Testing equality of two mean vectors and simultaneous confidence intervals in repeated measures with missing data

Kazuyuki Koizumi

Department of Mathematics,
Graduate School of Science,
Tokyo University of Science,
1-3, Kagurazaka, Shinjuku-ku,
Tokyo 162-8601, Japan.

Takashi Seo*

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 the exact test statistic for equality of two mean vectors in the intraclass correlation model with monotone missing data. Simultaneous confidence intervals for all contrasts of two mean vectors are derived by using the idea in Seo and Srivastava [6]. Finally, to evaluate the procedure proposed by in this paper, we investigate the power of a test statistic and the width of simultaneous confidence intervals.

AMS 2000 Mathematics Subject Classification. primary 62J15; secondary 62H15

Key Words and Phrases: Intraclass correlation model; Monotone missing data; Simultaneous confidence intervals for all contrasts; Two-sample problem; Power; Probability density function; Hotelling's T^2 statistic; Central F distribution; Non-central F distribution.

1 Introduction

In statistical data analysis, we frequently face the matrix of observations with missing values. EM algorithm by Dempster et al. [1] is a very general iterative algorithm for maximum likelihood estimation in incomplete-data problems. Since 1977, there have been many new uses of EM algorithm, as well as further work on its convergence properties (e.g., McLachran and Krishnan [4]). The procedure to obtain maximum likelihood estimates from the likelihood equation with missing observations by Newton-Raphson method is discussed in Srivastava [7].

However both EM algorithm and Srivastava's method are approximate procedures. In the case of one sample problem, Seo and Srivastava [6] has dirived the exact test statistic for the equality of mean components and the simulatanous confidence intervals for all contrasts in the intraclass correlation model with monotone missing data. When

*corresponding authors. *E-mail addresses:* j1106702@ed.kagu.tus.ac.jp (K. Koizumi), seo@rs.kagu.tus.ac.jp (T. Seo)

the missing observations are of the non-monotone type, they also have given the asymptotic simultaneous confidence intervals by usual maximum likelihood ratio method and an iterative numerical method in Srivastava [7] and Srivastava and Carter [8]. The maximum likelihood estimator for an intraclass correlation coefficient in a bivariate normal distribution, when some observations on either of the variables are missing, has been discussed by Konishi and Shimizu [3] and Minami and Shimizu [5].

In this paper, we discuss an extention to the procedure proposed by Seo and Srivastava [6]. We consider the testing for equality and the simultaneous confidence intervals of two mean vectors in repeated measures with monotone missing data. When the observations are complete, it is well known that Hotelling's T^2-statistic (see, Hotelling [2]) is used as the usual test statistic. Recently, when the missing observations are occur, Yu et al. [9] developed a pivotal quantity based on maximum likelihood estimators, and derived its approximate distribution to make inferential procedures. First, in Section 2, we give an unbiased estimation of unknown parameters σ^2 and ρ using the idea in Seo and Srivastava [6] in the case of one-sample problem. In Section 3, we derive the exact test statistic and the simultaneous confidence intervals which are an extension of the results in Seo and Srivastava [6]. In Section 4, to evaluate our procedure, the power of the test statistic are presented. Finally, in Section 5, we investigate the width of our simultaneous confidence intervals by numerical examinations.

2 Estimation of parameters

Let $\boldsymbol{x}_1^{(i)}, \boldsymbol{x}_2^{(i)}, \ldots, \boldsymbol{x}_{n^{(i)}}^{(i)}$ be the sample vectors from the i-th population ($i = 1, 2$). Also, we assume that

$$\boldsymbol{x}_1^{(i)}, \boldsymbol{x}_2^{(i)}, \ldots, \boldsymbol{x}_{n^{(i)}}^{(i)} \overset{i.i.d.}{\sim} N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}^{(i)}),$$

where $\boldsymbol{\mu}_i = (\mu_1^{(i)}, \mu_2^{(i)}, \ldots, \mu_p^{(i)})'$. Further we assume that $\boldsymbol{\Sigma}^{(i)} = \boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}$ has the uniform covariance structure, that is,

$$\boldsymbol{\Sigma} = \sigma^2 \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & & \vdots \\ \vdots & & \ddots & \rho \\ \rho & \cdots & \rho & 1 \end{bmatrix} = \sigma^2[(1-\rho)\boldsymbol{I}_p + \rho \mathbf{1}_p \mathbf{1}_p'],$$

where $\boldsymbol{I}_p$ is a $p \times p$ identity matrix, $\mathbf{1}_p = (1, 1, \ldots, 1)'$ is a p-vector and $\sigma^2,\ \rho \in [-1/(p-1), 1]$ are unknown parameters.

We consider the case when the missing observations are of the monotone-type. Our observations $\{x_{\ell j}^{(i)}\}$ can be written, without loss of generality, in the following form:

$$\begin{bmatrix} x_{11}^{(i)} & x_{12}^{(i)} & \cdots & \cdots & x_{1n_1^{(i)}}^{(i)} \\ \vdots & \vdots & \cdots & x_{2n_2^{(i)}}^{(i)} & * \\ \vdots & \vdots & \cdots & * & * \\ x_{p1}^{(i)} & \cdots & x_{pn_p^{(i)}}^{(i)} & * & * \end{bmatrix},$$

where $n_1^{(i)} \geq n_2^{(i)} \geq \cdots \geq n_p^{(i)}$ and "$*$" means missing component. We shall consider the case $n_1^{(i)} = n_2^{(i)}$ $(\equiv n^{(i)})$. To provide a test or simultaneous confidence intervals, we rewrite the observations in the following form:

$$\begin{bmatrix} x_{11}^{(i)} & x_{12}^{(i)} & \cdots & x_{1n^{(i)}}^{(i)} \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & x_{p_n^{(i)}n^{(i)}}^{(i)} \\ \vdots & x_{p_2^{(i)}2}^{(i)} & * & * \\ x_{p_1^{(i)}1}^{(i)} & * & * & * \end{bmatrix},$$

where $p \equiv p_1^{(i)} \geq p_2^{(i)} \geq \cdots \geq p_n^{(i)}$. Since $n_1^{(i)} = n_2^{(i)}$, we note that $p_n^{(i)} \geq 2$.

Writing $\boldsymbol{x}_j^{(i)} = (x_{1j}^{(i)}, x_{2j}^{(i)}, \ldots, x_{p_j^{(i)}j}^{(i)})'$, we find that

$$\boldsymbol{x}_j^{(i)} \overset{i.d.}{\sim} N_{p_j^{(i)}}(\boldsymbol{\mu}_{ij},\ \boldsymbol{\Sigma}_j),\ j = 1, 2, \ldots, n^{(i)},$$

where $\boldsymbol{\mu}_{ij} = (\mu_1^{(i)}, \mu_2^{(i)}, \ldots, \mu_{p_j^{(i)}}^{(i)})'$ and $\boldsymbol{\Sigma}_j = \sigma^2[(1-\rho)\boldsymbol{I}_{p_j^{(i)}} + \rho \mathbf{1}_{p_j^{(i)}} \mathbf{1}'_{p_j^{(i)}}]$. Let $\boldsymbol{C}_j^{(i)}$ be a $(p_j^{(i)} - 1) \times p_j^{(i)}$ matrix such that $\boldsymbol{C}_j^{(i)} \boldsymbol{C}_j^{(i)'} = \boldsymbol{I}_{p_j^{(i)}-1}$ and $\boldsymbol{C}_j^{(i)} \mathbf{1}_{p_j^{(i)}} = \mathbf{0}$. Clearly, then

$$\boldsymbol{y}_j^{(i)} = \boldsymbol{C}_j^{(i)} \boldsymbol{x}_j^{(i)} \sim N_{p_j^{(i)}-1}(\boldsymbol{C}_j^{(i)} \boldsymbol{\mu}_{ij}, \gamma^2 \boldsymbol{I}_{p_j^{(i)}-1}),$$

where $\gamma^2 \equiv \sigma^2(1-\rho)$. We shall write $\boldsymbol{C} : (p-1) \times p$ for $\boldsymbol{C}_1^{(i)}$, since $p_1^{(i)} \equiv p$. $\boldsymbol{C}$ has the following form:

$$\boldsymbol{C} = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & \cdots & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} & \cdots & 0 \\ \vdots & & \vdots & & \vdots \\ \frac{1}{\sqrt{p(p-1)}} & \frac{1}{\sqrt{p(p-1)}} & \frac{1}{\sqrt{p(p-1)}} & \cdots & -\frac{p-1}{\sqrt{p(p-1)}} \end{bmatrix}.$$

And put $n_{\ell+1} = \min_{i=1,2}\{n_{\ell+1}^{(i)}\}, \ell = 1, 2, \ldots, p-1$, then sample means of transformation data are given by

$$\overline{y}_{\ell\cdot}^{(i)} = \frac{1}{n_{\ell+1}} \sum_{j=1}^{n_{\ell+1}} y_{\ell j}^{(i)},$$

and an unbiased estimator of γ^2 is given by

$$f\widehat{\gamma}^2 = \sum_{i=1}^{2} \sum_{\ell=1}^{p-1} \sum_{j=1}^{n_{\ell+1}} \left(y_{\ell j}^{(i)} - \overline{y}_{\ell\cdot}^{(i)} \right)^2, f = 2\left(\sum_{\ell=1}^{p-1} n_{\ell+1} - p + 1 \right).$$

We note that $f\widehat{\gamma}^2/\gamma^2$ is distributed as χ^2 distribution with f degrees of freedom.

3 An exact test statistic and simultaneous confidence intervals

Let sample mean vectors for the i-th population be $\overline{\boldsymbol{y}}^{(i)} = (\overline{y}_{1\cdot}^{(i)}, \overline{y}_{2\cdot}^{(i)}, \ldots, \overline{y}_{p-1\cdot}^{(i)})'$, then we can obtain that

$$\begin{aligned}
\mathrm{E}(\overline{\boldsymbol{y}}^{(1)} - \overline{\boldsymbol{y}}^{(2)}) &= \boldsymbol{C}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \equiv \boldsymbol{C}\boldsymbol{\delta}_{12}, \\
\mathrm{Cov}(\overline{\boldsymbol{y}}^{(1)} - \overline{\boldsymbol{y}}^{(2)}) &= 2\gamma^2 \begin{bmatrix} n_2^{-1} & & \boldsymbol{0} \\ & \ddots & \\ \boldsymbol{0} & & n_p^{-1} \end{bmatrix} \equiv 2\gamma^2 \boldsymbol{V},
\end{aligned}$$

Therefore an exact test statistic T_{12}^2 is given by

$$T_{12}^2 \equiv \frac{(\overline{\boldsymbol{y}}^{(1)} - \overline{\boldsymbol{y}}^{(2)})' \boldsymbol{V}^{-1} (\overline{\boldsymbol{y}}^{(1)} - \overline{\boldsymbol{y}}^{(2)})}{2\widehat{\gamma}^2}. \tag{3.1}$$

$T_{12}^2/[2(p-1)]$ is distributed as F distribution with $p-1$ and f degrees of freedom under the null hypothesis.

Next, we consider the simultaneous confidence intervals of $\boldsymbol{a}'\boldsymbol{\delta}_{12}$ for $\boldsymbol{a} \in \mathbb{R}_p^* \equiv \mathbb{R}^p - \{\boldsymbol{0}\}$, $\boldsymbol{a}'\boldsymbol{1}_p = 0$. Since $\boldsymbol{a}'\boldsymbol{1}_p = 0$, we can choose $\widetilde{\boldsymbol{a}} \in \mathbb{R}_{p-1}^*$ such that $\boldsymbol{a}' = \widetilde{\boldsymbol{a}}'\boldsymbol{C}$. Therefore we can obtain the simultaneous confidence intervals of $\boldsymbol{a}'\boldsymbol{\delta}_{12}$ with simultaneous confidence level $1-\alpha$

$$\boldsymbol{a}'\boldsymbol{\delta}_{12} \in \left[\widetilde{\boldsymbol{a}}'(\overline{\boldsymbol{y}}^{(1)} - \overline{\boldsymbol{y}}^{(2)}) \pm \sqrt{E} \right], \ \forall \boldsymbol{a} \in \mathbb{R}_p^*, \tag{3.2}$$

where $E = 2t^2(\alpha)\widehat{\gamma}^2 \widetilde{\boldsymbol{a}}'\boldsymbol{V}\widetilde{\boldsymbol{a}}$ and $t^2(\alpha)$ is the upper 100α percentage point of T_{12}^2 statistic in (3.1).

4 Power of the test statistic

The power of a test statistic β is given by

$$\Pr\left(T_{12}^2 > t^2(\alpha) \mid \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2\right) = \beta. \tag{4.1}$$

Since T_{12}^2 statistic in (3.1) is essentially distributed as central F distribution under the null hypothesis, the distribution of T_{12}^2 in (3.1) under the alternative hypotheses is non-central F distribution with $p-1$ and f degrees of freedom and non-centrality parameter ξ^2. Non-centrality parameter ξ is given by

$$\xi = \sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{C}'(2\widehat{\gamma}^2\boldsymbol{V})^{-1}\boldsymbol{C}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}.$$

Therefore we can obtain the power by integrating probability density function of non-central F distribution. Put $\sigma^2 = 1,\ \rho = 0.5,\ \alpha = 0.05$, then the five patterns (1) $\sim$ (5) are as follows:

(1) complete case: $n_1 = n_2 = n_3 = n_4 = 40$

(2) incomplete case: $n_1 = n_2 = 40, n_3 = 30, n_4 = 20,\ |\mu_1^{(1)} - \mu_1^{(2)}| \neq 0$

(3) incomplete case: $n_1 = n_2 = 40, n_3 = 30, n_4 = 20,\ |\mu_3^{(1)} - \mu_3^{(2)}| \neq 0$

(4) incomplete case: $n_1 = n_2 = 40, n_3 = 30, n_4 = 20$, $|\mu_4^{(1)} - \mu_4^{(2)}| \neq 0$

(5) complete case: $n_1 = n_2 = n_3 = n_4 = 20$ (ignoring missing part of (2) $\sim$ (4)).

Five cases of (1) $\sim$ (5) are compared. Figure 1 presents the result of these calculations. The power in (4.1) of T_{12}^2 statistic in (3.1) is larger than the one of the earlier procedure

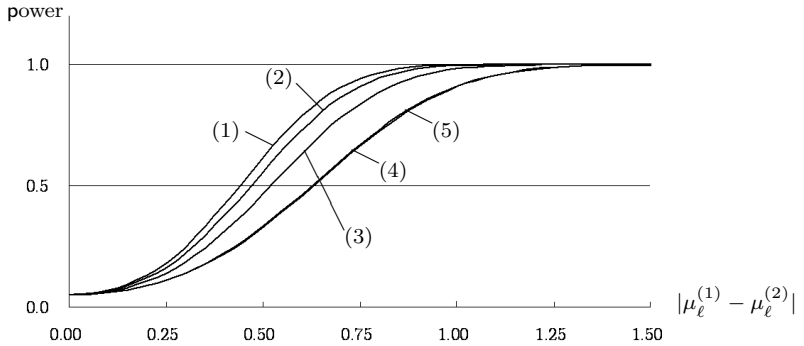

Figure 1. Power of the test statistic

ignoring missing part. In particular, when the first components of two mean vectors having the most large sample sizes are not the same, the power of the procedure proposed by in this paper is good. Even for the forth components the power of T_{12}^2 statistic in (3.1) is larger than the one of the earlier procedure.

5 Numerical example

Finally, to evaluate the procedure proposed by in this paper, we compare the width of simultaneous confidence intervals in (3.2). Parameters are as follows:

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \ \boldsymbol{\mu}_2 = \begin{pmatrix} m_1 \\ m_2 \\ m_3 \\ m_4 \end{pmatrix}, \ \sigma^2 = 1,\ 4,\ 9,\ \rho = 0.1,\ 0.5,\ 0.9,$$

where m_ℓ ($\ell = 1, 2, 3, 4$) are constants. At first, the three cases of (i), (ii) and (iii) are compared. (i), (ii) and (iii) are as follows:

(i) $n_1 = n_2 = n_3 = n_4 = 20$ (complete case)

$$\begin{bmatrix} x_{11}^{(i)} & x_{12}^{(i)} & \cdots & x_{1,20}^{(i)} \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ x_{41}^{(i)} & x_{42}^{(i)} & \cdots & x_{4,20}^{(i)} \end{bmatrix}.$$

(ii) $n_1 = n_2 = 40,\ n_3 = 30,\ n_4 = 20$ (incomplete case)

$$\begin{bmatrix} x_{11}^{(i)} & x_{12}^{(i)} & \cdots & x_{1,20}^{(i)} & \cdots & x_{1,30}^{(i)} & \cdots & x_{1,40}^{(i)} \\ \vdots & \vdots & & \vdots & & \vdots & & x_{2,40}^{(i)} \\ \vdots & \vdots & & \vdots & \cdots & x_{3,30}^{(i)} & * & * \\ x_{41}^{(i)} & x_{42}^{(i)} & \cdots & x_{4,20}^{(i)} & * & * & * & * \end{bmatrix}.$$

(iii) $n_1 = n_2 = n_3 = n_4 = 40$ (complete case)

$$\begin{bmatrix} x_{11}^{(i)} & x_{12}^{(i)} & \cdots & x_{1,20}^{(i)} & \cdots & x_{1,30}^{(i)} & \cdots & x_{1,40}^{(i)} \\ \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ x_{41}^{(i)} & x_{42}^{(i)} & \cdots & x_{4,20}^{(i)} & \cdots & x_{4,30}^{(i)} & \cdots & x_{4,40}^{(i)} \end{bmatrix}.$$

(ii)-m_ℓ, $(\ell = 1, 2, 3, 4)$ is the width of the simultaneous confidence intervals for changing values of m_ℓ. For example, the case of (ii)-m_3, we calculate the $\sqrt{E}$ by the $10,000$ Monte Carlo simulation for $m_1 = m_2 = m_4 = 0$, $m_3 \neq 0$. From Table 1, in the case of (ii)-m_4, we note that the width of the simultaneous confidence intervals are shorter than the ones of the earlier procedure (i). For the case of (ii)-m_1, our procedure is very near the complete case (iii). Since $\gamma^2 = \sigma^2(1-\rho)$, when the correlation coefficient ρ is large, the width of the simultaneous confidence intervals is short and when the variance σ^2 is large, the width of the simultaneous confidence intervals is long. Therefore our procedure is very useful for the case the observations are of the monotone-type missing.

σ^2	ρ	(i)	(ii)-m_4	(ii)-m_3	(ii)-m_1	(iii)
	0.1	0.849	0.778	0.669	0.598	0.597
1	0.5	0.633	0.580	0.498	0.446	0.445
	0.9	0.283	0.259	0.223	0.199	0.199
	0.1	3.398	3.114	2.675	2.392	2.387
4	0.5	2.532	2.321	1.993	1.783	1.779
	0.9	1.132	1.038	0.891	0.797	0.796
	0.1	7.645	7.006	6.018	5.383	5.372
9	0.5	5.698	5.222	4.485	4.012	4.003
	0.9	2.548	2.335	2.006	1.794	1.790

Table 1. the widths of the simultaneous confidence intervals

References

[1] Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society. Series B. Methodological*, **39**, 1–38.

[2] Hotelling, H. (1931). The generalization of Student's ratio, *The Annals of Mathematical Statistics*, **2**, 360–378.

[3] Konishi, S. and Shimizu, K. (1994). Maximum likelihood estimation of an intraclass correlation in a bivariate normal distribution with missing observations, *Communications in Statistics. Theory and Methods*, **23**, 1593–1604.

[4] McLachlan, G. J. and Krishnan, T. (1997). *The EM Algorithm and Extensions*, John Wiley, New York.

[5] Minami, M. and Shimizu, K. (1997). Estimation for a common intraclass correlation in a bivariate normal distributions with missing observations, *American Journal of Mathematical and Management Sciences*, **17**, 3–14.

[6] Seo, T. and Srivastava, M. S. (2000). Testing equality of means and simultaneous confidence intervals in repeated measures with missing data, *Biometrical Journal*, **42**, 981–993.

[7] Srivastava, M. S. (1985). Multivariate data with missing observations, *Communications in Statistics. A. Theory and Methods*, **14**, 775–792.

[8] Srivastava M. S. and Carter, E. M. (1986). The maximum likelihood method for non-response in sample survey, *Survey Methodology*, **12**, 61–72.

[9] Yu, J., Krishnamoorthy, K. and Pannala, M. K. (2006). Two-sample inference for normal mean vectors based on monotone missing data, *Journal of Multivariate Analysis*, **97**, 2162–2176.