Two-Sample Tests Based on Eigenstructures in the HDLSS Context

Aki Ishii

February 2017

Two-Sample Tests Based on Eigenstructures in the HDLSS Context

Aki Ishii
Doctoral Program in Mathematics

Submitted to the Graduate School of
Pure and Applied Sciences
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy in
Science
at the
University of Tsukuba

Contents

Preface

One of the features of modern data is the data has a high dimension and a low sample size. We call such data "HDLSS" or "large p, small n" data where $p/n \to \infty$; here p is the data dimension and n is the sample size. One can see HDLSS data in various areas of modern science such as genetic microarrays, medical imaging, finance, chemometrics and so on. When we analyze HDLSS data, how should we treat this type of data? We have a lot of theories and methodologies in multivariate analysis, however, we cannot apply multivariate analysis to HDLSS data without consideration because multivariate analysis is based on the large sample theory. We have to construct new theories and methodologies for HDLSS data.

Aoshima and Yata [2, 3] gave a broad perspective of high-dimensional statistical analysis such as given-bandwidth confidence region, two-sample test, test of equality of two covariance matrices, classification, variable selection, regression, pathway analysis and so on along with sample size determination to ensure prespecified accuracy for each inference. In addition, Aoshima and Yata [4, 5] gave review articles covering this field of research. Aoshima and Yata [7] developed the theory of asymptotic normality in order to ensure the accuracy for HDLSS data under mild conditions. As for the two-sample test, Aoshima and Yata [9] discussed the optimality of the two-sample test for HDLSS data and created new test procedures based on the eigenstructure of HDLSS data when $p \to \infty$ and $n \to \infty$. As for the classification problem, Aoshima and Yata [6] gave the distance-based classifier and developed the misclassification rate adjusted classification which controls misclassification rates. Aoshima and Yata [8] gave the geometric classifier which discriminates the classes by using the heteroscedasticity in addition to the difference of means. As for the pathway analysis, Yata and Aoshima [39, 41] considered tests of the correlation matrix. As for the noise of HDLSS data, the asymptotic behaviors of HDLSS data were studied by Hall et al. [18], Ahn et al. [1], and Yata and Aoshima [38] when $p \to \infty$ while n is fixed. They found several geometric representations of HDLSS data under some conditions. The HDLSS asymptotic study usually assumes either the normality as the population distribution or a ρ-mixing condition as the dependency of random variables in a sphered data matrix. See Jung and Marron [25]. In a more general framework, Yata and Aoshima [35] showed that the conventional principal component analysis (PCA) cannot give consistent estimators of eigenvalues and eigenvectors in the HDLSS context. In order to overcome this inconvenience, Yata and Aoshima [38] developed the noise-reduction (NR) methodology for Gaussian type HDLSS data. Moreover, Yata and Aoshima [36, 37, 40] created the cross-data-matrix (CDM) methodology for non-Gaussian type HDLSS data and in-

vestigated its asymptotic properties throughly when $p \to \infty$ and $n \to \infty$. Yata and Aoshima [42] considered the reconstruction of a low-rank signal matrix for HDLSS data by using the methods.

In this thesis, we consider the two-sample problem for HDLSS data when $p \to \infty$ while n is fixed. We investigate the eigenstructure of HDLSS data theoretically, and give new two-sample test procedures based on the eigenstructure of HDLSS data. This thesis consists of four chapters.

In Chapter 1, we consider the estimation of the first (largest) eigenvalue. We summarize the findings by Ishii et al. [19, 20] and Ishii [22]. The key point is the geometric representation of the noise space for HDLSS data. In Section 2, we introduce the geometric representation given by Ishii et al. [19]. In Section 3, we show that the conventional estimator does not work well for HDLSS data. According to Ishii et al. [20] and Ishii [22], we provide asymptotic properties of the estimators given by the NR method and the CDM method when $p \to \infty$ while n is fixed. In Section 4, we discuss the performance of the estimators numerically.

In Chapter 2, we consider applications of the first eigenvalue. We summarize the findings by Ishii et al. [20]. In Section 2, we construct the confidence interval of the first contribution ratio. We apply the result to actual microarray data sets. In Section 3, we consider the estimation of the first eigenvector. We show that the conventional estimator leads to the inconsistency in the HDLSS context. We give asymptotic properties of the NR estimator when $p \to \infty$ while n is fixed. In Section 4, we consider the estimation of the first PC score. In Section 5, we consider the one-sample test for a mean vector. Finally, we discuss the performance of the estimators numerically.

In Chapter 3, we consider the equality test of two covariance matrices. We summarize the findings by Ishii et al. [20] and Ishii [22]. In Section 2, we consider the equality test of the first eigenvalues between two classes. In Section 3, we consider the equality test of the first eigenspaces between two classes. By using the test procedure given in this section, one can check the validity of the assumption required in Chapter 4. In Section 4, we construct the equality test of two covariance matrices between two classes. By using the test procedure given in this section, one can distinguish two high-dimensional covariance matrices even when the sample sizes are fixed. Finally, we apply our test procedures to actual microarray data sets.

In Chapter 4, we consider the two-sample test for HDLSS data. We summarize the findings by Ishii [21, 22]. A lot of papers consider this premier problem, however, they usually assume the equality of two covariance matrices from technical reasons. We emphasize that assuming the equality of two covariance matrices is quite unrealistic in actual data analyses. We rather utilize the difference of the two covariance matrices and construct test procedures based on the eigenstructures. In Section 2, we introduce the test procedure given by Ishii [21] for Gaussian type HDLSS data. In Section 3, we introduce the test procedure given by Ishii [22] for non-Gaussian type HDLSS data. In Section 4, we discuss the performance of the test procedures numerically. Finally, we apply our test procedures to actual microarray data sets.

Acknowledgements

I would like to express my deepest appreciation to my supervisor, Professor Makoto Aoshima. He always encouraged me and gave his enthusiastic guidance, helpful support and patience to me. This dissertation would not have been possible without his advice and stimulating comments. He also provided me the tremendous research environment. I could engage in my research project strenuously.

I would also like to thank Associate Professor Kazuyoshi Yata for his variable comments.

Finally, I am deeply grateful to my parents for their patience and support.

Chapter 1

Estimation of the First Eigenvalue in the HDLSS Context

In this chapter, we consider estimation of the first eigenvalue in the HDLSS context. This chapter is organized by Ishii et al. [19, 20] and Ishii [22].

In Section 2, we consider geometric representations of HDLSS data. The asymptotic behaviors of HDLSS data were studied by Hall et al. [18], Ahn et al. [1], and Yata and Aoshima [38] when $p \to \infty$ while n is fixed. Hall et al. [18] discussed a geometric representation of high-dimensional data vectors themselves. On the other hand, Ahn et al. [1], and Yata and Aoshima [38] discussed geometric representations of HDLSS data in a dual space. In this section, we first consider the case when the population mean is known and introduce previous studies about geometric representations of HDLSS data in a dual space. Next, according to Ishii et al. [19], we give another geometric representation of HDLSS data in a dual space when the population mean is unknown.

In Section 3, we consider the estimation of the first eigenvalue of population covariance matrix. The first eigenvalue is quite important for high-dimensional data and it often becomes much larger than the other eigenvalues. We first show that the conventional estimator cannot estimate the first eigenvalue correctly in the HDLSS context. In order to overcome this inconvenience, we introduce two estimators given by using the NR method and the CDM method. We show that the NR estimator has asymptotic properties under a mild condition and so does the bias corrected CDM estimator under a more relaxed condition.

Finally, in Section 4, we summarize simulation studies and discuss the performances of the findings.

1 Introduction

Suppose we have a $p \times n$ data matrix, $\boldsymbol{X}_{(p)} = [\boldsymbol{x}_{1(p)}, ..., \boldsymbol{x}_{n(p)}]$, where $\boldsymbol{x}_{j(p)} = (x_{1j(p)}, ..., x_{pj(p)})^T$, $j = 1, ..., n$, are independent and identically distributed (i.i.d.) as a p-dimensional distribution with mean vector $\boldsymbol{\mu}_p$ and covariance matrix $\boldsymbol{\Sigma}_p$ (≥ 0). We assume $n \geq 4$. The eigen-decomposition of $\boldsymbol{\Sigma}_p$ is given by $\boldsymbol{\Sigma}_p = \boldsymbol{H}_p \boldsymbol{\Lambda}_p \boldsymbol{H}_p^T$, where $\boldsymbol{\Lambda}_p$ is a diagonal matrix of eigenvalues, $\lambda_{1(p)} \geq \cdots \geq \lambda_{p(p)} (\geq 0)$, and $\boldsymbol{H}_p = [\boldsymbol{h}_{1(p)}, ..., \boldsymbol{h}_{p(p)}]$ is an orthogonal matrix of the corresponding eigenvectors. Let $\boldsymbol{X}_{(p)} - [\boldsymbol{\mu}_p, ..., \boldsymbol{\mu}_p] = \boldsymbol{H}_p \boldsymbol{\Lambda}_p^{1/2} \boldsymbol{Z}_{(p)}$. Then, $\boldsymbol{Z}_{(p)}$ is a $p \times n$ sphered data matrix from a distribution with the zero mean and the identity covariance matrix. Here, we write $\boldsymbol{Z}_{(p)} = [\boldsymbol{z}_{1(p)}, ..., \boldsymbol{z}_{p(p)}]^T$ and $\boldsymbol{z}_{j(p)} = (z_{j1(p)}, ..., z_{jn(p)})^T$, $j = 1, ..., p$. Note that $E(z_{ji(p)} z_{j'i(p)}) = 0$ $(j \neq j')$ and $\mathrm{Var}(\boldsymbol{z}_{j(p)}) = \boldsymbol{I}_n$, where $\boldsymbol{I}_n$ is the n-dimensional identity matrix. Hereafter, the subscript p will be omitted for the sake of simplicity when it does not cause any confusion. We assume that the fourth moments of each variable in $\boldsymbol{Z}$ are uniformly bounded. Note that if $\boldsymbol{X}$ is Gaussian, z_{ij}s are i.i.d. as $N(0, 1)$, where $N(0, 1)$ denotes the standard normal distribution.

2 Geometric Representations in a Dual Space

In this section, we consider geometric representations for Gaussian-type HDLSS data when $p \to \infty$ while n is fixed.

2.1 When the population mean vector is known

We assume $\boldsymbol{\mu} = \boldsymbol{0}$ without loss of generality. Let us write the sample covariance matrix as $\boldsymbol{S}_o = n^{-1} \boldsymbol{X} \boldsymbol{X}^T$. Then, we define the $n \times n$ dual sample covariance matrix by $\boldsymbol{S}_{oD} = n^{-1} \boldsymbol{X}^T \boldsymbol{X}$. Let $\hat{\lambda}_{o1} \geq \cdots \geq \hat{\lambda}_{on} \geq 0$ be the eigenvalues of $\boldsymbol{S}_{oD}$. Then, we define the eigen-decomposition of $\boldsymbol{S}_{oD}$ by $\boldsymbol{S}_{oD} = \sum_{j=1}^n \hat{\lambda}_{oj} \hat{\boldsymbol{u}}_{oj} \hat{\boldsymbol{u}}_{oj}^T$, where $\hat{\boldsymbol{u}}_{oj}$ denotes a unit eigenvector corresponding to $\hat{\lambda}_{oj}$. Note that $\boldsymbol{S}_o$ and $\boldsymbol{S}_{oD}$ share the non-zero eigenvalues. We consider the following condition.

(A-i) $$\frac{\mathrm{tr}(\boldsymbol{\Sigma}^2)}{\mathrm{tr}(\boldsymbol{\Sigma})^2} = \frac{\sum_{s=1}^p \lambda_s^2}{(\sum_{s=1}^p \lambda_s)^2} \to 0, \quad p \to \infty.$$

Note that (A-i) is equivalent to the condition that $\lambda_1 / \mathrm{tr}(\boldsymbol{\Sigma}) \to 0$, $p \to \infty$. Then, when $\boldsymbol{X}$ is Gaussian or $\boldsymbol{Z}$ is ρ-mixing, Ahn et al. [1] and Jung and Marron [25] showed a geometric representation as follows:

$$\frac{n}{\mathrm{tr}(\boldsymbol{\Sigma})} \boldsymbol{S}_{oD} \xrightarrow{P} \boldsymbol{I}_n, \quad p \to \infty. \tag{2.1}$$

Let $\boldsymbol{w}_{oj} = \{n / \mathrm{tr}(\boldsymbol{\Sigma})\} \hat{\lambda}_{oj} \hat{\boldsymbol{u}}_{oj}$ and $\boldsymbol{R}_{on} = \{\boldsymbol{e}_n \in \boldsymbol{R}^n | \, ||\boldsymbol{e}_n|| = 1\}$. Yata and Aoshima [38] showed that

$$\boldsymbol{w}_{oj} \in \boldsymbol{R}_{on}, \quad j = 1, ..., n \tag{2.2}$$

in probability as $p \to \infty$. On the other hand, when $\boldsymbol{X}$ is non-Gaussian and $\boldsymbol{Z}$ is non-ρ-mixing, Yata and Aoshima [38] showed another geometric representation as follows:

$$\frac{n}{\mathrm{tr}(\boldsymbol{\Sigma})}\boldsymbol{S}_{oD} \xrightarrow{P} \boldsymbol{D}_n, \quad p \to \infty \tag{2.3}$$

where $\boldsymbol{D}_n$ is a diagonal matrix whose diagonal elements are of $O_P(1)$. Yata and Aoshima [38] considered a boundary condition between (2.1) and (2.3) as follows:

(A-ii) $$\frac{\mathrm{Var}(||\boldsymbol{x}_k - \boldsymbol{\mu}||^2)}{\mathrm{tr}(\boldsymbol{\Sigma})^2} = \frac{\sum_{r,s\geq 1}^{p} \lambda_r \lambda_s E\{(z_{rk}^2 - 1)(z_{sk}^2 - 1)\}}{(\sum_{s=1}^{p} \lambda_s)^2} \to 0, \quad p \to \infty.$$

Then, they gave the following result.

Theorem 2.1 (Yata and Aoshima [38])**.** *Assume (A-i). If the elements of $\boldsymbol{Z}$ satisfy (A-ii), we have (2.1) as $p \to \infty$. Otherwise, we have (2.3) as $p \to \infty$.*

2.2 When the population mean vector is unknown

Let us write the sample covariance matrix as $\boldsymbol{S} = (n-1)^{-1}(\boldsymbol{X} - \overline{\boldsymbol{X}})(\boldsymbol{X} - \overline{\boldsymbol{X}})^T = (n-1)^{-1}\sum_{j=1}^{n}(\boldsymbol{x}_j - \bar{\boldsymbol{x}})(\boldsymbol{x}_j - \bar{\boldsymbol{x}})^T$, where $\overline{\boldsymbol{X}} = [\bar{\boldsymbol{x}}, ..., \bar{\boldsymbol{x}}]$ and $\bar{\boldsymbol{x}} = \sum_{j=1}^{n} \boldsymbol{x}_j/n$. Then, we define the $n \times n$ dual sample covariance matrix by $\boldsymbol{S}_D = (n-1)^{-1}(\boldsymbol{X} - \overline{\boldsymbol{X}})^T(\boldsymbol{X} - \overline{\boldsymbol{X}})$. Let $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_{n-1} \geq 0$ be the eigenvalues of $\boldsymbol{S}_D$. Let us write the eigen-decomposition of $\boldsymbol{S}_D$ as $\boldsymbol{S}_D = \sum_{j=1}^{n-1} \hat{\lambda}_j \hat{\boldsymbol{u}}_j \hat{\boldsymbol{u}}_j^T$, where $\hat{\boldsymbol{u}}_j$ denotes a unit eigenvector corresponding to $\hat{\lambda}_j$. Note that $\boldsymbol{S}$ and $\boldsymbol{S}_D$ share the non-zero eigenvalues. Then, Ishii et al. [19] gave the following results.

Theorem 2.2. *Assume (A-i) and (A-ii). Then, we have as $p \to \infty$ that*

$$\frac{n-1}{tr(\boldsymbol{\Sigma})}\boldsymbol{S}_D \xrightarrow{P} \boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T,$$

where $\mathbf{1}_n = (1, ..., 1)^T$.

Proof. By using Chebyshev's inequality, for any $\tau > 0$, we have as $p \to \infty$ that

$$\begin{aligned} &P\Big(\Big|\frac{||\boldsymbol{x}_k - \boldsymbol{\mu}||^2}{\mathrm{tr}(\boldsymbol{\Sigma})} - 1\Big| > \tau\Big) \leq \tau^{-2}\frac{\sum_{r,s\geq 1}^{p} \lambda_r \lambda_s E\{(z_{rk}^2 - 1)(z_{sk}^2 - 1)\}}{\mathrm{tr}(\boldsymbol{\Sigma})^2} \to 0; \\ &P\Big(\Big|\frac{(\boldsymbol{x}_k - \boldsymbol{\mu})^T(\boldsymbol{x}_{k'} - \boldsymbol{\mu})}{\mathrm{tr}(\boldsymbol{\Sigma})}\Big| > \tau\Big) \leq \tau^{-2}\frac{\mathrm{tr}(\boldsymbol{\Sigma}^2)}{\mathrm{tr}(\boldsymbol{\Sigma})^2} \to 0 \quad (k \neq k') \end{aligned} \tag{2.4}$$

under (A-i) and (A-ii). Then, we have $(\boldsymbol{X} - [\boldsymbol{\mu}, ..., \boldsymbol{\mu}])^T(\boldsymbol{X} - [\boldsymbol{\mu}, ..., \boldsymbol{\mu}])/\mathrm{tr}(\boldsymbol{\Sigma}) \xrightarrow{P} \boldsymbol{I}_n$. We note that $(\boldsymbol{X} - [\boldsymbol{\mu}, ..., \boldsymbol{\mu}])(\boldsymbol{I}_n - \mathbf{1}_n\mathbf{1}_n^T/n) = \boldsymbol{X} - \overline{\boldsymbol{X}}$. Thus we write that

$$\boldsymbol{S}_D = \frac{(\boldsymbol{I}_n - \mathbf{1}_n\mathbf{1}_n^T/n)(\boldsymbol{X} - [\boldsymbol{\mu}, ..., \boldsymbol{\mu}])^T(\boldsymbol{X} - [\boldsymbol{\mu}, ..., \boldsymbol{\mu}])(\boldsymbol{I}_n - \mathbf{1}_n\mathbf{1}_n^T/n)}{n-1}.$$

Hence, we have that

$$\frac{(n-1)\boldsymbol{S}_D}{\mathrm{tr}(\boldsymbol{\Sigma})} \xrightarrow{P} \boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T.$$

It concludes the result. □

Corollary 2.1. *Let $\boldsymbol{w}_j = \{(n-1)/tr(\boldsymbol{\Sigma})\}\hat{\lambda}_j\hat{\boldsymbol{u}}_j$. Assume (A-i) and (A-ii). Then, we have that*

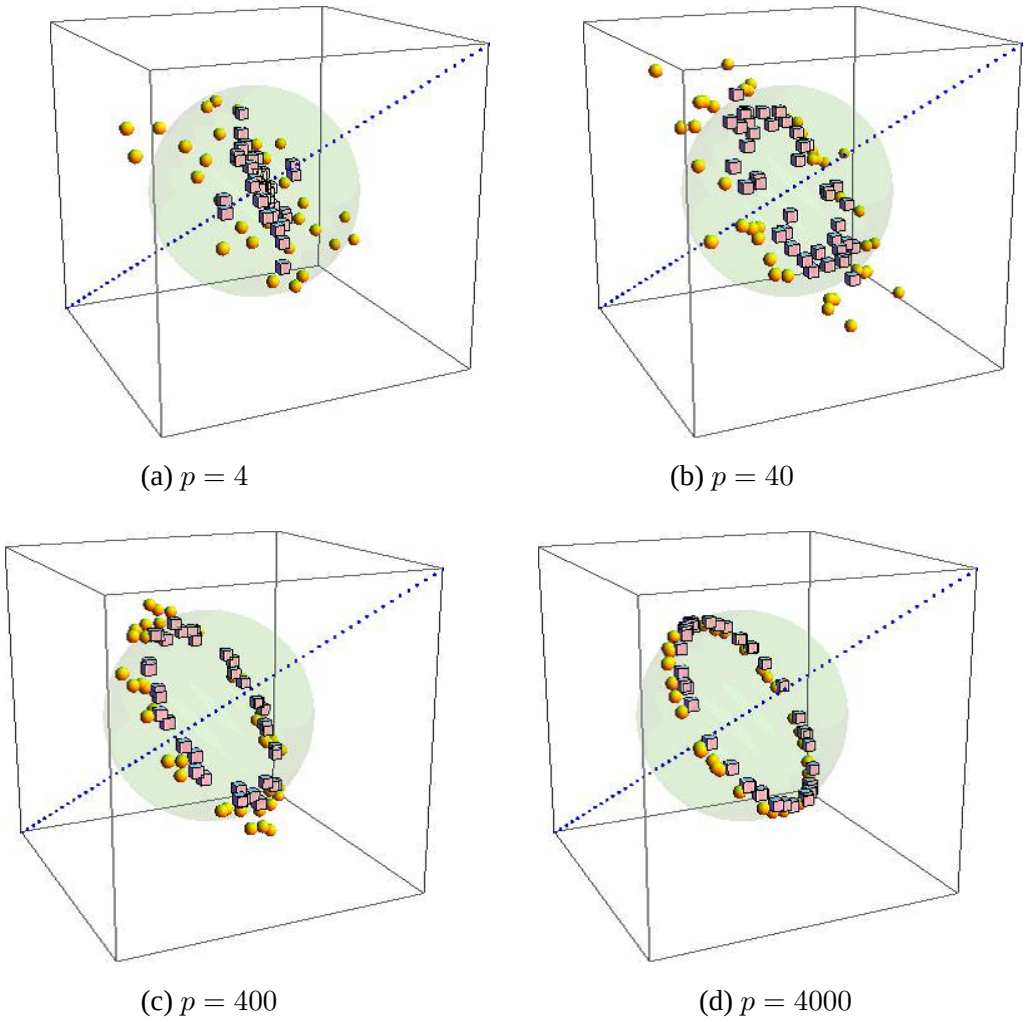
$$\frac{(n-1)\hat{\lambda}_i}{tr(\boldsymbol{\Sigma})} = \frac{(n-1)\hat{\boldsymbol{u}}_i^T\boldsymbol{S}_D\hat{\boldsymbol{u}}_i}{tr(\boldsymbol{\Sigma})} \xrightarrow{P} 1, \quad i = 1, ..., n-1;$$
$$\boldsymbol{w}_j \in \boldsymbol{R}_n, \quad j = 1, ..., n-1$$

in probability as $p \to \infty$, where $\boldsymbol{R}_n = \{\boldsymbol{e}_n \in \boldsymbol{R}^n |\ \boldsymbol{e}_n^T\mathbf{1}_n = 0,\ ||\boldsymbol{e}_n|| = 1\}$.

Proof. From Theorem 2.2 it follows that rank($\boldsymbol{S}_D$) $= n-1$ asymptotically. By noting that $\hat{\boldsymbol{u}}_i^T\mathbf{1}_n = 0$ with probability tending to 1 for $i = 1, ..., n-1$, it concludes the results. □

From Corollary 2.1 the eigenspace spanned by $\hat{\boldsymbol{u}}_i,\ i = 1, ..., n-1$, is close to the orthogonal complement of $\mathbf{1}_n$ in $\boldsymbol{R}^n$ as $p \to \infty$ and the direction of the eigenvectors is not uniquely determined. On the other hand, the eigenvalues become deterministic but there becomes no difference among them. For these reasons, it is difficult to estimate the eigenvalues and the eigenvectors by using $\boldsymbol{S}_D$ (or $\boldsymbol{S}$) in conventional PCA.

Let us observe a geometric representation given by Corollary 2.1. Now, we consider an easy example such as $\lambda_1 = \cdots = \lambda_p = 1$ and $n = 3$. In Fig. 1, we displayed scatter plots of 20 independent pairs of $\pm\boldsymbol{w}_j\ (j = 1, 2)$ that were generated from $N_p(\boldsymbol{\mu}, \boldsymbol{I}_p)$ for (a) $p = 4$, (b) $p = 40$, (c) $p = 400$ and (d) $p = 4000$. We denoted $\boldsymbol{w}_1$ by  and $\boldsymbol{w}_2$ by  . We also denoted $\mathbf{1}_n = (1, 1, 1)^T$ by the dotted line. We observed that all the plots of $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ gather on the surface of the orthogonal complement of $\mathbf{1}_n = (1, 1, 1)^T$ in $\boldsymbol{R}^3$ when p is large. Moreover, they appeared around the unit circle on the orthogonal complement of $\mathbf{1}_n = (1, 1, 1)^T$ in $\boldsymbol{R}^3$ as expected by Corollary 2.1.



(a) $p = 4$ (b) $p = 40$

(c) $p = 400$ (d) $p = 4000$

Figure 1. The geometric representation of 20 pairs of $\pm \boldsymbol{w}_j (j = 1, 2)$ from $N_p(\boldsymbol{\mu}, \boldsymbol{I}_p)$ when $p = 4, 40, 400$ and 4000. We denoted $\boldsymbol{w}_1$ by ●, $\boldsymbol{w}_2$ by ■ and $\mathbf{1}_n = (1, 1, 1)^T$ by the dotted line.

3 Estimation of the First Eigenvalue

In this section, we consider eigenvalue estimation and give asymptotic distributions for the first eigenvalue. In recent years, substantial work had been done on the asymptotic behavior of eigenvalues of the sample covariance matrix in the limit as $p \to \infty$, see Johnstone [24] and Paul [28] for Gaussian data and Baik and Silverstein [11] for non-Gaussian, i.i.d. data. Those literatures handled the cases when p and n increase at the same rate, i.e. $p/n \to c > 0$. The HDLSS asymptotic study usually assumes either the normality as the population distribution or a ρ-mixing condition as the dependency of random variables in a sphered data matrix. For instance, see Jung and Marron [25]. Yata and Aoshima [35, 40] succeeded in investigating the consistency properties of both eigenvalues and eigenvectors in a more general framework. Yata and Aoshima [38] gave consistent estimators of both the eigenvalues and eigenvectors together with

the principal component (PC) scores by a method called the *noise-reduction (NR) methodology*. Yata and Aoshima [36, 39] created the *cross-data-matrix (CDM) methodology* that provides a nonparametric method for non-Gaussian HDLSS data.

3.1 Conventional estimator

Usually, one uses eigenvalues and eigenvectors of the sample covariance matrix, $\boldsymbol{S} = (n-1)^{-1}(\boldsymbol{X} - \overline{\boldsymbol{X}})(\boldsymbol{X} - \overline{\boldsymbol{X}})^T$. Now, we recall the dual sample covariance matrix, $\boldsymbol{S}_D = (n-1)^{-1}(\boldsymbol{X} - \overline{\boldsymbol{X}})^T(\boldsymbol{X} - \overline{\boldsymbol{X}})$. Note that $\boldsymbol{S}$ and $\boldsymbol{S}_D$ share the non-zero eigenvalues. In actual data analyses, we use $\boldsymbol{S}_D$ to estimate the target eigenvalues because of its low computational cost. Let $\delta_i = \mathrm{tr}(\boldsymbol{\Sigma}^2) - \sum_{s=1}^{i} \lambda_s^2 = \sum_{s=i+1}^{p} \lambda_s^2$ for $i = 1, ..., p-1$. Then, we consider the following assumptions for the first eigenvalue:

(A-iii) $\dfrac{\delta_1}{\lambda_1^2} = o(1)$ as $p \to \infty$ when n is fixed; $\dfrac{\delta_{i_*}}{\lambda_1^2} = o(1)$ as $p \to \infty$ for some fixed i_* $(< p)$ when $n \to \infty$.

(A-iv) $\dfrac{\sum_{r,s\geq 2}^{p} \lambda_r \lambda_s E\{(z_{rk}^2 - 1)(z_{sk}^2 - 1)\}}{n\lambda_1^2} = o(1)$ as $p \to \infty$ either when n is fixed or $n \to \infty$.

Note that (A-iii) implies the conditions that $\lambda_2/\lambda_1 \to 0$ as $p \to \infty$ when n is fixed and $\lambda_{i_*+1}/\lambda_1 \to 0$ as $p \to \infty$ for some fixed i_* when $n \to \infty$. Also, note that (A-iv) holds when $\boldsymbol{X}$ is Gaussian and (A-iii) is met. See Remark 3.2.

Remark 3.1. For a spiked model such as

$$\lambda_j = a_j p^{\alpha_j} \ (j = 1, ..., m) \quad \text{and} \quad \lambda_j = c_j \ (j = m+1, ..., p)$$

with positive (fixed) constants, a_js, c_js and α_js, and a positive (fixed) integer m, (A-iii) holds under the condition that $\alpha_1 > 1/2$ and $\alpha_1 > \alpha_2$ when n is fixed. When $n \to \infty$, (A-iii) holds under $\alpha_1 > 1/2$ even if $\alpha_1 = \alpha_m$. See Yata and Aoshima [38] for the details.

Remark 3.2. For several statistical inferences of high-dimensional data, Bai and Saranadasa [10], Chen and Qin [13] and Aoshima and Yata [7] assumed a general factor model as follows:

$$\boldsymbol{x}_j = \boldsymbol{\Gamma}\boldsymbol{w}_j + \boldsymbol{\mu}$$

for $j = 1, ..., n$, where $\boldsymbol{\Gamma}$ is a $p \times r$ matrix for some $r > 0$ such that $\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T = \boldsymbol{\Sigma}$, and $\boldsymbol{w}_j$, $j = 1, ..., n$, are i.i.d. random vectors having $E(\boldsymbol{w}_j) = \boldsymbol{0}$ and $\mathrm{Var}(\boldsymbol{w}_j) = \boldsymbol{I}_r$. As for $\boldsymbol{w}_j = (w_{1j}, ..., w_{rj})^T$, assume that $E(w_{qj}^2 w_{sj}^2) = 1$ and $E(w_{qj} w_{sj} w_{tj} w_{uj}) = 0$ for all $q \neq s, t, u$. From Lemma 1 in Yata and Aoshima [40], one can claim that (A-iv) holds under (A-iii) in the factor model. Also, we note that the factor model naturally holds when $\boldsymbol{X}$ is Gaussian.

Let $\kappa = \mathrm{tr}(\boldsymbol{\Sigma}) - \lambda_1 = \sum_{s=2}^{p} \lambda_s$. Then, we have the following result.

Proposition 3.1. Under (A-iii) and (A-iv), it holds that

$$\frac{\hat{\lambda}_1}{\lambda_1} - ||\boldsymbol{z}_{o1}/\sqrt{n-1}||^2 - \frac{\kappa}{\lambda_1(n-1)} = o_p(1) \tag{3.1}$$

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

Proof. Let $\boldsymbol{P}_n = \boldsymbol{I}_n - \mathbf{1}_n\mathbf{1}_n^T/n$, where $\mathbf{1}_n = (1, ..., 1)^T$. Also, let $\boldsymbol{e}_n = (e_1, ..., e_n)^T$ be an arbitrary (random) n-vector such that $||\boldsymbol{e}_n|| = 1$ and $\boldsymbol{e}_n^T\mathbf{1}_n = 0$. We assume $\boldsymbol{\mu} = \mathbf{0}$ without loss of generality. We write that $\boldsymbol{X}^T\boldsymbol{X} = \sum_{s=1}^{i_*} \lambda_s \boldsymbol{z}_s\boldsymbol{z}_s^T + \sum_{s=i_*+1}^{p} \lambda_s \boldsymbol{z}_s\boldsymbol{z}_s^T$ for $i_* = 1$ when n is fixed, and for some fixed $i_*(\geq 1)$ when $n \to \infty$. Here, by using Markov's inequality, for any $\tau > 0$, under (A-iii) and (A-iv), we have that

$$\begin{aligned} &P\Big\{ \sum_{j=1}^{n} \Big(\sum_{s=i_*+1}^{p} \frac{\lambda_s(z_{sj}^2 - 1)}{n\lambda_1} \Big)^2 > \tau \Big\} \leq \frac{\sum_{r,s\geq 2}^{p} \lambda_r\lambda_s E\{(z_{rk}^2 - 1)(z_{sk}^2 - 1)\}}{\tau n\lambda_1^2} \to 0 \\ &\text{and } P\Big\{ \sum_{j\neq j'}^{n} \Big(\sum_{s=i_*+1}^{p} \frac{\lambda_s z_{sj}z_{sj'}}{n\lambda_1} \Big)^2 > \tau \Big\} \leq \frac{\delta_{i_*}}{\tau\lambda_1^2} \to 0 \end{aligned} \tag{3.2}$$

as $p \to \infty$ either when n is fixed or $n \to \infty$. Note that $\sum_{j=1}^{n} e_j^4 \leq 1$ and $\sum_{j\neq j'}^{n} e_j^2 e_{j'}^2 \leq 1$. Then, under (A-iii) and (A-iv), we have that

$$\begin{aligned} \Big| \sum_{j=1}^{n} e_j^2 \sum_{s=i_*+1}^{p} \frac{\lambda_s(z_{sj}^2 - 1)}{n\lambda_1} \Big| &\leq \Big\{ \sum_{j=1}^{n} e_j^4 \Big\}^{1/2} \Big\{ \sum_{j=1}^{n} \Big(\sum_{s=i_*+1}^{p} \frac{\lambda_s(z_{sj}^2 - 1)}{n\lambda_1} \Big)^2 \Big\}^{1/2} \\ &= o_p(1) \quad \text{and} \\ \Big| \sum_{j\neq j'}^{n} e_j e_{j'} \sum_{s=i_*+1}^{p} \frac{\lambda_s z_{sj}z_{sj'}}{n\lambda_1} \Big| &\leq \Big\{ \sum_{j\neq j'}^{n} e_j^2 e_{j'}^2 \Big\}^{1/2} \Big\{ \sum_{j\neq j'}^{n} \Big(\sum_{s=i_*+1}^{p} \frac{\lambda_s z_{sj}z_{sj'}}{n\lambda_1} \Big)^2 \Big\}^{1/2} \\ &= o_p(1) \end{aligned}$$

as $p \to \infty$ either when n is fixed or $n \to \infty$. Thus, we claim that

$$\boldsymbol{e}_n^T \frac{\boldsymbol{X}^T\boldsymbol{X}}{(n-1)\lambda_1} \boldsymbol{e}_n = \boldsymbol{e}_n^T \frac{\sum_{s=1}^{i_*} \lambda_s \boldsymbol{z}_s\boldsymbol{z}_s^T}{(n-1)\lambda_1} \boldsymbol{e}_n + \frac{\kappa}{(n-1)\lambda_1} + o_p(1) \tag{3.3}$$

from the fact that $\sum_{s=i_*+1}^{p} \lambda_s/\{(n-1)\lambda_1\} = \kappa/\{(n-1)\lambda_1\} + o(1)$ when $n \to \infty$. Note that $\boldsymbol{e}_n^T\boldsymbol{P}_n = \boldsymbol{e}_n^T$ and $\boldsymbol{P}_n\boldsymbol{z}_s = \boldsymbol{z}_{os}$ for all s. Also, note that $\boldsymbol{z}_{os}^T\boldsymbol{z}_{os'}/n = o_p(1)$ for $s \neq s'$ as $n \to \infty$ from the fact that $E\{(\boldsymbol{z}_{os}^T\boldsymbol{z}_{os'}/n)^2\} = o(1)$ as $n \to \infty$. Then, by noting that $P(\lim_{p\to\infty} ||\boldsymbol{z}_{o1}|| \neq 0) = 1$, $\liminf_{p\to\infty} \lambda_1/\lambda_2 > 1$ and $\boldsymbol{z}_{o1}^T\mathbf{1}_n = 0$, it holds that

$$\begin{aligned} \max_{\boldsymbol{e}_n} \Big\{ \boldsymbol{e}_n^T \frac{\sum_{s=1}^{i_*} \lambda_s \boldsymbol{z}_s\boldsymbol{z}_s^T}{(n-1)\lambda_1} \boldsymbol{e}_n \Big\} &= \max_{\boldsymbol{e}_n} \Big\{ \boldsymbol{e}_n^T \frac{\sum_{s=1}^{i_*} \lambda_s \boldsymbol{z}_{os}\boldsymbol{z}_{os}^T}{(n-1)\lambda_1} \boldsymbol{e}_n \Big\} \\ &= ||\boldsymbol{z}_{o1}/\sqrt{n-1}||^2 + o_p(1) \end{aligned} \tag{3.4}$$

as $p \to \infty$ either when n is fixed or $n \to \infty$. Note that $\hat{\boldsymbol{u}}_1^T \mathbf{1}_n = 0$ and $\hat{\boldsymbol{u}}_1^T \boldsymbol{P}_n = \hat{\boldsymbol{u}}_1^T$ when $\boldsymbol{S}_D \neq \boldsymbol{O}$. Then, from (3.3), (3.4) and $\boldsymbol{P}_n \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{P}_n/(n-1) = \boldsymbol{S}_D$, under (A-iii) and (A-iv), we have that

$$\hat{\boldsymbol{u}}_1^T \frac{\boldsymbol{S}_D}{\lambda_1} \hat{\boldsymbol{u}}_1 = \hat{\boldsymbol{u}}_1^T \frac{\boldsymbol{X}^T \boldsymbol{X}}{(n-1)\lambda_1} \hat{\boldsymbol{u}}_1 = ||\boldsymbol{z}_{o1}/\sqrt{n-1}||^2 + \frac{\kappa}{(n-1)\lambda_1} + o_p(1) \tag{3.5}$$

as $p \to \infty$ either when n is fixed or $n \to \infty$. It concludes the result. □

Remark 3.3. Jung et al. [26] gave a result similar to Proposition 3.1 when $\boldsymbol{X}$ is Gaussian, $\boldsymbol{\mu} = \mathbf{0}$ and n is fixed.

Now, we consider the asymptotic distribution of the conventional estimator, $\hat{\lambda}_1$ when $p \to \infty$ while n is fixed. As necessary, we consider the following assumption for the normalized first PC scores, z_{1j} $(= s_{1j}/\lambda_1^{1/2})$, $j = 1, ..., n$:

(A-v) $z_{1j},\ j = 1, ..., n$, are i.i.d. as $N(0, 1)$.

Note that $P(\lim_{p\to\infty} ||\boldsymbol{z}_{o1}|| \neq 0) = 1$ under (A-v) from the fact that $||\boldsymbol{z}_{o1}||^2$ is distributed as χ^2_{n-1}, where χ^2_ν denotes a random variable distributed as χ^2 distribution with ν degrees of freedom. From (3.1) we have the following result for the conventional estimator $\hat{\lambda}_1$.

Corollary 3.1. *Assume (A-v). If $\kappa/\lambda_1 = o(1)$ as $p \to \infty$, it holds that as $p \to \infty$*

$$(n-1)\frac{\hat{\lambda}_1}{\lambda_1} \Rightarrow \chi^2_{n-1}. \tag{3.6}$$

Proof. If $\kappa/\lambda_1 = o(1)$ as $p \to \infty$, from Proposition 3.1 it holds as $p \to \infty$ that $\hat{\lambda}_1/\lambda_1 = ||\boldsymbol{z}_{o1}/\sqrt{n-1}||^2 + o_p(1)$. Note that $||\boldsymbol{z}_{o1}||^2$ is distributed as χ^2_{n-1} under (A-v), where χ^2_{n-1} denotes a random variable distributed as χ^2 distribution with $n-1$ degrees of freedom. It concludes the result. □

Remark 3.4. Jung and Marron [25] gave (3.6) under different but still strict assumptions.

It holds that $E(||\boldsymbol{z}_{o1}/\sqrt{n-1}||^2) = 1$ and $||\boldsymbol{z}_{o1}/\sqrt{n-1}||^2 = 1 + o_p(1)$ as $n \to \infty$. If $\kappa/(n\lambda_1) = o(1)$ as $p \to \infty$ and $n \to \infty$, $\hat{\lambda}_1$ is a consistent estimator of λ_1. When n is fixed, the condition '$\kappa/\lambda_1 = o(1)$' is equivalent to '$\lambda_1/\mathrm{tr}(\boldsymbol{\Sigma}) = 1 + o(1)$' in which the contribution ratio of the first principal component is asymptotically 1. In that sense, '$\kappa/\lambda_1 = o(1)$' is quite strict condition in real high-dimensional data analyses. Hereafter, we assume $\liminf_{p\to\infty} \kappa/\lambda_1 > 0$.

3.2 Noise-reduction estimator

Yata and Aoshima [38] proposed a method for eigenvalue estimation called the *noise-reduction (NR) methodology* that was brought by the geometric representation in (2.2). When we apply the NR methodology to the case when $\boldsymbol{\mu}$ is unknown, the NR estimator of λ_i is given by

$$\tilde{\lambda}_i = \hat{\lambda}_i - \frac{\mathrm{tr}(\boldsymbol{S}_D) - \sum_{j=1}^{i} \hat{\lambda}_j}{n-1-i} \quad (i = 1, ..., n-2). \tag{3.7}$$

Note that $\tilde{\lambda}_i \geq 0$ for $i = 1, ..., n-2$. Also, note that the second term in (3.7) with $i = 1$ is an estimator of $\kappa/(n-1)$. See Lemma 2.1 in Chapter 2 for the details. Yata and Aoshima [38, 40] showed that $\tilde{\lambda}_i$ has several consistency properties when $p \to \infty$ and $n \to \infty$. On the other hand, Ishii et al. [19] gave asymptotic properties of $\tilde{\lambda}_1$ when $p \to \infty$ while n is fixed. The following theorem summarizes their findings:

Theorem 3.1 (Yata and Aoshima [40], Ishii et al. [19])**.** *Under (A-iii) and (A-iv), it holds that as* $p \to \infty$

$$\frac{\tilde{\lambda}_1}{\lambda_1} = \begin{cases} ||\boldsymbol{z}_{o1}/\sqrt{n-1}||^2 + o_p(1) & \textit{when } n \textit{ is fixed}, \\ 1 + o_p(1) & \textit{when } n \to \infty. \end{cases}$$

Under (A-iii) to (A-v), it holds that as $p \to \infty$

$$\begin{aligned} &(n-1)\frac{\tilde{\lambda}_1}{\lambda_1} \Rightarrow \chi^2_{n-1} && \textit{when } n \textit{ is fixed}, \\ &\sqrt{\frac{n-1}{2}}\Big(\frac{\tilde{\lambda}_1}{\lambda_1} - 1\Big) \Rightarrow N(0,1) && \textit{when } n \to \infty. \end{aligned}$$

Here, " $\Rightarrow$ *" denotes the convergence in distribution.*

Proof. When $n \to \infty$, we can claim the results from Theorems 4.1, 4.2 and Corollary 4.1 in Yata and Aoshima [40]. When n is fixed, by combining Proposition 3.1 with Lemma 2.1 in Chapter 2, we can claim the results because $||\boldsymbol{z}_{o1}||^2 = \sum_{k=1}^{n} z_{1k}^2 - n\bar{z}_1^2$ is distributed as χ^2_{n-1} under (A-v). □

Remark 3.5. Let $\mathrm{Var}(z_{1k}^2) = M_1$ $(< \infty)$ and assume $\liminf_{p\to\infty} M_1 > 0$. Note that $M_1 = 2$ if $z_{1j},\ j = 1, ..., n$, are i.i.d. as $N(0,1)$. When $p \to \infty$ and $n \to \infty$, Yata and Aoshima [40] showed that under (A-iii) and (A-iv)

$$\sqrt{\frac{n-1}{M_1}}\Big(\frac{\tilde{\lambda}_1}{\lambda_1} - 1\Big) \Rightarrow N(0,1).$$

On the other hand, if $\kappa/\lambda_1 = o(1)$ as $p \to \infty$, it holds as $p \to \infty$ and $n \to \infty$ that

$$\sqrt{\frac{n-1}{M_1}}\Big(\frac{\hat{\lambda}_1}{\lambda_1} - 1\Big) \Rightarrow N(0,1).$$

3.3 Bias corrected cross-data-matrix estimator

We consider the case when (A-iv) is not always met. In such cases, the NR methodology does not ensure the asymptotic properties. Yata and Aoshima [36] proposed a method called the *cross-data-matrix (CDM) methodology* to proceed with eigenvalue estimation even in such cases. Let $n_{(1)} = \lceil n/2 \rceil$ and $n_{(2)} = n - n_{(1)}$, where $\lceil x \rceil$ denotes the smallest integer $\geq x$. We divide the data matrix $\boldsymbol{X}$ into $\boldsymbol{X}_{(1)} = [\boldsymbol{x}_{(1)1}, ..., \boldsymbol{x}_{(1)n_{(1)}}]$ and $\boldsymbol{X}_{(2)} = [\boldsymbol{x}_{(2)1}, ..., \boldsymbol{x}_{(2)n_{(2)}}]$ at random. We define a cross data matrix with $\boldsymbol{X}_{(1)}$ and $\boldsymbol{X}_{(2)}$ by $\boldsymbol{S}_{D(1)} = \{(n_{(1)}-1)(n_{(2)}-1)\}^{-1/2}(\boldsymbol{X}_{(1)} - \overline{\boldsymbol{X}}_{(1)})^T(\boldsymbol{X}_{(2)} - \overline{\boldsymbol{X}}_{(2)})$, where $\overline{\boldsymbol{X}}_{(i)} = [\bar{\boldsymbol{x}}_{(i)}, ..., \bar{\boldsymbol{x}}_{(i)}]$ having p-vector $\bar{\boldsymbol{x}}_{(i)} = {n_{(i)}}^{-1}\sum_{j=1}^{n_{(i)}} \boldsymbol{x}_{(i)j}$ $(i = 1, 2)$. Let $r = n_{(2)} - 1$. We calculate the singular value decomposition of $\boldsymbol{S}_{D(1)}$ by $\boldsymbol{S}_{D(1)} = \sum_{j=1}^{r} \acute{\lambda}_j \acute{\boldsymbol{u}}_{(1)j}\acute{\boldsymbol{u}}_{(2)j}^T$, where $\acute{\lambda}_1 \geq \cdots \geq \acute{\lambda}_r (\geq 0)$ denote singular values of $\boldsymbol{S}_{D(1)}$, and $\acute{\boldsymbol{u}}_{(1)j}$ (or $\acute{\boldsymbol{u}}_{(2)j}$) denotes a unit left- (or right-) singular vector corresponding to $\acute{\lambda}_j$ $(j = 1, ..., r)$. Yata and Aoshima [36, 40] showed that $\acute{\lambda}_j$ enjoys several consistency properties to estimate λ_j without any assumptions about the population distribution when $p \to \infty$ and $n \to \infty$ even in the HDLSS context.

Let us write $\boldsymbol{X}_{(i)} - [\boldsymbol{\mu}, ..., \boldsymbol{\mu}] = \boldsymbol{H}\boldsymbol{\Lambda}^{1/2}\boldsymbol{Z}_{(i)}$, where $\boldsymbol{Z}_{(i)} = [\boldsymbol{z}_{(i)1}, ..., \boldsymbol{z}_{(i)p}]^T$ and $\boldsymbol{z}_{(i)j} = (z_{(i)j1}, ..., z_{(i)jn_{(i)}})^T$, $i = 1, 2;\ j = 1, ..., p$. Let $\boldsymbol{z}_{o(i)j} = \boldsymbol{z}_{(i)j} - (\bar{z}_{(i)j}, ..., \bar{z}_{(i)j})^T$, $j = 1, ..., p$, where $\bar{z}_{(i)j} = n_{(i)}^{-1}\sum_{k=1}^{n_{(i)}} z_{(i)jk}$ $(i = 1, 2; j = 1, ..., p)$. We assume $P(\lim_{p\to\infty} ||\boldsymbol{z}_{o(i)1}|| \neq 0) = 1,\ i = 1, 2$. We have that

$$\sqrt{(n_{(1)}-1)(n_{(2)}-1)}\boldsymbol{S}_{D(1)} = \lambda_1 \boldsymbol{z}_{o(1)1}\boldsymbol{z}_{o(2)1}^T + \sum_{j=2}^{p} \lambda_j \boldsymbol{z}_{o(1)j}\boldsymbol{z}_{o(2)j}^T. \tag{3.8}$$

Here, for any (i, j) element of $\sum_{j=2}^{p} \lambda_j \boldsymbol{z}_{o(1)j}\boldsymbol{z}_{o(2)j}^T$, it holds that as $p \to \infty$

$$\begin{aligned}&\frac{\mathrm{Var}\{\sum_{s=2}^{p} \lambda_s (z_{(1)si} - \bar{z}_{(1)s})(z_{(2)sj} - \bar{z}_{(2)s})\}}{\lambda_1^2} \\ &= \frac{(n_{(1)}-1)(n_{(2)}-1)}{n_{(1)}n_{(2)}} \frac{\mathrm{tr}(\boldsymbol{\Sigma}^2) - \lambda_1^2}{\lambda_1^2} \to 0\end{aligned}$$

under (A-iii). Hence, we can claim under (A-iii) that as $p \to \infty$

$$\frac{\sum_{j=2}^{p} \lambda_j \boldsymbol{z}_{o(1)j}\boldsymbol{z}_{o(2)j}^T}{\lambda_1} \xrightarrow{P} \boldsymbol{O}. \tag{3.9}$$

Let us observe (3.9) by using computer simulations. We took $n = 6$ samples from p-variate t-distribution, $t_p(\boldsymbol{0}, \boldsymbol{\Sigma}, 5)$, with mean $\boldsymbol{0}$, covariance matrix $\boldsymbol{\Sigma} = \mathrm{diag}(\lambda_1, \lambda_2, ..., \lambda_p)$ having $\lambda_1 = p^{2/3}$ and $\lambda_2 = \cdots = \lambda_p = 1$, and 5 degrees of freedom. We considered four cases: $p = 6, 60, 600, 6000$. For each case we calculated $\lambda_1^{-1}\{\sqrt{(n_{(1)}-1)(n_{(2)}-1)}\boldsymbol{S}_{D(1)} - \lambda_1 \boldsymbol{z}_{o(1)1}\boldsymbol{z}_{o(2)1}^T\} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_3)^T$, where $\boldsymbol{w}_i$s are 3×1 vectors. Note that $\boldsymbol{w}_j$s are orthogonal to $\boldsymbol{1}_3$ in view of (3.8). We plotted $\boldsymbol{w}_1$ (white triangle), $\boldsymbol{w}_2$ (black circle) and $\boldsymbol{w}_3$ (cross mark) twenty times on the compliment space of $\boldsymbol{1}_3$ in Fig. 2. One can observe $\boldsymbol{w}_i$s converge to zero when p is large as expected theoretically in (3.9).

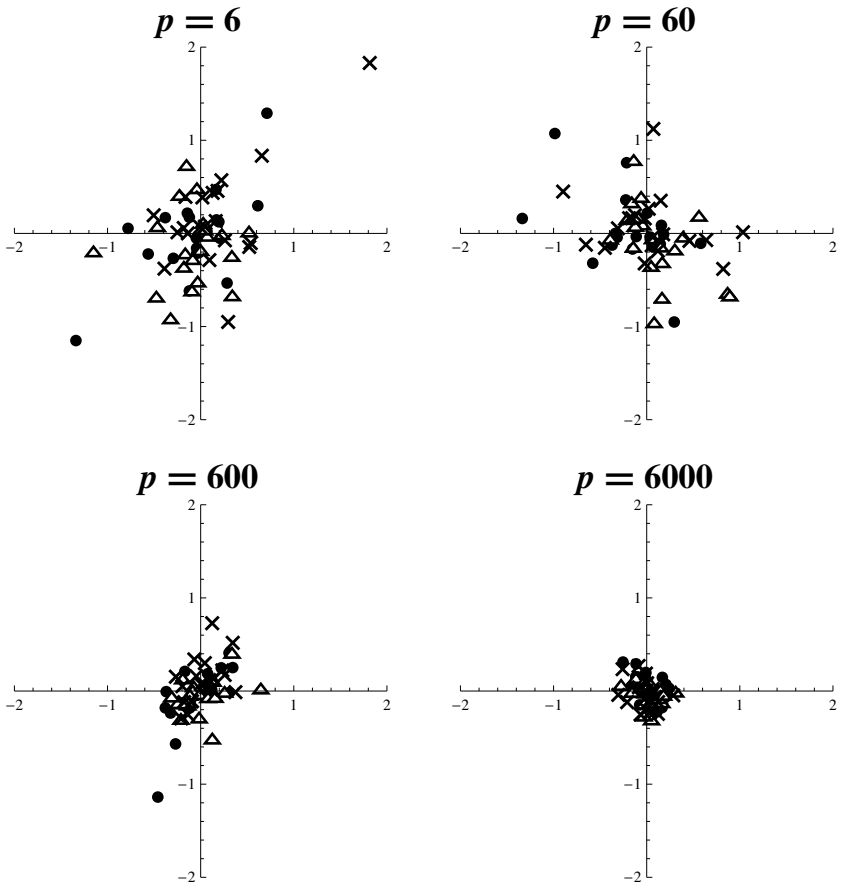


Figure 2. The behaviors of $\lambda_1^{-1}\{\sqrt{(n_{(1)}-1)(n_{(2)}-1)}\boldsymbol{S}_{D(1)} - \lambda_1 \boldsymbol{z}_{o(1)1}\boldsymbol{z}_{o(2)1}^T\} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_3)^T$ on the compliment space of $\mathbf{1}_3$. We plotted $\boldsymbol{w}_1$ (white triangle), $\boldsymbol{w}_2$ (black circle) and $\boldsymbol{w}_3$ (cross mark) when $n = 6$ samples are taken from $t_p(\mathbf{0}, \boldsymbol{\Sigma}, 5)$ with $\boldsymbol{\Sigma} = \mathrm{diag}(\lambda_1, \lambda_2, ..., \lambda_p)$ having $\lambda_1 = p^{2/3}$ and $\lambda_2 = \cdots = \lambda_p = 1$.

From (3.9) we have under (A-iii) that as $p \to \infty$

$$\begin{aligned}\frac{\acute{\lambda}_1}{\lambda_1} &= \acute{\boldsymbol{u}}_{1(1)}^T \frac{\boldsymbol{S}_{D(1)}}{\lambda_1} \acute{\boldsymbol{u}}_{1(2)} \\ &= \left(\acute{\boldsymbol{u}}_{1(1)}^T \boldsymbol{z}_{o(1)1} / \sqrt{n_{(1)}-1}\right)\left(\boldsymbol{z}_{o(2)1}^T \acute{\boldsymbol{u}}_{1(2)} / \sqrt{n_{(2)}-1}\right) + o_p(1). \end{aligned} \tag{3.10}$$

Then, we have the following result.

Theorem 3.2. *It holds under (A-iii) that as* $p \to \infty$

$$\frac{\acute{\lambda}_1}{\lambda_1} = \begin{cases} ||\boldsymbol{z}_{o(1)1}/\sqrt{n_{(1)}-1}||\,||\boldsymbol{z}_{o(2)1}/\sqrt{n_{(2)}-1}|| + o_p(1) & \textit{when } n \textit{ is fixed}, \\ 1 + o_p(1) & \textit{when } n \to \infty. \end{cases}$$

Proof. Let $\boldsymbol{e}_{(j)} = (e_{(j)1}, ..., e_{(j)n_{(j)}})^T$, $j = 1, 2$, be arbitrary unit $n_{(j)}$-vectors. From (3.8) and (3.9) we

have under (A-iii) that as $p \to \infty$

$$\lambda_1^{-1} \boldsymbol{e}_{(1)}^T \boldsymbol{S}_{D(1)} \boldsymbol{e}_{(2)} = \boldsymbol{e}_{(1)}^T \frac{\boldsymbol{z}_{o(1)1} \boldsymbol{z}_{o(2)1}^T}{\sqrt{(n_{(1)}-1)(n_{(2)}-1)}} \boldsymbol{e}_{(2)} + o_p(1).$$

Now we consider the first singular value of $\boldsymbol{S}_{D(1)}$. Then, it holds that as $p \to \infty$

$$\begin{aligned} \frac{\acute{\lambda}_1}{\lambda_1} &= \max \left\{ \boldsymbol{e}_{(1)}^T \left(\boldsymbol{z}_{o(1)1} / \sqrt{n_{(1)}-1} \right) \left(\boldsymbol{z}_{o(2)1}^T / \sqrt{n_{(2)}-1} \right) \boldsymbol{e}_{(2)} + o_p(1) \right\} \\ &= ||\boldsymbol{z}_{o(1)1} / \sqrt{n_{(1)}-1}|| ||\boldsymbol{z}_{o(2)1} / \sqrt{n_{(2)}-1}|| + o_p(1). \end{aligned} \tag{3.11}$$

Note that $||\boldsymbol{z}_{o(i)1} / \sqrt{n_{(i)}-1}|| = 1 + o_p(1),\ i = 1, 2$, when $p \to \infty$ and $n \to \infty$. Then, it concludes the result. □

Corollary 3.2. *It holds under (A-iii) and (A-v) that as* $p \to \infty$

$$\frac{\acute{\lambda}_1}{\lambda_1} \Rightarrow \sqrt{\frac{\chi^2_{(1)n_{(1)}-1}}{n_{(1)}-1}} \sqrt{\frac{\chi^2_{(2)n_{(2)}-1}}{n_{(2)}-1}}, \tag{3.12}$$

where " $\Rightarrow$ " denotes the convergence in distribution, and $\chi^2_{(i)n_{(i)}-1},\ i = 1, 2$, *are mutually independent random variables distributed as the chi-squared distribution with* $n_{(i)} - 1$, *degrees of freedom.*

Proof. Note that $||\boldsymbol{z}_{o(i)1}||^2 = \sum_{k=1}^{n_{(i)}} z^2_{(i)1k} - n_{(i)} \bar{z}^2_{(i)1}$ is distributed as $\chi^2_{n_{(i)}-1}$ for $i = 1, 2$, if $z_{(i)1k},\ k = 1, ..., n_{(i)}$, are i.i.d. as $N(0, 1)$. Thus we can conclude the result. □

From Corollary 3.2 we have that

$$E\left(\sqrt{\frac{\chi^2_{(1)n_{(1)}-1}}{n_{(1)}-1}} \sqrt{\frac{\chi^2_{(2)n_{(2)}-1}}{n_{(2)}-1}} \right) = \frac{c}{\sqrt{(n_{(1)}-1)(n_{(2)}-1)}}$$
$$\text{with } c = 2\Gamma\Big(\frac{n_{(1)}}{2}\Big)\Gamma\Big(\frac{n_{(2)}}{2}\Big)\Gamma\Big(\frac{n_{(1)}-1}{2}\Big)^{-1}\Gamma\Big(\frac{n_{(2)}-1}{2}\Big)^{-1},$$

where $\Gamma(\cdot)$ denotes the gamma function. Thus we can give a bias-correction of the CDM estimator for the first eigenvalue by

$$\acute{\lambda}_{1*} = \frac{\sqrt{(n_{(1)}-1)(n_{(2)}-1)}}{c} \acute{\lambda}_1. \tag{3.13}$$

Then, we have the following result.

Corollary 3.3. *It holds under (A-iii) and (A-v) that as* $p \to \infty$

$$\frac{\acute{\lambda}_{1*}}{\lambda_1} \Rightarrow \frac{1}{c} \sqrt{\chi^2_{(1)n_{(1)}-1}} \sqrt{\chi^2_{(2)n_{(2)}-1}} \ \textit{ and } \ E\Big(\frac{\acute{\lambda}_{1*}}{\lambda_1}\Big) \to 1.$$

Proof. From Corollary 3.2 and (3.13) we can conclude the result. □

4 Simulation Studies

In order to study the distributions of $\tilde{\lambda}_1$ and $\hat{\lambda}_1$, we used computer simulations. We set $\mathbf{\Sigma} = \text{diag}(\lambda_1, ..., \lambda_p)$ with $\lambda_1 = p^{2/3}$ and $\lambda_2 = \cdots = \lambda_p = 1$. We considered the cases of $p = 20$, 100, 500 and 2500 when (a) $n = 5$ and (b) $n = 25$. We generated $\boldsymbol{x}_j$, $j = 1, ..., n$, independently from the p-dimensional normal distribution, $N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$. Note that (A-iii) and (A-iv) hold, however, '$\kappa/\lambda_1 = o(1)$' does not hold. We denoted independent pseudorandom 2000 observations of $\tilde{\lambda}_1$ and $\hat{\lambda}_1$ by $\tilde{\lambda}_{1r}$ and $\hat{\lambda}_{1r}$ for $r = 1, ..., 2000$. In the end of the rth replication, we checked whether the event, $(n-1)\tilde{\lambda}_{1r}/\lambda_1 \leq a_{n-1}$, is true (or false) and defined $P_{ir} = 1$ (or 0) accordingly, where a_{n-1} is the upper 0.05 point of χ^2_{n-1}. We calculated $\overline{P}(0.95) = \sum_{r=1}^{2000} P_r/2000$ as an estimate of $P\{(n-1)\tilde{\lambda}_1/\lambda_1 \leq a_{n-1}\}$. Note that the standard deviation of the estimates is less than 0.011. As for $\hat{\lambda}_1$ as well, we calculated $\overline{P}(0.95) = \sum_{r=1}^{2000} P_r/2000$ similarly as an estimate of $P\{(n-1)\hat{\lambda}_1/\lambda_1 \leq a_{n-1}\}$.

In Fig. 3, we gave the histograms of $(n-1)\tilde{\lambda}_1/\lambda_1$ (left panel) and $(n-1)\hat{\lambda}_1/\lambda_1$ (right panel) together with $\overline{P}(0.95)$ for $p = 20$, 100, 500 and 2500 when (a) $n = 5$ and (b) $n = 25$. From Corollary 3.1 and Theorem 3.1, we displayed the asymptotic probability density of $(n-1)\tilde{\lambda}_1/\lambda_1$ (or $(n-1)\hat{\lambda}_1/\lambda_1$) and χ^2_{n-1}. We observed that the histograms of $(n-1)\tilde{\lambda}_1/\lambda_1$ become close to χ^2_{n-1} as p increases even when $n = 5$. On the other hand, the histograms of $(n-1)\hat{\lambda}_1/\lambda_1$ became separated from χ^2_{n-1} as p increases especially when $n = 5$. That is because the third term in (3.1) becomes large as p increases. The NR estimator, $\tilde{\lambda}_1$, gives a good approximation to the asymptotic distribution in such a case as well by removing the term as in (3.1).

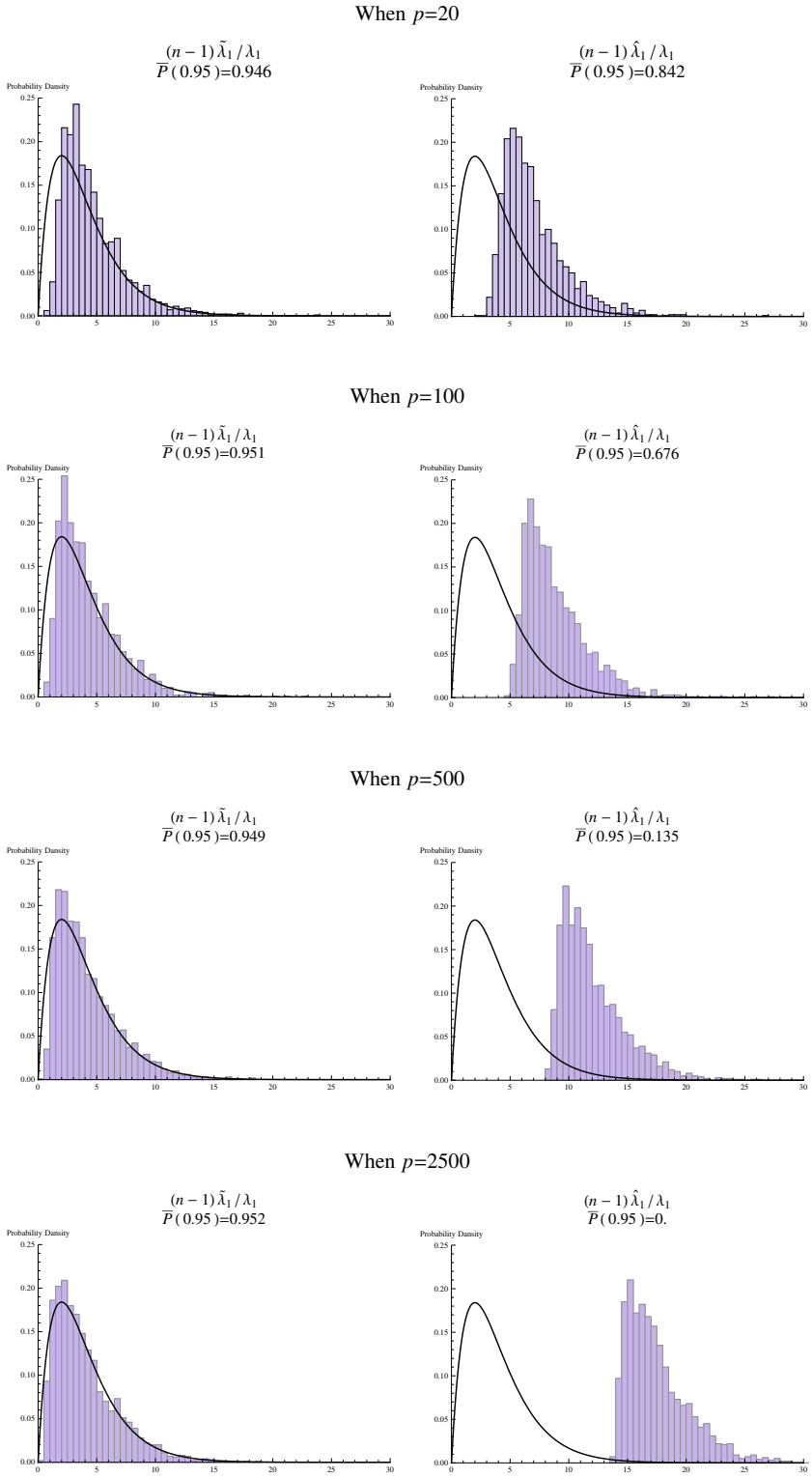


Figure 3. (a) The histograms of $(n-1)\tilde{\lambda}_1/\lambda_1$ (left panels) and $(n-1)\hat{\lambda}_1/\lambda_1$ (right panels) together with the probability density of χ^2_{n-1} and $\overline{P}(0.95)$ for $p = 20, 100, 500$ and 2500 when $n = 5$.

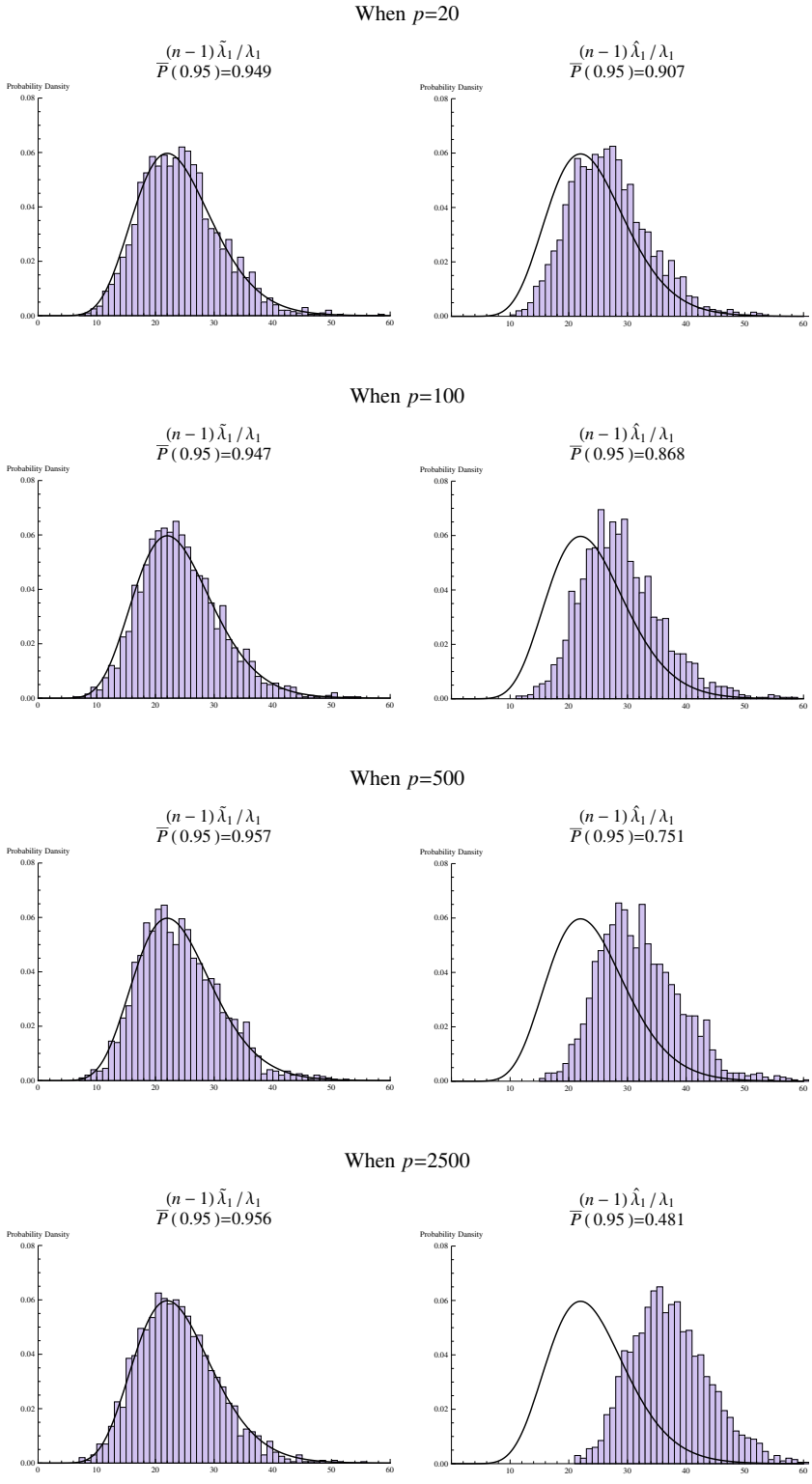


Figure 3. (b) The histograms of $(n-1)\tilde{\lambda}_1/\lambda_1$ (left panels) and $(n-1)\hat{\lambda}_1/\lambda_1$ (right panels) together with the probability density of χ^2_{n-1} and $\overline{P}(0.95)$ for $p = 20, 100, 500$ and 2500 when $n = 25$.

Chapter 2

Applications of the First Eigenvalue

In this chapter, we give several applications of the first eigenvalue. This chapter is organized by Ishii et al. [20].

In Section 1, we consider a confidence interval of the first contribution ratio. Since we analyzed the asymptotic behavior of noise space in Chapter 1, we can construct the confidence interval. We also apply the result to actual microarray data sets.

In Section 2, we consider the first eigenvector. As mentioned in Chapter 1, the first principal component contains the most important information for high-dimensional data. We give asymptotic properties of the conventional estimator and explain the reason why it behaves incorrectly in the HDLSS context. Instead, we apply the NR method to the first eigenvector. We give asymptotic properties of the NR estimator and show that it gives preferable performances.

In Section 3, we consider the first PC score. We give asymptotic properties of the NR estimator and show that it gives preferable performances. We also give a method to check the validity of the assumption required in Chapters 1–4.

In Section 4, we consider the one-sample test for a mean vector in the HDLSS context. We give a new test procedure based on the noise space.

Finally in Section 5, we summarize simulation studies of the findings.

1 Introduction

Suppose we have a $p \times n$ data matrix, $\boldsymbol{X}_{(p)} = [\boldsymbol{x}_{1(p)}, ..., \boldsymbol{x}_{n(p)}]$, where $\boldsymbol{x}_{j(p)} = (x_{1j(p)}, ..., x_{pj(p)})^T,\ j = 1, ..., n$, are independent and identically distributed (i.i.d.) as a p-dimensional distribution with a mean vector $\boldsymbol{\mu}_p$ and covariance matrix $\boldsymbol{\Sigma}_p\ (\geq \boldsymbol{O})$. We assume $n \geq 3$. The eigen-decomposition of $\boldsymbol{\Sigma}_p$ is given by $\boldsymbol{\Sigma}_p = \boldsymbol{H}_p \boldsymbol{\Lambda}_p \boldsymbol{H}_p^T$, where $\boldsymbol{\Lambda}_p$ =diag$(\lambda_{1(p)}, ..., \lambda_{p(p)})$ is a diagonal matrix of eigenvalues, $\lambda_{1(p)} \geq \cdots \geq \lambda_{p(p)} (\geq 0)$, and $\boldsymbol{H}_p = [\boldsymbol{h}_{1(p)}, ..., \boldsymbol{h}_{p(p)}]$ is an orthogonal matrix of the corresponding eigenvectors. Let $\boldsymbol{X}_{(p)} - [\boldsymbol{\mu}_p, ..., \boldsymbol{\mu}_p] = \boldsymbol{H}_p \boldsymbol{\Lambda}_p^{1/2} \boldsymbol{Z}_{(p)}$. Then, $\boldsymbol{Z}_{(p)}$ is a $p \times n$ sphered data matrix from a distribution with the zero mean and the identity covariance matrix. Let $\boldsymbol{Z}_{(p)} = [\boldsymbol{z}_{1(p)}, ..., \boldsymbol{z}_{p(p)}]^T$ and $\boldsymbol{z}_{i(p)} = (z_{i1(p)}, ..., z_{in(p)})^T,\ i = 1, ..., p$. Note that $E(z_{ij(p)} z_{i'j(p)}) = 0\ (i \neq i')$ and Var$(\boldsymbol{z}_{i(p)}) = \boldsymbol{I}_n$, where $\boldsymbol{I}_n$ is the n-dimensional identity matrix. The i-th true PC score of $\boldsymbol{x}_{j(p)}$ is given by $\boldsymbol{h}_{i(p)}^T (\boldsymbol{x}_{j(p)} - \boldsymbol{\mu}_p) = \lambda_{i(p)}^{1/2} z_{ij(p)}$ (hereafter called $s_{ij(p)}$). Note that Var$(s_{ij(p)}) = \lambda_{i(p)}$ for all i, j. Hereafter, the subscript p will be omitted for the sake of simplicity when it does not cause any confusion. Let $\boldsymbol{z}_{oi} = \boldsymbol{z}_i - (\bar{z}_i, ..., \bar{z}_i)^T,\ i = 1, ..., p$, where $\bar{z}_i = n^{-1} \sum_{k=1}^n z_{ik}$. We assume that λ_1 has multiplicity one in the sense that $\liminf_{p \to \infty} \lambda_1 / \lambda_2 > 1$. Also, we assume that $\limsup_{p \to \infty} E(z_{ij}^4) < \infty$ for all i, j and $P(\lim_{p \to \infty} ||\boldsymbol{z}_{o1}|| \neq 0) = 1$. Note that if $\boldsymbol{X}$ is Gaussian, z_{ij}s are i.i.d. as the standard normal distribution, $N(0, 1)$. Let $\delta_i = \text{tr}(\boldsymbol{\Sigma}^2) - \sum_{s=1}^i \lambda_s^2 = \sum_{s=i+1}^p \lambda_s^2$ for $i = 1, ..., p-1$. We consider the same assumptions in Chapter 1 for the first eigenvalue:

(A-i) $\dfrac{\delta_1}{\lambda_1^2} = o(1)$ as $p \to \infty$ when n is fixed; $\dfrac{\delta_{i_*}}{\lambda_1^2} = o(1)$ as $p \to \infty$ for some fixed $i_*\ (< p)$ when $n \to \infty$.

(A-ii) $\dfrac{\sum_{r,s \geq 2}^p \lambda_r \lambda_s E\{(z_{rk}^2 - 1)(z_{sk}^2 - 1)\}}{n \lambda_1^2} = o(1)$ as $p \to \infty$ either when n is fixed or $n \to \infty$.

Note that (A-i) implies the conditions that $\lambda_2 / \lambda_1 \to 0$ as $p \to \infty$ when n is fixed and $\lambda_{i_*+1} / \lambda_1 \to 0$ as $p \to \infty$ for some fixed i_* when $n \to \infty$. Also, note that (A-ii) holds when $\boldsymbol{X}$ is Gaussian and (A-i) is met. See Remark 3.2 in Chapter 1. Let $\kappa = \sum_{s=2}^p \lambda_s$. As mentioned in Chapter 1, we assume $\liminf_{p \to \infty} \kappa / \lambda_1 > 0$. As necessary, we consider the following assumption for the normalized first PC scores, $z_{1j}\ (= s_{1j} / \lambda_1^{1/2}),\ j = 1, ..., n$:

(A-iii) $z_{1j},\ j = 1, ..., n$, are i.i.d. as $N(0, 1)$.

Note that $P(\lim_{p \to \infty} ||\boldsymbol{z}_{o1}|| \neq 0) = 1$ under (A-iii) from the fact that $||\boldsymbol{z}_{o1}||^2$ is distributed as χ_{n-1}^2, where χ_ν^2 denotes a random variable distributed as χ^2 distribution with ν degrees of freedom. Let us write the sample covariance matrix as $\boldsymbol{S} = (n-1)^{-1} (\boldsymbol{X} - \overline{\boldsymbol{X}})(\boldsymbol{X} - \overline{\boldsymbol{X}})^T = (n-1)^{-1} \sum_{j=1}^n (\boldsymbol{x}_j - \bar{\boldsymbol{x}})(\boldsymbol{x}_j - \bar{\boldsymbol{x}})^T$, where $\overline{\boldsymbol{X}} = [\bar{\boldsymbol{x}}, ..., \bar{\boldsymbol{x}}]$ and $\bar{\boldsymbol{x}} = \sum_{j=1}^n \boldsymbol{x}_j / n$. Then, we define the $n \times n$ dual sample covariance matrix by $\boldsymbol{S}_D = (n-1)^{-1} (\boldsymbol{X} - \overline{\boldsymbol{X}})^T (\boldsymbol{X} - \overline{\boldsymbol{X}})$. Let $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_{n-1} \geq 0$ be the eigenvalues of $\boldsymbol{S}_D$. Let us write the eigen-decomposition of $\boldsymbol{S}_D$ as $\boldsymbol{S}_D = \sum_{j=1}^{n-1} \hat{\lambda}_j \hat{\boldsymbol{u}}_j \hat{\boldsymbol{u}}_j^T$, where $\hat{\boldsymbol{u}}_j = (\hat{u}_{j1}, ..., \hat{u}_{jn})^T$ denotes a unit eigenvector corresponding to $\hat{\lambda}_j$. Note that $\boldsymbol{S}$ and $\boldsymbol{S}_D$ share non-zero eigenvalues. Also, note that tr$(\boldsymbol{S})$ = tr$(\boldsymbol{S}_D)$.

Here, we emphasize that the first principal component is quite important for high-dimensional data because λ_1 often becomes much larger than the other eigenvalues as p increases in the sense that $\lambda_j / \lambda_1 \to 0$

as $p \to \infty$ for all $j \geq 2$. See Figure 1 in Yata and Aoshima (2013) or Table 1 in Section 2 for example. In other words, the first principal component contains much useful information about high-dimensional data sets. In addition, λ_1 and $\boldsymbol{h}_1$ can be accurately estimated for high-dimensional data by using the NR methodology even when n is fixed. It is likely that the first principal component is applicable to high-dimensional statistical inferences such as tests of mean vectors and covariance matrices. That is the reason why we focus on the first principal component.

2 Confidence Interval of the First Contribution Ratio

We consider a confidence interval for the contribution ratio of the first principal component. Let a and b be constants satisfying $P(a \leq \chi^2_{n-1} \leq b) = 1-\alpha$, where $\alpha \in (0,1)$. Then, from Theorem 3.1 in Chapter 1, under (A-i) to (A-iii), it holds that

$$\begin{aligned} &P\Big(\frac{\lambda_1}{\mathrm{tr}(\boldsymbol{\Sigma})} \in \Big[\frac{(n-1)\tilde{\lambda}_1}{b\kappa+(n-1)\tilde{\lambda}_1}, \frac{(n-1)\tilde{\lambda}_1}{a\kappa+(n-1)\tilde{\lambda}_1}\Big]\Big) \\ &= P\Big(a \leq (n-1)\frac{\tilde{\lambda}_1}{\lambda_1} \leq b\Big) = 1-\alpha+o(1) \end{aligned} \tag{2.1}$$

as $p \to \infty$ when n is fixed. We need to estimate κ in (2.1). Here, we give a consistent estimator of κ by $\tilde{\kappa} = (n-1)(\mathrm{tr}(\boldsymbol{S}_D) - \hat{\lambda}_1)/(n-2) = \mathrm{tr}(\boldsymbol{S}_D) - \tilde{\lambda}_1$. Then, we have the following results.

Lemma 2.1. *Under (A-i) and (A-ii), it holds that*

$$\frac{\tilde{\kappa}}{\kappa} = 1+o_p(1) \quad \textit{and} \quad \frac{\tilde{\kappa}}{\lambda_1} = \frac{\kappa}{\lambda_1}+o_p(1)$$

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

Proof. By using Markov's inequality, for any $\tau > 0$, under (A-i) and (A-ii), we have that

$$\begin{aligned} &P\Big\{\Big(\sum_{s=2}^{p}\frac{\lambda_s\{||\boldsymbol{z}_{os}||^2-(n-1)\}}{(n-1)\lambda_1}\Big)^2 > \tau\Big\} \\ &= P\Big\{\Big(\sum_{s=2}^{p}\frac{\lambda_s\{(n-1)\sum_{k=1}^{n}(z_{sk}^2-1)/n - \sum_{k\neq k'}^{n} z_{sk}z_{sk'}/n\}}{(n-1)\lambda_1}\Big)^2 > \tau\Big\} \\ &= O\Big\{\frac{\sum_{r,s\geq 2}^{p}\lambda_r\lambda_s E\{(z_{rk}^2-1)(z_{sk}^2-1)\}}{n\lambda_1^2}\Big\} + O\{\delta_1/(n\lambda_1)^2\} \to 0 \end{aligned}$$

as $p \to \infty$ either when n is fixed or $n \to \infty$. Thus it holds that $\mathrm{tr}(\boldsymbol{S}_D)/\lambda_1 = \kappa/\lambda_1 + ||\boldsymbol{z}_{o1}/\sqrt{n-1}||^2 + o_p(1)$ from the fact that $\mathrm{tr}(\boldsymbol{S}_D) = \lambda_1||\boldsymbol{z}_{o1}||^2/(n-1) + \sum_{s=2}^{p}\lambda_s||\boldsymbol{z}_{os}||^2/(n-1)$. Then, from Proposition 3.1 in Chapter 1 and $\liminf_{p\to\infty}\kappa/\lambda_1 > 0$, we can claim the results. □

Theorem 2.1. *Under (A-i) to (A-iii), it holds that*

$$P\Big(\frac{\lambda_1}{tr(\boldsymbol{\Sigma})} \in \Big[\frac{(n-1)\tilde{\lambda}_1}{b\tilde{\kappa}+(n-1)\tilde{\lambda}_1}, \frac{(n-1)\tilde{\lambda}_1}{a\tilde{\kappa}+(n-1)\tilde{\lambda}_1}\Big]\Big) = 1-\alpha+o(1) \tag{2.2}$$

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

Proof. From Theorem 3.1 in Chapter 1 and Lemma 2.1, under (A-i) to (A-iii), it holds that

$$\begin{aligned}
&P\Big(\frac{\lambda_1}{\mathrm{tr}(\boldsymbol{\Sigma})} \in \Big[\frac{(n-1)\tilde{\lambda}_1}{b\tilde{\kappa}+(n-1)\tilde{\lambda}_1}, \frac{(n-1)\tilde{\lambda}_1}{a\tilde{\kappa}+(n-1)\tilde{\lambda}_1}\Big]\Big) \\
&= P\Big(\frac{(n-1)\tilde{\lambda}_1}{b\tilde{\kappa}+(n-1)\tilde{\lambda}_1} \le \frac{\lambda_1}{\mathrm{tr}(\boldsymbol{\Sigma})} \le \frac{(n-1)\tilde{\lambda}_1}{a\tilde{\kappa}+(n-1)\tilde{\lambda}_1}\Big) \\
&= P\Big(\frac{a\tilde{\kappa}}{(n-1)\tilde{\lambda}_1} \le \frac{\kappa}{\lambda_1} \le \frac{b\tilde{\kappa}}{(n-1)\tilde{\lambda}_1}\Big) = P\Big(a \le (n-1)\frac{\tilde{\lambda}_1\kappa}{\lambda_1\tilde{\kappa}} \le b\Big) \\
&= 1-\alpha+o(1)
\end{aligned}$$

as $p \to \infty$ when n is fixed. It concludes the result. □

Remark 2.1. From Theorem 3.1 in Chapter 1 and Lemma 2.1, under (A-i) and (A-ii), it holds that $\mathrm{tr}(\boldsymbol{S}_D)/\mathrm{tr}(\boldsymbol{\Sigma}) = (\tilde{\kappa}+\tilde{\lambda}_1)/\mathrm{tr}(\boldsymbol{\Sigma}) = 1+o_p(1)$ as $p \to \infty$ and $n \to \infty$. We have that

$$\frac{\tilde{\lambda}_1}{\mathrm{tr}(\boldsymbol{S}_D)} = \frac{\lambda_1}{\mathrm{tr}(\boldsymbol{\Sigma})}\{1+o_p(1)\}.$$

Remark 2.2. The constants $(a,\ b)$ should be chosen for (2.2) to have the minimum length. If $\lambda_1/\kappa = o(1)$, the length of the confidence interval becomes close to $\{(n-1)\tilde{\lambda}_1/\tilde{\kappa}\}(1/a-1/b)$ under (A-i) and (A-ii) when $p \to \infty$ and n is fixed. Thus, we recommend to choose constants $(a,\ b)$ such that

$$\underset{a,b}{\mathrm{argmin}}(1/a-1/b) \quad \text{subject to } \ G_{n-1}(b)-G_{n-1}(a) = 1-\alpha,$$

where $G_{n-1}(\cdot)$ denotes the c.d.f. of χ^2_{n-1}.

We used gene expression data sets and constructed a confidence interval for the contribution ratio of the first principal component. The microarray data sets were as follows: Lymphoma data with $7129\ (=p)$ genes consisting of diffuse large B-cell (DLBC) lymphoma (58 samples) and follicular lymphoma (19 samples) given by Shipp et al. [29]; and prostate cancer data with $12625\ (=p)$ genes consisting of normal prostate (50 samples) and prostate tumor (52 samples) given by Singh et al. [30]. The data sets are given in Jeffery et al. [23]. We standardized each sample so as to have the unit variance. Then, it holds that $\mathrm{tr}(\boldsymbol{S})\ (=\mathrm{tr}(\boldsymbol{S}_D)) = p$, so that $\tilde{\lambda}_1+\tilde{\kappa} = p$. We gave estimates of the first five eigenvalues by $\hat{\lambda}_j$s and $\tilde{\lambda}_j$s in Table 1. We observed that the first eigenvalues are much larger than the others especially for prostate cancer data. We also observed

that $\hat{\lambda}_j$ was larger than $\tilde{\lambda}_j$ for $j = 1, ..., 5$, as expected theoretically from the fact that $\hat{\lambda}_j/\tilde{\lambda}_j \geq 0$ w.p.1 for all j. We considered an estimator of δ_1 by $\tilde{\delta}_1 = W_n - \tilde{\lambda}_1^2$ having W_n by (4) in Aoshima and Yata [7], where W_n is an unbiased and consistent estimator of tr($\mathbf{\Sigma}^2$). We calculated that $\tilde{\delta}_1/\tilde{\lambda}_1^2 = 0.163$ for DLBC lymphoma, $\tilde{\delta}_1/\tilde{\lambda}_1^2 = -0.082$ for follicular lymphoma, $\tilde{\delta}_1/\tilde{\lambda}_1^2 = -0.245$ for normal prostate and $\tilde{\delta}_1/\tilde{\lambda}_1^2 = -0.235$ for prostate tumor. From these observations, we concluded that these data sets satisfy (A-i). In addition, from Remark 4.1 given in Section 4, by using Jarque-Bera test, we could confirm that these data sets satisfy (A-iii) with the level of significance 0.05. Hence, from Theorem 2.1, we constructed a 95% confidence interval of the first contribution ratio for each data set by choosing $(a,\ b)$ as in Remark 2.2. The results are summarized in Table 2.

Table 1. Estimates of the first five eigenvalues by $\hat{\lambda}_j$s and $\tilde{\lambda}_j$s, for the microarray data sets.

	n	$\hat{\lambda}_1,\ \hat{\lambda}_2,\ \hat{\lambda}_3,\ \hat{\lambda}_4,\ \hat{\lambda}_5$	$\tilde{\lambda}_1,\ \tilde{\lambda}_2,\ \tilde{\lambda}_3,\ \tilde{\lambda}_4,\ \tilde{\lambda}_5$
Lymphoma data with $7129\ (= p)$ genes given by Shipp et al. [29]			
DLBC	58	1862, 564, 490, 398, 324	1768, 479, 412, 326, 257
Follicular	19	2476, 704, 614, 533, 369	2203, 457, 392, 333, 182
Prostate cancer data with $12625\ (= p)$ genes given by Singh et al [30]			
Normal	50	6760, 562, 426, 371, 304	6637, 450, 320, 271, 209
Prostate	52	6106, 687, 512, 462, 298	5976, 568, 401, 359, 199

Table 2. The 95% confidence interval (CI) of the first contribution ratio, together with $\tilde{\lambda}_1$ and $\tilde{\kappa}$, for the microarray data sets.

	(n, p)	CI	$\tilde{\lambda}_1$	$\tilde{\kappa}$
DLBC lymphoma	$(58, 7129)$	$[0.183, 0.322]$	1768	5361
Follicular lymphoma	$(19, 7129)$	$[0.178, 0.467]$	2203	4926
Normal prostate	$(50, 12625)$	$[0.422, 0.622]$	6637	5988
Prostate tumor	$(52, 12625)$	$[0.374, 0.569]$	5976	6649

3 First PC Direction Vector

In this section, we give asymptotic properties of the first PC direction in the HDLSS context. Let $\hat{\boldsymbol{H}} = [\hat{\boldsymbol{h}}_1, ..., \hat{\boldsymbol{h}}_p]$, where $\hat{\boldsymbol{H}}$ is a $p \times p$ orthogonal matrix of the sample eigenvectors such that $\hat{\boldsymbol{H}}^T \boldsymbol{S} \hat{\boldsymbol{H}} = \hat{\boldsymbol{\Lambda}}$ having

$\hat{\boldsymbol{\Lambda}} = \mathrm{diag}(\hat{\lambda}_1, ..., \hat{\lambda}_p)$. We assume $\boldsymbol{h}_i^T \hat{\boldsymbol{h}}_i \geq 0$ w.p.1 for all i without loss of generality. Note that $\hat{\boldsymbol{h}}_i$ can be calculated by $\hat{\boldsymbol{h}}_i = \{(n-1)\hat{\lambda}_i\}^{-1/2}(\boldsymbol{X} - \overline{\boldsymbol{X}})\hat{\boldsymbol{u}}_i$. First, we have the following result.

Lemma 3.1. *Under (A-i) and (A-ii), it holds that*

$$\hat{\boldsymbol{h}}_1^T \boldsymbol{h}_1 - \left(1 + \frac{\kappa}{\lambda_1 ||\boldsymbol{z}_{o1}||^2}\right)^{-1/2} = o_p(1)$$

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

Proofs. With the help of Proposition 3.1 in Chapter 1, under (A-i) and (A-ii), it holds that from (4.1) in the proof of Lemma 4.1

$$\begin{aligned}\boldsymbol{h}_1^T \hat{\boldsymbol{h}}_1 = \frac{\boldsymbol{h}_1^T(\boldsymbol{X} - \overline{\boldsymbol{X}})\hat{\boldsymbol{u}}_1}{\{(n-1)\hat{\lambda}_1\}^{1/2}} = \frac{\lambda_1^{1/2} \boldsymbol{z}_{o1}^T \hat{\boldsymbol{u}}_1}{\{(n-1)\hat{\lambda}_1\}^{1/2}} &= \frac{||\boldsymbol{z}_{o1}|| + o_p(n^{1/2})}{\{||\boldsymbol{z}_{o1}||^2 + \kappa/\lambda_1 + o_p(n)\}^{1/2}} \\ &= \frac{1}{\{1 + \kappa/(\lambda_1 ||\boldsymbol{z}_{o1}||^2)\}^{1/2}} + o_p(1)\end{aligned}$$

as $p \to \infty$ either when n is fixed or $n \to \infty$. It concludes the result. □

If $\kappa/(n\lambda_1) = o(1)$ as $p \to \infty$ and $n \to \infty$, $\hat{\boldsymbol{h}}_1$ is a consistent estimator of $\boldsymbol{h}_1$ in the sense that $\hat{\boldsymbol{h}}_1^T \boldsymbol{h}_1 = 1 + o_p(1)$. When n is fixed, $\hat{\boldsymbol{h}}_1$ is not a consistent estimator because $\liminf_{p\to\infty} \kappa/\lambda_1 > 0$. In order to overcome this inconvenience, we consider applying the NR methodology to the PC direction vector. Let $\tilde{\boldsymbol{h}}_i = \{(n-1)\tilde{\lambda}_i\}^{-1/2}(\boldsymbol{X} - \overline{\boldsymbol{X}})\hat{\boldsymbol{u}}_i$. From Lemma 3.1 we have the following result.

Theorem 3.1. *Under (A-i) and (A-ii), it holds that*

$$\tilde{\boldsymbol{h}}_1^T \boldsymbol{h}_1 = 1 + o_p(1)$$

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

Proof. With the help of Theorem 3.1 in Chapter 1, under (A-i) and (A-ii), we have that from (4.1) in the proof of Lemma 4.1

$$\boldsymbol{h}_1^T \tilde{\boldsymbol{h}}_1 = \frac{\boldsymbol{h}_1^T(\boldsymbol{X} - \overline{\boldsymbol{X}})\hat{\boldsymbol{u}}_1}{\{(n-1)\tilde{\lambda}_1\}^{1/2}} = \frac{||\boldsymbol{z}_{o1}|| + o_p(n^{1/2})}{\{||\boldsymbol{z}_{o1}||^2 + o_p(n)\}^{1/2}} = 1 + o_p(1)$$

as $p \to \infty$ either when n is fixed or $n \to \infty$. It concludes the result. □

Note that $||\tilde{\boldsymbol{h}}_1||^2 = \hat{\lambda}_1/\tilde{\lambda}_1 \geq 1$ w.p.1. We emphasize that $\tilde{\boldsymbol{h}}_1$ is a consistent estimator of $\boldsymbol{h}_1$ in the sense of the inner product even when n is fixed though $\tilde{\boldsymbol{h}}_1$ is not a unit vector. We give an application of $\tilde{\boldsymbol{h}}_1$ in Chapter 3. Let us introduce an illustrative example of Lemma 3.1. In Fig.1, the sphere represents the space of all possible sample eigenvectors with the first three eigenvectors as the coordinate axes. From Lemma 3.1 the angle of $\hat{\boldsymbol{h}}_1$ and $\boldsymbol{h}_1$ becomes $\pi/2$ in the worst case.

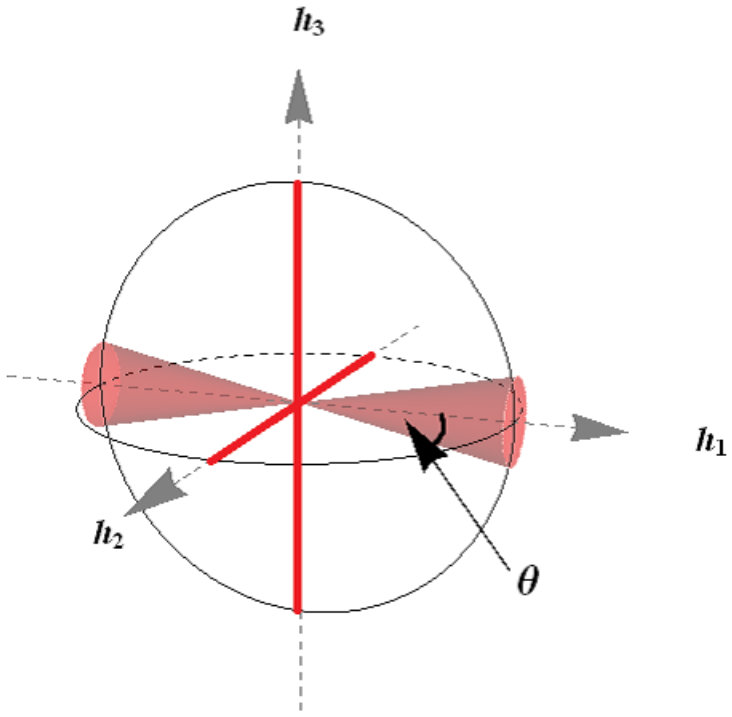


Figure 1. Geometric representation of the first PC direction. The sphere represents the space of possible sample eigenvectors. The first sample eigenvector, $\hat{\boldsymbol{h}}_1$, tends to lie in the red cone, with the θ angle. In the worst case, the angle becomes $\pi/2$ as represented by the red solid lines.

4 First PC Score

In this section, we give asymptotic properties of the first PC score in HDLSS context. We consider the first PC score that plays a decisive role for classification of HDLSS data. We note that the first PC score is given by $s_{1j} = \lambda_1^{1/2} z_{1j},\ j = 1, ..., n$. Let $z_{oij} = z_{ij} - \bar{z}_i$ for all i, j. Note that $\boldsymbol{z}_{oi} = (z_{oi1}, ..., z_{oin})^T$ for all i. First, we have the following result.

Lemma 4.1. *Under (A-i) and (A-ii), it holds that*

$$\hat{u}_{1j} = z_{o1j}/||\boldsymbol{z}_{o1}|| + o_p(1) \quad \textit{for } j = 1, ..., n$$

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

Proof. We note that $||\boldsymbol{z}_{o1}||^2/n = 1 + o_p(1)$ as $n \to \infty$. From (3.5) in Chapter 1, under (A-i) and (A-ii), we have that

$$\hat{\boldsymbol{u}}_1^T \boldsymbol{z}_{o1}/||\boldsymbol{z}_{o1}|| = 1 + o_p(1) \tag{4.1}$$

as $p \to \infty$ either when n is fixed or $n \to \infty$, so that $\hat{\boldsymbol{u}}_1^T \boldsymbol{z}_{o1} = ||\boldsymbol{z}_{o1}|| + o_p(n^{1/2})$. Thus, we can claim the result. □

Remark 4.1. From Lemma 4.1, by using $\hat{u}_{1j}$s and the test of normality such as Jarque-Bera test, one can check whether (A-iii) holds or not.

By applying the NR methodology to the first PC score, we obtain an estimate by $\tilde{s}_{1j} = \sqrt{(n-1)\tilde{\lambda}_1}\hat{u}_{1j},\ j = 1, ..., n$. A sample mean squared error of the first PC score is given by $\text{MSE}(\tilde{s}_1) = n^{-1}\sum_{j=1}^{n}(\tilde{s}_{1j} - s_{1j})^2$. Then, from Theorem 3.1 in Chapter 1 and Lemma 4.1, we have the following result.

Theorem 4.1. *Under (A-i) and (A-ii), it holds that*

$$\frac{1}{\sqrt{\lambda_1}}(\tilde{s}_{1j} - s_{1j}) = -\bar{z}_1 + o_p(1) \quad \textit{for } j = 1, ..., n$$

as $p \to \infty$ when n is fixed. Under (A-i) to (A-iii), it holds that

$$\sqrt{\frac{n}{\lambda_1}}(\tilde{s}_{1j} - s_{1j}) \Rightarrow N(0,1) \quad \textit{for } j = 1, ..., n; \quad \textit{and} \quad n\frac{\textit{MSE}(\tilde{s}_1)}{\lambda_1} \Rightarrow \chi^2_1$$

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

Proof. By combing Theorem 3.1 in Chapter 1 with Lemma 4.1, under (A-i) and (A-ii), we have that

$$\tilde{s}_{1j}/\sqrt{\lambda_1} = \hat{u}_{1j}\sqrt{(n-1)\tilde{\lambda}_1/\lambda_1} = \hat{u}_{1j}||\boldsymbol{z}_{o1}|| + o_p(1) = z_{o1j} + o_p(1)$$

as $p \to \infty$ when n is fixed. By noting that $z_{o1j} = z_{1j} - \bar{z}_1$ and $\bar{z}_1$ is distributed as $N(0, 1/n)$ under (A-iii), we have the results. □

Remark 4.2. The conventional estimator of the first PC score is given by $\hat{s}_{1j} = \sqrt{(n-1)\hat{\lambda}_1}\hat{u}_{1j},\ j = 1, ..., n$. From Theorems 8.1 and 8.2 in Yata and Aoshima [40], under (A-i) and (A-ii), it holds that as $p \to \infty$ and $n \to \infty$

$$\frac{\text{MSE}(\hat{s}_1)}{\lambda_1} = o_p(1) \ \text{ if } \kappa/(n\lambda_1) = o(1), \quad \text{and} \quad \frac{\text{MSE}(\tilde{s}_1)}{\lambda_1} = o_p(1).$$

5 One-Sample Test for the Mean Vector

In this section, we consider the following one-sample test for the mean vector:

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 \quad \text{vs.} \quad H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0 \tag{5.1}$$

where $\boldsymbol{\mu}_0$ is a candidate mean vector such as $\boldsymbol{\mu}_0 = \boldsymbol{0}$. Here, we have the following result.

Lemma 5.1. *Under (A-i), it holds that*

$$\frac{||\bar{\boldsymbol{x}} - \boldsymbol{\mu}||^2 - tr(\boldsymbol{S}_D)/n}{\lambda_1} = \bar{z}_1^2 - \frac{||\boldsymbol{z}_{o1}/\sqrt{n-1}||^2}{n} + o_p(1)$$

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

Proof. We write that

$$n||\bar{\boldsymbol{x}} - \boldsymbol{\mu}||^2 - \mathrm{tr}(\boldsymbol{S}_D) = \sum_{s=1}^{p} \lambda_s \Big(n\bar{z}_s^2 - \sum_{j=1}^{n} \frac{(z_{sj} - \bar{z}_s)^2}{n-1} \Big).$$

Then, from (3.2) in the proof of Proposition 3.1 in Chapter 1 and $n\bar{z}_1^2 - \sum_{j=1}^{n}(z_{sj} - \bar{z}_s)^2/(n-1) = \sum_{j \neq j'}^{n} z_{sj} z_{sj'}/(n-1)$ for all s, under (A-i), we have that

$$\{||\bar{\boldsymbol{x}} - \boldsymbol{\mu}||^2 - \mathrm{tr}(\boldsymbol{S}_D)/n\}/\lambda_1 = \bar{z}_1^2 - ||\boldsymbol{z}_{o1}/\sqrt{n-1}||^2/n + o_p(1)$$

as $p \to \infty$ when n is fixed. It concludes the result. □

Let

$$F_0 = \frac{n||\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0||^2 - \mathrm{tr}(\boldsymbol{S}_D)}{\tilde{\lambda}_1} + 1.$$

Note that $E(\tilde{\lambda}_1(F_0 - 1)/n) = ||\boldsymbol{\mu} - \boldsymbol{\mu}_0||^2$. Then, by combining Theorem 3.1 in Chapter 1 and Lemma 5.1, we have the following result.

Theorem 5.1. *Under (A-i) to (A-iii), it holds that*

$$F_0 \Rightarrow F_{1,n-1} \text{ under } H_0 \text{ in (5.1)}$$

as $p \to \infty$ when n is fixed, where F_{ν_1,ν_2} denotes a random variable distributed as F distribution with degrees of freedom, ν_1 and ν_2.

Proof. Under (A-iii), we note that $\bar{z}_1$ and $\boldsymbol{z}_{o1}$ are independent, and $n\bar{z}_1^2$ is distributed as χ_1^2. Then, from Theorem 3.1 in Chapter 1 and Lemma 5.1 we can conclude the result. □

For a given $\alpha \in (0, 1/2)$ we test (5.1) by

$$\text{rejecting } H_0 \Longleftrightarrow F_0 > F_{1,n-1}(\alpha),$$

where $F_{\nu_1,\nu_2}(\alpha)$ denotes the upper α point of F distribution with degrees of freedom, ν_1 and ν_2. Then, under (A-i) to (A-iii), it holds that

$$\text{size} = \alpha + o(1)$$

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

For the same gene expression data as in Section 2, we tested (5.1) with $\boldsymbol{\mu}_0 = \boldsymbol{0}$ and $\alpha = 0.05$. We observed that H_0 was rejected for all four data sets.

6 Simulation Studies

In this section, we summarize the findings in Chapter 2 by using computer simulations.

6.1 Confidence interval of the first contribution ratio

In order to study the performance of the confidence interval of the contribution ratio for the first principal component by (2.2), we used computer simulations. Our goal was to construct a 95% confidence interval by (2.2), so we set $\alpha = 0.05, a = \chi^2_{n-1}(0.975)$ and $b = \chi^2_{n-1}(0.025)$, where $\chi^2_\nu(\beta)$ denotes the upper β point of χ^2_ν. We consider the cases of $p = 20, 100, 500$ and 2500 when $n = 10$. We set $\mathbf{\Sigma} = \text{diag}(\lambda_1, ..., \lambda_p)$ with $\lambda_1 = p^{2/3}$ and $\lambda_2 = \cdots = \lambda_p = 1$. We considered $\boldsymbol{x}_j$, $j = 1, ..., n$, as z_{1j} being distributed as $N(0, 1)$ and $z_{ij},\ i = 2, ..., p$, being i.i.d. as $t_{p-1}(\mathbf{0}, I_{p-1}, 5)$, where z_{1j} and $(z_{2j}, ..., z_{pj})$ are independent. Note that (A-i) and (A-ii) hold, however '$\kappa/(n\lambda_1) = o(1)$' does not hold.

Independent pseudorandom 2000 ($= R$, say) observations of $\tilde{\lambda}_1$ and $\tilde{\kappa}$ were generated from the distribution. Let $\tilde{\lambda}_{1r}$ and $\tilde{\kappa}_r$ be the rth observation of $\tilde{\lambda}_1$ and $\tilde{\kappa}$ respectively, for $r = 1, ..., R$. Let us simply write $\tilde{\lambda}_1 = R^{-1}\sum_{r=1}^R \tilde{\lambda}_{1r}$ and $\tilde{\kappa} = R^{-1}\sum_{r=1}^R \tilde{\kappa}_r$. We also considered the Monte Carlo variability. Let $\text{var}(\tilde{\lambda}_1/\lambda_1) = (R-1)^{-1}\sum_{r=1}^R(\tilde{\lambda}_{1r} - \tilde{\lambda}_1)^2/\lambda_1^2$ and $\text{var}(\tilde{\kappa}/\kappa) = (R-1)^{-1}\sum_{r=1}^R(\tilde{\kappa}_r - \tilde{\kappa})^2/\kappa^2$. In the end of the rth replication, we checked whether $\lambda_1/\text{tr}(\mathbf{\Sigma})$ does (or does not) belong to the corresponding confidence interval and defined $P_r = 1$ (or 0) accordingly. Let $\overline{P}(0.95) = R^{-1}\sum_{r=1}^R P_r$, which estimates the target coverage probability, having its estimated standard error $s\{\overline{P}(0.95)\}$, where $s^2\{\overline{P}(0.95)\} = R^{-1}\overline{P}(0.95)(1 - \overline{P}(0.95))$. In Table 3, we gave $\overline{P}(0.95)$, $s\{\overline{P}(0.95)\}$, $\tilde{\lambda}_1/\lambda_1$, $\text{var}(\tilde{\lambda}_1/\lambda_1)$, $\tilde{\kappa}/\kappa$ and $\text{var}(\tilde{\kappa}/\kappa)$. We observed from Table 3 that $\overline{P}(0.95)$s become close to 0.95 as p increases. In addition, $\text{var}(\tilde{\lambda}_1/\lambda_1)$s become close to $\text{Var}(\chi^2_{n-1}/(n-1)) = 2/(n-1) \approx 0.222$ as p increases.

Table 3. The coverage probability of the first contribution ratio, $\bar{P}$(0.95), together with $\tilde{\lambda}_1/\lambda_1$, $\hat{\kappa}/\kappa$ and their standard errors in parentheses.

p	$\bar{P}$(0.95) ($s\{\overline{P}(0.95)\}$)	$\tilde{\lambda}_1/\lambda_1$ ($\text{var}(\tilde{\lambda}_1/\lambda_1)$)	$\tilde{\kappa}/\kappa$ ($\text{var}(\tilde{\kappa}/\kappa)$)
20	0.961 (0.00430)	1.032 (0.192)	0.973 (0.00245)
100	0.963 (0.00419)	1.053 (0.218)	0.993 (0.00113)
500	0.963 (0.00422)	1.025 (0.214)	0.997 (0.00050)
2500	0.957 (0.00453)	1.018 (0.221)	0.999 (0.00022)

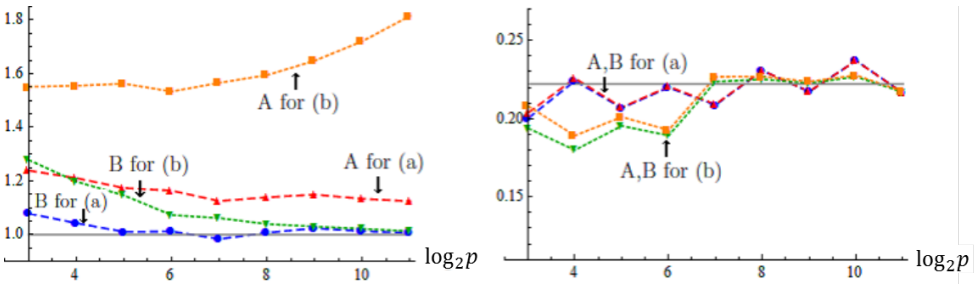
6.2 Comparison of the NR estimator and the conventional estimator

In this section, we compared the performance of $\tilde{\lambda}_1$, $\tilde{\boldsymbol{h}}_1$ and $\tilde{s}_{1j}$ with their conventional counterparts by Monte Carlo simulations. We set $p = 2^k,\ k = 3, ..., 11$ and $n = 10$. We considered two cases for λ_is:

(a) $\lambda_i = p^{1/i}$, $i = 1, ..., p$ and (b) $\lambda_i = p^{3/(2+2i)}$, $i = 1, ..., p$. Note that $\lambda_1 = p$ for (a) and $\lambda_1 = p^{3/4}$ for (b). Also, note that (A-i) holds both for (a) and (b). Let $p_* = \lceil p^{1/2} \rceil$, where $\lceil x \rceil$ denotes the smallest integer $\geq x$. We considered a non-Gaussian distribution as follows: $(z_{1j}, ..., z_{p-p_*j})^T$, $j = 1, ..., n$, are i.i.d. as $N_{p-p_*}(\mathbf{0}, \boldsymbol{I}_{p-p_*})$ and $(z_{p-p_*+1j}, ..., z_{pj})^T$, $j = 1, ..., n$, are i.i.d. as the p_*-variate t-distribution, $t_{p_*}(\mathbf{0}, \boldsymbol{I}_{p_*}, 10)$ with mean zero, covariance matrix $\boldsymbol{I}_{p_*}$ and degrees of freedom 10, where $(z_{1j}, ..., z_{p-p_*j})^T$ and $(z_{p-p_*+1j}, ..., z_{pj})^T$ are independent for each j. Note that (A-ii) and (A-iii) hold both for (a) and (b) from the fact that $\sum_{r,s\geq 2}^{p} \lambda_r \lambda_s E\{(z_{rk}^2 - 1)(z_{sk}^2 - 1)\} = 2\sum_{s=2}^{p-p_*} \lambda_s^2 + O(\sum_{r,s\geq p-p_*+1}^{p} \lambda_r \lambda_s) = o(\lambda_1^2)$.

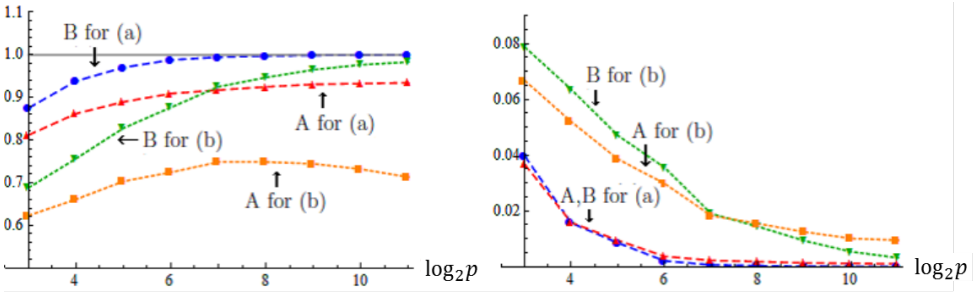
The findings were obtained by averaging the outcomes from 2000 ($= R$, say) replications. Under a fixed scenario, suppose that the r-th replication ends with estimates, ($\hat{\lambda}_{1r}$, $\hat{\boldsymbol{h}}_{1r}$, MSE$(\hat{s}_1)_r$) and ($\tilde{\lambda}_{1r}$, $\tilde{\boldsymbol{h}}_{1r}$, MSE$(\tilde{s}_1)_r$) $(r = 1, ..., R)$. Let us simply write $\hat{\lambda}_1 = R^{-1}\sum_{r=1}^{R} \hat{\lambda}_{1r}$ and $\tilde{\lambda}_1 = R^{-1}\sum_{r=1}^{R} \tilde{\lambda}_{1r}$. We also considered the Monte Carlo variability by var$(\hat{\lambda}_1/\lambda_1) = (R-1)^{-1}\sum_{r=1}^{R}(\hat{\lambda}_{1r} - \hat{\lambda}_1)^2/\lambda_1^2$ and var$(\tilde{\lambda}_1/\lambda_1) = (R-1)^{-1}\sum_{r=1}^{R}(\tilde{\lambda}_{1r} - \tilde{\lambda}_1)^2/\lambda_1^2$. Fig. 2 shows the behaviors of ($\hat{\lambda}_1/\lambda_1$, $\tilde{\lambda}_1/\lambda_1$) in the left panel and (var$(\hat{\lambda}_1/\lambda_1)$, var$(\tilde{\lambda}_1/\lambda_1)$) in the right panel for (a) and (b). We gave the asymptotic variance of $\tilde{\lambda}_1/\lambda_1$ by Var$\{\chi_{n-1}^2/(n-1)\} = 0.222$ from Theorem 3.1 in Chapter 1 and showed it by the solid line in the right panel. We observed that the sample mean and variance of $\tilde{\lambda}_1/\lambda_1$ become close to those asymptotic values as p increases.

Similarly, we plotted ($\hat{\boldsymbol{h}}_1^T\boldsymbol{h}_1$, $\tilde{\boldsymbol{h}}_1^T\boldsymbol{h}_1$) and (var$(\hat{\boldsymbol{h}}_1^T\boldsymbol{h}_1)$, var$(\tilde{\boldsymbol{h}}_1^T\boldsymbol{h}_1)$) in Fig. 3. Also, in Fig. 4, we plotted (MSE$(\hat{s}_1)/\lambda_1$, MSE$(\tilde{s}_1)/\lambda_1$) and (var(MSE$(\hat{s}_1)/\lambda_1$), var(MSE$(\tilde{s}_1)/\lambda_1$)). From Theorem 4.1 we gave the asymptotic mean of MSE$(\tilde{s}_1)/\lambda_1$ by $E(\chi_1^2/n) = 0.1$ and showed it by the solid line in the left panel of Fig. 4. We also gave the asymptotic variance of MSE$(\tilde{s}_1)/\lambda_1$ by Var$(\chi_1^2/n) = 0.02$ in the right panel of Fig. 4. Throughout, the estimators by the NR method gave good performances both for (a) and (b) when p is large. However, the conventional estimators gave poor performances especially for (b). This is probably because the bias of the conventional estimators, $\kappa/\{(n-1)\lambda_1\}$, is large for (b) compared to (a). See Proposition 3.1 in Chapter 1 for the details.



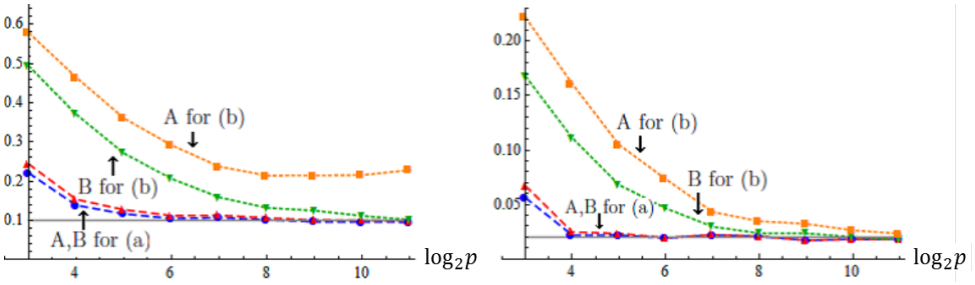
A: $\hat{\lambda}_1/\lambda_1$ and B: $\tilde{\lambda}_1/\lambda_1$ A: var$(\hat{\lambda}_1/\lambda_1)$ and B: var$(\tilde{\lambda}_1/\lambda_1)$

Figure 2. The values of A: $\hat{\lambda}_1/\lambda_1$ and B: $\tilde{\lambda}_1/\lambda_1$ are denoted by the dashed lines for (a) and by the dotted lines for (b) in the left panel. The values of A: var$(\hat{\lambda}_1/\lambda_1)$ and B: var$(\tilde{\lambda}_1/\lambda_1)$ are denoted by the dashed lines for (a) and by the dotted lines for (b) in the right panel. The asymptotic variance of $\tilde{\lambda}_1/\lambda_1$ was given by Var$\{\chi_{n-1}^2/(n-1)\} = 0.222$ and denoted by the solid line in the right panel.



A: $\hat{\boldsymbol{h}}_1^T \boldsymbol{h}_1$ and B: $\tilde{\boldsymbol{h}}_1^T \boldsymbol{h}_1$ A: var$(\hat{\boldsymbol{h}}_1^T \boldsymbol{h}_1)$ and B: var$(\tilde{\boldsymbol{h}}_1^T \boldsymbol{h}_1)$

Figure 3. The values of A: $\hat{\boldsymbol{h}}_1^T \boldsymbol{h}_1$ and B: $\tilde{\boldsymbol{h}}_1^T \boldsymbol{h}_1$ are denoted by the dashed lines for (a) and by the dotted lines for (b) in the left panel. The values of A: var$(\hat{\boldsymbol{h}}_1^T \boldsymbol{h}_1)$ and B: var$(\tilde{\boldsymbol{h}}_1^T \boldsymbol{h}_1)$ are denoted by the dashed lines for (a) and by the dotted lines for (b) in the right panel.



A: MSE$(\hat{s}_1)/\lambda_1$ and B: MSE$(\tilde{s}_1)/\lambda_1$ A: var(MSE$(\hat{s}_1)/\lambda_1$) and B: var(MSE$(\tilde{s}_1)/\lambda_1$)

Figure 4. The values of A: MSE$(\hat{s}_1)/\lambda_1$ and B: MSE$(\tilde{s}_1)/\lambda_1$ are denoted by the dashed lines for (a) and by the dotted lines for (b) in the left panel. The values of A: var(MSE$(\hat{s}_1)/\lambda_1$) and B: var(MSE$(\tilde{s}_1)/\lambda_1$) are denoted by the dashed lines for (a) and by the dotted lines for (b) in the right panel. The asymptotic mean and variance of MSE$(\tilde{s}_1)/\lambda_1$ were given by $E(\chi_1^2/n) = 0.1$ and Var$(\chi_1^2/n) = 0.02$ and denoted by the solid lines in both panels.

Chapter 3

Equality Tests of Two Covariance Matrices

In this chapter, we consider the test of equality of two covariance matrices in the HDLSS context. This chapter is organized by Ishii et al. [20] and Ishii [21].

Nowadays, it becomes more important to analyze covariance matrix structures in the HDLSS context. Even though there are a variety of tests to deal with covariance matrices when $p \to \infty$ and $n \to \infty$, there seem to be no tests available in the HDLSS context such as $p \to \infty$ while n is fixed. Some papers consider this problem only for the special covariance matrix, such as the identity matrix and the diagonal matrix. From these backgrounds we construct test procedures by using the asymptotic properties of the first eigenstructure.

In Section 2, we consider the equality of two first eigenvalues by using both of the NR method and the CDM method. We give asymptotic distributions under the null hypothesis when $p \to \infty$ while n is fixed.

In Section 3, we consider the equality of two first eigenspaces by using both of the NR method and the CDM method. By using our test procedures, one can check the validity of the assumption that is required in Chapter 4.

In Section 4, we consider the equality of two covariance matrices by using the NR method. We also apply our test procedure to actual microarray data sets and compare another test procedures given by Srivastava and Yanagihara [32].

Finally, in Section 5, we give a simulation study to check performances of our test procedures.

1 Introduction

Suppose we have two classes $\pi_i,\ i = 1, 2$, and define independent $p \times n_i$ data matrices, $\boldsymbol{X}_i = [\boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{in_i}],\ i = 1, 2$, from $\pi_i,\ i = 1, 2$, where $\boldsymbol{x}_{ij},\ j = 1, ..., n_i$, are independent and identically distributed (i.i.d.) as a p-dimensional distribution with a mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i\ (\geq \boldsymbol{O})$. We assume $n_i \geq 4,\ i = 1, 2$. The eigen-decomposition of $\boldsymbol{\Sigma}_i$ is given by $\boldsymbol{\Sigma}_i = \boldsymbol{H}_i \boldsymbol{\Lambda}_i \boldsymbol{H}_i^T$, where $\boldsymbol{\Lambda}_i = \text{diag}(\lambda_{1(i)}, ..., \lambda_{p(i)})$ having $\lambda_{1(i)} \geq \cdots \geq \lambda_{p(i)} (\geq 0)$ and $\boldsymbol{H}_i = [\boldsymbol{h}_{1(i)}, ..., \boldsymbol{h}_{p(i)}]$ is an orthogonal matrix of the corresponding eigenvectors. Let $\boldsymbol{X}_i - [\boldsymbol{\mu}_i, ..., \boldsymbol{\mu}_i] = \boldsymbol{H}_i \boldsymbol{\Lambda}_i^{1/2} \boldsymbol{Z}_i$ for $i = 1, 2$. Then, $\boldsymbol{Z}_i$ is a $p \times n_i$ sphered data matrix from a distribution with the zero mean and identity covariance matrix. Let $\boldsymbol{Z}_i = [\boldsymbol{z}_{1(i)}, ..., \boldsymbol{z}_{p(i)}]^T$ and $\boldsymbol{z}_{j(i)} = (z_{j1(i)}, ..., z_{jn_i(i)})^T,\ j = 1, ..., p$, for $i = 1, 2$. Note that $E(z_{jk(i)} z_{j'k(i)}) = 0\ (j \neq j')$ and $\text{Var}(\boldsymbol{z}_{j(i)}) = \boldsymbol{I}_{n_i}$, where $\boldsymbol{I}_{n_i}$ denotes the n_i-dimensional identity matrix. Also, note that if $\boldsymbol{X}_i$ is Gaussian, $z_{jk(i)}$s are i.i.d. as the standard normal distribution, $N(0, 1)$. We assume that the fourth moments of each variable in $\boldsymbol{Z}_i$ are uniformly bounded for $i = 1, 2$. Let $\boldsymbol{z}_{oj(i)} = \boldsymbol{z}_{j(i)} - (\bar{z}_{j(i)}, ..., \bar{z}_{j(i)})^T,\ j = 1, ..., p;\ i = 1, 2$, where $\bar{z}_{j(i)} = n_i^{-1} \sum_{k=1}^{n_i} z_{jk(i)}$. We assume that $P(\lim_{p\to\infty} ||\boldsymbol{z}_{o1(i)}|| \neq 0) = 1$ for $i = 1, 2$, where $|| \cdot ||$ denotes the Euclidean norm. We define $\overline{\boldsymbol{x}}_{in_i} = \sum_{j=1}^{n_i} \boldsymbol{x}_{ij}/n_i$ and $\boldsymbol{S}_{in_i} = \sum_{j=1}^{n_i} (\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_{in_i})(\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_{in_i})^T/(n_i - 1)$ for $i = 1, 2$. Let $\delta_{j(i)} = \text{tr}(\boldsymbol{\Sigma}_i^2) - \sum_{s=1}^{j(i)} \lambda_{s(i)}^2 = \sum_{s=j+1}^{p} \lambda_{s(i)}^2$ for $i = 1, 2;\ j = 1, ..., p-1$. We consider the same assumptions in Chapter 1 and 2 for the first eigenvalue of each π_i:

(A-i) $\dfrac{\delta_{1(i)}}{\lambda_{1(i)}^2} = o(1)$ as $p \to \infty$ when n_i is fixed; $\dfrac{\delta_{j_*(i)}}{\lambda_1^2} = o(1)$ as $p \to \infty$ for some fixed $j_*\ (< p)$ when $n_i \to \infty$.

(A-ii) $\dfrac{\sum_{r,s\geq 2}^{p} \lambda_{r(i)} \lambda_{s(i)} E\{(z_{rk(i)}^2 - 1)(z_{sk(i)}^2 - 1)\}}{n_i \lambda_{1(i)}^2} = o(1)$ as $p \to \infty$ either when n_i is fixed or $n_i \to \infty$.

As necessary, we also consider the assumption (A-iii) for each π_i in Chapter 1 and 2:

(A-iii) $z_{1j(i)},\ j = 1, ..., n_i$, are i.i.d. as $N(0, 1)$.

2 Equality Tests Using the First Eigenvalues

We consider the following test for the first eigenvalues:

$$H_0 : \lambda_{1(1)} = \lambda_{1(2)} \quad \text{vs.} \quad H_a : \lambda_{1(1)} \neq \lambda_{1(2)} \ \ (\text{or } \ H_b : \lambda_{1(1)} < \lambda_{1(2)}). \tag{2.1}$$

2.1 Gaussian type HDLSS data

We consider the test (2.1) for the Gaussian type HDLSS data in the sense that it holds (A-ii). Let $\tilde{\lambda}_{1(i)}$ be the estimate of $\lambda_{1(i)}$ by the NR methodology as in (3.7) in Chapter 1 for π_i. From Theorem 3.1 in Chapter 1 we have the following result.

Theorem 2.1. *Under (A-i) to (A-iii), it holds that*

$$\frac{\tilde{\lambda}_{1(1)}/\lambda_{1(1)}}{\tilde{\lambda}_{1(2)}/\lambda_{1(2)}} \Rightarrow F_{n_1-1,n_2-1}$$

as $p \to \infty$ when n_is are fixed.

Proof. From Theorem 3.1 in Chapter 1 $(n_i-1)\tilde{\lambda}_{1(i)}/\lambda_{1(i)}$ is distributed as $\chi^2_{n_i-1}$ when $p \to \infty$ while n_i is fixed for $i = 1, 2$. Note that $\boldsymbol{z}_{o1(1)}$ and $\boldsymbol{z}_{o1(2)}$ are independent. Then, it concludes the result. □

Let $F_1^{NR} = \tilde{\lambda}_{1(1)}/\tilde{\lambda}_{1(2)}$. For a given $\alpha \in (0, 1/2)$ we test (2.1) by

$$\text{accepting } H_a \Longleftrightarrow F_1^{NR} \notin [\{F_{n_2-1,n_1-1}(\alpha/2)\}^{-1}, F_{n_1-1,n_2-1}(\alpha/2)] \tag{2.2}$$

$$\text{or} \quad \text{accepting } H_b \Longleftrightarrow F_1^{NR} < \{F_{n_2-1,n_1-1}(\alpha)\}^{-1}. \tag{2.3}$$

Then, under (A-i) to (A-iii), it holds that

$$\text{size} = \alpha + o(1)$$

as $p \to \infty$ when n_is are fixed.

Now, we consider a test by the conventional estimator, $\hat{\lambda}_{1(i)}$. Let $\kappa_i = \text{tr}(\boldsymbol{\Sigma}_i) - \lambda_{1(i)} = \sum_{s=2}^{p} \lambda_{s(i)}$ for $i = 1, 2$. From Proposition 3.1 in Chapter 1, if $\kappa_i/\lambda_{1(i)} = o(1)$ for $i = 1, 2$, under (A-iii) it holds that

$$\frac{\hat{\lambda}_{1(1)}/\lambda_{1(1)}}{\hat{\lambda}_{1(2)}/\lambda_{1(2)}} \Rightarrow F_{n_1-1,n_2-1}$$

as $p \to \infty$ when n_is are fixed. As mentioned in Section 2 of Chapter 1, the condition '$\kappa_i/\lambda_{1(i)} = o(1)$ for $i = 1, 2$' is quite strict in real high-dimensional data analyses. See Table 2 in Chapter 2 for example. *Hereafter, we assume* $\liminf_{p\to\infty} \kappa_i/\lambda_{1(i)} > 0$ *for* $i = 1, 2$.

2.2 Non-Gaussian type HDLSS data

Now, we consider testing (2.1) when (A-ii) is not always met. Let $\acute{\lambda}_{1(i)}$ be the estimator of $\lambda_{1(i)}$ by using the CDM method. From Corollary 3.2 in Chapter 1 we have the following result.

Theorem 2.2. *Under (A-i) and (A-iii), it holds as $p \to \infty$ that*

$$\frac{\acute{\lambda}_{1(1)}/\lambda_{1(1)}}{\acute{\lambda}_{1(2)}/\lambda_{1(2)}} \Rightarrow \{F_{n_{(1)1}-1,n_{(1)2}-1} \times F_{n_{(2)1}-1,n_{(2)2}-1}\}^{1/2} \tag{2.4}$$

when n_is are fixed, where F_{ν_1,ν_2} denotes a random variable distributed as F-distribution with (ν_1, ν_2) degrees of freedom and $F_{n_{(1)1}-1,n_{(1)2}-1}$ and $F_{n_{(2)1}-1,n_{(2)2}-1}$ are mutually independent.

Proof. Similar to Theorem 2.1, the result is obtained from Corollary 3.2 in Chapter 1. □

Let $F_1^{CDM} = \acute{\lambda}_{1(1)}/\acute{\lambda}_{1(2)}$. For a given $\alpha \in (0, 1/2)$, let $g(\alpha)$ be the upper α point of (2.4). Then, one can test (2.1) by

$$\text{accepting } H_a \text{ in (2.1)} \Longleftrightarrow F_1^{CDM} \notin \{g(1-\alpha/2), g(\alpha/2)\} \tag{2.5}$$

$$\text{or} \quad \text{accepting } H_b \Longleftrightarrow F_1^{CDM} < g(1-\alpha). \tag{2.6}$$

Then, under (A-i) and (A-iii), it holds that as $p \to \infty$

$$size(F_1^{CDM}) = \alpha + o(1)$$

when n_is are fixed.

3 Equality Tests Using the First Eigenspace

In this section, we consider the equality test of the first eigenspaces. We consider the following test:

$$H_0 : (\lambda_{1(1)}, \boldsymbol{h}_{1(1)}) = (\lambda_{1(2)}, \boldsymbol{h}_{1(2)}) \text{ vs. } H_a : (\lambda_{1(1)}, \boldsymbol{h}_{1(1)}) \neq (\lambda_{1(2)}, \boldsymbol{h}_{1(2)}). \tag{3.1}$$

3.1 Gaussian type HDLSS data

Let $\tilde{\boldsymbol{h}}_{1(i)}$ be the estimator of the first PC direction for π_i by the NR methodology given in Section 2 of Chapter 2. We assume $\boldsymbol{h}_{1(i)}^T \tilde{\boldsymbol{h}}_{1(i)} \geq 0$ w.p.1 for $i = 1, 2$, without loss of generality. Here, we have the following result.

Lemma 3.1. *Under (A-i) to (A-iii), it holds as $p \to \infty$ that*

$$\tilde{\boldsymbol{h}}_{1(1)}^T \tilde{\boldsymbol{h}}_{1(2)} = \boldsymbol{h}_{1(1)}^T \boldsymbol{h}_{1(2)} + o_p(1)$$

either when n_is are fixed or $n_i \to \infty$.

Proof. Let $\boldsymbol{Z}_i = [\boldsymbol{z}_{1(i)}, ..., \boldsymbol{z}_{p(i)}]^T$ be a sphered data matrix of π_i for $i = 1, 2$, where $\boldsymbol{z}_{j(i)} = (z_{j1(i)}, ..., z_{jn_i(i)})^T$ for $j = 1, ..., p$. We assume $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{0}$ without loss of generality. Let $\beta_{st} = (\lambda_{s(1)}\lambda_{t(2)})^{1/2} \boldsymbol{h}_{s(1)}^T \boldsymbol{h}_{t(2)}$ for all s, t. Let $j_\star$ be a fixed constant such that $\sum_{s=j_\star+1}^p \lambda_{s(i)}^2/\lambda_{1(i)}^2 = o(1)$ as $p \to \infty$ for $i = 1, 2$. Note that $j_\star$ exists under (A-i). We write that

$$\begin{aligned}\boldsymbol{X}_1^T \boldsymbol{X}_2 = &\sum_{s,t \leq j_\star} \beta_{st} \boldsymbol{z}_{s(1)} \boldsymbol{z}_{t(2)}^T + \sum_{s,t \geq j_\star+1}^{p} \beta_{st} \boldsymbol{z}_{s(1)} \boldsymbol{z}_{t(2)}^T \\ &+ \sum_{s=j_\star+1}^{p} \sum_{t=1}^{j_\star} \beta_{st} \boldsymbol{z}_{s(1)} \boldsymbol{z}_{t(2)}^T + \sum_{s=1}^{j_\star} \sum_{t=j_\star+1}^{p} \beta_{st} \boldsymbol{z}_{s(1)} \boldsymbol{z}_{t(2)}^T.\end{aligned}$$

Note that

$$E\Big\{\Big(\sum_{s=j_\star+1}^{p}\sum_{t=1}^{j_\star}\beta_{st}z_{sk(1)}z_{tk'(2)}\Big)^2\Big\}$$
$$=\mathrm{tr}\Big(\sum_{s=j_\star+1}^{p}\lambda_{s(1)}\boldsymbol{h}_{s(1)}\boldsymbol{h}_{s(1)}^T\sum_{t=1}^{j_\star}\lambda_{t(2)}\boldsymbol{h}_{t(2)}\boldsymbol{h}_{t(2)}^T\Big)\leq j_\star\lambda_{j_\star+1(1)}\lambda_{1(2)}$$

for all k, k'. Also, note that

$$E\Big\{\Big(\sum_{s,t\geq j_\star+1}^{p}\beta_{st}z_{sk(1)}z_{tk'(2)}\Big)^2\Big\}=\mathrm{tr}\Big(\sum_{s=j_\star+1}^{p}\lambda_{s(1)}\boldsymbol{h}_{s(1)}\boldsymbol{h}_{s(1)}^T\sum_{t=j_\star+1}^{p}\lambda_{t(2)}\boldsymbol{h}_{t(2)}\boldsymbol{h}_{t(2)}^T\Big)$$
$$\leq\Big(\sum_{s=j_\star+1}^{p}\lambda_{s(1)}^2\sum_{t=j_\star+1}^{p}\lambda_{t(2)}^2\Big)^{1/2}$$

for all k, k'. Then, by using Markov's inequality, for any $\tau>0$, under (A-i), we have that

$$P\Big\{\sum_{k=1}^{n_1}\sum_{k'=1}^{n_2}\Big(\sum_{s=j_\star+1}^{p}\sum_{t=1}^{j_\star}\frac{\beta_{st}z_{sk(1)}z_{tk'(2)}}{(n_1n_2\lambda_{1(1)}\lambda_{1(2)})^{1/2}}\Big)^2>\tau\Big\}\to 0,$$
$$P\Big\{\sum_{k=1}^{n_1}\sum_{k'=1}^{n_2}\Big(\sum_{s=1}^{j_\star}\sum_{t=j_\star+1}^{p}\frac{\beta_{st}z_{sk(1)}z_{tk'(2)}}{(n_1n_2\lambda_{1(1)}\lambda_{1(2)})^{1/2}}\Big)^2>\tau\Big\}\to 0$$
$$\text{and } P\Big\{\sum_{k=1}^{n_1}\sum_{k'=1}^{n_2}\Big(\sum_{s,t\geq j_\star+1}^{p}\frac{\beta_{st}z_{sk(1)}z_{tk'(2)}}{(n_1n_2\lambda_{1(1)}\lambda_{1(2)})^{1/2}}\Big)^2>\tau\Big\}\to 0$$

as $p\to\infty$ either when n_i is fixed or $n_i\to\infty$ for $i=1,2$. Let $\boldsymbol{P}_{n_i}=\boldsymbol{I}_{n_i}-\mathbf{1}_{n_i}\mathbf{1}_{n_i}^T/n_i$, where $\mathbf{1}_{n_i}=(1,...,1)^T$. Also, let $\boldsymbol{e}_{n_i}=(e_1,...,e_{n_i})^T$ be an arbitrary (random) n_i-vector such that $||\boldsymbol{e}_{n_i}||=1$ and $\boldsymbol{e}_{n_i}^T\mathbf{1}_{n_i}=0$. Let $\nu_i=n_i-1$ for $i=1,2$. Similar to (3.3) in the proof of Proposition 3.1, it holds that

$$\frac{\boldsymbol{e}_{n_1}^T\boldsymbol{X}_1^T\boldsymbol{X}_2\boldsymbol{e}_{n_2}}{(\nu_1\nu_2\lambda_{1(1)}\lambda_{1(2)})^{1/2}}=\frac{\boldsymbol{e}_{n_1}^T\sum_{s,t\leq j_\star}\beta_{st}\boldsymbol{z}_{s(1)}\boldsymbol{z}_{t(2)}^T\boldsymbol{e}_{n_2}}{(\nu_1\nu_2\lambda_{1(1)}\lambda_{1(2)})^{1/2}}+o_p(1).$$

Note that $\boldsymbol{e}_{n_i}^T\boldsymbol{P}_{n_i}=\boldsymbol{e}_{n_i}^T$ and $\boldsymbol{P}_{n_i}\boldsymbol{z}_{1(i)}=\boldsymbol{z}_{o1(i)}$ for $i=1,2$, where $\boldsymbol{z}_{o1(i)}=\boldsymbol{z}_{1(i)}-(\bar{z}_{1(i)},...,\bar{z}_{1(i)})^T$ and $\bar{z}_{1(i)}=n_i^{-1}\sum_{k=1}^{n_i}z_{1k(i)}$. Also, note that $\boldsymbol{X}_i\boldsymbol{P}_{n_i}=(\boldsymbol{X}_i-\overline{\boldsymbol{X}}_i)$ for $i=1,2$, where $\overline{\boldsymbol{X}}_i=[\bar{\boldsymbol{x}}_i,...,\bar{\boldsymbol{x}}_i]$ and $\bar{\boldsymbol{x}}_i=\sum_{k=1}^{n_i}\boldsymbol{x}_{k(i)}/n_i$. Let $\hat{\boldsymbol{u}}_{1(i)}$ be the first (unit) eigenvector of $(\boldsymbol{X}_i-\overline{\boldsymbol{X}}_i)^T(\boldsymbol{X}_i-\overline{\boldsymbol{X}}_i)$ for $i=1,2$. Note that $\hat{\boldsymbol{u}}_{1(i)}^T\boldsymbol{P}_{n_i}=\hat{\boldsymbol{u}}_{1(i)}^T$ when $(\boldsymbol{X}_i-\overline{\boldsymbol{X}}_i)^T(\boldsymbol{X}_i-\overline{\boldsymbol{X}}_i)\neq\boldsymbol{O}$ for $i=1,2$. Then, under (A-i), we have that

$$\frac{\hat{\boldsymbol{u}}_{1(1)}^T(\boldsymbol{X}_1-\overline{\boldsymbol{X}}_1)^T(\boldsymbol{X}_2-\overline{\boldsymbol{X}}_2)\hat{\boldsymbol{u}}_{1(2)}}{(\nu_1\nu_2\lambda_{1(1)}\lambda_{1(2)})^{1/2}}=\frac{\hat{\boldsymbol{u}}_{1(1)}^T\sum_{s,t\leq j_\star}\beta_{st}\boldsymbol{z}_{os(1)}\boldsymbol{z}_{ot(2)}^T\hat{\boldsymbol{u}}_{1(2)}}{(\nu_1\nu_2\lambda_{1(1)}\lambda_{1(2)})^{1/2}}+o_p(1)\tag{3.2}$$

as $p\to\infty$ either when n_i is fixed or $n_i\to\infty$ for $i=1,2$. Note that $\tilde{\boldsymbol{h}}_{1(i)}=\{\nu_i\tilde{\lambda}_{1(i)}\}^{-1/2}(\boldsymbol{X}_i-\overline{\boldsymbol{X}}_i)\hat{\boldsymbol{u}}_{1(i)}$ for $i=1,2$. Also, note that $\boldsymbol{z}_{os(i)}^T\boldsymbol{z}_{os'(i)}/n_i=o_p(1)$ $(s\neq s')$ when $n_i\to\infty$ for $i=1,2$. Then, by combining (3.2) with Theorem 3.1 in Chapter 1 and (4.1) in the proof of Lemma 4.1 of Chapter 2, we can

claim the result. □

We note that under H_0 in (3.1)

$$(\lambda_{1(i)}\boldsymbol{h}_{1(i)})^T(\lambda_{1(j)}^{-1}\boldsymbol{h}_{1(j)}) = 1 \quad \text{for } i = 1, 2;\ j \neq i. \tag{3.3}$$

Hence, one may consider a test statistic such as $F_1^{NR}|\tilde{\boldsymbol{h}}_{1(1)}^T\tilde{\boldsymbol{h}}_{1(2)}|$ or $F_1^{NR}|\tilde{\boldsymbol{h}}_{1(1)}^T\tilde{\boldsymbol{h}}_{1(2)}|^{-1}$. From Theorem 2.1 and Lemma 3.1 $F_1^{NR}|\tilde{\boldsymbol{h}}_{1(1)}^T\tilde{\boldsymbol{h}}_{1(2)}|$ and $F_1^{NR}|\tilde{\boldsymbol{h}}_{1(1)}^T\tilde{\boldsymbol{h}}_{1(2)}|^{-1}$ are asymptotically distributed as F_{n_1-1,n_2-1}. Let $\tilde{h} = \max\{|\tilde{\boldsymbol{h}}_{1(1)}^T\tilde{\boldsymbol{h}}_{1(2)}|, |\tilde{\boldsymbol{h}}_{1(1)}^T\tilde{\boldsymbol{h}}_{1(2)}|^{-1}\}$. Note that $\tilde{h} \geq 1$ w.p.1. Then, in view of the power, we give a test statistic for (3.1) as follows:

$$F_2^{NR} = \frac{\tilde{\lambda}_{1(1)}}{\tilde{\lambda}_{1(2)}}\tilde{h}_* \ \ (= F_1^{NR}\tilde{h}_*),$$

where

$$\tilde{h}_* = \begin{cases} \tilde{h} & \text{if } \tilde{\lambda}_{1(1)} \geq \tilde{\lambda}_{1(2)}, \\ \tilde{h}^{-1} & \text{otherwise.} \end{cases}$$

From Lemma 3.1 we have the following result.

Theorem 3.1. *Under (A-i) to (A-iii), it holds as $p \to \infty$ that*

$$F_2^{NR} \Rightarrow F_{n_1-1,n_2-1} \text{ under } H_0 \text{ in (3.1)}$$

when n_is are fixed.

Proof. By combining Theorem 2.1 and (3.3), we can claim the result. □

From Theorem 3.1 we consider testing (3.1) by (2.2) with F_2^{NR} instead of F_1^{NR}. Then, the size becomes close to α as p increases.

3.2 Non-Gaussian type HDLSS data

Now, we consider testing (3.1) when (A-ii) is not always met. We estimate the first PC score by using the CDM method as follows: Let $n_{(1)i} = \lceil n_i/2 \rceil$ and $n_{(2)i} = n_i - n_{(1)i}$ for $i = 1, 2$. For each class we divide the data matrix $\boldsymbol{X}_i$ into $\boldsymbol{X}_{(1)i} : p \times n_{(1)i}$ and $\boldsymbol{X}_{(2)i} : p \times n_{(2)i}$ at random. Similar to Section 3.3 in Chapter 1, we construct the cross data matrix by using $\boldsymbol{X}_{(1)i}$ and $\boldsymbol{X}_{(2)i}$, and calculate the first singular value $\acute{\lambda}_{1(i)}$ and the corresponding unit left- (or right-) singular vector $\acute{\boldsymbol{u}}_{(1)1i}$ (or $\acute{\boldsymbol{u}}_{(2)1i}$) for each class. Similarly, let $\boldsymbol{z}_{o(j)1i}$ be the centered first PC vector for the jth division of class i. We assume $P(\lim_{p\to\infty}||\boldsymbol{z}_{o(j)1i}|| \neq 0) = 1$ for $i = 1, 2;\ j = 1, 2$. According to Yata and Aoshima [36], we also calculate $\acute{\boldsymbol{h}}_{(j)1i} = \{(n_{(j)i} - 1)\acute{\lambda}_{1(i)}\}^{-1/2}(\boldsymbol{X}_{(j)i} - \overline{\boldsymbol{X}}_{(j)i})\acute{\boldsymbol{u}}_{(j)1i}$ for $i = 1, 2;\ j = 1, 2$. Then, we have the following result.

Lemma 3.2. *Under (A-i), it holds that as $p \to \infty$*

$$\acute{\boldsymbol{u}}_{(j)1i} \xrightarrow{P} \boldsymbol{z}_{o(j)1i}/||\boldsymbol{z}_{o(j)1i}|| \quad \textit{for } \ i = 1, 2;\ j = 1, 2$$

when n_is are fixed.

Proof. We note that $\acute{\boldsymbol{u}}_{(j)1i}^T \mathbf{1}_{n_{(j)i}} = 0,\ i = 1, 2$, with probability tending to 1 under $P(\lim_{p\to\infty} ||\boldsymbol{z}_{o(j)1i}|| \neq 0) = 1$ for $i = 1, 2$ and $j = 1, 2$. Also, note that $\boldsymbol{z}_{o(j)1i}^T \mathbf{1}_{n_{(j)i}} = 0$ for $i = 1, 2$ and $j = 1, 2$. Hence, similar to Theorem 3.2 in Chapter 1, we have the result. □

From Lemma 3.2 one can check the validity of (A-iii) by applying the test of the normality such as the Jarque-Bera test to $\acute{\boldsymbol{u}}_{(j)1i}$ for the non-Gaussian type HDLSS data.

We also have the following results for the first PC direction vector.

Lemma 3.3. *Under (A-i), it holds that as* $p \to \infty$

$$\acute{\boldsymbol{h}}_{(j)11}^T \acute{\boldsymbol{h}}_{(j)12} = \left\{ \frac{(n_{(j')1} - 1)(n_{(j')2} - 1)||\boldsymbol{z}_{o(j)11}||^2 ||\boldsymbol{z}_{o(j)12}||^2}{(n_{(j)1} - 1)(n_{(j)2} - 1)||\boldsymbol{z}_{o(j')11}||^2 ||\boldsymbol{z}_{o(j')12}||^2} \right\}^{1/4} \boldsymbol{h}_{1(1)}^T \boldsymbol{h}_{1(2)} + o_p(1)$$

for $j = 1, 2;\ j \neq j'$*, when* n_i*s are fixed.*

Proof. Let $\nu_{(j)i} = n_{(j)i} - 1$ for $i = 1, 2$ and $j = 1, 2$. Similar to Lemma 3.1, under (A-i), we have that

$$\begin{aligned} &\frac{\acute{\boldsymbol{u}}_{(j)11}^T (\boldsymbol{X}_{(j)1} - \overline{\boldsymbol{X}}_{(j)1})^T (\boldsymbol{X}_{(j)2} - \overline{\boldsymbol{X}}_{(j)2}) \acute{\boldsymbol{u}}_{(j)12}}{\{\nu_{(j)1}\nu_{(j)2}\lambda_{1(1)}\lambda_{1(2)}\}^{1/2}} \\ &= \frac{\acute{\boldsymbol{u}}_{(j)11}^T \beta_{11} \boldsymbol{z}_{o(j)11} \boldsymbol{z}_{o(j)12}^T \acute{\boldsymbol{u}}_{(j)12}}{\{\nu_{(j)1}\nu_{(j)2}\lambda_{1(1)}\lambda_{1(2)}\}^{1/2}} + o_p(1) \end{aligned} \tag{3.4}$$

as $p \to \infty$ when $n_{(j)i}$ is fixed for $i = 1, 2$ and $j = 1, 2$. Note that $\acute{\boldsymbol{h}}_{(j)1i} = \{\nu_{(j)i}\acute{\lambda}_{1(i)}\}^{-1/2}(\boldsymbol{X}_{(j)i} - \overline{\boldsymbol{X}}_{(j)i})\acute{\boldsymbol{u}}_{(j)1i}$ for $i = 1, 2$ and $j = 1, 2$. By combining (3.4) with Theorem 3.2 in Chapter 1 and Lemma 3.2 for each π_i, we can conclude the result. □

Lemma 3.4. *Under (A-i), it holds that as* $p \to \infty$

$$(\acute{\boldsymbol{h}}_{(j)11}^T \acute{\boldsymbol{h}}_{(k)12})(\acute{\boldsymbol{h}}_{(j')11}^T \acute{\boldsymbol{h}}_{(k')12}) = \{\boldsymbol{h}_{1(1)}^T \boldsymbol{h}_{1(2)}\}^2 + o_p(1)$$

for $j, k = 1, 2;\ (j, k) \neq (j', k')$*, when* n_i*s are fixed.*

Proof. From Lemma 3.3 it concludes the result straightforwardly. □

Let $\acute{h} = \{(\acute{\boldsymbol{h}}_{(1)11}^T \acute{\boldsymbol{h}}_{(2)11})(\acute{\boldsymbol{h}}_{(2)11}^T \acute{\boldsymbol{h}}_{(2)12})\}^{1/2}$ and $\acute{h}_{\max} = \max\{\acute{h}, \acute{h}^{-1}\}$. From Theorem 2.2 and Lemma 3.4, we consider the test statistic:

$$F_2^{CDM} = \frac{\acute{\lambda}_{1(1)}}{\acute{\lambda}_{1(2)}} \acute{h}_*,$$

where

$$\acute{h}_* = \begin{cases} \acute{h}_{\max} & \text{when } \acute{\lambda}_{1(1)} \geq \acute{\lambda}_{1(2)}, \\ 1/\acute{h}_{\max} & \text{otherwise.} \end{cases}$$

Then, we have the following result.

Theorem 3.2. *Assume (A-iii). Under (A-i), it holds that*

$$F_2^{CDM} \Rightarrow \{F_{n_{(1)1}-1,n_{(1)2}-1} \times F_{n_{(2)1}-1,n_{(2)2}-1}\}^{1/2} \quad \textit{under } H_0 \textit{ in (3.1)}$$

as $p \to \infty$ when n_is are fixed.

Proof. From (3.3) by combining Theorem 2.2 with Lemma 3.4, we can get the result. □

From Theorem 3.2 we consider testing (3.1) by (2.5) with F_2^{CDM} instead of F_1^{CDM}. Then, the size becomes close to α as p increases.

4 Equality Test of Two Covariance Matrices

In this section, we consider equality test of two covariance matrices. We consider the following test:

$$H_0 : \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 \quad \text{vs.} \quad H_a : \boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2. \tag{4.1}$$

When $p \to \infty$ and n_is are fixed, one can estimate $\lambda_{1(i)}$s and $\boldsymbol{h}_{1(i)}$s by the NR methodology and the CDM methodology, however, one cannot estimate $\lambda_{j(i)}$s and $\boldsymbol{h}_{j(i)}$s for $j = 2, ..., p$. Instead, we consider estimating $\kappa_i = \sum_{s=2}^{p} \lambda_{s(i)}$s by using the NR methodology. As for the CDM methodology, we cannot estimate κ_is because they go to zero automatically. Then, we consider the test (4.1) by using the NR methodology for Gaussian-type HDLSS data. Let $\boldsymbol{S}_{D_i}$ be the dual sample covariance matrix for π_i. We estimate κ_i by $\tilde{\kappa}_i = \mathrm{tr}(\boldsymbol{S}_{D_i}) - \tilde{\lambda}_{1(i)}$ for $i = 1, 2$. From Lemma 2.1 in Chapter 2, under (A-i) and (A-ii) for each π_i, $\tilde{\kappa}_i$s are consistent estimators of κ_is in the sense that $\tilde{\kappa}_i/\kappa_i = 1 + o_p(1)$ as $p \to \infty$ when n_is are fixed. Let $\tilde{\gamma} = \max\{\tilde{\kappa}_1/\tilde{\kappa}_2, \tilde{\kappa}_2/\tilde{\kappa}_1\}$. Similar to F_2^{NR}, we give a test statistic for (4.1) as follows:

$$F_3^{NR} = \frac{\tilde{\lambda}_{1(1)}}{\tilde{\lambda}_{1(2)}} \tilde{h}_* \tilde{\gamma}_* \ (= F_2^{NR} \tilde{\gamma}_*),$$

where

$$\tilde{\gamma}_* = \begin{cases} \tilde{\gamma} & \text{if } \tilde{\lambda}_{1(1)} \geq \tilde{\lambda}_{1(2)}, \\ \tilde{\gamma}^{-1} & \text{otherwise.} \end{cases}$$

Then, we have the following result.

Theorem 4.1. *Under (A-i) to (A-iii) for each π_i, it holds that*

$$F_3^{NR} \Rightarrow F_{n_1-1,n_2-1} \text{ under } H_0 \text{ in (4.1)}$$

as $p \to \infty$ when n_is are fixed.

Proof. By combining Theorem 3.1 in Chapter 1, Lemmas 2.1 in Chapter 2 and 3.1, we can claim the result. □

From Theorem 4.1 we consider testing (4.1) by (2.2) with F_3^{NR} instead of F_1^{NR}. Then, the size becomes close to α as p increases.

We analyzed lymphoma data given by Shipp et al. [29] and prostate cancer data given by Singh et al. [30] which are the same gene expression data as in Section 2 in Chapter 2. When each sample is standardized, we note that $\tilde{\kappa}_1 \approx \tilde{\kappa}_2$ if $\lambda_{1(i)}/\kappa_i = o(1)$, $i = 1, 2$, since $\text{tr}(\boldsymbol{S}_{D_1}) = \text{tr}(\boldsymbol{S}_{D_2}) = p$, so that one loses information about the difference between κ_1 and κ_2. Hence, we did not standardize each sample. We set $\alpha = 0.05$. We considered two cases: (I) π_1 : DLBC lymphoma ($n_1 = 58$) and π_2 : follicular lymphoma ($n_2 = 19$) and (II) π_1 : normal prostate ($n_1 = 50$) and π_2 : prostate tumor ($n_2 = 52$). We compared the performance of F_3^{NR} with two other test statistics, Q_2^2 and T_2^2, by Srivastava and Yanagihara [32]. The results are summarized in Table 1. We observed that F_3^{NR} accepted H_a for (I) and H_0 for (II), namely, F_3^{NR} rejected H_0 in (4.1) for (I). On the other hand, Q_2^2 and T_2^2 did not work for these data sets because Q_2^2 and T_2^2 are established under the severe conditions that $0 < \lim_{p\to\infty} \text{tr}(\boldsymbol{\Sigma}^i)/p < \infty$ $(i = 1, ..., 4)$ and $p^{1/2}/n = o(1)$. As observed in Table 1, the conditions seem not to hold for these data sets. Hence, there is no theoretical guarantee for the results by Q_2^2 and T_2^2.

Table 1. Tests of $H_0 : \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ vs. $H_a : \boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$ with size 0.05 for two data sets: (I) lymphoma data with $p = 7129$ given by Shipp et al. [29] and (II) prostate cancer data with $p = 12625$ given by Singh et al. [30].

	H_a by F_3^{NR}	H_a by Q_2^2	H_a by T_2^2
(I) π_1: DLBC, π_2: Follicular	Accept	Accept	Reject
(II) π_1: Normal, π_2: Tumor	Reject	Reject	Reject

5 Simulation Study

We used computer simulations to study the performance of the test procedures by (2.2) with F_1^{NR} for (2.1), F_2^{NR} for (3.1) and F_3^{NR} for (4.1). We set $\alpha = 0.05$. Independent pseudo-random normal observations were generated from $\pi_i : N_p(\mathbf{0}, \boldsymbol{\Sigma}_i)$, $i = 1, 2$. We set $(n_1, n_2) = (15, 25)$. Let $\nu_i = n_i - 1$ for $i = 1, 2$. We

considered the cases: $p = 2^k,\ k = 4, ..., 12$, and

$$\boldsymbol{\Sigma}_i = \begin{pmatrix} \boldsymbol{\Sigma}_{i(1)} & \boldsymbol{O}_{2,p-2} \\ \boldsymbol{O}_{p-2,2} & \boldsymbol{\Sigma}_{i(2)} \end{pmatrix},\ i = 1, 2, \tag{5.1}$$

where $\boldsymbol{O}_{k,l}$ is the $k \times l$ zero matrix, $\boldsymbol{\Sigma}_{1(1)} = \mathrm{diag}(p^{3/4}, p^{1/2})$ and $\boldsymbol{\Sigma}_{1(2)} = (0.3^{|s-t|})$. When considered the alternative hypotheses, we set

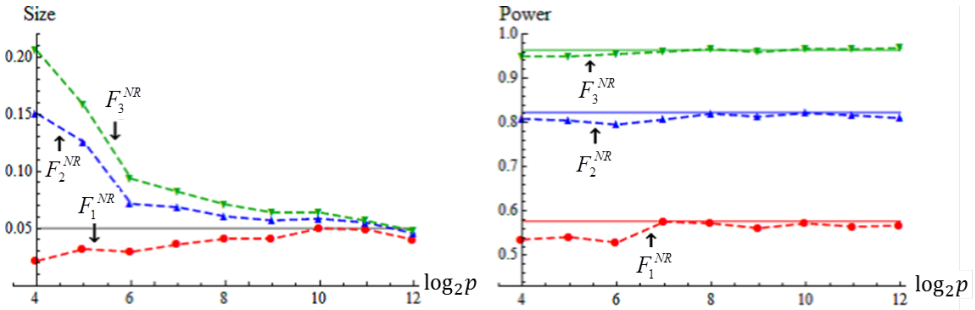
$$\boldsymbol{\Sigma}_{2(1)} = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix} \mathrm{diag}(3p^{3/4}, 1.5p^{1/2}) \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix} \tag{5.2}$$

and $\boldsymbol{\Sigma}_{2(2)} = 1.5(0.3^{|s-t|})$. Note that $\lambda_{1(2)}/\lambda_{1(1)} = 3$, $\kappa_2/\kappa_1 = 1.5$ and $\boldsymbol{h}_{1(1)}^T \boldsymbol{h}_{1(2)} = 1/\sqrt{2}$. Also, note that (A-i) to (A-iii) hold for each π_i. Let $h = \max\{|\boldsymbol{h}_{1(1)}^T \boldsymbol{h}_{1(2)}|, |\boldsymbol{h}_{1(1)}^T \boldsymbol{h}_{1(2)}|^{-1}\}$ and $\gamma = \max\{\kappa_1/\kappa_2, \kappa_2/\kappa_1\}$. From Lemmas 2.1 in Chapter 2 and 3.1, it holds that $\tilde{h} = h + o_p(1)$ and $\tilde{\gamma} = \gamma + o_p(1)$. Thus, from Theorems 2.1, 3.1 and 4.1, we obtained the asymptotic powers of F_1^{NR}, F_2^{NR} and F_3^{NR} with $(\tilde{h}_*, \tilde{\gamma}_*) = (h^{-1}, \gamma^{-1})$ as follows:

$$\begin{aligned} \text{Power}(F_1^{NR}) &= P\big\{(\lambda_{1(1)}/\lambda_{1(2)})f \notin [\{F_{n_2-1,n_1-1}(\alpha/2)\}^{-1}, F_{n_1-1,n_2-1}(\alpha/2)]\big\} = 0.577, \\ \text{Power}(F_2^{NR}) &= P\big\{h^{-1}(\lambda_{1(1)}/\lambda_{1(2)})f \notin [\{F_{n_2-1,n_1-1}(\alpha/2)\}^{-1}, F_{n_1-1,n_2-1}(\alpha/2)]\big\} = 0.823 \\ \text{and Power}(F_3^{NR}) &= P\big\{\gamma^{-1}h^{-1}(\lambda_{1(1)}/\lambda_{1(2)})f \notin [\{F_{n_2-1,n_1-1}(\alpha/2)\}^{-1}, F_{n_1-1,n_2-1}(\alpha/2)]\big\} = 0.963, \end{aligned}$$

where f denotes a random variable distributed as F distribution with degrees of freedom, $n_1 - 1$ and $n_2 - 1$. Note that Power(F_2^{NR}) and Power(F_3^{NR}) give lower bounds of the asymptotic powers when $\tilde{h}_* = h^{-1}$ and $\tilde{\gamma}_* = \gamma^{-1}$.

In Fig. 1, we summarized the findings obtained by averaging the outcomes from 4000 $(= R$, say) replications. Here, the first 2000 replications were generated by setting $\boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_1$ as in (5.1) and the last 2000 replications were generated by setting $\boldsymbol{\Sigma}_2$ as in (5.2). Let F_{ir}^{NR} $(i = 1, 2, 3)$ be the rth observation of F_i^{NR} for $r = 1, ..., 4000$. We defined $P_r = 1$ (or 0) when H_0 was falsely rejected (or not) for $r = 1, ..., 2000$, and H_a was falsely rejected (or not) for $r = 2001, ..., 4000$. We defined $\overline{\alpha} = (R/2)^{-1} \sum_{r=1}^{R/2} P_r$ to estimate the size and $1 - \overline{\beta} = 1 - (R/2)^{-1} \sum_{r=R/2+1}^{R} P_r$ to estimate the power. Their standard deviations are less than 0.011. When p is not sufficiently large, we observed that the sizes of F_2^{NR} and F_3^{NR} are quite higher than α. This is probably because $\tilde{h}_*$ (≥ 1) and $\tilde{\gamma}_*$ (≥ 1) are much larger than 1. Actually, the sizes became close to α as p increases. When p is large, F_3^{NR} gave excellent performances both for the size and power.



Sizes of F_1^{NR}, F_2^{NR} and F_3^{NR} Powers of F_1^{NR}, F_2^{NR} and F_3^{NR}

Figure 1. The values of $\overline{\alpha}$ are denoted by the dashed lines in the left panel and the values of $1-\overline{\beta}$ are denoted by the dashed lines in the right panel for F_1^{NR}, F_2^{NR} and F_3^{NR}. The asymptotic powers were given by Power$(F_1^{NR}) = 0.577$, Power$(F_2^{NR}) = 0.823$ and Power$(F_3^{NR}) = 0.963$ which were denoted by the solid lines in the right panel.

Chapter 4

Two-Sample Tests for HDLSS Data under the SSE Model

In this chapter, we consider two-sample tests for HDLSS data. This chapter is organized by Ishii [21] and Ishii [22].

We usually use the Hotelling's T^2 test statistic in Multivariate analysis. Since the sample covariance matrix is singular, one cannot use the test statistic in HDLSS context. For example, Dempster [15, 16], Srivastava [31] and Srivastava et al. [33] considered the two-sample test under the assumption that π_1 and π_2 are Gaussian. When π_1 and π_2 are non-Gaussian, Bai and Saranadasa [10] and Cai et al. [12] considered the two-sample test under homoscedasticity, $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$. Chen and Qin [13] and Aoshima and Yata [2, 7] considered the two-sample test under heteroscedasticity, $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$. Particularly, Aoshima and Yata [2] proposed a two-sample test procedure to ensure prespecified accuracies regarding both the size and power. We note that the above literatures considered constructing two-sample test procedures under the eigenvalue condition named the "non-strongly spiked eigenvalue (NSSE) model" by Aoshima and Yata [9]. Aoshima and Yata [9] also considered the other eigenvalue condition named the "strongly spiked eigenvalue (SSE) model". They proposed to develop high-dimensional inference not only for the NSSE model but also for the SSE model.

In this chapter, we focus on the SSE model and constructed test procedures when $p \to \infty$ while n_is are fixed.

In Section 2, we consider this problem for Gaussian type HDLSS data.

In Section 3, we constructed a test procedure for non-Gaussian type HDLSS data.

In Section 4, we show some simulation results.

In Section 5, we demonstrate the test procedure by using an actual microarray data set.

Finally, in Section 6, we give the concluding remark of this thesis.

1 Introduction

Suppose we have two classes $\pi_i,\ i = 1, 2$, and define independent $p \times n_i$ data matrices, $\boldsymbol{X}_i = [\boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{in_i}],\ i = 1, 2$, from $\pi_i,\ i = 1, 2$, where $\boldsymbol{x}_{ij},\ j = 1, ..., n_i$, are independent and identically distributed (i.i.d.) as a p-dimensional distribution with a mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i\ (\geq \boldsymbol{O})$. We assume $n_i \geq 4,\ i = 1, 2$. The eigen-decomposition of $\boldsymbol{\Sigma}_i$ is given by $\boldsymbol{\Sigma}_i = \boldsymbol{H}_i\boldsymbol{\Lambda}_i\boldsymbol{H}_i^T$, where $\boldsymbol{\Lambda}_i = \mathrm{diag}(\lambda_{1(i)}, ..., \lambda_{p(i)})$ having $\lambda_{1(i)} \geq \cdots \geq \lambda_{p(i)} (\geq 0)$ and $\boldsymbol{H}_i = [\boldsymbol{h}_{1(i)}, ..., \boldsymbol{h}_{p(i)}]$ is an orthogonal matrix of the corresponding eigenvectors. Let $\boldsymbol{X}_i - [\boldsymbol{\mu}_i, ..., \boldsymbol{\mu}_i] = \boldsymbol{H}_i\boldsymbol{\Lambda}_i^{1/2}\boldsymbol{Z}_i$ for $i = 1, 2$. Then, $\boldsymbol{Z}_i$ is a $p \times n_i$ sphered data matrix from a distribution with the zero mean and identity covariance matrix. Let $\boldsymbol{Z}_i = [\boldsymbol{z}_{1(i)}, ..., \boldsymbol{z}_{p(i)}]^T$ and $\boldsymbol{z}_{j(i)} = (z_{j1(i)}, ..., z_{jn_i(i)})^T,\ j = 1, ..., p$, for $i = 1, 2$. Note that $E(z_{jk(i)}z_{j'k(i)}) = 0\ (j \neq j')$ and $\mathrm{Var}(\boldsymbol{z}_{j(i)}) = \boldsymbol{I}_{n_i}$, where $\boldsymbol{I}_{n_i}$ denotes the n_i-dimensional identity matrix. Also, note that if $\boldsymbol{X}_i$ is Gaussian, $z_{jk(i)}$s are i.i.d. as the standard normal distribution, $N(0, 1)$. We assume that the fourth moments of each variable in $\boldsymbol{Z}_i$ are uniformly bounded for $i = 1, 2$. Let $\boldsymbol{z}_{oj(i)} = \boldsymbol{z}_{j(i)} - (\bar{z}_{j(i)}, ..., \bar{z}_{j(i)})^T,\ j = 1, ..., p;\ i = 1, 2$, where $\bar{z}_{j(i)} = n_i^{-1}\sum_{k=1}^{n_i} z_{jk(i)}$. We assume that $P(\lim_{p\to\infty} ||\boldsymbol{z}_{o1(i)}|| \neq 0) = 1$ for $i = 1, 2$, where $||\cdot||$ denotes the Euclidean norm. We define $\overline{\boldsymbol{x}}_{in_i} = \sum_{j=1}^{n_i} \boldsymbol{x}_{ij}/n_i$ and $\boldsymbol{S}_{in_i} = \sum_{j=1}^{n_i}(\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_{in_i})(\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_{in_i})^T/(n_i - 1)$ for $i = 1, 2$.

Aoshima and Yata [9] proposed the "non-strongly spiked eigenvalue (NSSE) model" defined by

$$\frac{\lambda_{1(i)}^2}{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)} \to 0 \ \text{ as } p \to \infty \text{ for } i = 1, 2. \tag{1.1}$$

However, (1.1) sometimes fails in actual high-dimensional analyses. See Aoshima and Yata [9] for the details. Aoshima and Yata [9] also proposed the "strongly spiked eigenvalue (SSE) model" defined by

$$\liminf_{p\to\infty} \left\{ \frac{\lambda_{1(i)}^2}{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)} \right\} > 0 \ \text{ for } i = 1 \text{ or } 2. \tag{1.2}$$

As for the SSE model, Ma et al. [27] considered a two-sample test for the factor model when $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$. Aoshima and Yata [9] considered the class of test statistics and constructed test procedures when $p \to \infty$ and $n_i s \to \infty$.

Ishii [21, 22] consider the following assumption:

(A-i) $$\frac{\mathrm{tr}(\boldsymbol{\Sigma}_i^2) - \lambda_{1(i)}^2}{\lambda_{1(i)}^2} = o(1), \quad p \to \infty.$$

The above eigenvalue model is regarded as a strongly spiked eigenvalue model which was proposed by Aoshima and Yata [9].

We consider the following test:

$$H_0 :\ \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \quad \text{vs.} \quad \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2 \tag{1.3}$$

We start with the following test statistic:

$$T_n = ||\overline{\boldsymbol{x}}_{1n_1} - \overline{\boldsymbol{x}}_{2n_2}||^2 - \sum_{i=1}^{2} \mathrm{tr}(\boldsymbol{S}_{in_i})/n_i.$$

Note that T_n was discussed by Chen and Qin [13] and Aoshima and Yata [2, 7] under the NSSE model. We consider T_n under the SSE model, (A-i). We assume the following assumption:

(A-ii) $\dfrac{\lambda_{1(1)}}{\lambda_{1(2)}} = 1 + o(1)$ and $\boldsymbol{h}_{1(1)}^T \boldsymbol{h}_{1(2)} = 1 + o(1)$ as $p \to \infty$.

(A-ii) means that the two classes share the first eigenspace when p is large. One can check (A-ii) by using the test statistics F_2^{NR} or F_2^{CDM} given in Chapter 3. As necessary, we also consider the same assumption (A-iii) for each π_i as in Chapter 1 and 2:

(A-iii) $z_{1j(i)},\ j = 1, ..., n_i$, are i.i.d. as $N(0, 1)$.

Let $n_{\min} = \min\{n_1, n_2\}$. Then, we have the following result for T_n.

Lemma 1.1. *Under H_0 in (1.3), (A-i) and (A-ii), it holds as $p \to \infty$ that*

$$\frac{T_n}{\lambda_{1(1)}} = (\bar{z}_{1(1)} - \bar{z}_{1(2)})^2 - \sum_{i=1}^{2} \frac{||\boldsymbol{z}_{o1(i)}||^2}{n_i(n_i - 1)} + o_p(n_{\min}^{-1})$$

either when $n_{\min}$ is fixed or $n_{\min} \to \infty$.

Proof. By using Chebyshev's inequality, for any $\tau > 0$, under (A-i), we have that for $i = 1, 2$

$$P\Big(\Big|\sum_{j \neq j'}^{n_i} \sum_{s=2}^{p} \frac{\lambda_{s(i)} z_{sj(i)} z_{sj'(i)}}{n_i(n_i - 1)}\Big| > \tau \lambda_{1(i)}/n_i\Big) = O\Big(\frac{\sum_{s=2}^{p} \lambda_{s(i)}^2}{\tau^2 \lambda_{1(i)}^2}\Big) \to 0 \tag{1.4}$$

as $p \to \infty$ either when n_i is fixed or $n_i \to \infty$. We write that

$$||\bar{\boldsymbol{x}}_{in_i} - \boldsymbol{\mu}_i||^2 - \frac{\mathrm{tr}(\boldsymbol{S}_{in_i})}{n_i} = \sum_{s=1}^{p} \lambda_{s(i)} \Big(\bar{z}_{s(i)}^2 - \frac{||\boldsymbol{z}_{os(i)}||^2}{n_i(n_i - 1)}\Big).$$

Here, $\bar{z}_{s(i)}^2 - ||\boldsymbol{z}_{os(i)}||^2/\{n_i(n_i - 1)\} = \sum_{j \neq j'}^{n_i} z_{sj(i)} z_{sj'(i)}/\{n_i(n_i - 1)\}$ for all i, s. Then, from (1.4) under (A-i), we have that

$$\frac{||\bar{\boldsymbol{x}}_{in_i} - \boldsymbol{\mu}_i||^2 - \mathrm{tr}(\boldsymbol{S}_{in_i})/n_i}{\lambda_{1(i)}} = \bar{z}_{1(i)}^2 - \frac{||\boldsymbol{z}_{o1(i)}||^2}{n_i(n_i - 1)} + o_p(n_i^{-1}) \tag{1.5}$$

as $p \to \infty$ either when n_i is fixed or $n_i \to \infty$. Let $\beta_{st} = (\lambda_{s(1)} \lambda_{t(2)})^{1/2} \times \boldsymbol{h}_{s(1)}^T \boldsymbol{h}_{t(2)}$ for all s, t. Then, we write that

$$\begin{aligned}
&(\overline{\boldsymbol{x}}_{1n_1} - \boldsymbol{\mu}_1)^T (\overline{\boldsymbol{x}}_{2n_2} - \boldsymbol{\mu}_2) \\
&= \sum_{s,t \geq 1}^{p} \beta_{st} \bar{z}_{s(1)} \bar{z}_{t(2)} = \beta_{11} \bar{z}_{1(1)} \bar{z}_{1(2)} + \sum_{s=2}^{p} \beta_{s1} \bar{z}_{s(1)} \bar{z}_{1(2)} \\
&\qquad + \sum_{t=2}^{p} \beta_{1t} \bar{z}_{1(1)} \bar{z}_{t(2)} + \sum_{s,t \geq 2}^{p} \beta_{st} \bar{z}_{s(1)} \bar{z}_{t(2)}.
\end{aligned} \tag{1.6}$$

Let $\mathbf{\Sigma}_{i*} = \sum_{s=2}^{p} \lambda_{s(i)} \boldsymbol{h}_{s(i)} \boldsymbol{h}_{s(i)}^T$ for $i = 1, 2$. Here, we have that

$$E\Big\{\Big(\sum_{s=2}^{p} \beta_{s1} \bar{z}_{s(1)} \bar{z}_{1(2)}\Big)^2\Big\} = \frac{\lambda_{1(2)} \boldsymbol{h}_{1(2)}^T \mathbf{\Sigma}_{1*} \boldsymbol{h}_{1(2)}}{n_1 n_2} \leq \frac{\lambda_{1(2)} \lambda_{2(1)}}{n_1 n_2};$$
$$E\Big\{\Big(\sum_{t=2}^{p} \beta_{1t} \bar{z}_{1(1)} \bar{z}_{t(2)}\Big)^2\Big\} = \frac{\lambda_{1(1)} \boldsymbol{h}_{1(1)}^T \mathbf{\Sigma}_{2*} \boldsymbol{h}_{1(1)}}{n_1 n_2} \leq \frac{\lambda_{1(1)} \lambda_{2(2)}}{n_1 n_2};$$
$$E\Big\{\Big(\sum_{s,t\geq 2}^{p} \beta_{st} \bar{z}_{s(1)} \bar{z}_{t(2)}\Big)^2\Big\} = \frac{\mathrm{tr}(\mathbf{\Sigma}_{1*} \mathbf{\Sigma}_{2*})}{n_1 n_2} \leq \frac{\sqrt{\mathrm{tr}(\mathbf{\Sigma}_{1*}^2)\mathrm{tr}(\mathbf{\Sigma}_{2*}^2)}}{n_1 n_2}.$$

Then, by using Chebyshev's inequality, for any $\tau > 0$, under (A-i) and (A-ii), it holds that

$$P\Big(\Big|\sum_{s=2}^{p} \beta_{s1} \bar{z}_{s(1)} \bar{z}_{1(2)}\Big| > \tau \lambda_{1(1)}/n_{\min}\Big) \leq \frac{\lambda_{1(2)} \lambda_{2(1)}}{\tau^2 \lambda_{1(1)}^2} \to 0;$$
$$P\Big(\Big|\sum_{t=2}^{p} \beta_{1t} \bar{z}_{1(1)} \bar{z}_{t(2)}\Big| > \tau \lambda_{1(1)}/n_{\min}\Big) \leq \frac{\lambda_{1(1)} \lambda_{2(2)}}{\tau^2 \lambda_{1(1)}^2} \to 0;$$
$$P\Big(\Big|\sum_{s,t\geq 2}^{p} \beta_{st} \bar{z}_{s(1)} \bar{z}_{t(2)}\Big| > \tau \lambda_{1(1)}/n_{\min}\Big) \leq \frac{\sqrt{\mathrm{tr}(\mathbf{\Sigma}_{1*}^2)\mathrm{tr}(\mathbf{\Sigma}_{2*}^2)}}{\tau^2 \lambda_{1(1)}^2} \to 0$$

as $p \to \infty$ either when n_i is fixed or $n_i \to \infty$ for $i = 1, 2$. Note that $\bar{z}_{1(1)} \bar{z}_{1(2)} = O_p(n_{\min}^{-1})$. Hence, from (1.6), under (A-i) and (A-ii), we have that

$$\begin{aligned} \frac{(\overline{\boldsymbol{x}}_{1n_1} - \boldsymbol{\mu}_1)^T (\overline{\boldsymbol{x}}_{2n_2} - \boldsymbol{\mu}_2)}{\lambda_{1(1)}} &= \frac{\beta_{11} \bar{z}_{1(1)} \bar{z}_{1(2)}}{\lambda_{1(1)}} + o_p(n_{\min}^{-1}) \\ &= \bar{z}_{1(1)} \bar{z}_{1(2)} + o_p(n_{\min}^{-1}) \end{aligned} \tag{1.7}$$

as $p \to \infty$ either when n_i is fixed or $n_i \to \infty$ for $i = 1, 2$. Here, we write that

$$\begin{aligned} ||\bar{\boldsymbol{x}}_{1n_1} - \bar{\boldsymbol{x}}_{2n_2}||^2 = &\sum_{i=1}^{2} ||\bar{\boldsymbol{x}}_{in_i} - \boldsymbol{\mu}_i||^2 - 2(\overline{\boldsymbol{x}}_{1n_1} - \boldsymbol{\mu}_1)^T (\overline{\boldsymbol{x}}_{2n_2} - \boldsymbol{\mu}_2) \\ &+ 2\boldsymbol{\mu}_{12}^T \{(\bar{\boldsymbol{x}}_{1n_1} - \boldsymbol{\mu}_1) - (\bar{\boldsymbol{x}}_{2n_2} - \boldsymbol{\mu}_2)\} + ||\boldsymbol{\mu}_{12}||^2. \end{aligned} \tag{1.8}$$

Then, by combining (1.5) with (1.7) and (1.8), we can get the result. □

Let $u_n = 1/n_1 + 1/n_2$. From Lemma 1.1, under H_0 in (1.3), (A-i) and (A-ii), we have that

$$\frac{1}{\lambda_{1(1)} u_n}\Big(T_n + \lambda_{1(1)} \sum_{i=1}^{2} \frac{||\boldsymbol{z}_{o1(i)}||^2}{n_i(n_i - 1)}\Big) = u_n^{-1}(\bar{z}_{1(1)} - \bar{z}_{1(2)}) + o_p(1) \tag{1.9}$$

as $p \to \infty$ either when $n_{\min}$ is fixed or $n_{\min} \to \infty$. Note that we assume that $E(z_{1k(i)}^4)$'s are uniformly bounded. Then, it holds that

$$u_n^{-1/2}(\bar{z}_{1(1)} - \bar{z}_{1(2)}) \Rightarrow N(0, 1)$$

as $n_{\min} \to \infty$ by Lyapunov's central limit theorem. Hence, from (1.9), it holds that as $p \to \infty$ and $n_{\min} \to \infty$

$$\frac{1}{\lambda_{1(1)} u_n} \Big(T_n + \lambda_{1(1)} \sum_{i=1}^{2} \frac{||\boldsymbol{z}_{o1(i)}||^2}{n_i(n_i-1)} \Big) \Rightarrow \chi_1^2 \tag{1.10}$$

under H_0 in (1.3), (A-i) and (A-ii), where χ_k^2 denotes a random variable distributed as χ^2 distribution with k degrees of freedom. On the other hand, under (A-iii), we note that $u_n^{-1/2}(\bar{z}_{1(1)} - \bar{z}_{1(2)})$ is distributed as $N(0,1)$ even when $n_{\min}$ is fixed. Hence, from (1.9), we have (1.10) as $p \to \infty$ when n_{min} is fixed under H_0 in (1.3) and (A-i) to (A-iii).

2 Gaussian Type HDLSS Data

In this section, we consider the test (1.3) for Gaussian- type HDLSS data. We also consider the following assumption:

(A-iv) $\dfrac{\sum_{r,s\geq 2}^{p} \lambda_{r(i)}\lambda_{s(i)} E\{(z_{rk(i)}^2 - 1)(z_{sk(i)}^2 - 1)\}}{n_i \lambda_{1(i)}^2} = o(1)$ as $p \to \infty$ either when n_i is fixed or $n_i \to \infty$.

Let $\nu = n_1 + n_2 - 2$. Let us write $\tilde{\lambda}_{1(i)}$ for $i = 1, 2$ as the NR estimator of $\lambda_{1(i)}$. Then, we have the following result.

Lemma 2.1. *Under (A-i) to (A-iv), it holds that as $p \to \infty$ when ν is fixed*

$$\frac{\sum_{i=1}^{2}(n_i - 1)\tilde{\lambda}_{1(i)}}{\lambda_{1(1)}} \Rightarrow \chi_\nu^2.$$

Under (A-i), (A-ii) and (A-iv), it holds that as $p \to \infty$ and $\nu \to \infty$

$$\frac{\sum_{i=1}^{2}(n_i - 1)\tilde{\lambda}_{1(i)}}{\nu \lambda_{1(1)}} = 1 + o_p(1).$$

In addition, from Theorem 3.1 in Chapter 1, we can estimate

$$\lambda_{1(1)} \sum_{i=1}^{2} \frac{||\boldsymbol{z}_{o1(i)}||^2}{n_i(n_i-1)}$$

in (1.9) by $\sum_{i=1}^{2} \tilde{\lambda}_{1(i)}/n_i$.

Proof. Under (A-iii), we note that $\boldsymbol{z}_{o1(1)}$ and $\boldsymbol{z}_{o1(2)}$ are independent, and $||\boldsymbol{z}_{o1(1)}||^2$ is distributed as $\chi_{n_i-1}^2$ for $i = 1, 2$. Hence, from Theorem 3.1 in Chapter 1 we can conclude the results. □

Then, Ishii [21] give a new test procedure by using the NR method. We consider the following test statistic.

$$F_{NR} = u_n^{-1} \frac{||\bar{\boldsymbol{x}}_{1n_1} - \bar{\boldsymbol{x}}_{2n_2}||^2 - \sum_{i=1}^{2}(\mathrm{tr}(\boldsymbol{S}_{in_i}) - \tilde{\lambda}_{1(i)})/n_i}{(n_1+n_2-2)^{-1}\sum_{i=1}^{2}(n_i-1)\tilde{\lambda}_{1(i)}} \tag{2.1}$$

Then, we have the following result.

Theorem 2.1. *Under (A-i) to (A-iv), it holds as $p \to \infty$ that*

$$F_{NR} \Rightarrow \begin{cases} F_{1,\nu} & \textit{when } \nu \textit{ is fixed}, \\ \chi_1^2 & \textit{when } \nu \to \infty. \end{cases}$$

Corollary 2.1. *Under (A-i), (A-ii) and (A-iv), it holds that as $p \to \infty$ and $n_{\min} \to \infty$*

$$F_{NR} \Rightarrow \chi_1^2 \textit{ under } H_0 \textit{ in (1.3).}$$

Proofs of Theorem 2.1. and Corollary 2.1. Under (A-iii), we note that $\bar{z}_{1(i)}$ and $\boldsymbol{z}_{o1(i)}$ are independent for $i = 1, 2$. By combining (1.10) with Theorem 3.1 in Chapter 1 and Lemma 2.1, we can conclude the results. □

Note that $\nu \to \infty$ either when $n_1 \to \infty$ or $n_2 \to \infty$. From Corollary 2.1 one can claim the result without (A-iii) if $n_{\min} \to \infty$ (i.e., $n_i \to \infty$ for $i = 1, 2$).

For a given $\alpha \in (0, 1/2)$ we test (1.3) by

$$\text{rejecting } H_0 \text{ in (1.3)} \Longleftrightarrow F_{NR} > F_{1,\nu}(\alpha), \tag{2.2}$$

where $F_{k_1,k_2}(\alpha)$ denotes the upper α point of F distribution with degrees of freedom, k_1 and k_2. Note that $F_{1,\nu}(\alpha) \to \chi_1^2(\alpha)$ as $\nu \to \infty$, where $\chi_k^2(\alpha)$ denotes the upper α point of χ^2 distribution with k degrees of freedom. Then, under the conditions in Theorem 2.1 (or Corollary 2.1), it holds that

$$\text{size} = \alpha + o(1)$$

as $p \to \infty$ either when ν is fixed or $\nu \to \infty$. Hence, one can use the test procedure even when n_is are fixed.

Next, we consider the power of the test by (2.2). Let $\boldsymbol{\mu}_{12} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$. Under H_1 in (1.3), we consider the following condition:

(A-v) $\dfrac{n_{\min}\boldsymbol{\mu}_{12}^T\boldsymbol{\Sigma}_i\boldsymbol{\mu}_{12}}{\lambda_{1(1)}^2} \to 0$ as $p \to \infty$ either when $n_{\min}$ is fixed or $n_{\min} \to \infty$.

Then, we have the following result.

Lemma 2.2. *Under (A-i), (A-ii) and (A-v), it holds that*

$$\frac{T_n}{\lambda_{1(1)}} = (\bar{z}_{1(1)} - \bar{z}_{1(2)})^2 - \sum_{i=1}^{2}\frac{||\boldsymbol{z}_{o1(i)}||^2}{n_i(n_i-1)} + \frac{||\boldsymbol{\mu}_{12}||^2}{\lambda_{1(1)}} + o_p(n_{\min}^{-1})$$

as $p \to \infty$ either when $n_{\min}$ is fixed or $n_{\min} \to \infty$.

Proof. We write that

$$T_n = \sum_{i=1}^{2} \left\{ ||\bar{\boldsymbol{x}}_{in_i} - \boldsymbol{\mu}_i||^2 - \frac{\mathrm{tr}(\boldsymbol{S}_{in_i})}{n_i} \right\} - 2(\bar{\boldsymbol{x}}_{1n_1} - \boldsymbol{\mu}_1)^T(\bar{\boldsymbol{x}}_{2n_2} - \boldsymbol{\mu}_2) + 2\boldsymbol{\mu}_{12}^T\{(\bar{\boldsymbol{x}}_{1n_1} - \boldsymbol{\mu}_1) - (\bar{\boldsymbol{x}}_{2n_2} - \boldsymbol{\mu}_2)\} + ||\boldsymbol{\mu}_{12}||^2. \tag{2.3}$$

Under (A-v), by using Chebyshev's inequality, for any $\tau > 0$ we have that

$$P\left(\boldsymbol{\mu}_{12}^T(\bar{\boldsymbol{x}}_{in_i} - \boldsymbol{\mu}_i) > \tau\lambda_{1(i)}/n_{\min}\right) = O\left(\frac{n_{\min}\boldsymbol{\mu}_{12}^T\boldsymbol{\Sigma}_i\boldsymbol{\mu}_{12}}{\tau^2\lambda_{1(i)}^2}\right) \to 0. \tag{2.4}$$

Then, by combining (2.4) with (2.3) and Lemma 1.1, it concludes the result. □

By using the above lemma, we have the following result.

Theorem 2.2. *Under (A-i) to (A-v), the test by (2.2) holds as $p \to \infty$ and $\nu \to \infty$ that*

$$Power = 1 - F_{\chi_1^2}\left(\chi_1^2(\alpha) - \frac{u_n^{-1}||\boldsymbol{\mu}_{12}||^2}{\lambda_{1(1)}}\right) + o(1)$$

, where $F_{\chi_1^2}(\cdot)$ denotes the cumulative distribution function of the chi-squared distribution with 1 degree of freedom.

Proof. Note that $F_{1,\nu}(\alpha) \to \chi_1^2(\alpha)$ as $\nu \to \infty$. From Lemmas 2.1 and 2.2, under (A-i) to (A-v), we have that as $p \to \infty$ and $\nu \to \infty$

$$\begin{aligned} &P\Big(u_n^{-1}\frac{T_n + \sum_{i=1}^2 \tilde{\lambda}_{1(i)}/n_i}{\nu^{-1}\sum_{i=1}^2 (n_i - 1)\tilde{\lambda}_{1(i)}} > F_{1,\nu}(\alpha)\Big) \\ &= P\Big(\chi_1^2 > \chi_1^2(\alpha) - \frac{||\boldsymbol{\mu}_{12}||^2}{u_n\lambda_{1(1)}} + o_p(1)\Big) \\ &= 1 - F_{\chi_1^2}\Big(\chi_1^2(\alpha) - \frac{||\boldsymbol{\mu}_{12}||^2}{u_n\lambda_{1(1)}}\Big) + o(1). \end{aligned}$$

It concludes the result. □

Remark 2.1. If $u_n^{-1}||\boldsymbol{\mu}_{12}||^2/\lambda_{1(1)} \to \infty$ as $p \to \infty$, the test by (2.2) holds under (A-i) to (A-v) that $Power = 1 + o_p(1)$ as $p \to \infty$ even when ν is fixed or $\nu \to \infty$.

3 Non-Gaussian Type HDLSS Data

We provide a new test procedure by using the CDM method when (A-iv) is not always met. Under (A-ii), we can estimate $\lambda_{1(1)}$ by using the CDM method as follows: We regard $\boldsymbol{X}_1 = [\boldsymbol{x}_{11}, ..., \boldsymbol{x}_{1n_1}]$ and

$\boldsymbol{X}_2 = [\boldsymbol{x}_{21}, ..., \boldsymbol{x}_{2n_2}]$ as $\boldsymbol{X}_{(1)}$ and $\boldsymbol{X}_{(2)}$ in the CDM method, respectively. We define the cross data matrix by $\boldsymbol{S}_{Dn_1} = \{(n_1-1)(n_2-1)\}^{-1/2}(\boldsymbol{X}_1 - \overline{\boldsymbol{X}}_1)^T(\boldsymbol{X}_2 - \overline{\boldsymbol{X}}_2)$. We define the singular value decomposition by $\boldsymbol{S}_{Dn_1} = \sum_{j=1}^{n_{\min}-1} \acute{\lambda}_{jn}\acute{\boldsymbol{u}}_{jn_1}\acute{\boldsymbol{u}}_{jn_2}^T$. Then, from Theorem 3.2 in Chapter 1, we have the following result.

Lemma 3.1. *Under (A-i) and (A-ii), it holds as $p \to \infty$ that*

$$\frac{\acute{\lambda}_{1n}}{\lambda_{1(1)}} = \begin{cases} ||\boldsymbol{z}_{o1(1)}/\sqrt{n_1-1}||||\boldsymbol{z}_{o1(2)}/\sqrt{n_2-1}|| + o_p(1) & \textit{when } n_i\textit{s are fixed} \\ 1 + o_p(1) & \textit{when } n_{\min} \to \infty. \end{cases}$$

Proof. Let $\nu_i = n_i - 1$. Also, let $\boldsymbol{P}_{n_i} = \boldsymbol{I}_{n_i} - n_i^{-1}\mathbf{1}_{n_i}\mathbf{1}_{n_i}^T$ for $i = 1, 2$. Similar to Lemma 3.1 in Chapter 3, it holds that

$$\frac{\boldsymbol{P}_{n_1}\boldsymbol{X}_1^T\boldsymbol{X}_2\boldsymbol{P}_{n_2}}{(\nu_1\nu_2\lambda_{1(1)}\lambda_{1(2)})^{1/2}} = \frac{\boldsymbol{P}_{n_1}\beta_{11}\boldsymbol{z}_{1(1)}\boldsymbol{z}_{1(2)}^T\boldsymbol{P}_{n_2}}{(\nu_1\nu_2\lambda_{1(1)}\lambda_{1(2)})^{1/2}} + o_p(1).$$

Note that $\boldsymbol{X}_i\boldsymbol{P}_{n_i} = (\boldsymbol{X}_i - \overline{\boldsymbol{X}}_i)$ and $\acute{\boldsymbol{u}}_{1n_i}^T\boldsymbol{P}_{n_i} = \acute{\boldsymbol{u}}_{1n_i}$ for $i = 1, 2$, when $(\boldsymbol{X}_1 - \overline{\boldsymbol{X}}_1)^T(\boldsymbol{X}_2 - \overline{\boldsymbol{X}}_2) \neq \boldsymbol{O}$. Then, under (A-i) and (A-ii), we have that

$$\begin{aligned} \frac{\acute{\boldsymbol{u}}_{1n_1}^T(\boldsymbol{X}_1 - \overline{\boldsymbol{X}}_1)^T(\boldsymbol{X}_2 - \overline{\boldsymbol{X}}_2)\acute{\boldsymbol{u}}_{1n_2}}{(\nu_1\nu_2)^{1/2}\lambda_{1(1)}} &= \frac{\acute{\boldsymbol{u}}_{1n_1}^T\beta_{11}\boldsymbol{z}_{o1(1)}\boldsymbol{z}_{o1(2)}^T\acute{\boldsymbol{u}}_{1n_2}}{(\nu_1\nu_2)^{1/2}\lambda_{1(1)}} + o_p(1) \\ &= \frac{||\boldsymbol{z}_{o1(1)}||||\boldsymbol{z}_{o1(2)}||}{(\nu_1\nu_2)^{1/2}} + o_p(1) \end{aligned}$$

as $p \to \infty$ either when n_i is fixed or $n_i \to \infty$ for $i = 1, 2$. Note that $||\boldsymbol{z}_{o1(i)}||/\nu_i^{1/2} = 1 + o_p(1)$ when $n_i \to \infty$ for $i = 1, 2$. Then, we can claim the result. □

From Lemmas 1.1 and 3.1 we consider the following test statistic:

$$F_{CDM} = u_n^{-1}T_n/\acute{\lambda}_{1n} + 1. \tag{3.1}$$

By using the above lemma, we have the following result.

Theorem 3.1. *Assume (A-iii). Under (A-i), (A-ii) and H_0 in (1.3), it holds that as $p \to \infty$*

$$F_{CDM} \Rightarrow \begin{cases} \dfrac{\nu_n\chi_1^2 - \{(n_1+n_2)\nu_n\}^{-1}\sum_{i\neq i'}^2 n_{i'}(n_{i'}-1)\chi_{n_i-1}^2}{\sqrt{\chi_{n_1-1}^2\chi_{n_2-1}^2}} + 1 & (3.2) \\ \qquad \textit{when } n_i\textit{s are fixed,} \\ \chi_1^2 \quad \textit{when } n_{min} \to \infty, & (3.3) \end{cases}$$

where $\nu_n = \sqrt{(n_1-1)(n_2-1)}$ and χ_1^2, $\chi_{n_1-1}^2$ and $\chi_{n_2-1}^2$ are mutually independent random variables distributed as the chi-squared distribution with degrees of freedom, 1, $n_1 - 1$ and $n_2 - 1$, respectively.

Proof. Under (A-iii), $u_n(\bar{z}_{1(1)} - \bar{z}_{1(2)})^2$ is distributed as χ_1^2. We note that $\bar{z}_{1(i)}$ and $\boldsymbol{z}_{o1(i)}$ $(i = 1, 2)$ are independent under (A-iii). Then, by combining Corollary 3.2 in Chapter 1 with Lemma 1.1 and Lemma 3.1, it concludes the result. □

Remark 3.1. When $p \to \infty$ and $n_{min} \to \infty$, the result in Theorem 3.1 holds without (A-iii).

For a given $\alpha \in (0, 1/2)$ let $f_{n_1,n_2}(\alpha)$ be the upper α point of (3.2). From Theorem 3.1 one can test (1.3) by

$$\text{rejecting } H_0 \text{ in (1.3)} \Longleftrightarrow F \geq f_{n_1,n_2}(\alpha). \tag{3.4}$$

Then, it holds under (A-i) to (A-iii) that

$$size = \alpha + o(1).$$

Next, we consider the power of the test by (3.4). From Lemma 2.2 we have the following result.

Theorem 3.2. *Under (A-i), (A-ii) and (A-v), the test by (3.4) holds that as $p \to \infty$ and $n_{min} \to \infty$*

$$Power = 1 - F_{\chi_1^2}\left(\chi_1^2(\alpha) - \frac{u_n^{-1}||\boldsymbol{\mu}_{12}||^2}{\lambda_{1(1)}}\right) + o(1),$$

where $F_{\chi_1^2}(\cdot)$ denotes the cumulative distribution function of the chi-squared distribution with 1 degree of freedom.

Proof. Note that $f_{n_1,n_2}(\alpha) \to \chi_1^2(\alpha)$ as $n_{\min} \to \infty$. From Remark 3.1 and Lemmas 3.1 and 2.2, under (A-i), (A-ii) and (A-v), we have that

$$\begin{aligned}
&P\left(\frac{u_n^{-1}T_n}{\acute{\lambda}_{1n}} + 1 > f_{n_1,n_2}(\alpha)\right) \\
&= P\left(\chi_1^2 > \chi_1^2(\alpha) - \frac{u_n^{-1}||\boldsymbol{\mu}_{12}||^2}{\lambda_{1(1)}} + o_p(1)\right) \\
&= 1 - F_{\chi_1^2}\left(\chi_1^2(\alpha) - \frac{u_n^{-1}||\boldsymbol{\mu}_{12}||^2}{\lambda_{1(1)}}\right) + o(1).
\end{aligned}$$

It concludes the result. □

Remark 3.2. If $u_n^{-1}||\boldsymbol{\mu}_{12}||^2/\lambda_{1(1)} \to \infty$ as $p \to \infty$, the test by (3.4) holds under (A-i), (A-ii), (A-iii) and (A-v) that Power$= 1 + o(1)$ as $p \to \infty$ either when n_is are fixed or $n_{min} \to \infty$.

Remark 3.3. In view of Theorem 3.2, one can consider the sample size determination so as to satisfy the probability requirement:

$$\text{Asymptotic power} \geq 1 - \beta \text{ whenever } ||\boldsymbol{\mu}_{12}||^2 \geq \Delta_0$$

for given $\beta \in (0, 1 - \alpha)$ and $\Delta_0 > 0$. If we consider minimizing the total sample size $n_1 + n_2$, one would obtain the following:

$$n_1 = n_2 = \frac{2\{\chi_1^2(\alpha) - \chi_1^2(1-\beta)\}\lambda_{1(1)}}{\Delta_0}.$$

One may estimate $\lambda_{1(1)}$ by using the bias-corrected CDM estimator in Chapter 1.

4 Simulation Studies

In this section, we summarize simulation studies of the findings by using computer simulations.

4.1 Gaussian type HDLSS data

We used computer simulations to study the performance of the test procedure by (2.2). We also checked the performance of the test procedure by

$$\text{rejecting } H_0 \Longleftrightarrow T_n/\widehat{K}^{1/2} > z_\alpha, \tag{4.1}$$

where z_α is a constant such that $P(N(0,1) > z_\alpha) = \alpha$ and

$$\widehat{K} = 2\sum_{i=1}^{2} \frac{W_{in_i}}{n_i(n_i-1)} + 4\frac{\text{tr}(\boldsymbol{S}_{1n_1}\boldsymbol{S}_{2n_2})}{n_1 n_2}$$

with $W_{in_i} = \{n_i(n_i-1)\}^{-1}\sum_{j\neq k}^{n_i}(\boldsymbol{x}_{ij}^T\boldsymbol{x}_{ik})^2 - 2\{n_i(n_i-1)(n_i-2)\}^{-1}\times\sum_{j\neq k\neq l}^{n_i}\boldsymbol{x}_{ij}^T\boldsymbol{x}_{ik}\boldsymbol{x}_{ik}^T\boldsymbol{x}_{il} + \{n_i(n_i-1)(n_i-2)(n_i-3)\}^{-1}\sum_{j\neq k\neq l\neq m}^{n_i}\boldsymbol{x}_{ij}^T\boldsymbol{x}_{ik}\boldsymbol{x}_{il}^T\boldsymbol{x}_{im}$. Here, W_{in_i} is an unbiased estimator of $\text{tr}(\boldsymbol{\Sigma}_i^2)$ given by Chen et al. [14]. See Srivastava et al. [34] for details of W_{in_i}. Note that Aoshima and Yata [2] and Yata and Aoshima [39] gave a different unbiased estimator of $\text{tr}(\boldsymbol{\Sigma}_i^2)$. From Theorems 1 and 2 in Chen and Qin [13] or Corollary 1 in Aoshima and Yata [9], under (1.1) and the factor model given in Remark 3.2 in Chapter 1, the test procedure by (4.1) has size $= \alpha + o(1)$ as $p \to \infty$ and $n_i \to \infty$, $i = 1, 2$. If (1.2) is met or n_is are fixed, we cannot claim "size$= \alpha + o(1)$" for the test procedure by (4.1).

We also considered the case when we use the conventional eigenvalue estimator, $\hat{\lambda}_{1(i)}$. Then, one can obtain the following test statistic:

$$\hat{F} = u_n^{-1}\frac{T_n + \sum_{i=1}^{2}\hat{\lambda}_{1(i)}/n_i}{\sum_{i=1}^{2}(n_i-1)\hat{\lambda}_{1(i)}}$$

and checked the performance of the test procedure by

$$\text{rejecting } H_0 \Longleftrightarrow \hat{F} > F_{1,n_1+n_2-2}(\alpha). \tag{4.2}$$

We set $\alpha = 0.05$, $\boldsymbol{\mu}_1 = \boldsymbol{0}$ and

$$\boldsymbol{\Sigma}_i = \begin{pmatrix} \boldsymbol{\Sigma}_{(1)} & \boldsymbol{O}_{2,p-2} \\ \boldsymbol{O}_{p-2,2} & c_i\boldsymbol{\Sigma}_{(2)} \end{pmatrix},\ i = 1, 2, \tag{4.3}$$

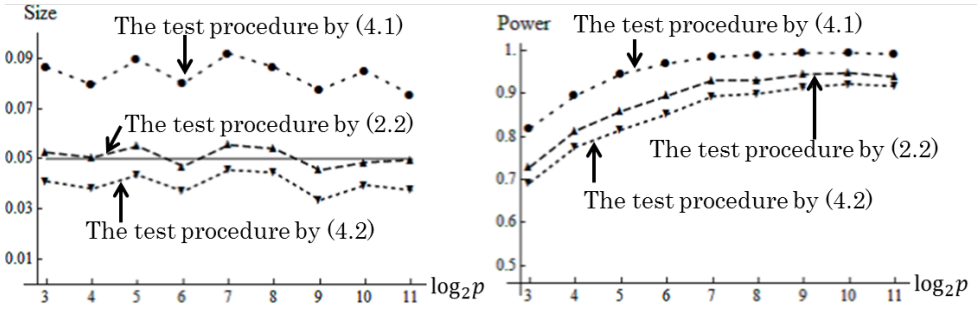
where $\boldsymbol{O}_{k,l}$ is the $k \times l$ zero matrix, $\boldsymbol{\Sigma}_{(1)} = \text{diag}(p^\tau, p^{1/2})$, $\boldsymbol{\Sigma}_{(2)} = (0.3^{|i-j|^{1/2}})$ and $(c_1, c_2) = (1, 1.5)$. Note that (A-i) is met when $\tau > 1/2$. Also, note that (A-ii) is met.

First, we considered the case when $p \to \infty$ while n_is are fixed. Independent pseudo-random observations were generated from $\pi_i : N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, $i = 1, 2$. We consider the following two cases:

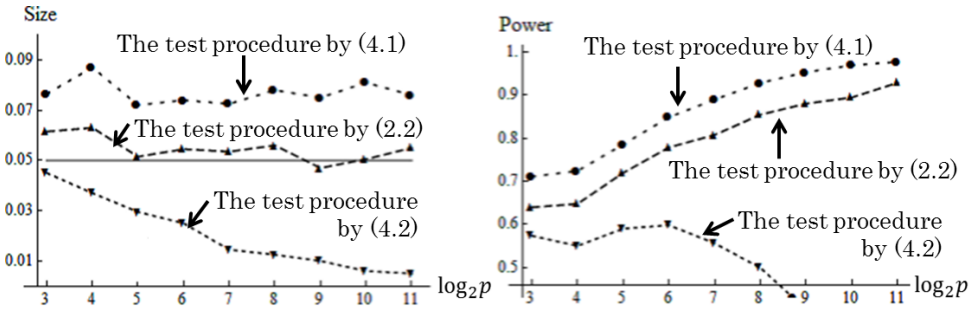
(a) $p = 2^s$ for $s = 3, ..., 11$, $(n_1, n_2) = (10, 15)$ and $\tau = 1$, and

(b) $p = 2^s$ for $s = 3, ..., 11$, $(n_1, n_2) = (10, 15)$ and $\tau = 2/3$.

We considered $\boldsymbol{\mu}_2 = \mathbf{0}$ for H_0 and $\boldsymbol{\mu}_2 = (0, ..., 0, 1, ..., 1)^T$ for H_1 whose last $\lceil p^\tau \rceil$ elements are 1, where $\lceil x \rceil$ denotes the smallest integer $\geq x$. For each case, we checked the performance as follows: We defined $P_r = 1$ (or 0) when H_0 was falsely rejected (or not) for $r = 1, ..., 2000$, and defined $\overline{\alpha} = \sum_{r=1}^{2000} P_r/2000$ to estimate the size. We also defined $P_r = 1$ (or 0) when H_1 was falsely rejected (or not) for $r = 1, ..., 2000$, and defined $1 - \overline{\beta} = 1 - \sum_{r=1}^{2000} P_r/2000$ to estimate the power. Note that their standard deviations are less than 0.011. In Fig. 1, we plotted $\overline{\alpha}$ (left panel) and $1 - \overline{\beta}$ (right panel) for the test procedure by (2.2) in each of (a) and (b). We also plotted them for the test procedure by (4.1) and (4.2) in each case.



(a) $p = 2^s$ for $s = 3, ..., 11$, $(n_1, n_2) = (10, 15)$ and $\tau = 1$.



(b) $p = 2^s$ for $s = 3, ..., 11$, $(n_1, n_2) = (10, 15)$ and $\tau = 2/3$.

Figure 1. The performances of the three test procedures by (2.2), (4.1) and (4.2). Independent pseudo-random observations were generated from $\pi_i : N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, $i = 1, 2$. The values of $\overline{\alpha}$ are denoted by the dashed lines in the left panels and the values of $1 - \overline{\beta}$ are denoted by the dashed lines in the right panels.

We observed that the test procedure by (2.2) gave better performances compared to (4.1) regarding the size. The size by (4.1) did not become close to α. This is probably because T_n does not hold the asymptotic normality under the SSE model, (1.2). One may think that (4.1) gave better performances compared to (2.2) regarding the power. This is because (4.1) cannot control the size under the SSE model. On the other hand, the test procedure by (4.2) gave quite bad performances for (b). The power was much lower than that of (2.2). The main reason must be that $\hat{\lambda}_{1(i)}$ was strongly inconsistent for (b).

Next, we considered the case when $n_i \to \infty$ for $i = 1, 2$. Independent pseudo-random observations were generated from $N_{p-p^*}(\mathbf{0}, \boldsymbol{I}_{p-p^*})$ for $(z_{1j(i)}, ..., z_{p-p^*j(i)})^T$, $j = 1, ..., n_i$, and from p^*-variate t-distribution, $t_{p^*}(\mathbf{0}, \boldsymbol{I}_{p^*}, \nu)$ for $(z_{p-p^*+1j(i)}, ..., z_{pj(i)})^T$, $j = 1, \cdots, n_i$, $i = 1, 2$. Let $p_* = \lceil p^{1/2} \rceil$. Note that (A-iv) holds from the fact that $\sum_{r,s\geq 2}^{p} \lambda_{r(i)}\lambda_{s(i)} E\{(z^2_{rk(i)} - 1)(z^2_{sk(i)} - 1)\} = 2\sum_{s=2}^{p-p_*} \lambda^2_{s(i)} + O(\sum_{r,s\geq p-p_*+1}^{p} \lambda_{r(i)}\lambda_{s(i)}) = o(\lambda^2_{1(i)})$ for $i = 1, 2$. We consider the following two cases:

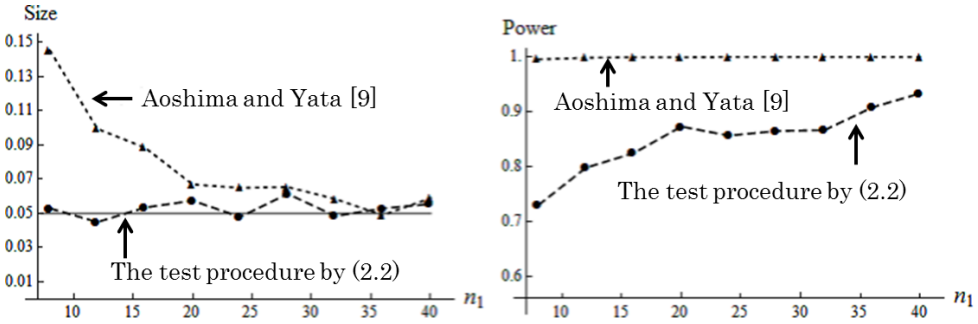
(c) $p = 200$, $n_1 = 4s$ for $s =$, $n_2 = 1.5n_1$ and $\tau = 3/4$, and

(d) $p = 1000$, $n_1 = 4s$ for $s = 2, ..., 10$, $n_2 = 1.5n_1$ and $\tau = 3/4$.

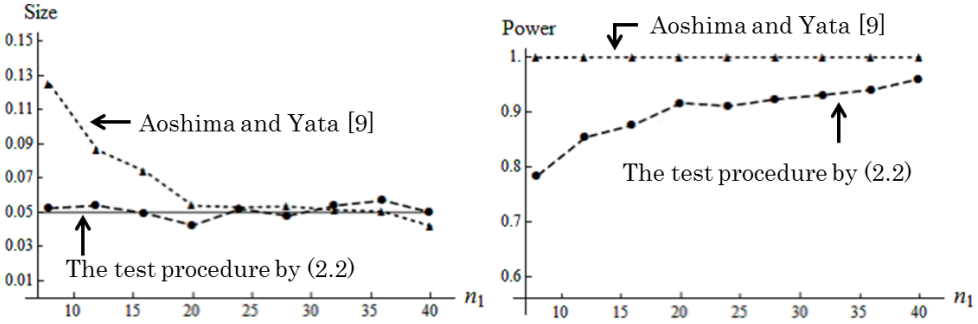
We set $\boldsymbol{\mu}_2 = (0, ..., 0, 1, ..., 1)^T$ for H_1 whose last $\lceil 5u_n\lambda_{1(1)} \rceil$ elements are 1 for each case. Note that $||\boldsymbol{\mu}_{12}||^2 = \lceil 5u_n\lambda_{1(1)} \rceil$ for H_1. Then, it holds that

$$F_{\chi^2_1}\{\chi^2_1(\alpha) - ||\boldsymbol{\mu}_{12}||^2/(u_n\lambda_{1(1)})\} = 0$$

for H_1. Thus from Theorem 2.2 the test by (2.2) has $Power = 1 + o(1)$ as $p \to \infty$ and $n_i \to \infty$, $i = 1, 2$. Similarly, we calculated $\overline{\alpha}$ and $1 - \overline{\beta}$. In Fig. 2, we plotted these values for the test procedures by (2.2) and (5.5) in Aoshima and Yata [9].



(c) $p = 200$, $n_1 = 4s$ for $s =$, $n_2 = 1.5n_1$ and $\tau = 3/4$.



(d) $p = 1000$, $n_1 = 4s$ for $s = 2, ..., 10$, $n_2 = 1.5n_1$ and $\tau = 3/4$.

Figure 2. The performances of the two test procedures by (2.2) and (5.5) in Aoshima and Yata [9]. Independent pseudo-random observations were generated from $N_{p-p^*}(\mathbf{0}, \boldsymbol{I}_{p-p^*})$ for $(z_{1j(i)}, ..., z_{p-p^*j(i)})^T$, $j = 1, ..., n_i$, and from p^*-variate t-distribution, $t_{p^*}(\mathbf{0}, \boldsymbol{I}_{p^*}, \nu)$ for $(z_{p-p^*+1j(i)}, ..., z_{pj(i)})^T$, $j = 1, \cdots, n_i$, $i = 1, 2$, where $p^* = \lceil p^{1/2} \rceil$. The values of $\overline{\alpha}$ are denoted by the dashed lines in the left panels and the values of $1 - \overline{\beta}$ are denoted by the dashed

lines in the right panels.

We observed that the test procedure by (2.2) gave better performances compared to the test procedure by Aoshima and Yata [9] regarding the size when n_is are very small. The test procedure by Aoshima and Yata [9] became close to α as n_is increase. In addition, the test procedure by Aoshima and Yata [9] gave better performances compared to (2.2) regarding the power. This is probably because the asymptotic variance of the test statistic by Aoshima and Yata [9] is smaller than $\mathrm{Var}(T_n)$ in this high-dimensional settings. See Section 5.1 in Aoshima and Yata [9] for the details. Hence, we recommend to use the test procedure by (2.2) for Gaussian type HDLSS data when n_is are very small (e.g. n_is are about 10) under the SSE model.

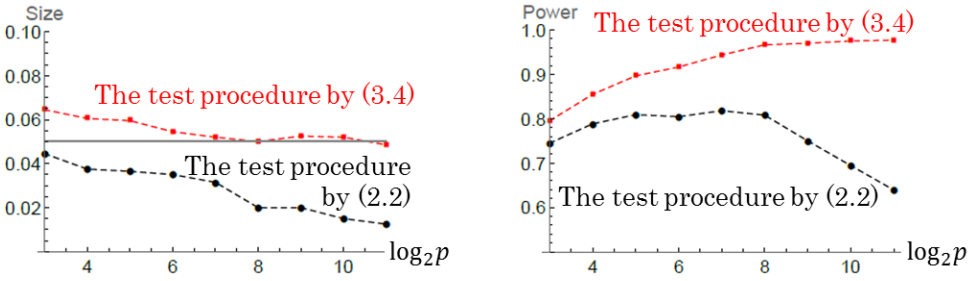
4.2 Non-Gaussian type HDLSS data

We used computer simulations to study the performance of the test procedure by (3.4). We set $\alpha = 0.05$, $\boldsymbol{\mu}_1 = \mathbf{0}$. We considered the same setting as (4.3) for $\boldsymbol{\Sigma}_i$ and set $\boldsymbol{\Sigma}_{(1)} = \mathrm{diag}(p^{3/4}, p^{1/2})$. Note that (A-i) is met.

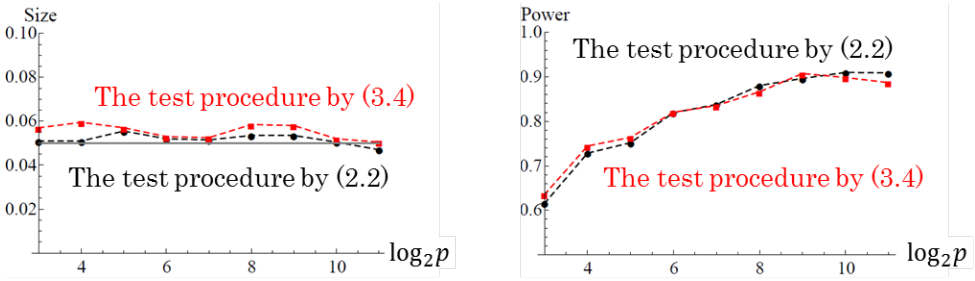
Independent pseudo-random observations were generated from $N_{p-p^*}(\mathbf{0}, \boldsymbol{I}_{p-p^*})$ for $(z_{1j(i)}, ..., z_{p-p^*j(i)})^T$, $j = 1, ..., n_i$, and from p^*-variate t-distribution, $t_{p^*}(\mathbf{0}, \boldsymbol{I}_{p^*}, \nu)$ for $(z_{p-p^*+1j(i)}, ..., z_{pj(i)})^T$, $j = 1, \cdots, n_i$, $i = 1, 2$. We considered three cases:

(a) $p = 2^s$ for $s = 3, ..., 11$, $p^* = p - 1$, $\nu = 5$ and $(n_1, n_2) = (10, 10)$,
(b) $p = 2^s$ for $s = 3, ..., 11$, $p^* = \lceil p^{1/2} \rceil$, $\nu = 10$ and $(n_1, n_2) = (12, 7)$, and
(c) $p = 1000$, $p^* = \lceil p^{1/2} \rceil$, $\nu = 10$ and $n_1 = n_2 = 3 + 6s$ for $s = 1, ..., 9$.

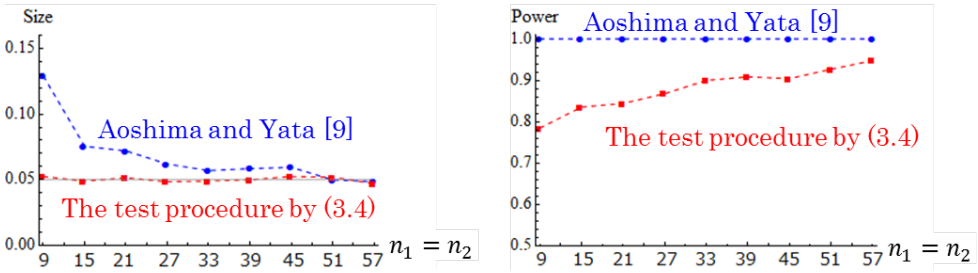
Note that (A-iv) is not satisfied for (a). We considered $\boldsymbol{\mu}_2 = \mathbf{0}$ for H_0 and $\boldsymbol{\mu}_2 = (0, ..., 0, 1, ..., 1)^T$ for H_1 whose last η elements are 1. We set $\eta = \lceil 1.5\lambda_{1(1)} \rceil$ for (a), $\eta = \lceil 1.4\lambda_{1(1)} \rceil$ for (b) and $\eta = \lceil 6\lambda_{1(1)} \rceil$ for (c), where $\lceil x \rceil$ denotes the smallest integer $\geq x$. For each case we checked the performance as follows: We defined $P_r = 1$ (or 0) when H_0 was falsely rejected (or not) for $r = 1, ..., 2000$, and defined $\overline{\alpha} = \sum_{r=1}^{2000} P_r/2000$ to estimate the size. We also defined $P_r = 1$ (or 0) when H_1 was falsely rejected (or not) for $r = 1, ..., 2000$, and defined $1 - \overline{\beta} = 1 - \sum_{r=1}^{2000} P_r/2000$ to estimate the power. Note that their standard deviations are less than 0.011. In Fig. 3, we plotted $\overline{\alpha}$ (left panel) and $1 - \overline{\beta}$ (right panel) for the test procedure by (3.4) in each of (a), (b) and (c). We also plotted them for the test procedure by (2.2) in (a) and (b), and for the test procedure by (5.5) in Aoshima and Yata [9] in (c). One can observe from (a), (b) and (c) that the test procedure by (3.4) gave good performances for large p even when n_is are fixed. Contrary to that, the test procedure by (2.2) gave a bad performance for (a) with respect to the power when p is large. This is probably because (A-iv) is not met in (a) when ν is small. On the other hand, it gave a good performance for (b) when p is large. The test procedure by Aoshima and Yata [9] gave a good performance when both p and n_is are large. We recommend to use the test procedure by (3.4) when the data is non-Gaussian and the sample size is quite small.



(a) $p = 2^s$ for $s = 3, ..., 11$, $p^* = p - 1$, $\nu = 5$ and $(n_1, n_2) = (10, 10)$.



(b) $p = 2^s$ for $s = 3, ..., 11$, $p^* = \lceil p^{1/2} \rceil$, $\nu = 10$ and $(n_1, n_2) = (12, 7)$.



(c) $p = 1000$, $p^* = \lceil p^{1/2} \rceil$, $\nu = 10$ and $n_1 = n_2 = 3 + 6s$ for $s = 1, ..., 9$.

Figure 3. The performances of the three test procedures by (3.4), (2.2) and (5.5) in Aoshima and Yata [9]. Independent pseudo-random observations were generated from $N_{p-p^*}(\mathbf{0}, \boldsymbol{I}_{p-p^*})$ for $(z_{1j(i)}, ..., z_{p-p^*j(i)})^T$, $j = 1, ..., n_i$ and from p^*-variate t-distribution, $t_{p^*}(\mathbf{0}, \boldsymbol{I}_{p^*}, \nu)$ for $(z_{p-p^*+1j(i)}, ..., z_{pj(i)})^T$, $j = 1, \cdots, n_i$, $i = 1, 2$. The values of $\overline{\alpha}$ are denoted by the solid lines in the left panels and the values of $1 - \overline{\beta}$ are denoted by the solid lines in the right panels.

5 Demonstration

In this section, we demonstrate the test procedure (3.4) by using actual microarray data set. We used acute myeloid leukemia data with 22283 $(= p)$ genes consisting of four classes: acute promyelocytic leukemia (APL) with t(15;17) (10 samples), acute myelogenous leukemia (AML) with inv(16) (4 samples), monocytic leukemia (ML) (7 samples) and nonmonocytic leukemia (NL) (22 samples). See Gutierrez et al. [17] for the details. The data set is available at NCBI Gene Expression Omunibus. First, we checked (A-iii) for

each class. As for each class, we divided the sample into two groups: the first $\lceil n_i/2 \rceil$ samples and the remaining samples. Then, we constructed the cross data matrix $\boldsymbol{S}_{D(i)}$ for each class. We calculated $\acute{\lambda}_{1(i)}$ and estimated $\delta = \sum_{s=2}^{p} \lambda_{s(i)}^2 / \lambda_{1(i)}^2$ by $\hat{\delta} = \{\mathrm{tr}(\boldsymbol{S}_{D(i)} \boldsymbol{S}_{D(i)}^T)\} / \acute{\lambda}_{1(i)}^2$. We had $\hat{\delta} = 0.013$ for APL, $\hat{\delta} = 0$ for AML, $\hat{\delta} = 0.171$ for ML and $\hat{\delta} = 0.034$ for NL. From these observations we concluded that each class satisfies (A-i). In addition, from Lemma 3.2, we could confirm that each class satisfies (A-iii) with the level of significance 0.05. We also checked (A-ii) for six pairs out of the four classes and tested (3.1) in Chapter 3 by using the test statistic F_2^{CDM} with the level of significance 0.05. We had P-values as 0.481 for (APL, AML), 0.187 for (APL, ML), 0.902 for (APL, NL), 0.52 for (AML, ML), 0.746 for (AML, NL) and 0.920 for (ML, NL). From these observations, we applied the two-sample test procedure (3.4) to all the cases. We tested (1.3) with the level of significance 0.05. Then, H_0 in (1.3) was rejected for (APL, ML), (APL, NL), (AML, NL) and (ML, NL). The results are summarized in Table 1.

Table 1. The upper 0.05 point, $f_{n_1,n_2}(0.05)$, of (3.4) and the value of F_{CDM} given by (3.1) for all the pairs from Gutierrez et al. [17]'s data sets having $p = 22283$.

	(APL, AML)	(APL, ML)	(APL, NL)	(AML, ML)	(AML, NL)	(ML, NL)
(n_1, n_2)	$(10, 4)$	$(10, 7)$	$(10, 22)$	$(4, 7)$	$(4, 22)$	$(7, 22)$
$f_{n_1,n_2}(0.05)$	5.94	4.91	4.38	6.05	5.76	4.71
F_{CDM}	2.39	13.22	12.26	4.31	6.04	19.37

6 Conclusion

As pointed out in Aoshima and Yata [9], we should choose a suitable test procedure reflected by the eigenstructures. In this thesis, we focused on the SSE model. In high-dimensional settings, it is unrealistic to assume the equality of the covariance matrices between the two classes. However, when analysing microarray data sets, we sometimes observe that the two covariance matrices share the first eigenspace at least. In such situations, we positively make use of the common eigenspace as a ground to compare the two class means. From this point of view, we provided two-sample test procedures by using both the NR method and the CDM method. Also, we discussed how to check the validity of the assumptions. Through the simulation studies, the proposed test procedures by (2.2) and (3.4) gave good performances when the dimension is large while the sample-size is quite small.

References

[1] 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.

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

[3] Aoshima, M. and Yata, K. (2011b). Authors' response, *Sequential Analysis* 30: 432-440.

[4] Aoshima, M. and Yata, K. (2013a). Invited review article: statistical inference in high-dimension, low-sample-size settings, *Sugaku* 65, 225-247.

[5] Aoshima, M. and Yata, K. (2013b). The JSS Research Prize Lecture: Effective methodologies for high-dimensional data, *J. Japan Statist. Soc. J* 43, 123-150.

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

[7] Aoshima, M. and Yata, K. (2015a). Asymptotic normality for inference on multisample, high-dimensional mean vectors under mild conditions. *Methodology and Computing in Applied Probability* 17: 419-439.

[8] Aoshima, M. and Yata, K. (2015b). Geometric classifier for multiclass, high-dimensional data. *Sequential Analysis*, Special Issue: Celebrating Seventy Years of Charles Stein's 1945 Seminal Paper on Two-Stage Sampling 34: 279-294.

[9] Aoshima, M. and Yata, K. (2016). Two-sample tests for high-dimension, strongly spiked eigenvalue models, *Statistica Sinica*, to appear.

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

[11] Baik, J. and Silverstein. (2006). Eigenvalues of large sample covariance matrices of spiked population models, *Journal of Multivariate Analysis* 97: 1382-1408.

[12] Cai, T. T., Liu, W. and Xia, Y. (2014). Two sample test of high dimensional means under dependence. *J. R. Statist. Soc. Ser. B* **76**, 349-372.

[13] Chen, S. X. and Qin, Y.-L. (2010a). A two-sample test for high-dimensional data with applications to gene-set testing. *Ann. Statist.* **38**, 808-835.

[14] Chen, S. X., Zhang, L.-X. and Zhong, P.-S. (2010b). Tests for high-dimensional covariance matrices. *J. Amer. Statist. Assoc.* **105**, 810-819.

[15] Dempster, A. P. (1958). A high dimensional two sample significance test. *Ann. Math. Statist.* **29**, 995-1010.

[16] Dempster, A. P. (1960). A significance test for the separation of two highly multivariate small samples. *Biometrics* **16**, 41-50.

[17] Gutiérrez, N.C., López-Pérez, R., Hernández, J.M., Isidro, I., González, B., Delgado, M., Fermiñán, E., García, J.L., Vázquez, L., González, M., and San Miguel, J.F. (2005). Gene expression profile reveals deregulation of genes with relevant functions in the different subclasses of acute myeloid leukemia. *Leukemia* 19: 402-409.

[18] 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.

[19] Ishii, A., Yata, K., and Aoshima, M. (2014). Asymptotic distribution of the largest eigenvalue via geometric representations of high-dimension, low-sample-size data. *Sri Lankan Journal of Applied Statistics*, Special Issue: Modern Statistical Methodologies in the Cutting Edge of Science (ed. Mukhopadhyay, N.), 81-94.

[20] Ishii, A., Yata, K., and Aoshima, M. (2016). Asymptotic properties of the first principal component and equality tests of covariance matrices in high-dimension, low-sample-size context. *Journal of Statistical Planning and Inference*, 170: 186-199.

[21] Ishii, A. (2016a). A two-sample test for high-dimension, low-sample-size data under the strongly spiked eigenvalue model. *Hiroshima Mathematical Journal*, to appear.

[22] Ishii, A. (2016b). A high-dimensional two-sample test for non-Gaussian data under a strongly spiked eigenvalue model, submitted.

[23] Jeffery, I.B., Higgins, D.G., Culhane, A.C., 2006. Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. BMC Bioinformatics 7, 359.

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

[25] Jung, S. and Marron, J.S. (2009). PCA consistency in high dimension, low sample size context, *Annals of Statistics* 37: 4104-4130.

[26] Jung, S., Sen A., and Marron, J.S. (2012). Boundary behavior in high dimension, low sample size asymptotics of PCA, *Journal of Multivariate Analysis* 109, 190-203.

[27] Ma, Y., Lan, W. and Wang, H. (2015). A high dimensional two-sample test under a low dimensional factor structure. *J. Multivariate Anal.* **140**: 162-170.

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

[29] Shipp, M.A., Ross, K.N., Tamayo, P., Weng, A.P., Kutok, J.L., Aguiar R.C., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G.S., Ray, T.S., Koval, M.A., Last, K.W., Norton, A., Lister, T.A., Mesirov, J., Neuberg, D.S., Lander, E.S., Aster, J.C., Golub, T.R., 2002. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nature Medicine 8, 68-74.

[30] Singh, D., Febbo, P.G., Ross, K., Jackson, D.G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A.A., D'Amico, A.V., Richie, J.P., Lander, E.S., Lo[Singh et al. (2002)]da, M., Kantoff, P.W., Golub, T.R., Sellers, W.R., 2002. Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 1, 203-209.

[31] Srivastava, M. S. (2007). Multivariate theory for analyzing high dimensional data. *J. Japan Statist. Soc.* **37**, 53-86.

[32] Srivastava, M.S., Yanagihara, H., (2010). Testing the equality of several covariance matrices with fewer observations than the dimension. J. Multivariate Anal. 101, 1319-1329.

[33] Srivastava, M. S., Katayama, S. and Kano, Y. (2013). A two sample test in high dimensional data. *J. Multivariate Anal.* **114**, 349-358.

[34] Srivastava, M. S., Yanagihara, H. and Kubokawa, T. (2014). Tests for covariance matrices in high dimension with less sample size. *J. Multivariate Anal.* **130**, 289-309.

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

[36] 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.

[37] Yata, K. and Aoshima, M. (2010b). Intrinsic dimensionality estimation of high-dimension, low sample size data with D-asymptotics, *Communications in Statistics. Theory and Methods, Special Issue Honoring Akahira, M. (ed. Aoshima, M.)* 39: 1511-1521.

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

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

[40] Yata, K. and Aoshima, M. (2013b). PCA consistency for the power spiked model in high-dimensional settings, *Journal of Multivariate Analysis* 122: 334-354.

[41] Yata, K. and Aoshima, M. (2016a). High-dimensional inference on covariance structures via the extended cross-data-matrix methodology, *Journal of Multivariate Analysis* 151: 151-166.

[42] Yata, K. and Aoshima, M. (2016b). Reconstruction of a high-dimensional low-rank matrix, *Electronic Journal of Statistics* 10: 895-917.