一様構造をもつ不完全データに対する平均ベクトルの多変量多重比較法について

東京理科大・理・院　小泉 和之
東京理科大・理　　瀬尾　隆

複数の多次元母集団において平均ベクトル間の比較を行う手法に多変量多重比較法がある．本報告では，多次元母集団から得られた標本にいくらかの欠測データが生じた場合の多変量多重比較法について考えるが，欠測データが生じたときの統計解析手法については，Dempster et al. (1977) の EM アルゴリズムがよく知られている．また，パラメータの MLE を，欠測データを含む標本から得られる尤度方程式を数値反復法によって解く方法が Srivastava (1985) で提案されている．さらに，そのような近似を用いずに平均成分の同等性を調べる方法が分散共分散行列に一様構造を仮定したもとで単調な欠測の場合に Seo and Srivastava (2000) などで議論され，2 つの平均ベクトルの同等性検定，同時信頼区間について，小泉，瀬尾 (2007) は Seo and Srivastava (2000) の方法を拡張し，ホテリングの T^2 型統計量を用いた正確な分布の議論とその検出力が数値的に調べられている．

ここでは，小泉，瀬尾 (2007) の方法を k 個の母集団の平均ベクトル間の対比較に対する同時信頼区間へ拡張することを考えた．

設定として，第 $i(i=1,2,\ldots,k)$ 母集団からの観測ベクトルを，$\boldsymbol{x}_1^{(i)},\boldsymbol{x}_2^{(i)},\ldots,\boldsymbol{x}_{n^{(i)}}^{(i)} \sim N_p(\boldsymbol{\mu}_i,\boldsymbol{\Sigma})$ とする．ここに，$\boldsymbol{\mu}_i=(\mu_1^{(i)},\mu_2^{(i)},\ldots,\mu_p^{(i)})'$ である．また，各母集団の分散共分散行列 $\boldsymbol{\Sigma}$ は同じであるとし，$\boldsymbol{\Sigma}$ には一様構造を仮定する．つまり，σ^2,ρ が未知のパラメータであるとし，$\boldsymbol{\Sigma}=\sigma^2[(1-\rho)\boldsymbol{I}_p+\rho\mathbf{1}_p\mathbf{1}_p']$, $\mathbf{1}_p=(1,1,\ldots,1)': p\times 1$ をみたしているとする．

得られた観測ベクトルを並べたデータ行列の ℓ 行目，j 列目の観測数を $n_\ell^{(i)}$ $(n_1^{(i)}=n_2^{(i)}\geq\cdots\geq n_p^{(i)})$，$p_j^{(i)}$ $(p\equiv p_1^{(i)}\geq p_2^{(i)}\geq\cdots\geq p_n^{(i)}\geq 2)$ とし，それぞれの観測ベクトルに対応する変換行列を $\boldsymbol{C}_j^{(i)}:(p_j^{(i)}-1)\times p_j^{(i)}$ $(p_j^{(i)}=p$ のとき，$\boldsymbol{C}$ と表す) とする．ただし，$\boldsymbol{C}_j^{(i)}$ は，$\boldsymbol{C}_j^{(i)}\boldsymbol{C}_j^{(i)'}=\boldsymbol{I}_{p_j^{(i)}-1}$，$\boldsymbol{C}_j^{(i)}\mathbf{1}_{p_j^{(i)}}=\mathbf{0}$ をみたす行列である．すると，$\boldsymbol{y}_j^{(i)}=\boldsymbol{C}_j^{(i)}\boldsymbol{x}_j^{(i)}$ は，

$$\boldsymbol{y}_j^{(i)}\sim N_{p_j^{(i)}-1}(\boldsymbol{C}_j^{(i)}\boldsymbol{\mu}_{ij},\gamma^2\boldsymbol{I}_{p_j^{(i)}-1}),\quad \gamma^2\equiv\sigma^2(1-\rho),\quad \boldsymbol{\mu}_{ij}=(\mu_1^{(i)},\mu_2^{(i)},\ldots,\mu_{p_j^{(i)}}^{(i)})'$$

となり，変換後のデータの標本平均，未知パラメータ γ^2 の不偏推定量は，

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

と導かれる．また，$\overline{\boldsymbol{y}}^{(i)}=(\overline{y}_{1\cdot}^{(i)},\overline{y}_{2\cdot}^{(i)},\ldots,\overline{y}_{p-1\cdot}^{(i)})'$ とおく．

まず $k=2$ とする．このとき標本平均ベクトルの差の期待値，分散共分散行列は，

$$\mathrm{E}(\overline{\boldsymbol{y}}^{(1)}-\overline{\boldsymbol{y}}^{(2)})=\boldsymbol{C}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\equiv\boldsymbol{C}\boldsymbol{\delta}_{12},$$

$$\mathrm{Cov}(\overline{\boldsymbol{y}}^{(1)}-\overline{\boldsymbol{y}}^{(2)})=\gamma^2\begin{bmatrix} n_2^{(1)^{-1}}+n_2^{(2)^{-1}} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & n_p^{(1)^{-1}}+n_p^{(2)^{-1}}\end{bmatrix}\equiv\gamma^2\boldsymbol{V}_{12}$$

となるので，$\boldsymbol{a}'\boldsymbol{\delta}_{12}(\boldsymbol{a} \in \mathbb{R}_p^* \equiv \mathbb{R}^p - \{\mathbf{0}\},\ \boldsymbol{a}'\mathbf{1}_p = 0)$ に対する同時信頼区間は，$\boldsymbol{a}' = \boldsymbol{d}'\boldsymbol{C}$ をみたす $\boldsymbol{d} \in \mathbb{R}_{p-1}^*$ に注意すれば，

$$\Pr\left(\boldsymbol{d}'(\overline{\boldsymbol{y}}^{(1)} - \overline{\boldsymbol{y}}^{(2)}) - \sqrt{E} \leq \boldsymbol{a}'\boldsymbol{\delta}_{12} \leq \boldsymbol{d}'(\overline{\boldsymbol{y}}^{(1)} - \overline{\boldsymbol{y}}^{(2)}) + \sqrt{E}\right) = 1 - \alpha$$

となることが小泉，瀬尾 (2007) によって導かれている．ここに，$E = T_{12}^2(\alpha)\hat{\gamma}^2\boldsymbol{d}'\boldsymbol{V}_{12}\boldsymbol{d}$ であり，$T_{12}^2(\alpha)$ はホテリングの T^2 型統計量の分布の上側 $100\alpha\%$ 点を表している．

次に $i = 1, 2, \ldots, k$ とする．このとき，平均ベクトル間の対比較に対する同時信頼区間，つまり，$\boldsymbol{a} \in \mathbb{R}_p^*,\ \boldsymbol{a}'\mathbf{1}_p = 0$ に対して，$\boldsymbol{a}'(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j),\ i < j,\ i, j = 1, 2, \ldots, k$ の同時信頼区間を考える．

先の結果を用いれば，

$$\Pr\left(T_{ij}^2 \leq g^2,\ i < j,\ i, j = 1, 2, \ldots, k\right) = 1 - \alpha \tag{1}$$

をみたす g を求めることにより，同時信頼区間は導かれる．このとき，(1) 式をみたす g は $T_{\max\cdot\mathrm{p}}^2 \equiv \max_{i<j}\{T_{ij}^2\}$ 統計量の上側 $100\alpha\%$ 点を $T_{\max\cdot\mathrm{p}}^2(\alpha)$ とすれば，

$$g^2 = T_{\max\cdot\mathrm{p}}^2(\alpha)$$

となる．従って，k 個の母集団の平均ベクトル $\boldsymbol{\mu}_i$ について，

$$\boldsymbol{a}'(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j),\ i < j,\ i, j = 1, 2, \ldots, k, \quad \text{for } {}^\forall\boldsymbol{a} \in \mathbb{R}_p^*,\ \boldsymbol{a}'\mathbf{1}_p = 0$$

の同時信頼区間は，$\boldsymbol{a}' = \boldsymbol{d}'\boldsymbol{C}$ に注意すれば，次のようになる．

$$\Pr\left(\boldsymbol{d}'(\overline{\boldsymbol{y}}^{(i)} - \overline{\boldsymbol{y}}^{(j)}) - \sqrt{L} \leq \boldsymbol{a}'\boldsymbol{\delta}_{ij} \leq \boldsymbol{d}'(\overline{\boldsymbol{y}}^{(i)} - \overline{\boldsymbol{y}}^{(j)}) + \sqrt{L}\right) = 1 - \alpha$$

ここに，$L = T_{\max\cdot\mathrm{p}}^2(\alpha)\hat{\gamma}^2\boldsymbol{d}'\boldsymbol{V}_{ij}\boldsymbol{d}$ である．よって，同時信頼区間に含まれている $T_{\max\cdot\mathrm{p}}^2$ 統計量の%点の評価が必要となる．そこで次の不等式

$$\Pr\left(T_{\max\cdot\mathrm{p}}^2 \leq g^2\right) \geq 1 - \sum_{i<j}\sum \Pr\left(T_{ij}^2 > g^2\right)$$

を用いれば，右辺の確率が $1 - \alpha$ となるようにすれば同時信頼区間の同時信頼係数は $1 - \alpha$ 以上となる．つまり，それぞれの T_{ij}^2 統計量に対して，

$$\Pr\left(T_{ij}^2 > g^2\right) = \frac{2}{k(k-1)}\alpha \equiv \alpha^*$$

であり，上式を満たすように g を計算すればよい．すると，それぞれの T_{ij}^2 統計量は本質的に F 分布に従っている (see, 小泉，瀬尾 (2007)) ことにより，g の値を求めることができる．以上のことをまとめると次のようになる．同時信頼係数 $1 - \alpha$ をもつ対比較に対する保守的な同時信頼区間は，

$$\Pr\left(\boldsymbol{d}'(\overline{\boldsymbol{y}}^{(i)} - \overline{\boldsymbol{y}}^{(j)}) - \sqrt{L_1} \leq \boldsymbol{a}'\boldsymbol{\delta}_{ij} \leq \boldsymbol{d}'(\overline{\boldsymbol{y}}^{(i)} - \overline{\boldsymbol{y}}^{(j)}) + \sqrt{L_1}\right) \geq 1 - \alpha$$

となることがわかった．ここに，$L_1 = t_1^2(\alpha^*)\hat{\gamma}^2\boldsymbol{d}'\boldsymbol{V}_{ij}\boldsymbol{d}$ であり，$t_1^2(\alpha^*)$ は T_{ij}^2 統計量の上側 $100\alpha^*\%$ 点である．