統計数理 (1993)
第41巻 第1号 25-31

誤差評価と情報量規準

統計数理研究所　石　黒　真木夫

(1992年12月　受付)

1. はじめに

情報量規準を使ったデータ解析をしているとつい忘れがちになるが，解析結果に誤差の見積りをつけてほしいといわれる事が多い．情報量規準の立場からも避けて通ることの出来ない問題である．

誤差に関する議論をまぎれなく進めるためには，何のどういう意味の誤差について語ろうとしているのかを明確にしておくことが重要である．一般に，ある確率変数 X の確率密度関数を $g(\cdot)$，その $g(\cdot)$ を近似する（と考えられる）パラメトリックモデルの族を $f(\cdot\,|\,\cdot)=\{f_k(\cdot\,|\,\theta_k)\colon k=1,2,\ldots,K\}$ で表すことにしよう．

この設定のもとでの誤差の議論には

(1) ある標本空間上の確率分布に関して定義される汎関数 $\mu[\cdot]$ の真の分布 $g(\cdot)$ に対する値 $\mu_*=\mu[g(\cdot)]$ の推定とその誤差評価について考察する場合，

(2) あるモデル $f_k(\cdot\,|\,\theta_k^*)$ を真の分布と考えて，パラメータ θ_k^* の推定法とその誤差評価について議論する場合，

などが区別される．つぎの章で情報量規準が (1) の立場での推定問題に貢献したものと持ち込んだ問題を論じ，第3章で (2) の立場における誤差評価に新しい光をあてる．本論文の中心は第3章にあり，分布のパラメータの推定誤差の見積りに対する情報量規準の観点からの意味付けとその利用法を論じる．

2. モデル選択と推定問題

ここでは，真の分布 $g(\cdot)$ に対するある汎関数 $\mu[\cdot]$ の値 $\mu_*=\mu[g(\cdot)]$ の推定とその誤差評価について考察する．

μ_* の推定法として最も簡単な方法は，あるモデル（f_k としよう）を固定しておいてなんらかの方法でそのパラメータの推定値 $\hat{\theta}_k(x)$ を求め，

$$\hat{\mu}_k(x)=\mu[f_k(\cdot\,|\,\hat{\theta}_k(x))]$$

をもって μ_* の推定値とする場合である．ある θ_k^* が存在して $g(\cdot)=f_k(\cdot\,|\,\theta_k^*)$ がなりたつものとして最尤法によって $\hat{\theta}_k$ を定めるのが代表的な方法である．最尤法を使った場合に $\hat{\mu}_k(x)$ の推定誤差を見積る方法はよく知られている．

つぎに推定手続きにモデル選択が組込まれた場合を考えよう．データ x が領域 I_k に落ちたときにモデル f_k が選択され，そのモデルのパラメータの推定値 $\hat{\theta}_k(x)$ を求める手続きもあらかじめ決められているものとする．$x\in I_k$ となる k を $\hat{k}(x)$ で表す．この手続きで決められる μ_*

の推定値 $\hat{\mu}_{\hat{k}(x)}(x)$ を $\hat{\mu}(x)$ で表すことにする．AIC 最小化法は，I_k を陽に与えるものではないが，このクラスの推定法として代表的なものである．

$\hat{\mu}(x)$ の推定誤差について考えよう．モデル f_k が選ばれる確率 P_k は

$$P_k=\int_{I_k} g(x)dx$$

で，$\hat{\mu}(x)$ の期待値と真値 μ_* のまわりの分散はそれぞれ

$$\int \hat{\mu}(x)g(x)dx=\sum_k \int_{I_k} \hat{\mu}_k(x)g(x)dx$$

と

$$\int (\hat{\mu}(x)-\mu_*)^2\, g(x)dx=\sum_k \int_{I_k} (\hat{\mu}_k(x)-\mu_*)^2\, g(x)dx$$

で与えられる．

AIC 最小化法の簡単な例を見てみよう．n 個の独立な測定値からなるデータ $x=(x_1, \ldots, x_n)^T$ にもとづいて x_j $(1\leq j\leq n)$ の分布 $g(\cdot)$ の期待値 $\mu_*=\mu[g(\cdot)]$ を推定する問題である．2つのモデル

$$\text{MODEL(0)}:\ x_j\sim N(0,1)$$
$$\text{MODEL(1)}:\ x_j\sim N(\mu,1)$$

の AIC の値は，それぞれ，

$$\text{AIC}_0=n\log 2\pi+\|x\|^2$$
$$\text{AIC}_1=n\log 2\pi+\|x\|^2-n\bar{x}^2+2$$

となる．ただし $\bar{x}$ は $x_1, \ldots, x_n$ の標本平均，また $\|x\|^2=x_1^2+x_2^2+\cdots+x_n^2$ である．この例の場合の $\{I_k\}$ は

$$I_0=\{x\mid n\bar{x}^2<2\}$$
$$I_1=\{x\mid n\bar{x}^2\geq 2\}$$

である．$x\in I_1$ となって MODEL(1) が選ばれた場合には $\bar{x}$ を μ_* の推定値として採用し，$x\in I_0$ となって MODEL(0) が選ばれた場合には 0 と推定するのである．

x の真の分布が $N(\mu_*, 1)$ であるとすると，$n=100$ のデータから求めた $\bar{x}$ の分布は図1上段のようになり（図は $\mu_*=0.1$ として描いてある），MODEL(0) が選ばれる確率は図1の中段のグラフの面積

$$P_0(\mu_*)=\int_{100t^2<2} \frac{10}{\sqrt{2\pi}}\, e^{-100(t-\mu_*)^2/2}dt$$

で与えられ，μ_* の AIC 最小化法による推定値 $\hat{\mu}_{\text{MAICE}}$ は図1下段に図式的に示したような分布を持つことになる．原点の矢印は $P_0(\mu_*)$ の重みがここに集中していることを示す．

$\hat{\mu}_{\text{MAICE}}$ の μ_* のまわりの分散は

$$V_{\text{MAICE}}(\mu_*)=P_0(\mu_*)\mu_*^2+\int_{100t^2>2} (t-\mu_*)^2\frac{10}{\sqrt{2\pi}}\, e^{-100(t-\mu_*)^2/2}dt$$

で計算される．μ_* の変動にともなう $V_{\text{MAICE}}(\mu_*)$ の変化は図2のようになる．

単純に $\bar{x}$ を μ_* の推定値としてそのまま採用した時の推定量 $\hat{\mu}_{\text{MLE}}$ の μ_* のまわりの分散 $V_{\text{MLE}}(\mu_*)$ は μ_* によらずに

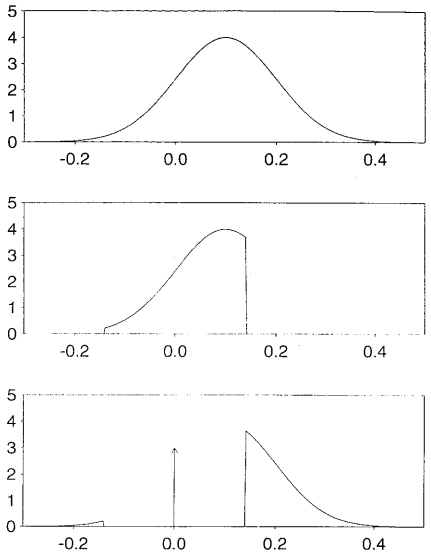

図1．$\bar{x}$ の分布と $\hat{\mu}_{\mathrm{MAICE}}$ の分布．

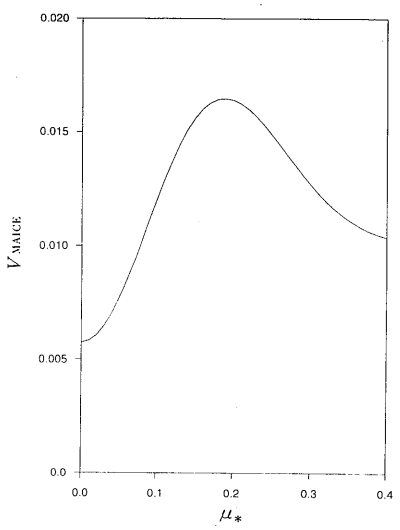

図2．$V_{\mathrm{MAICE}}(\mu_*)$．

$$V_{\mathrm{MLE}}(\mu_*)=0.01$$

であるから，$V_{\mathrm{MAICE}}(\mu_*)$ が $V_{\mathrm{MLE}}(\mu_*)$ よりよくなる場合があることがわかる．情報量規準を使ったモデル選択によって over-fitting を避けて推定誤差を小さくできる場合があるということである．単純ではあるが，情報量規準から（1）の立場での推定問題への貢献の典型的な例である．モデル選択という手法がかなり自由に使えるようになったことの効果は大きい．

この例は，しかしながら，（1）の場合の推定問題に情報量規準が難しい問題をもちこむ例にもなっている．モデル選択で MODEL(0) が選ばれて，0 が推定値として採用されたとき，その誤差はどの程度と見積るべきなのだろうか？ この問題は重要である．たとえば，医学的な量 μ_* が $|\mu_*|<\delta$ である事が健康であるか否かの判断に重要な意味を持っているものとしよう．$\hat{\mu}(x)=0$ である事は必ずしも $|\mu_*|<\delta$ を意味しない．この例の場合，図1を見て分るように $\hat{\mu}(x)=0$ がほぼ確実に $|\mu_*|<\delta$ を意味するのは $\delta>0.45$ の時である．この問題にたいする明確な答えが与えられたことは無いように思われる．ここでも結論をだすつもりは無いが，おそらく，

- MODEL(1) の誤差の見積りで代用するか，
- 図2のカーブの最大値を求めるか，
- μ_*に適当な事前分布を想定して図2のカーブの平均値を使うか，

のいずれかであろう．ここでとりあげたような単純な場合以外，2番目と3番目のような方法による見積りはかなりたいへんな仕事になるはずである．

3．パラメータの誤差評価の意味と情報量規準

この章では，第1章の（2）の場合におけるモデル $f_k(\cdot\,|\,\theta_k)$ のパラメータ θ_k の誤差評価について議論する．

パラメータの誤差評価のメリットの第1は，データ解析の成否の直観的な表現にある．パラメータの物理的な意味がはっきりしていて，その推定に関する要求が明確な場合，誤差の正確な見積りは解析の結果が満足すべきものであるか否かの絶対的な判断基準として使える．

パラメータの推定値に誤差幅をつけるもう一つの理由は，別の実験によるデータから得た結果との比較を可能にするためと考えられる．あるシステムのパラメータを継続的に推定して，システムに異常が起きたか否かを監視しようとするとき，新しい推定値が以前の推定値の「誤差幅」の中に落ちたのを見て安心するのは合理的である．

このような状況で，情報量規準を使うことによって，誤差幅がいらなくなるという考え方がある．もとのデータを残しておいて，そこに新しいデータをプールしてよいか否かを情報量規準で判断すれば，所期の目的を達成できるからというのがその論拠である．

ところで，最尤法によるパラメータの推定値の誤差の同時分布の分散共分散行列を

$$(*) \qquad V=-\left[\frac{\partial^2}{\partial\theta^2}\log f(x\mid\theta)\right]_{\hat\theta}^{-1}$$

で推定することが行なわれる．

最尤推定値 $\hat\theta$ と（＊）を将来のために記録にとどめておくということは

$$l'(\theta)=-\frac{1}{2}(\theta-\hat\theta)^T V^{-1}(\theta-\hat\theta)=\frac{1}{2}(\theta-\hat\theta)^T\left[\frac{\partial^2}{\partial\theta^2}\log f(x\mid\theta)\right]_{\hat\theta}(\theta-\hat\theta)$$

の関数形を記録にとどめておくことにほかならない．つまり，$\hat\theta$ を求める時に使った対数尤度関数の Taylor 展開の2次の項を残しておく事である．こう考えると誤差評価として（＊）を残しておくことの新しい役割が見えてくる．つまり，データが対数尤度関数の形で持っている情報を，ある程度形がくずれることを覚悟のうえで，縮約した形で保存しておくという役割である．

たとえば，データ x にもとづくパラメータの推定値 $\hat\theta(x)$ を

$$\log f(x\mid\theta)$$

の最大化で求めたとしよう．別のデータ y が x と同じ分布に従っていると考えてよいか否かを，情報量規準を使って解くには，θ の次元を k として

$$\mathrm{AIC}(x)=-2\log f(x\mid\hat\theta(x))+2k$$
$$\mathrm{AIC}(y)=-2\log f(y\mid\hat\theta(y))+2k$$
$$\mathrm{AIC}(x,y)=-2\{\log f(x\mid\hat\theta(x,y))+\log f(y\mid\hat\theta(x,y))\}+2k$$

を計算する．$\hat\theta(x,y)$ は

$$l(\theta\,;\,x,y)=\log f(x\mid\theta)+\log f(y\mid\theta)$$

の最大化で求める．

$$\mathrm{DAIC}(x,y)=\mathrm{AIC}(x)+\mathrm{AIC}(y)-\mathrm{AIC}(x,y)<0$$

なら x と y は異なる分布に従っていると考える．

さて，y を観測した時に，x が無くなっていたら，この解析はできない．しかし，$\hat\theta(x)$ と（＊），すなわち

$$l'(\theta\,;\,x)=\frac{1}{2}(\theta-\hat\theta(x))^T\left[\frac{\partial^2}{\partial\theta^2}\log f(x\mid\theta)\right]_{\hat\theta(x)}(\theta-\hat\theta(x))$$

が残っていれば，AIC(x) と AIC(x, y) を「疑似 AIC」

$$\mathrm{AIC}'(x)=2k$$
$$\mathrm{AIC}'(x, y)=-2l'(\hat{\theta}'(x, y))-2\log f(y\mid\hat{\theta}'(x, y))+2k$$

で代用し

$$\mathrm{DAIC}'(x, y)=\mathrm{AIC}'(x)+\mathrm{AIC}(y)-\mathrm{AIC}'(x, y)<0$$

なら，x と y は異なる分布に従っているとする方法が考えられる．ここで，$\hat{\theta}'(x, y)$ は

$$l'(\theta\,;\,x, y)=l'(\theta\,;\,x)+\log f(y\mid\theta)$$

の最大化で求めるのである．この方法を「最小疑似 AIC 法」と呼ぶことにしよう．

数値例. 回帰曲線の推定を考える．図 3 に示す曲線とそのまわりでのデータの散らばりの分散の推定が問題である．測定が 2 回行なわれた状況を想定する．図中では 2 回の測定結果が '1' と '2' で区別されている．実は '2' の方は図中の曲線（2 次曲線である）とはすこしずれた曲線のまわりにちらばらせてある．'1' と '2' はそれぞれ 30 個ある．

2 組のデータが残っていればそれらをプールすべきか否かを AIC 最小化で決められる．ここでは，この AIC 最小化法とデータ '1' に基づく 2 次回帰モデル

$$y=a_0+a_1x+a_2x^2+\varepsilon$$
$$\varepsilon\sim N(0, \sigma^2)$$

のパラメータ (σ^2, a_0, a_1, a_2) の推定値とその「誤差共分散」行列しか情報が残っていない場合を比較する実験を行なった．結果を表 1 にまとめてある．100 組のデータをシミュレーションで発生させて 2 つの方法によるプーリングに関する判定の結果を集計した分割表である．この表は，「疑似 AIC 最小化法」が AIC 最小化法とよく似たふるまいをすることを意味している．もともとの対数尤度関数が 2 次形式であれば，「疑似 AIC」を使ったモデル選択は本来の AIC を使ったものと全く同じになるはずである．分散 σ^2 の真値を既知とすれば対数尤度関数は 2 次形式になる．この例では，分散 σ^2 の推定を含めたために，対数尤度の 2 次近似が悪くなって表の非対角成分が 0 でなくなったと考えられる．

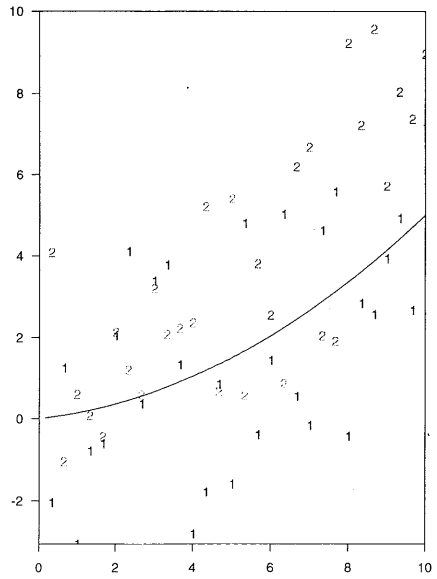

図 3. 2 組のデータの回帰分析.

この章の考えかたによると本質的なのは対数尤度関数の最大値を与える点における Hessian を残しておくことであって，その逆行列になんらかの，たとえば「誤差共分散」といった意味を付与する必要がなくなる．また，$f(x\mid\theta)$ が真の構造を捉えているか否かを気にする必要もなくなる．このような情報の保存が有効なのは対数尤度関数が 2 次近似できる場合である．そのような場合に，「誤差の見積り」が意味を持つと言って良いかもしれない．

対数尤度関数の 2 次近似が良いという条件は，

表1. 「最小疑似AIC法」とAIC最小化法の比較.

	$DAIC'(x, y)<0$	$DAIC'(x, y)\geq 0$
$DAIC(x, y)<0$	69	0
$DAIC(x, y)\geq 0$	4	27

かなりきついように見えるが，たとえば，最尤推定量の漸近正規性が成立する条件のもとで，データ数も十分にある場合には，少なくとも最尤推定量のゆらぎの範囲における対数尤度関数の2次近似は良いと考えられるから，この章の方法が有効であるための条件は見かけほどきついものではない.

もちろん，現在のように生データの保存が比較的容易な時代には情報の保存という役割をあまり重要視すべきではないだろう．それより，生データを保存しておいて積極的なモデリングが可能であるようにしておくことが重要である．しかし，古い論文などで，生データが失われて推定値とその誤差の見積りだけが与えられている場合に，前章に述べたような方法によって新しいデータによる結果と大雑把な比較をするといった使い道が考えられる.

4. 最 後 に

前章における

$$l'(\theta)+\log f(y\mid\theta)$$

の最大化によるθの推定は，形式的に，θに関する事前情報がベイズの意味の事前分布

$$C\exp\left\{\frac{1}{2}(\theta-\hat{\theta})^T\left[\frac{\partial^2}{\partial\theta^2}\log f(x\mid\theta)\right]_{\hat{\theta}}(\theta-\hat{\theta})\right\}$$

の形で与えられている場合の，事後分布のモードによるパラメータの推定にひとしい.

このことはベイズモデルと情報量規準の間に新しい接点があることを示唆しているのかも知れない．事前分布がデータと整合するか否かを情報量規準でチェックするということが出来ると面白い．今後の課題としたい.

謝 辞

この研究は，統計数理研究所の平成4年度の特定研究「データ解析支援システムの開発研究」の一部として行なわれた．癌研究会附属病院の松原敏樹博士と北里研究所の矢船明史博士による問題提起に負うところが多い．文中の図を描く道具を作るに当たっては統計数理研究所の丸山直昌博士に多くを御教示いただいた．Abstractの英文に関してはハワイ大学のWill Gersch博士のお世話になった．記して感謝の意を表する.

Proceedings of the Institute of Statistical Mathematics Vol. 41, No. 1, 25-31 (1993)

Error Estimation and Information Criterion

Makio Ishiguro

(The Institute of Statistical Mathematics)

In this paper, it is shown that the estimation errors of statistical inferences could be reduced by utilizing a model selection procedure based on Akaike Information Criterion, AIC. However the evaluation of the error of the Minimum AIC Estimate, MAICE, is difficult. The use of the error covariance matrix of MLE in the context of a statistical analysis using the AIC is discussed and a new "Quasi-AIC" is proposed.

Key words: AIC, MAICE, error estimation, model selection, MLE, information, Quasi-AIC.