発 表 要 旨

AIC とエントロピー最大化

統計数理研究所　赤　池　弘　次

は じ め に

統計的推論は統計的方法によって証拠から結論を導き出すものであり，統計的方法はデータから統計的な結論を生み出すためのプログラムと考えることができる．

計算機のプログラムの場合，良いプログラムはしばしば組織的であり，自然な形を持ち，単純である．エントロピー最大化原理というのは，統計的方法をひとつのプログラムとみなした場合，良いプログラムを作り出すためのひとつの原理あるいは原則である。

計算プログラムの作製経験者なら誰でも知っているように，最良のプログラムというものは通常存在しない。どんなに良く出来ているように見えるプログラムでも常に改良の可能性が残されている．プログラムとしての統計的方法あるいは手順の作製のためのエントロピー最大化原理も，より良い方法を生み出して行くための一般的な指針を与えるものであり，最良のものを一意的に定めるというような狭い性格の原理ではない．

エントロピー最大化原理の定義は後で与えるが，このシンポジウムの報告に見られるように，さまざまな領域での研究において自由に新しいモデルの開発を可能にしているという事実によって，この原理の有効性の証明が与えられるのである．

FPE から AIC へ

エントロピー最大化原理の展開を歴史的に眺めると，その出発点は1969年に時系列の自己回帰モデルの次数決定のために導入された最終予測誤差（Final Prediction Error ; FPE）にあることが分かる．

時系列データ $x_1, x_2, \cdots, x_N$ にもとづいて自己回帰モデル

$$x_n = a_0 + a_1 x_{n-1} + a_2 x_{n-2} + \cdots + a_M x_{n-M} + w_n$$

を決定しようとするとき，次数 M の決定が大きな問題となる．FPE は，モデルにもとづいて推定された係数を実際予測に用いた場合，期待される予測誤差の2乗の平均的な大きさの推定値として定義される．その性質の解析を通じて，必要以上に次数 M を増大させることは FPE の増大を招くことが明らかとなった．

もとの時系列データ x_n がベクトルである場合に対して FPE を拡張しようとすると種々の問題が生じる．まず第1に予測誤差の平均的な大きさを表現するひとつの数値として何を取ればよいかは自明ではない．多変量自己回帰モデルは，統計的最適制御系の設計等に多くの応用を見出し，このモデルの利用のためのプログラムパッケージ TIMSAC においては，MFPE, FPEC 等の統計量が次数決定のために利用されている．これらの統計量は，時系列にガウスモデルを想定し最尤法を適用する，ということを暗暗裡に認めて考案されたものである．

AIC の展開には 2 つの基本的な視点が要求される．1 つは FPE に見られる予測の視点で，統計的方法の効果を予測の立場から評価しようというものである．もう 1 つは FPE には見られなかったもので，予測は確率分布の形で与えられるものと考え，この予測の誤差の大きさを確率的エントロピーの概念を用いて測ろうとするものである．この予測とエントロピーという 2 つの視点の交叉する点に AIC が登場するわけである．

将来の観測値 y に対して，その分布を現在得られているデータ x の関数として $p(y\mid x)$ のように与え，これを予測分布と呼ぶことにしよう．統計的方法は，この場合 x から $p(y\mid x)$ を生み出すプログラムとして与えられる．例として y の分布が $p(y\mid\theta)$ の形を取る場合を考えると，データ x による θ の推定値を $\theta(x)$ とすれば，$p(y\mid x)=p(y\mid\theta(x))$ によって予測分布が定義される．更に一般化して，x の関数として θ 上の分布，$p(\theta\mid x)$ を与えれば，$p(y\mid x)=\int p(y\mid\theta)p(\theta\mid x)d\theta$ によって予測分布が定義される．後者の方がより一般的なプログラムであり，これはベイズ的方法を含むものとなる．

確率的エントロピーの概念は，もともと L. ボルツマンによって展開されたものであるが，その特徴はかならず 2 つの分布の間の関係を評価する量という形をとることである．

分布 $f(x)$ と $g(x)$ とがある場合，分布 $f(x)$ の分布 $g(x)$ に関するエントロピーは，

$$B(f\,;g)=-\int\frac{f(x)}{g(x)}\log\left(\frac{f(x)}{g(x)}\right)g(x)dx$$

によって与えられる．この量は統計学において基本的な役割を果すものであり，歴史的に見ても，K. ピアソンのカイ 2 乗統計量をはじめとして，常に最も成功的な結果に関連して登場することが認められる．またボルツマンの熱力学的エントロピーに関する研究成果は，M. プランクの量子論，A. アインシュタインの光量子の理論等の先がけをなすものであり，その研究の終局的な結果としてこの $B(f\,;g)$ が得られている．

エントロピー最大化原理は，この $B(f\,;g)$ と予測分布 $p(y\mid x)$ とを用いれば，$f(y)$ を y の真の分布として

"$\mathrm{E}_xB(f\,;p(\cdot\mid x))$ が大となるように $p(y\mid x)$ を作るプログラムを発展させること"

と表現される．ここで E_x はデータ x の分布に関する期待値を示す．実用上は $f(y)$ も E_x の構造も未知であるから，この原理を具体化するためには $\mathrm{E}_xB(f\,;\,p(\cdot\mid x))$ の測定をどうするかという問題を解かねばならない．

エントロピーの測定値としての対数尤度

パラメータ θ を持つ分布族 $\{g(\cdot\mid\theta)\}$ を考える．この分布族の中で，真の分布 $f(\cdot)$ に一番近いものを求めるためには

$$\operatorname*{Max}_{\theta} B(f\,;\,g(\cdot\mid\theta))$$

を考えればよい．いま E_x によって真の分布 $f(x)$ による平均を表わすものとすれば，

$$B(f\,;\,g(\cdot\mid\theta))=\mathrm{E}_x\log g(x\mid\theta)-\mathrm{E}_x\log f(x)$$

であるから，上記の探索は，

$$\operatorname*{Max}_{\theta}\mathrm{E}_x\log g(x\mid\theta)$$

で置き換えられることが分かる．

データ x が与えられた場合, $E_x \log g(x \mid \theta)$ の推定値として対数尤度 $\log g(x \mid \theta)$ を用いることにすれば, 最適な近似の探索は

$$\operatorname*{Max}_{\theta} \log g(x \mid \theta)$$

に置きかえられる. これが最尤法である. この考察によれば, $\log g(x \mid \theta)$ が $E_x \log g(x \mid \theta)$ の推定値として良いものであれば, 最尤法によって θ の良い推定値が得られることが分かる. R. A. フィッシャーは, x が同一の分布に従う独立な観測値の列 x_1, x_2, …, x_N で与えられる場合に対し, 真の分布 $f(x)$ が $\theta=\theta_0$ に対応する $g(x \mid \theta)$ によって与えられるものとして最尤法のすぐれた点を明らかにした.

一般に $g(\cdot \mid \theta)$ で与えられるモデルは常に真でないものと考えられるから, フィッシャーの議論では実用の場における最尤法の有効性の保証は得られない. エントロピーの視点を採用することによってはじめて, フィッシャーの推定論の局所性を超える発想が可能となるのである. 定義により AIC は

$$\mathrm{AIC}=(-2)\log \max_{\theta} g(x \mid \theta)+2\dim(\theta)$$

の形に与えられる. これは $p(y \mid x)=g(y \mid \theta(x))$ ($\theta(x)$ は θ の最尤推定値) とした場合の $(-2)E_x B(f\,;\,p(\cdot \mid x))$ の推定値となっている. したがって, エントロピー最大化＝AIC 最小化となる.

AIC の導入にはフィッシャーの漸近的な理論に相当する結果が想定されていたのであるが, エントロピー概念の明確な認識を通じてエントロピー最大化というより積極的な統計モデル評価の視点に導かれたわけである.

モデルの評価を平均対数尤度で行なうこととすれば, 真の分布にもとづく E_x の内容の如何にかかわらず, 対数尤度のより大きなモデルを選ぶという立場の合理性が認められることになる. この結果は, 各人の主観的な立場の差にもとづく E_x の相異の存在にも拘らず受け入れられるものであり, これが対数尤度に "客観性" を与えるのである.

確率分布の構成は常に主観的なものであり, その構造は利用上の目的と利用可能な情報とに依存して決定される. その意味で確率分布の主観から切り離された客観性を追求することは無意味である. しかし多くのモデルを想定してその対数尤度を比較検討した結果は, 他の人々に対して有効な情報となり得る. 客観性の概念はこの情報有効性の概念によって置き換えられねばならないわけである.

ベイズモデルの実用化

エントロピー最大化原理にもとづくベイズモデルの実用化は, この原理がわれわれを従来の統計学における推定論, 検定論の固定した図式による呪縛から解放するものであることを明らかにする.

フィッシャーによる推測確率 (fiducial probability) の理論は, ベイズ的方法を如何にして客観的なものとするかの試みである. 残念ながらその結果は失敗としか言いようがない. ネイマン・ピアソンの仮説検定の理論は, 第1種の過誤, 第2種の過誤の導入により, 事前確率の考えによらない "客観的" な方法を展開したものである. しかしながら, フィッシャーが繰り返し主張しているように, 一定の危険率の下での仮説の棄却を論ずるという形での検定の定式化は, 科学的研究の場における仮説の検討にはそぐわないものであることが明らかで, その意味では成功していない.

これらのすぐれた統計学者はベイズ的方法の良さを十分認識していた．しかしその実用化には成功しなかったのである．

この問題に対して，エントロピー最大化原理は容易に解答を与える．客観性あるいは情報有効性の獲得は，多くのモデルの比較結果を通じて実現される．ここではそれぞれのベイズモデルの対数尤度あるいはエントロピーの推定値が，客観性あるモデル比較上の情報を提供することになる．

ベイズモデルに不確定なパラメータが存在する場合にこれを最尤法によって決定することにすれば，

$$\text{ABIC} = (-2)\log \max \text{likelihood} + 2(\text{number of parameters})$$

によって情報量規準を定義することができる．ただし，ここに likelihood はベイズ模型の尤度で

$$\int g(x \mid \theta,\ S_1)p(\theta \mid S_2)d\theta$$

によって定義される．$g(x \mid \theta, S_1)$ はデータ分布，$p(\theta \mid S_2)$ は事前分布，S_1, S_2 は未知パラメータである．

ABIC の実用上の有効性は，このシンポジウムで報告される多くの例がこれを利用して満足すべき結果を与えていることから明らかである．ベイズ的方法の実用化という従来の統計学の枠組では解決し得なかった問題に対して組織的な接近を可能にするというこの事実は，エントロピー最大化原理が従来の統計学の枠組みを超える新しい視点を与えていることの証明である．

統計的モデルの幾何学

東京大学工学部　甘　利　俊　一

統計学は，確率的な機構にもとづいて発生したとみなせる観測データをもとに，もとの確率的な構造についての推論を行なう科学である．確率変数を x とし，その分布密度関数を $p(x)$ としよう．また，適当な正則条件を持つ密度関数全体の集合を S としよう．確率変数 x の k 個の実現値 $x_1, \cdots, x_k$ をもとにして，そのもとにある S の一つの要素を求めるのが推定の問題である．このとき，集合 S はどんな構造をしているのか，たとえば S の二つの要素 $p(x)$ と $q(x)$ の間には意味のある分離度が定義できるのかが問題となる．通常，集合 S の中から真の分布 p を探すには，S は広すぎてうまくいかない．このため，統計学者は統計的モデル M を考える．これは有限個のパラメータ $\theta = (\theta_1, \cdots, \theta_n)$ で指定される確率分布 $p(x, \theta)$ の集まりで，S の中に n 次元の部分空間として入っている．

M として何を採用するか，これはモデルを構成する問題であって，確率変数 x を生成する物理的機構など種々の先験的な知識に基づいて M を設定することになる．古典的な統計学はモデル M は正しい，すなわち真の分布 $p(x)$ は M の中に入っている，という仮定のもとで理論を展開する．しかし，通常 M は近似にすぎず，真の分布は M に入っていないことも多い．赤池の情報量基準は多数の合理的なモデルを考えて，どのモデルを選ぶべきかをデータから推論しようという，従来の枠を破る試みであった．モデル M を与えられたものとして考えるのではなく，このように一度対象化して考えるならば，統計的モデル M 自体の持つ性質を議論する必要