モデル選択の理論

慶大理工 柴 田 里 程

1. AIC と ABIC

AIC (Akaike's Information Criterion) については，ここで新めて紹介するまでもなく，Akaike (1971) の FPE (Final Prediction Error) を萌芽として Akaike (1973) の論文から数えても，今年で既に 10 年余の歴史を持つ様になった．この間に実際問題への適用のみならず，理論的な面からの評価も進み，いろいろな意味でのよさも明らかになってきた．もはや，"新しい波" は大きなうねりとなって統計学の世界におしよせているといっても過言ではないであろう．

一方，ABIC (Akaike's Bayesian Information Criterion) は同じモデル選択の問題に対して提案されたものであるが，その独特な Bayesian framework と歴史の浅さ [Akaike (1979, 1980, 1981) および Akaike and Ishiguro (1980)] が相まって未だに "新しい波" である．更に，実際の場面で AIC と ABIC をどう使い分けたらいいのか，結果がどの様に違うのか，といった点について，あまりはっきりした解答は得られていない．

ここでは，Non-Bayesian の立場から，この二つを比較してみる事によって，少しでも ABIC の理解を深めてみよう．

2. 比 較

問題を "parametric family $\mathcal{F}=\{f(y,\ \theta);\ \theta\in\Theta\}$ が与えられたときの θ の推定" と設定しよう．ここで，$y=(y_1,\ \ldots,\ y_K)$ は K 次元 observation vector であり，Θ についても簡単の為に $\Theta=\boldsymbol{R}^K$ と仮定する．

よく知られている様に，Θ の次元 K が大きくなると $\mathcal{F}$ 全体での最尤推定量 (m.l.e.) の効率は悪くなる．そこで "パラメーター空間 Θ に何らかの制限を加えた上での m.l.e. をいくつか考え，そのうちから最も適当な推定量を 1 つ選んで最終的な推定量としよう" というのがモデル選択の発想である．

制限の加え方には，いろいろな形があり得るが，ここでは次の 2 通りを考えることにする．

a) $\Theta=\bigcup_j \Theta(j),\ \Theta(j)=\{\theta;\ \theta_l=0,\ l\in j\}$

ここで，$j=\{j_1,\ \ldots,\ j_k\}$ は $\{1,\ \ldots,\ K\}$ の部分集合である．

b) $\Theta=\bigcup_\xi \Theta(\xi),\ \Theta(\xi)=\{\theta;\ \theta_l^2\leqq\xi_l,\ l=1,\ \ldots,\ K\}$

ここで，$\xi=(\xi_1,\ \ldots,\ \xi_K)$ は非負実数ベクトルである．

最初の a) の場合の j は，ちょうど AIC での選択の対象となるモデル

$$\mathcal{F}_j=\{f(y,\ \theta);\ \theta\in\Theta(j)\}$$

を規定していて，このモデルのもとでの m.l.e. $\hat{\theta}(j)$ は，パラメーター空間を $\Theta(j)$ に制限した m.l.e. に外ならない．

次の b) の場合，$\Theta(\xi)$ 上での m.l.e. は

$$L_\lambda(\theta)=\log f(y,\ \theta)-\frac{1}{2}\sum_{l=1}^{K}\lambda_l\theta_l^2,\quad \lambda=(\lambda_1,\ \ldots,\ \lambda_K)$$

としたとき，

$$\frac{\partial L_\lambda(\theta)}{\partial\theta_l}=0,\ \lambda_l\geqq 0,\ \sum_l\lambda_l(\theta_l^2-\xi_l)=0\qquad l=1,\ \ldots,\ K$$

の解として与えられる．上の式を，θ を媒介にした ξ と λ の対応 $\xi=\xi(\lambda)$ を与えているとみなすと，b) で $\Theta(\xi)$ の代わりに

$$\Theta(\lambda)=\{\theta_l^2\leqq\xi_l(\lambda),\ l=1,\ \ldots,\ K\}$$

を考えても同等である．その時，

$$\mathcal{F}_\lambda=\{f(y,\ \theta);\ \theta\in\Theta(\lambda)\}$$

は λ によって規定される 1 つのモデルであり，このモデルのもとでの m.l.e. は $L_\lambda(\theta)$ を最大にする $\hat{\theta}(\lambda)$ となる．

この様に考えると，AIC も ABIC も，それぞれ j,λ を選択する基準を与えるという共通点をもっている．それらの基準を個々の θ の値に依存しない様に構成している点でも共通しているが，一つの大きな違いは，j は離散量であるのに対し λ は連続量である点であろう．対比をはっきりさせるため，a), b) それぞれの場合に応じて，次の様な記号を導入しよう．

a)　$\mathrm{AIC}^\circ=-2\log\{f(y,\ \theta)I_{\Theta(j)}(\theta)\}$

$$\mathrm{AIC}_\alpha=-2\log\left[\max_\theta\{f(y,\ \theta)I_{\Theta(j)}(\theta)\}\times\exp\left\{-\frac{\alpha}{2}\dim(\Theta(j))\right\}\right]$$

ここで，$I_{\Theta(j)}(\theta)$ は $\Theta(j)$ の indicator function であり，$\dim(\Theta(j))$ は $\Theta(j)$ の次元である．又，α はあらかじめ定められた正数で，$\alpha=2$ なら AIC そのものである．

b)　$\mathrm{ABIC}^\circ=-2\log\{f(y,\ \theta)\Pi_\lambda(\theta)\}$

$$\mathrm{ABIC}=-2\log\left\{\int_\Theta f(y,\ \theta)\Pi_\lambda(\theta)d\theta\right\}$$

但し，

$$\Pi_\lambda(\theta)=\prod_{l=1}^{K}\left(\frac{\lambda_l}{2\pi}\right)^{1/2}\exp\left(-\frac{1}{2}\sum_l\lambda_l\theta_l^2\right),\quad \lambda=(\lambda_1\ldots\lambda_K),\ \lambda_l>0,\ l=1,\ \ldots,\ K\ \text{である．}$$

まず，AIC° を最小化して得られる θ の推定量が $\hat{\theta}(j)$ であることは明らかである．又，

$$\log\Pi_\lambda(\theta)=\frac{1}{2}\sum_l\log\left(\frac{\lambda_l}{2\pi}\right)-\frac{1}{2}\sum_l\lambda_l\theta_l^2\tag{2.1}$$

であるので，ABIC° を θ に関して最小化することと，先の $L_\lambda(\theta)$ を最大化する事は同等で，$\hat{\theta}(\lambda)$ が従う．$\Pi_\lambda(\theta)$ を一つの prior とみなすならば，$\hat{\theta}(\lambda)$ はこの prior のもとでの Bayes estimate になっている．これが Bayesian Criterion と名づけられたゆえんである．

次に，j あるいは λ の選択はそれぞれ AIC_α, ABIC を最小化する事によって得られる．いずれの基準も，推定量を導く時に使った AIC° あるいは ABIC° にもとづいて導かれている．個々の θ の値に依存しない量にするために，AIC_α では，最大値に補正項をかけて構成しているが，ABIC では，最大値ではなく，θ に関して積分することによって θ の値に無関係な量を得てい

る．ここで，最初に $L_\lambda(\theta)$ ではなく ABIC° を考えた効果は，(2.1) の右辺の第1項で，これが各 λ_l に関して単調増加である点である．この項がないと，ABIC の最小値が存在しなくなる．

又，再び Bayes の枠組で ABIC を解釈するならば，よく知られている様に，

$$-2\log(\text{marginal likelihood of }\lambda)$$

と解釈できる．

さて，ABIC° から ABIC を導くのに，AIC° から AIC_α を導いた時とのアナロジーでいえば，

$$-2\log[\max_\theta\{f(y,\theta)\Pi_\lambda(\theta)\}g(\lambda)]$$

の形の選択基準が考えられる．ここで，$g(\lambda)$ は AIC_α における補正項に対応し y に依存してもよい．$g(\lambda)$ の決め方としては，いろいろな基準があり得るが ABIC° の期待値の不偏推定になる様に決めるのも一つである．

しかし，a)，b) の場合の大きな違いは，先に述べた様に，j は離散量で，λ は連続量である点で，この様なアナロジーがうまく働くかどうかは保証の限りではない．今後の研究課題であろう．

3. 関 連 性

ここでは，次の様な簡単な例で，AIC と ABIC の関連を，もう少し詳しく調べてみよう．観測ベクトル $y=(y_1,\ldots,y_K)$ の各 y_l は $N(\theta_l,\sigma^2)$, $l=1,\ldots,K$ に従うものとする．又，σ^2 は既知とし便宜的に λ_l を λ_l/σ^2 におきかえて

$$\Pi_\lambda(\theta)=\prod_l\left(\frac{\lambda_l}{2\pi\sigma^2}\right)^{\frac{1}{2}}\exp\left(-\frac{1}{2}\sum_l\frac{\lambda_l\theta_l^2}{\sigma^2}\right)$$

と変形しておく．そうすると，この場合

a) $\hat\theta_l(j)=\begin{cases}0 & l\notin j\\ y_l & l\in j\end{cases}$, $l=1,\ldots,K$

b) $\hat\theta_l(\lambda)=\dfrac{y_l}{1+\lambda_l}$, $l=1,\ldots,K$

となることは，すぐわかる．

さて，AIC_α は

$$\text{AIC}_\alpha=\frac{1}{\sigma^2}\sum_l(y_l-\hat\theta_l(j))^2+\alpha\dim(\Theta(j))+K\log 2\pi\sigma^2=\frac{1}{\sigma^2}\sum_{l\in j}y_l^2+\alpha\dim(\Theta(j))+K\log 2\pi\sigma^2$$

となるので，j の選択は

$$\hat j=\{l\,;\,1\leqq l\leqq K,\ y_l^2>\alpha\sigma^2\}$$

となる．一方 ABIC は

$$\text{ABIC}=\sum_l\left\{\frac{\lambda_l}{1+\lambda_l}\frac{y_l^2}{\sigma^2}-\log\left(\frac{\lambda_l}{1+\lambda_l}\right)+\log 2\pi\sigma^2\right\}\tag{3.1}$$

となり，$\dfrac{\partial}{\partial\lambda_l}\text{ABIC}=\dfrac{\lambda_l(y_l^2-\sigma^2)-\sigma^2}{(1+\lambda_l)^2\lambda_l\sigma^2}$ であるので，$y_l^2\leqq\sigma^2$ のとき λ_l に関して単調減少，$y_l^2>\sigma^2$ のとき $\lambda_l=\lambda_l^*=\dfrac{\sigma^2}{y_l^2-\sigma^2}$ で最小になる．従って，λ の選択 $\hat\lambda=(\hat\lambda_1,\ldots,\hat\lambda_K)$ は

$$\hat{\lambda}_l=\begin{cases}\infty & y_l^2\leqq\sigma^2 \text{ の時}\\ \lambda_l^* & y_l^2>\sigma^2 \text{ の時}\end{cases}$$

である．

結局，選択を行なった後の推定量は

a)　$\hat{\theta}_l(\hat{j})=\begin{cases}0 & y_l^2\leqq\alpha\sigma^2 \text{ の時}\\ y_l & y_l^2>\alpha\sigma^2 \text{ の時}\end{cases}$

b)　$\hat{\theta}_l(\hat{\lambda})=\begin{cases}0 & y_l^2\leqq\sigma^2 \text{ の時}\\ \dfrac{y_l}{1+\lambda_l^*} & y_l^2>\sigma^2 \text{ の時}\end{cases}$

になり，いずれも一種の縮小推定量である．縮小量は，a) の場合，不連続で，b) の場合は連続であるが，次の図の様に後者は前者を補間したものとみなすこともできる．

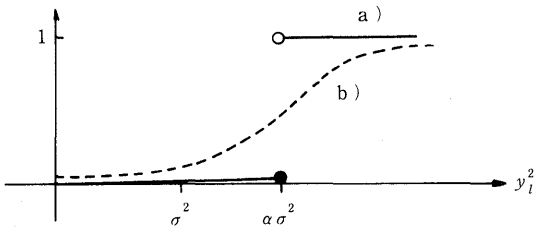

図 1.

ところで，ABIC は全ての λ_l が正である様な λ についてだけ定義されているが，これを 0 である λ_l も許す様に拡張したらどうなるであろうか？

$\lambda_l=0$ は θ_l に関して何らの制限もおかない事を意味するので，この場合 (3.1) の対応する項を

$$\mathrm{E}(-2\log f(y_l,\ \theta_l))=1+\log 2\pi\sigma^2$$

でおきかえてみよう．ここで $f(y_l, \theta_l)$ は平均 θ_l, 分散 σ^2 の正規密度である．すると $L=\{l\ ;\ \lambda_l>0\}$ として (3.1) は

$$\mathrm{ABIC}=\sum_{l\in L}\left(\frac{\lambda_l}{1+\lambda_l}\frac{y_l^2}{\sigma^2}-\log\left(\frac{\lambda_l}{1+\lambda_l}\right)+\log 2\pi\sigma^2\right)+\sum_{l\notin L}(1+\log 2\pi\sigma^2)$$

に拡張される．右辺の最初の和の各項は，$y_l^2\leqq\sigma^2$ の時，λ_l に関して単調減少で $y_l^2/\sigma^2+\log 2\pi\sigma^2$ に近づき，これは $1+\log 2\pi\sigma^2$ 以下なので，$\hat{\lambda}_l=\infty$ である．又，$y_l^2>\sigma^2$ の時は最小値は $1+\log\frac{y^2}{\sigma^2}+\log 2\pi\sigma^2$ で，$1+\log 2\pi\sigma^2$ より大きいので，$\hat{\lambda}_l=0$ とした方が ABIC は小さくなる．結局

$$\hat{\lambda}_l=\begin{cases}\infty & y_l^2\leqq\sigma^2 \text{ のとき}\\ 0 & y_l^2>\sigma^2 \text{ のとき}\end{cases}$$

で，最終的な推定量は

$$\hat{\theta}_l(\hat{\lambda})=\begin{cases}0 & y_l^2\leqq\sigma^2 \text{ のとき}\\ y_l & y_l^2>\sigma^2 \text{ のとき}\end{cases}$$

となり，これは $\alpha=1$ のときの $\hat{\theta}(\hat{j})$ と一致する．

従って，全て $\lambda_l>0$ な λ だけを許せば，縮小量は連続的であるが，$\lambda_l=0$ に対しても上の様に ABIC を定義すると，最終的な推定量は AIC_1 で選択したときの推定量と一致し，縮小量は不連続になることがわかった．

更に y がくり返し観測されるならば σ^2 の推定が可能で，各モデルのもとで σ^2 の値域に制限をおかなければ，$\tilde{\sigma}^2(j)$, $\tilde{\sigma}^2(\lambda)$ 共に full model $\mathcal{F}$ のもとでの最尤推定量 $\tilde{\sigma}^2$ になり，上の議論は σ^2 を $\tilde{\sigma}^2$, y_l^2 を $\overline{y_l^2}$ でおきかえてそのまま成り立つ．

参 考 文 献

Akaike, H. (1971). "Autoregressive model fitting for control," *Ann. Inst. Statist. Math.,* Vol. **23**, pp. 163-180.

Akaike, H. (1973). "Information theory and an extension of the maximum likelihood principle," pp. 267-281 in *2nd Int. Symposium on Information Theory,* ed. Petrov B.N. and Csaki, F., Akademia Kiado, Budapest.

Akaike, H. (1979). "Likelihood and the Bayes procedure," in *Trabajos de Estatistica.*

Akaike, H. and Ishiguro, M. (1980). "BAYSEA, A Bayesian seasonal adjustment program," *Computer Science Monographs,* **13**, Institute of Statistical Mathematics, Tokyo.

Akaike, H. (1980). "Seasonal adjustment by a Bayesian modeling," *J. Time Series Analysis,* Vol. **1**, pp. 1-13.

Akaike, H. (1981). "Likelihood of a model and information criteria," *J. Econometrics,* Vol. **16**, pp. 3-14.

討 論

講演後の討論では，まず AIC, ABIC の生みの親である赤池弘次氏から，"ABIC の定義を歪曲している" と次の様な反論があった．

まず ABIC は Bayesian の枠組で提案されたものであるから，問題の設定は "θ の推定" ではなくて，"posterior $p(\theta\mid y)$ の推定" である．

$$p(\theta\mid y)=\frac{f(y,\theta)\Pi_\lambda(\theta)}{\int f(y,\theta)\Pi_\lambda(\theta)d\theta}$$

に注意すれば，結局 "posterior の推定" は "λ の推定" である．

従って，$\mathrm{AIC}^\circ\xrightarrow[\theta]{}\min$ に対応するのは，$\mathrm{ABIC}\xrightarrow[\lambda]{}\min$ であり，$\mathrm{AIC}\xrightarrow[j]{}\min$ に対応するのは ABIC のもう一つ先の段階である $\mathrm{ABIC}+2\dim(\Lambda)\xrightarrow[\dim(\Lambda)]{}\min$ である．

この質問に対する筆者の答は，$\dim(\Lambda)$ を選択することも考えるならばこれもモデル選択の段階に含まれ，相変らず，推定量を導く段階と，モデル選択の段階の二つから成ることに変わりはない．そして "θ の推定" という問題設定のもとでは，あいかわらず

$$\mathrm{AIC}^\circ\xrightarrow[\theta]{}\min \qquad \mathrm{ABIC}^\circ\xrightarrow[\theta]{}\min$$
$$\mathrm{AIC}_a\xrightarrow[j]{}\min \qquad \mathrm{ABIC}+2\dim(\Lambda)\xrightarrow[\substack{\dim(\Lambda)\\ \lambda\in\Lambda}]{}\min$$

の対応は成り立つ．

もう一つの質問は，第2節で，AIC° から AIC_a を導いたアナロジーで，ABIC° から ABIC を最大値に基いて導いたらどうか？ という提案についてで，"安定性が悪い" という指摘であった．

この点に関しては，$g(\lambda)$をどう決めるかの問題とも関連し，今の段階では何とも言えない．今後に残された問題としておきたい．

ベイズ型コウホートモデル

統計数理研究所　中　村　　隆

1. コウホート分析

社会の変化を分析するとき，時勢による推移だけでなく，加齢による変化や世代による相違を区別することが重要である．コウホート分析は，継続的調査で得られる何らかの数量特性を年齢層×調査時点別に集計したデータ表から，年齢と時代効果，そして表の斜め方向に現われるコウホート（同時出生集団）の効果を分離しようとする方法である．

ただし，コウホート分析には識別問題が存在し，三効果を一義に分解することは原理的に不可能であるとされていた（Fienberg and Mason, 1979）．これに対し，中村（1982）は，識別問題を克服するためにパラメータの漸進的変化の条件を取り込んだベイズ型コウホートモデルを提案し，種々のコウホートデータに適用している．

2. コウホート表

コウホート分析の分析対象であるデータ表をコウホート表と呼ぶ．コウホート表は，まず，集計されたデータの性質にしたがって，比率型と数量型に分類できる．比率型は表のセルに回答比率を配したもの，数量型は平均子ども数のような数量を配したものである．それぞれ異なる標本抽出モデルを仮定する．

さらに，コウホート表はその形式にしたがって，標準表・一般表・特殊表に分けることができる．標準表とは，年齢×時代の表で調査間隔と年齢区分幅が一致しているものであり，同一コウホートが表の斜めのセルを渡るので分析が比較的容易である．一般表とは，調査間隔と年齢区分幅が一致しない場合である．標準表と一般表が継続調査で得られる年齢×時代の形式であるのに対し，特殊表はそれ以外の形式の表であり，一時点の回顧調査で得られるコウホート×年齢の表などがある．

3. コウホートモデル

ある数量特性x_{ij}を年齢×時代の形式に配するコウホート表データを考える．これに既知の関数gを用いて変換を施した数量y_{ij}を次のように分解する：

$$y_{ij}\equiv g(x_{ij})\cong\mu+\mu_i^A+\mu_j^P+\mu_k^C, \tag{1}$$

$$\sum\mu_i^A=\sum\mu_j^P=\sum\mu_k^C=0. \tag{2}$$

ここで，$\mu_i^A, \mu_j^P, \mu_k^C$はそれぞれ年齢・時代・コウホート効果のパラメータであり，μは総平均効果である．

(1)式は(2)式を考慮して次のように書くことができる：

$$\boldsymbol{y}=D\boldsymbol{\mu}. \tag{3}$$