Shao (2005) for the details of this result. From (5.4), it holds that $(n_i \delta / \sigma_{(i)j}) \geq 2(\log p)^{1+\zeta} (\sigma_{(1)j}^{1/2} + \sigma_{(2)j}^{1/2}) / \sigma_{(i)j}^{1/2} > 2(\log p)^{1+\zeta}$. Note that $\sqrt{2(\log p)^{1+\zeta}}/2 > \sqrt{2(\log p)^{1+\zeta/2}} = o(n_i^{1/2})$ as $p \rightarrow \infty$. Thus we have that

$$\begin{aligned} & P_{\boldsymbol{\theta}_j} \left(\left| \left(\frac{n_i}{\sigma_{(i)j}} \right)^{1/2} (\bar{x}_{ijn_i} - \mu_{ij}) \right| > \left(\frac{n_i}{\sigma_{(i)j}} \right)^{1/2} \sqrt{\delta}/2 \right) \\ & < P_{\boldsymbol{\theta}_j} \left(\left| \left(\frac{n_i}{\sigma_{(i)j}} \right)^{1/2} (\bar{x}_{ijn_i} - \mu_{ij}) \right| > \sqrt{2(\log p)^{1+\zeta/2}} \right) = \exp(-(\log p)^{1+\zeta/2}(1+o(1))) \\ & \leq \exp(-(\log p)^{1+\zeta/4}) = o(p^{-1}). \end{aligned} \quad (\text{A.13})$$

Then, it holds that

$$P_{\boldsymbol{\theta}_j} \left(|T_{j(\mathbf{n})} - (\mu_{1j} - \mu_{2j})| > \sqrt{\delta} \right) \leq \sum_{i=1}^2 P_{\boldsymbol{\theta}_j} \left(|\bar{x}_{ijn_i} - \mu_{ij}| > \sqrt{\delta}/2 \right) = o(p^{-1}). \quad (\text{A.14})$$

Thus by using Bonferroni's inequality, we claim that

$$P_{\boldsymbol{\theta}}(|\mathbf{D}^c \cap \widehat{\mathbf{D}}| = 0) \geq 1 - \sum_{j \in \mathbf{D}^c} P_{\boldsymbol{\theta}_j} \left(|T_{j(\mathbf{n})}| > \sqrt{\delta} \right) = 1 + o(1).$$

Next, we have for $j \in \mathbf{D}$ that

$$\begin{aligned} P_{\boldsymbol{\theta}_j} \left(|T_{j(\mathbf{n})}| > \sqrt{\delta} \right) & \geq P_{\boldsymbol{\theta}_j} \left(\left| |T_{j(\mathbf{n})} - (\mu_{1j} - \mu_{2j})| - |\mu_{1j} - \mu_{2j}| \right| > \sqrt{\delta} \right) \\ & \geq P_{\boldsymbol{\theta}_j} \left(|T_{j(\mathbf{n})} - (\mu_{1j} - \mu_{2j})| < |\mu_{1j} - \mu_{2j}| - \sqrt{\delta} \right) = 1 + o(1) \end{aligned}$$

by noting that $T_{j(\mathbf{n})} - (\mu_{1j} - \mu_{2j}) = o_p(1)$ and $|\mu_{1j} - \mu_{2j}| > \sqrt{\delta}$. Thus we have that $|\widehat{\mathbf{D}} \cap \mathbf{D}| = S(1 + o_p(1))$. The results follow. \square

Proof of Theorem 5.2. In a way similar to the proof of Theorem 5.1, we have for $j \in \mathbf{D}$ that

$$P_{\boldsymbol{\theta}_j} \left(|T_{j(m)}| > \sqrt{\delta} \right) = 1 + o(1).$$

Thus it holds that $|\tilde{\mathbf{D}} \cap \mathbf{D}| = S(1 + o_p(1))$. Let j_{i*} denote a subscript such that $\sigma_{(i)j_{i*}} = \max_{j \in \tilde{\mathbf{D}}} \sigma_{(i)j}$. We have as $m \rightarrow \infty$ that

$$\sqrt{\tilde{s}_i} \sum_{k=1}^2 \sqrt{\tilde{s}_k} \geq \sqrt{s_{ij_{i*}}} \sum_{k=1}^2 \sqrt{s_{kj_{k*}}} = \sqrt{\sigma_{(i)j_{i*}}} \sum_{k=1}^2 \sqrt{\sigma_{(k)j_{k*}}} (1 + o_p(1)).$$

Thus we claim that

$$N_i \geq \frac{2 \max\{(\log \tilde{S})^{1+\xi}, (\log p)^\varepsilon\}}{\delta} \sqrt{\sigma_{(i)j_{i*}}} \sum_{k=1}^2 \sqrt{\sigma_{(k)j_{k*}}} (1 + o_p(1)). \quad (\text{A.15})$$

Note that N_i and x_{ijm+l} ($j \in \tilde{\mathbf{D}}$) are independent for each l ($= 1, \dots, N_i$). Note that $\max\{(\log \tilde{S})^{1+\xi}, (\log p)^\varepsilon\} \rightarrow \infty$ as $p \rightarrow \infty$ either when $\tilde{S} = O(1)$ or $\tilde{S} \rightarrow \infty$. By combining (A.14) with (A.15), we have for $j \in \tilde{\mathbf{D}} \cap \mathbf{D}^c$ that

$$P_{\boldsymbol{\theta}}(|T_{j(\mathbf{N})}| > \sqrt{\delta} | \tilde{\mathbf{D}}) = o(1/\tilde{S}).$$

Thus we have that

$$P_{\boldsymbol{\theta}}(|\hat{\mathbf{D}} \cap \mathbf{D}^c| = 0) \geq 1 - E_{\boldsymbol{\theta}} \left(\sum_{j \in \hat{\mathbf{D}} \cap \mathbf{D}^c} P_{\boldsymbol{\theta}}(|T_{j(\mathbf{N})}| > \sqrt{\delta} | \tilde{\mathbf{D}}) \right) = 1 + o(1).$$

Next, from the fact that $\max\{(\log \tilde{S})^{1+\xi}, (\log p)^\varepsilon\} \rightarrow \infty$, it holds for $j \in \tilde{\mathbf{D}} \cap \mathbf{D}$ that

$$P_{\boldsymbol{\theta}_j}(|T_{j(\mathbf{N})}| > \sqrt{\delta}) = 1 + o(1).$$

Thus we have that $|\hat{\mathbf{D}} \cap \mathbf{D}| = S(1 + o_p(1))$. The results follow. \square

Proof of Theorem 7.1. By using Lemma A.2 in Fan and Fan (2008), we note that $\max_{j=1, \dots, p} |s_{(i)jn_i}/\sigma_{(i)j} - 1| = o_p(1)$ for $\log p = o(n_i)$. Then, similarly to (A.13)-(A.14), we have from (7.2) that

$$\begin{aligned} P_{\boldsymbol{\theta}}(|\mathbf{D}^c \cap \hat{\mathbf{D}}| = 0) &\geq 1 - \sum_{j \in \mathbf{D}^c} P_{\boldsymbol{\theta}_j} \left(\frac{|T_{j(\mathbf{n})}|}{\sqrt{s_{(1)jn_1} + s_{(2)jn_2}}} > \sqrt{\delta}/2 \right) \\ &= 1 - \sum_{j \in \mathbf{D}^c} P_{\boldsymbol{\theta}_j} \left(\frac{|T_{j(\mathbf{n})}|}{\sqrt{\sigma_{(1)j} + \sigma_{(2)j}}} > \sqrt{\delta}(1 + o_p(1))/2 \right) = 1 + o(1). \end{aligned} \quad (\text{A.16})$$

On the other hand, we have for $j \in \mathbf{D}$ that

$$P_{\boldsymbol{\theta}_j} \left(\frac{|T_{j(\mathbf{n})}|}{\sqrt{s_{(1)jn_1} + s_{(2)jn_2}}} \leq \sqrt{\delta}/2 \right) \leq P_{\boldsymbol{\theta}_j} \left(\frac{|T_{j(\mathbf{n})} - (\mu_1 - \mu_2)|}{\sqrt{\sigma_{(1)j} + \sigma_{(2)j}}} > \sqrt{\delta}(1 + o_p(1))/2 \right) = o(p).$$

Thus, it concludes the results similarly to (A.16). \square

Proof of Theorem 7.2. From the assumption that $\min_{j \in \mathbf{D}} (\sigma_{(1)j} + \sigma_{(2)j})^{-1} (\mu_{1j} - \mu_{2j})^2 \geq \delta$, it holds $\widehat{\mathbf{D}} = \mathbf{D}$ w.p.1. We first consider the case when $\mathbf{x}_0 \in \pi_1$. Note that $\text{tr}(\boldsymbol{\Sigma}_{1(S)} + \boldsymbol{\Sigma}_{2(S)}) \leq \|\boldsymbol{\mu}_{1(S)} - \boldsymbol{\mu}_{2(S)}\|^2 / \delta$. Thus it holds for $i = 1, 2$ that $\text{tr}(\boldsymbol{\Sigma}_{i(S)}^2) / (\|\boldsymbol{\mu}_{1(S)} - \boldsymbol{\mu}_{2(S)}\|^4 n_i) \leq \text{tr}(\boldsymbol{\Sigma}_{i(S)})^2 / (\|\boldsymbol{\mu}_{1(S)} - \boldsymbol{\mu}_{2(S)}\|^4 n_i) \rightarrow 0$ as $p \rightarrow \infty$. Let $\mathbf{x}_{(S)} = f_{\mathbf{D}}(\mathbf{x})$ for any vector $\mathbf{x} = (x_1, \dots, x_p)^T$. Then, we have for $i = 1, 2$ that

$$\begin{aligned} \text{Var}_{\boldsymbol{\theta}} \left(\frac{\|\bar{\mathbf{x}}_{in_i(S)} - \boldsymbol{\mu}_{i(S)}\|^2 - \text{tr}(\mathbf{S}_{in_i(S)})/n_i}{\|\boldsymbol{\mu}_{1(S)} - \boldsymbol{\mu}_{2(S)}\|^2} \right) &= O \left(\frac{\text{tr}(\boldsymbol{\Sigma}_{i(S)}^2)}{n_i^2 \|\boldsymbol{\mu}_{1(S)} - \boldsymbol{\mu}_{2(S)}\|^4} \right) \rightarrow 0; \\ \text{Var}_{\boldsymbol{\theta}} \left(\frac{(\bar{\mathbf{x}}_{1n_1(S)} - \boldsymbol{\mu}_{1(S)})^T (\bar{\mathbf{x}}_{2n_2(S)} - \boldsymbol{\mu}_{2(S)})}{\|\boldsymbol{\mu}_{1(S)} - \boldsymbol{\mu}_{2(S)}\|^2} \right) &= O \left(\frac{\text{tr}(\boldsymbol{\Sigma}_{1(S)} \boldsymbol{\Sigma}_{2(S)})}{n_1 n_2 \|\boldsymbol{\mu}_{1(S)} - \boldsymbol{\mu}_{2(S)}\|^4} \right) \rightarrow 0; \\ \text{Var}_{\boldsymbol{\theta}} \left(\frac{(\mathbf{x}_{0(S)} - \boldsymbol{\mu}_{1(S)})^T (\bar{\mathbf{x}}_{in_i(S)} - \boldsymbol{\mu}_{i(S)})}{\|\boldsymbol{\mu}_{1(S)} - \boldsymbol{\mu}_{2(S)}\|^2} \right) &= O \left(\frac{\text{tr}(\boldsymbol{\Sigma}_{i(S)}^2)}{n_i \|\boldsymbol{\mu}_{1(S)} - \boldsymbol{\mu}_{2(S)}\|^4} \right) \rightarrow 0; \\ \text{Var}_{\boldsymbol{\theta}} \left(\frac{(\mathbf{x}_{0(S)} - \boldsymbol{\mu}_{1(S)})^T (\boldsymbol{\mu}_{1(S)} - \boldsymbol{\mu}_{2(S)})}{\|\boldsymbol{\mu}_{1(S)} - \boldsymbol{\mu}_{2(S)}\|^2} \right) &= O \left(\frac{(\boldsymbol{\mu}_{1(S)} - \boldsymbol{\mu}_{2(S)})^T \boldsymbol{\Sigma}_{1(S)} (\boldsymbol{\mu}_{1(S)} - \boldsymbol{\mu}_{2(S)})}{\|\boldsymbol{\mu}_{1(S)} - \boldsymbol{\mu}_{2(S)}\|^4} \right) \rightarrow 0, \end{aligned}$$

where $\bar{\mathbf{x}}_{in_i(S)} = \sum_{l=1}^{n_i} \mathbf{x}_{il(S)} / n_i$ and $\mathbf{S}_{in_i(S)} = \sum_{l=1}^{n_i} (\mathbf{x}_{il(S)} - \bar{\mathbf{x}}_{in_i(S)}) (\mathbf{x}_{il(S)} - \bar{\mathbf{x}}_{in_i(S)})^T / (n_i - 1)$. Then, from the fact that $\widehat{\mathbf{D}} = \mathbf{D}$ w.p.1, we claim for $i = 1, 2$ that

$$\begin{aligned} \frac{\|\bar{\mathbf{x}}_{in_i} - \boldsymbol{\mu}_{i(S)}\|^2 - \text{tr}(\widehat{\mathbf{S}}_{in_i})/n_i}{\|\boldsymbol{\mu}_{1(S)} - \boldsymbol{\mu}_{2(S)}\|^2} &= o_p(1), \quad \frac{(\bar{\mathbf{x}}_{1n_1} - \boldsymbol{\mu}_{1(S)})^T (\bar{\mathbf{x}}_{2n_2} - \boldsymbol{\mu}_{2(S)})}{\|\boldsymbol{\mu}_{1(S)} - \boldsymbol{\mu}_{2(S)}\|^2} = o_p(1), \\ \frac{(\widehat{\mathbf{x}}_0 - \boldsymbol{\mu}_{1(S)})^T (\bar{\mathbf{x}}_{in_i} - \boldsymbol{\mu}_{i(S)})}{\|\boldsymbol{\mu}_{1(S)} - \boldsymbol{\mu}_{2(S)}\|^2} &= o_p(1), \quad \frac{(\widehat{\mathbf{x}}_0 - \boldsymbol{\mu}_{1(S)})^T (\boldsymbol{\mu}_{1(S)} - \boldsymbol{\mu}_{2(S)})}{\|\boldsymbol{\mu}_{1(S)} - \boldsymbol{\mu}_{2(S)}\|^2} = o_p(1). \end{aligned}$$

Hence, we have that

$$\begin{aligned} &\left(\left(\widehat{\mathbf{x}}_0 - \frac{\bar{\mathbf{x}}_{1n_1} + \bar{\mathbf{x}}_{2n_2}}{2} \right)^T (\bar{\mathbf{x}}_{2n_2} - \bar{\mathbf{x}}_{1n_1}) - \frac{\text{tr}(\widehat{\mathbf{S}}_{1n_1})}{2n_1} + \frac{\text{tr}(\widehat{\mathbf{S}}_{2n_2})}{2n_2} \right) / \|\boldsymbol{\mu}_{1(S)} - \boldsymbol{\mu}_{2(S)}\|^2 \\ &= -1/2 + o_p(1) < 0 \end{aligned}$$

w.p.1. For the case when $\mathbf{x}_0 \in \pi_2$, we obtain the result similarly. Thus the proof is completed. \square

Proof of Theorem 8.1. Let us write that

$$\begin{aligned} w_{i(1)} &= \sum_{j=1}^{n_{(1)}} \frac{(x_{j(*)} - \mu_*) \lambda_i^{1/2} z_{ij}}{n_{(1)} - 1} - \frac{n_{(1)}}{n_{(1)} - 1} (\bar{x}_{n_{(1)}(*)} - \mu_*) \lambda_i^{1/2} \sum_{j=1}^{n_{(1)}} z_{ij} / n_{(1)}, \\ w_{i(2)} &= \sum_{j=n_{(1)}+1}^n \frac{(x_{j(*)} - \mu_*) \lambda_i^{1/2} z_{ij}}{n_{(2)} - 1} - \frac{n_{(2)}}{n_{(2)} - 1} (\bar{x}_{n_{(2)}(*)} - \mu_*) \lambda_i^{1/2} \sum_{j=n_{(1)}+1}^n z_{ij} / n_{(2)}. \end{aligned}$$

Then, we can write that $\widehat{T}\boldsymbol{\sigma} = \sum_{i=1}^p w_{i(1)} w_{i(2)}$. When $\boldsymbol{\rho} = \mathbf{0}$, from the assumption, \mathbf{x}_j and $\mathbf{x}_{j(*)}$ are independent. Thus we have that $\text{Var}(w_{i(1)} w_{i(2)}) = ((n_{(1)} - 1)(n_{(2)} - 1))^{-1} \sigma_*^2 \lambda_i^2$.

Then, we claim for $i = 2, \dots, p$, that $E(w_{i(1)}w_{i(2)}|w_{i-1(1)}w_{i-1(2)}, \dots, w_{1(1)}w_{1(2)}) = 0$. Note that

$$P_{\theta} \left(\left| \sum_{j=1}^p \frac{w_{i(1)}^2 w_{i(2)}^2}{((n_{(1)} - 1)(n_{(2)} - 1))^{-1} \sigma_*^2 \text{tr}(\Sigma^2)} - 1 \right| > \tau \right) = O \left(\frac{\text{tr}(\Sigma^4)}{\text{tr}(\Sigma^2)^2} \right) + o(1) \rightarrow 0$$

as $p \rightarrow \infty$ and $n \rightarrow \infty$ for any $\tau > 0$. Hence, similarly to (A.11), from the martingale central limit theorem, the result follows. \square

ACKNOWLEDGMENTS

The authors would like to thank four anonymous referees for their valuable comments and careful reading which have led to improvements in our presentation and results of the paper. The authors are grateful to the Editor-in-Chief, Prof. Nitis Mukhopadhyay, for his helpful remarks on improving the layout of this manuscript. Research of the first author was partially supported by Grant-in-Aid for Scientific Research (B), Japan Society for the Promotion of Science (JSPS), under Contract Number 22300094.

REFERENCES

- Ahn, J., Marron, J. S., Muller, K. M., and Chi, Y.-Y. (2007). The High-Dimension, Low-Sample-Size Geometric Representation Holds Under Mild Conditions, *Biometrika* 94: 760-766.
- Aoshima, M. (2005). Statistical Inference in Two-Stage Sampling, *American Mathematical Society Translations Series* 215: 125-145.
- Aoshima, M., Chen, P., and Panchapakesan, S. (2003). Sequential Procedures for Selecting the Most Probable Multinomial Cell When a Nuisance Cell is Present, *Communications in Statistics-Theory & Methods* 32: 893-906.
- Aoshima, M. and Mukhopadhyay, N. (1998). Fixed-Width Simultaneous Confidence Intervals for Multinormal Means in Several Intraclass Correlation Models, *Journal of Multivariate Analysis* 66: 46-63.
- Aoshima, M. and Takada, Y. (2004). Asymptotic Second-Order Efficiency for Multivariate Two-Stage Estimation of a Linear Function of Normal Mean Vectors, *Sequential Analysis* 23: 333-353.
- Aoshima, M., Takada, Y., and Srivastava, M. S. (2002). A Two-Stage Procedure for Estimating a Linear Function of k Multinormal Mean Vectors When Covariance Matrices are Unknown, *Journal of Statistical Planning and Inference* 100: 109-119.
- Aoshima, M. and Yata, K. (2010). Asymptotic Second-Order Consistency for Two-Stage Estimation Methodologies and Its Applications, *Annals of Institute of Statistical Mathematics* 62: 571-600.
- Bai, Z. and Sarandasa, H. (1996). Effect of High Dimension: By an Example of a Two Sample Problem, *Statistica Sinica* 6: 311-329.
- Baik, J., Ben Arous, G., and Péché, S. (2005). Phase Transition of the Largest Eigenvalue for Non-Null Complex Covariance Matrices, *Annals of Probability* 33: 1643-1697.
- Baik, J. and Silverstein, J. W. (2006). Eigenvalues of Large Sample Covariance Matrices of Spiked Population Models, *Journal of Multivariate Analysis* 97: 1382-1408.
- Bradley, R. C. (2005). Basic Properties of Strong Mixing Conditions. A Survey and Some Open Questions, *Probability Surveys* 2: 107-144 (electronic).
- Chen, S. X. and Qin, Y.-L. (2010). A Two-Sample Test for High-Dimensional Data with Applications to Gene-Set Testing, *Annals of Statistics* 38: 808-835.

- Chiaretti, S., Li, X., Gentleman, R., Vitale, A., Vignetti, M., Mandelli, F., Ritz, J., and Foa, R. (2004). Gene Expression Profile of Adult T-cell Acute Lymphocytic Leukemia Identifies Distinct Subsets of Patients with Different Response to Therapy and Survival, *Blood* 103: 2771-2778.
- Drton, M. and Perlman, M. D. (2007). Multiple Testing and Error Control in Gaussian Graphical Model Selection, *Statistical Science* 22: 430-449.
- Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data, *Journal of American Statistical Association* 97: 77-87.
- Fan, J. and Fan, Y. (2008). High-Dimensional Classification Using Features Annealed Independence Rules, *Annals of Statistics* 36: 2605-2637.
- Ghosh, M., Mukhopadhyay, N., and Sen, P. K. (1997). *Sequential Estimation*, New York: Wiley.
- Hall, P., Marron, J. S., and Neeman, A. (2005). Geometric Representation of High Dimension, Low Sample Size Data, *Journal of Royal Statistical Society, Series B* 67: 427-444.
- Hall, P., Pittelkow, Y., and Ghosh, M. (2008). Theoretical Measures of Relative Performance of Classifiers for High Dimensional Data with Small Sample Sizes, *Journal of Royal Statistical Society, Series B* 70: 159-173.
- Ibragimov, I. A. (1975). A Note on the Central Limit Theorem for Dependent Random Variables, *Theory of Probability and Its Applications* 20: 135-141.
- Johnstone, I. M. (2001). On the Distribution of the Largest Eigenvalue in Principal Components Analysis, *Annals of Statistics* 29: 295-327.
- Jung S. and Marron J. S. (2009). PCA Consistency in High Dimension, Low Sample Size Context, *Annals of Statistics* 37: 4104-4130.
- Kolmogorov, A. N. and Rozanov, Y. A. (1960). On Strong Mixing Conditions for Stationary Gaussian Processes, *Theory of Probability and Its Applications* 5: 204-208.
- Kraft P., Schadt E., Aten J., and Horvath S. (2003). A Family-Based Test for Correlation between Gene Expression and Trait Values, *American Journal of Human Genetics* 72: 1323-1330.
- Meinshausen, N., Meier, L., and Bühlmann, P. (2009). p -Values for High-Dimensional Regression, *Journal of American Statistical Association* 104: 1671-1681.
- Mukhopadhyay, N. and Duggan, W. T. (1997). Can a Two-Stage Procedure Enjoy Second-Order Properties?, *Sankhyā, Series A* 59: 435-448.
- Mukhopadhyay, N. and Duggan, W. T. (1999). On a Two-Stage Procedure Having Second-Order Properties with Applications, *Annals of Institute of Statistical Mathematics* 51: 621-636.
- Paul, D. (2007). Asymptotics of Sample Eigenstructure for a Large Dimensional Spiked Covariance Model, *Statistica Sinica* 17: 1617-1642.
- Pollard, K. S., Dudoit, S., and van der Laan, M. J. (2005). Multiple Testing Procedures: R multtest Package and Applications to Genomics, *In Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, R. Gentleman, V. Carey, W. Huber, R. Irizarry, and S. Dudoit (eds.), pp. 249-271, New York: Springer.
- Saranadasa, H. (1993). Asymptotic Expansion of the Misclassification Probabilities of D-and A-criteria for Discrimination from Two High Dimensional Populations Using the Theory of Large Dimensional Random Matrices, *Journal of Multivariate Analysis* 46: 154-174.
- Sen, P. K., Tsai, M.-T., and Jou, Y.-S. (2007). High-Dimension, Low-Sample Size Perspectives in Constrained Statistical Inference: The SARSCov RNA Genome in Illustration, *Journal of American Statistical Association* 102: 686-694.

- Shao, Q. M. (2005). Self-Normalized Limit Theorems in Probability and Statistics, Manuscript, Personal communication.
- Srivastava, M. S. (2005). Some Tests Concerning the Covariance Matrix in High-Dimensional Data, *Journal of Japan Statistical Society* 35: 251-272.
- Srivastava, M. S. (2007). Multivariate Theory for Analyzing High Dimensional Data, *Journal of Japan Statistical Society* 37: 53-86.
- Srivastava, M. S. and Kubokawa, T. (2007). Comparison of Discrimination Methods for High Dimensional Data, *Journal of Japan Statistical Society* 37: 123-134.
- Stein, C. (1945). A Two-Sample Test for a Linear Hypothesis Whose Power is Independent of the Variance, *Annals of Mathematical Statistics* 16: 243-258.
- Wasserman, L. and Roeder, K. (2009). High-Dimensional Variable Selection, *Annals of Statistics* 37: 2178-2201.
- Yata, K. (2010). Effective Two-Stage Estimation for a Linear Function of High-Dimensional Gaussian Means, *Sequential Analysis* 29: 463-482.
- Yata, K. and Aoshima, M. (2009a). Double Shrink Methodologies to Determine the Sample Size via Covariance Structures, *Journal of Statistical Planning and Inference* 139: 81-99.
- Yata, K. and Aoshima, M. (2009b). PCA Consistency for Non-Gaussian Data in High Dimension, Low Sample Size Context, *Communications in Statistics-Theory & Methods*, Special Issue Honoring S. Zacks (ed., N. Mukhopadhyay) 38: 2634-2652.
- Yata, K. and Aoshima, M. (2010a). Effective PCA for High-Dimension, Low-Sample-Size Data with Singular Value Decomposition of Cross Data Matrix, *Journal of Multivariate Analysis* 101: 2060-2077.
- Yata, K. and Aoshima, M. (2010b). Intrinsic Dimensionality Estimation of High Dimension, Low Sample Size Data with d -asymptotics, *Communications in Statistics-Theory & Methods*, Special Issue Honoring M. Akahira (ed. M. Aoshima) 39: 1511-1521.
- Yata, K. and Aoshima, M. (2010c). Effective PCA for High-Dimension, Low-Sample-Size Data with Noise Reduction via Geometric Representations, *Journal of Multivariate Analysis*, revised.