AICとMDLとBIC

赤池　弘次

1．はじめに

AICは統計モデルをデータに基づいて比較するための相対的な評価量である．その基礎にある情報量の概念は，ボルツマンによる熱力学的エントロピーの研究にはじめて登場したもので，ひとつの確率分布から見て，もうひとつの確率分布がどれほど離れているかを測るものである．AICの導入により，ただひとつのモデルについての推定論，検定論の議論に集中していた伝統的な統計学の枠組を超えて，さまざまな科学的研究の現場で，新しい統計モデルの提案と比較検討を通じて研究活動の推進が試みられるようになった．

AICに関する誤解は数多いが，J. RissanenによるMDL (Minimum Description Length) 規準と，G. Schwarzによるベイズ理論的な解析に基づくBIC規準に関するものが，最も一般的であろう．本稿では，MDLあるいはBICが，AICを超える根拠を持つと考えるのは迷信に過ぎないことを示し，AICの本来の意味を再確認することにしたい．

2．AICの定義

データから有効な情報を得るために，データxの現われ方に対するわれわれの期待を確率分布$p(x|a)$の形で表現する．これが統計モデルである．通常このモデルは未定の変数(未知パラメータ)aを含む．データxが与えられたとき，$p(x|a)$の値をパラメータa(の与えるモデル)の尤度 (likelihood) と呼ぶ．最尤法 (the method of maximum likelihood) は，aの推定値として尤度$p(x|a)$を最大にする値 (最尤推定値) を採用する方法である．

R. A. Fisherの研究により，観測データxが実際に$p(x|a)$の形の確率分布に従って発生するとき，最尤法が優れた特性を示すことが示された．しかし，応用の場面では，データを生み出す確率的な構造が完全に分かっていることは無いから，Fisherの議論は，最尤法の実上用の根拠を与えない．

AICの導入には，確率分布$g(x)$から見て確率分布$f(x)$がどれだけ離れているかを測るために，情報量

$$I(f;g)=E_f \log f(X)-E_f \log g(X)$$

を採用する．ただしE_fはXが分布$f(x)$に従うものとしての期待値を示す．この量は非負の値をとり，0となるのはgがfに一致する時である．$I(f;g)$が小さいほど$g(x)$から見て$f(x)$が近いことになる．

$f(x)$を目的の分布，すなわち“真”の分布と考えれば，$E_f \log g(X)$が大きいほど$g(x)$は$f(x)$の良い近似となる．$E_f \log g(X)$の推定値として，データxによる対数尤度$\log g(x)$を採用すれば，$\log g(x)$が大きいほど良いモデルとみなされる．$g(x|a)$のように未知パラメータを含むモデルでは，$\log g(x|a)$を最大にするパラメータ値が最良とみなされ，最尤法が得られる．通常，aはベクトルである．

最尤推定値の与えるモデルの尤度は，パラメータの値をデータで調節することから，モデルの評価値としては高めに偏る．この点を考慮し，最尤法で決められたモデルの相対的な評価量として

AIC＝(－2)最大対数尤度＋2(パラメータ数)

が定義される．パラメータ数とは，最尤法で調節されたパラメータの数(aの次元)である．符号の関係からAICが小さいほど良いモデルと評価される．

3．AICの構造

$L(a|x)=\log p(x|a)$とすると，パラメータaの与えるモデルの“真”の評価は$E_f L(a|X)$で与えられる．E_fは，Xの“真の”分布$f(x)$に関する期待値を示す．aの次元をKとし，パラメータaの空間をA_Kとする．$f(x)=p(x|a_t)$と書けるものとし，$\Delta a=a-a_t$として，a_tの近傍で2次曲面近似

$$E_f L(a|X)=E_f L(a_t|X)-(1/2)\Delta a' M \Delta a \quad (1)$$

を採用する．記号'は転置を示す．最尤推定値a_oの近傍

あかいけ　ひろつぐ
〒302 取手市取手1－7－14－204

で，$L(a|x)$に対して(1)と同形の近似 $L(a|x)=L(a_o|x)-(1/2)\Delta a'M\Delta a$ が成立する状況を想定する．ただし，$\Delta a=a-a_o$である．これらの近似が有効に成立する範囲内でのパラメータの動きを考え，A_kを，a_tを原点とし，内積が$(a,b)=a'Mb$で与えられるベクトル空間として考察を進める．

aをA_kのk次元部分空間A_kに制約する場合の最尤推定値をa_{ko}とし，その評価として

$$2E_f L(a_t|X)-2E_f L(a_{ko}|X)=\|a_t-a_{ko}\|^2$$
$$=\|d_{kt}\|^2+\|a_{kt}-a_{ko}\|^2 \quad (2)$$

を採用する．ただし，$d_{kt}=a_t-a_{kt}$，a_{kt}はA_k内で$E_f L(a|X)$を最大にする値，$\|a\|^2=(a,a)$である．対応する対数尤度による評価は，$a_t\to a_o\to a_{ko}$という経路に沿って眺めれば，

$$2L(a_t|x)-2L(a_{ko}|x)=\|a_t-a_{ko}\|^2$$
$$=2(L(a_t|x)-L(a_o|x))$$
$$+2(L(a_o|x)-L(a_{ko}|x))$$
$$=-\|a_t-a_o\|^2+\|a_o-a_{ko}\|^2 \quad (3)$$

となる．$\Delta a=a_o-a_t$を，A_kへの射影ΔbとA_kの$(K-k)$次元直交補空間Cへの射影Δcとに分解すると，$\Delta a=\Delta b+\Delta c$，$a_{kt}-a_{ko}=-\Delta b$，$a_o-a_{ko}=(a_o-a_t)+(a_t-a_{kt})+(a_{kt}-a_{ko})=\Delta a+d_{kt}-\Delta b=d_{kt}+\Delta c$が成立する．

データxが互いに独立に同一分布に従うn個の観測値からなる場合，nが大となるとき，$\|\Delta a\|^2$，$\|\Delta b\|^2$は，それぞれ，漸近的に自由度K，kのカイ2乗分布に従う．これらの分布の平均値はそれぞれ，K，kである．$\|a_o-a_{ko}\|^2=\|d_{kt}+\Delta c\|^2$は，パラメータが$Ak$に属するという仮説の尤度比検定統計量で，漸近的に$\|d_{kt}\|^2+K-k$を平均値とする非心カイ2乗分布に従う．

これらの結果から，式(3)の右辺第1項を$-K$で置き換え，第2項に$-(K-k)+k=-K+2k$を加えれば，漸近分布の平均値が式(2)のそれと一致する．かくして，

$$2(L(a_{ko})-L(a_{ko}))-2K+2k$$
$$=-\mathrm{AIC}(K)+\mathrm{AIC}(k)$$

が，最大対数尤度の偏りを漸近的に修正した，評価式(2)の推定値となる．ただし，

$$\mathrm{AIC}(k)=-2L(a_{ko}|x)+2k$$

である．$-2E_f L(a_{ko}|X)$の代わりに，観測可能なAIC(k)を用いて，偏りのない相対的評価が可能になるわけである．KおよびA_kは，背景にある理想的な分布に対応するものと考えれば，明示される必要はない．対数尤度関数が上記の近似を許容する限り，AICは適用可能であり，分布形の異なるモデルの比較にも有効である．

4．ベイズモデルについて

ベイズモデルでは，データ分布$p(x|a)$のパラメータaに対して，事前分布$p(a)$を想定し，パラメータの推定値に代わり

$$p(a|x)=p(x|a)p(a)/p(x)$$

で定義されるパラメータの分布（事後分布）を利用する．ただし，$p(x)$はこのベイズモデルの尤度で

$$p(x)=\int p(x|a)p(a)\,da \quad (4)$$

で与えられる．

このような方法の予測の立場からの評価は，事前分布の選択を含め，情報量を利用して議論することができる［1,5］．利用目的に応じて，有効な情報が取り出せるように$p(a)$を選ぶことが大切で，このためには，当面の問題に対する十分な理解と知識が要求される．いくつかのベイズモデルがある場合には，尤度を比較して検討を進める．これらを総合してベイズモデルを構成することも可能であり，事前分布の未知パラメータ（ハイパーパラメータ）に最尤法を適用することも可能である．

ベイズモデルでは，データ分布$p(x|a)$のパラメータaの次元を低く保つ必要はない．データxの次元（データ長）よりも高い次元のパラメータを持つデータ分布による時系列の季節調整の実現が，このようなモデルの実用性を実証した［2,3］．AICの示した対数尤度の客観性あるいは間主観的（intersubjective）な性格が拠り所となって，複雑なベイズモデルの組織的な取扱いが可能になったと筆者は考えている［4］．

5．MDL規準について

1972年1月，R. Kalmanに招かれてフロリダでのシステム理論のシンポジウムに参加した筆者は，自己回帰モデルの次数決定に関連してAICの簡単な解説を行った．参加者の中でこの話に最も興味を示したのがJ. Rissanenである．その後，符号化あるいは複雑度(complexity)の視点から，モデル評価の議論を試みているが，MDL関係の仕事をまとめた書物［6］によって，その考え方の大要を追ってみよう．

MDL規準は，

MDL(k)

$$=-\log(p(x|a_o)p(a_o))+(k/2)\log n \quad (5)$$

のように定義される．ただし，a_oは最尤推定値，kはパ

ラメータ a の次元，n はデータの個数，すなわち x の長さである．モデルに基づく符号化によるデータ x の符号の長さは，$L(x,a)=L(x|a)+L(a)$ によって与えられる．$L(x|a)=-\log p(x|a)$，$L(a)=-\log p(a)-\Sigma d_j$ である．d_j は，$p(a)$ をヒストグラム状に離散化するための，a の j 一成分の離散化（粗い数値化）の単位幅である．Σ は $j=1,\cdots,k$ に対する総和を表わす．符号長を最短にするモデルに対する $L(x,a)$ の値は，a に関する最小化により

$$\min_a\{-\log p(x|a)-\log p(a)-\Sigma\log d_j\}$$

で与えられる．この時の a の値（事後分布のモードにあたる）を a_p とする．

離散化した a の中で $L(x|a)$ を最小にするものを a_d とすれば，離散化の誤差のために

$$L(x|a_d)=L(x|a_p)+(1/2)e'Me$$

のように符号の長さが増大する．ただし，$e=a_d-a_p$ である．ここで，e を $d=(d_1,d_2,\cdots,d_k)$ で置き換え，

$$(1/2)d'Ld-\Sigma\log d_j$$

を最小にするように d_j を決める．$M=nS$ と置けば，最適幅が $d_j=c_jn^{-1/2}$ の形で与えられる．このとき

$$\Sigma\log d_j=(k/2)\log n+O(1)$$

が成立し，n とともに増大する部分だけに注目すると

$$L(x,a_d)=-\log p(x|a_p)+(k/2)\log n$$

となり，MDL 規準が得られる．$p(a)$ が a_p の近傍で平坦であると仮定すれば，$a_p=a_o$（最尤推定値）となるから，式(5)の形になる．ただし，ここでは $\log p(a_o)$ は無視されている．

上記の導出の過程から，MDL 規準は，パラメータを離散化して符号化することを考えた結果，最尤推定値の近傍での対数尤度関数の動きが評価されて現われたものに過ぎないことが分かる．計算機の桁数が十分あるときに，パラメータの値を粗雑に区切り直すことは，統計理論の立場からは無意味である．したがって，MDL の根拠は統計学的に見れば無意味である．

Rissanen は，最終的には stochastic complexity なる概念を提案しているが，これはベイズモデルの対数尤度の符号を変えたものにすぎない．一般的なベイスモデルの適用に際しては，データは n 次元空間からのサイズ1のサンプルと考えるべきであるから，$\log n$ は無意味となる．結局，MDL は統計学的には無意味な状態にとどまっている．

符号化理論は，本来離散的な符号の取り扱いのために考案された Shannon のエントロピーが基礎にあり，連続的な変量への適用には注意を要する．記号の系列を生み出すに必要な計算機入力の長さ，すなわち符号長，を用いて，系列の確率を定義することを初めて提案したのは R.J.Solomonoff である．1974 年に筆者がハーバード大学で行ったセミナーに出席して AIC に興味を示し，1976 年の再会の際には，自身の方法を多項式の次数決定に適用したことを話してくれたが，結果は公表されていない．連続変量の取り扱いに内在する困難を示唆するものであろう．

6．BIC 規準について

1979 年の夏学期，筆者はスタンフォード大学統計学科で情報量規準とベイズモデルの利用について講義を行った．聴講者の中にいた G. Schwarz は，やがてベイズ的な枠組みに基づくモデル選択の規準（BIC）を導出し，AIC に最適性はありえないとする論文草稿が筆者あてに送られてきた．

この論文が Annals of Statisitics (Vol.6, 1978) に掲載されるに際して，筆者は Editor に手紙を送り，BIC によるモデル選択は，恣意的な事前分布に対応するものであり，最適性の主張には問題があることを指摘した．しかし何らの対策もとられず，BIC は AIC を否定するものとの誤解が広まることとなった．

複数のベイズモデルがあるとき，j 番目のモデルの尤度を $p(x|j)$ と表示すれば，このモデルの事後確率は，モデルの事前確率を $p(j)$ として

$$p(j|x)=p(x|j)p(j)/\Sigma p(x|j)p(j)$$

で与えられる．ただし Σ は j についての総和を示す．

$$\log p(j|x)=\log p(x|j)+\log p(j)+H$$

と表示すれば，データ数 n が大となるにつれ，右辺第1項の対数尤度が第2項に比べて支配的になり，これがモデルの相対評価の規準となる．

データ分布 $p(x|a)$ と事前分布 $p(a)$ で定義されるベイズモデルの尤度は式(4)により

$$p(x)=\int p(x|a)p(a)\,da$$

で与えられる．Schwartz は特定のモデルについて，パラメータ a が k 次元空間に制約される形の事前分布に従うとき，その対数尤度が

$$\log p(x|a_o)-(k/2)\log n+R$$

の形になることを示した．R は n とともに増大することのない部分であるから，これを無視することにして残りを (-2) 倍すれば

$$\mathrm{BIC}=(-2)\text{最大対数尤度}+k\log n$$

が得られる．

この結果は，AIC の定義における $2k$ を $k\log n$ で

置き換えるものであり，したがってAICによるモデル選択に最適性はあり得ないと言うのである．

Schwarzの取り扱ったベイズ的な構造では，モデルの事前確率$p(j)$とともに，パラメータの事前分布$p(a|j)$もデータ数nに無関係に一定とされる．これは，事前情報の完全利用を目指すベイズ的接近の立場からは恣意的な制約である．データの観測以前にnが未確定であれば，事前分布としてa，j，nの分布を

$$p(a, j, n)=p(a|j, n)p(j|n)p(n)$$

のように考え，データが与えられたときには，nで条件づけられた$p(a|j, n)p(j|n)$を事前分布として用いる．このように，nに依存する事前分布を用いることが，ベイズ的な立場からは自然なのである．

簡単な例について具体的に検討してみよう．データ$x=(x_1, x_2, \cdots, x_n)$について，各$x_i$は互いに独立に平均$a$，分散1の正規分布に従うものとする．$X=(1/n)\Sigma x_i$とすれば，$x$のデータ分布は

$$p(x|a)=C(x)\exp(-(n/2)(X-a)^2)$$

の形に書ける．$p(a|0)$として，$a=0$に確率1を与える分布，$p(a|1)$として，平均0，分散$4/n$の正規分布を考える．データの算術平均Xは，平均a，分散$1/n$の正規分布に従うから，後者の標準偏差は，Xの標準偏差の2倍である．

これらふたつのベイズモデルについて，その尤度を計算してみると，それぞれ，$C(x)\exp(-nX^2/2)$，$C(x)(5^{-1/2})\exp(-nX^2/10)$となる．これより，$nX^2>(5/4)\log 5(=2.01)$であれば，$p(a|1)$の与えるモデルの事後確率が$a=0$とするモデルのそれよりも大きくなる．これは，AICによる$a\neq 0$の判定条件，$nX^2>2$と，ほとんど完全に一致する．このモデルをk次元ベクトル観測値に拡張すれば，AICの一般的な形に対応する結果が得られる．

BICによる$a\neq 0$の判定条件は，$nX^2>\log n$，となる．これを，仮説$a=0$の検定と考えると，$n=7$の場合，AICによる判定とほぼ同じく，Xの0からの偏差が標準偏差の1.4倍程度で有意と判定されるが，$n=1000$になると，標準偏差の2.5倍でも有意とされず，有意性の判断基準がnに依存して変化することになる．BICの利用を受け入れる人は，精度が同じデータでも，nによって見方を変える立場に立っているのである．同形のMDLをはじめ，いわゆるモデルオーダーの決定に一致性(consistency)を示すその他の規準の利用者も同じである．

BICは，パラメータの値に比べて最尤推定値の誤差幅が極度に小さく，有意なパラメータと，そうでないものとが，容易に識別できる状況に対応するモデルから得られている．これに対してAICは，有意性がようやく認められる程度のパラメータの取り扱いに注目し，誤差の影響に埋没しそうになるところまでモデル化の可能性を追及しているのである．

7．おわりに

AICの導入により，Fisherが統計理論の対象外とした分布形指定の問題が，モデルの比較検討を通じて組織的に処理されるようになった．歴史的に見ると，モデルの提案が統計的解析の中心課題であることを明らかにしたことが，AICの主な貢献である．MDLとBICを廻って生じた誤解は，数式の具体的意味の理解の欠如が招く危険を示す，教訓的な事例を提供している．

参考文献

[1] Akaike, H.：A new look at the Bayes procedure, *Biometrika*, Vol. 65, 53-59 (1978).

[2] Akaike, H.：Likelihoood and the Bayes Procedure, *Bayesian Statistics*, Bernardo, J. M., DeDroot, M. H., Lindley, D. V. and Smith, A. F. M.(eds.), University Press, Valencia, Spain, 143-166 (1980).

[3] Akaike, H.：Seasonal adjustment by a Bayesian modeling, *Time Series Analysis*, Vol. 1, 1-13 (1980).

[4] Akaike, H. Prediction and entropy, *A Celebration of Statistics*, Atkinson, A. C. and Fienberg, S.E. (eds.), Springer-Verlag, New York, 1-24 (1985).

[5] 赤池弘次：事前分布の選択とその応用，ベイズ統計学とその応用，鈴木雪夫・国友直人編，東京大学出版会，81-98，1989.

[6] Rissanen, J.：*Stochastic Complexity in Statistical Inquiry*, World Scientific, Singapore (1989).