

Alternative two stage procedure for confidence region of mean vector

広島大学大学院・理学研究科 若木 宏文 (Hirofumi Wakaki)
 Graduate School of Science, Hiroshima University

1 はじめに

p 変量正規母集団 $N_p(\mu, \Sigma)$ の母平均 μ の信頼領域を構成する問題を考える. ここで分散共分散行列 Σ は未知とする. $(x_n)_{n=1,2,\dots}$ を $N_p(\mu, \Sigma)$ に従う独立な確率ベクトルの列とすると, 与えられた正値 d, α に対して

$$D = \{\mu : \|\mu - \bar{x}_N\| \leq d\}, \quad \bar{x}_N = \frac{1}{N} \sum_{i=1}^N x_i \tag{1.1}$$

なる信頼領域で, 信頼係数が $1 - \alpha$ 以上となるような N を導出したい.

$\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_p$ を Σ の固有値, $\chi_p^2(\alpha)$ を自由度 p のカイ二乗分布の上側 $100\alpha\%$ 点とすると, 任意の p 次元ベクトル y に対して

$$y'y \leq y'(\lambda_p \Sigma^{-1})y, \quad y'y \geq y'(\lambda_1 \Sigma^{-1})y, \tag{1.2}$$

であるから, $\frac{nd^2}{\lambda_p} \leq \chi_p^2(\alpha)$ ならば

$$\Pr\{\|\bar{x}_n - \mu\|^2 \leq d^2\} \leq \Pr\{n(\bar{x}_n - \mu)' \Sigma^{-1} (\bar{x}_n - \mu) \leq \frac{nd^2}{\lambda_p}\} \leq 1 - \alpha, \tag{1.3}$$

同様に $\frac{nd^2}{\lambda_1} \geq \chi_p^2(\alpha)$ ならば

$$\Pr\{\|\bar{x}_n - \mu\|^2 \leq d^2\} \geq 1 - \alpha \tag{1.4}$$

が成り立つ. したがって, (1.1) のような信頼領域で, 信頼係数が $1 - \alpha$ 以上となるために必要な標本数 N の最小値を n_0 とすると

$$\left\lceil \frac{\chi_p^2(\alpha)\lambda_p}{d^2} \right\rceil \leq n_0 \leq \left\lceil \frac{\chi_p^2(\alpha)\lambda_1}{d^2} + 1 \right\rceil \tag{1.5}$$

となり, Σ に関する情報が何もないとき, 初めから必要な標本数 N を決めることは不可能である. ただし, 記号 $[a]$ はガウス記号, すなわち, a を越えない最大の整数値を表す.

Healy (1956) の二段階標本抽出法は, 標本抽出を二段階に分けて第一段階で抽出された標本を用いて信頼領域の半径と信頼係数に関する条件をみたすような標本数を決定する方法である.

<第一段階> まず, m 個の標本 x_1, x_2, \dots, x_m を抽出し, 不偏分散共分散行列 $S_m = \frac{1}{m-1} \sum_{i=1}^m (x_i - \bar{x}_m)(x_i - \bar{x}_m)'$ を求める. $F_{p,m-p}(\alpha)$ を自由度 $p, m-p$ の F 分布の上側 $100\alpha\%$ 点とし, S_m の最大固有値 l_1 を用いて

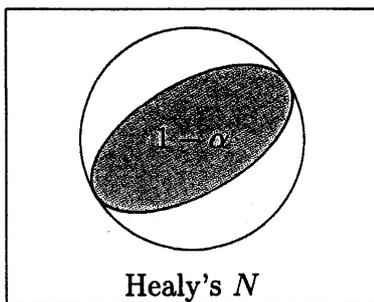
$$N = \max \left\{ m, \left\lceil \frac{a_{m,p}(\alpha)l_1}{d^2} + 1 \right\rceil \right\}, \quad a_{m,p}(\alpha) = \frac{p(m-1)}{m-p} F_{p,m-p}(\alpha) \tag{1.6}$$

と定義する。

<第二段階> $N > m$ ならば, $N - m$ 個の標本 $\mathbf{x}_{m+1}, \mathbf{x}_{m+2}, \dots, \mathbf{x}_N$ を追加して, (1.1) の信頼領域を作ると,

$$\Pr\{\boldsymbol{\mu} \in D\} \geq \Pr\{N(\bar{\mathbf{x}}_N - \boldsymbol{\mu})' S_m^{-1}(\bar{\mathbf{x}}_N - \boldsymbol{\mu}) \leq a_{m,p}(\alpha)\} \quad (1.7)$$

となるが, $\sqrt{N}(\bar{\mathbf{x}}_N - \boldsymbol{\mu})$ は S_m と独立に $N_p(0, \Sigma)$ に従うことから, $N(\bar{\mathbf{x}}_N - \boldsymbol{\mu})' S_m^{-1}(\bar{\mathbf{x}}_N - \boldsymbol{\mu})$ の分布の上側 $100\alpha\%$ 点が丁度 $a_{m,p}(\alpha)$ となり, D の信頼係数が $1 - \alpha$ 以上となることがわかる。



Healy の二段階法は $\lambda_2, \dots, \lambda_p$ が最大固有値 λ_1 と大きく異なるとき, 必要数以上の標本をとりやすくなる。左図は, $p = 2$ の場合の信頼領域と信頼係数の関係を表す。円は信頼領域 D , 円の内側の楕円は, $\{\boldsymbol{\mu} : N(\bar{\mathbf{x}}_N - \boldsymbol{\mu})' S_m^{-1}(\bar{\mathbf{x}}_N - \boldsymbol{\mu}) \leq a_{m,p}(\alpha)\}$ を表しており, $\boldsymbol{\mu}$ が楕円の内部に入る確率が $1 - \alpha$ となる。 S_m の小さい方の固有値を l_2 とするとき, l_1/l_2 が大きくなるほど円の中に占める楕円の割合が小さくなり, 信頼係数が必要以上に大きくなる。

2 変数の部分ベクトルが独立である場合

分散共分散行列に何らかの構造があることが分かっているとき, 前述のような標本数のむだを少なくできる場合がある。例えば

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \quad (\Sigma_{12} : q \times r) \quad (2.1)$$

と分割したとき, $\Sigma_{12} = 0$ であることが分かっているとき,

$$\sqrt{N}(\bar{\mathbf{x}}_N - \boldsymbol{\mu}) = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix}, \quad S_m = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \quad (2.2)$$

と分割して

$$l_* = \max\{l_{(1)}, l_{(2)}\}, \quad \begin{array}{l} l_{(1)} : A_{11} \text{ の最大固有値} \\ l_{(2)} : A_{22} \text{ の最大固有値} \end{array} \quad (2.3)$$

とすると,

$$N\|\bar{\mathbf{x}}_N - \boldsymbol{\mu}\|^2 = \|\mathbf{y}_1\|^2 + \|\mathbf{y}_2\|^2 \leq l_* (\mathbf{y}_1' A_{11}^{-1} \mathbf{y}_1 + \mathbf{y}_2' A_{22}^{-1} \mathbf{y}_2). \quad (2.4)$$

より,

$$\Pr\{\|\bar{\mathbf{x}}_N - \boldsymbol{\mu}\| \leq d\} \geq \Pr\left\{\mathbf{y}_1' A_{11}^{-1} \mathbf{y}_1 + \mathbf{y}_2' A_{22}^{-1} \mathbf{y}_2 \leq \frac{Nd^2}{l_*}\right\} \quad (2.5)$$

が成り立つ。 N が l_* の関数として定義される時、 $\Sigma_{12} = 0$ であることから $\mathbf{y}'_1 A_{11}^{-1} \mathbf{y}_1, \mathbf{y}'_2 A_{22}^{-1} \mathbf{y}_2, A_{11}, A_{22}$ は独立で、

$$\mathbf{y}'_1 A_{11}^{-1} \mathbf{y}_1 \sim \frac{q(m-1)}{m-q} F_{q, m-q}, \quad \mathbf{y}'_2 A_{22}^{-1} \mathbf{y}_2 \sim \frac{r(m-1)}{m-r} F_{r, m-r} \quad (2.6)$$

となり、 $T_{m,q,r} := \mathbf{y}'_1 A_{11}^{-1} \mathbf{y}_1 + \mathbf{y}'_2 A_{22}^{-1} \mathbf{y}_2$ の分布は、 m, q, r によって決まることがわかる。

$$N_0 = \max \left\{ m, \left\lceil \frac{b_{m,q,r}(\alpha) l_*}{d^2} + 1 \right\rceil \right\}, \quad b_{m,q,r}(\alpha) : T_{m,q,r} \text{ の上側 } 100\alpha\% \text{ 点} \quad (2.7)$$

として、 $N_0 - m$ 個の標本を追加し、

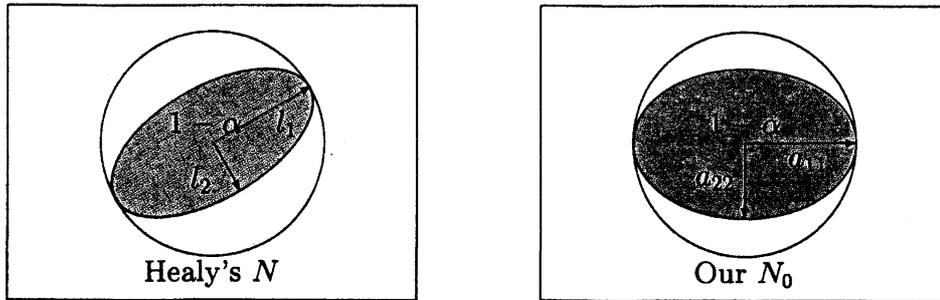
$$D_0 = \{ \mu : \|\mu - \bar{x}_{N_0}\| \leq d \}, \quad (2.8)$$

とすると、 D_0 の信頼係数も $1 - \alpha$ 以上となる。

下図は、 $p = 2$ の場合の Healy の方法による信頼係数と新しい方法による信頼係数を表す。 l_2 を S_m の小さい方の固有値とする。 $S_m = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ とすると $l_1 \geq l_{(1)} = a_{ii} \geq l_2$ ($i = 1, 2$) であるから

$$\frac{l_1}{l_2} \geq \frac{a_{11}}{a_{22}}, \quad \frac{l_1}{l_2} \geq \frac{a_{22}}{a_{11}}. \quad (2.9)$$

したがって、円に内接した楕円の面積は変数を分割した方が大きくなるので $b_{m,q,r}(\alpha) \geq a_{m,p}$ であることが予想される。



$p = 2, 3, 4$ と $m = 6, 8, 10, 15, 20, 30, 40$ について $a_{m,p}(0.05), b_{m,q,r}(0.05)$ を計算した結果を以下の表に示す。一様に $b_{m,p,q}(0.05) < a_{m,p}(0.05)$ となっており、その差は m が小さい程大きい。 $p = 4$ のとき、変数は、 $(q, r) = (1, 3), (2, 2)$ の二通りに分割することができるが、 $(q, r) = (2, 2)$ と分けた方が良いことがわかる。

表 1 : $p = 2$ の場合

m	6	8	10	15	20	30	40
$a_{m,2}(0.05)$	17.4	12.0	10.0	8.2	7.5	6.9	6.7
$b_{m,1,1}(0.05)$	11.2	9.3	8.3	7.4	7.0	6.7	6.4

表 2 : $p = 3$

m	6	8	10	15	20	30	40
$a_{m,3}(0.05)$	46.4	22.7	16.8	12.2	10.7	9.5	9.0
$b_{m,1,2}(0.05)$	20.9	14.8	12.6	10.6	9.6	8.9	8.7

表 3 : $p = 4$

m	6	8	10	15	20	30	40
$a_{m,4}(0.05)$	192.5	44.7	27.2	17.1	14.3	12.2	11.4
$b_{m,1,3}(0.05)$	48.8	25.2	19.0	14.2	12.5	11.4	10.8
$b_{m,2,2}(0.05)$	31.0	20.1	16.5	13.2	12.1	11.0	10.6

3 Σ について事前情報がない場合

$\Sigma_{12} = 0$ とは限らない場合, $T_{m,q,r}$ の分布は未知母数 Σ に依存するが (2.7) において, $b_{m,q,r}(\alpha)$ を $T_{m,q,r}$ の分布の上側 $100\alpha\%$ 点の Σ に関する上限で置き換えれば信頼係数が $1 - \alpha$ 以上となるように標本数を定めることができる.

一般性を失うことなく, $q \leq r$ と仮定する. $\rho_1^2, \rho_2^2, \dots, \rho_q^2$ ($1 \geq \rho_1 \geq \dots \geq \rho_q \geq 0$) を $\Sigma_{11}^{-1/2} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \Sigma_{11}^{-1/2}$ の固有値とし,

$$\begin{aligned} \Sigma_{11}^{-1/2} \Sigma_{12} \Sigma_{22}^{-1/2} &= H_1 [\text{diag}(\rho_1, \dots, \rho_q) \vdash O] H_2', \\ H_1' H_1 &= I_q, \quad H_2' H_2 = I_r \end{aligned} \quad (3.1)$$

と特異値分解して,

$$z_i = H_i \Sigma_{ii}^{-1/2} y_i, \quad B_{ij} = H_i' \Sigma_{ii}^{-1/2} A_{ij} \Sigma_{jj}^{-1/2} H_j \quad (i, j = 1, 2) \quad (3.2)$$

と変数変換すると,

$$T_{m,q,r} = z_1' B_{11}^{-1} z_1 + z_2' B_{22}^{-1} z_2 \quad (3.3)$$

と表される. z_i, B_{ij} の分布は,

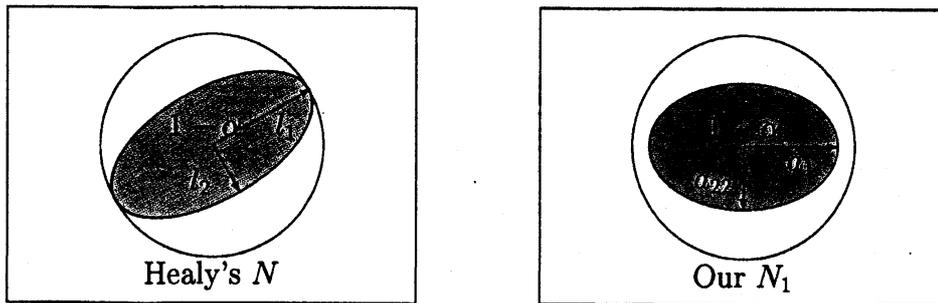
$$\begin{aligned} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} &\sim N_p(0, \Sigma_*), \quad \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} \sim W_p\left(m-1, \frac{1}{m-1} \Sigma_*\right), \\ \Sigma_* &= \begin{pmatrix} I_q & [\Lambda \vdash O] \\ \begin{bmatrix} \Lambda \\ O \end{bmatrix} & I_r \end{pmatrix}, \quad \Lambda = \text{diag}(\rho_1, \dots, \rho_q) \end{aligned} \quad (3.4)$$

であるから, $T_{m,q,r}$ の分布は, ρ_1, \dots, ρ_q のみに依存する. $c_{m,q,r}(\alpha; \rho_1, \dots, \rho_q)$ を $T_{m,q,r}$ の上側 $100\alpha\%$ 点とし,

$$\begin{aligned} N_1 &= \max \left\{ m, \left[\frac{c_{m,q,r}^*(\alpha) l_*}{d^2} + 1 \right] \right\}, \\ c_{m,q,r}^*(\alpha) &= \max_{1 \geq \rho_1 \geq \dots \geq \rho_q \geq 0} c_{m,q,r}(\alpha; \rho_1, \dots, \rho_q) \end{aligned} \quad (3.5)$$

と定めると、 N_1 個の標本平均に基づく信頼領域もまた、信頼係数 $1 - \alpha$ 以上であることが保証される。

次の図は $p = 2$ の場合に、Healy の N と、 N_1 と信頼係数との関係を示したものである。内側の楕円の長軸と短軸の比は、変数を分割した方が小さくなるのは独立と仮定した場合と同じであるが、上側 $100\alpha\%$ 点の、未知パラメータ ρ_1 に関する上限をとるため、 N_1 については、内側の楕円が、外円から離れて小さくなっており、その分標本数が無駄になってしまう。長軸と短軸の比が小さくなって楕円が円形に近づいたことによるメリットと、長径が短くなったことによるデメリットの差し引きによって Healy の N と我々の N_1 との優劣が決まる。



$\Gamma = \text{diag}(\sqrt{\frac{1-\rho_1}{1+\rho_1}}, \dots, \sqrt{\frac{1-\rho_q}{1+\rho_q}})$, $L = \begin{pmatrix} \frac{1}{2}(\Gamma^{-1} - \Gamma) \\ O \end{pmatrix} : r \times q$ とし、 $\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \sim N_p(\mathbf{0}, I_p)$,
 $\begin{pmatrix} T_1 & O \\ T_3 & T_2 \end{pmatrix} \begin{pmatrix} T_1 & O \\ T_3 & T_2 \end{pmatrix}' \sim W_p(m-1, I_p)$, ただし、 T_1, T_2 は上三角行列とすると、 $T_{m,q,r}$ の分布は

$$\{z_1'(T_1 T_1')^{-1} z_1 + (Lz_1 + z_2)' \{ (LT_1 + T_3)(LT_1 + T_3)' + T_2 T_2' \}^{-1} (Lz_1 + z_2)\} \quad (3.6)$$

の分布と一致する。 $p = 2q$ のとき、 $\rho_i \rightarrow 1 (i = 1, \dots, q)$ とすると

$$T_{m,q,q} \rightarrow 2(m-1)z_1'(T_1 T_1')^{-1} z_1 \sim \frac{2q(m-1)}{m-q} F_{q,m-q} \quad (\text{分布収束}) \quad (3.7)$$

となることから、 $c_{m,q,q}(\alpha; 1, \dots, 1) = \frac{2q(m-1)}{m-q} F_{q,m-q}(\alpha)$ であることがわかる。次の表4、表5は、 $\alpha = 0.05, p = 2, 4, m = 6, 8, 10, 15, 20, 30, 40$ に対して、 $c_{m,q,q}(\alpha; 1, \dots, 1)$ と、 $a_{m,p}(\alpha), b_{m,q,q}(\alpha) = c_{m,q,q}(\alpha; 0, \dots, 0)$ を比較したものである。 m が小さい場合には $c_{m,q,q}(\alpha; 1, \dots, 1) < a_{m,p}(\alpha)$ であるが、ある程度大きくなると、 $a_{m,p}(\alpha)$ の方が小さくなる。

表4 : $p = 2$

m	6	8	10	15	20	30	40
$a_{m,2}(0.05)$	17.4	12.0	10.0	8.2	7.5	6.9	6.7
$c_{m,1,1}(0.05; 1)$	13.2	11.2	10.2	9.2	8.8	8.4	8.2
$c_{m,1,1}(0.05; 0)$	11.2	9.3	8.3	7.4	7.0	6.7	6.4

表5 : $p = 4$

m	6	8	10	15	20	30	40
$a_{m,4}(0.05)$	192.5	44.7	27.2	17.1	14.3	12.2	11.4
$c_{m,2,2}(0.05; 1)$	34.7	24.0	20.0	16.4	15.0	13.8	13.3
$c_{m,2,2}(0.05; 0)$	31.0	20.1	16.5	13.2	12.1	11.0	10.6

4 数値実験

$c_{m,q,r}^*(\alpha)$ の数値計算は, $p = 2$ の場合でも困難である. ここでは, (3.6) を利用して, モンテカルロシミュレーションによって $c_{m,q,r}(0.05; \rho_1, \dots, \rho_q)$ の傾向を調べた. $p = 2, 3, 4$ とし, $(q, r) = (1, 1), (1, 2), (2, 2)$; $m = 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, 35, 40$ について以下の手順で計算を行った.

- (1) 疑似乱数を用いて, $(z_1, z_2, T_1, T_2, T_3)$ を 100,000 組発生させる.
- (2) $\gamma_i = \sqrt{\frac{1-\rho_i}{1+\rho_i}} = 0.0, 0.1, \dots, 1.0$ に対して, (3.6) により, $T_{m,q,r}$ を計算する.
- (3) 100,000 個の $T_{m,q,r}$ を小さい順に並べて $T_{(1)} < T_{(2)} < \dots < T_{(100,000)}$ とし, $c_{m,q,r}(0.05; \rho_1, \dots, \rho_q)$ を $T_{(95,000)}$ によって推定する.

表6, 表7は, $(q, r) = (1, 1), (1, 2)$ の場合の計算結果の抜粋である.

表6 : $(q, r) = (1, 1)$

$m \setminus \gamma$	0.0	0.1	0.2	0.3	0.4	0.5
5	15.62	<u>15.63</u>	15.6	15.31	14.95	14.49
6	<u>13.09</u>	13.00	12.89	12.54	12.16	11.91
7	<u>11.98</u>	11.95	11.79	11.49	11.12	10.78
8	<u>11.30</u>	11.23	11.02	10.79	10.45	10.09
9	<u>10.72</u>	10.66	10.47	10.17	9.82	9.48
10	<u>10.04</u>	9.97	9.83	9.54	9.24	8.95
20	<u>8.76</u>	8.69	8.53	8.27	7.97	7.65
30	<u>8.37</u>	8.31	8.13	7.88	7.57	7.26

表7 : $(q, r) = (1, 2)$

$m \setminus \gamma$	0.0	0.1	0.2	0.3	0.4	0.5
5	<u>32.43</u>	32.42	32.25	31.71	31.00	30.51
6	<u>22.90</u>	22.84	22.72	22.48	22.1	21.84
7	<u>19.20</u>	19.15	18.88	18.59	18.19	17.80
8	<u>16.72</u>	16.67	16.47	16.23	15.95	15.66
9	<u>15.30</u>	15.23	15.05	14.84	14.49	14.20
10	<u>14.52</u>	14.44	14.25	13.97	13.64	13.34
20	<u>11.28</u>	11.23	11.07	10.84	10.57	10.27
30	<u>10.44</u>	10.39	10.25	10.04	9.80	9.54

下線をつけた数値が, $\gamma_1 = 0.0, 0.1, \dots, 0.9, 1.0$ の中で最大であったものである. $(q, r) = (1, 1), m = 5$ の場合を除いて, $\gamma_1 = 0.0$ ($\rho_1 = 1$) のときに, 上側5%点は最大となっていることがわかる.

表8は, $(q, r) = (2, 2), m = 5$ の場合であるが, 上側5%点が最大となるのは, $\gamma_1 = \gamma_2 = 0.0$ ではなく, $\gamma_1 = \gamma_2 = 0.2$ ($\rho_1 = \rho_2 = 0.92$) のときである. $(q, r) = (2, 2)$ の場合, $m = 6, 7, 8$ までは, 同様に $(\gamma_1, \gamma_2) = (0, 0)$ 以外のところで上側5%点が最大となったが, $m \geq 9$ では, $(\gamma_1, \gamma_2) = (0, 0)$ で最大となった.

表8 : $(q, r) = (2, 2), m = 5$

$\gamma_1 \backslash \gamma_2$	0.0	0.1	0.2	0.3	0.4
0.0	51.41				
0.1	52.05	52.83			
0.2	53.09	53.27	<u>53.32</u>		
0.3	52.83	52.99	52.86	52.15	
0.4	52.54	52.18	51.96	51.45	51.84

$(q, r) = (1, 3)$ のときは, $m \leq 10$ では, $\gamma_1 = 0$ 以外で5%点が最大, $m \geq 15$ では, $\gamma_1 = 0$ で最大となった.

表9に, $p = 2, 3, 4$ のときの $c_{m,q,r}^*(0.05)$ と $a_{m,p}(0.05)$ を示す.

表9 : $a_{m,p}(0.05)$ と $c_{m,q,r}^*(0.05)$ の比較

m	5	6	7	8	9	10	15	20	30	40
$a_{m,2}(0.05)$	25.5	17.4	13.9	12.0	10.8	10.0	8.2	7.5	6.9	6.7
$c_{m,1,1}^*(0.05)$	15.6	13.9	12.0	11.3	10.7	10.0	9.2	8.8	8.4	8.2
$a_{m,3}(0.05)$	115.0	46.4	29.7	22.7	19.0	16.8	12.2	10.7	9.5	9.0
$c_{m,1,2}^*(0.05)$	32.4	22.9	19.2	16.7	15.3	14.5	12.0	11.3	10.4	10.1
$a_{m,4}(0.05)$	3593.0	192.5	72.9	44.7	33.2	27.2	17.1	14.3	12.2	11.4
$c_{m,2,2}^*(0.05)$	53.3	35.0	27.7	28.8	21.7	20.3	16.5	15.1	14.0	13.3
$c_{m,1,3}^*(0.05)$	122.0	50.0	32.6	25.3	21.4	20.6	15.7	14.0	12.7	12.2

$p = 2$ で $m \leq 10$, $p = 3$ で $m \leq 15$ なら, (3.5) で与えた N_1 が確率1で N より小さくなる. $p = 4$ の場合は, $(q, r) = (2, 2)$ と分割した場合には, $m = 15$ ぐらいまで, $(q, r) = (1, 3)$ と分割した場合には $m = 30$ ぐらいまでで, N_1 の方が確率1で小さい. N と N_1 の大小関係は, $a_{m,p}(\alpha), c_{m,q,r}^*(\alpha)$ だけでなく, 固有値 l_1 と l_* によっても左右される. 例えば $p = 2, m = 40$ のとき, $c_{m,1,1}^*(0.05)/a_{m,2}(0.05) = 1.22$ であるが, $S_m = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ とおくと, $a_{11} = a_{22}$ であって, 標本相関係数が, 0.22 以上であれば, $l_1/a_{11} \geq 1.22$ となり $N \geq N_1$ となる.

5 おわりに

前節で見たように, 変数を2組みに分割して (3.5) 式によって追加標本数を決める方法は第一段階での標本数 m を大きくとらない場合に有効となるが, l_1 や l_* は, 対応する Σ

の固有値の推定量として正の偏差を持っており、 m を小さく設定すると、 N や N_1 が平均的には大きくなってしまふことが予想される。指定された d と α に対して、 m をどう決めるかは、二段階法において未解決の問題である。

実際上の理由から m があらかじめ決まっているとき、 m に対してシミュレーション等で $a_{p,m}(\alpha)$ と $c_{p,q,r}(\alpha)$ を計算し、 $a_{p,m}(\alpha) \geq c_{p,q,r}(\alpha)$ となったならば、 N_1 を採用すればよいことがわかる。一方、 $a_{p,m}(\alpha) < c_{p,q,r}(\alpha)$ となった場合には、確率 1 で $N_1 > N$ が成り立つのではないし、期待標本数 $E(N_1)$ と $E(N)$ の大小も未知母数 Σ 次第である。 S_m から、 N と N_1 の両方を計算し、小さい方を採用するという方法では、残念ながら信頼係数が保証されなくなる。 $N_2 = \min\{N, N_1\}$ とすると

$$\begin{aligned} & \Pr\{\|\bar{x}_{N_2} - \mu\|^2 \leq d^2\} \\ &= \Pr\{\|\bar{x}_{N_2} - \mu\|^2 \leq d^2, N < N_1\} + \Pr\{\|\bar{x}_{N_2} - \mu\|^2 \leq d^2, N \geq N_1\} \\ &= \Pr\{N(\bar{x}_N - \mu)' S_m (\bar{x}_N - \mu) \leq N l_1^{-1} d^2, N \leq N_1\} \\ & \quad + \Pr\{N_1(\bar{x}_{N_1} - \mu)' S_m (\bar{x}_{N_1} - \mu) \leq N_1 l_1^{-1} d^2, N > N_1\} \end{aligned} \quad (5.1)$$

となり、 $N > N_1$ のとき $N_1 l_1^{-1} d^2 < \alpha_{m,p}(\alpha)$ なので (5.1) の第 1 項が、 $(1-\alpha) \Pr\{N \leq N_1\}$ より大きくても第 2 項が $(1-\alpha) \Pr\{N > N_1\}$ より小さくなるためである。

p が大きい場合に、変数を 3 組み以上に分割して S_m の各対角ブロックの最大固有値を用いて第二段階の標本数を定義することも考えられるが、この方法はあまり現実的ではない。例えば、 $p = 3$ として、1 次元ずつに分割して得られる統計量 $(T_{m,1,1,1})$ の分布は、3 つの相関係数に依存することになる。したがって、第 4 節のようにシミュレーションによって上側 $100\alpha\%$ 点の未知母数に関する上限を 0.1 きざみで計算しようとするとき、 11^3 個の点について 100,000 個の $T_{m,1,1,1}$ をソートしなければならず、計算量は膨大なものとなる。同じ理由から、変数を 2 組みに分割する場合も、 $(q, r) = (1, p-1)$ あるいは、 $(2, p-2)$ の分割が現実的である。 $E(N_1)$ の意味でどちらの分割が良いかは、未知母数に依存するが、表 9 の $p = 4$ の結果から予想すると、 $(q, r) = (1, p-1)$ ととる方が、 N を確実に改良できる m の範囲が広くなりそうである。

参考文献

Healy, W., C., Jr. Two-sample procedures in simultaneous estimation. *Ann. Math. Statist.* **27** (1956), 687-702.