エントロピーと同時分布推定

堀部安一

1. はじめに

周辺分布から同時分布は一意には定まらない——これはたいへん明らかなことなので，確率統計の教科書にも取り立てては出てこないほどである．たとえば二値0,1をとる変数 x_1, x_2, x_3 に関する三通りの2次元分布 $p_{(1,2)}, p_{(1,3)}, p_{(2,3)}$（$p_{(i,j)}$ は (x_i, x_j) の分布をあらわす）がつぎのように与えられているとしよう．

x_1x_2	$p_{(1,2)}(x_1,x_2)$	x_1x_3	$p_{(1,3)}(x_1,x_3)$	x_2x_3	$p_{(2,3)}(x_2,x_3)$
0 0	3/8	0 0	1/8	0 0	1/8
0 1	1/8	0 1	3/8	0 1	3/8
1 0	1/8	1 0	3/8	1 0	3/8
1 1	3/8	1 1	1/8	1 1	1/8

これらは，つぎの分布 $p=\{p(x_1,x_2,x_3)\}$ の周辺分布になっていることがわかる．

x_1	x_2	x_3	$p(x_1,x_2,x_3)$
0	0	0	t
0	0	1	$3/8-t$
0	1	0	$1/8-t$
0	1	1	t
1	0	0	$1/8-t$
1	0	1	t
1	1	0	$2/8+t$
1	1	1	$1/8-t$

$(0\leqq t\leqq 1/8)$

すなわち，p の (x_i, x_j)—周辺分布（これを $p_{/(i,j)}$ であらわす）をとると，

$$p_{/(i,j)}=p_{(i,j)},\ 1\leqq i<j\leqq 3,\ t\in\left[0,\ \frac{1}{8}\right].$$

多数の変数に関する同時分布を伴う数理的モデルでは，いくつかの低次元周辺分布だけから，同時分布を推定しなければならないことがおこりうる——限られたデータから直接信頼度高く得られるのは低次元分布だけだからである．

以下で述べる最大エントロピー的同時分布推定は，この問題に対する1つの興味ある方向と思われる．その際に，divergence とよばれる情報理論的な量 $D(p,q)=\sum p(x)\log(p(x)/q(x))$ の演ずる役割が基本的であることがわかる．

2. 記号など

N 個の変数 $x_1,\cdots,x_N$ があり，簡単のためどれも二値0,1をとるとし，われわれには $\binom{N}{2}$ 通りの2次元分布 p_E, $E=\{i,j\}$, $1\leqq i<j\leqq N$, のみがあたえられているとしよう．ここで p_E は，$E=\{i,j\}$ なら (x_i,x_j) の分布をあらわしている．添字 $1,\cdots,N$ を頂点とし，$E=\{i,j\}$ を頂点 i,j 間の辺とする N 頂点完全グラフ K_N において，各辺 E に分布 p_E が対応していると考えてもよい．1節の例では図1のようである．

N 次元分布 $p=(p(x))$, $x=(x_1,\cdots,x_N)$, は 2^N 個の成分をもつ（ベクトル）点とみられる．あらゆる N 次元分布の全体を Π とする．$p\in\Pi$ の (x_i, x_j)—周辺分布を $p_{/E}$, $E=\{i,j\}$, であらわし，$p_{/E}=p_E$ ならば p は p_E を**満たす**と言おう．すべての

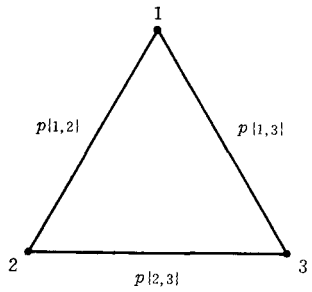

図 1 K_3

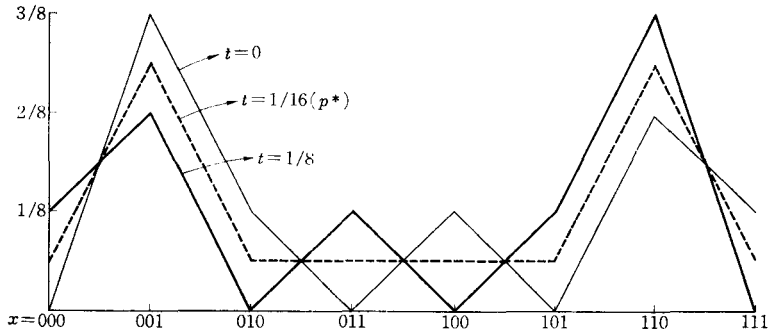

図 2

Eについて$p_{/E}=p_E$となっているpの全体をΠ_0とする($\Pi_0\subset\Pi$). Π_0が空だと意味がないので$\Pi_0\neq\phi$とする. 1節の例ではtを[0, 1/8]内で動かして得られる分布$(t, 3/8-t, \cdots, 1/8-t)$の全体が$\Pi_0$である.

$E=\{i,j\}$のとき$x=(x_1,\cdots,x_N)$の部分ベクトル(x_i,x_j)をx_Eであらわす. また, すべての辺E, すべてのx_Eについて$p_E(x_E)>0$としておこう——もしある$E=\{i,j\}$と$(x_i,x_j)=(a,b)$に対して$p_E(a,b)=0$なら, $(x_i,x_j)=(a,b)$なるすべてのxに対して$p(x)=0$となる$p\in\Pi$のみに話を限定すればよい.

問題は, Π_0に属すどの分布を推定分布とし, それをどのように見出せばよいかということになる.

3. エントロピー最大の分布

分布pのエントロピー$H(p)$は, その分布が《支配する》事象生起の不確定さの程度を測る量であって,

$$H(p)=-\sum_x p(x)\log p(x)$$

で与えられる. 対数の底は以後2とする.

『われわれは2次元分布族$\{p_E\}$しかわかっていない. したがって$\{p_E\}$を満たす分布のうち(すなわちΠ_0内の分布のうち)最も不確定さの大きい分布を推定分布とせざるをえない』という立場をとるならば,

$$H(p^*)=\max_{p\in\Pi_0} H(p)$$

なる$p^*\in\Pi_0$を推定分布とするのが自然であろう.

p^*は, $H(p)$が狭義凸関数[$-u\log u$の狭義凸性より]であり, Π_0が凸集合をなすことより, 一意に決まる. [Π_0の凸性は, $p, p'\in\Pi_0, \lambda+\lambda'=1$, $(\lambda p+\lambda' p')_{/E}=\lambda p_{/E}+\lambda' p'_{/E}=\lambda p_E+\lambda' p_E=p_E$ より]

1節の例でのpのエントロピーは,

$$\begin{aligned}H(p)=h(t)=&-3t\log t-(3/8-t)\log(3/8-t)\\&-3(1/8-t)\log(1/8-t)\\&-(2/8+t)\log(2/8+t).\end{aligned}$$

$\max_{p\in\Pi_0} H(p)=\max_{0\leq t\leq 1/8} h(t)=h(1/16)$より,

$$p^*=(1/16, 5/16, 1/16, 1/16, 1/16, 1/16, 5/16, 1/16).$$

図2では, この例について, 分布が折線であらわされていて, 折線群がΠ_0を代表的に示している.

4. $\{p_E\}$の不完全な利用

Π_0に属さない分布でもp^*に《近い》ものが, $\{p_E\}$を使って簡単に得られる場合がある. 分布qが分布pからみてどの程度違っているか, 離れているかを, divergence とよばれる量

$$D(p,q)=\sum_x p(x)\log\frac{p(x)}{q(x)}$$

で測ることにしよう. つねに$D(p,q)\geq 0$で, $=0$となるのは$p=q$のときに限るという性質がある. $D(p,q)$については付記にて2, 3述べることとし, ここでは, $D(p,q)$が小さくなればなるほどqはpによく《似た》分布になるということだけを記しておこう.

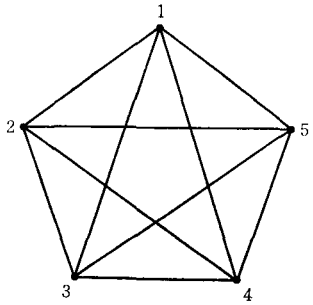

図 3 K_5

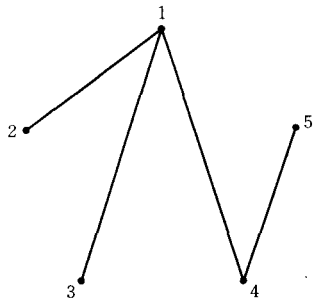

図 4 τ

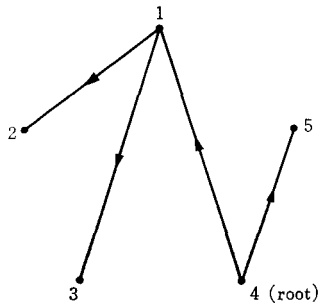

図 5 rooted tree

《情報》$\{p_E\}$ から同時分布 $p \in \Pi$ を構成する場合，$\{p_E\}$ の《利用程度》が大きくなれば，$D(p^*, p)$ は小さくなることを例でもって示してみたい．

K_5, $N=5$ を考えよう(図3)．

0) $\{p_E\}$ をまったく使わずに一様分布 $p_0, p_0(x)=2^{-5}$, を作ると，$D(p^*, p_0)=5-H(p^*)$.

1) $p(x_i, x_j)$, $1 \leqq i < j \leqq 5$ (混乱はおきないので，p_E の E など省略) より，その1次元周辺分布 $\{p(x_i)\}$, $1 \leqq i \leqq 5$ をとり，$p_1(x)=p(x_1)\cdots p(x_5)$ を作る．p^* が $\{p(x_i)\}$ を満たすことより $D(p^*, p_1)=H_1+\cdots+H_5-H(p^*)$. ここに，$H_i$ は $\{p(x_i)\}$ のエントロピー．$H_i \leqq \log 2=1$ だから明らかに $D(p^*, p_1) \leqq D(p^*, p_0)$.

2) たとえば $p_2(x)=p(x_1, x_2)p(x_3, x_4)p(x_5)$ を作る．p^* がこれら成分分布を満たすことより，

$$D(p^*, p_2)=H_{12}+H_{34}+H_5-H(p^*).$$

ここに，H_{ij} は $\{p(x_i, x_j)\}$ のエントロピー．$H_{ij} \leqq H_i+H_j$ であるから $D(p^*, p_2) \leqq D(p^*, p_1)$.

3) K_5 の完全木 (spanning tree) τ をとる (図4)．どれかの頂点たとえば4をrootとする τ の rooted tree を考える(図5)．root には $\{p(x_4)\}$ を，有向辺 (i, j) には $\{p(x_i, x_j)\}$ より得た $\{p(x_j|x_i)\}$ を対応させこれらの積によって分布 p_τ を作る：

$$p_\tau(x)=p(x_4)p(x_5|x_4)\,p(x_1|x_4)\,p(x_2|x_1)\,p(x_3|x_1)$$

これはたしかに確率分布であり root を変えると表現は変わっても分布としては同じものが得られることを容易に示すことができる．p^* が p_τ の成分分布を満たすことより，

$$D(p^*, p_\tau)=H_4+H_{5|4}+H_{1|4}+H_{2|1}+H_{3|1}-H(p^*).$$

ここに，$H_{j|i}$ は $\{p(x_j|x_i)\}$ に対する条件付エントロピー．さて x_i と x_j との相互情報量を I_{ij} とあらわすと，$I_{ij}=H_j-H_{j|i}=H_i-H_{i|j}$ だから，

$$D(p^*, p_\tau)=-(I_{12}+I_{13}+I_{14}+I_{45})+(H_1+\cdots+H_5)-H(p^*).$$

ここで右辺の I_{ij} の和の部分だけが完全木 τ のとり方に依存していて，τ の辺に対応して I_{ij} があらわれている．したがって，K_5 の各辺 $E=\{i, j\}$ について I_{ij} を計算しておき，これを E の《重み》とすれば，辺の重み和最大の完全木 τ_0 が簡単に見つかる([4])ので，$p_{\tau_0}=p_3$ とすれば任意の τ に対して $D(p^*, p_3) \leqq D(p^*, p_\tau)$ となる．したがって

$$D(p^*, p_2)=H_{12}+H_{34}+H_5-H(p^*)=-(I_{12}+I_{34})+(H_1+\cdots+H_5)-H(p^*) \geqq D(p^*, p_3)$$ となる．

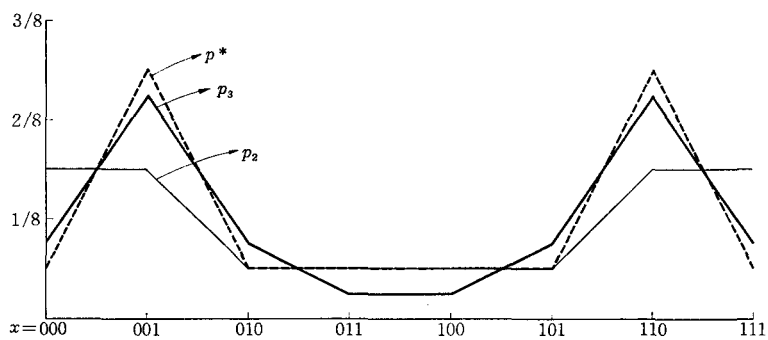

図 6

p_3 を求めるには10通りの p_E をすべて用いたが，p_3 自身には $N-1=4$ 個の p_E しか取り入れていないことに注意されたい．1節の例で，p_2 として $p(x_1,x_2)p(x_3)$ をとり，p_3 として，($I_{12}=I_{13}=I_{23}$ より)たとえば，$p(x_1)p(x_2|x_1)p(x_3|x_2)$ をとると，図2と対比して図6のようになる．

以上の計算では，p^* に関しては，それが $\{p_E\}$ を満たしているということしか使っていないので，p^* を，任意の $p\in\Pi_0$ でおきかえてもよい．したがって，divergence の意味で，p_0,p_1,p_2,p_3 の順に Π_0 に近くなっていると考えられる．

5. 反復計算

p^* を求めることは，凸集合 Π_0 上での凸関数 $H(p)$ の最大化問題にほかならないが，2^N 個もの変数 $p(x)$ を扱わなければならない．ここでは別の方法により，あるクラスに属する分布 p^0 を初期分布として，反復的に $p^0\to p^1\to\cdots\to p^*$ に至らしめることを考えよう．

前節の p_0,p_1,p_2,p_3 のように，積の形 $cq_1\cdots q_r$ をしていて，c は定数，各因子 q_k は $x_1,\cdots,x_N$ のうち高々2個の変数にしか依存しないようになっている分布を，初期分布 p^0 に選ぼう．p^* に近いという意味から，p_3 を p^0 にとるのもよい．このような p^0 に対しては，前節の具体的な計算からもわかるように，$D(p,p^0)=-\sum_x p(x)\log p^0(x)-H(p)$ において，$-\sum_x p(x)\log p^0(x)$ は，$p\in\Pi_0$ によらずに一定 c_0 であることに注意したい．

さて，p^n から p^{n+1} を求めるのに，p^n が(divergence の意味で)《最も満たしていない》p_E を p_{E_n} とする(greedy!)：

$$D(p_{E_n},p^n{}_{/E_n})=\max_E D(p_E,p^n{}_{/E}).$$

そこで p_{E_n} が満たされるように p^n に《押し込む》：

$$(1)\quad p^{n+1}(x)=\frac{p^n(x)}{p^n{}_{/E_n}(x_{E_n})}\cdot p_{E_n}(x_{E_n})$$

こうして，p^{n+1} が p_{E_n} を満たす分布になったことは明らかだが，そのつぎの段階でもこの p_{E_n} が満たされているという保証はない．

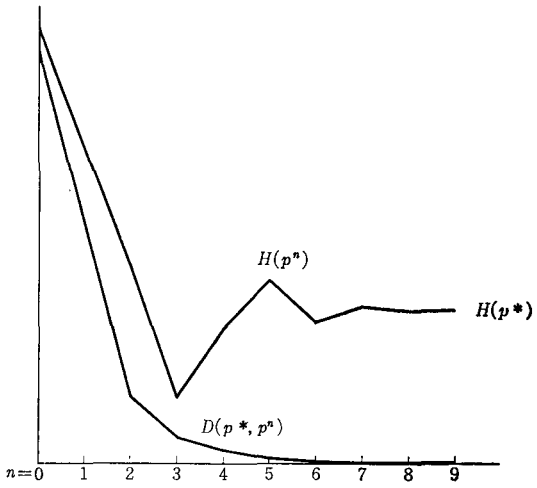

図 7

(1)によると，計算時間を喰うことはよしとするなら，この反復計算では常時 $O(N^2)$ の記憶量でよいことがわかる：$p^0(x)=2^{-N}$ の場合，各辺 E には2つの表 $\{p_E(x_E)\}$ と $\{r_E(x_E)\}$ が対応していると考え，初期では $r_E(x_E)$ はすべて1にセットされている．n 段階にきたとき，辺 E_n においては，

$$r_{E_n}(x_{E_n})\longleftarrow r_{E_n}(x_{E_n})\cdot\frac{p_{E_n}(x_{E_n})}{p^n{}_{/E_n}(x_{E_n})}$$

とし，他の辺 E では r_E は変更しないとすればよい．表 $\{p_E(x_E)\},\{r_E(x_E)\}$ は4項目を含むにすぎず，E は $\binom{N}{2}$ 個あるため記憶量は $O(N^2)$ である．

(1)より，ただちにつぎの式を得る：

任意の $p\in\Pi_0$ について

(2) $D(p,p^n)-D(p,p^{n+1})=D(p_{E_n},p^n{}_{/E_n})$.

したがって右辺分 (≥ 0) だけ，p^n は Π_0 に近づき p^{n+1} となることがわかり，$D(p,p^n)$ は単調減少し，ある値(c_p とおく)に収束する．

1節の例で，$p^0(x)=1/8$ としたときの，$H(p^n)$ と $D(p^*,p^n)$ の変化の様子が図7に示してある．

6. 収束性の証明

$p^n\to p^*$ $(n\to\infty)$ を示そう．Π は閉じた単体を形成しているから，$\{p^n\}$ から収束部分列 $\{p^{n_\nu}\}$ がとれ，$p^{n_\nu}\to\bar{p}$ $(\nu\to\infty)$ とする．(2)で，n を n_ν でおきかえ，$\nu\to\infty$ とすると，左辺は $\to 0$．ところ

が任意のEについて右辺$\geqq D(p_E, p^{n_\nu/E})$. したがって$\bar{p}_{/E}=p_E$. ゆえに$\bar{p}\in\Pi_0$.

(2)は任意の$p\in\Pi_0$について成立し，その右辺はpに依存しないので，

$$D(p^*,p^n)-D(p^*,p^{n+1})$$
$$=D(\bar{p},p^n)-D(\bar{p},p^{n+1}).$$

したがって，

$$D(p^*,p^0)-D(p^*,p^{n_\nu})$$
$$=D(\bar{p},p^0)-D(\bar{p},p^{n_\nu}).$$

ここで$\nu\to\infty$とすると，$D(p^*,p^0)-c_p*=D(\bar{p},p^0)$
前節のp^0に関する注意より，

$$D(p^*,p^0)=c_0-H(p^*),\ D(\bar{p},p^0)=c_0-H(\bar{p})$$

であるから，

$$H(\bar{p})-H(p^*)=c_p*$$

ところが，$H(p^*)$は最大エントロピーであるから$H(p^*)\geqq H(\bar{p})$. ゆえに$c_p*=0$.
これより$D(p^*,p^n)\to 0$，したがって$p^n\to p^*$.

7. おわりに

以上において，x_iは0，1しかとらないとしたが，一般の多値としても本質的変更はいらない．また，Eとしてあらゆる$\{i,j\}\subset\{1,\cdots,N\}$を考え，完全グラフをみてきたが，この仮定も必要なく一般の単純グラフを考えればよい．さらにEとして$\{1,\cdots,N\}$の任意の部分集合をとっても（このとき既知のp_Eは$|E|$次元分布となる），5，6節は記憶量の点を除けばそのままでよい．構造的にはハイパーグラフ的扱いになると思われるが皮相の見を越えていない．

付記——divergence について

$D(p,q)$は，情報理論的議論にしばしば現われて特異な役割をもつ量であり，informational divergence, discrimination information などとよばれることがある．pを《基準》として，それから，qがどの程度divergeしているか，違いがあるか，離れているか，どの程度判別しうるか，などをあらわす量である．この場を借りて，このような意味を裏づけるような事実を2，3取り上げて略述しておくのが適当と思われる．

1) $D(p,q)$はqについて凸関数である：

$$D(p,(1-\lambda)q+\lambda q')\leqq(1-\lambda)D(p,q)+\lambda D(p,q')$$

［これは$-\log u$の凸性より］．とくに$q'=p$ととると$D(p,\lambda p+(1-\lambda)q)\leqq(1-\lambda)D(p,q)$．ここで，$\lambda$を1に近づけると分布$\lambda p+(1-\lambda)q$は$p$にだんだん《似てくる》が，このことが$D(p,\lambda p+(1-\lambda)q)$が0に近くなることであらわされている．

2) variation $V(p,q)=\sum_x|p(x)-q(x)|$との関係は

$$D(p,q)\geqq\frac{1}{2}V^2(p,q)$$

（もっとシャープな評価については［5］）．

3) 分布pをもつ情報源を2元符号化する場合，各xに約$-\log p(x)$ビットを使えば，平均ほぼ$H(p)$ビットの《コンパクト》な符号化ができる．もし，pが不明でその推定分布qに対して同様に符号化を行なうと，平均$\sum_x p(x)(-\log q(x))$ビットになる．両者の差は$\sum_x p(x)(-\log q(x))-H(p)=D(p,q)$である．

4) いま分布pまたは分布qにしたがって独立な標本が次々と得られるとする．決定者はこの標本が，pにしたがって出現するのか，qにしたがうのかを，標本の《出かた》を観測しながら決定しなければならない．逐次決定理論によると，決定を誤まる確率を許容値に固定しておけば，pが真のとき，決定に至るまでに要する標本の大きさの期待値と$D(p,q)$との積は大体一定している．

参考文献

［1］ Brown, D. T. : A Note on Approximations of Discrete Probability Distributions, *Information and Control* **2**(1959), 386-392.

［2］ Forney, G. D. : Information Theory, Dept. Electrical Eng., Stanford Univ. Course Notes. EE 376, Winter 1972.

［3］ 国沢清典：エントロピー・モデル，日科技連出版社，1975.

［4］ Kruskal, J. B. Jr. : On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem, *Proc. Amer. Math. Soc.* **7**(1956), 48-50.

［5］ Toussaint, G. T. : Sharper Lower Bounds for Discrimination in Terms of Variations, *IEEE Trans. Information Theory*, **IT-21**(1975). 99-100.

（ほりべ・やすいち　静岡大学）