数理解析研究所講究録
第504巻 1983年 147-156

Conditional probability and I-divergence

東工大・理　菅原 昭博 (Akihiro Sugawara)

確率測度 P, Q に対して I-divergence (Kullback-Leibler information number) $I(P|Q)$ は

$$I(P|Q) = \begin{cases} \int \log \frac{dP}{dQ} \cdot dP & , \text{if } P \ll Q \\ +\infty & , \text{otherwise.} \end{cases}$$

と定義される。 また, 確率測度の族 $\mathcal{P}$ と確率測度 Q に対しては

$$I(\mathcal{P}|Q) = \begin{cases} \inf\{I(P|Q) ; P \in \mathcal{P}\} & , \mathcal{P} \neq \emptyset \\ +\infty & , \mathcal{P} = \emptyset \end{cases}$$

と定義する。

この小論では, 標本の数が大きい観測において, 二つの確率測度の分離度とされる I-divergence が, どのような働きをするかに視点をあわせる。 特に, 標本平均による条件付

確率測度と I-divergence の最小化がいかなるかかわりを持つかが示される。

§1 I-divergence と仮説検定

仮説検定における第一種の誤りの確率が，I-divergence によって説明されることをみる。

$(X, \mathcal{F})$ を標本空間（可測空間），その上の確率分布の族を $\{P_\theta ; \theta \in \Theta\}$ とする。母数空間を $\Theta = \{\theta_0, \theta_1\}$ として，次の単純仮説検定を考える。

$$\begin{cases} H_0 ; \theta = \theta_0 & (\text{帰無仮説}) \\ H_1 ; \theta = \theta_1 & (\text{対立仮説}) \end{cases}$$

仮説検定では，$(X, \mathcal{F})$ 上の $[0, 1]$ への可測写象を検定（test）と呼び，検定 φ が与えられたとき，

$$\alpha(\varphi) = E_{\theta_0}(\varphi) \ , \quad \beta(\varphi) = E_{\theta_1}(1-\varphi)$$

を，それぞれ第一種（第二種）の誤りの確率という。一般に有意水準 α の検定とは $\alpha(\varphi) \leq \alpha$ の検定で，その中で $\beta(\varphi)$ を最小にするものを Neyman-Pearson の検定基準を満たすという。検定 φ は，$x \in X$ が観測の結果得られたとき，確率 $\varphi(x)$ で仮説 H_0 を棄却，確率 $1-\varphi(x)$ で採択することを意

味する。

n 回の観測を行なうということは，上の仮説検定を標本空間を $(\mathcal{X}^{(n)}, \mathfrak{X}^{(n)})$，その上の確率分布を $\{P_\theta^{(n)} ; \theta \in \Theta\}$

$$\mathcal{X}^{(n)} = \prod_{i=1}^{n} \mathcal{X} \quad , \quad \mathfrak{X}^{(n)} = \bigotimes_{i=1}^{n} \mathfrak{X}$$

$$P_\theta^{(n)} = \bigotimes_{i=1}^{n} P_\theta$$

として考えることを意味する。

標本空間 $(\mathcal{X}^{(n)}, \mathfrak{X}^{(n)})$ における検定の全体を Φ_n とする，すなわち，

$$\Phi_n = \{ \varphi_n ; \mathcal{X}^{(n)} \to [0,1], \text{可測} \}.$$

また，α $(0 < \alpha < 1)$ に対して，

$$\beta_n(\alpha) = \inf \{ E_{\theta_1}(1-\varphi_n) ; \varphi_n \in \Phi_n, E_{\theta_0}(\varphi_n) \leq \alpha \}$$

とする。 Neyman-Pearson の定理によって，$\varphi_n^* \in \Phi_n$ で

$$\beta_n(\alpha) = E_{\theta_1}(1-\varphi_n^*), \ E_{\theta_0}(\varphi_n^*) = \alpha$$

を満たすものが存在することを注意する。

定理1. 任意の α，$0 < \alpha < 1$ に対して，

$$\lim_{n \to \infty} \frac{1}{n} \log \beta_n(\alpha) = -I(\theta_0 | \theta_1).$$

ここで，$I(\theta_0|\theta_1) = I(P_{\theta_0}|P_{\theta_1})$ である。

この定理は統計で Stein's Lemma として知られているものである．証明は Bahadur [1] が詳しい。これによって，第一種の誤りの確率を一定にしたときの，第二種の誤りの確率の限界が I-divergence によって与えられることが示されたことになる。

§2. Large deviations

$\{f_i ; i=1,2,\cdots\}$ を確率空間 $(X, \mathcal{X}, P)$ 上の独立な同一分布を持つ確率変数列とする．このとき，確率

$$P\left\{\frac{1}{n}\sum_{i=1}^{n} f_i \geq a\right\}$$

に注目する。平均 $E(f_1)$ が存在するとき，大数の法則により，

$$P\left\{\frac{1}{n}\sum_{i=1}^{n} f_i \geq a\right\} \to 0 \quad , \quad \text{if } a > E(f_1).$$

である。この収束の速さと I-divergence の関係をみる。

f を $(X, \mathcal{X}, P)$ 上の確率変数とする。f の m.g.f (moment generating function) を

$$\varphi(t) = \int \exp(t f(x))\, P(dx) \quad , \quad -\infty < t < +\infty .$$

で定義する。 φ は R 上の convex function で $\{t : \varphi(t)<\infty\}$ は 0 を含む区間である。 また, $P\{f=0\}<1$ かつ, $\varphi(b)<\infty$ $(\exists b>0)$ であれば,

(i) φ : strictly convex, continuous on $[0, b]$

(ii) $\varphi \in C^{\infty}(0, b)$

(iii) φ' : strictly increasing on $(0, b)$

(iv) $\varphi'(0+)=E(f)$, $\varphi'(b-)=E(f\cdot \exp bf)$.

以上はよく知られた m.g.f の性質である。

Chernoff [3] は m.g.f を用いて次を示した。

定理2. $\{f_i ; i=1, 2, \cdots\}$ は確率空間 $(X, \mathcal{F}, P)$ 上の独立同一分列で, f_i の m.g.f φ において, $\varphi(t)<+\infty$ $(\text{for } {}^{\forall} t\in[0, b))$ とする。 このとき,

$$a \in \left\{ \frac{\varphi'(t)}{\varphi(t)} : t\in(0, b) \right\}$$

に対して

$$P\left\{ \frac{1}{n}\sum_{i=1}^{n} f_i \geq a \right\} \leq \rho(a)^n,$$

$$\lim_{n\to\infty} \left(P\left\{ \frac{1}{n}\sum_{i=1}^{n} f_i \geq a \right\} \right)^{\frac{1}{n}} = \rho(a)$$

ここで,

$$\rho(a) = \inf_t \varphi(t)\exp(-ta)$$

である。

上記の $\rho(a)$ と I-divergence を結びつけた次の結果は，Sanov [7]，Hoeffding [6] による。

定理3.　確率分布の族 $\mathcal{P}$ を

$$\mathcal{P}=\left\{Q\,;\ \int f_i(x)\,Q(dx)\geq a_i\right\}$$

とすれば，

$$\rho(a)=\exp\left(-I(\mathcal{P}\mid P)\right)$$

である。

さて，上記の二つをまとめると，

$$\lim_{n\to\infty}\frac{1}{n}\log P\left\{\frac{1}{n}\sum_{i=1}^{n}f_i\geq a_i\right\}=-I(\mathcal{P}\mid P)$$

となる。Groenboom, Oosterhoff and Ruymgaart [5] は，この型の定理について，より一般的な場合について調らべている。

§3. Conditional probability and I-divergence.

前節までは標本の数が大きいときのある種の確率の収束の速さが，I-divergence で表わされることをみた。以下では，前の結果が条件付確率を持ち出すことにより，見通しがよくなることをみる。

$(X, \mathcal{F}, P)$ を確率空間とする。$A\in\mathcal{F}$, $P(A)\neq 0$

による条件付確率測度を

$$P_A(B)=P(B|A)\ ,\ B\in\mathfrak{X}$$

と表記する。

条件付確率測度はI-divergenceで特徴付けできる。すなわち次の定理が成立する。

定理4. $A\in\mathfrak{X}$, $P(A)\neq 0$ に対して

$$\mathcal{P}_A=\{Q;\ Q(A)=1\}$$

とすれば, P_A は $\mathcal{P}_A$ の中でI-divergenceを最小にする。

$$I(P_A|P)=I(\mathcal{P}_A|P)\ .$$

これは, $I(P_A|P)=-\log P(A)$ に注意すれば, 容易に証明することができる(文献[9])。

前節の定理を条件付確率を用いて書きかえる。ただし簡単のため, 有界性の条件を確率変数に付加する。

定理5. $(X,\mathfrak{X},P)$ を確率空間, $\{x_i;\ i=1,2,\cdots\}$ を, その上の自分自身に値をとる確率変数列で, 独立同一分布($Px_i^{-1}=P$)とする。f をX上の実確率変数で

$$\text{ess. inf}\, f(x_1)<a<\text{ess. sup}\, f(x_1)$$

とし，以後 $f(x_i)$ を f_i と書く．

$$\mathcal{P} = \{ Q ; \int f_1(x)\, Q(dx) \geq a \}$$

$$P_n(\cdot) = P(\cdot \mid \tfrac{1}{n} \textstyle\sum_{i=1}^{n} f_i \geq a)$$

とすれば，

$$\lim_{n\to\infty} \frac{1}{n} I(P_n | P) = I(\mathcal{P} | P)。$$

これは，$I(P_n|P) = -\log P\{\frac{1}{n}\sum_{i=1}^{n} f_i \geq a\}$ に注意すれば前定理そのものであるが，これによって，n 回の観測によって得られる情報量の平均値は，n が十分大きいとき，$I(\mathcal{P}|P)$ で近似されることがわかる。

ところで，我々は，$P^* \in \mathcal{P}$ が存在して

$$I(P^*|P) = I(\mathcal{P}|P)$$

を示すことができる。では，P_n と P^* の関係はどうなるのであろうか？

定理6.　任意の $A \in \mathfrak{X}$ に対して

$$P_n(A) \longrightarrow P^*(A) \quad (n \to \infty)。$$

証明の概略は，$(X, \mathfrak{X})$ 上の有界な実確率変数 g を持て来る。$g(x_i)$ を g_i と書き，$\bar{g}$ を

$$\bar{g} = \int g(x) P^*(dx).$$

として，任意の $\varepsilon > 0$ に対して

$$P_n\left(\left| \frac{1}{n} \sum_{i=1}^{n} g_i - \bar{g} \right| \geq \varepsilon \;\middle|\; \frac{1}{n} \sum_{i=1}^{n} f_i \geq a \right) \to 0$$

を示す，これから，

$$\int g P_m(dx) \to \int g P^*(dx).$$

となり，目的は達せられる。詳細は文献[9]にゆずる。

これにより，I-divergence を最小化することによって得られた確率測度 P^* は，標本平均による条件付確率測度の極木を意味していることが分かる。

REFERENCES.

[1] Bahadur,R.R.; 'Some Limit Theorems in Statistics', SIAM, Philadelphia (1971).

[2] Bahadur,R.R. and Zabell,S.L.; Large deviations of the sample mean in general vector spaces, Ann. Probability 7 587-621(1979).

[3] Chernoff,H.; A measure of asymptopic efficiency for tests of a hypothesis based on sums of observations, Ann. Math. Statist. 23, 493-507(1952).

[4] Csiszar,I.; I-divergence geometry of probability distributions and minimization problems, Ann. Probability 3, 146-158(1975).

[5] Groeneboom,R., Oosterhoff,J. and Ruymgaart,F.H.; Large deviation theorems for empirical probability measures, Ann. Probability 7, 553-586(1979).

[6] Hoeffding,W.; On probability of large deviations, Proc. Fifth Berkeley Symp. Math. Statist. Prob. 1, 203-219 (1967).

[7] Sanov,I.N.; On the probability of large deviations of random variables, Sel. Transl. Math. Statist. Prob. 1, 213-244(1967).

[8] Sugawara,A.; A mathmatical information channel with a non-commutative intermediate system (submitted).

[9] Sugawara,A.; Conditional probability and I-divergence, (under preparation).