Technical Report YK-048
June 30, 2020
Y. Karasawa

AIC（赤池情報量基準）を学ぶ

～良い統計的モデリングとその評価のための～

唐沢 好男

得られたデータからその元となる構造を推定する、あるいは、未来に起こることを予測する手法は統計的モデリングと呼ばれる。そのモデリングでは、正確さ（偏りとばらつき）、複雑さ（適用のしやすさやパラメータの数）、物理的意味との整合性などの種々の視点があり、出来具合にも優劣がある。そのようなモデルについて、よさの評価を行う手法としてAIC（Akaike Information Criterion: 赤池情報量基準）があり、以下の式で表される（詳しいことは本文で）。

$$\mathrm{AIC} = -2\times(\text{最大対数尤度}) + 2\times(\text{パラメータの数})$$

AIC は、赤池弘次博士が、数理統計理論を駆使して編み出したモデル選択指標のパラダイムである。無線通信、特に自然現象を相手にする電波伝搬においては統計的モデリングが重要であり、良いモデルがシステム設計の高信頼化・効率化に寄与する。上式で表される AIC はモデル評価の基本的考え方においても、評価式そのものものにおいてもシンプルでありながら、応用範囲が極めて広い。また、AIC を規範としてよいモデルを作ることもできる。本レポートでは、この AIC の入り口部分を解説する。「AIC の使い方だけ知っておけば十分」には満足せず、さりとて、「数学的に完璧に理解する」まとめ方は筆者にはできず、「直観的に理解できるレベル」を目標にまとめている。本レポートによって、AIC に興味を持った読者は、優れた専門書も多いので（例えば、[1]-[3]）、これらによって AIC の深い部分を学んで欲しい。

１．最小二乗近似の回帰曲線で何が問題か

図１のようなデータ（図中の Data：●プロット）があったとする。その傾向を捉えて、ベテランがエイヤッと引くカーブには職人芸をみる。でも、これは危うい。誰でもが共通にそれをなしうる科学的な方法が求められる。この目的に答える方法がガウスの編み出した最小二乗法（最小自乗法とも言われる）である。

真の特性 $g(x)$（図中の実線のカーブ）があり、x_i でのサンプル毎に誤差が乗り、y_i として得られたものが、図１のデータであったとする。すなわち、

$$y_i = g(x_i) + \varepsilon_i \quad (i = 1, 2, \cdots, n) \tag{1}$$

である。目的は真の特性 $y=g(x)$の形を知りたいのだが、手持ちの情報は、図 1 の Data として示した$(x_1,y_1), (x_2,y_2), \cdots, (x_n,y_n)$の n 組（図では n=11）のデータのみである。そこで、真の特性を予想するものとして、次式で与えられる K 次の多項式近似を考える。

$$f(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_k x^K \qquad (2)$$

この場合、次数 K と係数 a_0～a_K を定めることになる。次数を仮決めすると、係数 a_0～a_K は次式で求める最小自乗法で定められる（後に説明するように最尤推定でもある）。最小二乗法はモデルと実現値の残差の平方和を最小にする規範で、

$$J = \sum_{i=1}^{n} \left(f(x_i) - y_i \right)^2 \qquad (3)$$

の J が最小になるように $f(x)$を、すなわち a_0～a_K を定めればよい（その結果を $\hat{a}_0 \sim \hat{a}_K$ で表記）。最小二乗法による定め方は市販の数値計算ソフト（Excel, KaleidaGraph など）に組み込まれているので、ここではそれは述べない。このようにして定められたカーブ

$$\hat{y} = f(x \mid \hat{a}_0, \hat{a}_1, \hat{a}_2, \cdots, \hat{a}_K) \qquad (4)$$

は回帰曲線と呼ばれる。K=1 の場合は直線なので、回帰直線である。回帰曲線は、真のカーブ $g(x)$そのものを推定するのではなく、真のカーブが回帰曲線を中心に y 軸方向に正規分布すると予測する意味になる。

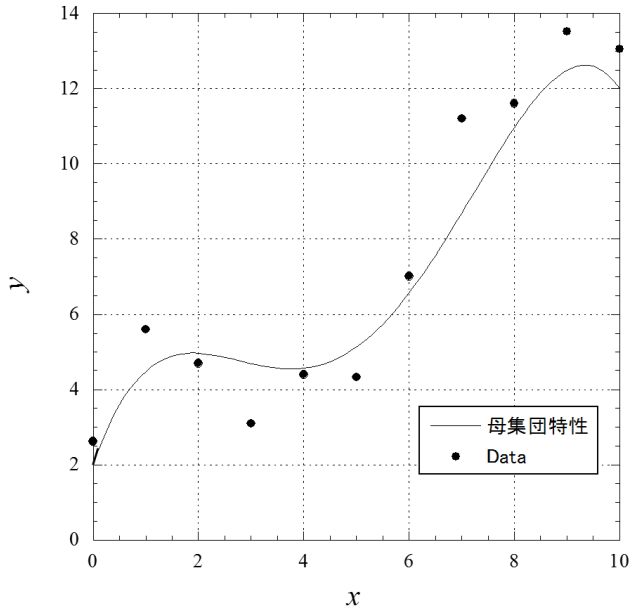

図 1　実線で表される母集団特性（真の特性 g(x)）に誤差を含んで現れる標本値（Data）の例

図２は、図１の Data （Data 1 と表記）に対して *K*=2, 4, 7 次で求めた回帰曲線を示している。図３は２種類の標本値（Data 1 と Data 2；Data 2 は Data 1 と誤差が独立な標本値）と Data 1 に対する回帰曲線との残差の平方和 J を示している。これより、回帰曲線を作った Data 1 に対しては次数を上げるほど *J* の値は小さくなってゆく様子が分かる。では、推定において次数は高ければ高いほどよいのであろうか？別の言葉で言うと、再度同じ現象からのデータが得られたとき最初のデータで求めた回帰曲線は、新たなデータに対しても有効に働いているのかどうか、である。図３の Data 2 をみるとむしろ次数が高くなると残差が増える傾向が読み取れる。（注：乱数を替えて行うと、特に Data 2 については、様々な結果が現われるので、図３は一例であるが傾向は類似している。）

これより、次数が低すぎるとデータの構造を適切に表現できないが、高すぎるとデータの偶然変動に過敏になり、将来の現象予測にも誤った知見を与えることになる。この例にみられるように、データへの適合度とモデルの複雑さ（次数選択問題）をどのように折り合いをつけるかが鍵になるが、そのことに対して最小二乗法は答えを与えてくれない。モデルの評価の視点で見ると、Data 1 では標本値でモデルを作り、その標本値でモデルの精度を評価しているから、精度において過大評価になることは、大いに予測できることである。

これから学ぶ AIC（Akaike Information Criterion: 赤池情報量基準）は、このようなときのモデル選択の判断基準を与えるものである。よいモデルを作るためには次数（パラメータ数）の選択が鍵になりそうなことが分かったが、では、AIC はこれにどのように答えてくれるのであろうか？

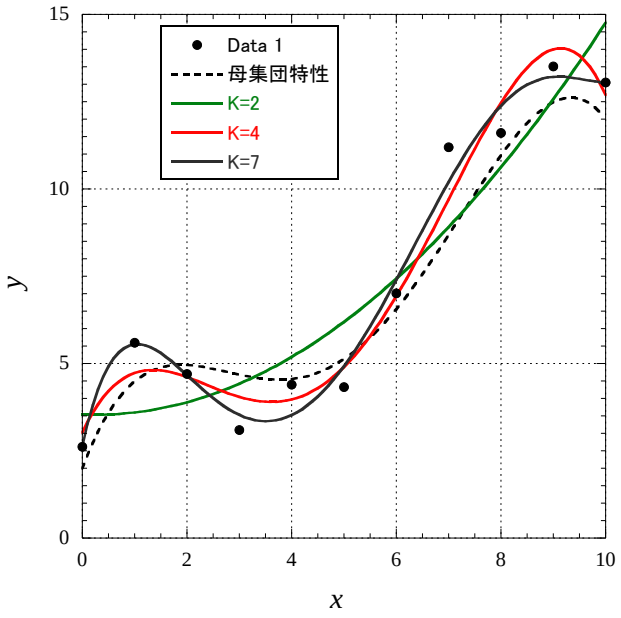

図２　標本値 Data 1（図１の Data と同じ）に対する回帰曲線（*K*=2, 4, 7）

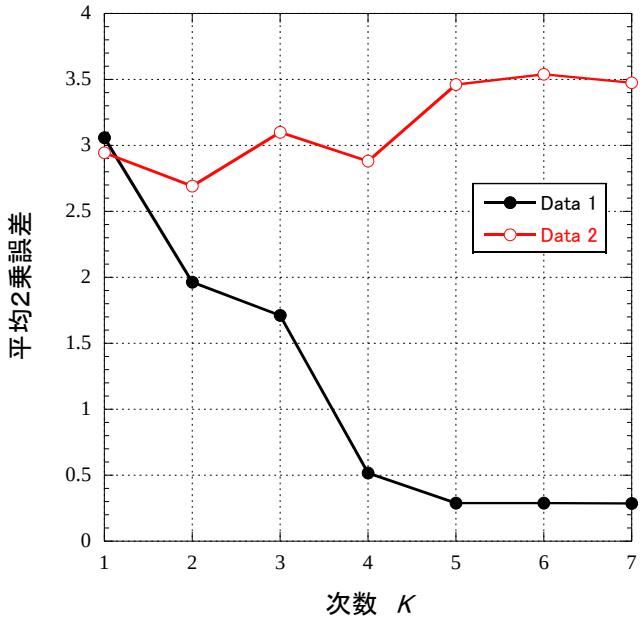

図３　Data 1 で作った回帰曲線に対する Data 1 と Data 2（Data 1 と誤差独立）の平均２乗残差の次数依存性

２．準備

第４節にて AIC を学ぶが、その理解に必要な用語（確率論や情報理論の教科書に出てくる用語）の定義を本節でまとめておく。

（１）確率分布

確率：p_i

確率 (probability)とは、偶然性を持つ事象について、その事象が起こることが期待される度合い、あるいは現れることが期待される割合のことを言い、0 から 1 までの値をとる。0 は絶対起こらない、1 は必ず起こるという極限値で、通常は、その中間の値になる。コイン投げでの表が出る確率は 1/2、サイコロでは、どの目の確率も 1/6 と言った値である。このように出現値が離散的なものであれば、上記のように、即座に数値で表すことができる。状態数が N のとき、p_i（i=1,2,…, N）の総和が 1 になる。

連続変数の確率密度関数：$f(x;\theta)$, $f(x|\theta)$

連続的に分布する量の確率は、離散状態数のような簡単なことにはゆかない。例えば、「人間の身長が 160cm である確率は？」と問われても、160.000…とぴったりした値は起こりえず、問いに意味がない。強いて言えば「確率 0」である。それに対して、「160cm から 165cm の間にある確率は？」、あるいは、「170cm 以上になる確率は？」と言う問いには、前提が明確であれば、適切に答えることができる。電波伝搬で取り上げられる物理量は、その大部分が、この例に属するアナログ量（＝連続分布）になる。

確率的規則に従って連続的に変化する物理量は確率変数と呼ばれる。変数 X がその変化範囲の中の微小範囲 dx(すなわち、$x- dx/2 \leq X < x+dx/2$)に存在する確率が、関数 $f(x)$ を用いて $f(x)dx$ と表わされるとき、$f(x)$を確率密度関数（probability density function: PDF）という。連続分布の場合は、確率密度と言う考え方が大事になる。確率密度は確率と違って、その値は0～∞の全範囲の値をとり得る。

変数 X の存在範囲を x_{min}～x_{max} とするとき、確率密度関数 $f(x)$には、連続分布・離散分布（デルタ関数の和で表される）を問わず、以下の性質がある。関数 $f(x)$が確率密度関数であるための必要条件とも言える。

$$f(x) \geq 0, \quad \int_{x_{\min}}^{x_{\max}} f(x)\,dx = 1 \tag{5}$$

確率分布のパラメータが例えば、θであるとき、$f(x; \theta)$のように ; で切り分けて表す。また、そのパラメータが具体的な値$\theta = \theta_0$ であるときの確率分布は $f(x|\theta_0)$のように | を用いる。

確率変数 X, Y の確率分布を $f(x), f(y)$とする。変数 X が微小範囲 $x-dx/2 \leq X < x+dx/2$)に、変数 Y が微小範囲 $y-dy/2 \leq Y < y+dy/2$)に同時に存在する確率が、関数 $f(x,y)$ を用いて $f(x,y)dxdy$ と表わされるとき、$f(x,y)$を結合確率密度関数と呼ぶ。二つの確率変数が無相関である場合、$f(x,y)=f(x)\,f(y)$であり、それぞれの分布の積になる。３変数以上も同様である。

（２）平均情報量（エントロピー）

AIC にはエントロピーと言う言葉は陽には現われないが、その前段となるカルバック・ライブラー情報量（KL 情報量；３節で説明）と深く関連しているので簡単に触れる。

離散確率分布のエントロピー

情報の量はビットを単位で表される。ビットはその事象（i）が起きる確率 p_i に対して、$-\log_2 p_i$ で与えられる。例えば、コイン投げでは $p_i=1/2$ であるので１ビット、サイコロでは $p_i=1/6$ なので 2.58 ビットである。滅多に起きないことが起きたことを知らせる情報は長いビット数になることを意味している。ビット数は、その情報に対する驚きの量であるともいえる。実際は、事象数 N に対して事象 i ごとに生起確率が異なるので、各情報量は以下のように表される。

$$I_i = -\log_2 p_i \quad \left(\sum_{i=1}^{N} p_i = 1 \right) \tag{6}$$

一つ一つの情報の集まりである情報源全体の情報量（＝平均情報量）は生起確率で重み付けされた情報量の全体であり以下の式になる。

$$H = \left\langle -\log_2 p_i \right\rangle = -\sum_{i=1}^{N} p_i \log_2 p_i \tag{7}$$

これは、**情報源のエントロピー**と呼ばれる（単位はビット）。式の形から明らかなように、p_i の一つが 1 で他がすべて 0 のとき、H は 0 となりそれ以外では正の値をとる。N=2 の場合、それぞれの生起確率を p_1, p_2 (=1-p_1)とすると、エントロピーは、p_1=0 と 1 で H=0, p_1 (=p_2)=0.5 で最大値 1 になる。状態数 N の場合に拡張すれば、p_1=p_2=⋯=p_N=1/N のとき、エントロピーが最大値 $\log_2 N$ になり、情報源として最も多くの情報を有する（＝冗長的な無駄なものがない）ことになる。

連続確率分布のエントロピー

上記は状態が離散的に生起する確率分布を述べたが、連続的に分布する場合の確率密度関数 $f(x)$に対するエントロピーをシャノンは以下の式で定義すると言う形で与えている（C.E. Shannon, BSTJ, 1948）。

$$H = -\int_{-\infty}^{\infty} f(x)\log f(x)\,dx \tag{8}$$

対数 log の底は、その定義を明確にして使えばなんでもよいが、以下では、e を底とする自然対数の意味で log を用いる。

離散分布との関係を調べる。連続確率変数 X を微小区間Δx 毎に分割し、それぞれの中心値を x_i で表す。区間 $x_i \pm \Delta x/2$ に入る確率を $f(x_i)\Delta x$ として、式(7)に適用すると、

$$\begin{aligned} H' &= \lim_{\Delta x\to 0}\left(-\sum_i f(x_i)\Delta x \log_2\{f(x_i)\Delta x\}\right) \\ &= \lim_{\Delta x\to 0}\left(-\sum_i f(x_i)\{\log_2 f(x_i)\}\Delta x\right) + \lim_{\Delta x\to 0}\left(-\sum_i f(x_i)\{\log_2 \Delta x\}\Delta x\right) \\ &\approx -\int f(x)\log_2 f(x)dx - \log_2 \Delta x \end{aligned} \tag{9}$$

と変形できるが、右辺第２項は$\Delta x \to 0$ で無限大になる。これは、連続分布の(7)式において無限個の状態数があるのと等価になり、この変換で得たエントロピーH'が無限大の値にたどり着くのは当然の帰結とも言える。しかし、式(9)からわかるように、右辺第２項の無限大項は確率分布とは無関係であり、種々の確率分布のエントロピーを比較したい場合には相殺されてしまう量である。このため、真に意味のある項は右辺の第１項ということになり、シャノンの定義式(8)となるわけである。離散分布の場合には、エントロピーは単位もビットであり、その値に意味があったが、連続分布の場合には、エントロピーの値そのものの意味は変質し、「エントロピーは非負である」の性質もなくなる（もちろん、(9)式の H'は正になる）。また、前述のとおり、対数の底も２である必要はなく、微積分が簡易になる自然対数とし、式(8)のように log =$\log_e$ を用いる。

一様分布（$a \le x \le b$）のエントロピーは

$$H = -\int_a^b \frac{1}{b-a} \log \frac{1}{b-a} dx = \log(b-a) \tag{10}$$

となり、0<*b*-*a*<1 で負になる。

平均値μ、標準偏差σの正規分布のエントロピーは

$$H = -\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \log\left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\right) dx$$

$$= \log_2 \sqrt{2\pi e \sigma^2} \tag{11}$$

である（導出は難しくない）。

正規分布は、２乗平均値がσ^2である確率密度関数の中で、最大エントロピー（(11)式）を持つ分布であり、**最大エントロピーの定理**と呼ばれる(*)。(*この証明は、例えば、中川聖一, 情報理論, p. 144, 近代科学社)

エントロピーは、式(8)の形から分かるように、確率変数 $X=x$ が $f(x)$で分布するときの -log $f(x)$の平均値である。確率変数 $X=x$ が $g(x)$で分布するときの-log $f(x)$の平均値を求めたいときは

$$H_{gf} = -\int_{-\infty}^{\infty} g(x) \log f(x) dx \tag{12}$$

と表される。この量は分布 g と f の**交差エントロピー**と呼ばれる。分布 g のエントロピー（式(8)で $f \rightarrow g$）を H_g と表すと、

$$H_g \le H_{gf} \tag{13}$$

が証明されており、等号は $g=f$ のときである。（証明は、例えば、３節で述べるカルバック・ライブラー情報量の式として、[2]の p.28）

（３）最尤推定

パラメータθをもつ連続型の確率分布 $f(x;\theta)$ からの無作為標本を考える。その n 個の標本値が $x_1, x_2, \ldots, x_n$ であったとする。このとき、θ の関数

$$L(\theta) = f(x_1 \mid \theta) \times f(x_2 \mid \theta) \times \cdots \times f(x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta) \tag{14}$$

に着目する。$L(\theta)$は $x_1, x_2, \ldots, x_n$ の関数と見ると、同時確率密度関数の形である。視点を変えてパラメータθの関数と見るとき、$L(\theta)$を**尤度関数**と呼ぶ。この時点で、分布のθ の値は分からないが、θ の値を変えて $L(\theta)$が最大になるθ、これを$\hat{\theta}$ とする、を見つける、すなわち

$$L(\hat{\theta}) = \max_{\theta} L(\theta) \tag{15}$$

となるθである。

このようにして尤度関数の最大値を定める方法を**最尤推定**（あるいは**最尤法**）と呼び、推定されたパラメータ値$\hat{\theta}$を**最尤推定値**、$L(\hat{\theta})$を**最大尤度**と言う。

最尤推定とは、その結果（ここでは$x_1, x_2, \ldots, x_n$が標本値として得られたこと）が起きる確率が最大になるパラメータ値を、最も有りそうなもの（最尤なもの）と推定する方法である。物事は起こるべくところに起きる、その現象が起きたということはその現象が最も起きやすかったから、という発想である。

尤度関数は式(14)により積の形で与えられるが、この対数をとった

$$l(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(x_i \mid \theta) \tag{16}$$

の方が扱いやすい。$l(\theta)$の最大値を与えるθは $L(\theta)$の最大値を与える最尤推定値$\hat{\theta}$と同じである。この$l(\theta)$を**対数尤度関数**、$l(\hat{\theta})$を**最大対数尤度**と言う。図４は上記一連の説明をまとめている。

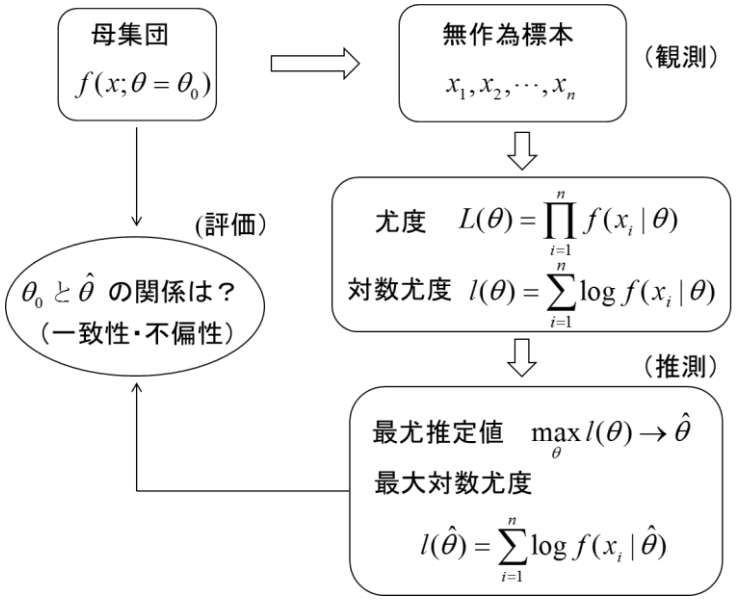

図４　最尤推定の仕組み（観測・推測・評価）

正規分布する母集団から無作為で抽出したn個の標本$(x_1, x_2, \ldots, x_n)$から求められる平均値と標準偏差の最尤推定値$\hat{\mu}, \hat{\sigma}$を求めてみよう。

対数尤度関数は

$$l(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{17}$$

であり、これを最大にするパラメータ値（$\hat{\mu},\hat{\sigma}^2$）は、以下の連立方程式で定められる。

$$\frac{\partial l(\mu,\sigma^2)}{\partial \mu}=\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)=0 \quad \rightarrow \quad \hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}x_i=\bar{x} \tag{18a}$$

$$\frac{\partial l(\mu,\sigma^2)}{\partial \sigma^2}=-\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2=0 \quad \rightarrow \quad \hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 \tag{18b}$$

このように、正規分布の場合には、平均値の最尤推定値 $\hat{\mu}$ は標本平均で、分散の最尤推定値 $\hat{\sigma}^2$ は標本分散で求められる便利な性質がある。ただし、分散の最尤推定値は、期待値を推定する不偏分散（(n-1)で割る形）とは係数が異なっていることに留意して欲しい。

種々の確率分布に対して最尤推定値が解析的に求められていると便利だが、それができるのは上述の正規分布や指数分布など限られている。例えばガンマ分布に対しては対数尤度関数の導関数までは求められるが、その後の連立方程式を解いて最尤推定値を求めようと思っても、特殊関数の非線形方程式になって閉形式の解が得られない。そのような場合には、コンピュータによる数値解法（ニュートン・ラフソン法などの反復法）によらなければならず、その点が不便である（この不便さは、AIC にも共通する）。

３．二つの確率分布の近さを測る尺度：カルバック・ライブラー情報量

一方を基準とする確率分布 $g(x)$、もう一方を評価したい確率分布 $f(x)$とする。この二つの分布の近さを定量的に評価する尺度が求めらのれる。たとえば、対数正規分布が、他の関数と組み合わせて使う場合の扱いにくさために、解析性に優れた、たとえば、ガンマ分布で置き換えたいというような場合である。あるいは、実測値の分布を得た場合に、それを特定の分布で近似したいというような場合である。

最も簡単な尺度は分布の距離を求めるノルムである。

$$\int\left|g(x)-f(x)\right|dx \qquad (L^1\ \text{ノルム}) \tag{19a}$$

$$\int\left\{g(x)-f(x)\right\}^2 dx \qquad (L^2\ \text{ノルム}) \tag{19b}$$

上記の評価で十分な場合もあるが、分布の値の大きい部分の差が結果を支配するので、確率の小さい部分も含めて全体を見たい場合には不向きである。これに対して、次式で定義されるカルバック・ライブラー情報量（Kullback-Leibler (KL) divergence ; KL 情報量）がある。

$$D(g;f)=\int_{-\infty}^{\infty}g(x)\log\left\{\frac{g(x)}{f(x)}\right\}dx \tag{20}$$

２つの分布 g と f が同じ場合に D=0、異なる場合に D>0 となり、ずれの程度に応じて値も大きくなる（$D\geq 0$ の証明は[2], p.28）。例えば、分布 g の置換目的であれば、置き換えたい

分布 f のパラメータ値を調整して、D が最も小さくなるように最適化するための規範として用いられる。直感的には二つの分布の距離のイメージであるが、上式で g と f を入れ替えたときの対称性が崩れているので、距離と呼びにくい事情がある。

式の形を見ると、比が求められるので、ノルム評価とは逆に確率の値の小さい部分の影響を大きく受けそうにも見えるが、$g(x)$がかかって値の小さい部分の影響は押さえられているので、バランスが保たれている。(20)式を書き直すと

$$\int_{-\infty}^{\infty} g(x)\log\left\{\frac{g(x)}{f(x)}\right\}dx=\int_{-\infty}^{\infty} g(x)\log g(x)\,dx-\int_{-\infty}^{\infty} g(x)\log f(x)\,dx \quad (21)$$

となる。右辺第１項に負符号をつけたものが確率分布 g のエントロピー、第２項の負符号を含んだものは交差エントロピーと呼ばれる量であり、(13)式そのものである。すなわち、

KL 情報量 = （確率分布 g と f の交差エントロピー）- （確率分布 g のエントロピー）
(22)

と整理される。KL 情報量は次節で述べる AIC の出発点にある。

KL 情報量は、電波伝搬に現われる確率分布の近似度評価でも有効である。筆者は無限積分を含んで計算が不便な Loo 分布（仲上・ライス分布の直接波成分に対数正規分布する変動が加わる伝搬環境を表すモデル）を、仲上m分布など複数の分布に近似する際の近似度の比較や適用範囲を調べるのに用いたが、有益な結果を得ることができた(*)。(* 唐沢研 HP 公開の技術レポート YK-044-rev)

４．AIC

第１節で統計的モデリングでの評価基準の必要性を、２節と３節で評価基準構築の下準備を行ってきた。本節が、本レポートの主要部分である。

（１）出発点と目的地：KL 情報量ではなぜだめか？

AIC 構築の背景をなす大きな流れは以下である。

１）未知の母集団 $g(x|\theta_0)$から無作為に抽出した一組の標本がある：$x_1, x_2, \ldots, x_n$

２）その仕組みを推定するモデル（＝評価したいモデル）がある：$f(x;\boldsymbol{\theta})$ （複数のパラメータを想定しベクトル$\boldsymbol{\theta}$で表記）

３）AIC でモデルを評価する（複数のモデルが有る場合には AIC の数値を比較して良い方のモデルを選択する）

もし、母集団の特性（確率分布 $g(x|\theta=\theta_0)$）が分かっていれば、モデルの良さは KL 情報量 $D(g,f)$で評価できる。しかし、それ（すなわち g）がわかっていないから、以下展開する AIC を生み出す土壌が存在している。すなわち、$g(x|\theta_0)$を知らずに、標本値だけで $f(x)$の良さをどう評価するか、と言う問題設定である。

（２）KL 情報量の中のモデル依存項に注目

KL 情報量は、式(21)より、

$$D(g,f)=\int_{-\infty}^{\infty} g(x)\log g(x)\,dx-\int_{-\infty}^{\infty} g(x)\log f(x)\,dx \tag{23}$$

で表されたが、右辺の第一項（確率分布 g のエントロピーの符号を反転させたもの）は、評価したい確率分布 f に関係しない項であり、種々のモデル f に対する比較をしたい場合には共通項として相殺されるものである。故に、右辺第２項の大小関係のみを調べることができれば、モデルの良さに関する比較評価が可能になる。すなわち、

$$D_0\equiv\int_{-\infty}^{\infty} g(x)\log f(x\,|\,\boldsymbol{\theta})\,dx=\left\langle \log f(x\,|\,\boldsymbol{\theta})\right\rangle_g \tag{24}$$

に着目すればよいということである。(21)式ではこの項に負符号をつけたものを交差エントロピーと呼んだが、次項ではこの量に着目する。しかし、この式にも未知の関数 $g(x)$が含まれているので、問題が解決されているわけではない。

（３）平均対数尤度と最大対数尤度：求めたいものと求められるもの

以下の２つの式を定義する。

$$la(\boldsymbol{\theta})\equiv nD_0=n\int_{-\infty}^{\infty}\log f(x\,|\,\boldsymbol{\theta})\,g(x)dx \tag{25}$$

$$l(\boldsymbol{\theta})\equiv\sum_{i=1}^{n}\log f(x_i\,|\,\boldsymbol{\theta}) \tag{26}$$

(25)式は前項で着目した(24)式と同じ形であるが、パラメータ $\boldsymbol{\theta}$ の関数として見ていて、平均対数尤度の n 倍値である。一方、(26)式は対数尤度（(16)式）である。

(26)式を最大にするパラメータの最尤推定値 $\hat{\boldsymbol{\theta}}$ を求め、(25), (26)式に代入すると

$$la(\hat{\boldsymbol{\theta}})=n\int_{-\infty}^{\infty}\log f\left(x\,|\,\hat{\boldsymbol{\theta}}(x_1,x_2,\cdots,x_n)\right)g(x)dx \tag{27}$$

$$l(\hat{\boldsymbol{\theta}})=\sum_{i=1}^{n}\log f\left(x_i\,|\,\hat{\boldsymbol{\theta}}(x_1,x_2,\cdots,x_n)\right)\quad\text{（最大対数尤度）} \tag{28}$$

である。(28)式は最大対数尤度、(27)式は、厳密には「パラメータを最尤推定値で固定した場合の平均対数尤度の n 倍値」であるが、以下ではこの値自体を「平均対数尤度 n 倍値」と呼ぶ。この二つについて、次のことが言える。

１）平均対数尤度 n 倍値 la は n が十分大きいとき、大数の法則により、

$$\lim_{n\to\infty} la(\hat{\boldsymbol{\theta}}) = n\int_{-\infty}^{\infty} \log f(x\,|\,\boldsymbol{\theta}_0)\, g(x)dx \tag{29}$$

となり、求めたい量に確率収束する。このため、平均対数尤度 n 倍値が求められれば、モデル間の比較評価が可能である。しかし、g を知らないのでこれを求めることはできない。

２）一方、最大対数尤度 l は標本値とモデルから計算できる。平均対数尤度 n 倍値と最大対数尤度の関係が分かれば、最大対数尤度からモデル評価が可能になる。ただし、n が十分大きい極限において両者が等しいのかどうかは式の上からは判然としない。

そこで、以下、最大対数尤度 $l(\hat{\boldsymbol{\theta}})$ と平均対数尤度 n 倍値 $la(\hat{\boldsymbol{\theta}})$ の違いを調べる。

（４）最大対数尤度 $l(\hat{\boldsymbol{\theta}})$ と平均対数尤度 n 倍値 $la(\hat{\boldsymbol{\theta}})$ の差

平均対数尤度 n 倍値 $la(\hat{\boldsymbol{\theta}})$ が分かれば、モデル評価に使える。しかし、それはわからず、最大対数尤度 $l(\hat{\boldsymbol{\theta}})$ なら計算できる。その差はどのくらいか、それを定量的に示したのが AIC の真髄である。そして、その差はパラメータの数で近似できるという驚くべき結果を得るのであるが、その話は次項で述べる。ここでは、計算機シミュレーションにより、両者の値を比較した結果を示す。

母集団の分布 g を独立な多重正規分布（平均値は 0）とする。モデル f も同様に、平均値 0 の多重正規分布とする。普通は g と f では分布形も異なるだろうが、ここでは差の存在を見たいだけが目的なので、状況は可能な限りシンプルにする。以下、必要な式を列記する。

$$g(x_1, x_2, \cdots, x_K\,|\,\sigma_1^2, \sigma_2^2, \cdots, \sigma_K^2) = \prod_{k=1}^{K} g_k(x_k\,|\,\sigma_k^2)$$

$$g_k(x_k\,|\,\sigma_k^2) = \frac{1}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{x_k^2}{2\sigma_k^2}\right)$$

$$f(x_1, x_2, \cdots, x_K\,|\,\hat{\sigma}_1^2, \hat{\sigma}_2^2, \cdots, \hat{\sigma}_K^2) = \prod_{k=1}^{K} f_k(x_k\,|\,\hat{\sigma}_k^2)$$

$$f_k(x_k\,|\,\hat{\sigma}_k^2) = \frac{1}{\sqrt{2\pi}\hat{\sigma}_k}\exp\left(-\frac{x_k^2}{2\hat{\sigma}_k^2}\right)$$

$$\hat{\sigma}_k^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_k(i)\right)^2$$

$$l(\hat{\sigma}_1^2, \hat{\sigma}_2^2, \cdots, \hat{\sigma}_K^2) = \log\prod_{i=1}^{n} f(x_{1,i}, x_{2,i}, \cdots, x_{K,i}\,|\,\hat{\sigma}_1^2, \hat{\sigma}_2^2, \cdots, \hat{\sigma}_K^2)$$

$$= \sum_{i=1}^{n}\sum_{k=1}^{K}\log f_k(x_{k,i}\,|\,\hat{\sigma}_k^2)$$

$$la(\hat{\sigma}_1^2,\hat{\sigma}_2^2,\cdots,\hat{\sigma}_K^2)=n\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\prod_{k=1}^{K}g_k(x_k\,|\,\hat{\sigma}_k^2)\sum_{k'=1}^{K}\log f_{k'}(x_{k'}\,|\,\hat{\sigma}_{k'}^2)\,dx_1dx_2\cdots dx_K$$

シミュレーションでは多重度 K=1, 2, 3 の三つ場合に対して、分散σ_1^2, σ_2^2, σ_3^2 は全て 1 で設定している。

図 5 は、K=1 と 3 について、n=100～200 とした場合の両者の差（*l*-*la*）をプロットしている。もとになる *l* や *la* の値が大きなものであり（K=1 で-200 程度）、差も大きくばらついていて、何か傾向が見えているとは言いにくい、また、n を増やしてもこの傾向は変わらない（＝ばらつきが収まることはない）ことを確認している。しかしこの中に、バイアスが隠れているのである。

このシミュレーションを乱数を変えて 100 回行い、この範囲の n の全数（n=100～200 の 101 組）に対して（すなわち 10,100 個で）平均をとると、K=1, 2, 3 に対して、それぞれ、0.9201, 2.1052, 3.1011 の値を得た。近似ではあるがパラメータ数 K そのものになっているのである。（上記数値は 1 セット（100 回試行）の一例であるが、これを繰り返しても、大体この程度であり、設定する分散の値（σ_1^2, σ_2^2, σ_3^2）を変えても同様の結果になることを確認している）。

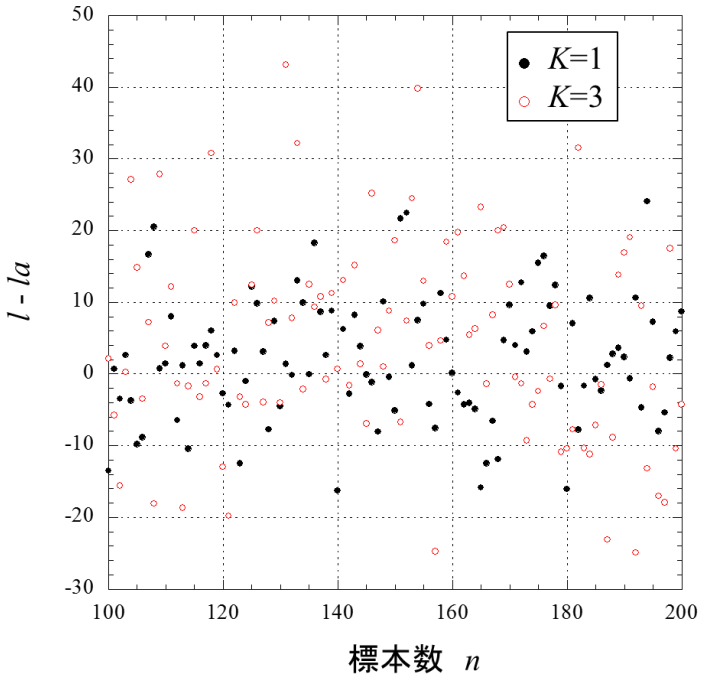

図 5　　最大対数尤度と平均対数尤度の差に関する計算機シミュレーション結果

（K=1 と 3、n=100～200）

（５）補正項（＝最大対数尤度 $l(\hat{\boldsymbol{\theta}})$ と平均対数尤度 n 倍値 $la(\hat{\boldsymbol{\theta}})$ の差）の理論的導出

前項（４）に示したシミュレーション結果では、最大対数尤度 $l(\hat{\boldsymbol{\theta}})$ と平均対数尤度 n 倍値 $la(\hat{\boldsymbol{\theta}})$ の差の平均値は、モデルのパラメータ値に近い値であることを予見することができた。しかしこれはあくまで一例であって、この量を定量的に明らかにすることが求められる。この解析が AIC の本丸になる。

補正値の理論解析は数理統計学の知識を必要とし、筆者には手に負えない部分がある。そこで、この導出は文献[1]（４章：ただし、正規分布の場合の解析に留まっている）や文献[2]（第３章：一般的な導出過程が示されている）をみて学んで欲しい。（ページを割いてその部分をここに再掲するより、オリジナルを読んでもらったほうがよいと判断）。以下、その概要のみを述べる。

補正項＝最大対数尤度 – 平均対数尤度 n 倍値

$$= l\left(\hat{\boldsymbol{\theta}}(x_1, x_2, \cdots, x_n)\right) - la\left(\hat{\boldsymbol{\theta}}(x_1, x_2, \cdots, x_n)\right)$$

$$= \sum_{i=1}^{n} \log f\left(x_i \mid \hat{\boldsymbol{\theta}}(x_1, x_2, \cdots, x_n)\right) - n\int_{-\infty}^{\infty} \log f\left(x \mid \hat{\boldsymbol{\theta}}(x_1, x_2, \cdots, x_n)\right) g(x)dx \quad (30)$$

であり、これを３つの部分に分解して、それぞれの期待値（<・>）を求める。すなわち、

$$\langle \text{補正項} \rangle = \left\langle l\left(\hat{\boldsymbol{\theta}}(x_1, x_2, \cdots, x_n)\right) - l\left(\boldsymbol{\theta}_0\right) \right\rangle + \left\langle l\left(\boldsymbol{\theta}_0\right) - la\left(\boldsymbol{\theta}_0\right) \right\rangle + \left\langle la\left(\boldsymbol{\theta}_0\right) - la\left(\hat{\boldsymbol{\theta}}(x_1, x_2, \cdots, x_n)\right) \right\rangle \quad (31)$$

である。このとき、右辺第２項が 0 になることは容易に導かれる。一方、右辺の第１項と第３項の導出については種々の近似が必要になる。その結果、最終的に、第１項と第３項については

$$\left\langle l\left(\hat{\boldsymbol{\theta}}(x_1, x_2, \cdots, x_n)\right) - l\left(\boldsymbol{\theta}_0\right) \right\rangle \approx \left\langle la\left(\boldsymbol{\theta}_0\right) - la\left(\hat{\boldsymbol{\theta}}(x_1, x_2, \cdots, x_n)\right) \right\rangle \approx \frac{\text{パラメータ数}}{2} \quad (32)$$

となる。結果として、

$$\text{最大対数尤度の補正項} \approx \text{パラメータ数} \quad (33)$$

である。ここまでシンプルな結果が得られるのなら、もっと直感的に説明ができるのではと思うが、筆者はその方法を知らない。

なお、この結果が得られるためには、モデル $f(x)$ が真の特性 $g(x)$ を含むかそれに近い関係（図５に示した例題はまさにそういう典型例）のときである等の条件もある[2]。通常は、真実とあまりにかけ離れたモデルを想定すると言うことは無いであろうから、(33)式は大部分のケースに適用できると期待してよい。

（６）AIC のまとめ

以上をまとめて AIC は以下のように整理されている。

$$\mathrm{AIC} = -2\times(\text{最大対数尤度}) + 2\times(\text{パラメータの数}) \tag{34}$$

$$\text{最大対数尤度}：l(\hat{\boldsymbol{\theta}}) = \sum_{i=1}^{n} \log f\left(x_i \mid \hat{\boldsymbol{\theta}}(x_1, x_2, \cdots, x_n)\right)$$

AIC では、この値が小さいほど良いモデル（＝選択すべきモデル）と言う規準を与えている。パラメータを多くすると対数尤度を大きくできるが、パラメータが増えること自体がペナルティであると解釈され、統計的モデリングにおいてはバランスの大事さ、すなわち、誤差が同じ程度ならパラメータ数の少ないモデルを選ぶべきということが強調される結果になっている。

図６は AIC による一連の評価ステップをまとめている。

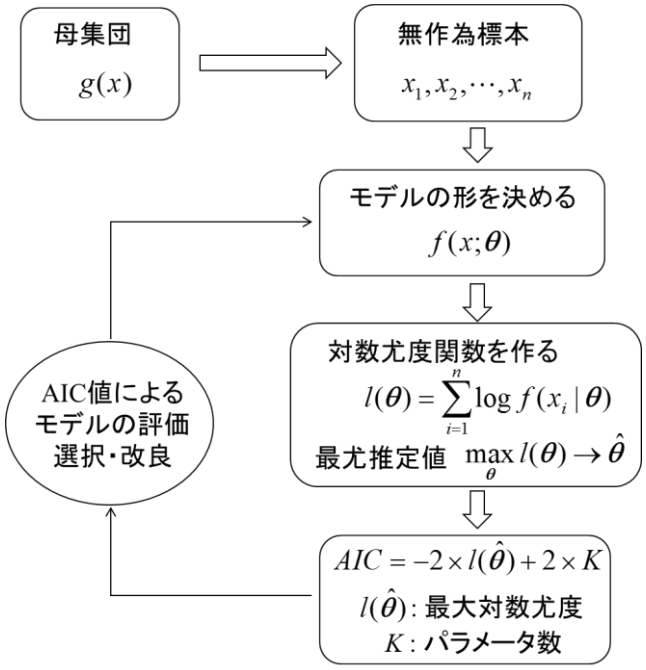

図６　AIC によるモデルの評価ステップ（評価・選択・改良）

５．AIC の応用例

得られた複数の標本値から、モデルとなる確率分布形を定め、そのパラメータの最尤推定値を得、そのモデルの良さを AIC で評価する方法は、図６で示した手順をそのまま実行すればよい。それを実際に行うときのネックは、モデルパラメーターの最尤推定値を標本値から定めるとき、それが解析的に求められない場合、コンピュータによる数値解法（ニュートン・ラフソン法などによる反復計算法）によらなければならない点である。正規分布やレイリー分布、指数分布のように解析解が得られる場合はよいが、ガンマ分布や仲上 m 分布な

ど大部分は数値解法になる。ここでは、回帰曲線との残差の分布が正規分布モデルに結び付けられる多項式近似モデルの例（＝1 節の図 1 で提起したその問題）を示す。

以下の式にしたがって生起される n 個の(x_i,y_i)標本値があったとする。

$$y_i = 2 + 4x_i - 1.8x_i^2 + 0.3x_i^3 - 0.015x_i^4 + \varepsilon_i \qquad (0 \le x \le 10 \text{ の } n \text{ 個}) \quad (35)$$

式中のε_iは標本に含まれる誤差を与える量で、平均値 0、分散 1 の標本毎に独立な正規分布（$N(0,1)$）である。1 節で示した Data とその母集団特性はこの式に基づいている。ここでは、n=10 と 100 で評価する。n=10 の例は図 1 で示しており、n=100 の場合を図 7 にプロットしている。

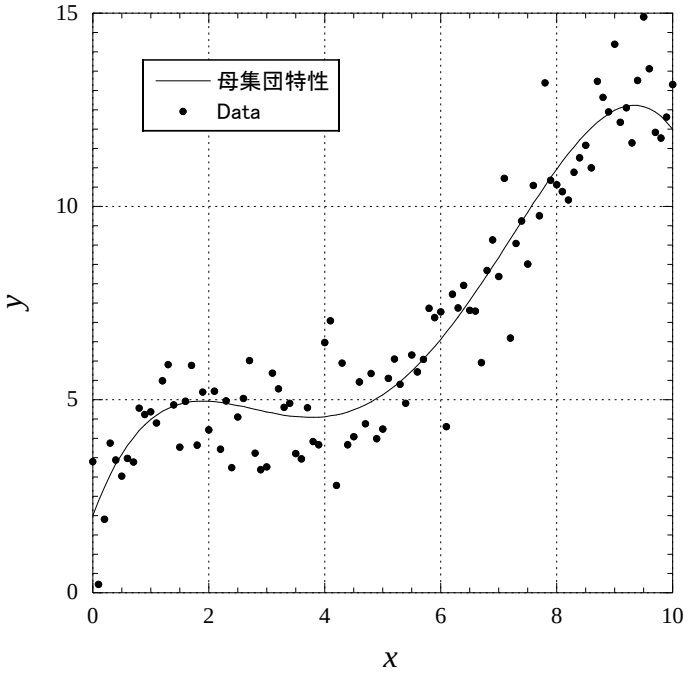

図 7　AIC 評価用データ（n=100）（n=10 は図 1 ）

このデータに対して以下の多項式モデルを評価する。

$$y = a_0 + a_1 x + a_2 x^2 + \cdots + a_K x^K + b \quad \left(f_b = N(0,\sigma^2) \right) \qquad (35)$$

この場合のパラメータは a_0～a_K と σ^2 の K+2 個である。基本的な考え方は、$a_0 + a_1 x + \cdots + a_K x^K$ を中心としてずれが正規分布するモデルのパラメータの最尤推定である。AIC によって、最適な次数（K の値）を見つけ、良いモデルを選び出す問題である。

正規分布の場合、対数尤度関数は(17)式であるが、以下に再掲する。

$$l(\boldsymbol{\theta}) = -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \sum_{k=0}^{K} a_i x_i^k\right)^2 \tag{36a}$$

$$\boldsymbol{\theta} = \left(a_0, a_1, a_2, \cdots, a_K, \sigma^2\right) \tag{36b}$$

係数 a_0～a_K の最尤推定値 $\hat{a}_0 \sim \hat{a}_K$ は尤度関数をそれぞれ a_i で偏微分し 0 と置いて得られる K+1 個の連立方程式から求められる。

$$\frac{\partial}{\partial a_i}\sum_{i=1}^{n}\left(y_i - \sum_{k=0}^{K} a_i x_i^k\right)^2 = 0 \ \ (i = 1, 2, \cdots, n) \ \Rightarrow \ \hat{a}_0 \sim \hat{a}_K \tag{37}$$

これは最小二乗法であるので、最小二乗法の機能が具備されている市販ソフト（Excel, KaleidaGraph など）で容易に求めることができる。また、誤差分散の最尤推定値は、

$$\frac{\partial l\left(\boldsymbol{\theta}(\hat{a}_0, \cdots, \hat{a}_K, \sigma^2)\right)}{\partial \sigma^2} = 0 \ \Rightarrow \ \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{k=0}^{K} \hat{a}_i x_i^k\right)^2 \tag{38}$$

となり、これらを(36a)式に代入すると、最大対数尤度は

$$l(\hat{\boldsymbol{\theta}}) = -\frac{n}{2}\log\left(2\pi\hat{\sigma}^2\right) - \frac{n}{2} \tag{39}$$

である。これより、AIC はパラメータ数が K+2 であるので次式となる。

$$AIC(K) = n\left(\log 2\pi + 1\right) + n\log\hat{\sigma}^2 + 2(K+2) \tag{40}$$

表 1 は、n=10 と 100 の一例（図 1 と図 7）について、K=1～7 のモデルに対する平均二乗残差（分散の最尤推定値 $\hat{\sigma}^2$）、最大対数尤度（$l(\hat{\boldsymbol{\theta}})$）、AIC 値（$AIC$）および最小値との差分（$\Delta AIC$）を示している。

同表より、n=11 では K=5 が、n=101 では K=4 が AIC 値の最小値になった。評価を繰り返してみても、標本値の数 n が大きくなると安定して K=4 を選択するが、標本数が少ない場合（n=11）では K=3, 4, 5 付近が選ばれる結果を得ている。母集団特性の次数を K=4 で与えているので、AIC が妥当な次数のモデルを選んでくれたと言える。

正規分布に帰着できるモデルの評価においては、回帰曲線との残差の二乗平均値は次数 K と共に小さくなるが、ある値を過ぎると変化ほとんどなくなる。この変化がない部分は最大対数尤度も変化が小さくなり、結果として、パラメータの増加に対するペナルティ分のみが AIC 値を押し上げる働きをする。このことから、二つのモデルが同程度の残差（＝標本値と回帰曲線との残差の二乗和）を与えるならば、パラメータの少ないほうが良いモデルであるとする原理が示されたのである。

表 1　*n*=11 と 101 の標本値に対するモデルの最大対数尤度と AIC 評価値

	$n=11$				$n=101$			
K	$\hat{\sigma}^2$	$l(\hat{\theta}^2)$	AIC	ΔAIC	$\hat{\sigma}^2$	$l(\hat{\theta}^2)$	AIC	ΔAIC
1	3.0594	-21.759	49.517	17.951	2.4465	-188.49	382.99	83.827
2	1.9648	-19.323	46.646	15.080	1.4410	-161.76	331.52	32.362
3	1.7104	-18.560	47.120	15.554	1.4363	-161.60	333.20	34.038
4	0.51739	-11.984	35.968	4.4022	1.0053	-143.58	299.16	0
5	0.28911	-8.7831	31.566	0	1.0041	-143.52	301.04	1.8795
6	0.28928	-8.7863	33.573	2.0067	0.99152	-142.88	301.77	2.6057
7	0.28748	-8.7520	35.504	3.9380	0.99111	-142.86	303.72	4.5634

６．むすびと参考文献

AIC が生まれるまでの背景と基本的な考え方、具体的な評価法について述べた。基礎理論は数学的に高度であるが、使う立場で見ると、非常に簡易で理に適った評価法であることがわかったと思う。それでも、一般的な分布からパラメータの最尤推定値を求める部分などに、コンピュータによる数値計算のお世話にならなければいけないが、正規分布を基本とするモデル化であれば、数式も解析的に解けて、非常に役立つ手法である。

情報量基準に基づく評価に本来使いたいもの（＝平均対数尤度）は母集団の確率分布（＝未知の分布）を含んでいて直接求めることはできないが、標本値とモデルによって求めることができる最大対数尤度は、一定の補正をすれば、平均対数尤度に替えて使える。その場合の補正量はモデルのパラメータ数である、と言う奇跡のような結果が得られたのである。AIC の真髄はまさにここにある。そしてその結果、二つのモデルが共に同程度の誤差を与えるならば、パラメータの少ないほうが良いモデルであるとする原理が示されたのである。まさに、**オッカムのかみそり**である。

以下、AIC に関する補足的なことをまとめておく（[2]の 3.5.3 項）。

１）我々のモデリングの目的は良いモデルを求めることであって、真のモデルを求めることではない。良いモデルとは、同じ現象が次に起こった場合でも、誤差の少ない安定した推定を行うことができるモデルである。真のモデルが分かっているのはコンピューターシミュレーションのときくらいである。現実の世界の真のモデルはパラメータ数が無限大（影響の大きいものから無視できるほどに小さいものまで合わせて）であるだろうとしかいえず、パラメータも結果に影響を及ぼす主要なものが選ばれていればそれで良い、と言う理解である。

２）AIC では、真の次数とモデルの次数が一致するとは限らない。データ数が少ない場合には、高次のモデルに見られる予測の不安定性を避けるためにも、低次のモデルを用いたほうが良い場合があることを示唆している。

３）AIC は複数のモデルを比べて、どれが最も良いモデルであるかを選ぶ相対評価ができるのであって、KL 情報量のような絶対的な意味での近似度評価はできない。その場合でも、

AIC をモデル最適化の規範（ゴールの方向を決める目的関数）にして、適応的にパラメータを追い込んで行くようなアダプティブアルゴリズムに組み込む手法には適している（AI（深層学習、ニューラルネット）による評価への応用など)。

４）情報量基準に基づくモデル評価手法は AIC を源流（1973～）にして様々に進化発展している。AIC では補正量をパラメータ数で決めているが、特に、この部分について、理論面で高度な一般化が行われている（TIC, GIC など[2])。また、ベイズ推定など、より進化した推定法も生まれている。しかし、本文でも述べたように、仮定するモデルが、真の特性を含むかそれに近いようなモデルであれば、AIC（＝補正項にパラメータ数を使う）で問題ないことが調べられている[2]。データとかけ離れた的外れなモデルでの評価では補正量が大きくなりすぎて、AIC は適当ではないが、我々が行う統計的モデリングでは、大部分が上記前提を満たしていると思うので、半世紀たった今においても AIC はその有用性を保ち続けている（決して時代遅れの情報量基準ではない)。

AIC の専門書や解説書も多数出版されているが、筆者がこのレポートを作成するに当たっては、以下の３冊を参考にした（＝３冊で勉強した)。

[1] 島谷健一郎, *フィールドデータによる統計モデリングと AIC*, ISM シリーズ : 進化する統計推理２, 近代科学社, 2012.
[2] 小西貞則, 北川源四郎, *情報量基準*, 朝倉書店, 2004,
[3] 赤池弘次他, *赤池情報量基準 AIC：モデリング・予測・知識発見*, 2007.

文献[1]は、フィールド科学者である著者が自然界の生物や林業を題材に統計モデリングの大切さを語っている。AIC を学びたいだけであれば、その第４章を読めばよいのであるが、全体を通して読み物としての楽しさ（生きた学問を語る熱さ）がある。文献[2]は AIC を含めた各種情報量基準を教科書として学ぶのに適している。AIC 構築の本丸である補正項についても丁寧な導出過程が示されている。文献[3]は、赤池博士が 2006 年に京都賞（第 22 回）を受賞したことを記念して、赤池博士自らと AIC 理論を進化・発展させてきたグループメンバーによってまとめられた本で、AIC が生まれる背景や思想、その後の発展についてまとめられている。筆者（唐沢）は[3], [1], [2]の順で読み進めたが、もう一度[3]に戻って読み直してみると、我々の分野（電波科学・情報通信）の研究にも共通する独創性の大切さを再認識する良い機会になり、目から鱗が落ちる思いがした。