Statistical inference on high-dimensional covariance structures under the SSE models

Aki Ishii[a], **Yumu Iwana**[b], **Kazuyoshi Yata**[c] **and Makoto Aoshima**[c]

[a] Department of Information Sciences, Tokyo University of Science
[b] Graduate School of Science and Technology, University of Tsukuba
[c] Institute of Mathematics, University of Tsukuba

Key words and phrases: ECDM; HDLSS; Noise-reduction method; SSE model

1 Introduction

One of the characteristics of the high-dimensional data is that the data dimension is much larger than the sample size. We call such data "high-dimension, low-sample-size (HDLSS)" or "large p, small n" data. Here, p is the data dimension and n is the sample size. Recently, Aoshima and Yata [3] created the two disjoint models: the strongly spiked eigenvalue (SSE) model and the non-SSE (NSSE) model. The SSE model is defined by

$$\liminf_{p\to\infty} \frac{\lambda_{\max}(\boldsymbol{\Sigma})}{\sqrt{\mathrm{tr}(\boldsymbol{\Sigma}^2)}} > 0,$$

where $\lambda_{\max}(\boldsymbol{\Sigma})$ is the largest eigenvalue of the covariance matrix, $\boldsymbol{\Sigma}$. On the other hand, the NSSE model is defined by

$$\frac{\lambda_{\max}(\boldsymbol{\Sigma})}{\sqrt{\mathrm{tr}(\boldsymbol{\Sigma}^2)}} \to 0, \quad p \to \infty.$$

In this talk, we focus on the SSE model and construct a new procedure for the correlation test. Suppose we take samples, $\boldsymbol{x}_j,\ j = 1,\dots,n$, of size $n\ (\geq 4)$, which are independent and identically distributed (i.i.d.) as a p-variate distribution. Here, we consider situations where the data dimension p is very high compared to the sample size n. Let $\boldsymbol{x}_j = (\boldsymbol{x}_{1j}^\top, \boldsymbol{x}_{2j}^\top)^\top$ and assume $\boldsymbol{x}_{ij} \in \mathbf{R}^{p_i}$, $i = 1, 2$, with $p_1 \in [1, p-1]$ and $p_2 = p - p_1$. We also assume that $\boldsymbol{x}_j$ has an unknown mean vector, $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^\top, \boldsymbol{\mu}_2^\top)^\top$, and unknown covariance matrix,

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{\Sigma}_* \\ \boldsymbol{\Sigma}_*^\top & \boldsymbol{\Sigma}_2 \end{pmatrix}\ (\geq \boldsymbol{O}),$$

that is, $\mathrm{E}(\boldsymbol{x}_{ij}) = \boldsymbol{\mu}_i$, $\mathrm{Var}(\boldsymbol{x}_{ij}) = \boldsymbol{\Sigma}_i$, $i = 1, 2$, and $\mathrm{Cov}(\boldsymbol{x}_{1j}, \boldsymbol{x}_{2j}) = \mathrm{E}(\boldsymbol{x}_{1j}\boldsymbol{x}_{2j}^\top) - \boldsymbol{\mu}_1\boldsymbol{\mu}_2^\top = \boldsymbol{\Sigma}_*$. Let σ_{ij} be the j-th diagonal element of $\boldsymbol{\Sigma}_i$ for $i = 1, 2;\ j = 1,\dots,p_i$, and assume $\sigma_{ij} > 0$ for all i, j. We denote the correlation coefficient matrix between $\boldsymbol{x}_{1j}$ and $\boldsymbol{x}_{2j}$ by $\mathrm{Corr}(\boldsymbol{x}_{1j}, \boldsymbol{x}_{2j}) = \boldsymbol{P}$, where $\boldsymbol{P} = \mathrm{diag}(\sigma_{11},\dots,\sigma_{1p_1})^{-1/2}\boldsymbol{\Sigma}_*\mathrm{diag}(\sigma_{21},\dots,\sigma_{2p_2})^{-1/2}$. Here, $\mathrm{diag}(\sigma_{i1},\dots,\sigma_{ip_i})$ denotes the diagonal matrix of elements, $\sigma_{i1},\dots,\sigma_{ip_i}$. Then, we consider testing the following hypotheses:

$$H_0 : \boldsymbol{P} = \boldsymbol{O} \quad \text{vs.} \quad H_1 : \boldsymbol{P} \neq \boldsymbol{O} \tag{1}$$

for high-dimensional settings. The test of the correlation coefficient matrix is a very important tool of pathway analysis or graphical modeling for high-dimensional data.

Aoshima and Yata [1] gave a test statistic for the test of correlation coefficients and Yata and Aoshima [7, 8] improved the test statistic by using the *extended cross-data-matrix (ECDM) methodology.* They gave asymptotic normality of the test statistic under the following model:

(A-i) $\min\left\{\frac{\lambda_{\max}(\boldsymbol{\Sigma}_1)}{\sqrt{\operatorname{tr}(\boldsymbol{\Sigma}_1^2)}}, \frac{\lambda_{\max}(\boldsymbol{\Sigma}_2)}{\sqrt{\operatorname{tr}(\boldsymbol{\Sigma}_2^2)}}\right\} \to 0, \quad p \to \infty.$

Note that (A-i) is one of the NSSE models.

2 Correlation test under the NSSE model

We consider the eigenvalue decomposition of $\boldsymbol{\Sigma}$ by $\boldsymbol{\Sigma} = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}^\top$, where $\boldsymbol{\Lambda}$ =diag$(\lambda_1, \dots, \lambda_p)$ having eigenvalues, $\lambda_1 \ge \cdots \ge \lambda_p \ge 0$, and $\boldsymbol{H}$ is an orthogonal matrix of the corresponding eigenvectors. Let $\boldsymbol{x}_j = \boldsymbol{H}\boldsymbol{\Lambda}^{1/2}\boldsymbol{z}_j + \boldsymbol{\mu}$, $j = 1, \dots, n$, where $\mathrm{E}(\boldsymbol{z}_j) = \boldsymbol{0}$ and $\mathrm{Var}(\boldsymbol{z}_j) = \boldsymbol{I}_p$. Here, $\boldsymbol{I}_p$ denotes the identity matrix of dimension p. Note that if $\boldsymbol{x}_j$ is Gaussian, the elements of $\boldsymbol{z}_j$ are i.i.d. as the standard normal distribution, $\mathcal{N}(0, 1)$. For $\boldsymbol{x}_j$, we consider the following model:

$$\boldsymbol{x}_j = \boldsymbol{\Gamma}\boldsymbol{w}_j + \boldsymbol{\mu}, \ j = 1, \dots, n, \tag{2}$$

where $\boldsymbol{\Gamma}$ is a $p \times q$ matrix for some $q > 0$ such that $\boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top = \boldsymbol{\Sigma}$, and $\boldsymbol{w}_j = (w_{1j}, \dots, w_{qj})^\top$, $j = 1, \dots, n$, are i.i.d. random vectors having $\mathrm{E}(\boldsymbol{w}_j) = \boldsymbol{0}$ and $\mathrm{Var}(\boldsymbol{w}_j) = \boldsymbol{I}_q$. Let $\boldsymbol{\Gamma} = (\boldsymbol{\Gamma}_1^\top, \boldsymbol{\Gamma}_2^\top)^\top$, where $\boldsymbol{\Gamma}_i = (\boldsymbol{\gamma}_{i1}, \dots, \boldsymbol{\gamma}_{iq})$ with $\boldsymbol{\gamma}_{ij}$s $\in \mathbf{R}^{p_i}$, $i = 1, 2$. Then, we have that $\boldsymbol{x}_{ij} = \boldsymbol{\Gamma}_i\boldsymbol{w}_j + \boldsymbol{\mu}_i$. Note that $\boldsymbol{\Sigma}_* = \boldsymbol{\Gamma}_1\boldsymbol{\Gamma}_2^\top = \sum_{r=1}^q \boldsymbol{\gamma}_{1r}\boldsymbol{\gamma}_{2r}^\top$. Also, note that (2) includes the case that $\boldsymbol{\Gamma} = \boldsymbol{H}\boldsymbol{\Lambda}^{1/2}$ and $\boldsymbol{w}_j = \boldsymbol{z}_j$. Let $\mathrm{Var}(w_{rj}^2) = M_r$, $r = 1, \dots, q$. We assume that $\limsup_{p\to\infty} M_r < \infty$ for all r. Similar to Aoshima and Yata [2] and Bai and Saranadasa [4], we assume

(A-ii) $\mathrm{E}(w_{rj}^2 w_{sj}^2) = \mathrm{E}(w_{rj}^2)\mathrm{E}(w_{sj}^2) = 1$ and $\mathrm{E}(w_{rj}w_{sj}w_{tj}w_{uj}) = 0$ for all $r \ne s, t, u$.

We also consider the following assumption instead of (A-ii) as necessary:

(A-iii) $\mathrm{E}(w_{r_1j}^{\alpha_1} w_{r_2j}^{\alpha_2} \cdots w_{r_vj}^{\alpha_v}) = \mathrm{E}(w_{r_1j}^{\alpha_1})\mathrm{E}(w_{r_2j}^{\alpha_2}) \cdots \mathrm{E}(w_{r_vj}^{\alpha_v})$ for all $r_1 \ne r_2 \ne \cdots \ne r_v \in [1, q]$ and $\alpha_i \in [1, 4]$, $i = 1, \dots, v$, where $v \le 8$ and $\sum_{i=1}^v \alpha_i \le 8$.

See Chen and Qin [5] and Zhong and Chen [9] about (A-iii).

Remark 1. *The assumption (A-iii) is naturally satisfied when $\boldsymbol{x}_j$ is Gaussian because the elements of $\boldsymbol{z}_j$ are independent and $M_r = 2$ for all r.*

Let $\Delta = \operatorname{tr}(\boldsymbol{\Sigma}_*\boldsymbol{\Sigma}_*^\top)(= \|\boldsymbol{\Sigma}_*\|_F^2)$, where $\|\cdot\|_F$ is the Frobenius norm. We introduce an unbiased estimator of Δ by the ECDM methodology. We define $n_{(1)} = \lceil n/2 \rceil$ and $n_{(2)} = n - n_{(1)}$, where $\lceil x \rceil$ denotes the smallest integer $\ge x$. Let

$$\boldsymbol{V}_{n(1)(k)} = \begin{cases} \{\lfloor k/2 \rfloor - n_{(1)} + 1, \dots, \lfloor k/2 \rfloor\} & \text{if } \ \lfloor k/2 \rfloor \ge n_{(1)}, \\ \{1, \dots, \lfloor k/2 \rfloor\} \cup \{\lfloor k/2 \rfloor + n_{(2)} + 1, \dots, n\} & \text{otherwise;} \end{cases}$$

$$\boldsymbol{V}_{n(2)(k)} = \begin{cases} \{\lfloor k/2 \rfloor + 1, \dots, \lfloor k/2 \rfloor + n_{(2)}\} & \text{if } \ \lfloor k/2 \rfloor \le n_{(1)}, \\ \{1, \dots, \lfloor k/2 \rfloor - n_{(1)}\} \cup \{\lfloor k/2 \rfloor + 1, \dots, n\} & \text{otherwise} \end{cases}$$

for $k = 3, \dots, 2n-1$, where $\lfloor x \rfloor$ denotes the largest integer $\le x$. Also, let $\#\boldsymbol{A}$ denote the number of elements in a set $\boldsymbol{A}$. Note that $\#\boldsymbol{V}_{n(l)(k)} = n_{(l)}$, $l = 1, 2$, $\boldsymbol{V}_{n(1)(k)} \cap \boldsymbol{V}_{n(2)(k)} = \boldsymbol{\emptyset}$ and $\boldsymbol{V}_{n(1)(k)} \cup \boldsymbol{V}_{n(2)(k)} = \{1, \dots, n\}$ for $k = 3, \dots, 2n-1$. It should be noted that

$$i \in \boldsymbol{V}_{n(1)(i+j)} \quad \text{and} \quad j \in \boldsymbol{V}_{n(2)(i+j)} \quad \text{for } i < j \ (\le n). \tag{3}$$

Let

$$\overline{\boldsymbol{x}}_{l(1)(k)} = n_{(1)}^{-1} \sum_{j \in \boldsymbol{V}_{n(1)(k)}} \boldsymbol{x}_{lj} \quad \text{and} \quad \overline{\boldsymbol{x}}_{l(2)(k)} = n_{(2)}^{-1} \sum_{j \in \boldsymbol{V}_{n(2)(k)}} \boldsymbol{x}_{lj}, \quad l = 1, 2$$

for $k = 3, \dots, 2n-1$. We consider the following quantity:

$$\widehat{\Delta}_{ij} = (\boldsymbol{x}_{1i} - \overline{\boldsymbol{x}}_{1(1)(i+j)})^\top (\boldsymbol{x}_{1j} - \overline{\boldsymbol{x}}_{1(2)(i+j)}) (\boldsymbol{x}_{2i} - \overline{\boldsymbol{x}}_{2(1)(i+j)})^\top (\boldsymbol{x}_{2j} - \overline{\boldsymbol{x}}_{2(2)(i+j)})$$

for all $i < j \ (\le n)$. Then, from (3), it holds that

(i) $\boldsymbol{x}_{li} - \overline{\boldsymbol{x}}_{l(1)(i+j)}$ and $\boldsymbol{x}_{lj} - \overline{\boldsymbol{x}}_{l(2)(i+j)}$ are independent for $l = 1, 2$;

(ii) $\mathrm{E}(\widehat{\Delta}_{ij}) = \Delta\{(n_{(1)} - 1)(n_{(2)} - 1)\}/(n_{(1)} n_{(2)})$

for all $i < j \ (\le n)$. Let $u_n = n_{(1)} n_{(2)} \{(n_{(1)} - 1)(n_{(2)} - 1)\}^{-1}$. Yata and Aoshima [8] proposed an unbiased estimator of Δ by

$$\widehat{T}_n = \frac{2u_n}{n(n-1)} \sum_{i<j}^{n} \widehat{\Delta}_{ij}.$$

Note that $\mathrm{E}(\widehat{T}_n) = \Delta$.

Let $m = \min\{p, n\}$ and $\delta = \sqrt{2\mathrm{tr}(\boldsymbol{\Sigma}_1^2)\mathrm{tr}(\boldsymbol{\Sigma}_2^2)}/n$. Yata and Aoshima [8] gave the following results.

Theorem 2.1 (Yata and Aoshima [8])**.** *Assume (A-i) and (A-ii). Under H_0 in (1), it holds that as $m \to \infty$,*

$$\mathrm{Var}(\widehat{T}_n) = \delta^2\{1 + o(1)\}.$$

Theorem 2.2 (Yata and Aoshima [8])**.** *Assume (A-i) and (A-iii). Under H_0 in (1), it holds that as $m \to \infty$,*

$$\frac{\widehat{T}_n}{\delta} \Rightarrow \mathcal{N}(0, 1),$$

where "$\Rightarrow$" denotes the convergence in distribution.

Yata and Aoshima [8] gave an estimator of $\mathrm{tr}(\boldsymbol{\Sigma}_i^2)$, $i = 1, 2$, by

$$W_{in} = \frac{2u_n}{n(n-1)} \sum_{r<s}^{n} \left\{ (\boldsymbol{x}_{ir} - \overline{\boldsymbol{x}}_{i(1)(r+s)})^\top (\boldsymbol{x}_{is} - \overline{\boldsymbol{x}}_{i(2)(r+s)}) \right\}^2.$$

Note that $\mathrm{E}(W_{in}) = \mathrm{tr}(\boldsymbol{\Sigma}_i^2)$. Let $\alpha \in (0, 1/2)$ be a prespecified constant. Also, let z_α be a constant such that $P\{\mathcal{N}(0,1) > z_\alpha\} = \alpha$. Yata and Aoshima [8] proposed testing (1) by

$$\text{rejecting } H_0 \Longleftrightarrow \frac{\widehat{T}_n}{\hat{\delta}} > z_\alpha, \tag{4}$$

where $\widehat{\delta} = n^{-1}(2W_{1n}W_{2n})^{1/2}$. Then, the test by (4) has

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

as $m \to \infty$ under the NSSE model (A-i) and (A-iii).

3 Correlation test under the SSE model

In this section, we assume p_1 is fixed. We also assume the following condition:

(C-i) $\dfrac{\lambda_{\max}(\boldsymbol{\Sigma}_2)}{\sqrt{\mathrm{tr}(\boldsymbol{\Sigma}_2^2)}} \to 1, \quad p_2 \to \infty.$

The model (C-i) is one of the SSE models and is called "uni-SSE model" in Ishii, Yata and Aoshima [6]. Under (C-i), we have the following result.

Theorem 3.1. *Assume (C-i) and some regularity conditions. Then, it holds that as* $m \to \infty$

$$\frac{n(\widehat{T}_n - \Delta)}{\lambda_{\max}(\boldsymbol{\Sigma}_2)} + \mathrm{tr}(\boldsymbol{\Sigma}_1) \Rightarrow \sum_{s=1}^{p_1} \lambda_{1s}\chi_{1s}^2,$$

where λ_{1s} *is the s-th eigenvalue of* $\boldsymbol{\Sigma}_1$, χ_{1s}^2 *stands for a chi-square random variable with* 1 *degree of freedom and* χ_{1s}^2, $s = 1, ..., p_1$ *are mutually independent.*

Acknowledgements

Research of the first author was partially supported by Grant-in-Aid for Young Scientists, Japan Society for the Promotion of Science (JSPS), under Contract Number 24K20749. Research of the third author was partially supported by Grant-in-Aid for Scientific Research (C), JSPS, under Contract Number 22K03412. Research of the fourth author was partially supported by Grants-in-Aid for Scientific Research (A) and Challenging Research (Exploratory), JSPS, under Contract Numbers 20H00576 and 22K19769.

References

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

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

[3] M. Aoshima, K. Yata, Two-sample tests for high-dimension, strongly spiked eigenvalue models. Statist. Sinica 28 (2018) 43–62.

[4] Z. Bai, H. Saranadasa, Effect of high dimension: by an example of a two sample problem, Statist. Sinica 6 (1996) 311–329.

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

[6] A. Ishii, K. Yata, M. Aoshima, Hypothesis tests for high-dimensional covariance structures. Ann. Inst. Statist. Math. 73 (2021) 599–622.

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

[8] Yata, K., Aoshima, M. (2016). High-dimensional inference on covariance structures via the extended cross-data-matrix methodology. J. Multivariate Anal. 151 (2016) 151–166.

[9] P.-S. Zhong, S.X. Chen, Tests for high-dimensional regression coefficients with factorial designs, J. Amer. Statist. Assoc. 106 (2011) 260–274.