Multiple Comparisons among Mean Vectors when the Dimension is Larger than the Total Sample Size

Sho Takahashi; Masashi Hyodo and Takahiro Nishiyama

Abstract — We consider multiple comparisons among mean vectors for high-dimensional data under the multivariate normality. The statistic based on Dempster trace criterion is given, and also its approximate upper percentile is derived by using Bonferroni's inequality. Finally, the accuracy of its approximate value is evaluated by Monte Carlo simulation.

Keyword: Asymptotic approximation; Bonferroni's inequality; Dempster trace criterion; high-dimensional data; multiple comparison.

1 Introduction

We consider the multiple comparisons among mean vectors. When the sample size is larger than the dimension, simultaneous confidence intervals among mean vectors have been discussed by Seo and Siotani (1992), Seo (1995) and so on. In this study, we consider the multiple comparisons among mean vectors for high-dimensional data.

Let $x_1^{(i)}, x_2^{(i)}, \ldots, x_N^{(i)}$ be $N$ independent sample vectors that have the multivariate normal distribution with mean vector $\mu^{(i)}$ and common covariance matrix $\Sigma$. Let the $i$-th sample mean vector and the $i$-th sample covariance matrix be

$$\bar{x}^{(i)} = \frac{1}{N} \sum_{j=1}^{N} x_j^{(i)},$$

$$S^{(i)} = \frac{1}{N-1} \sum_{j=1}^{N} (x_j^{(i)} - \bar{x}^{(i)})(x_j^{(i)} - \bar{x}^{(i)})^T,$$

respectively.

In general, the simultaneous confidence intervals for pairwise multiple comparisons among mean vectors with the confidence level $1 - \alpha$ are given by

$$a'(\mu^{(\ell)} - \mu^{(m)}) \in \left[ a'(\bar{x}^{(\ell)} - \bar{x}^{(m)}) \pm t \sqrt{2a'Sa/N} \right], \quad \forall a \in \mathbb{R}^p - \{0\}, \quad 1 \leq \ell < m \leq k,$$

where $S = (1/k) \sum_{i=1}^{k} S^{(i)}$, $\mathbb{R}^p - \{0\}$ is the set of any nonnull real $p$-dimensional vectors and the values $t (> 0)$ satisfies as follows:

$$T^2_{\text{max}, p} = \max_{1 \leq \ell < m \leq k} \left\{ T^2_{\ell m} \right\},$$

$$T^2_{\ell m} = \frac{N}{2} (y^{(\ell)} - y^{(m)})^T S^{-1} (y^{(\ell)} - y^{(m)}),$$

$$y^{(i)} = \bar{x}^{(i)} - \mu^{(i)}, \quad i = 1, 2, \ldots, k.$$
Here we consider the situation such that the dimension is larger than the total sample size. However, since $S$ is singular in such a situation, it is impossible to define $T^2_{km}$. So it is required to give another statistic instead of $T^2_{km}$. For high-dimensional data, Dempster trace criterion has been proposed by Dempster (1958). Also, Fujikoshi, Himeno and Wakaki (2004), Himeno (2007) and so on have discussed Dempster trace criterion for multivariate linear hypothesis. In this paper, we give the statistic based on Dempster trace criterion for multiple comparisons among mean vectors in high-dimensional framework.

2 The case of high-dimensional data

We discuss an extension to the case of high-dimensional data under the framework:

$$A1 : \quad n \to \infty, \quad p \to \infty, \quad \frac{p}{n} \to y \in [0, \infty).$$

Further, we assume that

$$A2 : \quad \text{tr} \Sigma^i = O(p), \quad i = 1, 2, \ldots, g.$$

Since $S$ is singular for high-dimensional data, we give the following statistic based on Dempster trace criterion.

$$D^2_{\text{max}, p} = \max_{1 \leq m < \infty} \{ D^2_{km} \},$$

$$D^2_{km} = \frac{p}{\tilde{\sigma}} \left( \frac{\text{tr} S_{km}}{\text{tr} S_{e}} - 1 \right),$$

where

$$S_{h}^{(r,m)} = \frac{N}{2} (y^{(r)} - y^{(m)})(y^{(r)} - y^{(m)})',$$

$$S_{e} = \sum_{i=1}^{k} \sum_{j=1}^{N} (x^{(r)}_{ij} - \bar{x}^{(r)})(x^{(m)}_{ij} - \bar{x}^{(m)}),$$

$$\tilde{\sigma} = \sqrt{\frac{2(r_1 + r_2) [\text{tr} S_{e}^2 / n^2 - (\text{tr} S_{e})^2 / n^2]}{p}},$$

and $n = k(N - 1)$, $r_1 = p_1$, $r_2 = p/n$.

In order to construct the actual simultaneous confidence intervals with the confidence level $1 - \alpha$, it is required to find the upper 100$\alpha$ percentiles of $D^2_{\text{max}, p}$ statistic. However, it is difficult to find the exact value. So we give an approximation based on Bonferroni's inequality for $D^2_{\text{max}, p}$ Statistic.

By Bonferroni's inequality for $\Pr \left[ D^2_{\text{max}, p} > z \right]$,

$$\Pr \left[ D^2_{\text{max}, p} > z \right] < \sum_{i=1}^{k-1} \sum_{m=1}^{k} \Pr \left[ D^2_{km} > z \right].$$

Therefore, the first order Bonferroni approximation $z_1$ is defined as

$$\sum_{i=1}^{k-1} \sum_{m=1}^{k} \Pr \left[ D^2_{km} > z_1 \right] = \alpha.$$

By using the same idea as Fujikoshi, Himeno and Wakaki (2004), we have the following theorem for the distribution of $D^2_{km}$.
Theorem 1 Under the framework $A_1$ and the assumption $A_2$, it holds that
\[ D_{2m}^2 \overset{d}{\to} N(0, 1). \]
Further, expanding the distribution of $D_{2m}^2$, we can obtain the following theorem.

Theorem 2 Under the framework $A_1$ and the assumption $A_2$, the distribution of $D_{2m}^2$ can be expanded as
\[
\Pr[D_{2m}^2 \leq z] = \Phi(z) - \phi(z) \left\{ \frac{1}{\sqrt{p}} \frac{\sqrt{2} \varepsilon_1}{\varepsilon_2} h_2(z) + \frac{1}{p} \frac{\varepsilon_2^3}{2 \varepsilon_2^5} h_3(z) + \frac{1}{2n} h_4(z) \right\} + o(p^{-\frac{3}{2}}),
\]
where $\Phi(z)$ and $\phi(z)$ are the distribution function of the standard normal distribution and the density function of the standard normal distribution, respectively, and $h_i(z)$ are the Hermite polynomials given by
\[
h_1(z) = z, \quad h_2(z) = z^2, \quad h_3(z) = z^3 - 3z, \quad h_5(z) = z^5 - 10z^3 + 15z.
\]
Also, by Cornish-Fisher expansion, its upper 100\(\alpha\) percentile can be expanded as
\[
z_{\alpha}(z) = z_{\alpha} + \frac{1}{\sqrt{p}} \frac{\sqrt{2} \varepsilon_1}{\varepsilon_2} (z_{\alpha}^2 - 1) + \frac{1}{p} \frac{\varepsilon_2^3}{2 \varepsilon_2^5} z_{\alpha} (z_{\alpha}^2 - 3) - \frac{2 \varepsilon_3}{9 \varepsilon_2^3} z_{\alpha} (2z_{\alpha}^2 - 5) + \frac{1}{2n} z_{\alpha}^2,
\]
where $z_{\alpha}$ is the upper 100\(\alpha\) percentile of the standard normal distribution, and
\[
\varepsilon_2 = \frac{n^2}{p(n+2)(n+1)} \left\{ \frac{\text{tr} S_i^2}{n^2} - \frac{\text{tr} S_e^2}{n^3} \right\},
\]
\[
\varepsilon_3 = \frac{1}{pn^3} \left\{ \text{tr} S_i^3 - \frac{3}{n} \text{tr} S_i^2 \text{tr} S_e + \frac{2}{n^2} \text{tr} S_e^3 \right\},
\]
\[
\varepsilon_4 = \frac{1}{pn^4} \left\{ b_1 (\text{tr} S_e)^4 + b_2 (\text{tr} S_i)^2 \text{tr} S_e + b_3 \text{tr} S_i \text{tr} S_e^2 + b_4 (\text{tr} S_e)^2 + b_5 (\text{tr} S_i)^2 \right\},
\]
\[
b_1 = \frac{n^6(5n + 6)}{(n + 6)(n + 4)(n + 2)(n + 1)(n - 1)(n - 2)(n - 3)},
\]
\[
b_2 = -\frac{2n^4(5n + 6)}{(n + 6)(n + 4)(n + 1)(n - 2)(n - 3)(n^2 + n - 2)},
\]
\[
b_3 = -\frac{4n^4(n^2 + n + 6)}{(n + 6)(n + 4)(n + 1)(n - 2)(n - 3)(n^2 + n - 2)},
\]
\[
b_4 = \frac{n^6(2n^2 + 3n + 6)}{(n + 6)(n + 4)(n + 1)(n - 2)(n - 3)(n^2 + n - 2)},
\]
\[
b_5 = \frac{n^6(n^2 + n + 2)}{(n + 6)(n + 4)(n + 1)(n - 2)(n - 3)(n^2 + n - 2)}.
\]

By using Theorem 2, we obtain the following corollary.

Corollary 3 Under the framework $A_1$ and the assumption $A_2$, the first order Bonferroni approximate upper 100\(\alpha\) percentiles of $D_{2m}^2$ that is, $z_{\alpha} = z_{\alpha}(\alpha)$ is given by
\[
z_{\alpha} = z_{\alpha'} + \frac{1}{\sqrt{p}} \frac{\sqrt{2} \varepsilon_1}{\varepsilon_2} (z_{\alpha'}^2 - 1) + \frac{1}{p} \frac{\varepsilon_2^3}{2 \varepsilon_2^5} z_{\alpha'} (z_{\alpha'}^2 - 3) - \frac{2 \varepsilon_3}{9 \varepsilon_2^3} z_{\alpha'} (2z_{\alpha'}^2 - 5) + \frac{1}{2n} z_{\alpha'}^2,
\]
where $\alpha' = \alpha/K$ and $K = k(k + 1)/2$. 

121
Therefore, by using $z_1$, we give the following approximate simultaneous confidence intervals for high-dimensional data.

$$a'(\mu^{(l)} - \mu^{(m)}) \in a'(\bar{x}^{(l)} - \bar{x}^{(m)}) \pm d_1 \sqrt{2a'(trS)a/N}, \quad \forall a \in \mathbb{R}^p - \{0\}, \quad 1 \leq l < m \leq k,$$

where $d_1^2 = 1 + (\lambda/p)z_1$.

3 Numerical examinations

We evaluate the accuracy of the obtained approximation by Monte Carlo simulation.

Acknowledgment

The research of the second author was partly supported by Japan Society for the Promotion of Science (JSPS). The research of the third author was partly supported by KAKENHI (21700311; Grant-in-Aid for Young Scientists (B)).

References


