Two-sample Tests of Sub-mean Vectors under Two-step Monotone Missing Data

Riku Hosonuma[1*], Tamae Kawasaki[2] and Takashi Seo[3]

[1*]Department of Applied Mathematics, Graduate School of Science, Tokyo University of Science, Tokyo, Japan.
[2]Department of Economics, College of Economics, Aoyama Gakuin University, Tokyo, Japan.
[3]Department of Applied Mathematics, Faculty of Science, Tokyo University of Science, Tokyo, Japan.

Abstract

This study proposes a novel test statistic for the two-sample problem involving a sub-mean vector under a two-step monotone missing data structure. Specifically, we derive the asymptotic expansion of the null distribution of the proposed statistic, obtain its distribution function, and provide an approximation to the upper percentiles. Furthermore, Bartlett and Bartlett-type corrections are developed to improve the chi-square approximation. Monte Carlo simulations confirm the accuracy of the proposed approximation and the effectiveness of the correction methods. A numerical example is also presented to illustrate the proposed procedure.

1 Introduction

The sub-mean vector problem was first investigated by Rao (1949), who introduced the well-known Rao's U-statistic for testing hypotheses on a sub-mean vector under multivariate normality. The Rao's U-statistic can be expressed using two Hotelling's T^2 statistics constructed from the mean vector of all components and that of a subset of components. It provides a procedure for testing hypotheses regarding a part of the mean vector while accounting for the remaining components. Siotani, Hayakawa and Fujikoshi (1985) introduced the null distribution of Rao's U-statistic.

Subsequently, several extensions of Rao's U-statistic have been developed. In Rencher (2012), a test for sub-mean vectors in two-sample problem was presented. Gupta, Xu and Fujikoshi (2003) investigated the distribution of Rao's U-statistic and derived an asymptotic expansion of its null distribution. Moreover, they investigated the sub-mean vector problem in the k-sample case and derived theoretical results for approximating the distribution of the generalized test statistic.

In addition to these developments, inference for sub-mean vectors in the two-sample setting has been studied. Kawasaki, Naito and Seo (2019) proposed a Hotelling's T^2-type test statistic and simultaneous confidence intervals for comparing two sub-mean vectors under complete data.

However, in various practical situations, complete observations are often unavailable and missing data frequently arise. This complicates statistical inference and highlights the need for appropriate methods for estimation and hypothesis testing. In the presence of missing data, several authors have developed methods for testing mean vectors; for example, Chang and Richards (2009), Krishnamoorthy and Yu (2012), and Yu, Krishnamoorthy, and Pannala (2006) provided such approaches.

In this line of research, Seko, Kawasaki and Seo (2011) and Seko, Yamazaki, and Seo (2012) studied the two-sample problem of testing the equality of mean vectors under a two-step monotone missing data structure. They proposed two types of test statistics: a likelihood ratio test statistic and a Hotelling's T^2-type test statistic. Related results for the one-sample mean vector under the same setting were developed by Seko, Yamazaki and Seo (2012). In this study, we adopt the T^2-type statistic to construct our test statistic, and utilize the maximum likelihood estimators derived therein.

Recently, Hosonuma, Kawasaki and Seo (2025) studied the sub-mean vector problem under two-step monotone missing data in the one-sample case and proposed test statistics for testing sub-mean vectors. They derived asymptotic expansions of the null distributions of the statistics and obtained approximations to the upper 100α percentiles.

The present study extends the results of Hosonuma, Kawasaki and Seo (2025) to the two-sample case. Specifically, we address the problem of testing the equality of sub-mean vectors under a two-step monotone missing data structure.

In the one-sample problem, the mean vector is assumed to be a known constant, whereas in the two-sample problem, the hypothesis is formulated as the equality of mean vectors across two populations. This leads to a difference in the interpretation of the hypothesis. Furthermore, the present framework accommodates potentially unequal sample sizes, adding further complexity. Therefore, the problem considered here should be regarded as a nontrivial generalization.

The test statistic was constructed by combining a Hotelling's T^2-type statistic with the standard Hotelling's T^2 statistic. By deriving its asymptotic expansion, we obtained a chi-square approximation to the null distribution. Based on this result, an approximation to the upper 100α percentile was derived. To improve the accuracy of the distributional approximation, we further considered Bartlett correction and Bartlett-type correction (Fujikoshi (2000)) for the proposed statistic. The performance of the proposed procedures was investigated through Monte Carlo simulations. In addition, their practical utility was illustrated through a numerical example.

2 Testing of sub-mean vectors

Consider two independent samples indexed by $i = 1, 2$. Let

$$\boldsymbol{x}_1^{(i)}, \ldots, \boldsymbol{x}_{N_1^{(i)}}^{(i)} \sim N_p(\boldsymbol{\mu}^{(i)}, \boldsymbol{\Sigma})$$

and

$$\boldsymbol{x}_{N_1^{(i)}+1}^{(i)}, \ldots, \boldsymbol{x}_{N^{(i)}}^{(i)} \sim N_{p_1+p_2}(\boldsymbol{\mu}_{(12)}^{(i)}, \boldsymbol{\Sigma}_{(12)(12)}).$$

where the mean vector and covariance matrix are partitioned as

$$\boldsymbol{\mu}^{(i)} = \begin{pmatrix} \boldsymbol{\mu}_1^{(i)} \\ \boldsymbol{\mu}_2^{(i)} \\ \boldsymbol{\mu}_3^{(i)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu}_{(12)}^{(i)} \\ \boldsymbol{\mu}_3^{(i)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu}_1^{(i)} \\ \boldsymbol{\mu}_{(23)}^{(i)} \end{pmatrix},$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} & \boldsymbol{\Sigma}_{13} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} & \boldsymbol{\Sigma}_{23} \\ \boldsymbol{\Sigma}_{31} & \boldsymbol{\Sigma}_{32} & \boldsymbol{\Sigma}_{33} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_{(12)(12)} & \boldsymbol{\Sigma}_{(12)3} \\ \boldsymbol{\Sigma}_{3(12)} & \boldsymbol{\Sigma}_{33} \end{pmatrix}.$$

The p-dimensional vector $\boldsymbol{x}_j^{(i)}, j = 1, \ldots, N_1^{(i)}$ ispartitioned as $\boldsymbol{x}_j^{(i)} = (\boldsymbol{x}_{1j}^{(i)}, \boldsymbol{x}_{2j}^{(i)}, \boldsymbol{x}_{3j}^{(i)})$, where $\boldsymbol{x}_{1j}^{(i)} : p_1 \times 1$ vector, $\boldsymbol{x}_{2j}^{(i)} : p_2 \times 1$ and $\boldsymbol{x}_{3j}^{(i)} : p_3 \times 1$. Similarly, for the $p_1 + p_2$-dimensional vector $\boldsymbol{x}_j^{(i)}, j = N_1^{(i)} + 1, \ldots, N^{(i)}$, $\boldsymbol{x}_j^{(i)} = (\boldsymbol{x}_{1j}^{(i)}, \boldsymbol{x}_{2j}^{(i)})$, where $\boldsymbol{x}_{1j}^{(i)} : p_1 \times 1$ vector and $\boldsymbol{x}_{2j}^{(i)} : p_2 \times 1$ vector.

Here, the observations are assumed to follow a two-step monotone missing data. Specifically, the first $N_1^{(i)}$ observations in sample i are fully observed p-dimensional vectors, whereas the remaining $N_2^{(i)} (= N^{(i)} - N_1^{(i)})$ observations include only the first $p_1 + p_2$ components. The third p_3 block of variables is missing for these observations.

The sample mean vectors for the fully observed data are defined as

$$\overline{\boldsymbol{x}}_{1F}^{(i)} = \frac{1}{N_1^{(i)}} \sum_{j=1}^{N_1^{(i)}} \boldsymbol{x}_{1j}^{(i)}, \overline{\boldsymbol{x}}_{2F}^{(i)} = \frac{1}{N_1^{(i)}} \sum_{j=1}^{N_1^{(i)}} \boldsymbol{x}_{2j}^{(i)}, \overline{\boldsymbol{x}}_{3F}^{(i)} = \frac{1}{N_1^{(i)}} \sum_{j=1}^{N_1^{(i)}} \boldsymbol{x}_{3j}^{(i)}.$$

Let $\overline{\boldsymbol{x}}_F^{(i)} = (\overline{\boldsymbol{x}}_{1F}^{(i)'}, \overline{\boldsymbol{x}}_{2F}^{(i)'}, \overline{\boldsymbol{x}}_{3F}^{(i)'})' = (\overline{\boldsymbol{x}}_{(12)F}^{(i)'}, \overline{\boldsymbol{x}}_{3F}^{(i)'})'$. For the partial observation

$$\overline{\boldsymbol{x}}_{1L}^{(i)} = \frac{1}{N_2^{(i)}} \sum_{j=N_1^{(i)}+1}^{N^{(i)}} \boldsymbol{x}_{1j}^{(i)}, \overline{\boldsymbol{x}}_{2L}^{(i)} = \frac{1}{N_2^{(i)}} \sum_{j=N_1^{(i)}+1}^{N^{(i)}} \boldsymbol{x}_{2j}^{(i)}.$$

Let $\overline{\boldsymbol{x}}_{(12)L}^{(i)} = (\overline{\boldsymbol{x}}_{1L}^{(i)'}, \overline{\boldsymbol{x}}_{2L}^{(i)'})'$. Further, define the total sample mean

$$\overline{\boldsymbol{x}}_{1T}^{(i)} = \frac{1}{N^{(i)}} \sum_{j=1}^{N^{(i)}} \boldsymbol{x}_{1j}^{(i)}, \overline{\boldsymbol{x}}_{2T}^{(i)} = \frac{1}{N^{(i)}} \sum_{j=1}^{N^{(i)}} \boldsymbol{x}_{2j}^{(i)}.$$

Let $\overline{\boldsymbol{x}}_{(12)T}^{(i)} = (\overline{\boldsymbol{x}}_{1T}^{(i)'}, \overline{\boldsymbol{x}}_{2T}^{(i)'})'$. The unbiased covariance matrices are defined as

$$\boldsymbol{S}_F = \frac{1}{N_1^{(1)} + N_1^{(2)} - 2} \sum_{i=1}^{2} \sum_{j=1}^{N_1^{(i)}} (\boldsymbol{x}_j^{(i)} - \overline{\boldsymbol{x}}_F^{(i)})(\boldsymbol{x}_j^{(i)} - \overline{\boldsymbol{x}}_F^{(i)})',$$

$$\boldsymbol{S}_L = \frac{1}{N_2^{(1)} + N_2^{(2)} - 2} \sum_{i=1}^{2} \sum_{j=N_1^{(i)}+1}^{N^{(i)}} (\boldsymbol{x}_j^{(i)} - \overline{\boldsymbol{x}}_{(12)L}^{(i)})(\boldsymbol{x}_j^{(i)} - \overline{\boldsymbol{x}}_{(12)L}^{(i)})'.$$

The problem of testing the equality of the sub-mean vector $\boldsymbol{\mu}_{(23)}^{(i)}$ between two-samples is considered under the assumption that the first component is common. Specifically, the hypothesis is formulated as

$$H_0 : \boldsymbol{\mu}_{(23)}^{(1)} = \boldsymbol{\mu}_{(23)}^{(2)} \ \text{ given } \ \boldsymbol{\mu}_1^{(1)} = \boldsymbol{\mu}_1^{(2)} \ \text{ vs. } \ H_1 : \boldsymbol{\mu}_{(23)}^{(1)} \neq \boldsymbol{\mu}_{(23)}^{(2)} \ \text{ given } \ \boldsymbol{\mu}_1^{(1)} = \boldsymbol{\mu}_1^{(2)}.$$

For testing the above hypothesis, the Hotelling's T^2-type statistic T_{Mp}^2 and Hotelling's T^2 statistic $T_{p_1}^2$, which have been derived in previous studies, are employed. The statistic T_{Mp}^2 is based on the two-step monotone missing data , while $T_{p_1}^2$ corresponds to the standard Hotelling statistic based on the complete data.

Using these statistics, we define

$$U = (M - p)\frac{T_{Mp}^2 - T_{p_1}^2}{M + T_{p_1}^2},$$

where

$$\begin{aligned}
T_{Mp}^2 =& (\widehat{\boldsymbol{\mu}}^{(1)} - \widehat{\boldsymbol{\mu}}^{(2)})'\widehat{\boldsymbol{\Gamma}}^{-1}(\widehat{\boldsymbol{\mu}}^{(1)} - \widehat{\boldsymbol{\mu}}^{(2)}),\\
T_{p_1}^2 =& \frac{N^{(1)}N^{(2)}}{N^{(1)} + N^{(2)}}(\overline{\boldsymbol{x}}_{1T}^{(1)} - \overline{\boldsymbol{x}}_{1T}^{(2)})'\boldsymbol{S}_{11}^{-1}(\overline{\boldsymbol{x}}_{1T}^{(1)} - \overline{\boldsymbol{x}}_{1T}^{(2)}),\\
S_{11} =& \frac{1}{M}\sum_{i=1}^{2}\sum_{j=1}^{N^{(i)}}(\boldsymbol{x}_{1j}^{(i)} - \overline{\boldsymbol{x}}_{1T}^{(i)})(\boldsymbol{x}_{1j}^{(i)} - \overline{\boldsymbol{x}}_{1T}^{(i)})',
\end{aligned}$$

and $M = N^{(1)} + N^{(2)} - 2$. The proposed statistic is constructed by replacing the Hotelling's T^2 statistics in Rao's U statistic with the corresponding Hotelling's T^2-type statistics adapted to the present missing data framework. Under the two-step monotone missing data structure, the maximum likelihood estimators of the mean vectors $\widehat{\boldsymbol{\mu}}^{(i)}$ and covariance matrix $\widehat{\boldsymbol{\Gamma}}$ have been derived by Seko, Kawasaki and Seo (2011). In this study, we employed these estimators in the construction of the statistic T_{Mp}^2. As shown by Seko, Kawasaki and Seo (2011), the resulting estimators can be expressed in terms of the sample means and covariance matrices. In particular, $\widehat{\boldsymbol{\mu}}^{(i)}$ and $\widehat{\boldsymbol{\Gamma}}$ are functions of the sample mean vectors and covariance matrices, respectively.

To approximate the distribution of the statistic U, we derive an asymptotic expansion of its distribution function under the null hypothesis. The asymptotic expansion is calculated as follows: $\gamma_j^{(i)} = (N_j^{(i)} - 1)/(N^{(i)} - 2) \to \delta_j^{(i)} \in (0,1), \lambda^{(i)} = (N_1^{(i)} + N_2^{(i)} - 1)/M, \lambda_1 = (N_1^{(1)} + N_1^{(2)} - 1)/M$, as $N_1^{(i)}, N_2^{(i)} \to \infty$. Let $G_k(x)$ denote the distribution function of a chi-squared distribution with k degrees. Subsequently, the distribution of U can be approximated as

$$\Pr(U \le x) = G_{p-p_1}(x) + \frac{1}{M}\sum_{j=0}^{2} \beta_j G_{p-p_1+2j}(x) + O(M^{-2}),$$

where

$$\begin{aligned}
\beta_0 =& \lambda^{(1)}\lambda^{(2)}\Big(\frac{\lambda^{(2)}}{\lambda^{(1)2}} + \frac{\lambda^{(1)}}{\lambda^{(2)2}} + \frac{1}{\lambda^{(1)}} + \frac{1}{\lambda^{(2)}}\Big)p_1 - \frac{1}{2}p_1p_2 \\
&+ \frac{1}{2}\Big[1 - \Big(\frac{1}{\gamma_1^{(1)}\lambda^{(1)}} + \frac{1}{\gamma_1^{(2)}\lambda^{(2)}}\Big)^{-1}\Big\{\frac{1}{\lambda^{(1)}\lambda^{(2)}\lambda_1} + \Big(\frac{\gamma_2^{(1)}}{\gamma_1^{(1)2}\lambda^{(1)2}} + \frac{\gamma_2^{(2)}}{\gamma_1^{(2)2}\lambda^{(2)2}}\Big) \\
&+ \frac{1}{\lambda^{(1)}\lambda^{(2)}\lambda_1}\Big(1 - \frac{\gamma_2^{(2)}}{\gamma_1^{(2)}}\lambda^{(1)} - \frac{\gamma_2^{(1)}}{\gamma_1^{(1)}}\lambda^{(2)}\Big)\Big\}\Big]p_1p_3 \\
&+ \Big\{\lambda^{(1)}\lambda^{(2)}\Big(\frac{\lambda^{(2)}}{\lambda^{(1)2}} + \frac{\lambda^{(1)}}{\lambda^{(2)2}} + \frac{1}{\lambda^{(1)}} + \frac{1}{\lambda^{(2)}}\Big) - \frac{1}{2}\Big\}p_2 \\
&- \frac{1}{4}p_2^2 - \frac{1}{2}\Big(\frac{1}{\gamma_1^{(1)}\lambda^{(1)}} + \frac{1}{\gamma_1^{(2)}\lambda^{(2)}}\Big)^{-1}\Big\{\Big(\frac{\gamma_2^{(1)}}{\gamma_1^{(1)2}\lambda^{(1)2}} + \frac{\gamma_2^{(2)}}{\gamma_1^{(2)2}\lambda^{(2)2}}\Big) \\
&+ \frac{1}{\lambda^{(1)}\lambda^{(2)}\lambda_1}\Big(1 - \frac{\gamma_2^{(2)}}{\gamma_1^{(2)}}\lambda^{(1)} - \frac{\gamma_2^{(1)}}{\gamma_1^{(1)}}\lambda^{(2)}\Big)\Big\}p_2p_3 \\
&+ \frac{1}{2\gamma_1^{(1)}\lambda^{(1)}\gamma_1^{(2)}\lambda^{(2)}\lambda_1^2}\Big\{\Big(\gamma_1^{(1)}\lambda^{(1)}\Big)^2 + \Big(\gamma_1^{(2)}\lambda^{(2)}\Big)^2\Big\}p_3,
\end{aligned}$$

$$\begin{aligned}
\beta_1 =& -\lambda^{(1)}\lambda^{(2)}\Big(\frac{\lambda^{(2)}}{\lambda^{(1)2}} + \frac{\lambda^{(1)}}{\lambda^{(2)2}} + \frac{1}{\lambda^{(1)}} + \frac{1}{\lambda^{(2)}}\Big)p_1 + \frac{1}{2}p_1p_2 \\
&- \frac{1}{2}\Big[1 - \Big(\frac{1}{\gamma_1^{(1)}\lambda^{(1)}} + \frac{1}{\gamma_1^{(2)}\lambda^{(2)}}\Big)^{-1}\Big\{\frac{1}{\lambda^{(1)}\lambda^{(2)}\lambda_1} + \Big(\frac{\gamma_2^{(1)}}{\gamma_1^{(1)2}\lambda^{(1)2}} + \frac{\gamma_2^{(2)}}{\gamma_1^{(2)2}\lambda^{(2)2}}\Big) \\
&+ \frac{1}{\lambda^{(1)}\lambda^{(2)}\lambda_1}\Big(1 - \frac{\gamma_2^{(2)}}{\gamma_1^{(2)}}\lambda^{(1)} - \frac{\gamma_2^{(1)}}{\gamma_1^{(1)}}\lambda^{(2)}\Big)\Big\}\Big]p_1p_3 \\
&- \lambda^{(1)}\lambda^{(2)}\Big(\frac{\lambda^{(2)}}{\lambda^{(1)2}} + \frac{\lambda^{(1)}}{\lambda^{(2)2}} + \frac{1}{\lambda^{(1)}} + \frac{1}{\lambda^{(2)}}\Big)p_2 \\
&+ \frac{1}{2}\Big(\frac{1}{\gamma_1^{(1)}\lambda^{(1)}} + \frac{1}{\gamma_1^{(2)}\lambda^{(2)}}\Big)^{-1}\Big\{\Big(\frac{\gamma_2^{(1)}}{\gamma_1^{(1)2}\lambda^{(1)2}} + \frac{\gamma_2^{(2)}}{\gamma_1^{(2)2}\lambda^{(2)2}}\Big)
\end{aligned}$$

$$- \frac{1}{\lambda^{(1)}\lambda^{(2)}\lambda_1}\Big(\frac{\gamma_2^{(2)}}{\gamma_1^{(2)}}\lambda^{(1)} + \frac{\gamma_2^{(1)}}{\gamma_1^{(1)}}\lambda^{(2)}\Big)\Big\}p_2p_3$$
$$- \frac{1}{2\gamma_1^{(1)}\lambda^{(1)}\gamma_1^{(2)}\lambda^{(2)}\lambda_1^2}\Big\{\Big(\gamma_1^{(1)}\lambda^{(1)}\Big)^2 + \Big(\gamma_1^{(2)}\lambda^{(2)}\Big)^2\Big\}p_3 - \frac{1}{4\gamma_1}p_3(p_3+2),$$

$$\beta_2 = \frac{1}{4}p_2(p_2+2) + \frac{1}{2}\Big(\frac{1}{\gamma_1^{(1)}\lambda^{(1)}} + \frac{1}{\gamma_1^{(2)}\lambda^{(2)}}\Big)^{-1}\frac{1}{\lambda^{(1)}\lambda^{(2)}\lambda_1}p_2p_3 + \frac{1}{4\lambda_1}p_3(p_3+2).$$

Next, we consider an approximation to the upper 100α percentiles of U. Let $u(\alpha)$ denote the upper 100α percentiles of the distribution of U. Using the asymptotic expansion of the distribution function obtained earlier, we can derive an approximation to $u(\alpha)$. Let $\chi_k^2(\alpha)$ denote the upper 100α percentiles of the chi-squared distribution with k degrees of freedom.

$$u(\alpha) = \chi^2_{p-p_1}(\alpha) + \frac{2}{M}\chi^2_{p-p_1}(\alpha)\Big\{\frac{\beta_2\chi^2_{p-p_1}(\alpha)}{(p-p_1)(p-p_1+2)} + \frac{\beta_1+\beta_2}{p-p_1}\Big\} + O(M^{-2})$$

The percentile $u(\alpha)$ can be approximated as

$$u^*(\alpha) = \chi^2_{p-p_1}(\alpha) + \frac{2}{M}\chi^2_{p-p_1}(\alpha)\Big\{\frac{\beta_2\chi^2_{p-p_1}(\alpha)}{(p-p_1)(p-p_1+2)} + \frac{\beta_1+\beta_2}{p-p_1}\Big\}.$$

To improve the accuracy of distributional approximations, we consider Bartlett and Bartlett-type corrections for the statistic U proposed in this study. The Bartlett correction is obtained by adjusting the expectation of the statistic to match the expectation of the limiting chi-square distribution to a higher order approximation.

From the asymptotic expansion of the distribution function, the expectation of U under the null hypothesis is denoted by

$$\mathrm{E}[U] = p - p_1 + \frac{2\beta_1 + 4\beta_2}{M} + O(M^{-2}).$$

Thus, the Bartlett correction is defined by

$$U_b = \Big\{1 - \frac{2\beta_1+4\beta_2}{M(p-p_1)}\Big\}U, \qquad \Big(M > \frac{2\beta_1+4\beta_2}{p-p_1}\Big).$$

This correction improves the chi-square approximation to the null distribution by eliminating the first-order bias in the expectation. In addition, a Bartlett-type correction following Fujikoshi (2000) is adopted, which further improves the approximation by reducing higher-order terms in the asymptotic expansion of the distribution.

The Bartlett-type correction is defined by

$$U_{bt} = (Mk_1 + k_2)\log\Big(1 + \frac{U}{Mk_1}\Big),$$

$$k_1 = \frac{1}{4\beta_2}(p_2+p_3)(p_2+p_3+2),$$
$$k_2 = -\frac{\beta_1+\beta_2}{2\beta_2}(p_2+p_3+2),$$

where $k_1 > 0$ and $Mk_1 + k_2 > 0$ is required. The effectiveness of these corrections is evaluated through Monte Carlo simulations in the next section.

3 Monte Carlo Simulation

To investigate the performance of the proposed approximation, we conducted Monte Carlo simulations. Specifically, we generated artificial two-step monotone missing data from $N_p(\mathbf{0}, \boldsymbol{I})$ for various combinations of dimension and sample size. The simulation was repeated 10^7 times. The following four testing procedures were compared.

1. Chi-square approximation
   Comparing the statistic U and the upper 100α percentile of the chi-squared distribution with $p - p_1$ degrees of freedom.

$$P_c = \Pr\{U > \chi^2_{p-p_1}(\alpha)\}$$

2. Proposed percentile approximation
   Comparison between the statistic U and the proposed approximate upper 100α percentile $u^*(\alpha)$, we obtain.

$$P_u = \Pr\{U > u^*(\alpha)\}$$

3. Bartlett correction
   The Bartlett correction U_b is compared with the chi-squared percentile.

$$P_b = \Pr\{U_b > \chi^2_{p-p_1}(\alpha)\}$$

4. Bartlett-type correction
   Comparison between the Bartlett-type correction U_{bt} and the chi-squared percentile indicates.

$$P_{bt} = \Pr\{U_{bt} > \chi^2_{p-p_1}(\alpha)\}$$

For each method, actual type I error rates was computed. The results are reported in Tables.

First, we considered the baseline case, where the dimension parameters are denoted by $p_1 = p_2 = p_3 = 2$. The sample sizes for the two groups were assumed to be equal: $N_1^{(1)} = N_2^{(1)} = N_1^{(2)} = N_2^{(2)}$. We set the significance levels $\alpha = 0.05$ and 0.01. The simulation results for this setting are shown in Table 1.

Next, we investigated the effect of different configurations of (p_1, p_2, p_3). Specifically, we considered the following three cases $(p_1, p_2, p_3) = (4, 2, 2), (2, 4, 2), (2, 2, 4)$.

Table 1 Upper percentiles and type I errors when $(p_1, p_2, p_3) = (2, 2, 2)$, and $\alpha = 0.05, 0.01$

$N_j^{(i)}$	$U(\alpha)$	$u^*(\alpha)$	$U_b(\alpha)$	$U_{bt}(\alpha)$	P_c	P_u	P_b	P_{bt}
$\alpha = 0.05$								
20	7.58	9.30	7.88	7.77	0.018	0.020	0.022	0.019
40	8.46	9.40	8.62	8.54	0.031	0.032	0.033	0.031
80	8.96	9.45	9.04	8.99	0.039	0.040	0.041	0.040
160	9.21	9.47	9.25	9.22	0.044	0.045	0.045	0.045
320	9.35	9.48	9.37	9.35	0.047	0.047	0.048	0.047
640	9.42	9.48	9.43	9.42	0.049	0.049	0.049	0.049
1280	9.45	9.49	9.46	9.45	0.049	0.049	0.049	0.049
2560	9.47	9.49	9.47	9.47	0.050	0.050	0.050	0.050
$\alpha = 0.01$								
20	10.49	13.88	10.91	10.52	0.002	0.001	0.003	0.002
40	11.73	13.58	11.96	11.69	0.005	0.004	0.005	0.004
80	12.47	13.43	12.59	12.42	0.007	0.006	0.007	0.007
160	12.85	13.35	12.91	12.82	0.008	0.008	0.008	0.008
320	13.07	13.32	13.10	13.05	0.009	0.009	0.009	0.009
640	13.17	13.30	13.19	13.16	0.010	0.009	0.010	0.010
1280	13.23	13.29	13.24	13.23	0.010	0.010	0.010	0.010
2560	13.26	13.28	13.26	13.25	0.010	0.010	0.010	0.010

As in the previous setting, the sample sizes were assumed to be equal: $N_1^{(1)} = N_2^{(1)} = N_1^{(2)} = N_2^{(2)}$. The significance level was fixed at $\alpha = 0.05$. The results are summarized in Table 2.

Next, we examined the effect of unequal sample sizes. In this case, the sample size ratio is denoted by $N_1^{(1)} : N_2^{(1)} : N_1^{(2)} : N_2^{(2)} = 1 : 1 : 2 : 2$. The simulation results are presented in Table 3.

Finally, we considered another unbalanced configuration of sample sizes: $N_1^{(1)} : N_2^{(1)} : N_1^{(2)} : N_2^{(2)} = 2 : 1 : 2 : 1$. The results are reported in Table 4.

Based on Tables 1–4, the following observations can be made. First, when p_1 is large, meaning that the dimension of the given component is high, the approximation accuracy using the upper $100\alpha\%$ percentile $u^*(\alpha)$ improves. Second, for all patterns considered, both U_b and U_{bt} were confirmed to improve the approximation accuracy. In particular, in many cases, the Bartlett correction U_b exhibits superior approximation accuracy compared to the others. Furthermore, for the same sample size, the approximation accuracy is better when $(p_1, p_2, p_3) = (4, 2, 2)$ than when $(p_1, p_2, p_3) = (2, 4, 2)$; this suggests that, even when the total dimension is the same, a larger p_1 leads to better approximation accuracy than a larger p_2.

We examined the power performance of the proposed test statistics. In addition to controlling type I error rates, it is important to evaluate the ability of the tests to detect departures from the null hypothesis has to be evaluated.

Table 5 presents the power of the test statistics U, U_b, and U_{bt} when $p_1 = p_2 = p_3 = 2$ and $N_1^{(1)} = N_2^{(1)} = N_1^{(2)} = N_2^{(2)} = 40$. Power is defined as $1 - \beta$, where β denotes the probability of a type II error. Specifically, $1 - \beta_U$, $1 - \beta_{U_b}$, and $1 - \beta_{U_{bt}}$ represent the power of U, U_b, and U_{bt}, respectively.

Table 2 Upper percentiles and type I errors, $N_1^{(1)} = N_2^{(1)} = N_1^{(2)} = N_2^{(2)}$, and $\alpha = 0.05$

$N_j^{(i)}$	$U(\alpha)$	$u^*(\alpha)$	$U_b(\alpha)$	$U_{bt}(\alpha)$	P_c	P_u	P_b	P_{bt}
$(p_1, p_2, p_3) = (4, 2, 2)$								
20	6.99	9.06	7.36	7.29	0.012	0.015	0.016	0.014
40	8.05	9.28	8.26	8.19	0.024	0.027	0.027	0.026
80	8.70	9.39	8.81	8.76	0.035	0.037	0.037	0.036
160	9.08	9.44	9.13	9.10	0.042	0.043	0.043	0.042
320	9.27	9.46	9.30	9.28	0.046	0.046	0.046	0.046
640	9.38	9.48	9.39	9.39	0.048	0.048	0.048	0.048
1280	9.43	9.48	9.44	9.44	0.049	0.049	0.049	0.049
2560	9.46	9.48	9.46	9.46	0.049	0.049	0.050	0.049
$(p_1, p_2, p_3) = (2, 4, 2)$								
20	9.86	12.79	10.12	9.99	0.013	0.012	0.016	0.013
40	11.03	12.69	11.17	11.05	0.025	0.024	0.027	0.025
80	11.73	12.64	11.81	11.73	0.035	0.035	0.037	0.035
160	12.14	12.62	12.18	12.14	0.042	0.042	0.043	0.042
320	12.37	12.60	12.38	12.36	0.046	0.046	0.046	0.046
640	12.47	12.60	12.48	12.46	0.048	0.048	0.048	0.048
1280	12.53	12.59	12.54	12.53	0.049	0.049	0.049	0.049
2560	12.56	12.59	12.57	12.56	0.050	0.049	0.050	0.049
$(p_1, p_2, p_3) = (2, 2, 4)$								
20	9.29	13.47	9.39	9.29	0.008	0.005	0.008	0.006
40	10.83	13.03	10.88	10.75	0.022	0.018	0.023	0.020
80	11.69	12.81	11.71	11.62	0.035	0.032	0.035	0.033
160	12.13	12.70	12.14	12.08	0.042	0.040	0.042	0.041
320	12.36	12.65	12.37	12.34	0.046	0.045	0.046	0.045
640	12.47	12.62	12.48	12.46	0.048	0.047	0.048	0.048
1280	12.53	12.61	12.53	12.53	0.049	0.049	0.049	0.049
2560	12.56	12.60	12.56	12.56	0.049	0.049	0.049	0.049

Based on Table 5, the overall trend in power remains essentially the same across all three statistics. That is, no substantial difference in power performance is observed among U, U_b, and U_{bt}, indicating that the correction methods do not materially affect the power while improving the approximation accuracy.

4 Numerical Example

In this section, we present a numerical example using the "Wine Quality Data Set"[1] from the UCI Machine Learning Repository to illustrate the proposed test statistics for the two-sample problem and consider two groups: red wine (Group 1) and white wine (Group 2).

We randomly selected $N^{(1)} = 40$ observations for Group 1 and $N^{(2)} = 40$ observations for Group 2. The total dimension $p = 7$ was partitioned into $p_1 = 2$ (density, pH), $p_2 = 3$ (alcohol, sulphates, chlorides), and $p_3 = 2$ (volatile acidity, citric acid). We tested the equality of the mean vectors for the last five variables ($p_2 + p_3 = 5$)

[1] https://archive.ics.uci.edu/dataset/186/wine+quality

Table 3 Upper percentiles and type I errors,
$N_1^{(1)} : N_2^{(1)} : N_1^{(2)} : N_2^{(2)} = 1 : 1 : 2 : 2$, and $\alpha = 0.05$

$N_1^{(1)}$	$U(\alpha)$	$u^*(\alpha)$	$U_b(\alpha)$	$U_{bt}(\alpha)$	P_c	P_u	P_b	P_{bt}
$(p_1, p_2, p_3) = (4, 2, 2)$								
20	7.62	8.79	7.96	7.88	0.019	0.027	0.023	0.021
40	8.46	9.16	8.64	8.58	0.031	0.036	0.034	0.032
80	8.95	9.33	9.04	9.00	0.039	0.042	0.041	0.040
160	9.21	9.41	9.26	9.24	0.044	0.046	0.045	0.045
320	9.35	9.45	9.37	9.36	0.047	0.048	0.048	0.047
640	9.41	9.47	9.43	9.42	0.048	0.049	0.049	0.049
1280	9.45	9.48	9.45	9.45	0.049	0.049	0.049	0.049
$(p_1, p_2, p_3) = (2, 4, 2)$								
20	10.59	12.55	10.84	10.71	0.020	0.021	0.023	0.021
40	11.49	12.58	11.63	11.53	0.032	0.032	0.034	0.032
80	12.01	12.59	12.08	12.02	0.040	0.040	0.041	0.040
160	12.29	12.59	12.32	12.29	0.045	0.045	0.045	0.045
320	12.44	12.59	12.45	12.44	0.047	0.047	0.048	0.047
640	12.52	12.59	12.52	12.52	0.049	0.049	0.049	0.049
1280	12.55	12.59	12.56	12.55	0.049	0.049	0.049	0.049
$(p_1, p_2, p_3) = (2, 2, 4)$								
20	10.19	13.22	10.32	10.19	0.015	0.011	0.016	0.014
40	11.35	12.92	11.41	11.30	0.029	0.026	0.030	0.028
80	11.96	12.76	12.00	11.93	0.039	0.037	0.040	0.038
160	12.27	12.67	12.29	12.25	0.044	0.043	0.045	0.044
320	12.43	12.63	12.44	12.42	0.047	0.046	0.047	0.047
640	12.52	12.61	12.52	12.51	0.049	0.048	0.049	0.049
1280	12.55	12.60	12.55	12.55	0.049	0.049	0.049	0.049

under the assumption that the mean vectors of the first two variables were equal. We applied the Box-Cox transformation to the data to satisfy the assumption of multivariate normality. To simulate a two-step monotone missing data structure, we artificially deleted 10 observations for the third block variables in each sample. Consequently, the sample sizes were $N_1^{(1)} = N_1^{(2)} = 30$ and $N_2^{(1)} = N_2^{(2)} = 10$.

The calculated test statistics are denoted by

$$U = 11.00, \quad U_b = 11.30, \quad U_{bt} = 10.94,$$

and the corresponding critical values for $\alpha = 0.05$ are

$$\chi^2_5(0.05) = 11.07, \quad u^*(0.05) = 11.29.$$

For $\alpha = 0.01$, the critical values are

$$\chi^2_5(0.01) = 15.09, \quad u^*(0.01) = 16.49.$$

For the significance level $\alpha = 0.05$, the original statistic U was compared with both $\chi^2_5(0.05)$ and $u(0.05)$. As $U < \chi^2_5(0.05)$ and $U < u^*(0.05)$, the null hypothesis is

Table 4 Upper percentiles and type I errors, $N_1^{(1)} : N_2^{(1)} : N_1^{(2)} : N_2^{(2)} = 2 : 1 : 2 : 1$, and $\alpha = 0.05$

$N_1^{(1)}$	$U(\alpha)$	$u^*(\alpha)$	$U_b(\alpha)$	$U_{bt}(\alpha)$	P_c	P_u	P_b	P_{bt}
$(p_1, p_2, p_3) = (4, 2, 2)$								
40	7.95	9.24	8.23	8.13	0.023	0.026	0.027	0.024
80	8.64	9.37	8.79	8.72	0.034	0.036	0.036	0.035
160	9.03	9.43	9.11	9.07	0.041	0.042	0.042	0.042
320	9.25	9.46	9.29	9.27	0.045	0.046	0.046	0.046
640	9.37	9.47	9.39	9.38	0.048	0.048	0.048	0.048
1280	9.42	9.48	9.43	9.43	0.049	0.049	0.049	0.049
2560	9.46	9.48	9.46	9.46	0.049	0.049	0.050	0.049
$(p_1, p_2, p_3) = (2, 4, 2)$								
40	10.83	12.79	11.02	10.86	0.023	0.021	0.025	0.022
80	11.62	12.69	11.72	11.61	0.034	0.032	0.035	0.033
160	12.08	12.64	12.13	12.06	0.041	0.040	0.042	0.041
320	12.32	12.62	12.35	12.32	0.045	0.045	0.046	0.045
640	12.45	12.60	12.46	12.44	0.047	0.047	0.048	0.047
1280	12.53	12.60	12.53	12.52	0.049	0.049	0.049	0.049
2560	12.56	12.59	12.56	12.56	0.049	0.049	0.049	0.049
$(p_1, p_2, p_3) = (2, 2, 4)$								
40	10.76	13.24	10.82	10.64	0.021	0.015	0.022	0.018
80	11.64	12.91	11.67	11.54	0.034	0.030	0.034	0.032
160	12.11	12.75	12.12	12.04	0.041	0.039	0.042	0.040
320	12.34	12.67	12.35	12.31	0.046	0.044	0.046	0.045
640	12.47	12.63	12.47	12.45	0.048	0.047	0.048	0.047
1280	12.53	12.61	12.53	12.52	0.049	0.048	0.049	0.049
2560	12.56	12.60	12.56	12.55	0.049	0.049	0.049	0.049

accepted. Similarly, for the Bartlett-type corrected statistic U_{bt}, the null hypothesis is accepted because $U_{bt} < \chi^2_5(0.05)$. However, the Bartlett corrected statistic U_b exceeds the chi-square critical value ($U_b > \chi^2_5(0.05)$), leading to rejection of the null hypothesis.

In contrast, for $\alpha = 0.01$, the null hypothesis is not rejected in all cases, as all statistics are significantly smaller than the corresponding critical values $\chi^2_5(0.01)$ and $u^*(0.01)$.

5 Conclusion

In this study, we investigated the two-sample test for the sub-mean vector under a two-step monotone missing data structure. Specifically, we proposed a new test statistic based on the structure of Rao's U-statistic.

To improve the accuracy of the test in small samples, we derived the asymptotic expansion of the proposed statistic and obtained the Bartlett and Bartlett-type corrected statistics, U_b and U_{bt}. Furthermore, we provided the approximate upper 100α percentiles of the null distribution, and evaluated the performance of the proposed procedures through Monte Carlo simulations.

Finally, the practicality of the proposed method was demonstrated through a numerical example.

Table 5 Simulated power functions of U, U_b, and U_{bt} when $(p_1, p_2, p_3) = (2, 2, 2)$

	$\alpha = 0.05$			$\alpha = 0.01$		
$\mu_{(23)}$	$1-\beta_U$	$1-\beta_{U_b}$	$1-\beta_{U_{bt}}$	$1-\beta_U$	$1-\beta_{U_b}$	$1-\beta_{U_{bt}}$
0	0.050	0.050	0.050	0.010	0.010	0.010
0.05	0.064	0.064	0.064	0.014	0.014	0.014
0.1	0.112	0.112	0.112	0.031	0.031	0.031
0.15	0.205	0.206	0.206	0.072	0.071	0.071
0.2	0.351	0.352	0.352	0.155	0.155	0.155
0.25	0.531	0.532	0.532	0.290	0.290	0.290
0.3	0.711	0.712	0.712	0.471	0.471	0.471
0.35	0.851	0.851	0.852	0.659	0.659	0.659
0.4	0.937	0.937	0.937	0.816	0.816	0.815
0.45	0.979	0.979	0.979	0.918	0.918	0.918
0.5	0.994	0.994	0.994	0.970	0.970	0.970
0.55	0.999	0.999	0.999	0.991	0.991	0.991
0.6	1.000	1.000	1.000	0.998	0.998	0.998
0.65	1.000	1.000	1.000	1.000	1.000	1.000
0.7	1.000	1.000	1.000	1.000	1.000	1.000
0.75	1.000	1.000	1.000	1.000	1.000	1.000

References

[1] Chang. W.-Y., and Richard. D. St. P. (2009). Finite-sample inference with monotone incomplete multivariate normal data I. *J. Multivariate Anal.*, **100**, no. 9, 1883-1899.

[2] Fujikoshi, Y. (2000). Transformations with improved chi-squared approximations. *J. Multivariate Anal.*, **72**, 249-263.

[3] Gupta, A. K., Xu, J., and Fujikoshi, Y. (2006). An asymptotic expansion of the distribution of Rao's U-statistic under a general condition. *J. Multivariate Anal.*, **97**, no. 2, 492–513.

[4] Hosonuma, R., Kawasaki, T., and Seo, T. (2025). On the extension of test statistics for the sub-mean vector under two-step monotone missing data. *Sankhyā B.* **87**, no. 2, 434–467.

[5] Kawasaki, T., Naito, T., and Seo, T. (2019). T^2 type test statistic and simultaneous confidence intervals for two sub-mean vectors. *Int. J. Stat. Probab.*, **9**, 1-8.

[6] Krishnamoorthy, K., and Yu, J. (2012). Multivariate Behrens-Fisher problem with missing data. *J. Multivariate Anal.*, **105**, 141-150.

[7] Rao, C. R. (1949). On some problems arising out of discrimination with multiple characters. *Sankhyā*, **9**, 343-366.

[8] Rencher, A. C., and Christensen, W. F. (2012). *Methods of Multivariate Analysis*, 2nd ed., Hoboken, NJ: Wiley.

[9] Seko, N., Kawasaki, T., and Seo, T. (2011). Testing equality of two mean vectors with two-step monotone missing data. *Amer. J. Math. Management Sci.*, **31**, no. 1-2, 117–135.

[10] Seko, N., Yamazaki, A. and Seo, T. (2012). Tests for mean vector with two-step monotone missing data. *SUT J. Math.*, **48**, 13–38.

[11] Siotani, M., Hayakawa, T. and Fujikoshi, Y. (1985). *Modern Multivariate Statistical Analysis: A Graduate Course and Handbook*. American Science Press Inc., Ohio.

[12] Yu . J.. Krishnamoorthy. K., and Pannala, K. M. (2006) . Two-sample inference for normal mean vectors based on monotone missing data. *SUT J. Math.*, **97**, no. 10, 2162-2176.