High-dimensional two-sample tests under strongly spiked eigenvalue models

Makoto Aoshima and Kazuyoshi Yata

University of Tsukuba

Abstract: We consider a new two-sample test for high-dimensional data under the strongly spiked eigenvalue (SSE) model. We provide a general test statistic as a function of positive-semidefinite matrices. We investigate the test statistic under the SSE model by considering strongly spiked eigenstructures and create a new effective test procedure for the SSE model.

Key words and phrases: Asymptotic normality, eigenstructure estimation, large p small n, noise reduction methodology, spiked model.

1. Introduction

A common feature of high-dimensional data is that the data dimension is high, however, the sample size is relatively low. This is the so-called "HDLSS" or "large p, small n" data, where p is the data dimension, n is the sample size and $p/n \to \infty$. Statistical inference on this type of data is becoming increasingly relevant, especially in the areas of medical diagnostics, engineering and other big data. Suppose we have independent samples of p-variate random variables from two populations, $\pi_i,\ i = 1, 2$, having an unknown mean vector $\boldsymbol{\mu}_i$ and unknown positive-definite covariance matrix $\boldsymbol{\Sigma}_i$ for each π_i. We do not assume that the population distributions are Gaussian. The eigen-decomposition of $\boldsymbol{\Sigma}_i$ $(i = 1, 2)$ is given by $\boldsymbol{\Sigma}_i = \boldsymbol{H}_i\boldsymbol{\Lambda}_i\boldsymbol{H}_i^T = \sum_{j=1}^{p} \lambda_{ij}\boldsymbol{h}_{ij}\boldsymbol{h}_{ij}^T$, where $\boldsymbol{\Lambda}_i = \mathrm{diag}(\lambda_{i1}, ..., \lambda_{ip})$ is a diagonal matrix of eigenvalues, $\lambda_{i1} \ge \cdots \ge \lambda_{ip} > 0$, and $\boldsymbol{H}_i = [\boldsymbol{h}_{i1}, ..., \boldsymbol{h}_{ip}]$ is an orthogonal matrix of the corresponding eigenvectors. Note that λ_{i1} is the largest eigenvalue of $\boldsymbol{\Sigma}_i$ for $i = 1, 2$. Having recorded i.i.d. samples, $\boldsymbol{x}_{ij}, j = 1, ..., n_i$, from each π_i, let $\boldsymbol{x}_{ij} = \boldsymbol{H}_i\boldsymbol{\Lambda}_i^{1/2}\boldsymbol{z}_{ij} + \boldsymbol{\mu}_i$, where $\boldsymbol{z}_{ij} = (z_{i1j}, ..., z_{ipj})^T$ is considered as a sphered data vector having the zero mean vector and identity covariance matrix. We assume that the fourth moments of each variable in $\boldsymbol{z}_{ij}$ are uniformly bounded. When $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, we simply omit the population index from $\boldsymbol{\Sigma}_i,\ \lambda_{ij}$s and $\boldsymbol{h}_{ij}$s. For example, we write the covariance matrix as $\boldsymbol{\Sigma}$ when $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$.

In this paper, we consider the two-sample test:

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \quad \text{vs.} \quad H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2. \tag{1.1}$$

Having recorded i.i.d. samples, $\boldsymbol{x}_{ij},\ j = 1, ..., n_i$, from each π_i, 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 $i = 1, 2$. We assume $n_i \geq 4$ for $i = 1, 2$. Hotelling's T^2-statistic is defined by

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}(\overline{\boldsymbol{x}}_{1n_1} - \overline{\boldsymbol{x}}_{2n_2})^T \boldsymbol{S}^{-1}(\overline{\boldsymbol{x}}_{1n_1} - \overline{\boldsymbol{x}}_{2n_2}),$$

where $\boldsymbol{S} = \{(n_1 - 1)\boldsymbol{S}_{1n_1} + (n_2 - 1)\boldsymbol{S}_{2n_2}\}/(n_1 + n_2 - 2)$. However, $\boldsymbol{S}^{-1}$ does not exist in the HDLSS context such as $p/n_i \to \infty,\ i = 1, 2$. In such situations, Dempster (1958, 1960) and Srivastava (2007) considered the test when π_1 and π_2 are Gaussian. When π_1 and π_2 are non-Gaussian, Bai and Saranadasa (1996) and Cai et al. (2014) considered the test under homoscedasticity, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$. On the other hand, Chen and Qin (2010) and Aoshima and Yata (2011, 2015) considered the "distance-based two-sample test" under heteroscedasticity, $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$. As discussed in Section 2 of Aoshima and Yata (2015), the distance-based two-sample test is quite flexible for high-dimension, non-Gaussian data. We note that those two-sample tests were constructed under the eigenvalue condition as follows:

$$\frac{\lambda_{i1}^2}{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)} \to 0 \ \text{ as } p \to \infty \text{ for } i = 1, 2. \tag{1.2}$$

However, if (1.2) is not met, one cannot use those two-sample tests. See Aoshima and Yata (2016) for the details. Aoshima and Yata (2016) called (1.2) the "non-strongly spiked eigenvalue (NSSE) model". On the hand, Aoshima and Yata (2016) considered the "strongly spiked eigenvalue (SSE) model" as follows:

$$\liminf_{p\to\infty} \left\{ \frac{\lambda_{i1}^2}{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)} \right\} > 0 \ \text{ for } i = 1 \text{ or } 2. \tag{1.3}$$

We emphasize that high-dimensional data often have the SSE model. See Fig. 1 in Yata and Aoshima (2013) and Section 8 in Aoshima and Yata (2016). For the SSE model, Katayama et al. (2013) considered a one-sample test when the population distribution is Gaussian. Ishii et al. (2016) considered the one-sample test for non-Gaussian cases. Ma et al. (2015) considered a two-sample test for the factor model when $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$.

In this paper, we propose a new effective test procedure for the SSE model. In Section 2, we provide a general test statistic as a function of positive-semidefinite matrices. We investigate the test statistic under the SSE model by considering strongly spiked eigenstructures. In Section 3, we create a new test procedure by estimating the eigenstructures for the SSE model.

2. Test statistic using eigenstructures

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

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

Let

$$\Psi_{i(s)} = \sum_{j=s}^{p} \lambda_{ij}^2 \quad \text{for } i = 1,2;\ s = 1, ..., p.$$

We consider the following model:

(A-i) For $i = 1,2$, there exists a positive fixed integer k_i such that $\lambda_{i1}, ..., \lambda_{ik_i}$ are distinct in the sense that $\liminf_{p\to\infty}(\lambda_{ij}/\lambda_{ij'} - 1) > 0$ when $1 \le j < j' \le k_i$, and λ_{ik_i} and λ_{ik_i+1} satisfy

$$\liminf_{p\to\infty} \frac{\lambda_{ik_i}^2}{\Psi_{i(k_i)}} > 0 \ \text{ and } \ \frac{\lambda_{ik_i+1}^2}{\Psi_{i(k_i+1)}} \to 0 \ \text{ as } p \to \infty.$$

Note that (A-i) implies (1.3), that is (A-i) is one of the SSE models. (A-i) is also a power spiked model given by Yata and Aoshima (2013). We consider the following test statistic with positive-semidefinite matrices, $\boldsymbol{A}_i,\ i = 1,2$, of dimension p:

$$T(\boldsymbol{A}_1, \boldsymbol{A}_2) = 2\sum_{i=1}^{2} \frac{\sum_{j<j'}^{n_i} \boldsymbol{x}_{ij}^T \boldsymbol{A}_i \boldsymbol{x}_{ij'}}{n_i(n_i-1)} - 2\overline{\boldsymbol{x}}_{1n_1}^T \boldsymbol{A}_1^{1/2} \boldsymbol{A}_2^{1/2} \overline{\boldsymbol{x}}_{2n_2}.$$

Let $\boldsymbol{I}_p$ denote the identity matrix of dimension p. Note that $T(\boldsymbol{I}_p, \boldsymbol{I}_p)$ is equivalent to the distance-based two-sample test. Let us write that $\boldsymbol{\mu}_{A_{12}} = \boldsymbol{A}_1^{1/2}\boldsymbol{\mu}_1 - \boldsymbol{A}_2^{1/2}\boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}_{i,A_i} = \boldsymbol{A}_i^{1/2}\boldsymbol{\Sigma}_i\boldsymbol{A}_i^{1/2}$, $i = 1,2$. Let $\Delta(\boldsymbol{A}_1, \boldsymbol{A}_2) = ||\boldsymbol{\mu}_{A_{12}}||^2$ and $K(\boldsymbol{A}_1, \boldsymbol{A}_2) = K_1(\boldsymbol{A}_1, \boldsymbol{A}_2) + K_2(\boldsymbol{A}_1, \boldsymbol{A}_2)$, where

$$K_1(\boldsymbol{A}_1, \boldsymbol{A}_2) = 2\sum_{i=1}^{2} \frac{\mathrm{tr}(\boldsymbol{\Sigma}_{i,A_i}^2)}{n_i(n_i-1)} + 4\frac{\mathrm{tr}(\boldsymbol{\Sigma}_{1,A_i}\boldsymbol{\Sigma}_{2,A_i})}{n_1 n_2}$$

and $K_2(\boldsymbol{A}_1, \boldsymbol{A}_2) = 4\sum_{i=1}^{2} \boldsymbol{\mu}_{A_{12}}^T \boldsymbol{\Sigma}_{i,A} \boldsymbol{\mu}_{A_{12}}/n_i$. Note that $E\{T(\boldsymbol{A}_1, \boldsymbol{A}_2)\} = \Delta(\boldsymbol{A}_1, \boldsymbol{A}_2)$ and $\mathrm{Var}\{T(\boldsymbol{A}_1, \boldsymbol{A}_2)\} = K(\boldsymbol{A}_1, \boldsymbol{A}_2)$. Let $\lambda_{\max}(\boldsymbol{B})$ denote the largest eigenvalue of any positive-semidefinite matrix, $\boldsymbol{B}$. We consider the following condition:

$$\frac{\{\lambda_{\max}(\boldsymbol{\Sigma}_{i,A_i})\}^2}{\mathrm{tr}(\boldsymbol{\Sigma}_{i,A_i}^2)} \to 0 \quad \text{as } p \to \infty \text{ for } i = 1, 2. \tag{2.1}$$

Then, Aoshima and Yata (2016) showed that as $m \to \infty$

$$\frac{T(\boldsymbol{A}_1, \boldsymbol{A}_2) - \Delta(\boldsymbol{A}_1, \boldsymbol{A}_2)}{\{K(\boldsymbol{A}_1, \boldsymbol{A}_2)\}^{1/2}} \Rightarrow N(0, 1) \tag{2.2}$$

under (2.1), $\limsup_{m\to\infty}\{\Delta(\boldsymbol{A}_1, \boldsymbol{A}_2)\}^2/K_1(\boldsymbol{A}_1, \boldsymbol{A}_2) < \infty$ and some regularity conditions. Here, "$\Rightarrow$" denotes the convergence in distribution and $N(0, 1)$ denotes a random variable distributed as the standard normal distribution.

We consider $\boldsymbol{A}_i$s as

$$\boldsymbol{A}_{i(k_i)} = \boldsymbol{I}_p - \sum_{j=1}^{k_i} \boldsymbol{h}_{ij}\boldsymbol{h}_{ij}^T = \sum_{j=k_i+1}^{p} \boldsymbol{h}_{ij}\boldsymbol{h}_{ij}^T \quad \text{for } i = 1, 2.$$

Note that $\boldsymbol{A}_{i(k_i)} = \boldsymbol{A}_{i(k_i)}^{1/2}$. Let $\boldsymbol{\Sigma}_{i*} = \boldsymbol{A}_{i(k_i)}^{1/2}\boldsymbol{\Sigma}_i\boldsymbol{A}_{i(k_i)}^{1/2} = \sum_{j=k_i+1}^{p} \lambda_{ij}\boldsymbol{h}_{ij}\boldsymbol{h}_{ij}^T$ for $i = 1, 2$. Then, it holds that $\mathrm{tr}(\boldsymbol{\Sigma}_{i*}^2) = \Psi_{i(k_i+1)}$ and $\lambda_{\max}(\boldsymbol{\Sigma}_{i*}) = \lambda_{k_i+1}$ for $i = 1, 2$, so that (2.1) is met when $\boldsymbol{A}_i = \boldsymbol{A}_{i(k_i)}$, $i = 1, 2$, under (A-i). Hence, for $\boldsymbol{A}_i = \boldsymbol{A}_{i(k_i)}$, $i = 1, 2$, we can claim (2.2) under (A-i) instead of (2.1). Hereafter, we simply write $T_* = T(\boldsymbol{A}_{1(k_1)}, \boldsymbol{A}_{2(k_2)})$, $\boldsymbol{\mu}_{i*} = \boldsymbol{A}_{i(k_i)}\boldsymbol{\mu}_i$ for $i = 1, 2$, $\Delta_* = \Delta(\boldsymbol{A}_{1(k_1)}, \boldsymbol{A}_{2(k_2)}) = ||\boldsymbol{\mu}_{1*} - \boldsymbol{\mu}_{2*}||^2$, $K_* = K(\boldsymbol{A}_{1(k_1)}, \boldsymbol{A}_{2(k_2)})$ and

$$K_{1*} = K_1(\boldsymbol{A}_{1(k_1)}, \boldsymbol{A}_{2(k_2)}) = 2\sum_{i=1}^{2} \frac{\mathrm{tr}(\boldsymbol{\Sigma}_{i*}^2)}{n_i(n_i - 1)} + 4\frac{\mathrm{tr}(\boldsymbol{\Sigma}_{1*}\boldsymbol{\Sigma}_{2*})}{n_1 n_2}.$$

Note that $\mathrm{tr}(\boldsymbol{\Sigma}_{i*}^2) = \Psi_{i(k_i+1)}$ for $i = 1, 2$. Let

$$x_{ijl} = \boldsymbol{h}_{ij}^T\boldsymbol{x}_{il} = \lambda_{ij}^{1/2} z_{ijl} + \mu_{i(j)} \quad \text{for all } i, j, l, \text{ where } \mu_{i(j)} = \boldsymbol{h}_{ij}^T\boldsymbol{\mu}_i.$$

Then, we write that

$$\begin{aligned} T_* =& 2\sum_{i=1}^{2} \frac{\sum_{l<l'}^{n_i}(\boldsymbol{x}_{il}^T\boldsymbol{x}_{il'} - \sum_{j=1}^{k_i} x_{ijl}x_{ijl'})}{n_i(n_i - 1)} \\ &- 2\frac{\sum_{l=1}^{n_1}\sum_{l'=1}^{n_2}(\boldsymbol{x}_{1l} - \sum_{j=1}^{k_1} x_{1jl}\boldsymbol{h}_{1j})^T(\boldsymbol{x}_{2l'} - \sum_{j=1}^{k_2} x_{2jl'}\boldsymbol{h}_{2j})}{n_1 n_2}. \end{aligned}$$

In order to use T_*, it is necessary to estimate x_{ijl}s and $\boldsymbol{h}_{ij}$s.

3. Test procedure using eigenstructures for the SSE model

In this section, we assume (A-i) and the following assumption for $\pi_i,\ i=1,2$:

(A-ii) $E(z_{isj}^2 z_{itj}^2)=E(z_{isj}^2)E(z_{itj}^2)$, $E(z_{isj}z_{itj}z_{iuj})=0$ and $E(z_{isj}z_{itj}z_{iuj}z_{ivj})=0$ for all $s\neq t,u,v$, with z_{ijl}s defined in Section 1.

When the π_is are Gaussian, (A-ii) naturally holds. First, we discuss estimation of the eigenvalues and eigenvectors in the SSE model.

3.1. Estimation of eigenvalues and eigenvectors

Throughout this section, we omit the subscript with regard to the population for the sake of simplicity. Let $\hat{\lambda}_1\ge\cdots\ge\hat{\lambda}_p\ge 0$ be the eigenvalues of $\boldsymbol{S}_n$. Let us write the eigen-decomposition of $\boldsymbol{S}_n$ as $\boldsymbol{S}_n=\sum_{j=1}^p\hat{\lambda}_j\hat{\boldsymbol{h}}_j\hat{\boldsymbol{h}}_j^T$, where $\hat{\boldsymbol{h}}_j$ denotes a unit eigenvector corresponding to $\hat{\lambda}_j$. We assume $\boldsymbol{h}_j^T\hat{\boldsymbol{h}}_j\ge 0$ w.p.1 for all j without loss of generality. Let $\boldsymbol{X}=[\boldsymbol{x}_1,...,\boldsymbol{x}_n]$ and $\overline{\boldsymbol{X}}=[\overline{\boldsymbol{x}}_n,...,\overline{\boldsymbol{x}}_n]$. Then, we define the $n\times n$ dual sample covariance matrix by

$$\boldsymbol{S}_D=(n-1)^{-1}(\boldsymbol{X}-\overline{\boldsymbol{X}})^T(\boldsymbol{X}-\overline{\boldsymbol{X}}).$$

Note that $\boldsymbol{S}_n$ and $\boldsymbol{S}_D$ share non-zero eigenvalues. 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 $\hat{\boldsymbol{h}}_j$ can be calculated by $\hat{\boldsymbol{h}}_j=\{(n-1)\hat{\lambda}_j\}^{-1/2}(\boldsymbol{X}-\overline{\boldsymbol{X}})\hat{\boldsymbol{u}}_j$. Let $\delta=\sum_{j=k+1}^p\lambda_j/(n-1)$. Let $m_0=\min\{p,n\}$. First, we have the following result.

Proposition 1 (Aoshima and Yata, 2016)**.** *Assume (A-i) and (A-ii). It holds for $j=1,...,k$, that as $m_0\to\infty$*

$$\frac{\hat{\lambda}_j}{\lambda_j}=1+\frac{\delta}{\lambda_j}+O_P(n^{-1/2})\quad \textit{and}\quad (\hat{\boldsymbol{h}}_j^T\boldsymbol{h}_j)^2=\Big(1+\frac{\delta}{\lambda_j}\Big)^{-1}+O_P(n^{-1/2}).$$

If $\delta/\lambda_j\to\infty$ as $m_0\to\infty$, $\hat{\lambda}_j$ and $\hat{\boldsymbol{h}}_j$ are strongly inconsistent in the sense that $\lambda_j/\hat{\lambda}_j=o_P(1)$ and $(\hat{\boldsymbol{h}}_j^T\boldsymbol{h}_j)^2=o_P(1)$. In order to overcome the curse of dimensionality, Yata and Aoshima (2012) proposed an eigenvalue estimation called the noise-reduction (NR) methodology, which was brought about by a geometric representation of $\boldsymbol{S}_D$. If one applies the NR methodology, the λ_js are estimated by

$$\tilde{\lambda}_j = \hat{\lambda}_j - \frac{\mathrm{tr}(\boldsymbol{S}_D) - \sum_{l=1}^{j} \hat{\lambda}_l}{n-1-j} \quad (j = 1, ..., n-2). \tag{3.1}$$

Note that $\tilde{\lambda}_j \geq 0$ w.p.1 for $j = 1, ..., n-2$, and the second term in (3.1) is an estimator of δ. When applying the NR methodology to the PC direction vector, one obtains

$$\tilde{\boldsymbol{h}}_j = \{(n-1)\tilde{\lambda}_j\}^{-1/2}(\boldsymbol{X} - \overline{\boldsymbol{X}})\hat{\boldsymbol{u}}_j \tag{3.2}$$

for $j = 1, ..., n-2$. Then, we have the following result.

Proposition 2 (Aoshima and Yata, 2016)**.** *Assume (A-i) and (A-ii). It holds for $j = 1, ..., k$, that as $m_0 \to \infty$*

$$\frac{\tilde{\lambda}_j}{\lambda_j} = 1 + O_P(n^{-1/2}) \quad \textit{and} \quad (\tilde{\boldsymbol{h}}_j^T \boldsymbol{h}_j)^2 = 1 + O_P(n^{-1}).$$

We note that $\tilde{\boldsymbol{h}}_j$ is a consistent estimator of $\boldsymbol{h}_j$ in terms of the inner product even when $\delta/\lambda_j \to \infty$ as $m_0 \to \infty$.

On the other hand, we note that $\boldsymbol{h}_j^T(\boldsymbol{x}_l - \boldsymbol{\mu}) = \lambda_j^{1/2} z_{jl}$ for all j, l. For $\hat{\boldsymbol{h}}_j$ and $\tilde{\boldsymbol{h}}_j$, we have the following result.

Proposition 3 (Aoshima and Yata, 2016)**.** *Assume (A-i) and (A-ii). It holds for $j = 1, ..., k$ $(l = 1, ..., n)$ that as $m_0 \to \infty$*

$$\lambda_j^{-1/2}\hat{\boldsymbol{h}}_j^T(\boldsymbol{x}_l - \boldsymbol{\mu}) = \frac{z_{jl} + (n-1)^{1/2}\hat{u}_{jl}\lambda_j^{-1}\delta\{1 + o_P(1)\}}{(1 + \lambda_j^{-1}\delta)^{1/2}} + O_P(n^{-1/2});$$

$$\lambda_j^{-1/2}\tilde{\boldsymbol{h}}_j^T(\boldsymbol{x}_l - \boldsymbol{\mu}) = z_{jl} + (n-1)^{1/2}\hat{u}_{jl}\lambda_j^{-1}\delta\{1 + o_P(1)\} + O_P(n^{-1/2}).$$

Let us consider the standard deviation of the above quantities. Note that $[\sum_{l=1}^{n}\{(n-1)^{1/2}\hat{u}_{jl}\delta/\lambda_j\}^2/n]^{1/2} = O(\delta/\lambda_j)$ and $\delta = O(p/n)$ for $\lambda_{k+1} = O(1)$. Hence, in Proposition 3, the inner products are very biased when p is large. Now, we explain the main reason why the inner products involve the large biased terms. Let $\boldsymbol{P}_n = \boldsymbol{I}_n - \boldsymbol{1}_n\boldsymbol{1}_n^T/n$, where $\boldsymbol{1}_n = (1, ..., 1)^T$. Note that $\boldsymbol{1}_n^T\hat{\boldsymbol{u}}_j = 0$ and $\boldsymbol{P}_n\hat{\boldsymbol{u}}_j = \hat{\boldsymbol{u}}_j$ when $\hat{\lambda}_j > 0$ since $\boldsymbol{1}_n^T\boldsymbol{S}_D\boldsymbol{1}_n = 0$. Also, when $\hat{\lambda}_j > 0$, note that

$$\{(n-1)\tilde{\lambda}_j\}^{1/2}\tilde{\boldsymbol{h}}_j = (\boldsymbol{X} - \overline{\boldsymbol{X}})\hat{\boldsymbol{u}}_j = (\boldsymbol{X} - \boldsymbol{M})\boldsymbol{P}_n\hat{\boldsymbol{u}}_j = (\boldsymbol{X} - \boldsymbol{M})\hat{\boldsymbol{u}}_j,$$

where $\boldsymbol{M} = [\boldsymbol{\mu}, ..., \boldsymbol{\mu}]$. Thus it holds that $\{(n-1)\tilde{\lambda}_j\}^{1/2}\tilde{\boldsymbol{h}}_j^T(\boldsymbol{x}_l - \boldsymbol{\mu}) = \hat{\boldsymbol{u}}_j^T(\boldsymbol{X} - \boldsymbol{M})^T(\boldsymbol{x}_l - \boldsymbol{\mu}) = \hat{u}_{jl}||\boldsymbol{x}_l - \boldsymbol{\mu}||^2 + \sum_{s=1(\neq l)}^n \hat{u}_{js}(\boldsymbol{x}_s - \boldsymbol{\mu})^T(\boldsymbol{x}_l - \boldsymbol{\mu})$, so that $\hat{u}_{jl}||\boldsymbol{x}_l - \boldsymbol{\mu}||^2$ is very biased since $E(||\boldsymbol{x}_l - \boldsymbol{\mu}||^2)/\{(n-1)^{1/2}\lambda_j\} \geq (n-1)^{1/2}\delta/\lambda_j$. Hence, one should not apply the $\hat{\boldsymbol{h}}_j$s or the $\tilde{\boldsymbol{h}}_j$s to the estimation of the inner product.

Here, we consider a bias-reduced estimation of the inner product. Let us write that

$$\hat{\boldsymbol{u}}_{jl} = (\hat{u}_{j1}, ..., \hat{u}_{jl-1}, -\hat{u}_{jl}/(n-1), \hat{u}_{jl+1}, ..., \hat{u}_{jn})^T$$

whose l-th element is $-\hat{u}_{jl}/(n-1)$ for all j, l. Note that $\hat{\boldsymbol{u}}_{jl} = \hat{\boldsymbol{u}}_j - (0, ..., 0, \hat{u}_{jl}n/(n-1), 0, ..., 0)^T$. Let

$$\tilde{\boldsymbol{h}}_{jl} = \{(n-1)\tilde{\lambda}_j\}^{-1/2}(\boldsymbol{X} - \overline{\boldsymbol{X}})\hat{\boldsymbol{u}}_{jl} \tag{3.3}$$

for all j, l. When $\hat{\lambda}_j > 0$, we note that $\{(n-1)\tilde{\lambda}_j\}^{1/2}\tilde{\boldsymbol{h}}_{jl} = (\boldsymbol{X} - \boldsymbol{M})\boldsymbol{P}_n\hat{\boldsymbol{u}}_{jl} = (\boldsymbol{X} - \boldsymbol{M})\hat{\boldsymbol{u}}_{j(l)}$ since $\boldsymbol{1}_n^T\hat{\boldsymbol{u}}_j = \sum_{l=1}^n \hat{u}_{jl} = 0$, where

$$\hat{\boldsymbol{u}}_{j(l)} = (\hat{u}_{j1}, ..., \hat{u}_{jl-1}, 0, \hat{u}_{jl+1}, ..., \hat{u}_{jn})^T + (n-1)^{-1}\hat{u}_{jl}\boldsymbol{1}_{n(l)} \quad \text{for } l = 1, ..., n.$$

Here, $\boldsymbol{1}_{n(l)} = (1, ..., 1, 0, 1, ..., 1)^T$ whose l-th element is 0. Thus it holds that

$$\begin{aligned}\{(n-1)\tilde{\lambda}_j\}^{1/2}\tilde{\boldsymbol{h}}_{jl}^T(\boldsymbol{x}_l - \boldsymbol{\mu}) &= \hat{\boldsymbol{u}}_{j(l)}^T(\boldsymbol{X} - \boldsymbol{M})^T(\boldsymbol{x}_l - \boldsymbol{\mu}) \\ &= \sum_{s=1(\neq l)}^n \{\hat{u}_{js} + (n-1)^{-1}\hat{u}_{jl}\}(\boldsymbol{x}_s - \boldsymbol{\mu})^T(\boldsymbol{x}_l - \boldsymbol{\mu}),\end{aligned}$$

so that the large biased term, $||\boldsymbol{x}_l - \boldsymbol{\mu}||^2$, has vanished. Then, we have the following result.

Proposition 4 (Aoshima and Yata, 2016)**.** *Assume (A-i) and (A-ii). It holds for $j = 1, ..., k$ $(l = 1, ..., n)$ that as $m_0 \to \infty$*

$$\lambda_j^{-1/2}\tilde{\boldsymbol{h}}_{jl}^T(\boldsymbol{x}_l - \boldsymbol{\mu}) = z_{jl} + \hat{u}_{jl} \times O_P\{(n^{1/2}\lambda_j)^{-1}\lambda_1\} + O_P(n^{-1/2}).$$

Note that $[\sum_{l=1}^n\{\hat{u}_{jl}\lambda_1/(n^{1/2}\lambda_j)\}^2/n]^{1/2} = \lambda_1/(\lambda_j n)$. The bias term is small when λ_1/λ_j is not large.

3.2. Test procedure using eigenstructures

Let $\tilde{x}_{ijl} = \tilde{\boldsymbol{h}}_{ijl}^T\boldsymbol{x}_{il}$ for all i, j, l, where $\tilde{\boldsymbol{h}}_{ijl}$s are defined by (3.3). From Propositions 2 and 4, we consider the following test statistic for (1.1):

$$\begin{aligned}\widehat{T}_* =& 2\sum_{i=1}^{2}\frac{\sum_{l<l'}^{n_i}(\boldsymbol{x}_{il}^T\boldsymbol{x}_{il'}-\sum_{j=1}^{k_i}\tilde{x}_{ijl}\tilde{x}_{ijl'})}{n_i(n_i-1)}\\ &-2\frac{\sum_{l=1}^{n_1}\sum_{l'=1}^{n_2}(\boldsymbol{x}_{1l}-\sum_{j=1}^{k_1}\tilde{x}_{1jl}\tilde{\boldsymbol{h}}_{1j})^T(\boldsymbol{x}_{2l'}-\sum_{j=1}^{k_2}\tilde{x}_{2jl'}\tilde{\boldsymbol{h}}_{2j})}{n_1n_2},\end{aligned}$$

where $\tilde{\boldsymbol{h}}_{ij}$s are defined by (3.2). Then, we have the following result.

Theorem 1 (Aoshima and Yata, 2016)**.** *Assume (A-i) and (A-ii). Assume also*

$$\limsup_{m\to\infty}\frac{\Delta_*^2}{K_{1*}}<\infty.$$

Then, it holds that as $m\to\infty$

$$\frac{\widehat{T}_*-\Delta_*}{K_*^{1/2}}\Rightarrow N(0,1)$$

under some regularity conditions.

Let z_c be a constant such that $P\{N(0,1)>z_c\}=c$ for $c\in(0,1)$. We note that $K_{1*}/K_*=1+o(1)$ as $m\to\infty$ under (A-i) and $\limsup_{m\to\infty}\Delta_*^2/K_{1*}<\infty$. Then, for given $\alpha\in(0,1/2)$, we consider testing the hypothesis in (1.1) by

$$\text{rejecting } H_0 \Longleftrightarrow \frac{\widehat{T}_*}{\widehat{K}_{1*}^{1/2}}>z_\alpha, \tag{3.4}$$

where $\widehat{K}_{1*}$ is defined in Section 5.2 of Aoshima and Yata (2016). Let power(Δ_*) denote the power of the test (3.4). Then, we have the following result.

Theorem 2 (Aoshima and Yata, 2016)**.** *Assume (A-i) and (A-ii). Then, the test (3.4) has as $m\to\infty$*

$$\textit{size}=\alpha+o(1)\quad \textit{and}\quad \textit{power}(\Delta_*)-\Phi\bigg(\frac{\Delta_*}{K_*^{1/2}}-z_\alpha\Big(\frac{K_{1*}}{K_*}\Big)^{1/2}\bigg)=o(1)$$

under some regularity conditions, where $\Phi(\cdot)$ denotes the cumulative distribution function of $N(0,1)$.

In general, k_is are unknown in $\widehat{T}_*$. See Section 6.2 in Aoshima and Yata (2016) for estimation of k_is.

Acknowledgements

The research of the first author was partially supported by Grants-in-Aid for Scientific Research (A) and Challenging Exploratory Research, Japan Society for the Promotion of Science (JSPS), under Contract Numbers 15H01678 and 26540010. The research of the second author was partially supported by Grant-in-Aid for Young Scientists (B), JSPS, under Contract Number 26800078.

References

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

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

Aoshima, M. and Yata, K. (2016). Two-sample tests for high-dimension, strongly spiked eigenvalue models. *Statist. Sinica*, revised (arXiv:1602.02491).

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

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.

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.

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

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

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**, 189-199.

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.

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.

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

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.

Yata, K. and 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: aoshima@math.tsukuba.ac.jp

Institute of Mathematics, University of Tsukuba, Ibaraki 305-8571, Japan.

E-mail: yata@math.tsukuba.ac.jp