

カテゴリー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×カテゴリーのデータから、それらの対象及びカテゴリーを数量化して、分類しようとする方法として Guttman [4] のカテゴリーカル主成分分析、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。ここではカテゴリー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ーカル正準相関分析（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×カテゴリーのパターン行列を

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

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 $\times$ カテゴリの行列

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

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

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

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 ($i=1, 2, \dots, m$) 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

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

に基づくことになる。従って

$$(1.5) \quad F = \text{diag}(f_1, f_2, \dots, f_m), \quad f_i = \sum_{j=1}^p p_i e_{ij}, \quad i=1, 2, \dots, m$$

$$G = \text{diag}(g_1, g_2, \dots, g_p), \quad g_j = \sum_{i=1}^m p_i e_{ij}, \quad j=1, 2, \dots, p$$

とすれば、最適なウェイトベクトル Y は

$$(1.6) \quad BY = P^2 G Y \quad \text{subject to} \quad Y' G Y = 1$$

の解として得られ、スコアベクトル X は

$$(1.7) \quad X = \frac{1}{P} F^{-1} E Y$$

として得られることになる。但し

$$(1.8) \quad B = E' F^{-1} E = (b_{jk}), \quad b_{jk} = \sum_{i=1}^m \frac{p_i e_{ij} e_{ik}}{f_i},$$

$$f_i = \sum_{j=1}^p e_{ij}, \quad i=1, 2, \dots, m; \quad j, k=1, 2, \dots, p$$

である。今、推定値には記号 $\hat{\cdot}$ を付けるものとするれば

$$(1.9) \quad \hat{p}_i = \frac{n_i}{n}, \quad \hat{f}_i = \sum_{j=1}^p \hat{p}_i e_{ij}, \quad \hat{g}_j = \sum_{i=1}^m \hat{p}_i e_{ij},$$

$$\hat{b}_{jk} = \sum_{i=1}^m \frac{\hat{p}_i e_{ij} e_{ik}}{\hat{f}_i}, \quad i=1, 2, \dots, m; \quad j, k=1, 2, \dots, p$$

である。従って、標本における最適なウェイト $\hat{Y}$ は

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

より決定され、最適スコア $\hat{X}$ は

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

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 ($r=1, 2, \dots, P$) 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

$$(2.1) \quad \begin{aligned} X &= (x_1, \dots, x_P), \quad Y = (y_1, \dots, y_P) \\ \Lambda &= \text{diag}(\lambda_1, \dots, \lambda_P) \end{aligned}$$

と定義し、同様に標本についても

$$(2.2) \quad \begin{aligned} \hat{X} &= (\hat{x}_1, \dots, \hat{x}_P), \quad \hat{Y} = (\hat{y}_1, \dots, \hat{y}_P) \\ \hat{\Lambda} &= \text{diag}(\hat{\lambda}_1, \dots, \hat{\lambda}_P) \end{aligned}$$

とおく。このとき、(1.6), (1.7) 式は

$$(2.3) \quad BY = GY\Lambda \quad \text{subject to} \quad Y'GY = I_P$$

$$(2.4) \quad X = F^{-1}EY\Lambda^{-\frac{1}{2}}$$

であり、(1.10), (1.11) 式は

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

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

となる。今、

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

により、 B^*, G^*, Y^*, Λ^* を定義すれば、(2.5)式から

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

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

となる。 δ -法により、 $*$ 印の2次以上の項を無視すれば、

$$(2.10) \quad BY + BY^* + B^*Y = GY\Lambda + GY\Lambda^* + GY^*\Lambda + G^*Y\Lambda$$

となる。左から Y' を乗じ、 $Y'GY = I_p$, $BY = GY\Lambda$ を使えば

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

が得られる。ここで、

$$(2.12) \quad Y^* = YZ$$

により、 Z を定義すれば、(2.11)式は、

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

となる。両辺の $r - r$ 要素を考えることにより

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

が得られる。さらに (2.13) 式の $s - r$ 要素を考えれば、

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

が得られる。また、(2.9) 式に上と同様の計算をすれば、

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

となる。この両辺の対角要素を考えれば、

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

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 $\hat{y}_r$ の漸近分散、共分散は次のようになる。

$$(3.1) \quad V(\hat{\lambda}_r) = \frac{1}{n} \sum_{i=1}^m p_i c_{ir}^2$$

$$\text{Cov}(\hat{y}_{jr}, \hat{y}_{kr}) = \frac{1}{n} \left\{ \sum_{i=1}^m p_i d_{ijr} d_{ikr} - \left(\sum_{i=1}^m p_i d_{ijr} \right) \left(\sum_{i=1}^m p_i d_{ikr} \right) \right\}$$

但し、

$$(3.2) \quad c_{ir} = \lambda_r \left(l_i x_{ir}^2 - \sum_{j=1}^p y_{jr}^2 e_{ij} \right)$$

$$d_{ijr} = \sum_{s=1}^p y_{js} c_{isr}$$

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

とする。

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

$$(3.3) \quad \lambda_r^* = \sum_{j=1}^p \sum_{k=1}^p y_{jr} y_{kr} \sum_{i=1}^m (\hat{p}_i - p_i) \frac{e_{ij} e_{ik}}{l_i} - \lambda_r \sum_{j=1}^p y_{jr}^2 \sum_{i=1}^m (\hat{p}_i - p_i) e_{ij}$$

$$= \sum_{i=1}^m \left(\sum_{j=1}^p \sum_{k=1}^p y_{jr} y_{kr} \frac{e_{ij} e_{ik}}{l_i} - \lambda_r \sum_{j=1}^p y_{jr}^2 e_{ij} \right) (\hat{p}_i - p_i)$$

$$= \sum_{i=1}^m c_{ir} (\hat{p}_i - p_i) \quad (\because (2.6), (3.2) \text{ による})$$

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

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

$$\text{Cov}(\hat{p}_i - p_i, \hat{p}_h - p_h) = -\frac{1}{n} p_i p_h$$

である。従って、

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

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、

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

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

となる。さて、(2.12) 式より

$$(3.8) \quad y_r^* = Y Z_r \quad \text{但し } Z = (Z_1, \dots, Z_p).$$

従って

$$\begin{aligned}
 (3.9) \quad y_{jr}^* &= \sum_{s=1}^p y_{js} Z_{sr} = \sum_{s=1}^p y_{js} \sum_{i=1}^m c_{isr} (\hat{p}_i - p_i) \\
 &= \sum_{i=1}^m \left(\sum_{s=1}^p y_{js} c_{isr} \right) (\hat{p}_i - p_i) = \sum_{i=1}^m d_{ijr} (\hat{p}_i - p_i)
 \end{aligned}$$

となる。(3.4) を用いて前と同様の計算をすることによつて、

$$\begin{aligned}
 (3.10) \quad \text{Cov}(\hat{y}_{jr}, \hat{y}_{kr}) &= E(y_{jr}^* y_{kr}^*) \\
 &= \frac{1}{n} \left\{ \sum_{i=1}^m p_i d_{ijr} d_{ikr} - \left(\sum_{i=1}^m p_i d_{ijr} \right) \left(\sum_{i=1}^m p_i d_{ikr} \right) \right\}
 \end{aligned}$$

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、 $\hat{y}_2$ の漸近共分散と、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

カテゴリー 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、

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

となる。これより、(3.1)を用いて漸近分散を計算すれば、

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

$$V(\hat{\lambda}_3) = 0.0110$$

が得られる。また、ウェイトの漸近共分散は、

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

と計算される。

§4-2. 林の例

林[5]の例については

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

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

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

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, 統計学彙報, 第13巻第1号, 1 - 12 (訂正, 統計学彙報, 第13巻第2号, 135 - 137).

[3] Doob, J. L. (1935), The limiting distributions of certain statistics, Ann. Math. Statist. 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. Council, New York

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

[6] 丘本正・遠藤秀敏, Basic properties of categorical canonical correlation analysis. 投稿中