An equality test of high-dimensional covariance matrices under the SSE model

Kazuyoshi Yata
Institute of Mathematics
University of Tsukuba

Aki Ishii
Department of Information Sciences
Tokyo University of Science

Makoto Aoshima
Institute of Mathematics
University of Tsukuba

Abstract

In this paper, we consider an equality test of high-dimensional covariance matrices under the strongly spiked eigenvalue (SSE) model. We introduce an eigenvalue model called the "strongly spiked eigenvalue (SSE) model" which was proposed by Aoshima and Yata (2018). We give a new test procedure based on the spiked eigenstructures.

1 Introduction

Suppose we have two classes $\pi_i,\ i=1,2$. We define independent $d\times n_i$ data matrices, $\boldsymbol{X}_i=[\boldsymbol{x}_{i1},...,\boldsymbol{x}_{in_i}],\ i=1,2,$ for $\pi_i,\ i=1,2,$ 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 $n_i\geq 4,\ i=1,2$. The eigen-decomposition of $\boldsymbol{\Sigma}_i$ is given by

$$\boldsymbol{\Sigma}_i=\boldsymbol{H}_i\boldsymbol{\Lambda}_i\boldsymbol{H}_i^T,$$

where $\boldsymbol{\Lambda}_i=\mathrm{diag}(\lambda_{1(i)},...,\lambda_{d(i)})$ 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

$$\boldsymbol{Z}_i = [\boldsymbol{z}_{1(i)}, ..., \boldsymbol{z}_{d(i)}]^T \text{ and } \boldsymbol{z}_{j(i)} = (z_{j1(i)}, ..., z_{jn_i(i)})^T, \ j = 1, ..., d$$

for $i = 1, 2$. Note that $E(z_{jk(i)} z_{j'k(i)}) = 0 \ (j \neq j')$ and $\mathrm{Var}(\boldsymbol{z}_{j(i)}) = \boldsymbol{I}_{n_i}$, 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 $i = 1, 2$. Also, we consider the following assumption:

(A-i) $E(z_{qj(i)}^2 z_{sj(i)}^2) = 1$ and $E(z_{qj(i)} z_{sj(i)} z_{tj(i)} z_{uj(i)}) = 0$ for all $q \neq s, t, u$.

This kind of assumption was made by Bai and Saranadasa (1996), Chen and Qin (2010) and Aoshima and Yata (2011). We note that (A-i) naturally holds when $\boldsymbol{X}_i$ is Gaussian.

We consider a test problem as follows:

$$H_0 : \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 \quad \text{vs.} \quad H_1 : \boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2. \tag{1}$$

Schott (2007) gave a test procedure when $d/n_i \to c_i \in [0, \infty)$ and normal distribution. Aoshima and Yata (2011) gave a test procedure based on the quantity of $\mathrm{tr}(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2)$. They also discussed sample size determination so as to have prespecified size and power simultaneously. Li and Chen (2012) and Endo et al. (2018) considered the test problem by using the quantity of $\mathrm{tr}\{(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2)^2\}$. The above literatures discussed asymptotic properties of their test procedures when $d \to \infty$ and $n_i \to \infty$ under the following eigenvalue condition:

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

Aoshima and Yata (2018) called (2) the "non-strongly spiked eigenvalue (NSSE) model". On the other hand, Ishii et al. (2016) investigated asymptotic properties of the first principal component and considered the test problem (1) when $d \to \infty$ while n_is are fixed under the following eigenvalue condition:

$$\liminf_{d \to \infty} \left\{ \frac{\lambda_{1(i)}^2}{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)} \right\} > 0 \quad \text{for } i = 1 \text{ or } 2. \tag{3}$$

Aoshima and Yata (2018) called (3) the "strongly spiked eigenvalue (SSE) model" and showed that high-dimensional data often have the SSE model. Ishii (2017a, b) considered two-sample tests under the SSE model when $d \to \infty$ while n_is are fixed. The SSE model is very difficult to handle because of the influence of strongly spiked noise. Aoshima and Yata (2018) created a data-transformation technique for two-sample tests which transforms the SSE model to the NSSE model. In this paper, we focus on the SSE model and give a new test procedure for (1) by using a different approach from the data-transformation technique.

In Section 2, we introduce the test statistic given by Li and Chen (2012). We emphasize that one should construct test procedures by considering the eigenstructure of high-dimensional data. In Section 3, we give a new test procedure under the SSE model.

2 Performance of the earlier test procedure under the SSE model

In this section, we investigate the performance of the test procedure given by Li and Chen (2012). For (1) they assumed

$$\mathrm{tr}(\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_j\boldsymbol{\Sigma}_k\boldsymbol{\Sigma}_l) = o\{\mathrm{tr}(\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_j)\mathrm{tr}(\boldsymbol{\Sigma}_k\boldsymbol{\Sigma}_l)\} \tag{4}$$

for any i, j, k and $l \in \{1, 2\}$. Note that (4) is one of the NSSE models. They proposed a test statistic as follows:

$$U = A_{n_1} + A_{n_2} - 2\mathrm{tr}\left(\boldsymbol{S}_{1n_1}\boldsymbol{S}_{2n_2}\right), \tag{5}$$

where $\boldsymbol{S}_{in_i}$ is the sample covariance matrix having $E(\boldsymbol{S}_{in_i}) = \boldsymbol{\Sigma}_i$ and

$$\begin{aligned} A_{n_i} =& \frac{1}{n_i(n_i-1)}\sum_{j\neq k}^{n_i}(\boldsymbol{x}_{ij}^T\boldsymbol{x}_{ik})^2 - \frac{2}{n_i(n_i-1)(n_i-2)}\sum_{j\neq k\neq l}^{n_i}\boldsymbol{x}_{ik}^T\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^T\boldsymbol{x}_{il} \\ &+ \frac{1}{n_i(n_i-1)(n_i-2)(n_i-3)}\sum_{j\neq k\neq l\neq l'}^{n_i}\boldsymbol{x}_{ij}^T\boldsymbol{x}_{ik}\boldsymbol{x}_{il}^T\boldsymbol{x}_{il'}. \end{aligned}$$

Note that U is an unbiased estimator of

$$||\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2||_F^2 = \mathrm{tr}\{(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2)^2\} \ (= \Delta, \text{ say}).$$

In this paper, we consider the divergence condition such as $d \to \infty$, $n_1 \to \infty$ and $n_2 \to \infty$, which is equivalent to

$$m \to \infty, \quad \text{where } \ m = \min\{d, n_1, n_2\}.$$

Under (4) and some regularity conditions, they showed the following asymptotic result:

$$\frac{U - \Delta}{\mathrm{Var}(U)^{1/2}} \Rightarrow N(0, 1) \text{ as } m \to \infty. \tag{6}$$

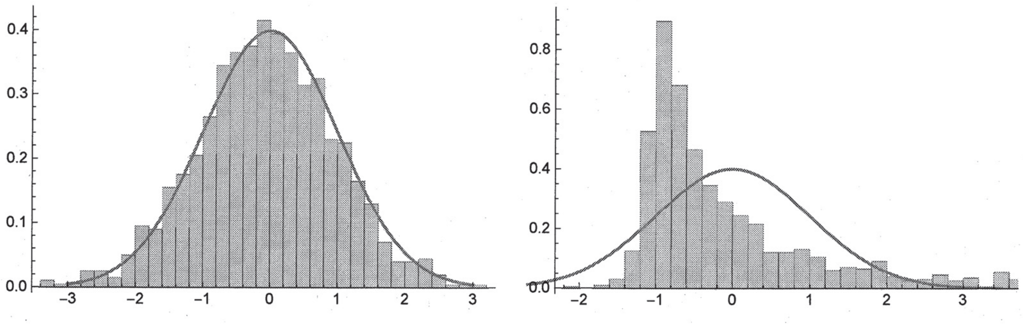
Here, " $\Rightarrow$ " denotes the convergence in distribution and $N(0, 1)$ denotes a random variable distributed as the standard normal distribution.

Let us show a toy example about the asymptotic null distribution of U in (6). We set $d = 2048$ and $n_1 = n_2 = 100$. We assumed $N_d(\boldsymbol{0}, \boldsymbol{\Sigma})$ for each class under $H_0: \ \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$. Let us write a $k \times l$ zero matrix by $\boldsymbol{O}_{k,l}$. We set

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{(1)} & \boldsymbol{O}_{2,d-2} \\ \boldsymbol{O}_{d-2,2} & \boldsymbol{\Sigma}_{(2)} \end{pmatrix} \text{ having } \boldsymbol{\Sigma}_{(2)} = (0.3^{|i-j|})$$

and considered two cases:

$$\text{(i) } \boldsymbol{\Sigma}_{(1)} = \mathrm{diag}(d^{1/3}, d^{1/6}) \quad \text{and} \quad \text{(ii) } \boldsymbol{\Sigma}_{(1)} = \mathrm{diag}(d^{1}, d^{1/2}).$$



(i) When $\mathbf{\Sigma}_{(1)} = \mathrm{diag}(d^{1/3}, d^{1/6})$

(ii) When $\mathbf{\Sigma}_{(1)} = \mathrm{diag}(d^{1}, d^{1/2})$

Figure 1: The histograms of (normalized) U for a NSSE model (in the left panel) and for a SSE model (in the right panel). The solid line denotes the p.d.f. of $N(0,1)$.

Note that (i) is a NSSE model and (ii) is a SSE model. We generated independent pseudorandom observations from each class and calculated (normalized) U 1000 times. In Fig.1, we gave histograms for (i) and (ii). One can observe that U does not converge to $N(0,1)$ for (ii). In order to overcome this inconvenience, we modify U under a SSE model and newly construct a test procedure for the SSE model in Section 3.

3 Modification of U under a SSE model

We assume the following assumption for the eigenvalues:

(A-ii) $$\frac{\sum_{s=2}^{d} \lambda_{s(i)}^2}{\lambda_{1(i)}^2} = o(1) \text{ as } d \to \infty \text{ for } i = 1, 2.$$

Note that (A-ii) is one of the SSE models. Also, note that (A-ii) implies the conditions that $\lambda_{2(i)}/\lambda_{1(i)} \to 0$ and $\lambda_{1(i)}^2/\mathrm{tr}(\mathbf{\Sigma}_i^2) \to 1$ as $d \to \infty$. For a spiked model as

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

with positive (fixed) constants, $a_{j(i)}$s, $c_{j(i)}$s and $\alpha_{j(i)}$s, and a positive (fixed) integer k_i, (A-ii) holds when $\alpha_{1(i)} > 1/2$ and $\alpha_{1(i)} > \alpha_{2(i)}$.

In addition, we consider the following condition:

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

For all i, j, $E\{(z_{1j(i)}^2 - 1)^2\} = 2$ under (A-iii). Let

$$K = 2\lambda_{1(1)}^2/n_1 + 2\lambda_{1(2)}^2/n_2.$$

We have the following result.

Lemma 1 (Ishii et al., 2017). *Under (A-i) to (A-iii) and H_0, it holds that as $m \to \infty$*

$$U = \left(\sum_{j=1}^{n_1} \frac{\lambda_{1(1)}(z_{1j(1)}^2 - 1)}{n_1} - \sum_{k=1}^{n_2} \frac{\lambda_{1(2)}(z_{1k(2)}^2 - 1)}{n_2} \right)^2 - K + o_p(K).$$

Let

$$T = U/K + 1.$$

Then, we have an asymptotic distribution of T under H_0.

Theorem 1 (Ishii et al., 2017). *Under (A-i) to (A-iii) and H_0, it holds that as $m \to \infty$*

$$T \Rightarrow \chi_1^2.$$

Here, χ_ν^2 denotes a random variable distributed as a χ^2 distribution with ν degrees of freedom.

Since $\lambda_{1(i)}$s are unknown, we need to estimate them. It is well known that the sample eigenvalues get too much noise for high-dimensional data. See Jung and Marron (2009), Yata and Aoshima (2009), Ishii et al. (2016) and Shen et al. (2016) for the details. We consider estimating $\lambda_{1(i)}$s by using the *noise-reduction (NR) methodology* given by Yata and Aoshima (2012). We denote the dual matrix of $\boldsymbol{S}_{in_i}$ by $\boldsymbol{S}_{Dn_i}$ and define its eigen-decomposition as follows:

$$\begin{aligned} \boldsymbol{S}_{iD} &= (n_i - 1)^{-1}(\boldsymbol{X}_i - \overline{\boldsymbol{X}}_i)^T(\boldsymbol{X}_i - \overline{\boldsymbol{X}}_i) \\ &= \sum_{s=1}^{n_i - 1} \hat{\lambda}_{s(i)} \hat{\boldsymbol{u}}_{s(i)} \hat{\boldsymbol{u}}_{s(i)}^T, \end{aligned}$$

where $\overline{\boldsymbol{X}}_i = [\bar{\boldsymbol{x}}_i, ..., \bar{\boldsymbol{x}}_i]$ and $\bar{\boldsymbol{x}}_i = n_i^{-1} \sum_{j=1}^{n_i} \boldsymbol{x}_{ij}$ for $i = 1, 2$. If one uses the NR method, $\lambda_{j(i)}$s are estimated by

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

Note that $\tilde{\lambda}_{j(i)} \geq 0$ w.p.1 for $j = 1, ..., n_i - 2$. Yata and Aoshima (2012, 2013) showed that $\tilde{\lambda}_{j(i)}$ has consistency properties when $d \to \infty$ and $n_i \to \infty$. On the other hand, Ishii et al. (2016) gave asymptotic properties of $\tilde{\lambda}_{1(i)}$ when $d \to \infty$ while n_i is fixed. Let $s_{1(i)} = \sum_{j=1}^{n_i}(z_{1j(i)} - \bar{z}_{1(i)})^2/(n_i - 1)$ for $i = 1, 2$, where $\bar{z}_{1(i)} = n_i^{-1} \sum_{j=1}^{n_i} z_{1j(i)}$.

Theorem 2 (Yata and Aoshima, 2013 and Ishii et al., 2016). *Under (A-i) and (A-ii), it holds that as $d \to \infty$*

$$\frac{\tilde{\lambda}_{1(i)}}{\lambda_{1(i)}} = \begin{cases} s_{1(i)} + o_p(1) & \textit{when } n_i \textit{ is fixed,} \\ 1 + o_p(1) & \textit{when } n_i \to \infty. \end{cases}$$

Under (A-i) to (A-iii), it holds that as $d \to \infty$

$$(n_i - 1)\frac{\tilde{\lambda}_{1(i)}}{\lambda_{1(i)}} \Rightarrow \chi^2_{n_i - 1} \qquad \textit{when } n_i \textit{ is fixed,}$$

$$\textit{and} \quad \sqrt{\frac{n_i - 1}{2}}\left(\frac{\tilde{\lambda}_{1(i)}}{\lambda_{1(i)}} - 1\right) \Rightarrow N(0,1) \quad \textit{when } n_i \to \infty.$$

Let $\widetilde{K} = 2\tilde{\lambda}^2_{1(1)}/n_1 + 2\tilde{\lambda}^2_{1(2)}/n_2$ and

$$\widetilde{T} = U/\widetilde{K} + 1.$$

We have the following result.

Theorem 3 (Ishii et al., 2017). *Under (A-i) to (A-iii) and H_0, it holds that as $m \to \infty$*

$$\widetilde{T} \Rightarrow \chi^2_1.$$

We consider testing (1) for a given $\alpha \in (0, 1/2)$ by

$$\text{rejecting } H_0 \Longleftrightarrow \widetilde{T} \geq c_1(\alpha), \tag{7}$$

where $c_1(\alpha)$ denotes the upper α point of χ^2_1. Then, under (A-i) to (A-iii), it holds that as $m \to \infty$

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

See Ishii et al. (2017) for the asymptotic power of the test procedure by (7).

4 Simulation

We compared the performance of the test by $\widetilde{T}$ with the test by U in numerical simulations. Independent pseudo-random observations were generated from $N_d(\mathbf{0}, \boldsymbol{\Sigma}_i)$. We set $\alpha = 0.05$ and

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

where $\boldsymbol{O}_{k,l}$ is the $k \times l$ zero matrix. We set

$$\boldsymbol{\Sigma}_{1(1)} = \text{diag}(d^{2/3}, d^{1/2}) \text{ and } \boldsymbol{\Sigma}_{1(2)} = (0.3^{|i-j|^{1/3}}).$$

As for the alternative hypothesis, we considered $\boldsymbol{\Sigma}_2 = 2\boldsymbol{\Sigma}_1$. Note that (A-i) to (A-iii) are met. We set $(n_1, n_2) = (\lceil 3d^{1/2} \rceil, \lceil 4d^{1/2} \rceil)$ and $d = 2^s$ for $s = 5, ..., 10$, where $\lceil x \rceil$ denotes the smallest integer $\geq x$.

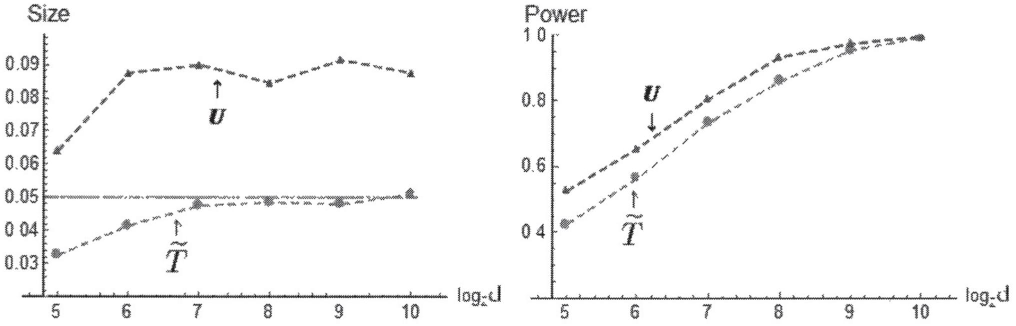


Figure 2: The performances of the tests by $\widetilde{T}$ and U. We set $(n_1, n_2) = (\lceil 3d^{1/2} \rceil, \lceil 4d^{1/2} \rceil)$ and $d = 2^s$ for $s = 5, ..., 10$. The values of $\overline{\alpha}$ are denoted by the dashed lines in the left panel and the values of $1 - \overline{\beta}$ are denoted by the dashed lines in the right panel.

We checked the performance by 2000 replications. 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. 2, we plotted $\overline{\alpha}$ (left panel) and $1 - \overline{\beta}$ (right panel).

One can observe that the test by $\widetilde{T}$ gave preferable performances. On the other hand, the test by U gave a bad performance with respect to the size. Remember that U was constructed under (2). We emphasize that it is very important to select a suitable test procedure depending on the eigenstructure.

Acknowledgements

Research of the first author was partially supported by Grant-in-Aid for Scientific Research (C), Japan Society for the Promotion of Science (JSPS), under Contract Number 18K03409. Research of the second author was partially supported by Grants-in-Aid for Young Scientists, JSPS, under Contract Number 18K18015. Research of the third author was partially supported by Grants-in-Aid for Scientific Research (A) and Challenging Research (Exploratory), JSPS, under Contract Numbers 15H01678 and 17K19956.

References

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

[2] Aoshima, M., Yata, K., 2018. Two-sample tests for high-dimension, strongly spiked eigenvalue models. Stat. Sin. 28, 43-62.

[3] Bai, Z., Saranadasa, H., 1996. Effect of high dimension: By an example of a two sample problem. Stat. Sin. 6, 311-329.

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

[5] Endo, K., Yata, K., Aoshima, M., 2018. A test for high-dimensional covariance matrices via the extended cross-data-matrix methodology. RIMS Kokyuroku, submitted.

[6] Ishii, A., Yata, K., Aoshima, M., 2016. Asymptotic properties of the first principal component and equality tests of covariance matrices in high-dimension, low-sample-size context. J. Stat. Plan. Inference 170, 186-199.

[7] Ishii, A., Yata, K., Aoshima, M., 2017. Equality tests of high-dimensional covariance matrices under the strongly spiked eigenvalue model, submitted.

[8] Ishii, A., 2017a. A two-sample test for high-dimension, low-sample-size data under the strongly spiked eigenvalue model. Hiroshima Math. J. 47, 273-288.

[9] Ishii, A., 2017b. A high-dimensional two-sample test for non-Gaussian data under a strongly spiked eigenvalue model. J. Japan Statist. Soc. 47, 273-291.

[10] Jung, S., Marron, J.S., 2009. PCA consistency in high dimension, low sample size context. Ann. Statist. 37, 4104-4130.

[11] Li, J., Chen, S.X., 2012. Two sample tests for high-dimensional covariance matrices. Ann. Statist. 40, 908-940.

[12] Schott, J.R., 2007. A test for the equality of covariance matrices when the dimension is large relative to the sample sizes. Comput. Statist. Data Anal. 51, 6535-6542.

[13] Shen, D., Shen, H., Zhu, H., Marron, J.S., 2016. The statistics and mathematics of high dimension low sample size asymptotics. Stat. Sin. 26, 1747-1770.

[14] Yata, K., Aoshima, M., 2009. PCA consistency for non-Gaussian data in high dimension, low sample size context. Commun. Statist. Theory Methods, Special Issue Honoring Zacks, S. (ed. Mukhopadhyay, N.) 38, 2634-2652.

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

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

Institute of Mathematics
University of Tsukuba
Ibaraki 305-8571
Japan
E-mail address: yata@math.tsukuba.ac.jp