数理解析研究所講究録
第269巻 1976年 83-97

統計モデルと情報量

統計数理研究所　　赤池弘次

1. 何故統計的な見方が必要か

1.1 統計学は近似の学問である

統計学は近似の技術に関する学問であるという見方がある。与えられた問題に対し、限られたデータによって与えられる情報をもとに、適当な規準について最良の解を求めることがその主な課題であるという意味からである。

1.2 数値解析は近似の技術を取扱う

一方、数値解析も近似の理論そのものであると見ることができる。理論的な結果に対応する数値を求めることは、いわばひとつの観測法を与えるものである。観測には観測手段の分解能の問題、あるいは雑音の問題、がつきものである。したがってここには観測あるいは測定の精度に関する本質的に統計的な問題があるはずである。

2. 情報抽出の手段としての統計モデル

2.1 情報の有限性

数学的な表現においては無限の情報を含んでいるはずの対象の動きも，現実の観測の過程を通じて見るときは実質的に有限個のデータに対応する情報しか与えることができない。たとえば，連続的に変動する現象についても有限個の時点での観測値しか得られず，観測の精度も量子化誤差によっておさえられてしまうわけである。

2.2 環境の記述

制御しきれない誤差が存在することを知りながら，なおかつ最も妥当とみなされる決定を下さなければならないことは，すべての現実的な問題の処理に共通の現象である。工学あるいは技術に関連するほとんどすべての問題にこの現象はつきものである。数値解析においても，実際の数を扱うことになるとここに技術的な問題が発生する。制御しきれない誤差の発生する環境を適切にとらえ，表現し，共通な言語によって共通の経験化をはかることが，技術の組織的な発展のために極めて必要なこととなる。統計的な環境についてはその統計的な表現が必要である。適切な表現は（平均的に）成功にみちびき，不適切な表現は全く実用性の無い結果を与える。既得の知識と，データの与える情報とを有効に結合させ，データの与える情報を適切に抽出する手がかりを与えるのが

統計モデルである。

3. 統計的モデルの評価規準としてのエントロピー

3.1 情報と予測

情報とは通常何らかの新しい知識を与えるものと理解される。情報によって，これまで漠然としかわからなかったものが，より明らかに見てとられるようになると考えられる。この漠然さは，そのときの知識にもとづいて何等かの判断をしようとする場合の不確かさであるといえる。こうして，将来のできごとに関する予測の不確からしさの減少に役立つものとして情報を捉えることの自然さを見ることができる。将来の観測値の現われ方をその統計的な分布によって表現する場合，情報はこの分布を可能な限りの精度で知るために利用されるべきであると考えられる。統計的な現象について，データから得られる情報は，将来の観測値の分布を推定するために利用され，推定法の良否は，推定された分布による真の分布の近似の平均的な良否によって判定されると見る立場がここに成立する。

3.2 分布の類似度としてのエントロピー

ある確率分布が密度関数 $g(x)$ を持つ場合，この分布の確率論的な意味でのエントロピーは一般に

$$-\int g(x) \log g(x)\, dx \qquad (3.1.1)$$

によって定義される。ここで dx を $f(x)dx$ とおきかえると $g(x)$ は $g(x)/f(x)$ とおきかえられ

$$-\int \frac{g(x)}{f(x)} \log\left(\frac{g(x)}{f(x)}\right) f(x)\, dx \qquad (3.1.2)$$

が得られる。これを $g(x)$（によって定義される分布）の $f(x)$ に関するエントロピーと呼ぶことにする。この量は $f(x)$ によって与えられる分布に従う変量を多数回独立に観測した場合，その観測値の標本としての分布が $g(x)$ に近いものとなる確率の対数に比例しているものとみなすことができる。おおまかにいえば，$f(x)$ に従う変量の観測結果として $g(x)$ という分布が得られる確率の対数（の常数倍されたもの）を与えているものといえる。Boltzmann によってはじめて与えられた熱力学的エントロピーの確率論的解釈は，まさにこのようなものであった。[1] この解釈に従うとき，(3.1.2) は $f(x)$ による $g(x)$ の近似の程度を測る量として極めて自然なものであることがわかる。(3.1.2) の符号を反転したもの，いわゆるネゲントロピー，は Kullback の情報量として知られている。[2] この量は $g(x)$ と $f(x)$ の与える分布が一致しない限り常に正である。(3.1.2) の定義によるエントロピーは，したがって常に

0または負の値を取る。

3.3 最尤法とエントロピー

Fisherによって導入された尤度（likelihood）の概念は，統計学におけるもっとも基本的な概念のひとつである。$f(x)$がパラメータθによって規定され，$f(x|\theta)$と表わされる場合を考える。この分布から独立な観測値$x_1, x_2, \cdots, x_N$が得られたとき，θの尤度は

$$L(\theta) = \prod_{i=1}^{N} f(x_i|\theta) \qquad (3.3.1)$$

によって定義される。このとき$x_1, x_2, \cdots, x_N$は固定され，$L(\theta)$はθの関数とみなされる。観測値$x_1, x_2, \cdots, x_N$にもとづいて"真のθ"を推定する量として，$L(\theta)$を最大にするようなθの値をとることにするのが最尤法（the method of maximum likelihood）の原理である。この原理は，"はっきりした最適化に対する考慮にもとづいていないにもかかわらず，種々の問題で満足すべき方法に導くことに成功した"[3]とみなされてきた。

ここで対数尤度$\log L(\theta)$を考えると

$$\frac{1}{N}\log L(\theta) = \frac{1}{N}\sum_{i=1}^{N}\log f(x_i|\theta) \qquad (3.3.2)$$

となることから，Nが無限に大となるとき，(3.3.2)の右辺が

$$E \log f(X|\theta) = \int g(x) \log f(x|\theta)\, dx$$

に近づくことが期待される。$g(x)$は真の分布を与えるものである。$E \log g(X) - E \log f(X|\theta)$ が Kullback の情報量として $f(x|\theta)$ による $g(x)$ の近似の程度を表わすものであったから，$E \log f(X|\theta)$ が大なる程，$f(x|\theta)$ は $g(x)$ に良く類似していることになる。あるいは $f(x|\theta)$ から見た $g(x)$ のエントロピーが大となると表現しても良い。こうして Fisher の導入した最尤法の原理は，エントロピーの推定量として，$E \log f(X|\theta)$ の代りに $(1/N)\Sigma \log f(x_i|\theta)$ を用い，知ることのできない $E \log g(x)$ の部分は無視して，このエントロピーの推定量に関して最良と考えられる θ の値 $\hat{\theta}$ を"真の θ"の推定量として採用することを提案したものとみなすことができる。

3.4 最尤法の限界

前節に述べた立場から最尤法を見る場合，その限界がたちまち明らかなものとなる。最尤法の有効性は，まったく $(1/N)\Sigma \log f(x_i|\theta)$ の $E \log f(X|\theta)$ の推定量としての有効性にかかっている。また $f(x|\theta)$ はわれわれが勝手にえらんだものであって，これが真の $g(x)$ の良い近似を与えることのできるものであるか否かについて別に十分な検討が必要で

ある。これは統計的方法適用上の難点であると同時に，また積極的な既得の知識の利用の可能性を示している。問題あるいは対象に関する知識を，$f(x|\theta)$ の選択に最大限に利用することが必要なわけである。

今 N 箇の独立な観測値 (y_i, x_i) $(i=1, 2, \cdots, N)$ が得られたとする。y_i と x_i との間に多項式によって表わされる関係

$$y_i = a_0 + a_1 x_i + a_2 x_i^2 + \cdots + a_{p-1} x_i^{p-1} + z_i \qquad (3.4.1)$$

があるものとする。ここに z_i は，x_i とは独立に平均0分散1の正規分布 $N(0,1)$ に従っているものとする。このとき $\theta = (a_0, a_1, \cdots a_{p-1})$ の対数尤度は

$$-\frac{N}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{N}\{y_i - (a_0 + a_1 x_i + \cdots + a_{p-1} x_i^{p-1})\}^2 \qquad (3.4.2)$$

によって与えられる。これを最大にする θ は最小2乗法によって求められる。今ここで（3.4.1）のようなモデルを認めることとして，p の値が知られていない場合を考える。このとき p は0とある一定の値 g との間にあることは明らかなものとする。そこで p の値を0からはじめて次第に増加させながら g まで変化させ，そのそれぞれの p の値について（3.4.2）を最大にし，その値を $\log L(\hat{\theta}_p)$ と表わすことにする。このとき $L(\hat{\theta}_p)$ が p の増大とともに単調に増大することは明ら

かである。すなわち次数の高いモデルが常により高い尤度を示す。最大尤度の原理に従うとすれば常に最高次のモデルが採用されることとなる。この結果が，通常の多項式のあてはめに際して期待される結果でないことはほとんど明らかである。過度に高次のモデルをあてはめることは，極めて"信頼性の低い"，すなわち再現性の無い結果を与えることは経験的に知られている。

統計的方法によって情報を有効な形にとり出すために，統計的モデルが利用される。モデル中の自由なパラメータの数と，データの数との比が，"信頼性のある"結果を与えるか否かを決定する。対数尤度がエントロピーの推定量を与えるものとして利用される場合，この推定量の"精度"が，モデル中のパラメータの数によって影響される。この点を無視した最大尤度原理の適用が上述の最高次モデルを最良とする結果に導いたわけである。

3.5 AIC によるモデルの選択[4,5]

$f(x|\theta)$ において，$\theta = (\theta_1, \theta_2, \cdots \theta_L)'$ とする。$g(x) = f(x|\theta_0)$ であって，$\theta_0 = (\underline{\theta_1}, \underline{\theta_2}, \cdots, \underline{\theta_p}, 0, 0, \cdots, 0)$ であるとする。このとき独立な観測値 $x_1, x_2 \cdots, x_N$ にもとづく θ の対数尤度は

$$\sum_{i=1}^{N} \log f(x_i|\theta)$$

によって与えられる。とくに $\theta_{k+1}=\theta_{k+2}=\cdots=\theta_L=0$ と制約されたθの中で最大尤度を与えるものを ${}_k\hat{\theta}$ で示すことにする。このときkベクトルでは ${}_k\hat{\theta}$ を用いて定義される分布 $f(x|{}_k\hat{\theta})$ について，漸近的に

$$E\log f(X|{}_k\hat{\theta}) = E\log f(X|\theta_0) - \frac{k}{2N}$$

という関係が成立する。一方対数尤度については，$2\sum_{i=1}^{N}\log f(x_i|{}_k\hat{\theta})$ が $2NE\log f(X|\theta_0)$ の推定量として漸近的にkだけ過大な値を与える。これらの結果を総合すると $f(x|{}_k\hat{\theta})$ の $f(x|\theta_0)$ に対する平均的な近似度を規定する $E\log f(X|{}_k\hat{\theta})$（Eは X と ${}_k\hat{\theta}$ の両方に関する平均を示している）の推定量として，$(1/N)\{\Sigma\log f(x_i|{}_k\hat{\theta})-k\}$ を用いることが妥当であることがわかる。統計学において良く用いられている対数尤度比検定統計量との対比から，モデル $f(x|{}_k\hat{\theta})$ の良さの評価規準として

$$AIC = -2\sum_{i=1}^{N}\log f(x_i|{}_k\hat{\theta}) + 2k$$

が導入される。言葉で表現すれば

AIC = (−2)(maximum log likelihood) + 2(number of free parameters)

となる。この値が小さいモデルが良いモデルとみなされるわけである。もしk<pの場合には，Nの増大にともないAICは無限に増大する。実用上，kの増大にともなう平均的な誤差

の増大と，エントロピーの意味で評価されるモデルの近似度の向上とが釣合うようなkでAICが最小となることが期待される。このAIC最小化によって決定されるモデルを最小AIC推定（minimum AIC estimate，略してMAICE）と呼ぶことにする。

MAICEの統計的特性はまだ極めておおまかにしか知られていない。上述のようなモデル選択の場合が最もAICの微細な動きに依存する場合で，異った形の $f(x|\theta)$ 間の比較の場合などは，通常その何れかがきわだって悪い結果を与える場合が多く，このときには判断は極めて容易なものとなる。kの増加にともなうAICの変化がAICの最小値を与えるkで明瞭な下降から上昇への変化を示さないような場合が，MAICE適用上の困難がある場合である。

4. 事前情報の利用

4.1 MAICEの限界とJames-Stein推定量

次のようなモデルを考える。

$$y(n) = a_0 x_0(n) + a_1 x_1(n) + \cdots + a_{2L} x_{2L}(n) + z(n) \quad n = 1, 2, \cdots, N,$$

ただし $x_0(n) = \frac{1}{\sqrt{N}}$, $x_{2k-1}(n) = \sqrt{\frac{2}{N}} \cos\left(2\pi \frac{k}{N} n\right)$, $x_{2k}(n) = \sqrt{\frac{2}{N}} \sin\left(2\pi \frac{k}{N} n\right)$

とし，$z(n)$ は互に独立に平均0分散1の正規分布に従うものとする。a_m の最尤推定値 $\hat{a}_m$ は

$$\hat{a}_m = \sum_{n=1}^{N} y(n)\, x_m(n) \qquad m = 0, 1, \cdots, 2L$$

によって与えられ，モデルが正しい場合平均 a_m，分散1の正規分布に従い互に独立となる。今，$y(n)$ は十分良い近似式のあてはめの残差分である場合を考える。このような場合には，a_m がすべて小さな値を取るであろう。Lがあらかじめ十分大きく取られていたとし，$a_{2k+1}=a_{2k+2}=\cdots=a_{2L}=0$ と想定してMAICEを利用することを考える。もしすべての a_m が1ないし2程度の大きさを持っていたとすると，この場合平均的には常に最高次数のモデルを採用することがエントロピーの意味で最良の結果を与えることがわかる。したがってMAICEの適用は損失を招くだけである。このような場合，James-Steinの推定量

$$\tilde{a}_m=\left\{1-\frac{(2L-1)}{\sum_{m=0}^{2L}\hat{a}_m^2}\right\}\hat{a}_m \qquad m=0,1,\cdots,2L$$

が，$\hat{a}_m\ (m=0,1,\cdots,2L)$ より良い（$E\Sigma[\tilde{a}_m-a_m]^2$ が $E\Sigma[\hat{a}_m-a_m]^2$ より小さい）ことが知られている。

この場合 $\hat{a}_m\ (m=0,1,\cdots,2L)$ の分布に注目し，その真の分布（平均 a_m，分散1の互に独立な正規分布）の，$\hat{a}_m$ を平均とし，分散1で互に独立な正規分布に関するエントロピーを求めると $-(\frac{1}{2})\Sigma(\hat{a}_m-a_m)^2$ となる。$\hat{a}_m$ の真の分布の，平均 θ_m，分散1，の互に独立な正規分布 $(m=0,1,\cdots,2L)$ に関するエントロピーは $\{(2L+1)/2\}[1-\{1/(2L+1)\}E\Sigma(\hat{a}_m-\theta_m)^2]$ によって与えられる。ここでEを取外したものをこのエントロピーの

推定値とみなすことにすると

$$B=\{(2L+1)/2\}\left[1-\{1/(2L+1)\}\sum_{m=0}^{2L}(\hat{a}_m-\theta_m)^2\right]$$

が得られる。この推定エントロピーBを最大にするようなθ_mを求めることは，対数尤度$-(\frac{1}{2})\Sigma\{(\hat{a}_m-\theta_m)^2+\log 2\pi\}$を最大にすることと同等で$\hat{\theta}_m=\hat{a}_m$が得られる。Bにおいて$\theta_m=\hat{a}_m$とおいてみるとB=(2L+1)/2が得られる。これは明らかにエントロピー（負の値をとる）の推定量としては無意味な量で，T度本来のエントロピー$-(\frac{1}{2})\Sigma(\hat{a}_m-a_m)^2$の平均値の符号を反転したものに一致する。この結果はAICの議論の場合と同様，エントロピーの推定量としての対数尤度の精度の限界を示すものである。この場合$E\Sigma(\rho\hat{a}_m-a_m)^2$が小さな値となるような$\rho$を$\hat{a}_m$ $(m=0,1,\cdots,2L)$の関数として求めようとすると，最終的にΣa_m^2の良い推定量を求めることが問題となる。この1箇のパラメータを決定するために，$\hat{a}_m$ $(m=0,1,\cdots,2L)$の対数尤度が利用される。この対数尤度が，このパラメータによって規定される分布に関するエントロピーの推定量として有効なものであれば，ふたたび最尤法の原理が有効に適用できるわけである。ここに統計的モデルが，エントロピーの推定量としての対数尤度を媒介として，データに含まれる情報の抽出を可能ならしめる機構が明らかに見られる。このよ

うにして得られるSは，James-Steinの推定量に類似の結果を与えることが知られている。[5] この結果は，James-Stein推定量の有効性の根拠に対するひとつの解釈を与えるもの，とみなすことができる。

4.2 モデル設定の重要性

James-Steinの推定量は，すべてのa_mが0に近いときに有効に働く。MAICEは，いくつかのa_mが著しく0と異なり，他のa_mがその推定量$\hat{a}_m$の誤差に比して0に近いような場合に有効に働く。それぞれの方法の特徴が生かされるような場面に適用されることが必要である。これには問題の理論的な解析結果にもとづく知識の適切な利用が要求されるわけで，ここにいわゆる事前情報の利用の問題がある。既得の知識を最も有効に利用するように統計的モデルを設定することによってはじめて，データに含まれる情報を最も有効に抽出することができるわけである。数値解析に関連する実例については，田辺国士氏の報告を参照されたい。

参考文献

1) Boltzmann, L. (1877). Uber die Beziehung zwischen dem zweiten Hauptzatz der mechanischen Warmetheorie und der Wahrscheinlichkeitsrechnung respective dem Satzen uber das Warmegleichgewicht. Wiener Berichte, 76, 373-435.

2) Kullback, S. (1959). Information and Statistics, New York : Wiley.

3) Lehmann, E. L. (1959). Testing Statistical Hypotheses, New York : Wiley.

4) Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In 2nd International Symposium on Information Theory (B. N. Petrov and F. Csaki, eds), PP. 267-281. Budapest : Akademiai Kiado.

5) Akaike, H. (1974). A new look at the statistical model identification. IEEE Trans. Automat. Contr., AC-19, 716-723.

6) Akaike, H. (1975). An extension of the method of maximum likelihood and the Stein's problem. Research Memo. No. 84, The Institute of Statistical Mathematics.