Asymptotic distribution of the likelihood ratio test statistic for equality of two covariance matrices with two-step monotone missing data

Yuko Fukumoto[a], Nobumichi Shutoh[a,b,*] and Takashi Seo[c]

[a] *Department of Mathematical Information Science,*
Graduate School of Science, Tokyo University of Science,
1-3, Kagurazaka, Shinjuku-ku, Tokyo 162-8601, Japan
[b] *Research Fellow of the Japan Society for the Promotion of Science*
[c] *Department of Mathematical Information Science,*
Faculty of Science, Tokyo University of Science,
1-3, Kagurazaka, Shinjuku-ku, Tokyo 162-8601, Japan

Abstract

In this paper, we consider a test for the equality of covariance matrices in two sample problem based on 2-step monotone missing data via likelihood ratio criterion. Further, by using the Bartlett correction, we derive modified likelihood ratio test (LRT) statistic. Finally we investigate the asymptotic behavior of the distribution of test statistics by Monte Carlo simulations.

Key Words and Phrases: Monotone missing data; Likelihood ratio test; Asymptotic distribution; Monte Carlo simulation.

1 Introduction

In multivariate analysis based on the data set observed from more than two populations, we may be interested in the assumption for covariance matrices because the types of the

*Corresponding author. nobumichi.shutoh@gmail.com

procedures depend on the equality of them. For example, we use Hotelling's two-sample T^2 statistic under $\Sigma^{(1)} = \Sigma^{(2)}$, on the other hand, under $\Sigma^{(1)} \neq \Sigma^{(2)}$, we use method of Bennett (1951) or Welch's test for testing equality of means vectors in complete data. Thus we consider the test for the equality of covariance matrices.

The most famous scheme of the considered test is LRT. For instance, based on complete data, Nagao (1967) obtained the modified LR critical region and indicated monotonicity of the modified LRT for a covariance matrix. The modified LRT statistic is considered for testing equality of covariance structure for complete data in one sample (see, e.g., Anderson (2003)). Furthermore, the similar procedure for two sample problem could be also derived.

The LRT for equality of two covariance matrices based on complete data has been already considered. First of all, we review the tests based on complete data, i.e., the p-dimensional observation vectors $\boldsymbol{x}_j^{(i)}$ from $\Pi^{(i)}$ $(j = 1, \ldots, N_1^{(i)},\ i = 1, 2)$. Now we consider the LRT for equality of two covariance matrices based on complete data for a special case. Henceforth we consider two populations $\Pi^{(i)} : N_p(\boldsymbol{\mu}^{(i)}, \Sigma^{(i)})$ for $i = 1, 2$ and testing the hypothesis

$$H_0 : \Sigma^{(1)} = \Sigma^{(2)} = I \quad \text{vs.} \quad H_1 : \Sigma^{(1)} \neq \Sigma^{(2)}.$$

In this case, we can provide the LRT statistic for testing H_0 in two sample problem:

$$-2\ln\Lambda_1' \;=\; \sum_{i=1}^{2} -2\ln\Lambda_1^{(i)\prime},$$

where

$$\begin{aligned}
\Lambda_1^{(i)\prime} &= \Big(\frac{e}{N_1^{(i)}-1}\Big)^{\frac{1}{2}(N_1^{(i)}-1)p} \mid V_1^{(i)} \mid \mathrm{etr}\Big\{-\frac{1}{2}V_1^{(i)}\Big\}, \\
V_1^{(i)} &= \sum_{j=1}^{N_1^{(i)}} \big(\boldsymbol{x}_j^{(i)} - \overline{\boldsymbol{x}}^{(i)}\big)\big(\boldsymbol{x}_j^{(i)} - \overline{\boldsymbol{x}}^{(i)}\big)', \\
\overline{\boldsymbol{x}}^{(i)} &= \frac{1}{N_1^{(i)}} \sum_{j=1}^{N_1^{(i)}} \boldsymbol{x}_j^{(i)}.
\end{aligned}$$

Then, under H_0, the asymptotic distribution of $-2\ln\Lambda_1'$ for large $N_1^{(i)}$ is χ^2 distribution with $p(p+1)$ degrees of freedom. Therefore, the hypothesis H_0 is rejected if $-2\ln\Lambda_1' > \chi^2_{p(p+1),\alpha}$, where $\chi^2_{p(p+1),\alpha}$ is the upper $100\alpha\%$ point of χ^2 distribution with $p(p+1)$ degrees of freedom.

Next, we propose the modified LRT statistic as follows:

$$-2\ln\Lambda_1^* = \sum_{i=1}^{2} -2\rho^{(i)}\ln\Lambda_1^{(i)\prime},$$

where $\rho^{(i)} = 1 - (2p^2 + 3p - 1)/\{6(N_1^{(i)} - 1)(p + 1)\}$. Then, under H_0, the asymptotic distribution of $-2\ln\Lambda_1^*$ is also χ^2 distribution with $p(p+1)$ degrees of freedom. Therefore, the hypothesis H_0 is rejected if $-2\ln\Lambda_1^* > \chi^2_{p(p+1),\alpha}$.

Recently, some authors relaxed the assumptions of the data set in the test for the covariance matrices. For example, Schott (2001) considered a Wald statistic under elliptical distributions. He proposed the test for equality of covariance matrices in $K(\geq 2)$ sample problem. Hao and Krishnamoorthy (2001) provided the modified LRT statistic for the null hypothesis $\Sigma = \Sigma_0 = I$ for 2-step monotone missing data. Further, Chang and Richards (2010) provided that for the null hypothesis $\Sigma = \Sigma_0$ for 2-step monotone missing data. These two paper dealt with one sample problem.

In this paper, based on 2-step monotone missing data, we consider the test for equality of covariance matrices in two sample problem. As it turns out, we derive LRT statistic for testing equality of covariance matrices in more complicate setting for the data set. Using the simulation studies, we investigate the asymptotic properties of the proposed test statistics.

This rest of this paper is organized as follows. In Section 2, we review the MLEs under the hypothesis (see, e.g., Anderson and Olkin (1985), Shutoh et al. (2011)). In Section 3, we develop the expression for LRT statistic. In Section 4, we derive modified LRT statistic via the Bartlett correction. Finally, we give numerical results in order to investigate the asymptotic properties of test statistics by Monte Carlo simulations in Section 5.

2 MLEs based on two-step monotone missing data

We assume distribution of observation vector:

$$
\begin{aligned}
\boldsymbol{x}_j^{(i)} = \begin{pmatrix} \boldsymbol{x}_{1j}^{(i)} \\ \boldsymbol{x}_{2j}^{(i)} \end{pmatrix} &\sim N_p(\boldsymbol{\mu}^{(i)}, \Sigma^{(i)}) \quad (j = 1, \ldots, N_1^{(i)},\ i = 1, 2), \\
\boldsymbol{x}_{1j}^{(i)} &\sim N_{p_1}(\boldsymbol{\mu}_1^{(i)}, \Sigma_{11}^{(i)}) \quad (j = N_1^{(i)} + 1, \ldots, N^{(i)},\ i = 1, 2),
\end{aligned}
$$

respectively, where $\boldsymbol{x}_{\ell j}^{(i)}$ $(\ell = 1, 2)$ denotes a p_ℓ-dimensional partitioned vector of $\boldsymbol{x}_j^{(i)}$ and $p = p_1 + p_2$. Further, $\boldsymbol{\mu}^{(i)}$ and $\Sigma^{(i)}$ are partitioned according to blocks of data set, i.e.,

$$
\boldsymbol{\mu}^{(i)} = \begin{pmatrix} \boldsymbol{\mu}_1^{(i)} \\ \boldsymbol{\mu}_2^{(i)} \end{pmatrix}, \quad \Sigma^{(i)} = \begin{pmatrix} \Sigma_{11}^{(i)} & \Sigma_{12}^{(i)} \\ \Sigma_{21}^{(i)} & \Sigma_{22}^{(i)} \end{pmatrix}.
$$

$\boldsymbol{\mu}_1^{(i)}$ is p_1-dimensional vector, $\boldsymbol{\mu}_2^{(i)}$ is p_2-dimensional vector, $\Sigma_{11}^{(i)}$ is $p_1 \times p_1$ matrix, $\Sigma_{12}^{(i)} = \Sigma_{21}^{(i)\prime}$ is $p_1 \times p_2$ matrix and $\Sigma_{22}^{(i)}$ is $p_2 \times p_2$ matrix, respectively.

In general, $\boldsymbol{x}_{1j}^{(i)}$ and $\boldsymbol{x}_{2j}^{(i)}$ are not independent. So we consider the transformation of the observation vector $\boldsymbol{x}_j^{(i)}$ dewoted by $\boldsymbol{y}_j^{(i)} = (\boldsymbol{y}_{1j}^{(i)\prime}, \boldsymbol{y}_{2j}^{(i)\prime})'$, where

$$
\boldsymbol{y}_j^{(i)} = \begin{pmatrix} \boldsymbol{y}_{1j}^{(i)} \\ \boldsymbol{y}_{2j}^{(i)} \end{pmatrix} = \begin{pmatrix} I_{p_1} & 0 \\ -\Sigma_{21}^{(i)}\Sigma_{11}^{(i)-1} & I_{p_2} \end{pmatrix} \begin{pmatrix} \boldsymbol{x}_{1j}^{(i)} \\ \boldsymbol{x}_{2j}^{(i)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_{1j}^{(i)} \\ \boldsymbol{x}_{2j}^{(i)} - \Sigma_{21}^{(i)}\Sigma_{11}^{(i)-1}\boldsymbol{x}_{1j}^{(i)} \end{pmatrix}.
$$

Then, $\boldsymbol{y}_{1j}^{(i)}$ and $\boldsymbol{y}_{2j}^{(i)}$ are mutually independent and are distributed as

$$
\begin{aligned}
\boldsymbol{y}_{1j}^{(i)} &\sim N_{p_1}(\boldsymbol{\eta}_1^{(i)}, \Psi_{11}^{(i)}) \quad (j = 1, \ldots, N_1^{(i)},\ i = 1, 2), \\
\boldsymbol{y}_{2j}^{(i)} &\sim N_{p_2}(\boldsymbol{\eta}_2^{(i)}, \Psi_{22}^{(i)}) \quad (j = N_1^{(i)} + 1, \ldots, N^{(i)},\ i = 1, 2),
\end{aligned}
$$

respectively, where

$$
\begin{aligned}
\boldsymbol{\eta}^{(i)} &= \begin{pmatrix} \boldsymbol{\eta}_1^{(i)} \\ \boldsymbol{\eta}_2^{(i)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu}_1^{(i)} \\ \boldsymbol{\mu}_2^{(i)} - \Psi_{21}^{(i)}\boldsymbol{\mu}_1^{(i)} \end{pmatrix}, \\
\Psi^{(i)} &= \begin{pmatrix} \Psi_{11}^{(i)} & \Psi_{12}^{(i)} \\ \Psi_{21}^{(i)} & \Psi_{22}^{(i)} \end{pmatrix} = \begin{pmatrix} \Sigma_{11}^{(i)} & \Sigma_{11}^{(i)-1}\Sigma_{12}^{(i)} \\ \Sigma_{21}^{(i)}\Sigma_{11}^{(i)-1} & \Sigma_{22\cdot1}^{(i)} \end{pmatrix}, \\
\Sigma_{22\cdot1}^{(i)} &= \Sigma_{22}^{(i)} - \Sigma_{21}^{(i)}\Sigma_{11}^{(i)-1}\Sigma_{12}^{(i)}.
\end{aligned}
$$

In other words, we can write the probability density function of $\boldsymbol{y}_{1j}^{(i)}$ and $\boldsymbol{y}_{2j}^{(i)}$ as follows:

$$
\begin{aligned}
\phi_1^{(i)}(\boldsymbol{y}_{1j}^{(i)}) &= \frac{1}{(2\pi)^{\frac{p_1}{2}} \mid \Psi_{11}^{(i)} \mid^{\frac{1}{2}}} \exp\Big\{ -\frac{1}{2}\big(\boldsymbol{y}_{1j}^{(i)} - \boldsymbol{\eta}_1^{(i)}\big)' \Psi_{11}^{(i)^{-1}} \big(\boldsymbol{y}_{1j}^{(i)} - \boldsymbol{\eta}_1^{(i)}\big)\Big\}, \\
\phi_2^{(i)}(\boldsymbol{y}_{2j}^{(i)}) &= \frac{1}{(2\pi)^{\frac{p_2}{2}} \mid \Psi_{22}^{(i)} \mid^{\frac{1}{2}}} \exp\Big\{ -\frac{1}{2}\big(\boldsymbol{y}_{2j}^{(i)} - \boldsymbol{\eta}_2^{(i)}\big)' \Psi_{22}^{(i)^{-1}} \big(\boldsymbol{y}_{2j}^{(i)} - \boldsymbol{\eta}_2^{(i)}\big)\Big\}.
\end{aligned}
$$

Therefore, the likelihood function to obtain MLEs of $\boldsymbol{\eta}^{(1)}, \boldsymbol{\eta}^{(2)}, \Psi^{(1)}$ and $\Psi^{(2)}$ has the following form:

$$
\begin{aligned}
&\mathrm{L}(\boldsymbol{\eta}^{(1)}, \boldsymbol{\eta}^{(2)}, \Psi^{(1)}, \Psi^{(2)}) \\
&= \prod_{i=1}^{2} \left(\prod_{j=1}^{N^{(i)}} \phi_1^{(i)}(\boldsymbol{y}_{1j}^{(i)}) \prod_{j=1}^{N_1^{(i)}} \phi_1^{(i)}(\boldsymbol{y}_{2j}^{(i)}) \right) \\
&= \text{Const.} \times \prod_{i=1}^{2} \Bigg[|\Psi_{11}^{(i)}|^{-\frac{1}{2}N^{(i)}} |\Psi_{22}^{(i)}|^{-\frac{1}{2}N_1^{(i)}} \\
&\qquad \times \exp\Big\{ -\frac{1}{2} \sum_{j=1}^{N^{(i)}} \big(\boldsymbol{y}_{1j}^{(i)} - \boldsymbol{\eta}_1^{(i)}\big)' \Psi_{11}^{(i)^{-1}} \big(\boldsymbol{y}_{1j}^{(i)} - \boldsymbol{\eta}_1^{(i)}\big)\Big\} \\
&\qquad\quad \times \exp\Big\{ -\frac{1}{2} \sum_{j=1}^{N_1^{(i)}} \big(\boldsymbol{y}_{2j}^{(i)} - \boldsymbol{\eta}_2^{(i)}\big)' \Psi_{22}^{(i)^{-1}} \big(\boldsymbol{y}_{2j}^{(i)} - \boldsymbol{\eta}_2^{(i)}\big)\Big\} \Bigg].
\end{aligned}
$$

If we define the sample mean vectors

$$
\begin{aligned}
\overline{\boldsymbol{x}}_{1T}^{(i)} &= \frac{1}{N^{(i)}} \sum_{j=1}^{N^{(i)}} \boldsymbol{x}_{1j}^{(i)}, \quad \overline{\boldsymbol{x}}_{1F}^{(i)} = \frac{1}{N_1^{(i)}} \sum_{j=1}^{N_1^{(i)}} \boldsymbol{x}_{1j}^{(i)}, \\
\overline{\boldsymbol{x}}_{2F}^{(i)} &= \frac{1}{N_1^{(i)}} \sum_{j=1}^{N_1^{(i)}} \boldsymbol{x}_{2j}^{(i)}, \quad \overline{\boldsymbol{x}}_{1L}^{(i)} = \frac{1}{N_2^{(i)}} \sum_{j=N_1^{(i)}+1}^{N^{(i)}} \boldsymbol{x}_{1j}^{(i)},
\end{aligned}
$$

we can express the MLEs under H_1 as follows:

$$
\widehat{\boldsymbol{\eta}}^{(i)} = \begin{pmatrix} \widehat{\boldsymbol{\eta}}_1^{(i)} \\ \widehat{\boldsymbol{\eta}}_2^{(i)} \end{pmatrix} = \begin{pmatrix} \overline{\boldsymbol{x}}_{1T}^{(i)} \\ \overline{\boldsymbol{x}}_{2F}^{(i)} - \widehat{\Psi}_{21}^{(i)} \overline{\boldsymbol{x}}_{1F}^{(i)} \end{pmatrix}, \quad \widehat{\Psi}^{(i)} = \begin{pmatrix} \widehat{\Psi}_{11}^{(i)} & \widehat{\Psi}_{12}^{(i)} \\ \widehat{\Psi}_{21}^{(i)} & \widehat{\Psi}_{22}^{(i)} \end{pmatrix},
$$

where

$$\begin{aligned}
\widehat{\Psi}_{11}^{(i)} &= \frac{1}{N^{(i)}}\sum_{j=1}^{N^{(i)}}\big(\boldsymbol{x}_{1j}^{(i)}-\overline{\boldsymbol{x}}_{1T}^{(i)}\big)\big(\boldsymbol{x}_{1j}^{(i)}-\overline{\boldsymbol{x}}_{1T}^{(i)}\big)',\\
\widehat{\Psi}_{12}^{(i)} &= \Big\{\sum_{j=1}^{N_1^{(i)}}\big(\boldsymbol{x}_{1j}^{(i)}-\overline{\boldsymbol{x}}_{1F}^{(i)}\big)\big(\boldsymbol{x}_{1j}^{(i)}-\overline{\boldsymbol{x}}_{1F}^{(i)}\big)'\Big\}^{-1}\Big\{\sum_{j=1}^{N_1^{(i)}}\big(\boldsymbol{x}_{1j}^{(i)}-\overline{\boldsymbol{x}}_{1F}^{(i)}\big)\big(\boldsymbol{x}_{2j}^{(i)}-\overline{\boldsymbol{x}}_{2F}^{(i)}\big)'\Big\},\\
\widehat{\Psi}_{22}^{(i)} &= \frac{1}{N_1^{(i)}}\Bigg[\sum_{j=1}^{N_1^{(i)}}\big(\boldsymbol{x}_{2j}^{(i)}-\overline{\boldsymbol{x}}_{2F}^{(i)}\big)\big(\boldsymbol{x}_{2j}^{(i)}-\overline{\boldsymbol{x}}_{2F}^{(i)}\big)'-\Big\{\sum_{j=1}^{N_1^{(i)}}\big(\boldsymbol{x}_{2j}^{(i)}-\overline{\boldsymbol{x}}_{2F}^{(i)}\big)\big(\boldsymbol{x}_{1j}^{(i)}-\overline{\boldsymbol{x}}_{1F}^{(i)}\big)'\Big\}\\
&\qquad\times\Big\{\sum_{j=1}^{N_1^{(i)}}\big(\boldsymbol{x}_{1j}^{(i)}-\overline{\boldsymbol{x}}_{1F}^{(i)}\big)\big(\boldsymbol{x}_{1j}^{(i)}-\overline{\boldsymbol{x}}_{1F}^{(i)}\big)'\Big\}^{-1}\Big\{\sum_{j=1}^{N_1^{(i)}}\big(\boldsymbol{x}_{1j}^{(i)}-\overline{\boldsymbol{x}}_{1F}^{(i)}\big)\big(\boldsymbol{x}_{2j}^{(i)}-\overline{\boldsymbol{x}}_{2F}^{(i)}\big)'\Big\}\Bigg].
\end{aligned}$$

In the same way, we can obtain the MLEs under H_0 as follows:

$$\widetilde{\boldsymbol{\eta}}^{(i)} = \begin{pmatrix}\widetilde{\boldsymbol{\eta}}_1^{(i)}\\ \widetilde{\boldsymbol{\eta}}_2^{(i)}\end{pmatrix} = \begin{pmatrix}\overline{\boldsymbol{x}}_{1T}^{(i)}\\ \overline{\boldsymbol{x}}_{2F}^{(i)}-\widetilde{\Psi}_{21}\overline{\boldsymbol{x}}_{1F}^{(i)}\end{pmatrix},\quad \widetilde{\Psi} = \begin{pmatrix}\widetilde{\Psi}_{11} & \widetilde{\Psi}_{12}\\ \widetilde{\Psi}_{21} & \widetilde{\Psi}_{22}\end{pmatrix},$$

where

$$\begin{aligned}
\widetilde{\Psi}_{11} &= \frac{1}{N}\sum_{i=1}^{2}\sum_{j=1}^{N^{(i)}}\big(\boldsymbol{x}_{1j}^{(i)}-\overline{\boldsymbol{x}}_{1T}^{(i)}\big)\big(\boldsymbol{x}_{1j}^{(i)}-\overline{\boldsymbol{x}}_{1T}^{(i)}\big)',\\
\widetilde{\Psi}_{12} &= \Big\{\sum_{i=1}^{2}\sum_{j=1}^{N_1^{(i)}}\big(\boldsymbol{x}_{1j}^{(i)}-\overline{\boldsymbol{x}}_{1F}^{(i)}\big)\big(\boldsymbol{x}_{1j}^{(i)}-\overline{\boldsymbol{x}}_{1F}^{(i)}\big)'\Big\}^{-1}\Big\{\sum_{i=1}^{2}\sum_{j=1}^{N_1^{(i)}}\big(\boldsymbol{x}_{1j}^{(i)}-\overline{\boldsymbol{x}}_{1F}^{(i)}\big)\big(\boldsymbol{x}_{2j}^{(i)}-\overline{\boldsymbol{x}}_{2F}^{(i)}\big)'\Big\},\\
\widetilde{\Psi}_{22} &= \frac{1}{N_1}\Bigg[\sum_{i=1}^{2}\sum_{j=1}^{N_1^{(i)}}\big(\boldsymbol{x}_{2j}^{(i)}-\overline{\boldsymbol{x}}_{2F}^{(i)}\big)\big(\boldsymbol{x}_{2j}^{(i)}-\overline{\boldsymbol{x}}_{2F}^{(i)}\big)'-\Big\{\sum_{i=1}^{2}\sum_{j=1}^{N_1^{(i)}}\big(\boldsymbol{x}_{2j}^{(i)}-\overline{\boldsymbol{x}}_{2F}^{(i)}\big)\big(\boldsymbol{x}_{1j}^{(i)}-\overline{\boldsymbol{x}}_{1F}^{(i)}\big)'\Big\}\\
&\qquad\times\Big\{\sum_{i=1}^{2}\sum_{j=1}^{N_1^{(i)}}\big(\boldsymbol{x}_{1j}^{(i)}-\overline{\boldsymbol{x}}_{1F}^{(i)}\big)\big(\boldsymbol{x}_{1j}^{(i)}-\overline{\boldsymbol{x}}_{1F}^{(i)}\big)'\Big\}^{-1}\Big\{\sum_{i=1}^{2}\sum_{j=1}^{N_1^{(i)}}\big(\boldsymbol{x}_{1j}^{(i)}-\overline{\boldsymbol{x}}_{1F}^{(i)}\big)\big(\boldsymbol{x}_{2j}^{(i)}-\overline{\boldsymbol{x}}_{2F}^{(i)}\big)'\Big\}\Bigg].
\end{aligned}$$

Then we have the MLEs under H_1:

$$\widehat{\boldsymbol{\mu}}^{(i)} = \begin{pmatrix}\widehat{\boldsymbol{\mu}}_1^{(i)}\\ \widehat{\boldsymbol{\mu}}_2^{(i)}\end{pmatrix} = \begin{pmatrix}\overline{\boldsymbol{x}}_{1T}^{(i)}\\ \overline{\boldsymbol{x}}_{2F}^{(i)}-\widehat{\Psi}_{21}^{(i)}(\overline{\boldsymbol{x}}_{1F}^{(i)}-\overline{\boldsymbol{x}}_{1T}^{(i)})\end{pmatrix},$$

$$\widehat{\Sigma}^{(i)} = \begin{pmatrix}\widehat{\Sigma}_{11}^{(i)} & \widehat{\Sigma}_{12}^{(i)}\\ \widehat{\Sigma}_{21}^{(i)} & \widehat{\Sigma}_{22}^{(i)}\end{pmatrix} = \begin{pmatrix}\widehat{\Psi}_{11}^{(i)} & \widehat{\Psi}_{11}^{(i)}\widehat{\Psi}_{12}^{(i)}\\ \widehat{\Psi}_{21}^{(i)}\widehat{\Psi}_{11}^{(i)} & \widehat{\Psi}_{22}^{(i)}+\widehat{\Psi}_{21}^{(i)}\widehat{\Psi}_{11}^{(i)}\widehat{\Psi}_{12}^{(i)}\end{pmatrix},$$

and the MLEs under H_0:

$$\widetilde{\boldsymbol{\mu}}^{(i)} = \begin{pmatrix} \widetilde{\boldsymbol{\mu}}_1^{(i)} \\ \widetilde{\boldsymbol{\mu}}_2^{(i)} \end{pmatrix} = \begin{pmatrix} \overline{\boldsymbol{x}}_{1T}^{(i)} \\ \overline{\boldsymbol{x}}_{2F}^{(i)} - \widetilde{\Psi}_{21}(\overline{\boldsymbol{x}}_{1F}^{(i)} - \overline{\boldsymbol{x}}_{1T}^{(i)}) \end{pmatrix},$$

$$\widetilde{\Sigma} = \begin{pmatrix} \widetilde{\Sigma}_{11} & \widetilde{\Sigma}_{12} \\ \widetilde{\Sigma}_{21} & \widetilde{\Sigma}_{22} \end{pmatrix} = \begin{pmatrix} \widetilde{\Psi}_{11} & \widetilde{\Psi}_{11}\widetilde{\Psi}_{12} \\ \widetilde{\Psi}_{21}\widetilde{\Psi}_{11} & \widetilde{\Psi}_{22} + \widetilde{\Psi}_{21}\widetilde{\Psi}_{11}\widetilde{\Psi}_{12} \end{pmatrix}.$$

3 Likelihood ratio test statistic

In this section, we develop the expression for the LRT statistic for Σ. We obtain the likelihood ratio

$$\Lambda = \prod_{i=1}^{2} |\, \widehat{\Sigma}_{11}^{(i)} \,|^{\frac{N^{(i)}}{2}} |\, \widehat{\Sigma}_{22\cdot 1}^{(i)} \,|^{\frac{N_1^{(i)}}{2}} \frac{\exp\Big\{ -\dfrac{1}{2}\displaystyle\sum_{j=1}^{N^{(i)}} (\boldsymbol{x}_{1j}^{(i)} - \widetilde{\boldsymbol{\eta}}_1^{(i)})'(\boldsymbol{x}_{1j}^{(i)} - \widetilde{\boldsymbol{\eta}}_1^{(i)}) \Big\}}{\exp\Big\{ -\dfrac{1}{2}\displaystyle\sum_{j=1}^{N^{(i)}} (\boldsymbol{x}_{1j}^{(i)} - \widehat{\boldsymbol{\eta}}_1^{(i)})'\widehat{\Psi}_{11}^{(i)^{-1}}(\boldsymbol{x}_{1j}^{(i)} - \widehat{\boldsymbol{\eta}}_1^{(i)}) \Big\}}$$

$$\times \frac{\exp\Big\{ -\dfrac{1}{2}\displaystyle\sum_{j=1}^{N_1^{(i)}} (\boldsymbol{x}_{2j}^{(i)} - \widetilde{\boldsymbol{\eta}}_2^{(i)})'(\boldsymbol{x}_{2j}^{(i)} - \widetilde{\boldsymbol{\eta}}_2^{(i)}) \Big\}}{\exp\Big\{ -\dfrac{1}{2}\displaystyle\sum_{j=1}^{N_1^{(i)}} (\boldsymbol{x}_{2j}^{(i)} - \widehat{\boldsymbol{\eta}}_2^{(i)})'\widehat{\Psi}_{22}^{(i)^{-1}}(\boldsymbol{x}_{2j}^{(i)} - \widehat{\boldsymbol{\eta}}_2^{(i)}) \Big\}},$$

where

$$\exp\Big\{ -\frac{1}{2}\sum_{j=1}^{N^{(i)}} (\boldsymbol{x}_{1j}^{(i)} - \widetilde{\boldsymbol{\eta}}_1^{(i)})'(\boldsymbol{x}_{1j}^{(i)} - \widetilde{\boldsymbol{\eta}}_1^{(i)}) \Big\} = \exp\Big\{ -\frac{1}{2}\sum_{j=1}^{N^{(i)}} (\boldsymbol{x}_{1j}^{(i)} - \overline{\boldsymbol{x}}_{1T}^{(i)})'(\boldsymbol{x}_{1j}^{(i)} - \overline{\boldsymbol{x}}_{1T}^{(i)}) \Big\},$$

$$\exp\Big\{ -\frac{1}{2}\sum_{j=1}^{N_1^{(i)}} (\boldsymbol{x}_{2j}^{(i)} - \widetilde{\boldsymbol{\eta}}_2^{(i)})'(\boldsymbol{x}_{2j}^{(i)} - \widetilde{\boldsymbol{\eta}}_2^{(i)}) \Big\} = \exp\Big\{ -\frac{1}{2}\sum_{j=1}^{N_1^{(i)}} (\boldsymbol{x}_{2j}^{(i)} - \overline{\boldsymbol{x}}_{2F}^{(i)})'(\boldsymbol{x}_{2j}^{(i)} - \overline{\boldsymbol{x}}_{2F}^{(i)}) \Big\},$$

$$\exp\Big\{ -\frac{1}{2}\sum_{j=1}^{N^{(i)}} (\boldsymbol{x}_{1j}^{(i)} - \widehat{\boldsymbol{\eta}}_1^{(i)})'\widehat{\Psi}_{11}^{(i)^{-1}}(\boldsymbol{x}_{1j}^{(i)} - \widehat{\boldsymbol{\eta}}_1^{(i)}) \Big\} = \exp\left(\frac{1}{2}N^{(i)}p_1\right),$$

$$\exp\Big\{ -\frac{1}{2}\sum_{j=1}^{N_1^{(i)}} (\boldsymbol{x}_{2j}^{(i)} - \widehat{\boldsymbol{\eta}}_2^{(i)})'\widehat{\Psi}_{22}^{(i)^{-1}}(\boldsymbol{x}_{2j}^{(i)} - \widehat{\boldsymbol{\eta}}_2^{(i)}) \Big\} = \exp\left(\frac{1}{2}N_1^{(i)}p_2\right).$$

Hence, the likelihood ratio can be expressed as

$$
\begin{aligned}
\Lambda &= \prod_{i=1}^{2}\Bigg[\mathrm{e}^{\frac{1}{2}(N^{(i)}p_1+N_1^{(i)}p_2)}\mid\widehat{\Sigma}_{11}^{(i)}\mid^{\frac{N^{(i)}}{2}}\mid\widehat{\Sigma}_{22\cdot1}^{(i)}\mid^{\frac{N_1^{(i)}}{2}}\\
&\quad\times\mathrm{etr}\Big\{-\frac{1}{2}\Big(\sum_{j=1}^{N^{(i)}}(\boldsymbol{x}_{1j}^{(i)}-\overline{\boldsymbol{x}}_{1T}^{(i)})'(\boldsymbol{x}_{1j}^{(i)}-\overline{\boldsymbol{x}}_{1T}^{(i)})+\sum_{j=1}^{N_1^{(i)}}(\boldsymbol{x}_{2j}^{(i)}-\overline{\boldsymbol{x}}_{2F}^{(i)})'(\boldsymbol{x}_{2j}^{(i)}-\overline{\boldsymbol{x}}_{2F}^{(i)})\Big)\Big\}\Bigg].
\end{aligned}
$$

Then we define

$$
\begin{aligned}
V^{(i)} &= \sum_{j=1}^{N^{(i)}}(\boldsymbol{x}_{1j}^{(i)}-\overline{\boldsymbol{x}}_{1T}^{(i)})(\boldsymbol{x}_{1j}^{(i)}-\overline{\boldsymbol{x}}_{1T}^{(i)})',\\
V_{\ell m}^{(i)} &= \sum_{j=1}^{N_1^{(i)}}(\boldsymbol{x}_{\ell j}^{(i)}-\overline{\boldsymbol{x}}_{\ell F}^{(i)})(\boldsymbol{x}_{mj}^{(i)}-\overline{\boldsymbol{x}}_{mF}^{(i)})' \quad (\ell, m = 1, 2),
\end{aligned}
$$

and, henceforth, we consider the likelihood ratio statistic with replacing $N^{(i)}$ by $n^{(i)}$ and $N_1^{(i)}$ by $n_1^{(i)}$:

$$
\begin{aligned}
\Lambda' &= \prod_{i=1}^{2}\Bigg[\Big(\frac{e}{n^{(i)}}\Big)^{\frac{n^{(i)}p_1}{2}}\mid V^{(i)}\mid^{\frac{n^{(i)}}{2}}\mathrm{etr}\Big\{-\frac{1}{2}\big(V^{(i)}\big)\Big\}\\
&\quad\times\Big(\frac{e}{n_1^{(i)}}\Big)^{\frac{n_1^{(i)}p_2}{2}}\mid V_{22}^{(i)}-V_{21}^{(i)}V_{11}^{(i)^{-1}}V_{12}^{(i)}\mid^{\frac{n_1^{(i)}}{2}}\mathrm{etr}\Big\{-\frac{1}{2}\big(V_{22}^{(i)}-V_{21}^{(i)}V_{11}^{(i)^{-1}}V_{12}^{(i)}\big)\Big\}\\
&\quad\times\mathrm{etr}\Big\{-\frac{1}{2}\big(V_{21}^{(i)}V_{11}^{(i)^{-1}}V_{12}^{(i)}\big)\Big\}\Bigg]\\
&= \prod_{i=1}^{2}\Bigg[\Lambda_{11}^{(i)'}\Lambda_{22\cdot1}^{(i)'}\mathrm{etr}\Big\{-\frac{1}{2}\big(V_{21}^{(i)}V_{11}^{(i)^{-1}}V_{12}^{(i)}\big)\Big\}\Bigg],
\end{aligned}
$$

where

$$
\begin{aligned}
\Lambda_{11}^{(i)'} &= \Big(\frac{e}{n^{(i)}}\Big)^{\frac{n^{(i)}p_1}{2}}\mid V^{(i)}\mid^{\frac{n^{(i)}}{2}}\mathrm{etr}\Big\{-\frac{1}{2}\big(V^{(i)}\big)\Big\},\\
\Lambda_{22\cdot1}^{(i)'} &= \Big(\frac{e}{n_1^{(i)}}\Big)^{\frac{n_1^{(i)}p_2}{2}}\mid V_{22}^{(i)}-V_{21}^{(i)}V_{11}^{(i)^{-1}}V_{12}^{(i)}\mid^{\frac{n_1^{(i)}}{2}}\mathrm{etr}\Big\{-\frac{1}{2}\big(V_{22}^{(i)}-V_{21}^{(i)}V_{11}^{(i)^{-1}}V_{12}^{(i)}\big)\Big\},\\
n^{(i)} &= N^{(i)}-1, \quad n_1^{(i)} = N_1^{(i)}-p_1-1.
\end{aligned}
$$

Thus, we can obtain the LRT statistic:

$$
-2\ln\Lambda' = \sum_{i=1}^{2}\Big\{-2\ln\Lambda_{11}^{(i)'}-2\ln\Lambda_{22\cdot1}^{(i)'}+\mathrm{tr}\big(V_{21}^{(i)}V_{11}^{(i)^{-1}}V_{12}^{(i)}\big)\Big\}.
$$

As, $N^{(i)} \to \infty$ and $N_1^{(i)} \to \infty$, test statistic is asymptotically distributed as

$$-2\ln\Lambda' \sim \chi^2_{p(p+1)},$$

under H_0. Therefore, if $-2\ln\Lambda' > \chi^2_{p(p+1),\alpha}$, the hypothesis H_0 is rejected.

4 Modified likelihood ratio test statistic

Now, we propose the modified LRT statistic as

$$-2\ln\Lambda^* = \sum_{i=1}^{2}\Big\{-2\rho_1^{(i)}\ln\Lambda_{11}^{(i)\prime} - 2\rho_2^{(i)}\ln\Lambda_{22\cdot1}^{(i)\prime} + \mathrm{tr}\big(V_{21}^{(i)}V_{11}^{(i)^{-1}}V_{12}^{(i)}\big)\Big\},$$

where

$$\rho_1^{(i)} = 1 - \frac{2p_1^2+3p_1-1}{6n^{(i)}(p_1+1)}, \quad \rho_2^{(i)} = 1 - \frac{2p_2^2+3p_2-1}{6n_1^{(i)}(p_2+1)}.$$

As, $N^{(i)} \to \infty$ and $N_1^{(i)} \to \infty$, test statistic is asymptotically distributed as

$$-2\ln\Lambda^* \sim \chi^2_{p(p+1)},$$

under H_0. Therefore, if $-2\ln\Lambda^* > \chi^2_{p(p+1),\alpha}$, the hypothesis H_0 is rejected.

5 Simulation studies

In this section, we present the simulation results under various setting of dimension and sample sizes in order to investigate the asymptotic behavior of the proposed test statistics. For convenience, we prepare data sets whose sample sizes are equal in all the simulation. Let $M_1 \equiv N_1^{(1)} = N_1^{(2)}$ and $M_2 \equiv N_2^{(1)} = N_2^{(2)}$ be the sample size for complete data and that for missing data, respectively. In all the tables, we list the probability

$$\Pr\left(-2\ln T > \chi^2_{f,\alpha}\right),$$

where $-2\ln T$ is LRT statistic and $f = p(p+1)$.

Table 1 Comparison of the LRT statistics with 2-step monotone missing data when $M_1 \to \infty$ and M_2 : fix.

			LRT			modified LRT		
p	M_1	M_2	(1,3)	(2,2)	(3,1)	(1,3)	(2,2)	(3,1)
4	10	10	0.0995	0.0701	0.0679	0.0535	0.0507	0.0503
	20	10	0.0669	0.0588	0.0604	0.0504	0.0500	0.0503
	50	10	0.0559	0.0533	0.0554	0.0502	0.0504	0.0500
	100	10	0.0522	0.0519	0.0520	0.0501	0.0503	0.0503
			(2,6)	(4,4)	(6,2)	(2,6)	(4,4)	(6,2)
8	10	10	0.6076	0.2703	0.2699	0.1398	0.0756	0.0555
	20	10	0.1418	0.0883	0.0954	0.0544	0.0516	0.0512
	50	10	0.0725	0.0617	0.0674	0.0502	0.0501	0.0506
	100	10	0.0599	0.0555	0.0585	0.0505	0.0502	0.0500

At first, we compared the LRT statistic and modified LRT statistic under 2-step monotone missing data, when $p = 4((p_1, p_2) = (1, 3), (2, 2), (3, 1))$, $8((p_1, p_2) = (2, 6), (4, 4), (6, 2))$, $M_1 = 10, 20, 50, 100$, $M_2 = 10$ and $\alpha = 0.05$. In particular, under small number of observations, it can be observed that modified LRT statistic has better accuracy than LRT statistic. In contrast, Table 1 also implies that large dimensionality results in poorer approximations of modified LRT.

Table 2 Comparison of the LRT statistics with 2-step monotone missing data when $M_1 \to \infty$, $M_2 \to \infty$ and $M_1/M_2 = 1$.

			LRT			modified LRT		
p	M_1	M_2	(1,3)	(2,2)	(3,1)	(1,3)	(2,2)	(3,1)
4	10	10	0.0999	0.0707	0.0679	0.0531	0.0506	0.0502
	20	20	0.0668	0.0576	0.0576	0.0501	0.0499	0.0505
	50	50	0.0554	0.0526	0.0527	0.0500	0.0504	0.0500
	100	100	0.0526	0.0508	0.0517	0.0503	0.0499	0.0498
			(2,6)	(4,4)	(6,2)	(2,6)	(4,4)	(6,2)
8	10	10	0.6077	0.2704	0.1635	0.1393	0.0754	0.0555
	20	20	0.1413	0.0845	0.0821	0.0551	0.0515	0.0505
	50	50	0.0724	0.0597	0.0602	0.0503	0.0504	0.0505
	100	100	0.0593	0.0545	0.0550	0.0506	0.0497	0.0500

Table 2 lists the results under both of M_1 and M_2 are large, when $p = 4((p_1, p_2) = (1,3), (2,2), (3,1)), 8((p_1, p_2) = (2,6), (4,4), (6,2)), M_1 = 10, 20, 50, 100, M_2 = 10, 20, 50, 100$ and $\alpha = 0.05$. The simulations conducted in Table 2 implies that the proposed approximation is useful under $M_1 \to \infty$, $M_2 \to \infty$ and $M_1/M_2 = 1$.

Table 3 Comparison of the LRT statistics with complete data and 2-step monotone missing data.

			LRT	modified LRT
p	M_1	M_2	(1,3)	(1,3)
4	10	0	0.2895	0.1263
	20	0	0.1284	0.0786
	50	0	0.0738	0.0599
	100	0	0.0611	0.0546
			(2,6)	(2,6)
8	10	0	0.9733	0.4968
	20	0	0.4226	0.1209
	50	0	0.1334	0.0689
	100	0	0.0833	0.0582

			LRT	modified LRT
p	M_1	M_2	(1,3)	(1,3)
4	10	10	0.0995	0.0535
	20	10	0.0669	0.0504
	50	10	0.0559	0.0502
	100	10	0.0522	0.0501
			(2,6)	(2,6)
8	10	10	0.6076	0.1398
	20	10	0.1418	0.0544
	50	10	0.0725	0.0502
	100	10	0.0599	0.0505

In Table 3, we compare the results for complete data and two-step monotone missing data, where $p = 4((p_1, p_2) = (1,3))$, $8((p_1, p_2) = (2,6))$, $M_1 = 10, 20, 50, 100$, $M_2 = 10$ and $\alpha = 0.05$. However, when we use complete data, we put $M_2 = 0$. The results in Table 3 indicate that modified LRT statistic in the case of 2-step monotone missing data has the better performance.

6 Conclusion and future problems

This paper provided the modified LRT statistic for equality of two covariance matrices for 2-step monotone missing data.

By the performed simulation studies, we compared the result of the LRT statistic and the modified LRT statistic under several settings of dimensionality and sample sizes. The simulation studies listed in Tables 1 and 2 indicated that the result of modified LRT statistic improved the accuracy of the result derived by LRT statistic based on 2-step monotone missing data. Table 3 indicated that the result of modified LRT statistic based on 2-step monotone missing data had better accuracy than the approximations derived by modified LRT statistic based on complete data.

For large p, we consider that it will be needed for better approximation to be provided. Furthermore, we consider that we extend the test for the hypothesis :

$$H : \Sigma^{(1)} = \Sigma^{(2)} \quad \text{vs.} \quad A \neq H,$$

and develop the expression for the modified LRT statistic.

References

[1] Anderson, T. W. (2003). *An introduction to multivariate statistical analysis.*(3rd eds.), John Wiley & Sons, Inc., Hoboken, New Jersey.

[2] Anderson, T. W. and Olkin, I. (1985). Maximum-likelihood estimation of the parameters of a multivariate normal distribution. *Linear Algebra Appl.*, **70**, 147–171.

[3] Bennett, B. M. (1951). Note on a solution of the generalized Behrens-Fisher problem. *Ann. Inst. Statist. Math.*, Tokyo **2**, 87–90.

[4] Chang, W. Y. and Richards, D. St. P. (2010). Finite-sample inference with monotone incomplete multivariate normal data, II. *J. Multivariate Anal.* **101**, 603–620.

[5] Hao, J. and Krishnamoorthy, K. (2001). Inferences on a normal covariance matrix and generalized variance with monotone missing data. *J. Multivariate Anal.*, **78**, 62–82.

[6] Nagao, H (1967). Monotonicity of the modified likelihood ratio test for a covariance matrix. *J. Sci. Hirosima Univ. Ser. A-I.*, **13**, (1967). 147–150.

[7] Schott, J. R. (2001). Some tests for the equality of covariance matrices. *J. Statist. Plann. Inference.*, **94**, (2001), 25–36.

[8] Shutoh, N., Hyodo, M. and Seo, T. (2011). An asymptotic approximation for EPMC in linear discriminant analysis based on two-step monotone missing samples. *J. Multivariate Anal.*, **102**, 252–263.