数理解析研究所講究録
第197巻 1974年 1-11

On a non-parametric discriminant procedure

阪大 基礎工 小口光代

2変数の2標本 location 問題に対して、S.K. Chatterjee と P.K. Sen (1964) が提供した ある rank-sum statistics を判別問題に応用して、判別関数を作ることを試み、その場合の判別方式が consistency が成立していることを示した。又、標本の大きさについての考察を加えた。

§1. 序

次のような2変量判別問題を考える。π_i $(i=1,\cdots,k)$ を、k個の異った母集団とする。(X_i, Y_i) は、それぞれ、分布関数 $F_i(x,y)$ をもつ確率変数のk個の組とする。$F_i(x,y)$ は π_i における分布関数を表わし、連続関数であるとする。各 i $(i=1,\cdots,k)$ について、この $F_i(x,y)$ については、何等の知識もなくて、母集団 π_i から取りあつめたことが事前に判っている大きさ n_i の標本値を持つだけとする。その標本値を $(x_{i1}, y_{i1}), \cdots, (x_{in_i}, y_{in_i})$ とする。そして、他の母集団

π_0において、連続な分布関数 $F_0(x,y)$ をもって分布している確率変数 (X_0, Y_0) がある。この母集団 π_0 は、$\pi_1, \cdots, \pi_k$ の k 個の母集団のうちのいずれか1つと一致していることが解っている。π_0 から、又 n_0 の標本組を得るとして、その標本組を、$(x_{01}, y_{01}), \cdots, (x_{0n_0}, y_{0n_0})$ とする。これらの得られた観測値の組を比較して、π_0 は、どの母集団と一致しているかを判定したい。

このような、non-parametric な判別問題は、最初、Fix と Hodges (1951, 1952) によって試みられた。Stoller (1954) は、1変数の場合の2母集団への判別を行っている。Das Gupta (1964) は、p 変量の場合の k 個の母集団への判別を、経験分布関数の間の、Kolmogorov の距離を用いて行ない、その判別方法の consistency を示している。多変数の場合、tolerance region を用いることを試みている Quessenberry と Gessaman (1968) の論文、又、1変数2母集団への判別で、2種の判別誤差の最大のものを、与えた値でおさえるような sequential な方法を導入した Woinsky と Kurz (1969) の論文は興味深いものである。その他に、Ryzin (1966), Hudimoto (1963, 1964, 1968), Pelto (1969) 等、いくつかの論文がある。

§2. 判別方式.

母集団 π_i $(i=1,\cdots,k)$ における分布関数 $F_i(x,y)$ は、次の条件をみたすものとする。

(1) $F_1(x,y), \cdots, F_k(x,y)$ は、連続関数である。

(2). ある単調関数 $g(x)$ が存在して、$Y_1=g(X_1), \cdots, Y_k=g(X_k)$ となる確率は、それぞれ0である。

(3) $i\neq j=1,\cdots,k$ に対して、

$$P(X_i>X_j)+P(X_i\leq X_j)$$

$$P(Y_i>Y_j)+P(Y_i\leq Y_j)$$

観測値の大きさ $n_0, n_1, \cdots n_k$ は、$N_i=n_0+n_i$ $(i=1,\cdots k)$ とすると、$n_0/N_i \longrightarrow \lambda_i$ $(0<\lambda_i<1)$ $(i=1,\cdots,k)$ なる条件をみたすように大きくなるとする。

各 i $(=1,\cdots,k)$ に対して、観測値の組、

π_0 : $(x_{01}, y_{01}), \cdots, (x_{0n_0}, y_{0n_0})$

π_i : $(x_{i1}, y_{i1}), \cdots, (x_{in_i}, y_{in_i})$

を比較する。$(x_{01}, \cdots, x_{0n_0}, x_{i1}, \cdots, x_{in_i})$ を大きいものから順に並べて、その順位を $(r_{1i}, \cdots, r_{n_0i}, r_{n_0+1}\cdots r_{N_ii})$ とする。同じく $(y_{01}, \cdots, y_{0n_0}, y_{i1}\cdots y_{in_i})$ を大きいものから順に並べて、その順位を $(s_{1i}, \cdots, s_{n_0i}, s_{n_0+1i}\cdots, s_{N_ii})$ とする。 $S_{1i}=\sum_{\alpha=1}^{n_0} r_{\alpha i}$, $S_{2i}=\sum_{\alpha=1}^{n_0} s_{\alpha i}$

$$C_i = \frac{12}{N_i(N_i^2-1)} \sum_{\alpha=1}^{N_i} \left(r_{\alpha i} - \frac{N_i+1}{2}\right)\left(s_{\alpha i} - \frac{N_i+1}{2}\right)$$

とおき、

$$R_i = \frac{12}{(1-C_i^2)\, n_0 n_i (N_i+1)} \left[\left(S_{1i} - \frac{n_0(N_i+1)}{2}\right)^2 + \left(S_{2i} - \frac{n_0(N_i+1)}{2}\right)^2 \right.$$

$$\left. - 2C_i \left(S_{1i} - \frac{n_0(N_i+1)}{2}\right)\left(S_{2i} - \frac{n_0(N_i+1)}{2}\right) \right]$$

とおく。この統計量に基づいて、

$R_i = \min_{1 \le j \le k} R_j$ ならば、 $\pi_0 = \pi_i$

と判別する判別方式を考える。

§3. 判別方式の consistency

$$\Delta(x) = \begin{cases} 1 & : \ x > 0 \\ 0 & : \ x = 0 \\ -1 & : \ x < 0 \end{cases} \quad \text{とする。}$$

$$M_{j1}(1,1) = \sum_{\alpha}^{n_0} \sum_{\beta}^{n_j} \Delta(x_{0\alpha} - x_{j\beta})^2 \Big/ n_0 n_j$$

$$M_{j2}(1,1) = \sum_{\alpha \neq \alpha'}^{n_0} \sum_{\beta}^{n_j} \Delta(x_{0\alpha} - x_{j\beta})\, \Delta(x_{0\alpha'} - x_{j\beta}) \Big/ n_0(n_0-1) n_j$$

$$M_{j3}(1,1) = \sum_{\alpha}^{n_0} \sum_{\beta \neq \beta'}^{n_j} \Delta(x_{0\alpha} - x_{j\beta})\, \Delta(x_{0\alpha} - x_{j\beta'}) \Big/ n_0 n_j (n_j-1)$$

$$M_{j4}(1,1) = \sum_{\alpha\neq\alpha'}^{n_0}\sum_{\beta\neq\beta'}^{n_j} \Delta(x_{0\alpha}-x_{j\beta})\Delta(x_{0\alpha'}-x_{j\beta'}) \Big/ n_0 n_j (n_0-1)(n_j-1)$$

$$M_{j1}(2,2) = \sum_{\alpha}^{n_0}\sum_{\beta}^{n_j} \Delta(y_{0\alpha}-y_{j\beta})^2 \Big/ n_0 n_j$$

$$M_{j2}(2,2) = \sum_{\alpha\neq\alpha'}^{n_0}\sum_{\beta}^{n_j} \Delta(y_{0\alpha}-y_{j\beta})\Delta(y_{0\alpha'}-y_{j\beta}) \Big/ n_0(n_0-1)n_j$$

$$M_{j3}(2,2) = \sum_{\alpha}^{n_0}\sum_{\beta\neq\beta'}^{n_j} \Delta(y_{0\alpha}-y_{j\beta})\Delta(y_{0\alpha}-y_{j\beta'}) \Big/ n_0 n_j (n_j-1)$$

$$M_{j4}(2,2) = \sum_{\alpha\neq\alpha'}^{n_0}\sum_{\beta\neq\beta'}^{n_j} \Delta(y_{0\alpha}-y_{j\beta})\Delta(y_{0\alpha'}-y_{j\beta'}) \Big/ n_0 n_j (n_0-1)(n_j-1)$$

$$M_{j1}(1,2) = \sum_{\alpha}^{n_0}\sum_{\beta}^{n_j} \Delta(x_{0\alpha}-x_{j\beta})\Delta(y_{0\alpha}-y_{j\beta}) \Big/ n_0 n_j$$

$$M_{j2}(1,2) = \sum_{\alpha\neq\alpha'}^{n_0}\sum_{\beta}^{n_j} \Delta(x_{0\alpha}-x_{j\beta})\Delta(y_{0\alpha'}-y_{j\beta}) \Big/ n_0(n_0-1)n_j$$

$$M_{j3}(1,2) = \sum_{\alpha}^{n_0}\sum_{\beta\neq\beta'}^{n_j} \Delta(x_{0\alpha}-x_{j\beta})\Delta(y_{0\alpha}-y_{j\beta'}) \Big/ n_0 n_j (n_j-1)$$

$$M_{j4}(1,2) = \sum_{\alpha\neq\alpha'}^{n_0}\sum_{\beta\neq\beta'}^{n_j} \Delta(x_{0\alpha}-x_{j\beta})\Delta(y_{0\alpha'}-y_{j\beta'}) \Big/ n_0 n_j (n_0-1)(n_j-1)$$

とおくと R_j は.

$$R_j = \frac{3}{(1-c_j^2)\,n_0 n_j (N_j+1)} \Big[n_0 n_j \left(M_{j1}(1,1) + M_{j1}(2,2) - 2c_j M_{j1}(1,2) \right)$$

$$+n_0(n_0-1)n_j\left(M_{j2}(1,1)+M_{j2}(2,2)-2C_jM_{j2}(1,2)\right)$$
$$+n_0n_j(n_j-1)\left(M_{j3}(1,1)+M_{j3}(2,2)-2C_jM_{j3}(1,2)\right)$$
$$+n_0n_j(n_0-1)(n_j-1)\left(M_{j4}(1,1)+M_{j4}(2,2)-2C_jM_{j4}(1,2)\right)]$$

と、U-統計量を使って書き直せる。U-統計量の性質より

$A_{j1}=M_{j1}(1,1)+M_{j1}(2,2)-2C_jM_{j1}(1,2)$, $A_{j2}=M_{j2}(1,1)+M_{j2}(2,2)-2C_jM_{j2}(1,2)$, $A_{j3}=M_{j3}(1,1)+M_{j3}(2,2)-2C_jM_{j3}(1,2)$, $A_{j4}=M_{j4}(1,1)+M_{j4}(2,2)-2C_jM_{j4}(1,2)$ は、それぞれ、

$$\overline{A}_{j1}=2\left(1-\overline{P}_j\int_{E^2}(2F_j(x,\infty)-1)(2F_j(\infty,y)-1)\,dF_0(x,y)\right),$$

$$\overline{A}_{j2}=\int_{E^1}(2F_0(x,\infty)-1)^2\,dF_j(x,\infty)-2\overline{P}_j\int_{E^2}(2F_0(x,\infty)-1)(2F_0(\infty,y)-1)$$
$$dF_j(x,y)+\int_{E^1}(2F_0(\infty,y)-1)^2\,dF_j(\infty,y),$$

$$\overline{A}_{j3}=\int_{E^1}(2F_j(x,\infty)-1)^2\,dF_0(x,\infty)+\int_{E^1}(2F_j(\infty,y)-1)^2\,dF_0(\infty,y)$$
$$-2\overline{P}_j\int_{E^2}(2F_j(x,\infty)-1)(2F_j(\infty,y)-1)\,dF_0(x,y),$$

$$\overline{A}_{j4}=\left[\int_{E^1}(2F_j(x,\infty)-1)\,dF_0(x,\infty)-\overline{P}_j\int_{E^1}(2F_j(\infty,y)-1)\,dF_0(\infty,y)\right]^2$$
$$+\left[\int_{E^1}(2F_j(\infty,y)-1)\,dF_0(\infty,y)\right]^2(1-\overline{P}_j^2)$$

に、$\pi_0=\pi_1$ の下で、確率収束することがいえる。ここで、C_jについては、Chatterjee と Sen の論文[9]で、

$$\overline{F}_j = 3 \int_{E^2} (2\overline{F}_j(x,\infty)-1)(2\overline{F}_j(\infty,y)-1)\, d\overline{F}_j(x,y)$$

ただし、$\overline{F}_j(x,y) = \chi_j F_0(x,y) + (1-\chi_j) F_j(x,y)$ である。

とすると、C_j は $\overline{r}_j$ に確率収束することが証明されている。

ここで、$\overline{r}_j = \pm 1$ となるのは、単調関数 $g(x)$ が存在して、$Y_j = g(X_j)$, $Y_0 = g(X_0)$ と書ける場合、かつ、その場合に限る。

以上により、条件 (2), (3) の下では、

$$R_j = \frac{3A_{j1}}{(1-C_j^2)(N_j+1)} + \frac{3(n_0-1)A_{j2}}{(1-C_j^2)(N_j+1)} + \frac{3(n_j-1)A_{j3}}{(1-C_j^2)(N_j+1)} + \frac{3(n_0-1)(n_j-1)}{(1-C_j^2)(N_j+1)} A_{j4}$$

の第1項が0，第2項が $\frac{3\overline{A}_{j2}\chi_j}{(1-\overline{r}_j^2)}$ 第3項が $\frac{3\overline{A}_{j3}(1-\chi_j)}{(1-\overline{r}_j^2)}$

第4項は $\frac{3\overline{A}_{j4}}{(1-\overline{r}_j^2)}$ に確率収束するが、$\frac{(n_0-1)(n_j-1)}{N_j+1}$ が…∞に

なり、任意の定数 c に対して、$\pi_0 = \pi_i$ の下では、

$$\mathrm{Prob}(R_j > c) \longrightarrow 1 \qquad (j \neq i)$$

となる。

一方、R_i は、$F_i(x,y)$ の下では、その極限分布として、自由度2の χ^2-分布を持つことが [9] で証明されている。

よって $\pi_0 = \pi_i$ の下では、

$$\mathrm{Prob.}\left(R_i = \min_{1 \le j \le k} R_j\right) \longrightarrow 1$$

となり、この判別方式は、consistent である。

§4. 標本の大きさについて.

先に述べた判別方式は、1変数の場合には、

$$R_i = \frac{12}{n_0 n_i (N_i+1)} \left(S_i - \frac{n_0(N_i+1)}{2} \right)^2 \qquad i=1,\cdots,k$$

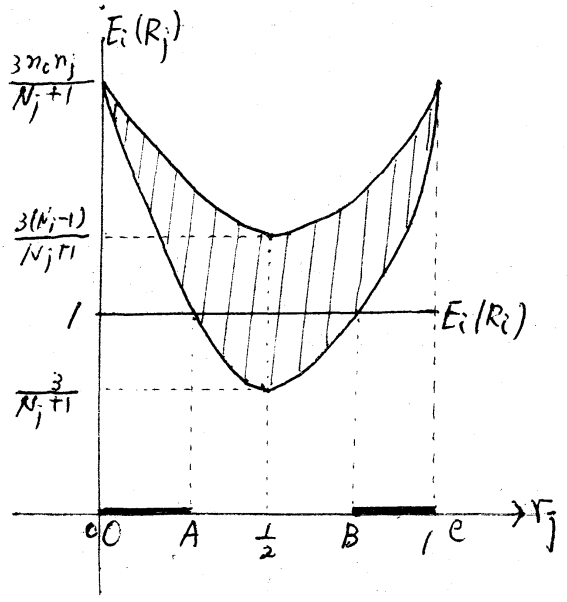
となる。S_iは、$(x_{01}\cdots x_{0n_0}, x_{i1}\cdots x_{in_i})$の順位$(r_{01}\cdots r_{0n_0}, r_{i1}, \cdots, r_{iN_i})$の最初の$n_0$個の和である。

$i \neq j$のとき、R_jを分布関数$F_i(x)$の下で平均をとると、

$$\frac{3}{N_j+1}\left[4(n_0 n_j - 1)(r_j^2 - r_j) + n_0 n_j\right] \leqq E_i(R_j) \leqq$$

$$\frac{3}{N_j+1}\left[4(n_0-1)(n_j-1)(r_j^2-r_j) + n_0 n_j\right]$$

となる。ここに $R_j = \int_{E^1} F_j(x)\, dF_i(x)$. よって、$R_j$の分布関数$F_i(x)$の下での平均値は、下図の斜線の部分にある。



A, Bの座標は、

$\frac{1}{2} - \sqrt{\frac{N_j-2}{12(n_0n_j-1)}}$, $\frac{1}{2} + \sqrt{\frac{N_j-2}{12(n_0n_j-1)}}$

である。

$E_i(R_i) = 1$ であるから、平均的にみれば、図のOABCの部分にr_jが落ちるように、標本の大きさ(n_0, n_j)を決めればよい。

思われる。これをすべての $j=1,\cdots,k$ に行なって、標本の大きさ $(n_0, n_1, \cdots, n_k)$ を決めるのが適当であろう。non-parametric な場合、γ_j は未知であるが、標本の大きさが大きくなるにつれ、OA, BC の長さが大きくなるので、十分大きな標本の大きさをとれば、未知の γ_j が、この区間に落ちる確率も大きくなる。これは、又、5.3で証明された consistency からも当然である。

参考文献

[1]. E. Fix and J. L. Hodges (1951)

Discriminatory Analysis, Non-Parametric Discrimination: Consistency Properties. USAF. school of aviation Medicine, Report No.4

[2]. E. Fix and J. L. Hodges (1952)

Discriminatory Analysis, Non-Parametric Discrimination: Small Sample Performance. USAF. school of aviation Medicine, Report No.11

[3]. S. Das-Gupta (1964)

Non-Parametric Classification Rules. Sankhyā A 24. p.p. 25-30

[4]. H. Hudimoto (1963)

分類について - I. 二群への分類 AISM. p.p. 31-38

[5]. H. Hudimoto (1964)

On a Distribution-free Two-Way Classification. AISM. 16 p.p. 247-253

[6]. H. Hudimoto (1968)

On the Empirical Bayes Procedure (1). AISM. 20 p.p. 169-185

[7]. C. R. Pelto (1969)

Adaptive Non-Parametric Classification. Technometrics 11.

[8]. C. P. Queenberry and M. P. Gassaman (1968) Non-Parametric Discrimination Using Tolerance regions. AMS. 39 p.p. 664-673

[9]. J. Van Ryzin (1966)

Bayes Risk Consistency of Classification Procedures Using Density

Estimation. Sankhyā A 26. p.p. 261-270

[10]. S. K. Chatterjee and P. K. Sen (1964)

Non-Parametric Tests for the Bivariate Two-Sample Location Problems.

CSAB 14 p.p. 18-58

[11]. D. S. Stoller (1954)

Univariate Two-Population Distribution-Free Descrimination.

TASA 49 p.p. 770-777

[12]. M. N. Woinsky and L. Kurz (1969)

Sequential Non-Parametric Two-Way Classification with a Prescribed

Maximum Asymptotic Error Probability. AMS. 40 445-455.

//