数理解析研究所講究録
第 1999 巻 2016 年 28-35

Tests of Mean Vectors in High-Dimension, Low-Sample-Size Context

Aki Ishii

Graduate School of Pure and Applied Sciences, University of Tsukuba, Ibaraki, Japan

Abstract: A common feature of high-dimensional data is that the data dimension is high, however, the sample size is relatively low. We call such data HDLSS data. In this paper, we consider a new one-sample test and two-sample test for high-dimensional data under the strongly spiked eigenvalue (SSE) model. We focus on the asymptotic properties of the first principal component to provide new test procedures. We consider HDLSS asymptotic theories as the dimension grows for both the cases when the sample size is fixed and the sample size goes to infinity. We introduce the noise-reduction (NR) methodology and provide asymptotic properties of the largest-eigenvalue estimation. We apply the NR method to the one-sample test and two-sample test. Finally, we give simulation studies and discuss the performance of the new one-sample test procedure.

Keywords: HDLSS; Large p, small n; Noise-reduction methodology; One-sample test; Two-sample test.

1 Introduction

In this paper, we consider the one-sample test and the two-sample test for high-dimensional data. The problem of testing mean vectors has been studied by a lot of papers, however, it is still necessary to study these problems under more suitable conditions for actual high-dimensional data .

Suppose we have two independent $d \times n_i$ data matrices, $\boldsymbol{X}_i = [\boldsymbol{x}_{ij}, ..., \boldsymbol{x}_{in_i}]$, $i = 1, 2$, where $\boldsymbol{x}_{ij}$, $j = 1, ..., n_i$, are independent and identically distributed (i.i.d.) as a d-dimensional distribution with a mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$ ($\geq \boldsymbol{O}$). We assume $n_i \geq 3$, $i = 1, 2$. The eigen-decomposition of $\boldsymbol{\Sigma}_i$ is given by $\boldsymbol{\Sigma}_i = \boldsymbol{H}_i \boldsymbol{\Lambda}_i \boldsymbol{H}_i^T$, where $\boldsymbol{\Lambda}_i = \text{diag}(\lambda_{1(i)}, ..., \lambda_{d(i)})$ having $\lambda_{1(i)} \geq \cdots \geq \lambda_{d(i)} (\geq 0)$ and $\boldsymbol{H}_i = [\boldsymbol{h}_{1(i)}, ..., \boldsymbol{h}_{d(i)}]$ 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 $d \times n_i$ sphered data matrix from a distribution with the zero mean and identity covariance matrix. Let $\boldsymbol{Z}_i = [\boldsymbol{z}_{1(i)}, ..., \boldsymbol{z}_{d(i)}]^T$ and $\boldsymbol{z}_{j(i)} = (z_{j1(i)}, ..., z_{jn_i(i)})^T$, $j = 1, ..., d$, 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 $\boldsymbol{I}_{n_i}$ is the n_i-dimensional identity matrix. Let $\boldsymbol{z}_{oj(i)} = \boldsymbol{z}_{j(i)} - (\bar{z}_{j(i)}, ..., \bar{z}_{j(i)})^T$, $j = 1, ..., d$; $i = 1, 2$, where $\bar{z}_{j(i)} = n_i^{-1} \sum_{k=1}^{n_i} z_{jk(i)}$. 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 $\limsup_{d \to \infty} E(z_{jk(i)}^4) < \infty$ for all i, j, k, and $P(\lim_{d \to \infty} ||\boldsymbol{z}_{o1(i)}|| \neq 0) = 1$ for $i = 1, 2$. As necessary, we consider the following assumption for $z_{1k(i)}$, $k = 1, ..., n_i$:

(A-i) $z_{1k(i)}$, $k = 1, ..., n_i$, are i.i.d. as $N(0, 1)$ for $i = 1, 2$.

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 us write the eigen-decomposition of $\boldsymbol{S}_{in_i}$ as $\boldsymbol{S}_{in_i} = \sum_{j=1}^{d} \hat{\lambda}_{j(i)} \hat{\boldsymbol{h}}_{j(i)} \hat{\boldsymbol{h}}_{j(i)}^T$, where $\hat{\boldsymbol{h}}_{j(i)}$ denotes a unit eigenvector corresponding to $\hat{\lambda}_{j(i)}$.

A famous method to test for mean vectors is Hotelling's T^2 test, however, one cannot use the test statistic in the HDLSS context such as $n_i/d \to 0,\ i = 1,2$. In order to overcome such situations, Dempster [7, 8] and Srivastava [12] considered the two-sample test when the populations π_1 and π_2 are Gaussian. When π_1 and π_2 are non-Gaussian, Bai and Saranadasa [4] and Cai et al. [5] considered the test under the homoscedasticity, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, and Chen and Qin [6] and Aoshima and Yata [1, 2] considered the test under the heteroscedasticity, $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$. We note that those two-sample tests were constructed under the eigenvalue condition as follows:

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

Aoshima and Yata [3] called (1.1) the "non-strongly spiked eigenvalue (NSSE) model". On the other hand, Aoshima and Yata [3] considered the "strongly spiked eigenvalue (SSE) model" as follows:

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

For the SSE model, Katayama et al. [10] considered a one-sample test when $\boldsymbol{x}_{ij}$s are Gaussian. Ishii et al. [9] considered the one-sample test for non-Gaussian cases. Ma et al. [11] considered a two-sample test for the factor model. Aoshima and Yata [3] gave two-sample tests by considering eigenstructures when $d \to \infty$ and $n_i \to \infty,\ i = 1,2$. In this paper, we disscuss a one-sample test and a two-sample test for the SSE model when $d \to \infty$ while n_is are fixed.

In Section 2, we introduce the noise-reduction (NR) methodology and provide asymptotic distribution of the largest-eigenvalue estimation in the HDLSS context. Then, we apply the NR method to the one-sample test for the SSE model in Section 3. In Section4, we consider the two-sample test for the SSE model and give a new test procedure in the HDLSS context. In Section 5, we give simulation studies and discuss the performance of the new test procedure.

2 Asymptotic Properties of the Largest Eigenvalue

In this section, we provide asymptotic properties of the largest eigenvalue. We introduce a method for eigenvalue estimation called the *noise-reduction (NR) methodology* that was proposed by Yata and Aoshima [14]. See Sections 2 and 3 in Yata and Aoshima [14] for the details. When we apply the NR methodology, the NR estimator of $\lambda_{j(i)}$ is given by

$$\tilde{\lambda}_{j(i)} = \hat{\lambda}_{j(i)} - \frac{\mathrm{tr}(\boldsymbol{S}_{in_i}) - \sum_{k=1}^{j} \hat{\lambda}_{k(i)}}{n_i - 1 - j} \quad (j = 1, ..., n_i - 2).$$

Note that $\tilde{\lambda}_{j(i)} \geq 0$ for $j = 1, ..., n_i - 2$. Yata and Aoshima [14, 15] showed that $\tilde{\lambda}_{j(i)}$ has several consistency properties when $d \to \infty$ and $n_i \to \infty$. In this paper, we focus on the largest eigenvalue, $\tilde{\lambda}_{1(i)}$, that has the most important information in data analyses. We assume the following conditions for the largest eigenvalue:

(A-ii) $$\frac{\mathrm{tr}(\boldsymbol{\Sigma}_i^2) - \lambda_{1(i)}^2}{\lambda_{1(i)}^2} = \frac{\sum_{j=2}^{d} \lambda_{j(i)}^2}{\lambda_{1(i)}^2} = o(1) \text{ as } d \to \infty \text{ for } i = 1,2;$$

(A-iii) $$\frac{\sum_{r,s\geq 2}^{d} \lambda_{r(i)}\lambda_{s(i)} E\{(z_{rk(i)}^2 - 1)(z_{sk(i)}^2 - 1)\}}{\lambda_{1(i)}^2} = o(1) \text{ as } d \to \infty \text{ for } i = 1,2.$$

Note that (A-ii) is one of the SSE model (1.2). We also note that (A-ii) implies the condition that $\lambda_{2(i)}/\lambda_{1(i)} \to 0$ as $d \to \infty$. Note that (A-iii) is naturally satisfied when X_i is Gaussian and (A-ii) is met.

Remark 2.1. For a spiked model such as

$$\lambda_{j(i)} = a_{ij} d^{\alpha_{ij}} \ (j = 1, ..., m_i) \quad \text{and} \quad \lambda_{j(i)} = c_{ij} \ (j = m_i + 1, ..., d)$$

with positive and fixed constants, a_{ij}s, c_{ij}s and α_{ij}s, and a positive and fixed integer m_i, (A-ii) holds under the conditions that $\alpha_{i1} > 1/2$ and $\alpha_{i1} > \alpha_{i2}$. See Yata and Aoshima [14] for the details.

Remark 2.2. For several statistical inferences of high-dimensional data, Bai and Saranadasa [4], Chen and Qin [6] and Aoshima and Yata [2] assumed a general factor model as follows:

$$\boldsymbol{x}_{ij} = \boldsymbol{\Gamma}_i \boldsymbol{w}_{ij} + \boldsymbol{\mu}_i$$

for $j = 1, ..., n_i$, where $\boldsymbol{\Gamma}_i$ is a $d \times r_i$ matrix for some $r_i > 0$ such that $\boldsymbol{\Gamma}_i \boldsymbol{\Gamma}_i^T = \boldsymbol{\Sigma}_i$, and $\boldsymbol{w}_{ij}$, $j = 1, ..., n_i$, are i.i.d. random vectors having $E(\boldsymbol{w}_{ij}) = \boldsymbol{0}$ and $\text{Var}(\boldsymbol{w}_{ij}) = \boldsymbol{I}_{r_i}$. As for $\boldsymbol{w}_{ij} = (w_{1j(i)}, ..., w_{rj(i)})^T$, assume that $E(w_{qj(i)}^2 w_{sj(i)}^2) = 1$ and $E(w_{qj(i)} w_{sj(i)} w_{tj(i)} w_{uj(i)}) = 0$ for all $q \neq s, t, u$. From Lemma 1 in Yata and Aoshima [15], one can claim that (A-iii) holds under (A-ii) in the factor model. Also, we note that the factor model naturally holds when $\boldsymbol{X}_i$ is Gaussian.

Then, Ishii et al. [9] gave the following theorem.

Theorem 2.1 ([9]). *Under (A-ii) and (A-iii), it holds that as $d \to \infty$*

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

for $i = 1, 2$. Under (A-i) to (A-iii), it holds that as $d \to \infty$ when n_i is fixed

$$(n_i - 1)\frac{\tilde{\lambda}_{1(i)}}{\lambda_{1(i)}} \Rightarrow \chi^2_{n_i - 1} \quad \textit{for } i = 1, 2.$$

3 One-Sample Test for SSE Model

In this section, we consider the one-sample test in the high-dimensional context. We consider the following test:

$$H_0 : \boldsymbol{\mu}_i = \boldsymbol{0} \quad \text{vs.} \quad H_1 : \boldsymbol{\mu}_i \neq \boldsymbol{0}, \tag{3.1}$$

Bai and Saranadasa [4] proposed a test statistic:

$$T_{BS} = n_i ||\overline{\boldsymbol{x}}_{in_i}||^2 - \text{tr}(\boldsymbol{S}_{in_i}). \tag{3.2}$$

Srivastava and Du [13] proposed a test statistic:

$$T_S = n_i \overline{\boldsymbol{x}}_{in_i}^T \boldsymbol{D}_i^{-1} \overline{\boldsymbol{x}}_{in_i}, \tag{3.3}$$

where $\boldsymbol{D}_i = \text{diag}(s_{11(i)}, ..., s_{dd(i)})$ and $s_{jj(i)}$, $j = 1, ..., d$ are the diagonal elements of $\boldsymbol{S}_{in_i}$. They gave the asymptotic normality of T_{BS} or T_S under H_0 in (3.1) for the NSSE model (1.1). On the other hand, Katayama et al. [10] gave the asymptotic distribution of T_{BS} and T_S for the SSE model (1.2) when $\boldsymbol{X}_i$ is Gaussian.

Now, we consider a new one-sample test for the SSE model by using the asymptotic properties of the largest eigenvalue. By considering T_{BS} in (3.2) under (A-ii), we have the following result.

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

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

either when n_i is fixed or $n_i \to \infty$.

By using the NR method, we consider the following test statistic:

$$F_1 = \frac{n_i||\overline{\boldsymbol{x}}_{in_i}||^2 - \mathrm{tr}(\boldsymbol{S}_{in_i})}{\tilde{\lambda}_{1(i)}} + 1.$$

Note that $E(\tilde{\lambda}_{1(i)}(F_1 - 1)/n_i) = ||\boldsymbol{\mu}_i||^2$. Then, by combining Theorem 2.1 and Lemma 3.1, Ishii et al. [9] gave the following result.

Theorem 3.1 ([9]). *Under (A-i) to (A-iii), it holds as $d \to \infty$ that*

$$F_1 \Rightarrow \begin{cases} F_{1,n_i-1} & \textit{when } n_i \textit{ is fixed,} \\ \chi_1^2 & \textit{when } n_i \to \infty, \end{cases}$$

under H_0 in (3.1), where F_{ν_1,ν_2} denotes a random variable distributed as F distribution with degrees of freedom, ν_1 and ν_2; and χ_ν^2 denotes a random variable distributed as χ^2 distribution with ν degrees of freedom.

For a given $\alpha \in (0, 1/2)$ we test (3.1) by

$$\text{rejecting } H_0 \Longleftrightarrow F_1 > F_{1,n_i-1}(\alpha), \tag{3.4}$$

where $F_{\nu_1,\nu_2}(\alpha)$ denotes the upper α point of F distribution with degrees of freedom, ν_1 and ν_2. Note that $F_{1,n_i-1}(\alpha) \to \chi_1^2(\alpha)$ as $n_i \to \infty$, where $\chi_\nu^2(\alpha)$ denotes the upper α point of χ^2 distribution with ν degrees of freedom. Then, under (A-i) to (A-iii), it holds as $d \to \infty$ that

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

either when n_i is fixed or $n_i \to \infty$.

4 Two-Sample Test for SSE Model

In this section, we consider the two-sample test in the high-dimensional context. Now, we consider the following test:

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \quad \text{vs.} \quad H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2. \tag{4.1}$$

We assume the following assumption:

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

Remark 4.1. Note that (A-iv) is not a general condition for high-dimensional data, so that it is necessary to check. See Lemma 4.1 in Ishii et al. [9] for checking the condition in actual data analyses.

Let

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

Note that $E(T_n) = ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||^2$ and

$$\operatorname{Var}(T_n) = \sum_{i=1}^{2} \frac{\operatorname{tr}(\boldsymbol{\Sigma}_i^2)}{n_i(n_i-1)} + 4\frac{\operatorname{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)}{n_1 n_2} + 4\sum_{i=1}^{2} \frac{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_i (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{n_i}.$$

By using Theorem 1 in Chen and Qin [6] or Theorem 4 in Aoshima and Yata [2], we can claim that as $d \to \infty$ and $n_i \to \infty, i = 1, 2$

$$\frac{T_n}{\operatorname{Var}(T_n)^{1/2}} \Rightarrow N(0,1)$$

under H_0 in (4.1), (1.1) and some regularity conditions.

We consider an asymptotic distribution of T_n under the SSE models. We have the following results.

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

$$\frac{T_n}{\lambda_{1(1)}} = (\bar{z}_{1(1)} - \bar{z}_{1(2)})^2 - \sum_{i=1}^{2} \frac{||\boldsymbol{z}_{o1(i)}/\sqrt{n_i - 1}||^2}{n_i} + o_p(1) \quad \text{under } H_0 \text{ in (4.1)}$$

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

Let $\nu = n_1 + n_2 - 2$. From Theorem 2.1, we have the following result.

Lemma 4.2. *Under (A-i) to (A-iv), it holds as $d \to \infty$ when n_is are fixed that*

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

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

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

Let

$$F_2 = u_n \frac{T_n + \sum_{i=1}^{2} \tilde{\lambda}_{1(i)}/n_i}{\sum_{i=1}^{2}(n_i - 1)\tilde{\lambda}_{1(i)}},$$

where $u_n = \nu(1/n_1 + 1/n_2)^{-1}$. Then, by combining Lemmas 4.1 with 4.2, we have the following theorem.

Theorem 4.1. *Under (A-i) to (A-iv), it holds as $d \to \infty$ that*

$$F_2 \Rightarrow \begin{cases} F_{1,\nu} & \text{when } \nu \text{ is fixed,} \\ \chi_1^2 & \text{when } \nu \to \infty \end{cases}$$

under H_0 in (4.1).

For a given $\alpha \in (0, 1/2)$ we test (4.1) by

$$\text{rejecting } H_0 \Longleftrightarrow F_2 > F_{1,\nu}(\alpha), \tag{4.2}$$

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

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

as $d \to \infty$ either when ν is fixed or $\nu \to \infty$.

5 Simulation Studies

In order to compare the performances of the one-sample test procedures, we used computer simulations. We consider the test (3.1). In this simulation, we compared the test procedure (3.4) to T_{BS} in (3.2) and the test procedures given by Katayama et al. [10]. We set $\alpha = 0.05$ and $\boldsymbol{\Sigma}_i = (\boldsymbol{I}_d + d^{-1}\mathbf{1}_d\mathbf{1}_d^T)/2$, where $\mathbf{1}_d = (1, ..., 1)^T$. For such a situation, Katayama et al. [10] gave the follwoing test procedures:

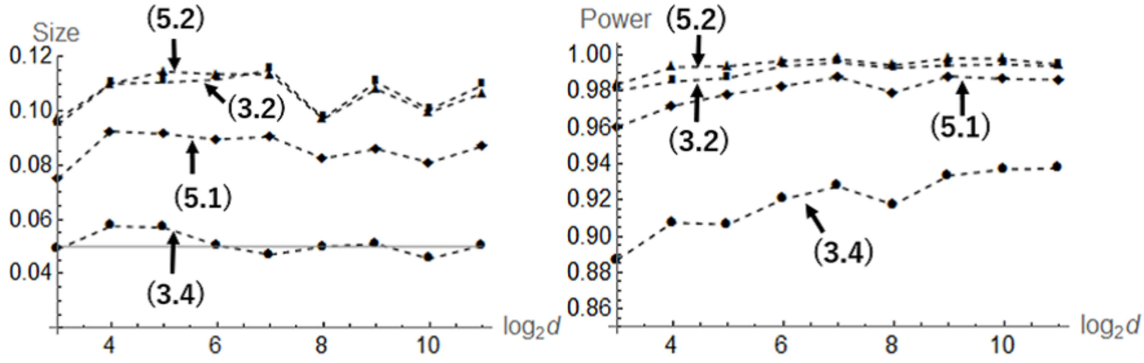
$$\text{rejecting } H_0 \Longleftrightarrow \frac{T_{BS}}{\sqrt{\widehat{\text{tr}(\boldsymbol{\Sigma}^2)}}} + 1 > \chi_1^2(\alpha), \tag{5.1}$$

$$\text{rejecting } H_0 \Longleftrightarrow \frac{T_S - d(n_i - 1)/(n_i - 3)}{\sqrt{\widehat{\text{tr}(\boldsymbol{R}_i^2)}}} + 1 > \chi_1^2(\alpha), \tag{5.2}$$

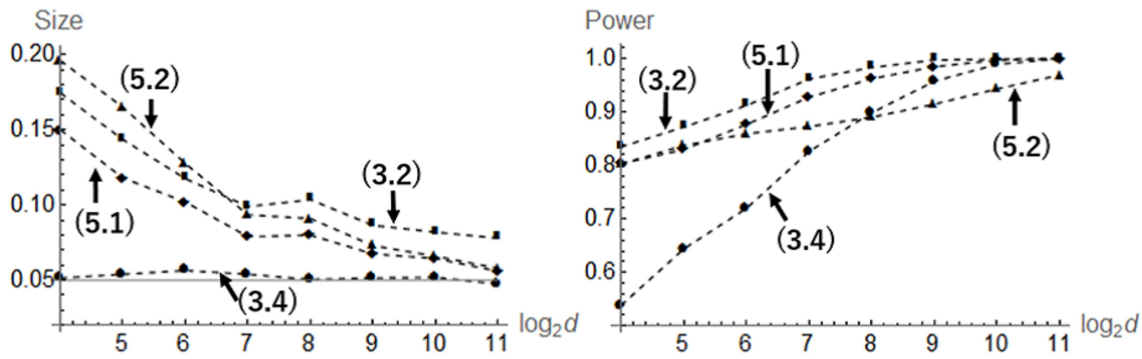
where $\widehat{\text{tr}(\boldsymbol{\Sigma}_i^2)}$ and $\widehat{\text{tr}(\boldsymbol{R}_i^2)}$ are the consistent estimators of $\text{tr}(\boldsymbol{\Sigma}_i^2)$ and $\text{tr}(\boldsymbol{R}_i^2)$, $\boldsymbol{R}_i$ is the population correlation matrix, given in Katyama et al. [10]. We considered the case $\boldsymbol{X}_i$ is Gaussian. Note that (A-i) to (A-iii) hold. We considered two cases (I) $d = 2^k (k = 3, ...11)$ and $n_i = 10$; and (II) $d = 2^k (k = 4, ...11)$ and $n_i = \lceil d^{1/2} \rceil$, where $\lceil x \rceil$ denotes the smallest integer $\geq x$. In order to check the size, we set (a) $\boldsymbol{\mu}_i = 0$ for each case. As for the power, we set (b) $\boldsymbol{\mu}_i = (1, ..., 1, 0, ..., 0)$ whose first $d/2$ elements are 1 for (I); and first $\lceil 3.8d/n_i \rceil$ elements are 1 for (II).

The findings were obtained by averaging the outcomes from 2000 ($= R$, say) replications. We defined $P_r = 1$ (or 0) when H_0 in (3.1) was falsely rejected (or not) for $r = 1, ..., 2000$ in (a) and defined $\overline{\alpha} = \sum_{r=1}^{R} P_r/R$ to estimate the size. We also defined $P_r = 1$ (or 0) when H_1 in (3.1) was falsely rejected (or not) for $r = 1, ..., 2000$ in (b) and defined $1 - \overline{\beta} = 1 - \sum_{r=1}^{R} P_r/R$ to estimate the power. Note that their standard deviations are less than 0.011. In Fig. 1, we plotted $\overline{\alpha}$ in the left panels and $1 - \overline{\beta}$ in the right panels for (I) and (II).

Throughout, the original test procedure T_{BS} in (3.2) does not give a good performance in terms of the size. It is probably because T_{BS} does not hold the asymptotic normality when (1.1) is not met. On the other hand, the tests (5.1) and (5.2) do not give good performances in terms of the size when n_i is small. We observed that the power of (3.2), (5.1) and (5.2) gave better performances compared to that of (3.4) in (I). This is because (3.2), (5.1) and (5.2) cannot control the size. In the case of (II), the size of (5.1) and (5.2) become close to α slowly as both the d and n_i are large. Contrary to that, (3.4) showed a quite good performance in terms of the size even when n_i is small. It should be noted that high-dimensional data often have SSE model and the sample size is quite small. Thus, we conclude that if one can assume (A-ii), we recommend to use the new test procedure (3.4).



(I) $d = 2^k (k = 3, ...11)$ and $n_i = 10$.



(II) $d = 2^k (k = 4, ...11)$ and $n_i = \lceil d^{1/2} \rceil$, where $\lceil x \rceil$ denotes the smallest integer $\geq x$.

Figure 1. We compared the test procedure (3.4) to (3.2), (5.1) and (5.2). We set $\alpha = 0.05$ and $\boldsymbol{X}_i$ is Gaussian. The values of $\overline{\alpha}$ are denoted by the dashed lines in the left panels and $1 - \overline{\beta}$ are denoted by the dashed lines in the right panels.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Professor Makoto Aoshima, for his enthusiastic guidance and helpful support to my research project. I would also like to thank Assistant Professor, Kazuyoshi Yata, for his valuable suggestions.

References

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

[2] Aoshima, M. and Yata, K. (2015). Asymptotic normality for inference on multisample, high-dimensional mean vectors under mild conditions. *Methodol. Comput. Appl. Probab.* **17**, 419-439.

[3] Aoshima, M. and Yata, K. (2016). Two-sample tests for high-dimension, strongly spiked eigenvalue models, arXiv: 1602.02491.

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

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

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

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

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

[9] 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. *J. Statist. Plan. Infer.* **170**, 186-199.

[10] Katayama, S., Kano, Y. and Srivastava, M. S. (2013). Asymptotic distributions of some test criteria for the mean vector with fewer observations than the dimension. *J. Multivariate Anal.* **116**, 410-421.

[11] 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.

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

[13] Srivastava, M. S. and Du, M.(2008). A test for the mean vector with fewer observations than the dimension. *J. Multivariate Anal.* **99**, 386-402.

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

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

Graduate School of Pure and Applied Sciences, University of Tsukuba,
Ibaraki 305-8571, Japan
E-mail address: `ishii-akitk@math.tsukuba.ac.jp`

筑波大学・数理物質科学研究科　　石井　晶