

Testing homogeneity of high-dimensional covariance matrices under a low-dimensional factor structure

Takahiro Nishiyama^a, Masashi Hyodo^b, Shoma Yamashita^c, Nobumichi Shutoh^d

^a Faculty of Science, Tokyo University of Science

^b Faculty of Economics, Kanagawa University

^c Graduate School of Science, Kobe University

^d Laboratory of Medical Data Sciences, Kobe Pharmaceutical University
/ Graduate School of Science, Kobe University

1. Introduction

Suppose that a p -dimensional random sample of size n_g is represented by $\mathbf{x}_{g1}, \dots, \mathbf{x}_{gn_g}$, where the observations are independently drawn from population Π_g ($g = 1, \dots, k$) with mean vector $\boldsymbol{\mu}_g = E(\mathbf{x}_{gi})$ and covariance matrix $\boldsymbol{\Sigma}_g = \text{cov}(\mathbf{x}_{gi})$. We consider the k -sample hypothesis

$$\mathcal{H} : \boldsymbol{\Sigma}_1 = \dots = \boldsymbol{\Sigma}_k \quad \text{versus} \quad \mathcal{A} : \neg \mathcal{H}. \quad (1)$$

Note that hypothesis (1) is equivalent to

$$\mathcal{H} : \|\boldsymbol{\Delta}_{gh}\|_F^2 = 0 \quad \text{for all } g < h,$$

where $\boldsymbol{\Delta}_{gh} = \boldsymbol{\Sigma}_g - \boldsymbol{\Sigma}_h$ and $\|\cdot\|_F$ denotes the Frobenius norm. There are several approaches for testing the equality of high-dimensional covariance matrices. Among them, Li and Chen [5] proposed a pioneering two-sample test based on the asymptotic normality of an unbiased estimator of $\|\boldsymbol{\Delta}_{gh}\|_F^2$. Their asymptotic theory relies on the weak dependence condition

$$\text{tr}(\boldsymbol{\Sigma}_g \boldsymbol{\Sigma}_h \boldsymbol{\Sigma}_{g'} \boldsymbol{\Sigma}_{h'}) = o\{\text{tr}(\boldsymbol{\Sigma}_g \boldsymbol{\Sigma}_h) \text{tr}(\boldsymbol{\Sigma}_{g'} \boldsymbol{\Sigma}_{h'})\}, \quad g, h, g', h' \in \{1, \dots, k\}, \quad p \rightarrow \infty, \quad (2)$$

which is also known as the boundary condition on the maximum eigenvalue. Several extensions of Li and Chen's methodology have subsequently been proposed. For the general k -sample problem, Zhang et al. [7] extended their approach to testing the equality of covariance matrices under essentially the same weak dependence condition.

To overcome these limitations, Aoshima and Yata [2] introduced the strongly spiked eigenvalue (SSE) model, under which Ishii et al. [4] proposed a two-sample covariance matrix test. An alternative strategy is to assume a low-dimensional latent factor model. For high-dimensional covariance matrix testing, Hyodo et al. [3] developed a two-sample test for the equality of covariance matrices under a low-dimensional latent factor structure. These procedures removes the restrictive condition (2) and remains valid even when the covariance matrices possess strong latent dependence. In this study, we extend the two-sample methodology of Hyodo et al. [3] to the general k -sample problem and develop a new test procerure for (1).

2. Main results

2.1. Data generation model and asymptotic framework

The data generation model of population Π_g is assumed to be a latent factor model, which is expressed as

$$\mathbf{x}_g = \boldsymbol{\mu}_g + \mathbf{B}\mathbf{f}_g + \boldsymbol{\epsilon}_g. \quad (3)$$

Here, $\boldsymbol{\mu}_g \in \mathbb{R}^p$ is the population mean vector, and the elements of the $p \times d$ ($d < p$) non-random matrix $\mathbf{B} = (\mathbf{b}_1, \dots, \mathbf{b}_p)^\top$ are referred to as factor loadings. $\mathbf{f}_g \in \mathbb{R}^d$ and $\boldsymbol{\epsilon}_g \in \mathbb{R}^p$ are random vectors for common factors and specific factors, respectively. We assume that $\mathbf{f}_g$ and $\boldsymbol{\epsilon}_g$ are independent of each other and that their mean vectors are $E(\mathbf{f}_g) = \mathbf{0}$ and $E(\boldsymbol{\epsilon}_g) = \mathbf{0}$, respectively. Also, we assume that $\text{cov}(\mathbf{f}_g) = \boldsymbol{\Phi}_g$ is a positive definite matrix and $\text{cov}(\boldsymbol{\epsilon}_g) = \boldsymbol{\Psi}_g = \text{diag}(\psi_{g1}, \dots, \psi_{gp})$. Let $\mathbf{z}_g = (z_{g1}, \dots, z_{gd})^\top = \boldsymbol{\Phi}_g^{-1/2} \mathbf{f}_g$ and $\mathbf{e}_g = (e_{g1}, \dots, e_{gp})^\top = \boldsymbol{\Psi}_g^{-1/2} \boldsymbol{\epsilon}_g$. Furthermore, we assume that z_{gi} are iid with $E(z_{gi}) = 0$, $E(z_{gi}^2) = 1$, and $E(z_{gi}^4) = \kappa_z < \infty$, and e_{gj} are iid with $E(e_{gj}) = 0$, $E(e_{gj}^2) = 1$, and $E(e_{gj}^4) = \kappa_e < \infty$ for $g \in \{1, \dots, k\}$, $i \in \{1, \dots, d\}$, and $j \in \{1, \dots, p\}$. These assumptions along with equation (3) imply that

$$E(\mathbf{x}_g) = \boldsymbol{\mu}_g, \quad \text{cov}(\mathbf{x}_g) = \mathbf{B}\boldsymbol{\Phi}_g\mathbf{B}^\top + \boldsymbol{\Psi}_g := \boldsymbol{\Sigma}_g. \quad (4)$$

For the asymptotic evaluation, we impose the following regularity conditions:

- (A1) Let $n = n_1 + \dots + n_k$. Suppose that $p = p(n_1, \dots, n_k)$ is a function of $n_1, \dots, n_k$ such that $p \rightarrow \infty$ as $n_1, \dots, n_k \rightarrow \infty$, and that p and n are of the same order, i.e.,

$$0 < \liminf \frac{p}{n} \leq \limsup \frac{p}{n} < \infty.$$

Moreover, assume that

$$\frac{n_g}{n} \rightarrow \gamma_g \in (0, 1), \quad g \in \{1, \dots, k\}.$$

The number of factors d and the number of groups k are fixed.

- (A2) Let $\psi_{g\max} = \max_{1 \leq j \leq p} \psi_{gj}$, $g \in \{1, \dots, k\}$. Assume that $\limsup_{p \rightarrow \infty} \psi_{g\max} < \infty$, $g \in \{1, \dots, k\}$.

There exist positive definite symmetric matrices $\mathbf{B}_*$ and $\boldsymbol{\Phi}_{g*}$, $g \in \{1, \dots, k\}$, such that

$$\left\| \frac{1}{p} \mathbf{B}^\top \mathbf{B} - \mathbf{B}_* \right\|_F = o(1), \quad \|\boldsymbol{\Phi}_g - \boldsymbol{\Phi}_{g*}\|_F = o(1), \quad g \in \{1, \dots, k\}.$$

- (A3) $\mathbf{f}_{gi} \sim \mathcal{N}_d(\mathbf{0}, \boldsymbol{\Phi}_g)$.

2.2. Test procedure

We first define the following random variable T :

$$T = \sum_{g < h}^k \frac{n_g n_h}{n} \left\{ \frac{\widehat{\|\Delta_{gh}\|_F^2}}{p^2} + \sum_{\ell \in \{g, h\}} \frac{\text{tr}(\mathbf{A}_{\ell\ell}^2) + \{\text{tr}(\mathbf{A}_{\ell\ell})\}^2}{n_\ell} \right\}.$$

Here,

$$\widehat{\|\Delta_{gh}\|_F^2} = \sum_{\ell \in \{g, h\}} \widehat{\|\Sigma_\ell\|_F^2} - 2\text{tr}(\mathbf{S}_g \mathbf{S}_h),$$

where

$$\begin{aligned} \widehat{\|\Sigma_\ell\|_F^2} &= \frac{n_\ell - 1}{n_\ell(n_\ell - 2)(n_\ell - 3)} [(n_\ell - 1)(n_\ell - 2)\text{tr}(\mathbf{S}_\ell^2) + \{\text{tr}(\mathbf{S}_\ell)\}^2 - n_\ell K_\ell], \\ \mathbf{S}_\ell &= \frac{1}{n_\ell - 1} \sum_{i=1}^{n_\ell} (\mathbf{x}_{\ell i} - \bar{\mathbf{x}}_\ell)(\mathbf{x}_{\ell i} - \bar{\mathbf{x}}_\ell)^\top, \quad K_\ell = \frac{1}{n_\ell - 1} \sum_{i=1}^{n_\ell} \|\mathbf{x}_{\ell i} - \bar{\mathbf{x}}_\ell\|^4. \end{aligned}$$

Theorem 1. Under $\mathcal{H}$ and (A1)–(A3), T is asymptotically distributed as

$$WC = 2 \left\{ \sum_{i=1}^d \lambda_i(\mathbf{A})^2 \chi_{ii}^2 + \sum_{j=2}^d \sum_{i=1}^{j-1} \lambda_i(\mathbf{A}) \lambda_j(\mathbf{A}) \chi_{ij}^2 \right\},$$

where χ_{ij}^2 's are mutually independent, chi-square distributed random variables with $k - 1$ degree of freedom, $\lambda_i(\cdot)$ is the i -th largest eigenvalue, and $\mathbf{A} = \mathbf{\Phi}_*^{1/2} \mathbf{B}_* \mathbf{\Phi}_*^{1/2}$. Here, $\mathbf{\Phi}_* = \gamma_1 \mathbf{\Phi}_{1*} + \cdots + \gamma_k \mathbf{\Phi}_{k*}$.

To estimate the number of factors d , we explore the idea of Wang [6] and Ahn and Horenstein [1]. Here, we focus on the criteria function, which was originally proposed by Ahn and Horenstein [1]:

$$ER(i) = \frac{\lambda_i(\mathbf{S})}{\lambda_{i+1}(\mathbf{S})},$$

where $ER(\cdot)$ refers to the eigenvalue ratio and $\mathbf{S} = \{1/(n - k)\} \sum_{g=1}^k (n_g - 1) \mathbf{S}_g$. The estimator of d is i such that maximizes $ER(i)$. That is,

$$\hat{d} = \arg \max_{1 \leq i \leq i_{\max}} ER(i), \quad (5)$$

where $i_{\max}$ is the prespecified upper bound of i such that $i_{\max} \in (d, \lfloor c \min(p, n) \rfloor - d - 1]$ for some $c \in (0, 1]$. We estimate the unknown parameters $\text{tr}(\mathbf{A}_{gg})$ and $\text{tr}(\mathbf{A}_{gg}^2)$ in T as

$$\widehat{\text{tr}(\mathbf{A}_{gg})} = \frac{\sum_{i=1}^{\hat{d}} \lambda_i(\mathbf{S}_g)}{p} \quad \text{and} \quad \widehat{\text{tr}(\mathbf{A}_{gg}^2)} = \frac{\sum_{i=1}^{\hat{d}} \{\lambda_i(\mathbf{S}_g)\}^2}{p^2},$$

respectively. Using these estimators, the proposed test statistic is defined as

$$\hat{T} = \sum_{g < h}^k \frac{n_g n_h}{n} \left\{ \frac{\widehat{\|\Delta_{gh}\|_F^2}}{p^2} + \sum_{\ell \in \{g, h\}} \frac{\widehat{\text{tr}(\mathbf{A}_{\ell\ell}^2)} + \{\widehat{\text{tr}(\mathbf{A}_{\ell\ell})}\}^2}{n_\ell} \right\}$$

Theorem 2. Under $\mathcal{H}$ and (A1)–(A3), $\hat{T}$ is asymptotically distributed as WC .

Based on Theorem 2, we propose the following approximate significance test for $\mathcal{H}$. The procedure consists of the following four steps:

1. For each $g \in \{1, \dots, k\}$, draw n_g observations from population G_g and calculate $\hat{d}$, $\hat{T}$, $\widehat{\lambda_i(\mathbf{A})} = p^{-1} \lambda_i(\mathbf{S})$.
2. Draw a sample of $\hat{d}(\hat{d} + 1)/2$ independently and χ^2 -distributed random variables to obtain

$$\widehat{WC} = 2 \left\{ \sum_{i=1}^{\hat{d}} \widehat{\lambda_i(\mathbf{A})}^2 \chi_{ii}^2 + \sum_{j=2}^{\hat{d}} \sum_{i=1}^{j-1} \widehat{\lambda_i(\mathbf{A})} \widehat{\lambda_j(\mathbf{A})} \chi_{ij}^2 \right\}.$$

3. Repeat step 2 until a Monte Carlo estimate of the distribution for the random variable $\widehat{WC}$ and its $(1 - \alpha)$ -quantile $\hat{t}_\alpha$ are obtained.
4. Conduct the test with nominal significance level α as follows:

$$\text{Reject } \mathcal{H} \stackrel{\text{def}}{\iff} \hat{T} > \hat{t}_\alpha. \quad (6)$$

2.3. Aspects of power

Theorem 3. Under the fixed alternative $\sum_{g < h} \|\Delta_{gh}\|_F^2 / p^2 \asymp 1$ and (A1)–(A3), the asymptotic power function is equal to 1.

From Theorem 3, the asymptotic power is equal to 1 under every fixed alternative. This is the standard notion of consistency for a statistical test, namely, that its asymptotic power converges to 1 under every fixed alternative. Therefore, asymptotic power under fixed alternatives is not sufficiently informative for distinguishing between competing tests. To overcome this limitation, we consider the following local alternative.

$\mathcal{A}_L$: Let δ be a constant greater than 0. For $g, h \in \{1, \dots, k\}$, $g < h$, there exists a $d \times d$ matrix Θ_{gh} such that the following condition is met:

$$\|n^\delta(\Phi_g - \Phi_h) - \Theta_{gh}\|_F = o(1) \quad (n, p \rightarrow \infty),$$

where $\Theta_{gh} \neq \mathbf{0}$ for at least one pair (g, h) . In addition, for $g, h \in \{1, \dots, k\}$, $g < h$, $\|\Psi_g - \Psi_h\|_F = o(p^{1/2})$ ($n, p \rightarrow \infty$) is met.

Theorem 4. Under (A1)–(A3) and $\mathcal{A}_L$ with $\delta \geq 1/2$, $\hat{T}$ is asymptotically distributed as

$$\sum_{g < h} (\sqrt{\gamma_h} \mathbf{w}_{g*} - \sqrt{\gamma_g} \mathbf{w}_{h*} + \delta_{gh})^\top (\mathbf{B}_* \otimes \mathbf{B}_*) (\sqrt{\gamma_h} \mathbf{w}_{g*} - \sqrt{\gamma_g} \mathbf{w}_{h*} + \delta_{gh}), \quad (7)$$

where $\mathbf{w}_{g*}$ s are mutually independent and $\mathbf{w}_{g*} \sim \mathcal{N}_{d^2}(\mathbf{0}, \mathbf{C}_g)$. Here,

$$\delta_{gh} = n^{1/2-\delta} \sqrt{\gamma_g \gamma_h} \text{vec}(\Theta_{gh}), \quad \mathbf{C}_g = (\mathbf{I}_{d^2} + \mathbf{K}_{d^2})(\Phi_{g*} \otimes \Phi_{g*}).$$

Applying this theorem, we obtain the following corollary of the asymptotic power under local alternative $\mathcal{A}_L$.

Corollary 1. *Under (A1)–(A3) and $\mathcal{A}_L$,*

$$\text{power} \rightarrow \begin{cases} \alpha, & \text{under } \delta > 1/2, \\ 1 - G(t_\alpha), & \text{under } \delta = 1/2, \\ 1, & \text{under } 0 < \delta < 1/2. \end{cases}$$

Here, t_α denotes the $(1 - \alpha)$ -quantile of the weighted chi-square distribution WC in Theorem 1, and $G(\cdot)$ is the cdf of (7).

3. Numerical studies

We examine the size and power of test (6) in a finite sample and dimension by Monte Carlo simulations.

References

- [1] Ahn, S. C. and Horenstein, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica* **81**, 1203–1227.
- [2] Aoshima, M. and Yata, K. (2018). Two-Sample tests for high-dimension, strongly spiked eigenvalue models. *Statistica Sinica* **28**, 43–62.
- [3] Hyodo, M., Nishiyama, T., Watanabe, H., Nakagawa, T., Tahata, K. (2025). Equality tests of covariance matrices under a low-dimensional factor structure. *J. Multivariate Anal.* **206**, 105397.
- [4] Ishii, A., Yata, K., Aoshima, M. (2019). Equality tests of high-dimensional covariance matrices under the strongly spiked eigenvalue model. *J. Stat. Plan. Inference* **202**, 99–111.
- [5] Li, J., Chen, S. X. (2012). Two sample tests for high-dimensional covariance matrices. *Ann. Statist.* **40**, 908–940.
- [6] Wang, H. (2012). Factor profiled sure independence screening. *Biometrika* **99**, 15–28.
- [7] Zhang, C., Bai, Z., Hu, J., Wang, C. (2018). Multi-sample test for high-dimensional covariance matrices. *Comm. Statist. Theory Methods* **47**, 3161–3177.