数理解析研究所講究録
第46巻 1968年 62-87

経験的ベーズ手法
(複合型決定問題を含む)

統計数理研　鈴木義一郎

§1. 序

この報告の目的は、表題に掲げた主題に関するH. Robbinsの一連の論文[1],[2],[3],[4](主に[1],[3])の簡単なスケッチと、これらを総括的に議論した筆者[5]の要旨の紹介である。

論文[1]の重要な意義は、通例妥当と考えられてきたミニマックスな法則に対して有力な競争相手の決定法則を構成し、従来数学的、理論的面に於てのみ重視されてきた"ベーズ的考え"を統計的、実際的面に取り入れ、いわゆる"経験的ベーズ手法(empirical Bayes approach)"と呼ばれる一連の理論の展開のきっかけを与えたことに在ろう。経験的ベーズという言葉は[1],[2]の論文に於ては用いられていないが、彼等の思考の根底にそのような概念が潜んでいたことは、Robbinsが続いて発表した論文[3],[4]から十分推測できる。尚これら一連の論文の扱っている主題が同じものであることは、§4

の一般的定式化（詳しくは拙著[5]）を眺められれば明らかとなろう。

従来のベーズ推論は、純理論統計家（統計学の中に問題を求めそれを数学的に処理している研究家）の間で諸概念（許容性とか完全性）の説明の為に便宜的に用いられていたが、実践家や実践的統計家からは（先験分布を勝手に与えて了うことから）実用上問題にならぬとそっぽを向かれてきた。純理論家の多くは実際問題を直視せず、Robbinsの提唱したこの経験的ベーズの立場を受け容れようとしない。理論家と実践家との間のこのような溝を埋めるのに“経験的ベーズ”は恰好な主題であると考えられるので、ぜひ論文[1],[3]の一読をお勧めする。これらに関連した文献については[5]の参考文献を参照されたい。

参考文献

[1] H. Robbins, Proc. 2nd Berkeley Symp. (1951), 131－148.

[2] ――――, Ann. Math. Statist. 26 (1955), 37－51.

[3] ――――, Proc. 3rd Berkeley Symp. (1955), 157－163.

[4] ――――, Ann. Math. Statist. 33 (1964), 1－20.

[5] G. Suzuki, Research Memorandum No.19 (1968) (Inst. Statist. Math.)

§2. Robbins の論文[1],[2]のスケッチ

H. Robbins は論文[1]に於て次のような問題を提示した。観測可能な確率変数 X は、分散が1，平均が θ の正規分布に従うものとする。ここでパラメータ θ の値は未知であるが、-1 か $+1$ のいづれかの値しかとり得ないものとする。今、n 個の独立な（無関係な）観測 $(X_1, X_2, \cdots, X_n)$ が与えられているとする。ここで X_i は正規分布 $N(\theta_i, 1)$ に従うことが判っているが、θ_i の値は -1 か $+1$ かは判らない。しかし、その値がいづれであるかの決定（decision）を強いられている。目的は間違った決定（θ の値が $+1$ のときに -1 と判定する場合と丁度反対の場合の2種類ある）を下す割合の期待値をできるだけ小さくするような決定法則を見出すことである。（2種類の失敗を同じ比重で扱うという仮定や分布が正規族であるといった仮定は除き得ることを後の論文[2]で示している。）

このような複合型決定問題、即ち観測値ベクトル $\underline{X} = (X_1, X_2, \cdots, X_n)$ からパラメータベクトル $\underline{\theta} = (\theta_1, \theta_2, \cdots, \theta_n)$ を推定する問題に於ける決定法則（decision rule）D とは，$u_i(\underline{X})$ の確率で $\theta_i = +1$（$1 - u_i(\underline{X})$ の確率で $\theta_i = -1$）と定めるような n 組の函数 $\underline{u} = (u_1, u_2, \cdots, u_n)$ を対応させることである。さて、真のパラメータベクトルが $\underline{\theta} = (\theta_1, \theta_2, \cdots, \theta_n)$ のとき、

このような決定Dを用いてこおむるリスク（間違った分類の割合の期待値）は

(1) $$r(D,\underline{\theta}) = E\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\frac{1+\theta_i}{2} - sgn(\theta_i)\,u_i(\underline{X})\right]\middle|\,\underline{\theta}\right\}$$

$$= p(\underline{\theta}) - \frac{1}{n}\sum_{i=1}^{n} sgn(\theta_i)\int u_i(\underline{x})\,\phi(\underline{x},\underline{\theta})\,d\underline{x}$$

と表わせる。但し、

$$p(\underline{\theta}) = \frac{1}{2n}\sum_{i=1}^{n}(1+\theta_i),$$

$$\phi(\underline{x},\underline{\theta}) = \frac{1}{(2\pi)^{n/2}}\,e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta_i)^2}.$$

従って真のパラメータベクトル $\underline{\theta}=(\theta_1,\theta_2,\cdots,\theta_n)$ に対する最尤推定量は

$$\theta_i = sgn(X_i),\quad i=1,2,\cdots,n$$

で与えられる。この決定法則を $\tilde{D}$ と表わそう。対応するu-函数は

$$u_i(\underline{x}) = u(x_i) = \begin{cases} 1, & x_i>0, \\ 0, & x_i<0, \end{cases}\quad i=1,2,\cdots,n$$

であるから、容易に

$$r(\tilde{D},\underline{\theta}) = \Phi(-1) = 0.1587$$

となることが示され（Φは標準正規分布の分布函数）、而もこの法則が唯一のミニマックス解となることも判る。

ある決定法則Dがsimpleであるというのは、対応するu-函数が

$$u_i(\underline{x}) = u(x_i),\quad i=1,2,\cdots,n$$

の形をしている場合である。simpleな法則Dに対するリスク(1)は

(2) $$r(D,\underline{\theta}) = p(\underline{\theta}) - \int \{p(\underline{\theta})\varphi(x-1) - [1-p(\underline{\theta})]\varphi(x+1)\} u(x)\,dx$$

の如く $p(\underline{\theta})$ の一次函数として書き表わせる。(φ は標準正規分布の密度函数) 任意の λ ($0 \leqq \lambda \leqq 1$) に対して

$$u(x) = u_\lambda(x) = \begin{cases} 1 & \lambda\varphi(x-1) - (1-\lambda)\varphi(x+1) > 0 \\ 0 & \lambda\varphi(x-1) - (1-\lambda)\varphi(x+1) \leqq 0 \end{cases}$$

と置く。この u_λ に対応する決定法則 D_λ は

$$\theta_i = \operatorname{sgn}\left(x_i - \frac{1}{2}\log\frac{1-\lambda}{\lambda}\right),\quad i=1,2,\cdots,n$$

で、リスクは

$$r(D_\lambda, \underline{\theta}) = p(\underline{\theta})\,\alpha(\lambda) + [1-p(\underline{\theta})]\,\beta(\lambda)$$

で与えられる。ここで

$$\alpha(\lambda) = \Phi\left(-1+\frac{1}{2}\log\frac{1-\lambda}{\lambda}\right),\quad \beta(\lambda) = \Phi\left(-1-\frac{1}{2}\log\frac{1-\lambda}{\lambda}\right).$$

特に $D_{\frac{1}{2}} \equiv \tilde{D}$ であるから、$p(\underline{\theta}) = \frac{1}{2}$ でない限り確に $\tilde{D}$ より良いsimpleな決定法則 $D_{p(\underline{\theta})}$ がある筈で、そのリスクは

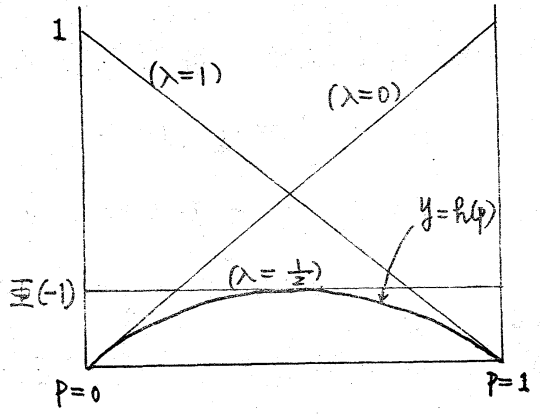
(3) $$h(p(\underline{\theta})) = p(\underline{\theta})\,\alpha(p(\underline{\theta})) + [1-p(\underline{\theta})]\,\beta(p(\underline{\theta}))$$

である。実際曲線 $h(p)$ は、直線群

$$\{\, y = \alpha(\lambda)p + \beta(\lambda)(1-p);\ 0 \leqq \lambda \leqq 1 \,\}$$

の包絡曲線で、直線 $y = \alpha(\frac{1}{2})p + \beta(\frac{1}{2})(1-p) \equiv \Phi(-1)$ の下方に在ることが判る。

しかし $p(\theta)$ の値が既知というのは稀であって、多くの場合このような決定法則を得ることはできない。唯 $p(\theta)$ を推定することが可能なときには $\lambda = \hat{p}$（$p(\theta)$ の推定量）と置いた決定法則 R_λ が利用でき、而も $\hat{p}$ が比較的良い推定量であれば、この決定法則の方が $\tilde{D}$ より（少くともサンプル数が大きいとき）良いと期待できるかも知れない。但し推定量 $\hat{p}$ が確率変数、従って λ が x の函数になっているから、このようにして得られる決定法則はもはや simple ではなくなる。



さて、θ の最尤推定量を用いて $p(\theta)$ に対する最尤推定量

$$\hat{p} = \hat{p}(x) = \frac{1}{n}\{\text{no. of } i \text{ for which } x_i > 0\}$$

を得る。これは

$$E\{\hat{p}\,|\,\theta\} = [1 - 2\text{亜}(-1)]\,p(\theta) + \text{亜}(-1)$$

であるから $p(\theta) = \frac{1}{2}$ のとき以外不偏でない。そこで

$$z = [\hat{p} - \text{亜}(-1)] \,/\, [1 - 2\text{亜}(-1)]$$

と修正すれば不偏推定量になる。z の分散は

$$V\{z\,|\,\theta\} = \frac{\text{亜}(-1)[1 - \text{亜}(-1)]}{n[1 - 2\text{亜}(-1)]^2} \fallingdotseq \frac{0.29}{n}.$$

更に小さな分散の不偏推定量は

$$V = \frac{1}{2}(1+\bar{X}),\quad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

で与えられる。実際初等的計算によって

$$E\{V|\underline{\theta}\} = p(\underline{\theta}),\quad V\{V|\underline{\theta}\} = \frac{1}{4n} = \frac{0.25}{n}$$

となることも示される。唯 $0 \leqq p(\underline{\theta}) \leqq 1$ であるから、これを

$$\tilde{V} = \begin{cases} 0 & V \leqq 0 \\ V & 0 < V < 1 \\ 1 & V \geqq 1 \end{cases}$$

のように修正した方が妥当であろう。但しこの $\tilde{V}$ には多少の偏りがある。さてこの $\tilde{V}$ を $p(\underline{\theta})$ に対する推定量としてとったとき、対応する決定法則 D^* は

(4) $$\theta_i = \mathrm{sgn}(X_i - x^*),\quad i = 1, 2, \cdots, n$$

で与えられる。ここで

$$x^* = \begin{cases} +\infty, & \bar{X} \leqq -1, \\ \frac{1}{2}\log\frac{1-\bar{X}}{1+\bar{X}}, & -1 < \bar{X} < 1, \\ -\infty, & \bar{X} \geqq 1. \end{cases}$$

この D^* は勿論 simple ではないが、リスクは

$$r(D^*, \underline{\theta}) = h(p(\underline{\theta}), n)$$

で与えられる。ここで

$$h(p, n) = p\,\Phi(-2p\sqrt{n}) + (1-p)\,\Phi(-2(1-p)\sqrt{n}) + \int_{-2(1-p)\sqrt{n}}^{2p\sqrt{n}} g(x; p, n)\,\varphi(x)\,dx,$$

$$g(x;p,n)=p\Phi\left[\sqrt{\frac{n}{n-1}}\left(-1-\frac{x}{\sqrt{n}}+\frac{1}{2}\log\frac{1-p+x/2\sqrt{n}}{p-x/2\sqrt{n}}\right)\right]$$
$$+(1-p)\Phi\left[\sqrt{\frac{n}{n-1}}\left(-1-\frac{x}{\sqrt{n}}-\frac{1}{2}\log\frac{1-p+x/2\sqrt{n}}{p-x/2\sqrt{n}}\right)\right].$$

明らかに

$$\lim_{n\to\infty} h(p,n)=h(p).$$

$h(p)$ は(3)で与えられた函数であるから、

$$\lim_{n\to\infty} r(D^*,\underline{\theta})=h(p(\underline{\theta}))\leqq\Phi(-1).$$

従って(4)で与えられる(non-simple な)法則 D^* がミニマックスな決定法則 $\tilde{D}$ より劣るとは何人も断言できないだろう。

Hannan-Robbins は論文[2]に於て、この議論をもっと一般的に記述した。まず θ のとり得る値は便宜上0又は1とし(パラメータ空間は勝手な抽象空間の部分集合で良く元の個数のみが本質的である)、2種類の間違った判断にウェイトをつけた。即ち $\underline{\theta}=(\theta_1,\theta_2,\cdots,\theta_n)$ に対して $\underline{d}=(d_1,d_2,\cdots,d_n)$ という決定をする(d_i, $1-d_i$ の確率で夫々 $\theta_i=1$, $\theta_i=0$, $i=1,2,\cdots,n$ と判断する)ときの損失を

$$W(\underline{d},\underline{\theta})=\frac{1}{n}\sum_{i=1}^{n}\left[a\theta_i(1-d_i)+b(1-\theta_i)d_i\right]$$

で与える。(a, b は適当な正数で、$a=b=1$ の場合が[1]で扱われている。)

前と同様な観測値ベクトル $\underline{X}=(X_1,X_2,\cdots,X_n)$ が与えられている。各 X_i は独立で分布函数として $F(\cdot,\theta_i)$ をもっている、

但し $\theta_i = 0$ 又は 1, $i=1,2,\cdots,n$ とする。

$$\mu_\theta(B) = \int_B dF(x,\theta), \quad \theta = 0, 1,$$

$$\mu(B) = \mu_0(B) + \mu_1(B), \quad B\text{は一次元ボレル集合}$$

によって有限な測度 μ を定義し、μ_θ の μ に関する一般化された密度を $f(\cdot,\theta)$ と表わす（$\theta = 0,1$）。真のパラメータベクトルが $\underline{\theta} = (\theta_1, \theta_2, \cdots, \theta_n)$ のとき決定函数 $\underline{u} = (u_1, u_2, \cdots, u_n)$ を用いて（$u_i(\underline{x})$, $1-u_i(\underline{x})$ の確率で夫々 $\theta_i = 1$, $\theta_i = 0$ と判断して）こおむるリスクは

$$r(\underline{u}, \underline{\theta}) = E\{W(\underline{u}(\underline{X}), \underline{\theta}) | \underline{\theta}\}$$

$$= \frac{1}{n}\int\left[\sum_{i=1}^{n}\{a\theta_i(1-u_i(\underline{x})) + b(1-\theta_i)u_i(\underline{x})\}\right] f(\underline{x},\underline{\theta})\, d\mu^n$$

である。ここで $f(\underline{x},\underline{\theta}) = \prod_{i=1}^{n} f(x_i,\theta_i)$ は真のパラメータが $\underline{\theta}$ のときの直積測度 μ^n に関する $\underline{X} = (X_1, X_2, \cdots, X_n)$ の一般化された直積密度である。特に simple な $\underline{u} = (u, u, \cdots, u)$ に対しては、

(5) $$r(u,\underline{\theta}) = p(\underline{\theta}) - \int [ap(\underline{\theta})f(x,1) - b(1-p(\underline{\theta}))f(x,0)]\, u(x)\, d\mu$$

で与えられる。但し

$$p(\underline{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\theta_i .$$

従って

$$u_\lambda(x) = \begin{cases} 1, & x \in C[\lambda] = \{x : a\lambda f(x,1) > b(1-\lambda)f(x,0)\}, \\ 0, & x \overline{\in} C[\lambda] \end{cases}$$

と置くとき、(5) は simple な $u_{p(\underline{\theta})}$ によって最小値

$$\phi(p(\underline{\theta})) = p(\underline{\theta}) - ap(\underline{\theta})\mu_1(C[p(\underline{\theta})]) + b(1-p(\underline{\theta}))\mu_0(C[p(\underline{\theta})])$$

が与えられる。

さて $p(\theta)$ の値が未知の場合このような simple な決定函数は定められないが、$p(\theta)$ に対するある一致推定量 $\hat{p}_n = \hat{p}_n(\underline{x})$ が与えられているとき（[2]では $p(\theta)$ に対する不偏推定量のある部分クラスの中で分散を最小にするものの求め方についても記されている）、

$$(6) \qquad u_i^*(\underline{x}) = \begin{cases} 1, & x_i \in C[\hat{p}_n(\underline{x})], \\ 0, & x_i \in \bar{C}[\hat{p}_n(\underline{x})], \end{cases} \quad i = 1, 2, \cdots, n$$

によって定義される non-simple な函数 $\underline{u}^* = (u_1^*, u_2^*, \cdots, u_n^*)$ に対して次のような結果が成立する。

“任意の $\varepsilon > 0$ に対し

$$P_r\{W(\underline{u}^*(\underline{x}), \underline{\theta}) \leq \Phi(p(\underline{\theta})) + \varepsilon, \text{ for all } n \geq n_0(\varepsilon)\} \geq 1 - \varepsilon$$

となるような整数 $n_0(\varepsilon)$ が得られる。従って、

$$r(\underline{u}^*, \underline{\theta}) \leq \Phi(p(\underline{\theta})) + \varepsilon, \text{ for all } n \geq n_1(\varepsilon)$$

となるような整数 $n_1(\varepsilon)$ も選ぶことができる。”

即ち (6) で与えられる決定函数は、サンプル数が充分大であれば、simple な決定函数のクラスの中に於る最良なものと同じ程度に良い決定を為し得ることが示された。彼等は尚、simple なものを含むより広い（各段階での決定を置換してもリスクを変えないようなものの）クラスの中での最良のものと比較しても決定函数 (6) が見劣りしないことを示している。

§3. Robbinsの論文[3],[4]のスケッチ

Xを離散的な値をとる確率変数（連続的な値をとる場合も平行した議論ができる）、Λをある（未知の）先験分布

$$G(\lambda) = Pr\{\Lambda \leq \lambda\}$$

に従う確率変数とする。$\Lambda = \lambda$を与えたときのXの条件付分布

$$p(x|\lambda) = Pr\{X = x \mid \Lambda = \lambda\}$$

は判っているものとする。従ってXの条件付でない分布は

$$P_G(x) = Pr\{X = x\} = \int p(x|\lambda)\, dG(\lambda)$$

で与えられる。Λに対する$\varphi(X)$の形の推定量の自乗平均誤差は

$$E\{\varphi(X) - \Lambda\}^2 = E\{E[\varphi(X) - \Lambda]^2 \mid \Lambda\}$$
$$= \sum_x \int p(x|\lambda)[\varphi(x) - \lambda]^2 dG(\lambda)$$

で与えられ、Xを与えたときのΛの条件付期待値

(1) $$\varphi_G(X) = E\{\Lambda \mid X\}$$

によって最小になる。ところで

$$\varphi_G(x) = \int p(x|\lambda)\lambda\, dG(\lambda) \Big/ \int p(x|\lambda)\, dG(\lambda)$$

であるから、分布Gが判ればベーズ推定量(1)を作ることができる。しかし多くの場合分布Gは未知であるから、このような推定量を作ることができない。それで従来は、不偏性とかミニマックス基準とかのGに関する情報を必要としない制

約条件のみを課して推定量を定めてきた。

しかし次のような場合には事情が一寸異る。確率変数の組の同一分布に従う独立系列 $\{(X_i, \Lambda_i), i=1,2,\cdots\}$ が与えられているとする。各 (X_i, Λ_i) の分布の構造は最初に述べた確率変数の組 (X, Λ) と同一で、Λ_i の方は観測可能でない確率変数とする。$X_1, X_2, \cdots, X_n$ に関する観測値を全部利用して、Λ_n に対する推定量を作りたい。Robbins [3] の提唱した問題というのは、次のようなものである。

"$X_1, X_2, \cdots, X_{n-1}$ に関する情報を函数形を定めるのに用いて、(1) で与えられる φ_G に確率収束するようなランダムな函数列 $\{\hat{\varphi}_n(\cdot) = \hat{\varphi}_n(X_1, X_2, \cdots, X_{n-1}; \cdot), n=2,3\cdots\}$ を構成できるか？"

一般的には否定的答えしか出ない。何故ならそのようなことが可能となるのは、原則として未知の先験分布 G に対する一致推定量 G_n が得られねばならず、而も Λ_i を直接観測できない限り X_i の観測値を通して為されなければならない。これは mixture より mixing を求める問題、言い換えると第一種フレッドホルム型積分方程式を解く問題に帰着され、無条件では解けないからである。(少くとも mixture と mixing との間の一対一対応 —— Teicher : Ann. Math. Stat. 34 (1963), 1265-9 —— の条件が必要である。唯一対一対応がないからといってさじを投げるのも早計である —— Robbins [4; 85]。)

Robbins [3, 4] は核 $p(x|\lambda)$ が指数型分布族の場合については肯定的答えのでることを例示した。まず条件付密度が

$$p(x|\lambda) = \lambda^x f(x) h(\lambda) \quad , x = 0, 1, 2, \cdots$$

の形をしている場合（ポアソン分布や負の二項分布）を考える。

$$\int p(x|\lambda)\, \lambda\, dG(\lambda) = \int \lambda^{x+1} f(x) h(\lambda)\, dG(\lambda)$$
$$= \frac{f(x)}{f(x+1)} p_G(x+1)$$

であるから、(1) の函数に対して

$$\varphi_G(x) = f(x)\, p_G(x+1) \,/\, f(x+1)\, p_G(x)$$

なる関係が得られる。そこで各 n に対し

$$p_n(x) = \frac{1}{n}\{\text{no. of } X_1, \cdots, X_n \text{ which are equal to } x\}, \quad x = 0, 1, \cdots$$

を定義すると、これは $p_G(\cdot)$ に対する一致推定量となる。

従って

$$\hat{\varphi}_n(x) = \hat{\varphi}_n(X_1, X_2, \cdots, X_{n-1}; x)$$
(2)
$$= f(x)\, p_{n-1}(x+1) \,/\, f(x+1)\, p_{n-1}(x)$$

がベーズ推定量 (1) に対する一つの一致推定量を与える；

(3) $$E[\hat{\varphi}_n(X_n) - \Lambda_n]^2 \longrightarrow E[\varphi_G(X) - \Lambda]^2.$$

特にポアソン核

$$p(x|\lambda) = e^{-\lambda} \frac{\lambda^x}{x!} \quad , x = 0, 1, \cdots \; ; \lambda > 0$$

で先験分布 G がガムマタイプ

$$G'(\lambda) = \frac{h^b}{\Gamma(b)} \lambda^{b-1} e^{-h\lambda} \quad , \lambda, b, h > 0$$

の場合を考えてみる。ベーズ推定量 (1) は

$$\varphi_G(X) = [X+b]/[1+k]$$

で、その自乗平均誤差は

$$E[\varphi_G(X) - \Lambda]^2 = b/k(1+k).$$

一方通常の（最小分散不偏）推定量はX自身で、誤差は

$$E[X-\Lambda]^2 = E\Lambda = b/k$$

と計算される。従ってこのような場合、(2)で構成される推定量 $\hat{\varphi}_n(X_n)$ が通例のポアソン推定量 X_n より良くなる場合が十分期待できる（少くとも漸近的には(3)の関係から前者の方が後者よりefficientになる）。

最後にラプラシアンタイプ

$$p(x|\lambda) = e^{\lambda x} f(x) k(\lambda)$$

の場合を考える。微分と積分の演算の交換が許されるものとすると

$$\begin{aligned}\varphi_G(x) P_G(x) &= \int p(x|\lambda)\lambda\, dG(\lambda) \\ &= f(x)\int \lambda e^{\lambda x} k(\lambda)\, dG(\lambda) \\ &= f(x)\frac{d}{dx}\{P_G(x)/f(x)\}\end{aligned}$$

であるから

$$\varphi_G(x) = \frac{d}{dx}\log\left\{\frac{P_G(x)}{f(x)}\right\}$$

なる関係を得る。この場合にも $P_n(x)$ を適当ななめらかさを持つように修正した $\hat{P}_n(x)$ で $P_G(x)$ を近似することにより、(3)の関係が成立するような(2)に対応する $\hat{\varphi}_n(\cdot)$ が得られる。

Robbins は更に[4]に於て行動空間（action space）を導入して、このようなベーズ推定やベーズ型仮説検定問題等を含むような一般論を展開した。しかし[1],[2]の複合型決定問題を含む程一般的ではない。彼は漸近的にベーズ決定法則に（平均的に）近づくような（例えば(3)の関係が成立するようなもので、かかる性質をasymptotically optimalと呼んでいる）ランダムな決定函数を得る為の条件を示しているが、それは殆んど定義を少し書き直したようなものに過ぎない。更に先験分布（mixing）の推定問題にも触れているが、一般的な場合にはやはり見るべき結果を与えていない。唯パラメータ空間が有限個の元から成る場合については、有限個の函数のクラス $\{f(\cdot,\theta_1),\cdots,f(\cdot,\theta_K)\}$ から互に直交する成分をとり出し、これを用いて先験分布を推定するという有効な手法を示した。論文[4]で定式化されたものは、次の§4で述べるより一般の問題に含まれるので、これ以上の直接的紹介は省略する。

§4. 一般的定式化

$\mathcal{X}$, $\mathcal{Y}$, A を夫々標本，パラメータ，行動空間とする。更に $\mathcal{X}$ 上の確率測度の集合

$$\{\nu(\cdot,y): y\in\mathcal{Y}\}$$

が与えられていて、ある σ- 有限な測度 μ によって dominate されている。μ に関する R-N. 微係数を

$$\frac{d\nu(\cdot, y)}{d\mu} = f(\cdot, y)$$

と表わす。$w(a, y)$ は与えられた $A \times \mathcal{Y}$ 上の有界可測函数とする（非負値でなくとも構わない）。$\mathcal{X}$ から A の中への写像 d を決定函数と呼び、このような d の集合を D で表わす。$D \times \mathcal{Y}$ 上にリスク函数

(1) $$r(d, y) = \int_{\mathcal{X}} w(d(x), y) f(x, y) \mu(dx)$$

が定義される。これを用いて $\mathcal{Y}$ 上の距離

$$\delta(y_1, y_2) = \sup_{d \in D} |r(d, y_1) - r(d, y_2)|$$

を定義する。この δ の意味での $\mathcal{Y}$ の開集合をすべて含むような最小の σ- 体を $\mathcal{Y}$ と表わす。全く同様にして $\mathcal{D}$ も定義される。空間 $\mathcal{Y}$ が可分ならば、函数 (1) は $\mathcal{D} \times \mathcal{Y}$-可測である。

$\mathcal{M}_0$ を $\mathcal{Y}$ 上の確率測度の全体とし、$D \times \mathcal{M}_0$ 上に

$$r_1(d, \xi) = \int_{\mathcal{Y}} r(d, y) \xi(dy)$$

なるリスク函数、更に $\mathcal{M}_0$ 上にも

$$r_1(\xi) = \inf_{d \in D} r_1(d, \xi)$$

なるベーズリスクを定義する。後で $\mathcal{M}_0$ の元を推定する問題を考えるが、直接得られる推定量は一般に確率 1 で $\mathcal{M}_0$ に属さない。$\mathcal{M}_0$ へ入るよう修正できるが、リスクを最小にするという観点からはそのような方法が最良とは言えない。そこでベーズの

概念をもっと拡張して置く。$\mathcal{M}$を$\mathcal{Y}$上の有限な有符号測度の全体とする。任意の$\nu \in \mathcal{M}$に対して、$\nu = \nu^+ - \nu^-$（ν^+, ν^-は有限測度）なる表現が一通りにできるから、$D \times \mathcal{M}$上のリスク函数を

$$r^*(d,\nu) = \int_Y r(d,y)\,|\nu|(dy) = |\nu|(Y)\, r_1(d,\tilde{\nu})$$

で与えることにする。但し

$$|\nu| = \nu^+ + \nu^-, \quad \tilde{\nu} = [|\nu|(Y)]^{-1}|\nu| \ (\in \mathcal{M}_0).$$

$|\nu| = |\nu'|$となるν（同一リスクを与えるν）の集り全体で$\mathcal{M}$の一個の元と見做すことにする。$\mathcal{M}$上のベーズリスク及び距離も次のように定義して置く。

$$r^*(\eta) = \inf_{d \in D} r^*(d,\eta),$$

$$\delta^*(\eta_1,\eta_2) = \sup_{d \in D} |r^*(d,\eta_1) - r^*(d,\eta_2)|.$$

次に$\mathcal{Y}$上のある確率測度mを一つ選んで固定する。任意の$p \geq 1$に対し、$L^p = L^p(Y,m)$を

$$\|g\|_p = \left\{\int_Y |g(y)|^p m(dy)\right\}^{1/p} < \infty$$

となるような$\mathcal{Y}$上の実数値函数gの全体として

$$\mathcal{M}^p = \mathcal{M}^p(Y,m) = \{\nu \in \mathcal{M} : \nu \ll m, \ \frac{d\nu}{dm} \in L^p\}$$

なる$\mathcal{M}$の部分クラスを定義する。$\mathcal{M}_0^p$についても同様に定義する。wは有界であるから

$$\overline{w}(y) = \sup_{a \in A} |w(a,y)|$$

なる函数が定義できて、$\overline{w} \in L^\infty = L^\infty(Y,m)$, $\frac{1}{p} + \frac{1}{\infty} = 1$, である。

$\mathcal{M}^p$ に次のような距離を入れる。

$$\delta_p(\eta_1, \eta_2) = \|w\|_q \cdot \left\| \frac{d\eta_1}{dm} - \frac{d\eta_2}{dm} \right\|_p .$$

Schwartz の不等式から明らかなように

$$\delta^*(\eta_1, \eta_2) \leqq \delta_p(\eta_1, \eta_2)$$

である。即ち δ_p の方が δ^* より強い位相である。

$\mathcal{M}^p$ より D の中への可測写像 $v = v(\eta)$ を decision procedure と呼ぶ。ある $\varepsilon \geqq 0$ に対して

$$r^*(v(\eta), \eta) - r^*(\eta) \leqq \varepsilon , \qquad \text{for all } \eta \in \mathcal{M}^p$$

となる procedure v を ε-optimal と言い、v^ε と表わす。

$\mathcal{Y}$ の σ-体 $\mathcal{B}$ に、確率測度 m を用いて

$$\rho(E, F) = m(E \triangle F) = m(E - F) + m(F - E) , \quad E, F \in \mathcal{B}$$

な距離を導入する。以下距離空間 $(\mathcal{B}, \rho)$ が可分であると仮定する。位相解析の定理（例えば Zaanen: Linear Analysis P.74）によって、この仮定の下で $(\mathcal{M}^p, \delta_p)$ が可分になる。（$\mathcal{Y}$ の距離よの意味での可分性のみからではこの性質は導れない。）

次に $\mathcal{M}^p$ に於ける任意の系列 $\underline{\eta} = \{\eta_n\}$ を考えて、

$$\eta_n^{(0)} = \eta_n , \qquad n = 1, 2, \cdots$$

$$\eta_n^{(i)} = \frac{1}{n} \sum_{k=1}^{n} \eta_k^{(i-1)} , \quad n = 1, 2, \cdots ; \ i = 1, 2, \cdots$$

なる系列を順次定義する。あるに対し

$$\lim_{n \to \infty} \delta_p(\eta_n^{(i)}, \eta) = 0$$

となる $\eta \in \mathcal{M}^p$ が唯一つ存在するとき、$\underline{\eta} = \{\eta_n\}$ は (H, i)-収束

すると言い、$\xi \in H^{(i)}$ と書くことにする。明らかに $H^{(0)} \subset H^{(1)} \subset \cdots$ である。i 次の極限を $\xi^{(i)}$ と表わす。

さて以上の準備の下で、一般的な決定問題を定式化しよう。$(\Omega, \mathcal{S}, P)$ を普々の基本確率空間とする。この空間より標本空間 $\mathcal{X}$ およびパラメータ空間 $\mathcal{Y}$ の中への可測写像 $X = X(\omega)$, $Y = Y(\omega)$ を夫々ランダムサンプル，ランダムパラメータと呼ぶ。$\{(X_n, Y_n), n=1,2,\cdots\}$ をランダムサンプル，ランダムパラメータの組の独立な系列とする。Y_n の分布は $\xi_n \in \mathcal{M}_0^P$（同一分布でなくとも良い），$Y_n = y$ を与えたときの X_n の(条件付)分布は $\nu(\cdot, y)$（ν は μ に関する微分 $f(\cdot, y)$）で与えられる。次に、$\xi = \{\xi_n\} \in H^{(i)}$ となる最小の整数を i_0 とする（かかる有限な i_0 は存在するものと仮定して）。各 $d \in D$ に対し次のような確率変数列を順次定義する。

$$W_n^{(0)}(d) = w(d(X_n), Y_n), \quad n = 1, 2, \cdots,$$

$$W_n^{(i)}(d) = \frac{1}{n} \sum_{k=1}^{n} W_k^{(i-1)}(d), \quad n = 1, 2, \cdots; \ i = 1, \cdots, i_0.$$

容易に判るように

$$\mathrm{E}\{W_n^{(i_0)}(d)\} = r^*(d, \xi_n^{(i_0)}). \tag{2}$$

この値を最小にするような $d \in D$ を見出すことが望ましいが、分布 $\xi_n^{(i_0)}$ に関する完全な情報が与えられていない場合には不可能である。しかし、n 番目の決定を考す前に $X_1, X_2, \cdots, X_n$

に関する観測が与えられている場合は常である。そこでこのような先験情報を活用して

$$\delta_p(\hat{\eta}_n, \eta_n^{(i_0)}) \xrightarrow{P} 0$$

となる推定量 $\hat{\eta}_n = \hat{\eta}_n(X_n)$ ($\eta^{(i_0)}$ に対する一致推定量) が与えられたとすると、(2) を最小にする d ($\eta_n^{(i_0)}$ に対するベーズ) と同じ程度良い"ランダム"な決定法則を作ることができるのではないかと期待するのは自然であろう。実際次の結果が得られる。

定理1. $L^p = L^p(Y, m)$ に於る $g_n^{(i_0)} = \frac{d\eta_n^{(i_0)}}{dm}$ に対する推定量 $\hat{g}_n = \hat{g}_n(X_n)$ が、関係

(3) $$E\{\|\hat{g}_n - g_n^{(i_0)}\|_p\} \leqq K n^{-\Delta} \quad (K, \Delta \text{ はある正定数})$$

を満足するものとする。

$$\hat{\eta}_n(E) = \int_E \hat{g}_n(y)\, m(dy), \quad E \in \mathcal{Y}$$

によって $\mathcal{M}^p$ に於る推定量を定義する。$\varepsilon_n = \|w\|_q\, n^{-\Delta}$ と置いて

$$\hat{d}_n = V^{\varepsilon_n}(\hat{\eta}_n), \quad n = 1, 2, \cdots$$

により"ランダム"な決定函数の系列を定義すると、

$$E\{r^*(\hat{d}_n, \eta_n^{(i_0)})\} \leqq r^*(\eta_n^{(i_0)}) + (K+1)\|w\|_q\, n^{-\Delta}.$$

従って、任意に固定された $\varepsilon > 0$ に対し

$$P\{r^*(\hat{d}_n, \eta_n^{(i_0)}) - r^*(\eta_n^{(i_0)}) \geqq \varepsilon\} \leqq \varepsilon^{-1}(K+1)\|w\|_q\, n^{-\Delta}.$$

即ち

(4) $$r^*(\hat{d}_n, \eta_n^{(i_0)}) - r^*(\eta_n^{(i_0)}) \xrightarrow{P} 0.$$

[注1.] ε-optimal な decision procedure の作り方について簡単に触れて置く。距離空間 $(\mathcal{M}^P, \delta_P)$ が可分であるから、任意の $\varepsilon>0$ に対して次の関係を満すような高々可附番個の $\mathcal{M}^P$ の（空でない）可測部分集合の系列 $\{\mathcal{M}_n, n=1,2,\cdots\}$ がとれる:

$$\bigcup_{n=1}^{\infty}\mathcal{M}_n=\mathcal{M}^P,\quad \mathcal{M}_i\cap\mathcal{M}_j=\phi\ (i\neq j),$$

$$\xi_1,\xi_2\in\mathcal{M}_n \text{ なら } \delta_P(\xi_1,\xi_2)\leq\frac{\varepsilon}{3},\ n=1,2,\cdots.$$

次に各 $\mathcal{M}_n$ から任意に一つ ξ_n をとり出し、

$$r^*(d_n,\xi_n)-r^*(\xi_n)\leq\frac{\varepsilon}{3}$$

となるような $d_n\in D$ を一つ選んで

$$v(\xi)=d_n,\quad \text{for all } \xi\in\mathcal{M}_n,\ n=1,2,\cdots$$

なる procedure を定義すると、これが ε-optimal なものであることは容易に証明され、而も v の値域が D の元の高々可附番個のものから成っていることも明らかである。

[注2.] (4) の関係を成立せしめるようなランダムな決定列 $\hat{\underline{d}}=\{\hat{d}_n\}$ も $\underline{\xi}=\{\xi_n\}\in H^{(i_0)}$ に対して (H,i_0)-optimal であると呼ぶことにすれば、§2 で紹介した複合型決定問題（最適な simple 決定法則に近似を与える問題）は $(H,1)$-optimal なものを見出す問題に還元される。尚その際、Y_n (Nature) の与える分布 (strategy) ξ_n は点 θ_n で全質量を与える一点分布 (pure strategy) であると考えれば完全に一致する。又§3で記述したベーズ推論の場合には、$(H,0)$-optimal なものを

見出す問題になる（この場合各ζ_nは同一分布$g=\frac{d\eta}{d\mu}$に従っているものと考えれば良い）。

定理1によって、(3)の関係を満すような推定量$\hat{g}_n=\hat{g}_n(X_n)$を作れば問題は完結する。そこで$g_n^{(i_0)}$（$i_0=1$の場合）に対する推定を考えてみよう。（一般のi_0にしたまゝでも平行した議論が可能であり、又$g_n=g$（$n=1,2,\cdots$）と置くことにより未知の唯一の先験分布gの推定に於る結果も含むことになる。）更に（直交展開が使える等の）種々の観点から、$p=2$の場合を考えるのが妥当であろう。

即ち問題は、独立で観測可能な確率変数の系列$\{X_n,\ n=1,2,\cdots\}$が与えられていて、各X_nは（必ずしも同一でない）分布

(5) $$f(x,g_n)=\int_Y f(x,y)g_n(y)m(dy),\quad n=1,2,\cdots$$

に従っているものとする。更に核函数$f(\cdot,\cdot)$は直積空間$\mathcal{X}\times\mathcal{Y}$で可測で、Identifiabilityの条件

$$m\{y\in\mathcal{Y}:g_i(y)\neq g_j(y)\}>0 \text{ ならば } \mu\{x\in\mathcal{X}:f(x,g_i)\neq f(x,g_j)\}>0$$

を満しているとする。

最初にパラメータ空間が有限個から成る場合$\mathcal{Y}=\{y_1,y_2,\cdots,y_k\}$について考える。$m$は各点に$1/k$の質量を与える離散型測度にとる。

$$g_n(y_j)=k\alpha_j^{(n)},\quad j=1,2,\cdots,k;\ n=1,2,\cdots$$

$$f(x, y_j) = \beta_j(x), \quad j = 1, 2, \cdots, k$$

と置くとき、X_nの分布(5)は

(6) $$f(x; \underline{\alpha}^{(n)}) = \sum_{j=1}^{k} \alpha_j^{(n)} \beta_j(x)$$

と表わされる。

$\mathcal{P}_k = \{\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_k\}$が$\underline{\beta}_k = (\beta_1, \cdots, \beta_k)$に関する測度空間$(\mathcal{X}, \mu)$の正則分割であるとは

a) 各$\mathcal{X}_i$はμ-可測, $\mathcal{X} = \bigcup_{i=1}^{k} \mathcal{X}_i$, $\mu(\mathcal{X}_i \cap \mathcal{X}_j) = 0$ $(i \neq j)$

b) 行列 $D_k = D_k(\mathcal{P}_k; \underline{\beta}_k) = \left[\int_{\mathcal{X}} \beta_i(x) \chi_j(x) \mu(dx)\right]$が正則、但し$\chi_j$は集合$\mathcal{X}_j$の特性函数である。

このような正則分割を用いて

(7) $$\varphi_i(x) = \sum_{j=1}^{k} d_{ij}^{*} \chi_j(x), \quad i = 1, 2, \cdots, k$$

なる函数が定義できる。d_{ij}^{*}はD_kの逆行列の(i, j)-元とする。

定理2. 独立な確率変数列$\{X_n, n = 1, 2, \cdots\}$は夫々(6)の分布に従うものとする。(7)の函数を用いて

$$\hat{\alpha}_i^{(n)} = \hat{\alpha}_i^{(n)}(X_n) = \frac{1}{n} \sum_{l=1}^{n} \varphi_i(X_l), \quad i = 1, 2, \cdots, k$$

なる推定量を作れば、

$$E\{\hat{\alpha}_i^{(n)}\} = \bar{\alpha}_i^{(n)} = \frac{1}{n} \sum_{l=1}^{n} \alpha_i^{(l)}, \quad i = 1, 2, \cdots, k,$$

(8) $$\sum_{i=1}^{k} V\{\hat{\alpha}_i^{(n)}\} = E\left\{\sum_{i=1}^{k} (\hat{\alpha}_i^{(n)} - \bar{\alpha}_i^{(n)})^2\right\} \leq d_k^{*2} k / n,$$

但し$(d_k^{*})^{-1}$は行列式$|D_k|$の絶対値である。

(8)の不等式より

$$E\{\|\hat{g}_n - g_n^{(l)}\|_2^2\} = E\left\{k \sum_{i=1}^{k} (\hat{\alpha}_i^{(n)} - \bar{\alpha}_i^{(n)})^2\right\} \leq k^2 d_k^{*2} / n$$

であるから、結局

$$E\{\|\hat{g}_n - g_n^{(1)}\|_2\} \leqq \left[E\{\|\hat{g}_n - g_n^{(1)}\|_2^2\}\right]^{1/2} \leqq k d_k^* n^{-1/2}.$$

従ってこの場合には、$K = k d_k^*$, $\Delta = \frac{1}{2}$ として (3) の関係を満すような推定量が作れた。

[注3.] (7) で与えられる φ_i 及び (6) に従う確率変数 X_n に対して

$$E\{\varphi_i(X_n)\} = \alpha_i^{(n)}$$

が成立するが、Robbins [4; §7] は一次独立な函数系 $\{\beta_1, \cdots, \beta_k\}$ より正規直交系 $\{\beta_1^*, \cdots, \beta_k^*\}$ を作りだして、$\beta_i^*(X_n)$ を $\alpha_i^{(n)}$ の不偏推定として用いた。φ_i と β_i^* との間には

$$\varphi_i(x) = \beta_i^*(x) + \tilde{\varphi}_i(x) \quad a.e.\,\mu, \quad i = 1, 2, \cdots, k$$

なる関係が成立する。ここで $\tilde{\varphi}_i$ は $\{\beta_1, \cdots, \beta_k\}$ によって span される $L^2(\mathcal{X}, \mu)$ の部分空間に直交する成分である。つまり、Robbins の直交化の方法は $\tilde{\varphi}_i(x) = 0$ $a.e.\,\mu$, $i = 1, 2, \cdots, k$ なる $\varphi_k = (\varphi_1, \cdots, \varphi_k)$ を用いたに過ぎない。如何なる $\tilde{\varphi}_i$ をとるのが（分散を小さくする等の意味に於て）最適かという問題は、そう単純には解決されそうにない。従って作り方が単純であるという見地からは、吾々の提示した方法の方にやゝ利があると言えるかも知れない。又完全に直交化する方法は、解析的取扱いを要易にするが、有限でない場合の吾々の truncation に依る方法に対しては用いることができない。

最後に一般の $L^2 = L^2(Y, m)$ の場合の結果を述べよう。この場合には少々きつい条件が附加されるので、実用上未だ難点が残ろう。$g_n \in L^2$ であるから適当な正規直交系 $\{\psi_j, j=1,2,\cdots\}$ を用いて

$$g_n(y) = \sum_{j=1}^{\infty} \alpha_j^{(n)} \psi_j(y), \quad n=1,2,\cdots$$

の如く展開できる。従って X_n の分布 (5) は

$$f(x;\alpha^{(n)}) = \sum_{j=1}^{\infty} \alpha_j^{(n)} \beta_j(x), \quad n=1,2,\cdots \tag{9}$$

$$\beta_j(x) = \int_Y f(x,y)\psi_j(y)\, m(dy), \quad j=1,2,\cdots$$

と表わせる。

次のような条件を置く。ある非負値定数 γ 及び $\mathcal{X}$ の分割列 $\{\mathcal{P}_k, k=2,3,\cdots\}$ が与えられていて次の条件を満す。即ち、各分割 $\mathcal{P}_k$ が $\beta_k = (\beta_1,\cdots,\beta_k)$ に関する正則分割で、而も整数 k_0 と正定数 C とが存在して

$$d_k^* = |\det D_k(\mathcal{P}_k;\beta_k)|^{-1} \leqq C k^{\gamma}, \quad k = k_0, k_0+1, \cdots$$

更に (9) に表れる係数が、ある $\delta > \gamma$ に対し次の条件を満す。

$$|\alpha_j^{(n)}| = \left|\frac{1}{n}\sum_{\ell=1}^{n} \alpha_j^{(\ell)}\right| \leqq B j^{-(\frac{1}{2}+\delta)}, \quad j=1,2,\cdots \text{ (B は正定数)}.$$

このような仮定の下で、次の一般的命題が成立する。

定理 3. 各 n に対し

$$k = k(n) = [n^{\frac{1}{1+2\delta}}] \qquad ([\cdot] \text{はガウス記号})$$

と置く。(7) で与えられる $\varphi_n = (\varphi_1, \cdots, \varphi_k)$ を用いて

$$\hat{g}_n(y) = \sum_{j=1}^{k(n)} \hat{\alpha}_j^{(n)} \psi_j(y) = \sum_{j=1}^{k(n)} \left[\frac{1}{n}\sum_{\ell=1}^{n} \varphi_j(X_\ell)\right] \psi_j(y)$$

なる推定量を作ると

$$E\{\|\hat{g}_n - g_n^{(1)}\|_2^2\} = E\{\sum_{j=1}^{k}(\hat{\alpha}_j^{(n)} - \bar{\alpha}_j^{(n)})^2 + \sum_{j=k+1}^{\infty}\bar{\alpha}_j^{(n)2}\} = O(n^{-\frac{2(\delta-\gamma)}{1+2\delta}}).$$

従って

$$E\{\|\hat{g}_n - g_n^{(1)}\|_2\} = O(n^{-\frac{\delta-\gamma}{1+2\delta}}).$$

[注4.] 条件式 (10) は技巧的でやゝ不自然な感じを与えるので、各 β_i が非負値函数（例えば密度函数）である場合に一つの十分条件を示してみよう。$\mathcal{X} = \sum_{i=1}^{\infty}\mathcal{X}_i$, $\mu(\mathcal{X}_i \cap \mathcal{X}_j) = 0$, $i \neq j$ なる可附番個の分割が次の条件を満しているとする。

$$\det M_{k_0} \geq \frac{1}{Ck_0^{\delta}}$$

$$\int_{\mathcal{X}_k}\beta_k(x)\mu(dx) \geq 1 - \frac{1}{2C}\left\{\frac{1}{(k-1)^{\delta}} - \frac{1}{k^{\delta}}\right\}, \quad k = k_0+1, k_0+2, \cdots$$

但し

$$M_k = \left(\int_{\mathcal{X}_j}\beta_i(x)\mu(dx)\right)_{i,j=1,2,\cdots,k}, \quad k = k_0, k_0+1, \cdots$$

となる正定数 C 及び整数 k_0 が存在する。"

まず Hadamard の不等式を用いて

(11) $$\det M_k \leq \frac{1}{Ck^{\delta}}, \quad k = k_0, k_0+1, \cdots$$

なる関係が示される。

$$P_k^* = \{\mathcal{X}_1, \cdots, \mathcal{X}_{k-1}, \sum_{i=k}^{\infty}\mathcal{X}_i\}, \quad k = 2, 3, \cdots$$

なる分割列を考えると、β_i の非負値性と M_k の定義より

$$\det D_k(P_k^*) \geq \det M_{k-1} - \frac{1}{C}\left\{\frac{1}{(k-1)^{\delta}} - \frac{1}{k^{\delta}}\right\}.$$

これと (11) の不等式より (10) の関係を得る。