北海道大学工学部研究報告
第97号（昭和55年）

Bulletin of the Faculty of Engineering,
Hokkaido University, No. 97 (1980)

潜在クラス分析における誤差の評価

佐藤義治* 須川和明* 河口至商*

（昭和54年9月29日受理）

On an Error Estimation in the Latent Class Analysis

Yoshiharu SATO Kazuaki SUGAWA Michiaki KAWAGUCHI

(Received September 29, 1979)

Abstract

On the Latent Class Analysis, M. Okamoto pointed out that most of the solutions, e.g. Gibson method, Green method and modified Green method, are unstable in the following sense that the errors of the latent parameters are extremely large in comparison with the observational errors.

In this paper, the errors of the parameters are estimated from the different point of view, that is, from the differential geometrical structure of the statistical parameter space. And it will be shown that this instability is caused by the relative expansion and contraction between the observable parameter space and the latent parameter space rather than the solutions.

1. 序　　論

潜在クラス分析の解の安定性について丘本・磯貝[1]は次のような指摘を行なっている。潜在クラス分析の解法であるGreen法，Gibson法，修正Green法により推定される解はデータに含まれる誤差に比較して，非常に大きな誤差を含むこと，また，Signature，Stratifierの選択の仕方による誤差が存在すること等である。さらにGibson法による誤差の評価も行なっている。

本研究は，データの誤差が解に及ぼす影響について，一般的に潜在クラス分析のモデルに含まれるパラメータからなる空間を通して検討したものである。

データの誤差に比較して解の誤差に大きな変動が生ずるのは，データの空間と潜在パラメータからなる空間の相対的な伸縮が大きな原因のひとつであり，その大きさの各点の近傍での変化は空間の法曲率及びその平均曲率の大きさで評価することができることを示す。実際の計算実験を質問項目4，クラス数2の場合について具体的に述べる。

2. 潜在クラス分析のモデル

ひとつの集団がいくつかの潜在的な等質部分集団から構成されていると予想される場合がある。潜在クラス分析とは，このような集団に対して適当な数の2値的な質問（Yes又はNoと応答する）を行なうことによって，潜在的な部分集団の確率構造を推定しようとするものである。

* 情報工学専攻　情報数理工学第一講座

質問項目の数を k とするとき，応答パターンを $(x_1, x_2, \cdots, x_k)$ と表わすことにすれば（ここで，質問 i に Yes と応答した時は $x_i=1$，No と応答した時は $x_i=0$ とする。）応答パターンは 2^k 個ある。$(x_1, x_2, \cdots, x_k)$ と応答される確率を $P(x_1, \cdots, x_k)=\phi^\lambda(\lambda=1, \cdots, 2^k)$ とすれば，$\sum_\lambda \phi^\lambda=1$ であり，$(x_1, \cdots, x_k)$ は ϕ^λ をパラメータとする多項分布をすることが知られている。ここで考えている集団が m 個の等質な部分集団から構成されていると仮定し，更に，部分集団内では k 個の質問項目に対する応答が確率的に独立であると仮定（局所独立性の仮定）すれば，潜在クラス分析の基本的な方程式が得られる。つまり，各部分集団の構成比を $w^t(t=1, 2, \cdots, m;\ \sum_t w^t=1)$ 部分集団 t において質問項目 i に Yes と応答する確率を π_i^t と表わすことにすれば，ϕ^λ は

$$\phi^\lambda = P(x_1, \cdots, x_k) = \sum_{t=1}^{m} w^t \prod_{i=1}^{k} (\pi_i^t)^{x_i}(1-\pi_i^t)^{1-x_i} \qquad (\lambda = 1, 2, \cdots, 2^k) \tag{2.1}$$

となる。N 個の個体から成る集団に対して，データとして得られた各応答パターンの反応数を n^λ $(\lambda=1, 2, \cdots, 2^k;\ \sum n^\lambda=N)$ とするとき，$\hat{\phi}^\lambda=n^\lambda/N$ は ϕ^λ の最尤推定値であるから，この $\hat{\phi}^\lambda$ を基にして (2.1) 式を満たす $\hat{w}^\lambda$, $\hat{\pi}_i^t$ を推定することができる。

潜在パラメータと呼ばれる w^t, π_i^t を推定する方法は，いくつか提案されているが，その多くは，(2.1) 式と同値でモーメント方程式と呼ばれる式を用いている。$P(x_1, \cdots, x_k)$ の周辺確率を $P(x_i=1)=p_i$, $P(x_i=1,\ x_j=1)=p_{ij}$ 等と表わせば，モーメント方程式とは，

$$\begin{aligned}
&\sum_{t=1}^{m} w^t = 1 \\
&p_i = \sum_{t=1}^{m} w^t \pi_i^t \\
&p_{ij} = \sum_{t=1}^{m} w^t \pi_i^t \pi_j^t \\
&p_{ijk} = \sum_{t=1}^{m} w^t \pi_i^t \pi_j^t \pi_k^t \\
&\cdots\cdots\cdots\cdots \\
&\cdots\cdots\cdots\cdots
\end{aligned} \tag{2.2}$$

であり，上と同様に n^λ からの推定値 $\hat{p}_i$, $\hat{p}_{ij}$ 等から潜在パラメータ $\hat{w}_t$, $\hat{\pi}_i^t$ を推定することができる。特に Green, Gibson, Anderson 等の推定法では，局所独立性の仮定を 3 次までにして，3 次までのモーメント方程式だけを用いて潜在パラメータを推定している。

3. 潜在パラメータの空間

潜在パラメータの空間を考えるにあたって以下では，質問項目の数を 4，潜在クラスの数を 2 とする。このとき応答パターンの数は $2^4=16$ であり，その確率 $\phi^\lambda(\lambda=1, 2, \cdots, 16)$ は $\sum_{\lambda=1}^{16} \phi^\lambda=1$ であるから，独立なものは 15 である。第 2 章で述べたように ϕ^λ は多項分布のパラメータであるから，ϕ^λ のパラメータ空間に Fisher の情報行列を計量テンソルとするように計量を定義すれば，この空間は 15 次元の Riemann 空間となる。このとき計量テンソルは，

$$g_{\lambda\mu} = \begin{cases} \dfrac{1}{\phi^\lambda} + \dfrac{1}{\phi^{16}} & (\lambda = \mu) \\ \dfrac{1}{\phi^{16}} & (\lambda \neq \mu) \end{cases} \tag{3.1}$$

であり，正の定曲率空間であることが知られている[2]。

潜在クラスの数が2で，質問項目の数が4の場合の潜在パラメータは w^1, w^2, π_i^1, π_i^2 $(i=1, \cdots, 4)$ であり，$w^1+w^2=1$ であるから独立なものは全部で9である。以後の記述を簡便にするために，次の様な記号を用いることにする。

$$\begin{aligned} \theta^i &= \pi_i^1 \qquad (i=1, 2, \cdots, 4) \\ \theta^{i+4} &= \pi_i^2 \qquad (i=1, 2, \cdots, 4) \\ \theta^9 &= w^1 \end{aligned} \tag{3.2}$$

さらに $\bar{\theta}^i=1-\theta^i$ $(i=1, 2, \cdots, 9)$ と表わすことにする。この記号によって前章の基本方程式(2.1)は次の様に表わせる。

$$\begin{aligned} \phi^1 &= \theta^1\theta^2\theta^3\theta^4\theta^9+\theta^5\theta^6\theta^7\theta^8\bar{\theta}^9 \\ \phi^2 &= \theta^1\theta^2\theta^3\bar{\theta}^4\theta^9+\theta^5\theta^6\theta^7\bar{\theta}^8\bar{\theta}^9 \\ \phi^3 &= \theta^1\theta^2\bar{\theta}^3\theta^4\theta^9+\theta^5\theta^6\bar{\theta}^7\theta^8\bar{\theta}^9 \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \phi^{15} &= \bar{\theta}^1\bar{\theta}^2\bar{\theta}^3\theta^4\theta^9+\bar{\theta}^5\bar{\theta}^6\bar{\theta}^7\theta^8\bar{\theta}^9 \end{aligned} \tag{3.3}$$

上の式から，潜在パラメータ θ^i $(i=1, 2, \cdots, 9)$ のパラメータ空間は，多項分布のパラメータ空間の部分空間となっていることが判る。(3.3)式のヤコビアンを $B_i^\lambda \equiv \partial\phi^\lambda/\partial\theta^i$ とすれば，これによって部分空間に計量が誘導され，潜在パラメータの空間もこの計量によって Riemann 空間となる。この場合の誘導された計量テンソルは，

$$g_{ij} = B_i^\lambda B_j^\mu g_{\lambda\mu} = \sum_{\alpha=1}^{16} \frac{B_i^\alpha B_j^\alpha}{\phi^\alpha} \tag{3.4}$$

であり，その逆行列 g^{ij} から $B_\lambda^i=g^{ij}g_{\lambda\mu}B_j^\mu$ が計算できる。さらにオイラー・スカウテンの曲率テンソル

$$H_{ij}^\lambda \equiv \nabla_i B_j^\lambda = \frac{\partial B_j^\lambda}{\partial\theta^i} - \left\{ \begin{matrix} l \\ ij \end{matrix} \right\} B_l^\lambda + B_i^\mu \left\{ \begin{matrix} \lambda \\ \mu\nu \end{matrix} \right\} B_j^\nu \tag{3.5}$$

から平均曲率ベクトル $H^\lambda=g^{ij}H_{ij}^\lambda/9$ と平均曲率 $H=g_{\lambda\mu}H^\lambda H^\mu$ が計算できる。また同様にリーマン・クリストッフェルの曲率テンソル

$$R_{ijkl} = R_{\lambda\mu\nu\omega}B_i^\lambda B_j^\mu B_k^\nu B_l^\omega + H_{\alpha jk}H_{il}^\alpha - H_{\beta jl}H^\beta{}_{ik} \tag{3.6}$$

や，スカラー曲率 $R=R_{ijkl}g^{il}g^{jk}$ 等も合せて，これらの量によって潜在パラメータの空間の局所的な性質を知ることができる。

このような観点から，潜在パラメータの最適な推定値とは，データから推定された $\hat{\phi}^\lambda$ に対して，$\hat{\phi}^\lambda$ と $\phi^\lambda=\phi^\lambda(\theta^i)$ の距離を最小にするような $\hat{\theta}^i$ であると考えることができる。

4. 誤差の評価

母数の真値を

$$\phi_0^\lambda = \phi_0^\lambda(\theta_0^i) \tag{4.1}$$

とするとき，これにある誤差 (n^λ) を含むデータを

$$\phi^\lambda = \phi_0^\lambda + n^\lambda \tag{4.2}$$

とする。この誤差によって生ずる潜在パラメータの誤差を (s^i) すなわち

$$\hat{\theta}^i = \theta_0^i + s^i \tag{4.3}$$

とすると，s^i は

$$s^i = B^i_\lambda n^\lambda \tag{4.4}$$

として評価することができる。g_{ij} は $g_{\lambda\mu}$ の誘導計量であるから，

$$g_{\lambda\mu}n^\lambda n^\mu \geqq g_{ij}s^i s^j \tag{4.5}$$

なる関係がある。

つぎに，これら2つの誤差 (n^λ), (s^i) の平均2乗誤差を考える。各平均2乗誤差は

$$\begin{aligned} \mathrm{AED} &= \left\{\frac{1}{15}\,\delta_{\lambda\mu}n^\lambda n^\mu\right\}^{\frac{1}{2}} = \left\{\frac{1}{15}\sum_\lambda (n^\lambda)^2\right\}^{\frac{1}{2}} \\ \mathrm{AEP} &= \left\{\frac{1}{9}\,\delta_{ij}s^i s^j\right\}^{\frac{1}{2}} = \left\{\frac{1}{9}\sum_i (s^i)^2\right\}^{\frac{1}{2}} \end{aligned} \tag{4.6}$$

と表わされる。

これらの大きさは，空間の metric 構造 $g_{ij}=B^\lambda_i B^\mu_j g_{\lambda\mu}$ によって定まるものであり，

$$\frac{\mathrm{AEP}}{\mathrm{AED}} \approx \frac{(\sqrt{g})^{\frac{1}{15}}}{(\sqrt{\bar{g}})^{\frac{1}{9}}} \tag{4.7}$$

なる関係にある。ただし，g, $\bar{g}$ はそれぞれ

$$g = \det(g_{\lambda\mu}), \qquad \bar{g} = \det(g_{ij}) \tag{4.8}$$

であり，すなわち，2乗誤差の比は各空間の相対的伸縮に密接に関係するものである。潜在クラス分析の場合には，

$$(\sqrt{g})^{\frac{1}{15}}/(\sqrt{\bar{g}})^{\frac{1}{9}} \gg 1 \tag{4.9}$$

なる関係にあり，潜在パラメータ空間 (θ^i) は (ϕ^λ) の部分空間として入っているのであるが，その入り方が十分に縮んで入っていることになる。したがって，ごく微少な n^λ をとっても，s^i が相対的に大きな2乗誤差を持つことになる。2乗誤差の比 AEP/AED の各点の近傍での変化は，$g_{ij}=B^\lambda_i B^\mu_j g_{\lambda\mu}$ であることから，B^λ_i の共変微分係数であるオイラー・スカウテンの曲率テンソル $H^\lambda_{ij}=\nabla_j B^\lambda_i$ によって，法曲率あるいはその平均曲率の大きさで評価することができる。

この関係は，潜在クラス分析の解法とは無関係であるから，一般的に言えることであり，勿論項目数3，クラス数2の場合，すなわち ϕ^λ と θ^i の次元が一致する場合にも生ずるものである。

5. 計 算 例

質問項目数を4，クラスの数を2とする。(ϕ^λ) からなる空間の次元は15次元であるから，真値 $\phi^\lambda_0=\phi^\lambda_0(\theta^i_0)$ に15個の独立な誤差 (n^λ) をつぎの意味で長さ一定として加えたものをデータ ϕ^λ と考える。各 (n^λ) の大きさは，

$$g_{\lambda\mu}n^\lambda n^\mu = \sum_{\lambda=1}^{16}\frac{(\phi^\lambda_0-\phi^\lambda)^2}{\phi^\lambda_0} = C \tag{5.1}$$

なる C をサンプルが1000個のときの χ^2-統計量の95%点をとる値とする。この場合には，自由度15であるから，$C=0.007261$ である。一方，各 n^λ は $\sum_{\lambda=1}^{16}\phi^\lambda=\sum_{\lambda=1}^{16}(\phi^\lambda_0+n^\lambda)=1$ より $\sum_{\lambda=1}^{16}n^\lambda=0$ なる条件を満たさなければならない。一般に m 次元ベクトルでこのような性質を持つものは $(1/\sqrt{m}, \cdots, 1/\sqrt{m})$ に垂直なベクトルとなり，

$$
\begin{matrix}
\left(\sqrt{\frac{m-1}{m}},\ -\frac{1}{\sqrt{m(m-1)}},\ \cdots\cdots,\ -\frac{1}{\sqrt{m(m-1)}}\right) \\
\left(-\frac{1}{\sqrt{m(m-1)}},\ \sqrt{\frac{m-1}{m}},\ \cdots\cdots,\ -\frac{1}{\sqrt{m(m-1)}}\right) \\
\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
\left(-\frac{1}{\sqrt{m(m-1)}},\ -\frac{1}{\sqrt{m(m-1)}},\ \cdots\cdots,\ \sqrt{\frac{m-1}{m}}\right)
\end{matrix}
\qquad (5.2)
$$

とし求められるが，ここではこれらのベクトルを，それぞれ45°回転したものを用いている。

特徴的ないくつかの点での数値をまとめたものが次の表である。

表—1

$C=0.007261$
$H=233.58$
$K=10.38$

	.2 .8	.4 .7	.1 .3	.5 .6	.3 .2	AEP	AED
1	.07 .93	.23 .67	.06 .26	.46 .59	.49 .18	.097	.0042
2	−.57 1.57	−.23 .57	−.41 .22	.26 .56	.52 .25	.39	.0034
3	−.59 1.59	−.11 .56	−.39 .22	.22 .56	.45 .27	.38	.0039
4	−.34 1.34	−.10 .63	−.26 .25	.47 .55	.32 .27	.28	.0032
5	−.18 1.18	−.08 .67	.12 .21	.36 .58	.35 .25	.21	.0036
6	.86 .14	.81 .85	.46 .39	.61 .66	.22 .13	.30	.0046
7	−.02 1.02	.12 .69	.06 .25	.48 .57	.35 .23	.12	.0032
8	−0.5 1.05	.15 .67	−.02 .26	.52 .54	.45 .19	.14	.0049
9	.48 .52	.83 .70	.08 .36	.53 .63	.29 .18	.17	.0038
10	−.13 1.13	.09 .65	−.08 .25	.41 .60	.40 .21	.17	.0048
11	.21 .79	.32 .72	.20 .27	.50 .60	.28 .19	.046	.0037
12	−.66 1.66	−.17 .52	−.48 .23	.38 .51	.53 .24	.41	.0046
13	.29 .71	.42 .77	.29 .28	.57 .60	.26 .21	.095	.0029
14	−.80 1.80	−.07 .44	−.59 .22	.25 .56	.58 .25	.46	.0035
15	.81 .19	.74 .84	.46 .36	.65 .64	.18 .16	.28	.0026

表—2

$C=0.002167$
$K=7.28$

	.3 .7	.2 .1	.3 .8	.5 .3	AEP	AED
1	.47 .53	.13 .09	.26 .94	.41 .30	.095	.0048
2	.22 .78	.16 .11	.08 .83	.49 .33	.090	.0042
3	.68 .32	.17 .06	.87 .83	.48 .21	.26	.0034
4	.04 .96	.19 .15	−.08 .78	.53 .35	.17	.0042
5	−.12 1.12	.31 .12	−.03 .63	.68 .35	.23	.0032
6	.76 .24	.12 .07	.51 1.03	.26 .27	.32	.0034
7	.37 .63	.16 .10	.35 .83	.45 .32	.045	.0066

6. 結　　論

データの誤差とそれによる潜在パラメータの誤差を平均2乗誤差で見ると，非常に大きな差が生じることは，潜在クラス分析の解法よりもむしろモデルそのものによるものと考えられる。すなわち，潜在クラス分析のモデル式の構造により，潜在パラメータの空間が縮んだ形で，データすなわち反応パターンの確率からなるパラメータ空間に入っているためである。

潜在クラス分析を実際に応用する場合には，ここで述べたことを考慮の上，データの大きさ，質問項目の数，クラスの数等を適当に選んで行なう必要がある。特に，項目数の減少に関しては丘本・磯貝[3]によるブロックモデルが提案されている。

参 考 文 献

1) 丘本　正，磯貝恭史：SDA 研究会報告集 (昭 53), pp. 15–19.
2) Yoshizawa, T.: Memorandom TYH-2 at Dep. Statist. Harvard Univ. (1971), pp. 158–187.
3) 丘本　正，磯貝恭史：SDA 研究会報告集 (昭 53), pp. 20–23.
4) McHugh, R. B.: Psychometrika **21** (1956), pp. 331–347.
5) Rao, C. R.: Bull. Cal. Math. Soc. **37** (1945), pp. 81–91.