A Test for Mean Vector and Simultaneous Confidence Intervals with Three-step Monotone Missing Data

Ayaka Yagi and Takashi Seo

Department of Mathematical Information Science
Tokyo University of Science
1-3, Kagurazaka, Shinjuku-ku, Tokyo, 162-8601 Japan

Abstract

In this paper, we consider the problem of testing for mean vector and simultaneous confidence intervals when the data have three-step monotone pattern missing observations. The maximum likelihood estimators of the mean vector and the covariance matrix with a three-step monotone missing data pattern are presented based on the derivation of Jinadasa and Tracy (1992). An approximate upper percentile of Hotelling's T^2 type statistic to test mean vector is proposed. Approximate simultaneous confidence intervals for any and all linear compounds of the mean and testing equality of mean components are also obtained. Finally the accuracy of the approximation is investigated by Monte Carlo simulation.

Key Words and Phrases: T^2 type statistic; Maximum likelihood estimator; Three-step monotone missing data

1. Introduction

We deal with the problem of testing for mean vector with three-step monotone missing data:

$$\begin{pmatrix} x_{11} & \cdots & x_{1p_3} & x_{1,p_3+1} & \cdots & x_{1p_2} & x_{1,p_2+1} & \cdots & x_{1p_1} \\ \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ x_{n_1 1} & \cdots & x_{n_1 p_3} & x_{n_1,p_3+1} & \cdots & x_{n_1 p_2} & x_{n_1,p_2+1} & \cdots & x_{n_1 p_1} \\ x_{n_1+1,1} & \cdots & x_{n_1+1,p_3} & x_{n_1+1,p_3+1} & \cdots & x_{n_1+1,p_2} & * & \cdots & * \\ \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ x_{n_2 1} & \cdots & x_{n_2 p_3} & x_{n_2,p_3+1} & \cdots & x_{n_2 p_2} & * & \cdots & * \\ x_{n_2+1,1} & \cdots & x_{n_2+1,p_3} & * & \cdots & * & * & \cdots & * \\ \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ x_{n_3 1} & \cdots & x_{n_3 p_3} & * & \cdots & * & * & \cdots & * \end{pmatrix},$$

where $p = p_1 > p_2 > p_3 > 0$ and $n_1 > p$, and " $*$ " indicates a missing observation. That is, we have complete data set for n_1 observations with p_1 dimensions and two incomplete data sets which have n_2 observations with p_2 dimensions and n_3 observations with p_3 dimensions. Further $\boldsymbol{x}$ be distributed as $N_p(\boldsymbol{\mu}, \Sigma)$, and let $\boldsymbol{x}_i = (\boldsymbol{x})_i$, be the vector of the first p_i elements of $\boldsymbol{x}$. Then $\boldsymbol{x}_i(= (x_1, x_2, \ldots, x_{p_i})')$ is distributed as $N_{p_i}(\boldsymbol{\mu}_i, \Sigma_i)$, $i = 1, 2, 3$, where $\boldsymbol{\mu}_i = (\boldsymbol{\mu})_i = (\mu_1, \ldots, \mu_{p_i})'$ and Σ_i is the principal submatrix of $\Sigma(= \Sigma_1)$ of order $p_i \times p_i$.

Let $(\Sigma_i)_j$ be the principal submatrix of Σ_i of order $p_j \times p_j$, $1 \leq i < j \leq 3$. We define

$$\Sigma_{i+1} = (\Sigma_1)_{i+1}, \quad \Sigma_1 = \Sigma = \begin{pmatrix} \Sigma_{i+1} & \Sigma_{i+1,2} \\ \Sigma'_{i+1,2} & \Sigma_{i+1,3} \end{pmatrix}$$

and

$$\Sigma_i = \begin{pmatrix} \Sigma_{i+1} & \Sigma_{(i,2)} \\ \Sigma'_{(i,2)} & \Sigma_{(i,3)} \end{pmatrix}, \quad i = 1, 2, 3.$$

For example, we can express Σ_1 as

$$\Sigma_1 = \begin{pmatrix} \overbrace{\Sigma_2}^{p_2} & \overbrace{\Sigma_{22}}^{p_1 - p_2} \\ \Sigma'_{22} & \Sigma_{23} \end{pmatrix} \begin{matrix} \}p_2 \\ \}p_1 - p_2 \end{matrix} \quad \text{or} \quad \Sigma_1 = \begin{pmatrix} \overbrace{\Sigma_3}^{p_3} & \overbrace{\Sigma_{32}}^{p_1 - p_3} \\ \Sigma'_{32} & \Sigma_{33} \end{pmatrix} \begin{matrix} \}p_3 \\ \}p_1 - p_3 \end{matrix} .$$

Also, we have

$$\Sigma_2 = \begin{pmatrix} \overbrace{\Sigma_3}^{p_3} & \overbrace{\Sigma_{(2,2)}}^{p_2 - p_3} \\ \Sigma'_{(2,2)} & \Sigma_{(2,3)} \end{pmatrix} \begin{matrix} \}p_3 \\ \}p_2 - p_3 \end{matrix} ,$$

where Σ_2 is an upper left submatrix of Σ_1.

If $\boldsymbol{x}_{ij}$ denotes the jth observation on $\boldsymbol{x}_i$, then the three-step monotone missing data set is the form of

$$\left(\begin{array}{ccccccc} & & \boldsymbol{x}'_{11} & & & & \\ & & \vdots & & & & \\ & & \boldsymbol{x}'_{1n_1} & & & & \\ \hline & \boldsymbol{x}'_{2,n_1+1} & & & * & \cdots & * \\ & \vdots & & & \vdots & & \vdots \\ & \boldsymbol{x}'_{2n_2} & & & * & \cdots & * \\ \hline \boldsymbol{x}'_{3,n_2+1} & * & \cdots & * & * & \cdots & * \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ \boldsymbol{x}'_{3n_3} & * & \cdots & * & * & \cdots & * \end{array} \right).$$

Such a data set is called a three-step monotone missing data pattern. As for a k-step monotone sample or a k-step monotone missing data pattern, see Bhargava (1962), Srivastava and Carter (1983), Little and Rubin (1987) and Srivastava (2002) and so on.

Jinadasa and Tracy (1992) obtained closed form expressions for the maximum likelihood estimators of the mean vector and the covariance matrix of the multivariate normal distribution in the case of k-step monotone missing data. In particular, Anderson (1957) and Anderson and Olkin (1985) considered the two-step monotone missing data pattern. Kanda and Fujikoshi (1998) discussed the distribution of the MLEs in the case of k-step monotone missing data. In this paper, we consider the problem of testing $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ vs. $H_1 \neq H_0$ when the data have a monotone pattern missing observations. In the case of two-step monotone missing data, Krishnamoorthy and Pannala (1999) gave a Hotelling's T^2 type statistic and its approximate distribution. Chang and Richards (2009) and Seko, Yamazaki and Seo (2012) discussed the Hotelling's T^2 type statistic and gave some properties and its another approximation procedure.

We propose a Hotelling's T^2 type statistic and its approximate upper percentile in the case of three-step monotone missing data, similar to the one in the case of two-step monotone missing data . Our approximation procedure is essentially based on the ones given in Seko, Yamazaki and Seo (2012). In Section 2, we present the maximum likelihood estimators of the mean vector and the covariance matrix with a three-step monotone missing data using the notations and same derivation by Jinadasa and Tracy (1992). These results are simple and useful in order to get a Hotelling's T^2 type statistic and the covariance for the maximum likelihood estimator of the mean vector. In Section 3, we give the T^2 type statistic of testing for mean vector and its approximate upper percentile. In Section 4, we discuss the simultaneous confidence intervals for any and all linear compounds of the mean and testing equality of mean components. We also give some simulation results.

2. MLEs of $\boldsymbol{\mu}$ and Σ

Let the MLEs of $\boldsymbol{\mu}$ and Σ denote by $\widehat{\boldsymbol{\mu}}$ and $\widehat{\Sigma}$. If the data have three-step monotone pattern missing observations, then we have the following theorem by the same derivation of Jinadasa and Tracy (1992).

Theorem 1. *If the data have three-step monotone pattern missing observations, then the maximum likelihood estimator of the mean vector is given by*

$$\widehat{\boldsymbol{\mu}} = \overline{\boldsymbol{x}}_1 + \widehat{T}_2 \boldsymbol{d}_2 + \widehat{T}_2 \widehat{T}_3 \boldsymbol{d}_3,$$

where

$$\overline{\boldsymbol{x}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \boldsymbol{x}_{ij}, \ i = 1, 2, 3,$$

$$\boldsymbol{d}_2 = \frac{n_2}{N_3} \left\{ \overline{\boldsymbol{x}}_2 - (\overline{\boldsymbol{x}}_1)_2 \right\}, \ \boldsymbol{d}_3 = \frac{n_3}{N_4} \left[\overline{\boldsymbol{x}}_3 - \frac{1}{N_3} \left\{ n_1 (\overline{\boldsymbol{x}}_1)_3 + n_2 (\overline{\boldsymbol{x}}_2)_3 \right\} \right],$$

$$\widehat{T}_2 = \begin{pmatrix} I_{p_2} \\ \widehat{\Sigma}'_{(1,2)} \widehat{\Sigma}_2^{-1} \end{pmatrix}, \ \widehat{T}_3 = \begin{pmatrix} I_{p_3} \\ \widehat{\Sigma}'_{(2,2)} \widehat{\Sigma}_3^{-1} \end{pmatrix},$$

$$N_{i+1} = \sum_{j=1}^{i} n_j, \ i = 1, 2, 3,$$

and then the maximum likelihood estimator of the covariance matrix is given by

$$\widehat{\Sigma} = \frac{1}{N_2} E_1 + \frac{1}{N_3} G_2 \left[E_2 + \frac{N_2 N_3}{n_2} \boldsymbol{d}_2 \boldsymbol{d}_2' - \frac{n_2}{N_2} L_{11} \right] G_2'$$
$$+ \frac{1}{N_4} G_2 G_3 \left[E_3 + \frac{N_3 N_4}{n_3} \boldsymbol{d}_3 \boldsymbol{d}_3' - \frac{n_3}{N_3} L_{21} \right] G_3' G_2',$$

where

$$E_i = \sum_{j=1}^{n_i} (\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_i)(\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_i)', \ i = 1, 2, 3,$$

$$G_2 = \begin{pmatrix} I_2 \\ L_{12}' L_{11}^{-1} \end{pmatrix}, \ G_3 = \begin{pmatrix} I_3 \\ L_{22}' L_{21}^{-1} \end{pmatrix},$$

$$L_1 = E_1, \ L_2 = L_{11} + E_2 + \frac{N_2 N_3}{n_2} \boldsymbol{d}_2 \boldsymbol{d}_2',$$

$$L_i = \begin{pmatrix} L_{i1} & L_{i2} \\ L_{i2}' & L_{i3} \end{pmatrix}, \ i = 1, 2.$$

Figure 1 shows the data set with a three-step monotone missing data pattern which are used to calculate $\boldsymbol{d}_2$ and $\boldsymbol{d}_3$, respectively.

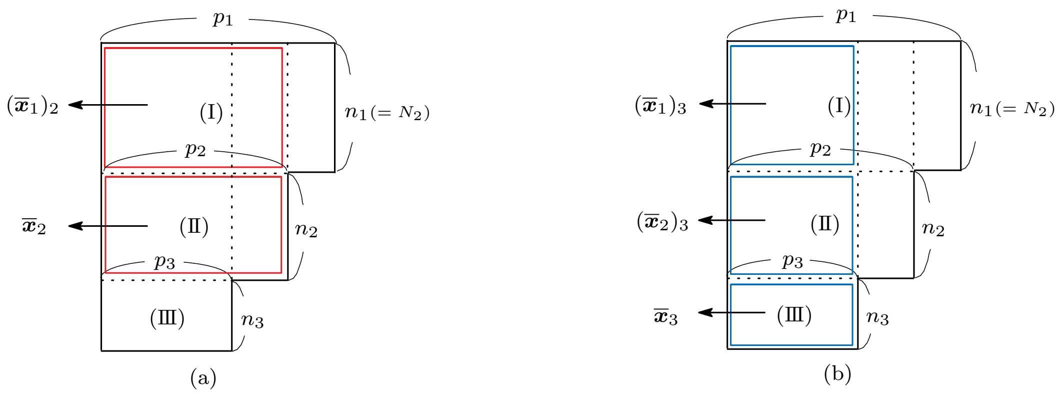

Figure 1: (a) the data to calculate $\boldsymbol{d}_2$ and (b) the data to calculate $\boldsymbol{d}_3$.

The results of MLEs coincide with which Kanda and Fujikoshi (1998) derived by the conditional approach. In this paper, we present the MLEs in the case of three-step monotone missing data in order to get a Hotelling's T^2 type statistic for testing of mean vector.

3. T^2 type statistic

In this section, we consider the following the hypothesis test with three-step monotone missing data:

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 \text{ vs. } H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0,$$

where $\boldsymbol{\mu}_0$ is known. Without loss of generality, we can assume that $\boldsymbol{\mu}_0 = \mathbf{0}$. To test the hypothesis H_0, an usual Hotelling's T^2 type statistic are given by

$$T^2 = \widehat{\boldsymbol{\mu}}' \, \widehat{\Gamma}^{-1} \widehat{\boldsymbol{\mu}},$$

where $\widehat{\boldsymbol{\mu}} = \overline{\boldsymbol{x}}_1 + \widehat{T}_2 \boldsymbol{d}_2 + \widehat{T}_2 \widehat{T}_3 \boldsymbol{d}_3$ and $\widehat{\Gamma}$ is an estimator of $\Gamma = \mathrm{Cov}(\widehat{\boldsymbol{\mu}})$. In order to discuss the distribution of T^2, the covariance matrix of $\widehat{\boldsymbol{\mu}}$, $\mathrm{Cov}(\widehat{\boldsymbol{\mu}})$ is a key quantity. However, we use the Hotelling's T^2 type statistic with $\widehat{\mathrm{Cov}}(\widetilde{\boldsymbol{\mu}})$ instead of $\widehat{\mathrm{Cov}}(\widehat{\boldsymbol{\mu}})$ since $\widehat{\mathrm{Cov}}(\widehat{\boldsymbol{\mu}})$ is complicated and the value of $\widehat{\mathrm{Cov}}(\widetilde{\boldsymbol{\mu}})$ may be almost the same that of $\widehat{\mathrm{Cov}}(\widehat{\boldsymbol{\mu}})$.

Therefore, we adopt that $\widetilde{T}^2 = \widehat{\boldsymbol{\mu}}' \, \widetilde{\Gamma}^{-1} \widehat{\boldsymbol{\mu}}$, where $\widetilde{\Gamma} = \widehat{\mathrm{Cov}}(\widetilde{\boldsymbol{\mu}})$ and $\widetilde{\boldsymbol{\mu}} = \overline{\boldsymbol{x}}_1 + T_2 \boldsymbol{d}_2 + T_2 T_3 \boldsymbol{d}_3$.

Then we have $\mathrm{Cov}(\widetilde{\boldsymbol{\mu}}) = \mathrm{E}[\widetilde{\boldsymbol{\mu}}\widetilde{\boldsymbol{\mu}}'] - \boldsymbol{\mu}\boldsymbol{\mu}'$, where

$$\begin{aligned}\mathrm{E}[\widetilde{\boldsymbol{\mu}}\widetilde{\boldsymbol{\mu}}'] =& \mathrm{E}[(\overline{\boldsymbol{x}}_1 + T_2\boldsymbol{d}_2 + T_2T_3\boldsymbol{d}_3)(\overline{\boldsymbol{x}}_1 + T_2\boldsymbol{d}_2 + T_2T_3\boldsymbol{d}_3)'] \\ =& \mathrm{E}[\overline{\boldsymbol{x}}_1\overline{\boldsymbol{x}}_1' + \overline{\boldsymbol{x}}_1\boldsymbol{d}_2'T_2' + \overline{\boldsymbol{x}}_1\boldsymbol{d}_3'T_3'T_2' + T_2\boldsymbol{d}_2\overline{\boldsymbol{x}}_1' + T_2\boldsymbol{d}_2\boldsymbol{d}_2'T_2' \\ &+ T_2\boldsymbol{d}_2\boldsymbol{d}_3'T_3'T_2' + T_2T_3\boldsymbol{d}_3\overline{\boldsymbol{x}}_1' + T_2T_3\boldsymbol{d}_3\boldsymbol{d}_2'T_2' + T_2T_3\boldsymbol{d}_3\boldsymbol{d}_3'T_3'T_2'].\end{aligned}$$

Further, using the following results,

$$\mathrm{E}[\overline{\boldsymbol{x}}_1\overline{\boldsymbol{x}}_1'] = \frac{1}{n_1}\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}', \quad \mathrm{E}[\overline{\boldsymbol{x}}_1\boldsymbol{d}_2'] = -\frac{n_2}{N_2N_3}\begin{pmatrix}\Sigma_2\\ \Sigma_{22}'\end{pmatrix}, \quad \mathrm{E}[\overline{\boldsymbol{x}}_1\boldsymbol{d}_3'] = -\frac{n_3}{N_3N_4}\begin{pmatrix}\Sigma_3\\ \Sigma_{32}'\end{pmatrix},$$
$$\mathrm{E}[\boldsymbol{d}_2\boldsymbol{d}_2'] = \frac{n_2}{N_2N_3}\Sigma_2, \quad \mathrm{E}[\boldsymbol{d}_2\boldsymbol{d}_3'] = \frac{n_2n_3(n_1-n_2)}{n_1N_3^2N_4}\begin{pmatrix}\Sigma_3\\ \Sigma_{(2,2)}'\end{pmatrix}, \quad \mathrm{E}[\boldsymbol{d}_3\boldsymbol{d}_3'] = \frac{n_3}{N_3N_4}\Sigma_3.$$

As a result, we can obtain

$$\widetilde{\Gamma} = \widehat{\mathrm{Cov}}(\widetilde{\boldsymbol{\mu}}) = \frac{1}{N_2}\widehat{\Sigma}_1 - \frac{n_2}{N_2N_3}\widehat{U} - \frac{(n_1^2 - n_1n_2 + 2n_2^2)n_3}{N_2N_3^2N_4}\widehat{V},$$

where

$$\widehat{U} = \begin{pmatrix}\widehat{\Sigma}_2\\ \widehat{\Sigma}_{22}'\end{pmatrix}\widehat{T}_2', \ \widehat{V} = \begin{pmatrix}\widehat{\Sigma}_3\\ \widehat{\Sigma}_{32}'\end{pmatrix}\widehat{T}_3'\widehat{T}_2'.$$

Therefore, we can provide T^2 type statistic. We note that, under H_0, the T^2 type statistic is asymptotically distributed as χ^2 distribution with p degrees of freedom when $n_1, N_4 \to \infty$ with $n_1/N_4 \to \delta \in (0, 1]$. However, it is noted that χ^2 approximation is not a good approximation to the upper percentile of the T^2 type statistic when the sample is not large. Using the same idea for two-step monotone missing data in Seko, Yamazaki and Seo (2012), we propose the approximate upper percentile of $\widetilde{T}^2$ statistic since it is difficult to find the exact upper percentiles of $\widetilde{T}^2$ statistic.

Theorem 2. *Suppose that the data have three-step monotone pattern missing observations. Then the approximate upper 100α percentile of the $\widetilde{T}^2$ statistic is given by*

$$t_{YS}^2 = T_{n_1,\alpha}^2 - \frac{n_2p_2 + n_3p_3}{(n_2+n_3)p_1}(T_{n_1,\alpha}^2 - T_{N_4,\alpha}^2),$$

where

$$T_{N_4,\alpha}^2 = \frac{(N_4-1)p_1}{N_4-p_1}F_{p_1,N_4-p_1,\alpha}, \quad T_{n_1,\alpha}^2 = \frac{(n_1-1)p_1}{n_1-p_1}F_{p_1,n_1-p_1,\alpha}$$

and $F_{p,q,\alpha}$ is the upper 100α percentile of the F distribution with p and q degrees of freedom.

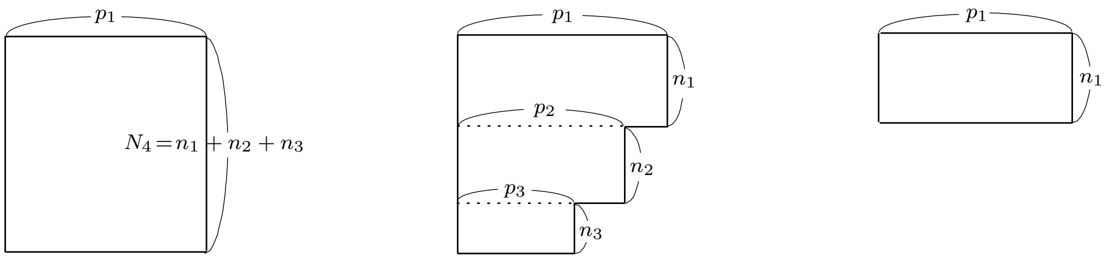

Figure 2: Approximation to Upper Percentiles of T^2

Figure 2 shows that the upper percentiles of T^2 may be between $T^2_{N_4,\alpha}$ and $T^2_{n_1,\alpha}$, where $T^2_{N_4,\alpha}$ and $T^2_{n_1,\alpha}$ are calculated from $N_4 \times p_1$ complete data set (left side) and $n_1 \times p_1$ complete data set (right side), respectively.

4. Simultaneous confidence intervals and testing equality of mean components

We consider the simultaneous confidence intervals for any and all linear compounds of the mean when the data have three-step monotone missing observations. Using the approximate upper percentiles of $\widetilde{T}^2$ in Section 3, for any nonnull vector $\boldsymbol{a} = (a_1, \ldots, a_p)'$, the approximate simultaneous confidence intervals for $\boldsymbol{a}'\boldsymbol{\mu}$ are given by

$$\boldsymbol{a}'\widehat{\boldsymbol{\mu}} - \sqrt{\boldsymbol{a}'\widetilde{\Gamma}\boldsymbol{a}\, t^2_{YS}} \le \boldsymbol{a}'\boldsymbol{\mu} \le \boldsymbol{a}'\widehat{\boldsymbol{\mu}} + \sqrt{\boldsymbol{a}'\widetilde{\Gamma}\boldsymbol{a}\, t^2_{YS}}, \quad \forall \boldsymbol{a} \in \boldsymbol{R}^p - \{\boldsymbol{0}\}.$$

As for a two-step monotone missing data, see Seko, Yamazaki and Seo (2012). Further we consider the testing equality of mean components for the case of a three-step monotone missing data, that is

$$H_0 : \mu_1 = \cdots = \mu_p \text{ vs. } H_1 \ne H_0.$$

In this case, let $\boldsymbol{y}_{ij} = C_i \boldsymbol{x}_{ij}, i = 1, 2, 3, j = 1, 2, \ldots, n_i$, where C_i is a $(p_i - 1) \times p_i$ matrix such that $C_i \boldsymbol{1} = \boldsymbol{0}$ and $C_i C_i' = I_{p_i - 1}$, then $\boldsymbol{y}_{ij}$'s are distributed as $N_{p_i-1}(C_i \boldsymbol{\mu}_i,\ I_{p_i-1})$ since, without loss of generality, we may assume that $\Sigma = I$ when we consider the T^2 type statistic with a monotone missing data. Hence the Hotelling's T^2 type statistic is given by

$$\widetilde{T}^2_c = \widehat{\boldsymbol{\mu}}^{*\prime} \widetilde{\Gamma}^{*-1} \widehat{\boldsymbol{\mu}}^*,$$

where $\widehat{\boldsymbol{\mu}}^*$ is the MLE of $\boldsymbol{\mu}^* = C_1\boldsymbol{\mu}$ and $\widetilde{\Gamma}^* = \widehat{\mathrm{Cov}}(\widetilde{\boldsymbol{\mu}}^*)$, $\widetilde{\boldsymbol{\mu}}^* = C\widetilde{\boldsymbol{\mu}}$. Therefore, essentially, using the same values of t^2_{YS} derived in Theorem 2 for the approximate upper percentile of $\widetilde{T}^2_c$ statistic, we can test for equality of mean components with a three-step monotone missing data.

5. Simulation studies

We compute the upper percentiles of the T^2 type statistic with a three-step monotone missing data by Monte Carlo simulation. The Monte Carlo simulation was 1,000,000 simulations for selected values of $p_i, n_i, i = 1, 2, 3$ and α. It is interesting to see how the approximations are close to the exact upper percentiles. Relating to this problem, some simulation results are given in Tables 1 $\sim$ 3. Computations are made for two cases ;

$$\begin{aligned}\text{Case I }:\ & (p_1, p_2, p_3) = (6, 4, 2),\ (12, 8, 4),\\ & n_1 = 30, 50, 100, 200, 300,\\ & n_2, n_3 = 10, 20,\\ & \alpha = 0.01, 0.05,\end{aligned}$$

where the sets of (n_1, n_2, n_3) are all combinations of n_1, n_2 and n_3.

$$\begin{aligned}\text{Case II }:\ & (p_1, p_2, p_3) = (12, 8, 4),\\ & (n_1, n_2, n_3) = (30m, 10m, 10m), \quad m = 1(1)5, 8, 12,\\ & \alpha = 0.01, 0.05.\end{aligned}$$

Tables 1 and 2 present the simulated upper percentiles of T^2, t^2_{simu}, the approximate upper percentiles of T^2, t^2_{YS}, and the upper percentiles of χ^2 distribution with p degrees of freedom, χ^2_p for Case I. It may be noted from Tables 1 and 2 that the simulated values are closer to the upper percentiles of χ^2 distribution when the sample size n_1 becomes large. Therefore, we note that the chi-squared approximation χ^2_p is not good for cases when n_i is small. However, it is seen that the proposed approximation t^2_{YS} is very good even for cases when n_1 is not large. Tables 1 and 2 also present the simulated coverage probabilities for t^2_{YS} and χ^2_p for Case I. It may be noted from Tables 1 and 2 that the simulated coverage probabilities for $t^2_{\rm YS}$, ${\rm CP}(t^2_{\rm YS})$ are very close to the nominal level $1-\alpha$ even for cases when n_1 is small. It conclusion, even for small samples, our approximation procedure is very accurate.

Tables 3 presents the values of $t^2_{\rm simu}$, $t^2_{\rm YS}$ and χ^2_p, respectively for Case II. Also, the values of ${\rm CP}(t^2_{\rm YS})$ and ${\rm CP}(\chi^2_p)$ for Case II are given in Table 3. We note that the missing rate of the data sets in Case II is a constant $(=0.2)$, where

$$\text{the missing rate} = \frac{N_4 p_1}{\displaystyle\sum_{i=1}^{3} n_i p_i} - 1.$$

It seems from Table 3 that our approximation is very good even when (n_1, n_2, n_3) is small. In conclusion, we have developed the approximate upper percentiles of T^2 type statistic for the test of mean vector in case of a three-step monotone missing data. The proposed approximate values can be calculated easily and the approximation is much better than the χ^2 approximation even when the sample size is small.

Table 1: The simulated and the approximate values for T^2, χ^2 approximation, and the simulated coverage probabilities when $(p_1, p_2, p_3) = (6, 4, 2)$

Sample size			Upper percentile			Coverage probability	
n_1	n_2	n_3	t^2_{simu}	t^2_{YS}	χ^2_p	CP(t^2_{YS})	CP(χ^2_p)
			$\alpha = 0.05$				
30	10	10	17.32	16.82	12.59	0.944	0.859
50	10	10	15.17	14.99	12.59	0.947	0.900
100	10	10	13.85	13.76	12.59	0.949	0.926
200	10	10	13.23	13.18	12.59	0.949	0.938
300	10	10	13.02	12.98	12.59	0.949	0.942
30	20	10	17.02	16.36	12.59	0.942	0.864
50	20	10	14.95	14.78	12.59	0.948	0.905
100	20	10	13.76	13.70	12.59	0.949	0.928
200	20	10	13.21	13.16	12.59	0.949	0.939
300	20	10	13.00	12.97	12.59	0.950	0.942
30	10	20	16.95	16.72	12.59	0.947	0.868
50	10	20	14.95	14.92	12.59	0.950	0.905
100	10	20	13.75	13.74	12.59	0.950	0.928
200	10	20	13.19	13.17	12.59	0.950	0.939
300	10	20	13.01	12.98	12.59	0.950	0.942
30	20	20	16.74	16.35	12.59	0.945	0.870
50	20	20	14.74	14.75	12.59	0.950	0.909
100	20	20	13.68	13.68	12.59	0.950	0.930
200	20	20	13.18	13.15	12.59	0.949	0.939
300	20	20	12.99	12.97	12.59	0.950	0.943
			$\alpha = 0.01$				
30	10	10	25.01	24.13	16.81	0.988	0.944
50	10	10	21.11	20.85	16.81	0.989	0.968
100	10	10	18.88	18.76	16.81	0.990	0.980
200	10	10	17.85	17.78	16.81	0.990	0.986
300	10	10	17.50	17.45	16.81	0.990	0.987
30	20	10	24.30	23.31	16.81	0.988	0.948
50	20	10	20.73	20.51	16.81	0.989	0.970
100	20	10	18.75	18.66	16.81	0.990	0.981
200	20	10	17.77	17.75	16.81	0.990	0.986
300	20	10	17.48	17.44	16.81	0.990	0.987
30	10	20	24.54	23.96	16.81	0.989	0.948
50	10	20	20.82	20.74	16.81	0.990	0.970
100	10	20	18.68	18.72	16.81	0.990	0.981
200	10	20	17.80	17.76	16.81	0.990	0.986
300	10	20	17.49	17.45	16.81	0.990	0.987
30	20	20	24.14	23.31	16.81	0.988	0.951
50	20	20	20.40	20.45	16.81	0.990	0.972
100	20	20	18.63	18.63	16.81	0.990	0.982
200	20	20	17.86	17.74	16.81	0.990	0.986
300	20	20	17.44	17.43	16.81	0.990	0.987

Table 2: The simulated and the approximate values for T^2, χ^2 approximation, and the simulated coverage probabilities when $(p_1, p_2, p_3) = (12, 8, 4)$

Sample size			Upper percentile			Coverage probability	
n_1	n_2	n_3	t^2_{simu}	t^2_{YS}	χ^2_p	CP(t^2_{YS})	CP(χ^2_p)
			$\alpha = 0.05$				
30	10	10	42.20	38.25	21.03	0.927	0.611
50	10	10	30.25	29.34	21.03	0.942	0.787
100	10	10	24.98	24.76	21.03	0.948	0.883
200	10	10	22.94	22.82	21.03	0.948	0.920
300	10	10	22.25	22.21	21.03	0.949	0.931
30	20	10	40.81	36.19	21.03	0.920	0.629
50	20	10	29.56	28.57	21.03	0.941	0.799
100	20	10	24.75	24.56	21.03	0.948	0.887
200	20	10	22.85	22.77	21.03	0.949	0.921
300	20	10	22.24	22.18	21.03	0.949	0.932
30	10	20	41.45	38.01	21.03	0.930	0.627
50	10	20	29.82	29.10	21.03	0.944	0.796
100	10	20	24.84	24.68	21.03	0.948	0.886
200	10	20	22.90	22.80	21.03	0.949	0.921
300	10	20	22.26	22.20	21.03	0.949	0.931
30	20	20	40.44	36.37	21.03	0.925	0.638
50	20	20	29.18	28.47	21.03	0.944	0.807
100	20	20	24.59	24.50	21.03	0.949	0.891
200	20	20	22.81	22.75	21.03	0.949	0.922
300	20	20	22.22	22.17	21.03	0.949	0.932
			$\alpha = 0.01$				
30	10	10	60.29	53.41	26.22	0.982	0.763
50	10	10	40.07	38.73	26.22	0.988	0.904
100	10	10	32.03	31.72	26.22	0.989	0.962
200	10	10	28.93	28.83	26.22	0.990	0.979
300	10	10	28.03	27.93	26.22	0.990	0.983
30	10	20	59.45	53.11	26.22	0.983	0.775
50	10	20	39.53	38.37	26.22	0.988	0.910
100	10	20	31.76	31.59	26.22	0.990	0.963
200	10	20	28.86	28.80	26.22	0.990	0.979
300	10	20	27.97	27.92	26.22	0.990	0.983
30	20	10	58.35	50.10	26.22	0.979	0.779
50	20	10	39.03	37.55	26.22	0.987	0.913
100	20	10	31.77	31.41	26.22	0.989	0.964
200	20	10	28.84	28.75	26.22	0.990	0.979
300	20	10	27.95	27.90	26.22	0.990	0.983
30	20	20	57.88	50.48	26.22	0.981	0.786
50	20	20	38.58	37.41	26.22	0.988	0.917
100	20	20	31.49	31.31	26.22	0.990	0.966
200	20	20	28.75	28.72	26.22	0.990	0.980
300	20	20	27.87	27.88	26.22	0.990	0.984

Table 3: The simulated and the approximate values for T^2, χ^2 approximation, and the simulated coverage probabilities when $(p_1, p_2, p_3) = (12, 8, 4)$

Sample size			Upper percentile			Coverage probability	
n_1	n_2	n_3	t^2_{simu}	t^2_{YS}	χ^2_p	CP(t^2_{YS})	CP(χ^2_p)
			$\alpha = 0.05$				
30	10	10	42.20	38.25	21.03	0.927	0.611
60	20	20	27.43	27.04	21.03	0.946	0.839
90	30	30	24.62	24.67	21.03	0.951	0.890
120	40	40	23.43	23.64	21.03	0.953	0.912
150	50	50	22.79	23.06	21.03	0.953	0.923
240	80	80	21.83	22.25	21.03	0.955	0.938
360	120	120	21.38	21.83	21.03	0.956	0.945
			$\alpha = 0.01$				
30	10	10	60.29	53.41	26.22	0.982	0.763
60	20	20	35.75	35.18	26.22	0.989	0.937
90	30	30	31.55	31.58	26.22	0.990	0.965
120	40	40	29.72	30.04	26.22	0.991	0.975
150	50	50	28.83	29.19	26.22	0.991	0.980
240	80	80	27.43	28.00	26.22	0.992	0.985
360	120	120	26.82	27.38	26.22	0.992	0.988

Acknowledgements

The second author's research was in part supported by Grant-Aid for Scientific Research (C) (23500360).

References

[1] Anderson, T. W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association*, **52**, 200–203.

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

[3] Chang, W. -Y. and Richards, D. St. P. (2009). Finite-sample inference with monotone incomplete multivariate normal data, I. *Journal of Multivariate Analysis*, **100**, 1883–1899.

[4] Jinadasa, K. G. and Tracy, D. S. (1992). Maximum likelihood estimation for multivariate normal distribution with monotone sample. *Communication in Statistics – Theory and Methods*, **21**, 41–50.

[5] Kanda, T. and Fujikoshi, Y. (1998). Some basic properties of the MLE's for a multivariate normal distribution with monotone missing data. *American Journal of Mathematical and Management Sciences*, **18**, 161–190.

[6] Krishnamoorthy, K. and Pannala, M. K. (1999). Confidence estimation of a normal mean vector with incomplete data. *Canadian Journal of Statistics*, **27**, 395–407.

[7] Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data, 2nd ed.*, Wiley, New York.

[8] Seko, N., Yamazaki, A. and Seo, T. (2012). Tests for mean vector with two-step monotone missing data. *SUT Journal of Mathematics*, **48**, 13–36.

[9] Srivastava, M. S. (2002). *Methods of Multivariate Statistics.* Wiley, New York.

[10] Srivastava, M. S. and Carter, E. M. (1983). *An Introduction to Applied Multivariate Statistics*, North-Holland, New York.