SUT Journal of Mathematics
Vol. 55, No. 2 (2019), 139–154

Test for equality of generalized variance in high-dimensional and large sample settings

Takatoshi Sugiyama, Masashi Hyodo, Hiroki Watanabe, Shin-ichi Tsukada and Takashi Seo

(Received June 27, 2019; Revised September 27, 2019)

Abstract. Generalized variance (GV), proposed by Wilks [8], is a one-dimensional measure of multidimensional scatter. It plays an important role in both theoretical and applied research on analyzing big data. This article examines the problem of testing equality of generalized variances of k multivariate normal populations in high-dimensional and large sample settings. The conventional likelihood-ratio test statistic reveals a serious bias as dimensions increase. We present a new test statistic that eliminates this bias, and propose an asymptotic approximation-based test. The likelihood-ratio test statistic can be interpreted as an estimator of criteria related to Jensen's inequality. Our test statistic is obtained by appropriately estimating this criteria in high-dimensional and large sample settings. In addition, our proposed test is valid not only in high dimensional settings but also in large sample settings. We also obtain the asymptotic non-null distribution of the proposed test in high-dimensional and large sample settings. Finally, we study the finite sample and dimension behavior of this test through Monte Carlo simulations.

AMS 2010 *Mathematics Subject Classification.* 62H12, 62E30.

Key words and phrases. Big data analysis, generalized variance, asymptotic approximation-based test, high-dimensional data.

§1. Introduction

For $i \in \{1, 2, \ldots, k\}$ $(k \geq 2)$, let $\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iN_i}$ be p-dimensional random vectors from the i-th multivariate normal population. We denote the i-th population mean vector by $\boldsymbol{\mu}_i$, the i-th population covariance matrix by $\boldsymbol{\Sigma}_i$, and the i-th multivariate normal population by $\mathcal{N}_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, respectively. Let $\mathbf{x}_{ij}$ for $j \in \{1, 2, \ldots, N_i\}$, $i \in \{1, 2, \ldots, k\}$ be distributed as $\mathcal{N}_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, and the random vectors

$$\mathbf{x}_{11}, \mathbf{x}_{12}, \ldots, \mathbf{x}_{1N_1}, \mathbf{x}_{21}, \mathbf{x}_{22}, \ldots, \mathbf{x}_{2N_2}, \ldots, \mathbf{x}_{k1}, \mathbf{x}_{k2}, \ldots, \mathbf{x}_{kN_k}$$

are mutually independent. Many studies examine inference related to covariance matrices in the large sample settings, see, for example, Fujikoshi [2], Nagao [3], Nagao and Sugiura [6, 7].

As data collection technology evolves, high-dimensional data are becoming increasingly available. Suitable improvements in classical multivariate analysis are necessary to tackle high-dimensional data. In this article, we focus on the problem of testing equality of generalized variances of k multivariate normal populations in high-dimensional and large sample settings. Generalized variance (GV), proposed by Wilks [8], is a one-dimensional measure of multidimensional scatter. GVs are very useful for big data variability. Thus, it plays an important role in both theoretical and applied research. Our primary interest is to test the following hypothesis:

$$\mathcal{H} : |\mathbf{\Sigma}_1| = |\mathbf{\Sigma}_2| = \cdots = |\mathbf{\Sigma}_k| \quad \text{vs.} \quad \mathcal{A} : \neg\mathcal{H}.$$

We discuss the mathematical relationship between the two hypotheses $\mathcal{H} : |\mathbf{\Sigma}_1| = \cdots = |\mathbf{\Sigma}_k|$ and $\widetilde{\mathcal{H}} : \mathbf{\Sigma}_1 = \cdots = \mathbf{\Sigma}_k$. Note that if the hypothesis $\widetilde{\mathcal{H}}$ holds then the hypothesis $\mathcal{H}$ holds. However, its converse does not hold. For example, we set $\mathbf{\Sigma}_1 = \mathrm{diag}(1/2, 1, 1, \ldots, 1)$ and $\mathbf{\Sigma}_2 = \mathrm{diag}(1, 1/2, 1, \ldots, 1)$, then $|\mathbf{\Sigma}_1| = |\mathbf{\Sigma}_2|$ holds, but $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ does not hold. We will mention the application of two hypothesis tests. It is well-known that testing $\widetilde{\mathcal{H}}$ vs. $\mathcal{A} : \neg\widetilde{\mathcal{H}}$ is used for pretesting to decide whether to adopt linear discriminant analysis or quadratic discriminant analysis. On the other hand, the testing $\mathcal{H}$ vs. $\mathcal{A} : \neg\mathcal{H}$ can be used for checking whether the term $\ln|\mathbf{\Sigma}_i|$ in the quadratic discriminant function can be reduced.

Let $n_i = N_i - 1$ for $i \in \{1, 2, \ldots, k\}$. When $p \leq \min\{n_1, n_2, \ldots, n_k\}$, the likelihood ratio test can be used for the above hypothesis. Let

$$n = N - k = \sum_{i=1}^{k} n_i, \ N = \sum_{i=1}^{k} N_i,$$
$$\overline{\mathbf{x}}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} \mathbf{x}_{ij}, \ \mathbf{S}_i = \frac{1}{n_i}\sum_{j=1}^{N_i} (\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)(\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)^\top.$$

Then, the likelihood-ratio test statistic for $\mathcal{H}$ is

$$T_L = Np \ln \left(\sum_{i=1}^{k} \frac{N_i}{N} \left| \frac{n_i}{N_i} \mathbf{S}_i \right|^{1/p} \right) - \sum_{i=1}^{k} N_i \ln \left| \frac{n_i}{N_i} \mathbf{S}_i \right|.$$

This statistic was derived by Najarzadeh [4].

In the large sample framework, the likelihood-ratio test statistic T_L is a

naive estimator of the following criteria:

$$\Lambda_L = Np\left[\ln\left\{\sum_{i=1}^{k}\frac{N_i}{N}\exp\left(\frac{\ln|\boldsymbol{\Sigma}_i|}{p}\right)\right\} - \sum_{i=1}^{k}\frac{N_i}{N}\frac{\ln|\boldsymbol{\Sigma}_i|}{p}\right].$$

We note that Λ_L is a special case of the following criteria, related to Jensen's inequality. For any $\mathbf{c} = (c_1, c_2, \ldots, c_k)^\top$, such that $c_i > 0$ for $i \in \{1, 2, \ldots, k\}$ and $\mathbf{c}^\top \mathbf{1}_k = 1$, we define

$$\Lambda(\mathbf{c}) = Np\left[\ln\left\{\sum_{i=1}^{k}c_i\exp\left(\frac{\ln|\boldsymbol{\Sigma}_i|}{p}\right)\right\} - \sum_{i=1}^{k}c_i\frac{\ln|\boldsymbol{\Sigma}_i|}{p}\right].$$

Note that if $\mathbf{c} = (N_1/N, N_2/N, \ldots, N_k/N)^\top =: \mathbf{c}_L$, then $\Lambda(\mathbf{c}) = \Lambda_L$. Further, from Jensen's inequality, $\Lambda(\mathbf{c}) \geq 0$ and $\Lambda(\mathbf{c}) = 0$ holds if and only if $\mathcal{H}$ holds. Therefore, we understand that $\Lambda(\mathbf{c})$ is reasonable to classify null and alternative hypotheses.

Our goal is to propose a valid test based on criteria $\Lambda(\mathbf{c})$ under high-dimensional settings. T_L/n is a consistent estimator of Λ_L/n under large sample settings, but its validity is broken if the dimension is large. Using $|\mathbf{S}_i|$ to estimate $|\boldsymbol{\Sigma}_i|$ causes a fall in validity. In this study, using the appropriate estimator of $|\boldsymbol{\Sigma}_i|$ under high-dimensional settings, we can estimate $\Lambda(\mathbf{c})$. Specifically, we estimate $|\boldsymbol{\Sigma}_i|$ contained in $\Lambda(\mathbf{c})$ by

$$\widehat{|\boldsymbol{\Sigma}_i|} = |\mathbf{S}_i|\prod_{\ell=1}^{p}\frac{n_i}{n_i - \ell + 1}.$$

Using this estimator, we construct an estimator of $\Lambda(\mathbf{c})$ as follows:

$$L_H(\mathbf{c}) = Np\left\{\ln\left(\sum_{i=1}^{k}c_i\widehat{|\boldsymbol{\Sigma}_i|}^{1/p}\right) - \sum_{i=1}^{k}c_i\ln\widehat{|\boldsymbol{\Sigma}_i|}^{1/p}\right\}$$

and use it as a test statistic. Section 2.1 introduces a good relationship between $L_H(\mathbf{c})$ and $\Lambda(\mathbf{c})$ under high-dimensional settings. Applying the result of Cai et al. [1], we also show that the test statistic $L_H(\mathbf{c})$ converges in distribution to quadratic forms involving normal random vectors (see Theorem 1 for details). From this result, the null asymptotic distribution is described as a weighted sum of chi-squares with respect to $\mathbf{c}$; therefore, it is difficult to handle in practical use. Fortunately, we can solve this problem by choosing a suitable $\mathbf{c}$ (see Section 2.2 for details). Specifically, by choosing $\mathbf{c}$ so that the null asymptotic distribution of constant multiple of statistic $\Lambda(\mathbf{c})$ is a chi-squared distribution, we propose

$$T_H = \hat{q}\sum_{j=1}^{k}\hat{\psi}_j^2 L_H(\mathbf{c}_H),$$

where $\hat{q} = p/n$, $\mathbf{c}_H = (\sum_{j=1}^{k} \hat{\psi}_j^2)^{-1}(\hat{\psi}_1^2, \hat{\psi}_2^2, \ldots, \hat{\psi}_k^2)$. Here, $\hat{\psi}_i = \{-\ln(1 - p/n_i)\}^{-1/2}$.

The remainder of the paper is organized as follows. Section 2 presents the new test statistic, its asymptotic null and non-null distributions, and the asymptotic approximation-based test. Section 3 presents an empirical analysis of the null and non-null distribution of the proposed test statistic. Finally, Section 4 concludes the paper.

§2. Main results

2.1. Estimation of criteria $\Lambda(\mathbf{c})$ in high-dimensional and large sample settings

First, we investigate the asymptotic relationship between T_L and Λ_L under the following asymptotic regime.

(A0) The dimension p is fixed and each $n_i = n_i(n)$ grows as a function of n, such that n_i also tends to infinity. Furthermore, $\lim_{n\to\infty} n_i(n)/n = r_i$ for some $0 < r_i < 1$.

From consistency of $\mathbf{S}_i$, under (A0),

$$\frac{T_L}{n} = \frac{\Lambda_L}{n} + o_p(1), \tag{2.1}$$

as $n \to \infty$.

The statistic T_L has a good property like (2.1) under the large sample framework; however, when the dimension increases, this property is not always preserved. We consider the asymptotic properties of T_L under the following high-dimensional setting.

(A1) Each $n_i = n_i(n)$ grows as a function of n, such that n_i also tends to infinity and $\lim_{n\to\infty} n_i(n)/n = r_i$ for some $0 < r_i < 1$. Let $n_0(n) = \min\{n_1(n), n_2(n), \ldots, n_k(n)\}$. $p = p(n)$ grows as a function of n as long as $p(n) < n_0(n)$, such that p also tends to infinity and $\lim_{n\to\infty} p(n)/n_0(n) = q_0$ for some $0 \le q_0 < 1$.

(A2) Each $\ln|\mathbf{\Sigma}_i|/p$ grows as a function of p, such that $\lim_{p\to\infty} \ln|\mathbf{\Sigma}_i|/p = a_i$ for some $0 \le a_i < \infty$.

Examples of covariance matrices that satisfy (A2) are those with the following spike model:

$$\lambda_j(\mathbf{\Sigma}_i) = O(p),\ j = 1, \ldots, m\ (< \infty),\ \lambda_j(\mathbf{\Sigma}_i) = c_{ij} \in (0, \infty),\ j = m+1, \ldots, p.$$

In fact, $\ln|\mathbf{\Sigma}_i|/p = O(\ln p/p) = o(1)$ as $p \to \infty$.

Under (A1), $\lim_{n\to\infty} p(n)/n = q$ for some $0 \le q < 1$. Note that

$$\begin{aligned} \mathrm{E}\left(\frac{\ln|\mathbf{S}_i|}{p}\right) =& \frac{\ln|\mathbf{\Sigma}_i|}{p} - \ln\left(\frac{n_i}{2}\right) + \frac{1}{p}\sum_{\ell=1}^{p}\psi\left(\frac{n_i-\ell+1}{2}\right) \\ =& \frac{\ln|\mathbf{\Sigma}_i|}{p} + \frac{1}{p}\sum_{\ell=1}^{p}\ln\left(1-\frac{\ell}{n_i}\right) + O\left(\frac{1}{pn_i}\right), \end{aligned} \tag{2.2}$$

$$\begin{aligned} \mathrm{var}\left(\frac{\ln|\mathbf{S}_i|}{p}\right) =& \frac{1}{p^2}\sum_{\ell=1}^{p}\psi'\left(\frac{n_i-\ell+1}{2}\right) \\ =& \frac{1}{p^2}\sqrt{-2\ln\left(1-\frac{p}{n_i}\right)} + O\left(\frac{1}{p^2 n_i}\right), \end{aligned} \tag{2.3}$$

where $\psi(x) = \frac{\partial}{\partial z}\ln\Gamma(z)|_{z=x}$ is the digamma function with the gamma function $\Gamma(z)$. From (2.2) and (2.3), under (A1),

$$\frac{\ln|\mathbf{S}_i|}{p} = \frac{\ln|\mathbf{\Sigma}_i|}{p} + \frac{\sum_{\ell=1}^{p}\ln(1-\ell/n_i)}{p} + O\left(\frac{1}{pn}\right) + O_p\left(\frac{1}{p}\right), \tag{2.4}$$

as $n \to \infty$. From (2.4) and the continuous mapping theorem,

$$\frac{T_L}{np} = \frac{\Lambda^*}{np} + o_p(1) = \frac{\Lambda_L}{np} + O(1) + o_p(1),$$

as $n \to \infty$. Here,

$$\begin{aligned} \Lambda^* =& Np\left[\ln\left\{\sum_{i=1}^{k}\frac{N_i}{N}|\mathbf{\Sigma}_i|^{1/p}\prod_{\ell=1}^{p}\left(1-\frac{\ell}{n_i}\right)^{1/p}\right\}\right. \\ & \left. - \sum_{i=1}^{k}\frac{N_i}{N}\left(\ln|\mathbf{\Sigma}_i|^{1/p} + \frac{\sum_{\ell=1}^{p}\ln(1-\ell/n_i)}{p}\right)\right]. \end{aligned}$$

Thus, the likelihood-ratio test statistic T_L has an asymptotic bias to Λ_L in high-dimensional settings, except to $q_0 = 0$.

To estimate $\Lambda(\mathbf{c})$ without the asymptotic bias in the high-dimensional setting, we prepare a consistent estimator of $\ln|\mathbf{\Sigma}_i|/p$. We construct $\ln|\widehat{\mathbf{\Sigma}_i}|/p$ as an estimator of $\ln|\mathbf{\Sigma}_i|/p$. Note that under (A1) and (A2),

$$\frac{\ln|\widehat{\mathbf{\Sigma}_i}|}{p} = \frac{\ln|\mathbf{\Sigma}_i|}{p} + o_p(1), \tag{2.5}$$

as $n \to \infty$. By changing $\ln |\mathbf{\Sigma}_i|/p$ in $\Lambda(\mathbf{c})$ to $\ln \widehat{|\mathbf{\Sigma}_i|}/p$, we propose the following statistic:

$$L_H(\mathbf{c}) = Np \left\{ \ln \left(\sum_{i=1}^{k} c_i \widehat{|\mathbf{\Sigma}_i|}^{1/p} \right) - \sum_{i=1}^{k} c_i \ln \widehat{|\mathbf{\Sigma}_i|}^{1/p} \right\}.$$

We note that under (A1) and (A2),

$$\frac{L_H(\mathbf{c})}{np} = \frac{\Lambda(\mathbf{c})}{np} + o_p(1), \tag{2.6}$$

as $n \to \infty$. We obtain this result directly from the continuous mapping theorem and (2.5). This results shows that it is reasonable to use $L_H(\mathbf{c})$ in high-dimensional settings.

2.2. Test statistic and their asymptotic null and non-null distributions

In this section, we determine an appropriate $\mathbf{c}$ and propose an approximate size α test based on asymptotic theory. In addition, we obtain the asymptotic distribution of the non-null distribution of the proposed test statistic.

First, we investigate the asymptotic null distribution of $L_H(\mathbf{c})$ under (A1). We prepare the following lemma, given by Cai et al. [1] to derive the asymptotic distribution of $L_H(\mathbf{c})$.

Lemma 1 (Cai et al. [1])**.** *Let* $\mathbf{W}_i \sim \mathcal{W}(n_i, \mathbf{\Sigma}_i)$ *and* $\mathbf{S}_i = \mathbf{W}_i/n_i$*. Under (A1),*

$$\frac{\ln |\mathbf{S}_i| - \sum_{\ell=1}^{p} \ln(1 - \ell/n_i) - \ln |\mathbf{\Sigma}_i|}{\sqrt{-2\ln(1 - p/n_i)}} \rightsquigarrow \mathcal{N}(0, 1),$$

as $n \to \infty$*. Here, "*$\rightsquigarrow$*" denotes convergence in distribution.*

By applying Lemma 1, we obtain the asymptotic null distribution of $L_H(\mathbf{c})$ in the following theorem.

Theorem 1. *Let* $\mathbf{z}$ *be a random vector according to multivariate normal distribution* $\mathcal{N}(\mathbf{0}, \mathbf{I}_k)$*. Define each component of the* $k \times k$ *matrix* $\mathbf{H}_{\mathcal{H}}$ *as follows:*

$$(\mathbf{H}_{\mathcal{H}})_{ij} = \begin{cases} -\frac{c_i}{q}(1 - c_i)\ln(1 - \frac{q}{r_i}), & i = j, \\ -\frac{c_i c_j}{q}\sqrt{-\ln(1 - \frac{q}{r_i})}\sqrt{-\ln(1 - \frac{q}{r_j})}, & i \neq j. \end{cases}$$

Then, under (A1) and $\mathcal{H}$*,* $L_H(\mathbf{c}) \rightsquigarrow \mathbf{z}^\top \mathbf{H}_{\mathcal{H}} \mathbf{z}$*, as* $n \to \infty$*.*

Proof. Let $\mathrm{z}_i = (\ln|\widehat{\boldsymbol{\Sigma}_i}| - \ln|\boldsymbol{\Sigma}_i|)/\sqrt{-2\ln(1-p/n_i)}$. From the mutually independence of $\mathrm{z}_1, \mathrm{z}_2, \ldots, \mathrm{z}_k$ and Lemma 1, under (A1),

$$\mathbf{z} = (\mathrm{z}_1, \mathrm{z}_2, \ldots, \mathrm{z}_k)^\top \rightsquigarrow \mathcal{N}_k(\mathbf{0}, \mathbf{I}_k), \tag{2.7}$$

as $n \to \infty$. By using $\mathbf{z}$, under $\mathcal{H}$ and (A1), $L_H(\mathbf{c})$ is expanded as follows:

$$\begin{aligned} L_H(\mathbf{c}) =& \frac{1}{q}\sum_{i=1}^{k} c_i\{-\ln(1-q/r_i)\}\mathrm{z}_i^2 - \frac{1}{q}\left(\sum_{i=1}^{k} c_i\sqrt{-\ln(1-q/r_i)}\mathrm{z}_i\right)^2 + o_p(1) \\ =& \mathbf{z}^\top \mathbf{H}_{\mathcal{H}}\mathbf{z} + o_p(1), \end{aligned}$$

as $n \to \infty$. This result and (2.7) prove Theorem 1. □

Next, we discuss the selection of $\mathbf{c}$ and propose a new test statistic T_H. From Theorem 1, we understand that the null asymptotic distribution of $L_H(\mathbf{c})$ depends on vector $\mathbf{c}$. We can choose $\mathbf{c}$ so that the null asymptotic distribution of the constant multiple of statistic $L_H(\mathbf{c})$ is a chi-square distribution with $k-1$ degrees of freedom. Let $\psi_i = \{-\ln(1-q/r_i)\}^{-1/2}$ for $i \in \{1, 2, \ldots, k\}$, and let

$$\mathbf{c}_H^* = \frac{1}{\sum_{i=1}^{k}\psi_i^2}\left(\psi_1^2, \psi_2^2, \ldots, \psi_k^2\right)^\top.$$

Then, the matrix $\mathbf{H}_{\mathcal{H}}$ in Theorem 1 is denoted as follows:

$$\mathbf{H}_{\mathcal{H}} = \frac{1}{q\sum_{i=1}^{k}\psi_i^2}(\mathbf{I}_k - \mathbf{b}\mathbf{b}^\top),$$

where

$$\mathbf{b} = \frac{1}{\sqrt{\sum_{i=1}^{k}\psi_i^2}}(\psi_1, \psi_2, \ldots, \psi_k)^\top.$$

Using Theorem 1, under (A1) and $\mathcal{H}$,

$$L_H(\mathbf{c}_H^*) \rightsquigarrow \left(q\sum_{i=1}^{k}\psi_i^2\right)^{-1}\mathbf{z}^\top(\mathbf{I}_k - \mathbf{b}\mathbf{b}^\top)\mathbf{z},$$

as $n \to \infty$. Here, the quadratic form $\mathbf{z}^\top(\mathbf{I}_k - \mathbf{b}\mathbf{b}^\top)\mathbf{z}$ follows a special distribution. Note that $\mathbf{I}_k - \mathbf{b}\mathbf{b}^\top$ is idempotent matrix and $\mathrm{rank}(\mathbf{I}_k - \mathbf{b}\mathbf{b}^\top) = k-1$. Thus, from Cochran's theorem, $\mathbf{z}^\top(\mathbf{I}_k - \mathbf{b}\mathbf{b}^\top)\mathbf{z} \sim \chi_{k-1}^2$. Replacing ψ_i by $\hat{\psi}_i$ in $q\sum_{i=1}^{k}\psi_i^2 L_H(\mathbf{c}_H^*)$, we propose T_H as a test statistic. Note that $\hat{\psi}_i \to \psi_i$ as $n \to \infty$ for $i \in \{1, 2, \ldots, k\}$ under (A1), we get the following corollary.

Corollary 1. *Under (A1) and $\mathcal{H}$, $T_H \rightsquigarrow \chi_{k-1}^2$, as $n \to \infty$.*

We investigate the behavior of the proposed test statistic T_H in the large sample setting (A0). Under (A0), $\hat{q}\sum_{i=1}^{k}\hat{\psi}_i^2 = 1 + o(1)$, $(\mathbf{c}_H)_i = N_i/N + o(1)$, and $\widehat{|\mathbf{\Sigma}_i|} = |\mathbf{S}_i| + o_p(1)$. Here, $(\mathbf{c}_H)_i$ denotes the i-th element of $\mathbf{c}_H$. Hence, under (A0),

$$\frac{T_H}{n} = \frac{T_L}{n} + o_p(1),$$

that is, T_H/n is asymptotically equivalent to the likelihood-ratio test statistic divided by n under the large sample framework.

Theorem 2. *Under (A0) and $\mathcal{H}$, $T_H \rightsquigarrow \chi^2_{k-1}$, as $n \to \infty$.*

Proof. Let $\mathrm{y}_i = (\ln\widehat{|\mathbf{\Sigma}_i|} - \ln|\mathbf{\Sigma}_i|)/\sqrt{2p/n_i}$. From the mutually independence of $\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_k$ and Corollary 1 in Cai et al. [1], under (A0),

$$\mathbf{y} = (\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_k)^\top \rightsquigarrow \mathcal{N}_k(\mathbf{0}, \mathbf{I}_k), \tag{2.8}$$

as $n \to \infty$. By using $\mathbf{z}$, under $\mathcal{H}$ and (A1), L_H is expanded as follows:

$$T_H = \sum_{i=1}^{k}\mathrm{z}_i^2 - \left(\sum_{i=1}^{k}\sqrt{r_i}\mathrm{z}_i\right)^2 + o_p(1) = \mathbf{y}^\top(\mathbf{I}_k - \mathbf{r}\mathbf{r}^\top)\mathbf{y} + o_p(1), \tag{2.9}$$

as $n \to \infty$. Here,

$$\mathbf{r} = (\sqrt{r_1}, \sqrt{r_2}, \ldots, \sqrt{r_k})\,.$$

Note that $\mathbf{I}_k - \mathbf{r}\mathbf{r}^\top$ is idempotent matrix and $\mathrm{rank}(\mathbf{I}_k - \mathbf{r}\mathbf{r}^\top) = k - 1$. Thus, from Cochran's theorem and (2.8), $\mathbf{y}^\top(\mathbf{I}_k - \mathbf{r}\mathbf{r}^\top)\mathbf{y} \rightsquigarrow \chi^2_{k-1}$, as $n \to \infty$. This result and expansion (2.9) prove Theorem 2. □

From Theorem 2, the asymptotic null distribution of T_H is invariant even under the large sample settings (A0), that is, the proposed method is also valid in the large sample settings.

Third, we propose an approximate test of size α by using T_H. From Corollary 1, we propose an approximate test of size α as follows:

$$T_H > \chi^2_{k-1}(\alpha) \iff \text{reject } \mathcal{H},$$

where $\chi^2_{k-1}(\alpha)$ is the upper 100α percentile of chi-square distribution with $k-1$ degrees of freedom.

Finally, we obtain asymptotic non-null distribution of test statistic T_H by applying Lemma 1.

Theorem 3. *Let $\mathbf{z}$ be a random vector according to multivariate normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I}_k)$, and let*

$$\mathbf{H}_{\mathcal{A}} = \frac{1}{q\sum_{j=1}^{k}\psi_j^2\exp(a_j)}\mathbf{D}_{\mathcal{A}}(\mathbf{I}_k - \mathbf{b}_{\mathcal{A}}\mathbf{b}_{\mathcal{A}}^\top)\mathbf{D}_{\mathcal{A}},$$

where

$$\mathbf{D}_{\mathcal{A}} = \mathrm{diag}\left(\sqrt{\exp(a_1)}, \sqrt{\exp(a_2)}, \ldots, \sqrt{\exp(a_k)}\right),$$
$$\mathbf{b}_{\mathcal{A}} = \frac{1}{\sqrt{\sum_{i=1}^k \psi_i^2 \exp(a_i)}} \left(\psi_1\sqrt{\exp(a_1)}, \psi_2\sqrt{\exp(a_2)}, \ldots, \psi_k\sqrt{\exp(a_k)}\right)^\top,$$
$$(\boldsymbol{\gamma})_i = \sqrt{2}N\left(\frac{\psi_i \exp(a_i)}{\sum_{j=1}^k \psi_j^2 \exp(a_j)} - \frac{\psi_i}{\sum_{j=1}^k \psi_j^2}\right).$$

Then, under (A1) and (A2),

$$T_H \rightsquigarrow q\sum_{j=1}^k \psi_j^2\{\Lambda(\mathbf{c}_H^*) + \boldsymbol{\gamma}^\top\mathbf{z} + \mathbf{z}^\top\mathbf{H}_{\mathcal{A}}\mathbf{z}\},$$

as $n \to \infty$.

Proof. Let $\mathrm{z}_i = (\ln|\widehat{\boldsymbol{\Sigma}_i}| - \ln|\boldsymbol{\Sigma}_i|)/\sqrt{-2\ln(1-p/n_i)}$. From mutually independence of $\mathrm{z}_1, \mathrm{z}_2, \ldots, \mathrm{z}_k$ and Lemma 1, under (A1),

$$\mathbf{z} = (\mathrm{z}_1, \mathrm{z}_2, \ldots, \mathrm{z}_k)^\top \rightsquigarrow \mathcal{N}_k(\mathbf{0}, \mathbf{I}), \tag{2.10}$$

as $n \to \infty$. By using $\mathbf{z}$, under (A1) and (A2), $L_H(\mathbf{c}_H^*)$ is expanded as follows:

$$\begin{aligned} L_H(\mathbf{c}_H^*) =& \Lambda(\mathbf{c}_H^*) + \sqrt{2}N\sum_{i=1}^k \left(\frac{\psi_i|\boldsymbol{\Sigma}_i|^{1/p}}{\sum_{j=1}^k \psi_j^2|\boldsymbol{\Sigma}_j|^{1/p}} - \frac{\psi_i}{\sum_{j=1}^k \psi_j^2}\right)\mathrm{z}_i \\ &+ \frac{1}{q\sum_{j=1}^k \psi_j^2|\boldsymbol{\Sigma}_j|^{1/p}}\sum_{i=1}^k |\boldsymbol{\Sigma}_i|^{1/p}\mathrm{z}_i^2 \\ &- \frac{1}{q}\left(\frac{1}{\sum_{j=1}^k \psi_j^2|\boldsymbol{\Sigma}_j|^{1/p}}\sum_{i=1}^k \psi_i|\boldsymbol{\Sigma}_i|^{1/p}\mathrm{z}_i\right)^2 + o_p(1) \\ =& \Lambda(\mathbf{c}_H^*) + \boldsymbol{\gamma}^\top\mathbf{z} + \mathbf{z}^\top\mathbf{H}_{\mathcal{H}}\mathbf{z} + o_p(1). \end{aligned}$$

This result and (2.10) prove Theorem 3. □

§3. Simulation studies

First, we investigate the accuracy of the approximation of the null and non-null distributions using Monte Carlo simulations. We can generate Monte Carlo samples

$$t_H^{(1)}, t_H^{(2)}, \ldots, t_H^{(B)}$$

of test statistic T_H by repeating the following procedure B times.

1 We generate independent sample

$$\boldsymbol{z}_{11}^{(b)}, \boldsymbol{z}_{12}^{(b)}, \ldots, \boldsymbol{z}_{1N_1}^{(b)}, \boldsymbol{z}_{21}^{(b)}, \boldsymbol{z}_{22}^{(b)}, \ldots, \boldsymbol{z}_{2N_2}^{(b)}, \ldots, \boldsymbol{z}_{k1}^{(b)}, \boldsymbol{z}_{k2}^{(b)}, \ldots, \boldsymbol{z}_{kN_k}^{(b)}$$

drawn from $\mathcal{N}(\mathbf{0}, \mathbf{I}_k)$, and calculate $\boldsymbol{x}_{ij}^{(b)} = \boldsymbol{\Sigma}_i^{1/2} \boldsymbol{z}_{ij}^{(b)}$ for

$$j \in \{1, 2, \ldots, N_i\},\ i \in \{1, 2, \ldots, k\}.$$

Let

$$\boldsymbol{X}^{(b)} = (\boldsymbol{x}_{11}^{(b)}, \boldsymbol{x}_{12}^{(b)}, \ldots, \boldsymbol{x}_{1N_1}^{(b)}, \boldsymbol{x}_{21}^{(b)}, \boldsymbol{x}_{22}^{(b)}, \ldots, \boldsymbol{x}_{2N_2}^{(b)}, \ldots, \boldsymbol{x}_{k1}^{(b)}, \boldsymbol{x}_{k2}^{(b)}, \ldots, \boldsymbol{x}_{kN_k}^{(b)}).$$

2 For the sample $\boldsymbol{X}^{(b)}$, we calculate the realized value of T_H, which is denoted as $t_H^{(b)}$.

Using the probability expression given in Theorem 3, we can generate Monte Carlo samples of $q \sum_{j=1}^{k} \psi_j^2 \{\Lambda(\mathbf{c}_H^*) + \boldsymbol{\gamma}^\top \mathbf{z} + \mathbf{z}^\top \mathbf{H}_{\mathcal{A}} \mathbf{z}\}$

$$\tilde{t}_H^{(1)}, \tilde{t}_H^{(2)}, \ldots, \tilde{t}_H^{(B)}$$

by repeating the following procedure B times.

1 We generate independent sample $\boldsymbol{z}^{(b)}$ drawn from $\mathcal{N}(\mathbf{0}, \mathbf{I}_k)$. Estimate r_i, q, and a_i contained in $\boldsymbol{\gamma}$ and $\mathbf{H}_{\mathcal{A}}$ by $\hat{r}_i = n_i/n$, $\hat{q} = p/n$, $\hat{a}_i = \ln |\boldsymbol{\Sigma}_i|/p$, respectively. We denote estimated version $\boldsymbol{\gamma}$ and $\mathbf{H}_{\mathcal{A}}$, as $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\mathbf{H}}_{\mathcal{A}}$, respectively.

2 For the sample $\boldsymbol{z}_b$, we calculate

$$\tilde{t}_H^{(b)} = \hat{q} \sum_{j=1}^{k} \hat{\psi}_j^2 \{\Lambda(\mathbf{c}_H) + \widehat{\boldsymbol{\gamma}}^\top \boldsymbol{z} + \boldsymbol{z}^\top \widehat{\mathbf{H}}_{\mathcal{A}} \boldsymbol{z}\}.$$

In all simulations, we set $B = 10^5$. We implement the above-mentioned procedure with some parameter settings. In all of these simulations, without any loss of generality, we suppose that $\boldsymbol{\mu}_i = \mathbf{0}$ for $i \in \{1, 2, \ldots, k\}$. We set the following model for the covariance structure:

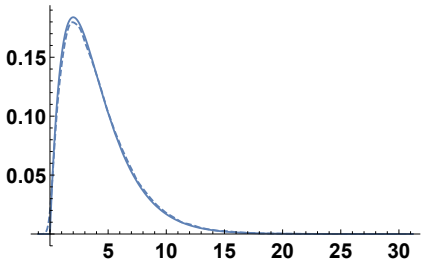
C-I $:(\boldsymbol{\Sigma}_i)_{\ell m} = 0.5^{|\ell - m|}$,

C-II $:(\boldsymbol{\Sigma}_i)_{\ell m} = \left(1 + \frac{2i}{n^\eta}\right) \times 0.5^{|\ell - m|}$, where $\eta \in \{0.3, 0.5, 0.7\}$.

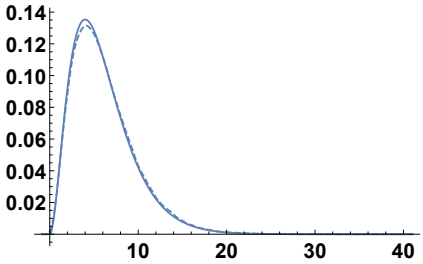
If we set C-I, then $\mathcal{H}$ meets, and if we set C-II, then $\mathcal{A}$ meets. In C-II, the setting with larger η is closer to the null hypothesis $\mathcal{H}$. We also set each sample size $N_i = p + 20i$ for $i \in \{1, 2, \ldots, k\}$. In C-I, for any combination

of $p \in \{2, 50, 100\}$ and $k \in \{3, 5\}$, we compare the smoothed histogram of $t_H^{(1)}, t_H^{(2)}, \ldots, t_H^{(B)}$ with the density function of the chi-squared distribution with $k-1$ degrees of freedom. These results are shown in Figures 1-3. In each figure, we represent the smoothed histogram of $t_H^{(1)}, t_H^{(2)}, \ldots, t_H^{(B)}$ with dashed lines and the chi-squared density function with solid lines, respectively. From Figures 2 and 3, we confirmed that the null distribution of T_H can be well approximated by a chi-square distribution with $k-1$ degrees of freedom when the dimension is large. Furthermore, Figure 1 shows that the approximation accuracy is not significantly degraded even when the dimension is small. These behaviors are consistent with Corollary 1 and Theorem 2. In C-II, for any combination of $p \in \{2, 50, 100\}$, $k \in \{3, 5\}$, and $\eta \in \{0.3, 0.5, 0.7\}$, we compare the smoothed histogram of $t_H^{(1)}, t_H^{(2)}, \ldots, t_H^{(B)}$ with one of $\tilde{t}_H^{(1)}, \tilde{t}_H^{(2)}, \ldots, \tilde{t}_H^{(B)}$. In each figure, we represent the smoothed histogram of $t_H^{(1)}, t_H^{(2)}, \ldots, t_H^{(B)}$ with dashed lines, the smoothed histogram of $\tilde{t}_H^{(1)}, \tilde{t}_H^{(2)}, \ldots, \tilde{t}_H^{(B)}$ with solid lines, and the chi-squared density function with dotted lines, respectively. These results are shown in Figures 4-9. In all figures, $\tilde{t}_H^{(b)}$'s histogram is very close to $t_H^{(b)}$'s histogram. These behaviors are consistent with Theorem 3. Additionally, $\tilde{t}_H^{(b)}$'s histogram and $t_H^{(b)}$'s histogram deviate from the chi-square distribution with $k-1$ degrees of freedom as η decreases, that is, the power of the proposed test increases as η becomes smaller. $\tilde{t}_H^{(b)}$'s histogram and $t_H^{(b)}$'s histogram also deviate from the chi-square distribution with $k-1$ degrees of freedom, as the number of population k and dimension p increase. From these results, we can confirm the natural behavior of the power of the proposed test.

Next, we investigate the size and power of propose test. We use C-I as a setting for calculating empirical sizes, and C-II as a setting for calculating empirical power. We also set the sample size, dimension, and number of populations as in the first experiment, and set the nominal significance level as $\alpha \in \{0.01, 0.05, 0.10\}$. The empirical sizes, calculated with 10^5 replications, are listed in Tables 1. The empirical sizes are close to nominal value α, but still tended to exceed the nominal value α. The empirical powers, calculated with 10^5 replications, are listed in Tables 2 and 3 for $k = 5$ and $k = 7$, respectively. In these tables, we denote empirical power and power obtained using the distribution expression in Theorem 3, by EP and AP, respectively. We confirm that the power tends to decrease as η increased. Also, the power tends to increase as the dimension p and number of groups k increases. These trends are natural. We also confirmed that the approximate power (AP) and the empirical power (EP) are close. Thus, we confirmed that the asymptotic result in Theorem 3 works even for a finite sample.

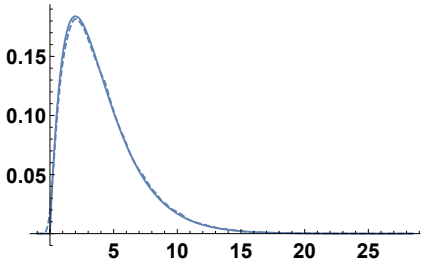


$k = 5$

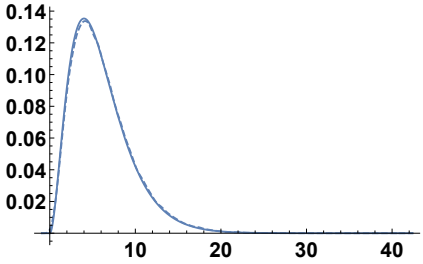


$k = 7$

Figure 1: $p = 2$

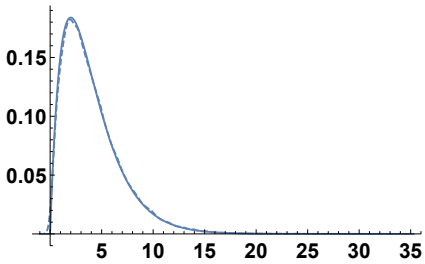


$k = 5$

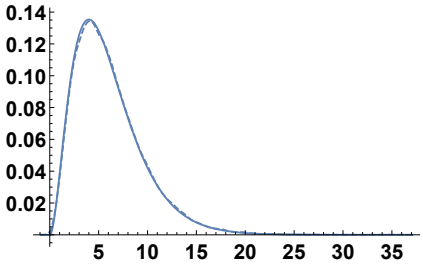


$k = 7$

Figure 2: $p = 50$

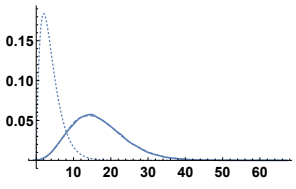


$k = 5$

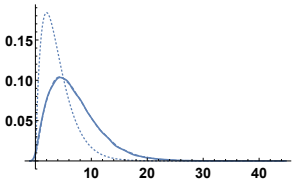


$k = 7$

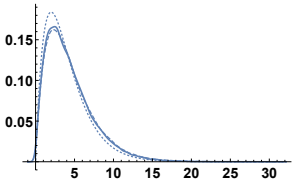
Figure 3: $p = 100$



$\eta = 0.3$

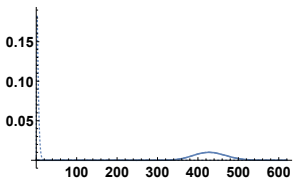


$\eta = 0.5$

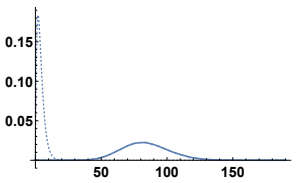


$\eta = 0.7$

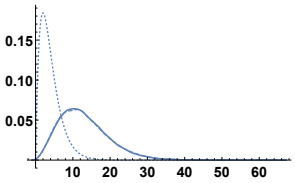
Figure 4: $p = 2, k = 5$



$\eta = 0.3$

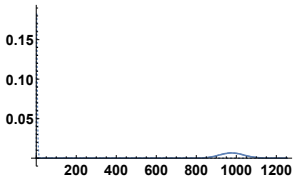


$\eta = 0.5$

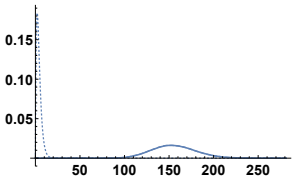


$\eta = 0.7$

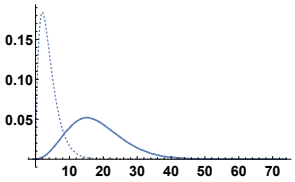
Figure 5: $p = 50, k = 5$



$\eta = 0.3$

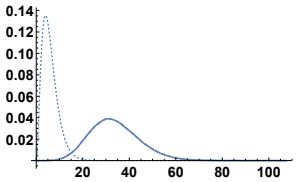


$\eta = 0.5$

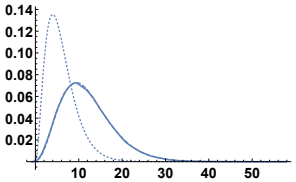


$\eta = 0.7$

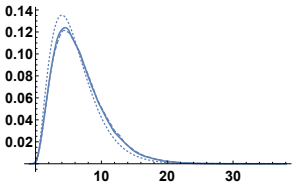
Figure 6: $p = 100, k = 5$



$\eta = 0.3$

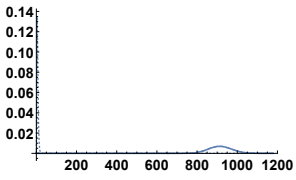


$\eta = 0.5$

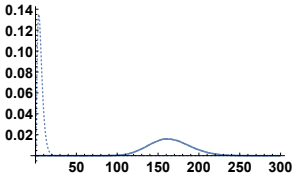


$\eta = 0.7$

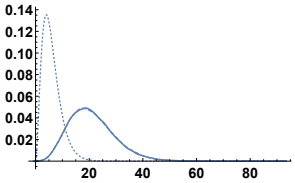
Figure 7: $p = 2, k = 7$



$\eta = 0.3$

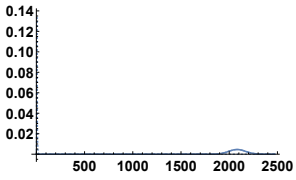


$\eta = 0.5$

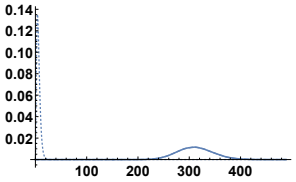


$\eta = 0.7$

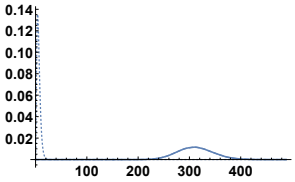
Figure 8: $p = 50, k = 7$



$\eta = 0.3$



$\eta = 0.5$



$\eta = 0.7$

Figure 9: $p = 100, k = 7$

Table 1: The empirical size.

	$k=5$			$k=7$		
$\alpha\backslash p$	$p=2$	$p=50$	$p=100$	$p=2$	$p=50$	$p=100$
0.01	0.011	0.011	0.010	0.011	0.011	0.011
0.05	0.054	0.052	0.052	0.053	0.053	0.052
0.10	0.107	0.104	0.104	0.106	0.106	0.104

Table 2: The empirical power and approximate power when $k=5$.

		$\eta=0.3$		$\eta=0.5$		$\eta=0.7$	
p	α	EP	AP	EP	AP	EP	AP
2	0.01	0.648	0.647	0.095	0.092	0.018	0.016
	0.05	0.844	0.842	0.250	0.245	0.075	0.070
	0.10	0.909	0.907	0.367	0.361	0.139	0.130
50	0.01	1.000	1.000	1.000	1.000	0.425	0.422
	0.05	1.000	1.000	1.000	1.000	0.661	0.661
	0.10	1.000	1.000	1.000	1.000	0.768	0.767
100	0.01	1.000	1.000	1.000	1.000	0.693	0.695
	0.05	1.000	1.000	1.000	1.000	0.865	0.865
	0.10	1.000	1.000	1.000	1.000	0.921	0.921

Table 3: The empirical power and approximate power when $k=7$.

		$\eta=0.3$		$\eta=0.5$		$\eta=0.7$	
p	α	EP	AP	EP	AP	EP	AP
2	0.01	0.965	0.964	0.193	0.190	0.020	0.018
	0.05	0.993	0.992	0.404	0.401	0.081	0.077
	0.10	0.997	0.996	0.535	0.532	0.146	0.140
50	0.01	1.000	1.000	1.000	1.000	0.657	0.655
	0.05	1.000	1.000	1.000	1.000	0.842	0.840
	0.10	1.000	1.000	1.000	1.000	0.906	0.905
100	0.01	1.000	1.000	1.000	1.000	0.917	0.918
	0.05	1.000	1.000	1.000	1.000	0.975	0.976
	0.10	1.000	1.000	1.000	1.000	0.989	0.989

§4. Conclusion

In this study, we proposed an asymptotic approximation-based test for the equality of generalized variances of k multivariate normal populations in high-dimensional and large sample settings. In recent years, Najarzadeh [4] proposed a likelihood-ratio test statistic for this testing problem, and proposed a reasonable approximation test in a large sample setting. Our test is an improvement of the likelihood-ratio statistic that has validity in high dimensional settings. Using the asymptotic results in Cai et al. [1], we obtained null and non-null asymptotic distributions of the proposed test statistic. The features of the proposed test statistic are valid not only for high-dimensional and large sample settings but also for large sample settings. Furthermore, under several parameter settings, we studied the finite sample and dimension behavior of this test statistic through Monte Carlo simulations. The simulation results confirmed that our asymptotic results work well as approximations in a finite sample and dimension. We demonstrate that the proposed test is novel in that it works in a wider range than the conventional method. On the two-sample problem, we also can consider naive statistic $\ln(|\widehat{\mathbf{\Sigma}_1}|/|\widehat{\mathbf{\Sigma}_2}|)$. Future studies can discuss the power comparison between tests, using this statistic and the proposed test.

Acknowledgments

We are grateful to the Editor-in-Chief and reviewer for many valuable comments and helpful suggestions, which have led to an improved version of this paper. We would also like to express our gratitude to Professor Yasunori Fujikoshi for many valuable comments and discussions. The second author's research was supported in part by a Grant-in-Aid for Young Scientists (B) (17K14238) from the Japan Society for the Promotion of Science.

References

[1] Cai, T. T., Liang, T. and Zhou, H. H. (2015). Law of log determinant of sample covariance matrix and optimal estimation of differential entropy for high-dimensional Gaussian distributions, *J. Multivariate Anal.* **137**, 161–172.

[2] Fujikoshi, Y. (1968). Asymptotic expansion of the distribution of the generalized variance in the non-central case, *Journal of Science of the Hiroshima University, Series A-I (Mathematics)* **32**, 293–299.

[3] Nagao, H. (1973). Asymptotic expansions of the distributions of Bartlett's test and sphericity test under the local alternatives, *Ann. Inst. Statist. Math.* **25**, 407–422.

[4] Najarzadeh, D. (2017). Testing equality of generalized variances of k multivariate normal populations, *Comm. Statist. Simulation Comput.* **46**, 6414–6423.

[5] Sugiura, N. (1969). Asymptotic expansions of the distributions of the likelihood ratio criteria for covariance matrix, *Ann. Math. Statist.* **40**, 2051–2063.

[6] Sugiura, N. and Nagao, H. (1968). Unbiasedness of some test criteria for the equality of one or two covariance matrices, *Ann. Math. Statist.* **39**, 1686–1692.

[7] Sugiura, N. and Nagao, H. (1969). On Bartlett's test and Lehmann's test for homogeneity of variances, *Ann. Math. Statist.*, **40**, 2018–2032.

[8] Wilks, S.S. (1932). Certain generalizations in the analysis of variances, *Biometrika* **24**, 471–494.

Takatoshi Sugiyama
Department of Applied Mathematics, Tokyo University of Science
1-3, Kagurazaka, Shinjuku-ku, Tokyo 162-8601, Japan
E-mail: sugiyamatakatoshi@gmail.com

Masashi Hyodo
Department of Mathematical Sciences, Graduate School of Science, Osaka Prefecture University
1-1 Gakuen-cho, Naka-ku, Sakai, Osaka 599-8531, Japan
E-mail: hyodo_h@yahoo.co.jp

Hiroki Watanabe
Department of Health Sciences, Oita University of Nursing and Health Sciences
2944-9, Megusuno, Oita City, Oita 870-1201, Japan
E-mail: watahiro919@gmail.com

Shin-ichi Tsukada
School of Education, Meisei University
2-1-1, Hodokubo, Hino, Tokyo 191-8506 Japan
E-mail: tsukada@ed.meisei-u.ac.jp

Takashi Seo
Department of Applied Mathematics, Tokyo University of Science
1-3, Kagurazaka, Shinjuku-ku, Tokyo 162-8601, Japan
E-mail: seo@rs.tus.ac.jp