数理解析研究所講究録
第197巻 1974年 12-22

12

カテゴリカル正準相関分析の漸近理論

大阪大学　基礎工学部　丘本　正
和歌山大学　教育学部　遠藤秀樹

§0. 序

対象×カテゴリのデータから、それらの対象及びカテゴリを数量化して、分類しようとする方法としてGuttman［4］のカテゴリカル主成分分析、林［5］の数量化理論第3類があり、これらは共に実際面において有用な手段と考えられるが、さらに解の信頼性、あるいは安定性の為の理論の欲しいところである。これに対し、青山［2］は数量化第3類について解の標本分散を不等式によって抑え、解の誤差評価をしている。ここではカテゴリカル主成分分析と数量化第3類の一般化であるカテゴリカル正準相関分析（丘本・遠藤［6］参照）について、1つの自然な確率モデルを想定して、解の漸近分散を求めることを問題とする。

§1. 記号及び確率モデルの導入

対象×カテゴリのパターン行列を

(1.1) (e_{ij}), $e_{ij} \geq 0$, $i=1,2,\cdots,n$; $j=1,2,\cdots,p$

とする。丘本・遠藤[6]より、同じ反応パターンを示した対象については、その行をプールしてもよいから、今m個の異なるタイプがあり、第iタイプに属した対象の数をn_i ($\sum_{i=1}^{m} n_i = n$)とすれば、解はタイプ×カテゴリの行列

(1.2) $(n_i e_{ij})$ $i=1,2,\cdots m$; $j=1,2,\cdots,p$

から求められる。また、全体を正の定数で除してもよいから、結局

(1.3) $\hat{E} = \left(\frac{n_i}{n} e_{ij}\right)$ $i=1,2,\cdots,m$; $j=1,2,\cdots,p$

から計算すればよい。さてこのデータの信頼性を測るため、次のようなモデルを考えよう。対象の無限母集団において、第iタイプの母集団比率をp_i ($i=1,2,\cdots,m$)とし、n個の対象がこの母集団からランダムに得られたものとする。このとき母集団におけるカテゴリカル正準相関分析は、パターン行列

(1.4) $E = (p_i e_{ij})$ $i=1,2,\cdots,m$; $j=1,2,\cdots,p$

に基づくことになる。従がって

(1.5)
$$F = diag(f_1, f_2, \cdots, f_m),\ f_i = \sum_{j=1}^{p} p_i e_{ij},\ i = 1, 2, \cdots, m$$
$$G = diag(g_1, g_2, \cdots, g_p),\ g_j = \sum_{i=1}^{m} p_i e_{ij},\ j = 1, 2, \cdots, p$$

とすれば、最適なウェイトベクトル y は

(1.6)
$$By = \rho^2 G y \quad \text{subject to} \quad y'Gy = 1$$

の解として得られ、スコアベクトル x は

(1.7)
$$x = \frac{1}{\rho} F^{-1} E y$$

として得られることになる。但し

(1.8)
$$B = E'F^{-1}E = (b_{jk}),\ b_{jk} = \sum_{i=1}^{m} \frac{p_i e_{ij} e_{ik}}{l_i},$$
$$l_i = \sum_{j=1}^{p} e_{ij},\ i = 1, 2, \cdots, m;\ j, k = 1, 2, \cdots, p$$

である。今、推定値には記号へを付けるものとすれば

(1.9)
$$\hat{p}_i = \frac{n_i}{n},\ \hat{f}_i = \sum_{j=1}^{p} \hat{p}_i e_{ij},\ \hat{g}_j = \sum_{i=1}^{m} \hat{p}_i e_{ij},$$
$$\hat{b}_{jk} = \sum_{i=1}^{m} \frac{\hat{p}_i e_{ij} e_{ik}}{l_i},\ i = 1, 2, \cdots, m;\ j, k = 1, 2, \cdots, p$$

である。従がって、標本における最適なウェイト $\hat{y}$ は

(1.10)
$$\hat{B}\hat{y} = \hat{\rho}^2 \hat{G} \hat{y} \quad \text{subject to} \quad \hat{y}'\hat{G}\hat{y} = 1$$

より決定され、最適スコア $\hat{x}$ は

(1.11)　$\hat{x} = \frac{1}{\hat{\rho}} \hat{F}^{-1} \hat{E} \hat{y}$

より求められる。さて、δ－法（Anderson［1］, Doob［3］参照）により、$\hat{\rho}^2 = \hat{\lambda}$, $\hat{y}$ の漸近分散を求めてみよう。

§2．δ－法

$p \leq m$ を仮定する。第 r 母集団固有値 λ_r（$r = 1, 2, \cdots, p$）に対応するスコアベクトルを x_r，ウェイトベクトルを y_r とし、行列 X, Y, Λ を

$$(2.1) \qquad \begin{aligned} X &= (x_1, \cdots, x_p), \quad Y = (y_1, \cdots, y_p) \\ \Lambda &= diag(\lambda_1, \cdots, \lambda_p) \end{aligned}$$

と定義し、同様に標本についても

$$(2.2) \qquad \begin{aligned} \hat{X} &= (\hat{x}_1, \cdots, \hat{x}_p), \quad \hat{Y} = (\hat{y}_1, \cdots, \hat{y}_p) \\ \hat{\Lambda} &= diag(\hat{\lambda}_1, \cdots, \hat{\lambda}_p) \end{aligned}$$

とおく。このとき、（1.6），（1.7）式は

$$(2.3) \qquad B Y = G Y \Lambda \qquad subject\ to\ \ Y' G Y = I_p$$

$$(2.4) \qquad X = F^{-1} E Y \Lambda^{-\frac{1}{2}}$$

であり、（1.10），（1.11）式は

(2.5) $\hat{B}\hat{\Upsilon} = \hat{G}\hat{\Upsilon}\hat{\Lambda}$ subject to $\hat{\Upsilon}'\hat{G}\hat{\Upsilon} = I_p$

(2.6) $\hat{X} = \hat{F}^{-1}\hat{E}\hat{\Upsilon}\hat{\Lambda}^{-\frac{1}{2}}$

となる。今、

(2.7) $\hat{B} = B + B^*,\ \hat{G} = G + G^*,\ \hat{\Upsilon} = \Upsilon + \Upsilon^*,\ \hat{\Lambda} = \Lambda + \Lambda^*$

により、$B^*, G^*, \Upsilon^*, \Lambda^*$を定義すれば、(2.5)式から

(2.8) $(B+B^*)(\Upsilon+\Upsilon^*) = (G+G^*)(\Upsilon+\Upsilon^*)(\Lambda+\Lambda^*)$

(2.9) $(\Upsilon+\Upsilon^*)'(G+G^*)(\Upsilon+\Upsilon^*) = I_p$

となる。δ-法により、*印の2次以上の項を無視すれば、

(2.10) $B\Upsilon + B\Upsilon^* + B^*\Upsilon = G\Upsilon\Lambda + G\Upsilon\Lambda^* + G\Upsilon^*\Lambda + G^*\Upsilon\Lambda$

となる。左からΥ'を乗じ、$\Upsilon'G\Upsilon = I_p$, $B\Upsilon = G\Upsilon\Lambda$を使えば

(2.11) $\Upsilon'B\Upsilon^* + \Upsilon'B^*\Upsilon = \Lambda^* + \Upsilon'G\Upsilon^*\Lambda + \Upsilon'G^*\Upsilon\Lambda$

が得られる。ここで、

(2.12) $\Upsilon^* = \Upsilon Z$

により、Zを定義すれば、(2.11)式は、

(2.13) $$\Lambda Z + Y'B^*Y = \Lambda^* + Z\Lambda + Y'G^*Y\Lambda$$

となる。両辺の $r-r$ 要素を考えることにより

(2.14) $$\begin{aligned}\lambda_r^* &= (Y'B^*Y)_{rr} - \lambda_r (Y'G^*Y)_{rr} \\ &= \sum_{j=1}^{p}\sum_{k=1}^{p} y_{jr}y_{kr}b_{jk}^* - \lambda_r \sum_{j=1}^{p} y_{jr}^2 g_j^*\end{aligned}$$

が得られる。さらに(2.13)式の $s-r$ 要素を考えれば、

(2.15) $$\begin{aligned}z_{sr} &= (\lambda_s - \lambda_r)^{-1}\{\lambda_r (Y'G^*Y)_{sr} - (Y'B^*Y)_{sr}\} \\ &= (\lambda_s - \lambda_r)^{-1}\left(\lambda_r \sum_{j=1}^{p} y_{jr}y_{js}g_j^* - \sum_{j=1}^{p}\sum_{k=1}^{p} y_{jr}y_{ks}b_{jk}^*\right)\end{aligned}$$

が得られる。また、(2.9)式に上と同様の計算をすれば、

(2.16) $$Z + Z' + Y'G^*Y = 0$$

となる。この両辺の対角要素を考えれば、

(2.17) $$z_{rr} = -\frac{1}{2}(Y'G^*Y)_{rr} = -\frac{1}{2}\sum_{j=1}^{p} y_{jr}^2 g_j^*$$

が得られる。

§3. 標本固有値、最適ウェイトの漸近分散

[定理] 第 r 標本固有値 $\hat{\lambda}_r$、最適ウェイト $\hat{g}_r$ の漸近分散、共分散は次のようになる。

(3.1)
$$V(\hat{\lambda}_r) = \frac{1}{n}\sum_{i=1}^{m} p_i c_{ir}^2$$
$$Cov(\hat{y}_{jr}\ \hat{y}_{kr}) = \frac{1}{n}\left\{\sum_{i=1}^{m} p_i d_{ijr} d_{ikr} - \left(\sum_{i=1}^{m} p_i d_{ijr}\right)\left(\sum_{i=1}^{m} p_i d_{ikr}\right)\right\}$$

但し、

$$c_{ir} = \lambda_r\left(l_i x_{ir}^2 - \sum_{j=1}^{p} y_{jr}^2 e_{ij}\right)$$

(3.2)
$$d_{ijr} = \sum_{s=1}^{p} y_{js} c_{isr}$$
$$c_{isr} = \begin{cases} \dfrac{1}{\lambda_s - \lambda_r}\left(\lambda_r \sum_{j=1}^{p} y_{jr} y_{js} e_{ij} - \sqrt{\lambda_s \lambda_r}\, l_i x_{is} x_{ir}\right) & \text{if } s \neq r \\ -\dfrac{1}{2}\sum_{j=1}^{p} y_{jr}^2 e_{ij} & \text{if } s = r \end{cases}$$

とする。

(証明)　(2.14)式の b_{jk}^*, g_j^* と(2.7)式の定義に従がって $\hat{p}_i$ で表わせば、

(3.3)
$$\lambda_r^* = \sum_{j=1}^{p}\sum_{k=1}^{p} y_{jr} y_{kr} \sum_{i=1}^{m} (\hat{p}_i - p_i)\frac{e_{ij} e_{ik}}{l_i} - \lambda_r \sum_{j=1}^{p} y_{jr}^2 \sum_{i=1}^{m} (\hat{p}_i - p_i) e_{ij}$$
$$= \sum_{i=1}^{m}\left(\sum_{j=1}^{p}\sum_{k=1}^{p} y_{jr} y_{kr} \frac{e_{ij} e_{ik}}{l_i} - \lambda_r \sum_{j=1}^{p} y_{jr}^2 e_{ij}\right)(\hat{p}_i - p_i)$$
$$= \sum_{i=1}^{m} c_{ir}(\hat{p}_i - p_i) \qquad (\because (2.6), (3.2) \text{による})$$

ところで $(n_1, n_2, \cdots, n_m)$ は多項分布であるから、

(3.4)
$$E(\hat{p}_i - p_i) = 0 \qquad V(\hat{p}_i - p_i) = \frac{1}{n} p_i (1 - p_i)$$
$$Cov(\hat{p}_i - p_i,\ \hat{p}_h - p_h) = -\frac{1}{n} p_i p_h$$

である。従がって、

$$(3.5)\qquad V(\hat{\lambda}_r) = E(\lambda_r^{*2}) = \sum_{i=1}^{m}\sum_{h=1}^{m} c_{ir}c_{hr}E(\hat{p_i}-p_i)(\hat{p}_h-p_h)$$
$$= \sum_{i=1}^{m} c_{ir}^2 \frac{1}{n} p_i(1-p_i) - \sum_{i\neq h} c_{ir}c_{hr}\frac{1}{n}p_i p_h$$
$$= \frac{1}{n}\sum_{i=1}^{m} p_i c_{ir}^2 .$$

ここで、$\sum_{i=1}^{m} p_i c_{ir} = 0$ が成立することに注意しておく。さらに、(2.15) (2.17)式より、z_{sr}、z_{rr} を $\hat{p}_i$ で表わせば、

$$(3.6)\qquad z_{sr} = \sum_{i=1}^{m}\frac{1}{\lambda_s-\lambda_r}\left(\lambda_r\sum_{j=1}^{p} y_{jr}y_{js}e_{ij} - \sum_{j=1}^{p}\sum_{k=1}^{p} y_{js}y_{kr}\frac{e_{ij}e_{ik}}{l_i}\right)(\hat{p}_i-p_i)$$
$$= \sum_{i=1}^{m} c_{isr}(\hat{p}_i-p_i)$$

$$(3.7)\qquad z_{rr} = \sum_{i=1}^{m}\left(-\frac{1}{2}\sum_{j=1}^{p} y_{jr}^2 e_{ij}\right)(\hat{p}_i-p_i) = \sum_{i=1}^{m} c_{irr}(\hat{p}_i-p_i)$$

となる。さて、(2.12)式より

$$(3.8)\qquad y_r^* = Y z_r \qquad \text{但し } z = (z_1, \cdots, z_p) .$$

従がって

$$(3.9)\qquad y_{jr}^* = \sum_{s=1}^{p} y_{js}z_{sr} = \sum_{s=1}^{p} y_{js}\sum_{i=1}^{m} c_{isr}(\hat{p}_i-p_i)$$
$$= \sum_{i=1}^{m}\left(\sum_{s=1}^{p} y_{js}c_{isr}\right)(\hat{p}_i-p_i) = \sum_{i=1}^{m} d_{ijr}(\hat{p}_i-p_i)$$

となる。(3.4)を用いて前と同様の計算をすることによって、

$$(3.10)\qquad Cov(\hat{y}_{jr}, \hat{y}_{kr}) = E(y_{jr}^{*} y_{kr}^{*})$$
$$= \frac{1}{n}\left\{\sum_{i=1}^{m} P_i d_{ijr} d_{ikr} - \left(\sum_{i=1}^{m} P_i d_{ijr}\right)\left(\sum_{i=1}^{m} P_i d_{ikr}\right)\right\}$$

が得られる。

§4．数値例

前節の結果に基づき、青山［2］の例について、$\hat{\lambda}_2$．$\hat{\lambda}_3$の漸近分散、$\hat{y}_2$の漸近共分散を，また林［5］の例については$\hat{\lambda}_2$，$\hat{\lambda}_3$の漸近分散を計算すれば，下のようになる。

§4－1．青山の例

カテゴリ / タイプ	1	2	3	4	5	人数
1	1	1	0	1	0	3
2	1	0	1	0	0	4
3	0	0	1	1	1	3

これについて，$\hat{\lambda}$，$\hat{y}$．$\hat{x}$を求めれば，

$$\hat{\lambda}_1 = 1 \quad \hat{y}_1' = \frac{\sqrt{65}}{13}(1, 1, 1, 1, 1) \qquad \hat{x}_1' = \frac{\sqrt{650}}{130}(1, 1, 1)$$
$$\hat{\lambda}_2 = \frac{10}{21} \quad \hat{y}_2' = \frac{\sqrt{42}}{42}(3, 7, -3, 0, -7) \qquad \hat{x}_2' = \frac{\sqrt{5}}{3}(1, 0, -1)$$
$$\hat{\lambda}_3 = \frac{8}{21} \quad \hat{y}_3' = \frac{\sqrt{2730}}{546}(-6, 7, -6, 7, 7) \qquad \hat{x}_3' = \frac{2\sqrt{65}}{39}\left(1, -\frac{9}{4}, 1\right)$$

となる。これより，（3．1）を用いて漸近分散を計算すれば，

$$V(\hat{\lambda}_2) = 0.00278$$
$$V(\hat{\lambda}_3) = 0.01110$$

が得られる。また，ウェイトの漸近共分散は，

$$E(y_2^* y_2^{*\prime}) = \begin{pmatrix} 0.16516 & -0.06106 & 0.16283 & -0.06378 & -0.06650 \\ & 0.03115 & -0.06650 & 0.02480 & 0.01845 \\ & & 0.16516 & -0.06378 & -0.06106 \\ & & & 0.02480 & 0.02480 \\ & & & & 0.03115 \end{pmatrix}$$

と計算される。

§4－2．林の例

林［5］の例については

$$\hat{\lambda}_2 = 0.3262 \qquad \hat{\lambda}_3 = 0.2737$$

$$\hat{y}_2' = (1.1502,\ 0.5313,\ 0.4984,\ 0.3998,\ 0.0110,$$
$$-0.2191,\ -0.2958,\ -0.3725,\ -0.5368,\ -2.1033)$$

$$\hat{y}_3' = (0.8380,\ -0.3122,\ 0.2410,\ 0.5587,\ -0.3505,$$
$$-0.4163,\ 0.0822,\ -0.2958,\ -0.2520,\ 2.5907)$$

と計算され，これより

$$V(\hat{\lambda}_2) = 0.00278$$

$$V(\hat{\lambda}_3) = 0.00532$$

と求められる。

参考文献

［1］ Anderson, T.W. (1951), The asymptotic distribution of certain characteristic roots and vectors, Second

Berkeley Symp., 1 103 - 130.

[2] 青山博次郎 (1965), ダミー変数と数量化への応用, 統数研彙報, 第13巻第1号, 1 - 12 (訂正, 統数研彙報, 第13巻第2号, 135 - 137).

[3] Doob, J. L. (1935), The limiting distributions of certain statistics, Ann. Math. Statis. 6, 160 - 169.

[4] Guttman, L. (1941), The quantification of a class of attributes : A theory and method of scal construction, The Prediction of Personal Adjustment (ed. P. Horst), 319 - 348, Soc. Sci Res. Conal, New York

[5] 林知己夫 (1956). 数量化理論とその応用例 (II), 統数研彙報, 第4巻第2号, 19 - 30

[6] 立本正・遠藤秀機, Basic properties of categorical canonical correlation analysis, 投稿中