Journal of Multivariate Analysis 132 (2014) 183–196

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Equivalence testing of mean vector and covariance matrix for multi-populations under a two-step monotone incomplete sample

Shin-ichi Tsukada

School of Education, Meisei University, 2-1-1 Hodokubo, Hino, Tokyo, 191-8506, Japan

ARTICLE INFO

Article history:
Received 2 April 2013
Available online 3 September 2014

AMS 2000 subject classifications:
62H10
62E20

Keywords:
Mean vector
Covariance matrix
Likelihood ratio criterion
Wald-type criterion
Monotone incomplete data

ABSTRACT

This paper investigates the hypothesis testing of a mean vector and covariance matrix for multi-populations in the context of two-step monotone incomplete data drawn from $N_{p+q}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, a multivariate normal population with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Three null hypotheses are considered, and the likelihood ratio criterion and Wald-type criterion are derived. On the basis of numerical simulations, the test that employs the Wald-type criterion is recommended.

1. Introduction

Missing data often appear in practical data analysis. Therefore, rather than conventional statistical analyses, methods that identify missing data should be employed. These methods have been studied by many authors, which include Anderson and Olkin [3], Srivastava [10], Little [8], Kanda and Fujikoshi [7], and Chang and Richards [4,5].

In this paper, three hypothesis tests are considered in the context of two-step monotone incomplete data drawn from $N_{p+q}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, a $(p + q)$-dimensional multivariate normal population with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. To begin, suppose that the data are composed of N mutually independent observations consisting of a random sample of n complete observations and $N - n$ additional observations on $\boldsymbol{x}$ alone. That is,

$$\begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{y}_1 \end{pmatrix}, \begin{pmatrix} \boldsymbol{x}_2 \\ \boldsymbol{y}_2 \end{pmatrix}, \ldots, \begin{pmatrix} \boldsymbol{x}_n \\ \boldsymbol{y}_n \end{pmatrix}, \begin{pmatrix} \boldsymbol{x}_{n+1} \\ * \end{pmatrix}, \ldots, \begin{pmatrix} \boldsymbol{x}_N \\ * \end{pmatrix} \sim N_{p+q}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{1}$$

where $\boldsymbol{x}$ is a $p \times 1$ vector, $\boldsymbol{y}$ is a $q \times 1$ vector, and the symbol $*$ denotes the missing data. The data in (1) are usually referred to as two-step monotone incomplete data; these data show the simplest pattern with the missing data. The maximum likelihood estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ can be explicitly expressed and are given by Anderson and Olkin [3]. Properties of the maximum likelihood estimator are discussed by Kanda and Fujikoshi [7] and Chang and Richards [4,5].

Hao and Krishnamoorthy [6] considered a hypothesis test for a k-step monotone incomplete sample where the covariance matrix is equal to a specified matrix, and the mean vector and the covariance matrix are equal to a given vector and matrix, respectively. They derived the likelihood ratio criterion and asymptotic null distribution. Provost [9] considered the mutual

E-mail address: tsukada@ed.meisei-u.ac.jp.

http://dx.doi.org/10.1016/j.jmva.2014.08.005
0047-259X/

independence of a covariance matrix for a two-step monotone incomplete sample, and derived the likelihood ratio criterion as well. More recently, Chang and Richards [5] derived the likelihood ratio criterion for the hypothesis testing of a mean vector and covariance matrix for a two-step monotone incomplete sample. In this paper, the following hypothesis tests, where the mean vector and covariance matrix are equivalent for multi-populations under a two-step monotone incomplete sample, are considered:

$$\begin{aligned}
&H_{01}: \boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_g, && && H_{11}: H_{01} \text{ is not valid},\\
&H_{02}: \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_g, && \boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_g, && H_{12}: H_{02} \text{ is not valid},\\
&H_{03}: \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_g, && && H_{13}: H_{03} \text{ is not valid},
\end{aligned}$$

where $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are the population mean vector and population covariance matrix, respectively, of the group i. We derive the likelihood ratio criterion and the Wald-type criterion using the maximum likelihood estimator or the unbiased estimator by Tsukada [11]. These estimators are also used to investigate the size and power of the test.

To begin, preliminary results are described in Section 2. Sections 3–5 discuss hypothesis testing for the null hypotheses H_{01}, H_{02}, and H_{03}, respectively. The necessary criteria for testing the hypotheses are derived in these sections as well. The size and power of the tests that use these criteria are compared by performing numerical simulations. These results are presented in Section 6.

2. Preliminary results

When data are chosen according to (1), the sample means are defined as

$$\bar{\boldsymbol{x}} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i, \qquad \bar{\boldsymbol{x}}_1 = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i, \qquad \bar{\boldsymbol{x}}_2 = \frac{1}{N-n}\sum_{i=n+1}^{N}\boldsymbol{x}_i, \qquad \bar{\boldsymbol{y}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{y}_i, \tag{2}$$

and the corresponding matrices of sums of squares and products are given by

$$\begin{aligned}
&\boldsymbol{W}_{11}^{(1)} = \sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_1)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_1)', \qquad \boldsymbol{W}_{12}^{(1)} = \boldsymbol{W}_{21}^{(1)'} = \sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_1)(\boldsymbol{y}_i - \bar{\boldsymbol{y}})',\\
&\boldsymbol{W}_{22}^{(1)} = \sum_{i=1}^{n}(\boldsymbol{y}_i - \bar{\boldsymbol{y}})(\boldsymbol{y}_i - \bar{\boldsymbol{y}})', \qquad \boldsymbol{W}^{(2)} = \sum_{i=1}^{N}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})' - \boldsymbol{W}_{11}^{(1)}.
\end{aligned} \tag{3}$$

The maximum likelihood estimators $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$, obtained by Anderson [1], are as follows:

$$\begin{aligned}
\hat{\boldsymbol{\mu}}_1 &= \bar{\boldsymbol{x}}, \qquad \hat{\boldsymbol{\mu}}_2 = \bar{\boldsymbol{y}} - (1-\tau)\boldsymbol{W}_{21}^{(1)}\boldsymbol{W}_{11}^{(1)-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2),\\
\hat{\boldsymbol{\Sigma}}_{11} &= \frac{1}{N}\left(\boldsymbol{W}_{11}^{(1)} + \boldsymbol{W}^{(2)}\right), \qquad \hat{\boldsymbol{\Sigma}}_{12} = \hat{\boldsymbol{\Sigma}}_{21}' = \hat{\boldsymbol{\Sigma}}_{11}\boldsymbol{W}_{11}^{(1)-1}\boldsymbol{W}_{12}^{(1)},\\
\hat{\boldsymbol{\Sigma}}_{22} &= \frac{1}{n}\boldsymbol{W}_{22\cdot 1}^{(1)} + \hat{\boldsymbol{\Sigma}}_{21}\hat{\boldsymbol{\Sigma}}_{11}^{-1}\hat{\boldsymbol{\Sigma}}_{12},
\end{aligned} \tag{4}$$

where $\tau = (N-n)/N$. Kanda and Fujikoshi [7] derived the expectation of maximum likelihood estimators $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ and pointed out that $\hat{\boldsymbol{\Sigma}}$ is not unbiased. Tsukada [11] proposed the following unbiased estimator for $\boldsymbol{\Sigma}$,

$$\begin{aligned}
\tilde{\boldsymbol{\Sigma}}_{11} &= \frac{N}{N-1}\hat{\boldsymbol{\Sigma}}_{11}, \qquad \tilde{\boldsymbol{\Sigma}}_{12} = \tilde{\boldsymbol{\Sigma}}_{21}' = \frac{N}{N-1}\hat{\boldsymbol{\Sigma}}_{12},\\
\tilde{\boldsymbol{\Sigma}}_{22} &= \frac{N}{N-1}\hat{\boldsymbol{\Sigma}}_{22} - c_0\hat{\boldsymbol{\Sigma}}_{22\cdot 1},
\end{aligned} \tag{5}$$

and showed that the risk of the unbiased estimator is smaller than that of the maximum likelihood estimator with regard to Stein's loss function, where $\hat{\boldsymbol{\Sigma}}_{22\cdot 1} = \hat{\boldsymbol{\Sigma}}_{22} - \hat{\boldsymbol{\Sigma}}_{21}\hat{\boldsymbol{\Sigma}}_{11}^{-1}\hat{\boldsymbol{\Sigma}}_{12}$ and

$$c_0 = \frac{(N-n)(p+1)(p+2) - n(N-n)}{(N-1)(n-p-2)(n-p-1)}.$$

The asymptotic distribution of the estimators is given by the following theorems.

Theorem 2.1 (*Chang and Richards [4]*). *Suppose $n, N \to \infty$ with $0 < n/N \leq 1$. For $n > q+2$, the asymptotic distribution of $\hat{\boldsymbol{\mu}}$ is*

$$N_{p+q}\left(\boldsymbol{\mu}, \frac{1}{N}\boldsymbol{\Sigma} + \frac{\tau(n-2)}{n(n-p-2)}\begin{pmatrix}\boldsymbol{O} & \boldsymbol{O}\\ \boldsymbol{O} & \boldsymbol{\Sigma}_{22\cdot 1}\end{pmatrix}\right).$$