統計数理 (1993)
第41巻 第2号 131-147

モデルの信頼集合と地図によるモデル探索

東京大学* 下 平 英 寿

(1993年11月 受付)

1. はじめに

AICに基づくモデル比較は，広く応用され成果をあげている．その中でも代表的な，AIC最小化法によるモデル選択は，候補となるモデルについてAICの値を計算し，その値を最小にするモデルをひとつ選ぶものである．しかしモデル探索の初期の段階では，AICを比較的小さくするモデルを複数選ぶ方がよい．実際，AICの第1項に相当する対数尤度は大きな分散をもつので，データ数に対してモデルの候補の数が多い場合には，AIC最小のモデルがその他のモデルに対して有意に良いといえなくなる．すなわち，AICの差が，その標準誤差に対して有意に大きくなければ，モデル選択は信頼できないものとなる．

本稿ではAICの代わりに，より精密な，竹内 (1976) によるTICを用いる．これについて2章で述べる．3章では，AICの差の分散について調べる．4章では，その分散の推定量を用い，多重比較の方法を利用して，モデルの信頼集合を構成する．モデル信頼集合とは，候補となるモデルのうち一番良いものを，与えられた有意水準で含むような集合である．信頼集合の計算結果は，各モデルの P-値として与えられる．ただし，モデルの P-値とは，そのモデルが信頼集合に含まれるような有意水準の最大値である．実際の P-値の計算では近似を行なっているので，得られる結果は名目上の P-値である．

ここで行なおうとしているモデル選択では，まず，候補となる各モデルの P-値を計算する．そしてモデルを P-値の大きい順に並べ，与えられた有意水準以上の P-値をもつモデルを，モデル信頼集合の要素として選ぶ．モデルの候補の数に比べてデータ数が少ない場合，モデル信頼集合は大きくなる傾向がある．これは，データ数が十分でない場合，AICの値はあまり信頼できなくなることに対応する．このような場合，モデル信頼集合にどういう傾向のモデルが含まれているかを調べることが重要になる．

5章では，モデル信頼集合の傾向を調べるために，モデル間の距離を定義し，モデル地図を描く．モデル間距離としては，AICの差の分散（の推定値）の平方根を用いる．これは真の分布に依存して変わるので，「その真の分布から見た」地図（の推定）が描かれることになる．この地図を見ることによって，各モデル毎に推定された分布の，確率分布の空間における相対的な位置関係を把握できる．これにより，真の分布がどういうモデルに近いかといったことについて，示唆を得ることができる．6章では，重回帰分析の変数選択問題について，いくつかの数値例をあげる．

* 計数工学科：〒113 文京区本郷7-3-1.

2. 情報量規準

実ベクトル x を確率変数とし，x に関する二つの確率密度関数 q と p の遠さをあらわすための規準として，$D(q, p)=\int q(x)d(x, p)dx$ を用いる．例えば，$d(x, p)=-\log p(x)$ とおくと，$p(x)$ によらない定数項 $H(q)=-\int q(x)\log q(x)dx$ を除いて，$D(q, p)$ は Kullback-Leibler (K-L) の情報量になる．モデルを比較する時は，$D(q, p)$ の差をとってから用いるので，$H(q)$ の分の違いはモデル選択には影響しない．確率変数 x は真の密度関数 $q(x)$ に従い，その独立な観測値 $x_1, \ldots, x_n$ が得られたとする．n をデータ数とする．経験分布を $\hat{q}(x)=(1/n)\sum_{t=1}^{n}\delta(x-x_t)$ と書く（図1）．

本稿で「モデル」とは，実ベクトル θ を母数とする密度関数 $p(x\mid\theta)$ のこととする．Θ を適当な開集合とし，$\theta\in\Theta$ とする．本稿を通して，モデルに関する適当な正則条件は満たされているものとする．一つのモデルを与えた時，最適な母数を $\theta^*=\arg\inf_{\theta\in\Theta}D(q, p(\theta))$ で定義する．本稿では真の分布 q がモデル $p(\cdot)$ に含まれることを仮定しないので，一般に，$p^*(x)=p(x\mid\theta^*)$ と $q(x)$ は一致しない．$\hat{\theta}=\arg\inf_{\theta\in\Theta}D(\hat{q}, p(\theta))$ は θ^* の推定量になる．特に，K-L 情報量の時は，最尤推定になっている．推定量 $\hat{\theta}$ は最適母数 θ^* の一致推定量で，次のように漸近正規分布していると仮定する．$\hat{\theta}\sim AN(\theta^*, (1/n)H^{*-1}G^*H^{*-1})$．ただし，行列 G^* と H^* の成分を $g_{ij}^*=E\{\partial_i d(x, \theta^*)\partial_j d(x, \theta^*)\}$, $h_{ij}^*=E\{\partial_i\partial_j d(x, \theta^*)\}$ とする．ここで $\partial_i=\partial/\partial\theta_i$ である．本稿を通して，この漸近正規性を仮定する．

モデル $p(\cdot)$ の良さを，推定した分布 $\hat{p}(x)=p(x\mid\hat{\theta})$ が $q(x)$ から平均的にどれだけ遠いかで定義する．つまり，

$$\text{risk}\,(q, p(\cdot))=E\{D(q, p(\hat{\theta}))\}$$

をモデルの良さとして用いる．$\hat{\theta}$ は観測データ $x_1, \ldots, x_n$ の関数であるから，確率変数であることに注意する．実際には真の分布 q を知らないので，この risk を計算することはできない．そこで，観測データ $\hat{q}(x)$ より計算できる risk の推定量として情報量規準

$$\text{TIC}(\hat{q}, p(\cdot))=D(\hat{q}, \hat{p})+(1/n)\text{tr}G^*H^{*-1}$$

を用いる．実際には第2項 $\text{tr}\,G^*H^{*-1}$ を適当な一致推定量で置き換えて計算する．例えば，$\text{tr}\,\hat{G}\hat{H}^{-1}$ などで推定する．ただし，行列 $\hat{G}$ と $\hat{H}$ の各成分を，$\hat{g}_{ij}=\hat{E}\{\partial_i d(x, \hat{\theta})\partial_j d(x, \hat{\theta})\}$, $\hat{h}_{ij}=\hat{E}\{\partial_i\partial_j d(x, \hat{\theta})\}$ とする．ここで，$\hat{E}(\cdot)$ と $\hat{V}(\cdot)$ は経験分布 $\hat{q}(x)$ を用いた期待値と分散を表

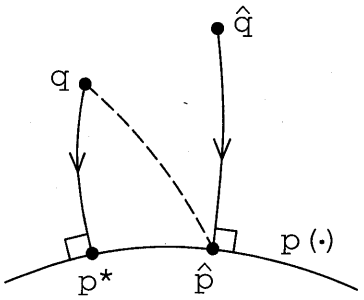

図1．確率分布の空間の幾何学的関係．q が真の分布，$\hat{q}$ が経験分布，p^* が最適な分布，$\hat{p}$ が実際の推定，$\text{risk}(q, p(\cdot))=E\{D(q, \hat{p})\}$, $\text{TIC}(\hat{q}, p(\cdot))=D(\hat{q}, \hat{p})+(1/n)\text{tr}G^*H^{*-1}$.

す．このように，tr $\hat{G}\hat{H}^{-1}$ を使って第2項を推定したものは，一般に情報量規準TIC (Takeuchi's modification of AIC) と呼ばれるものに等価である（竹内 (1976, 1983)). 本稿で扱う漸近的な議論では，tr $\hat{G}\hat{H}^{-1}-$tr $G^*H^{*-1}=O_p(n^{-0.5})$ の違いは影響せず，高次の誤差項に含まれてしまう．従って，本稿で与える TIC に関する関係式は，その第2項をどちらの定義で与えても成り立つ．

また，$q(x)$ が特定の分布族やモデルに含まれることを仮定するとさらに良い推定量を構成できる．たとえば，$D(q, p)$ として K-L 情報量を用いた時，$q(x)$ が $p(x \mid \theta^*)$ に十分近いという仮定をおくと，$G^* \approx H^*$ と近似でき，結果として，AIC $(\hat{q}, p(\cdot))=D(\hat{q}, \hat{p})+(1/n)\dim\theta$ を得る．これは一般に情報量規準 AIC (Akaike's information criterion) と呼ばれているものに等価である (Akaike (1974)). AIC の方が TIC より情報量規準の第2項 tr G^*H^{*-1} を効率良く推定しているが，実際には $G^*=H^*$ とおくための仮定が厳密になり立つことはなく，AIC は bias をもって推定することになるので，注意が必要である．いずれにしても，tr G^*H^{*-1} の推定の効率は，本稿での漸近的な議論には影響しない．実際の計算ではデータ数 n は有限なので，分散の小さい AIC を使うか，bias の小さい TIC を使うかを，場合に応じて考える必要があろう．以下の議論では，$q(x)$ に関して特別な仮定を設けないので，AIC ではなく TIC を用いる．

さて，TIC は risk の推定量である．すなわち，漸近的に，

$$E\{\mathrm{TIC}(\hat{q}, p(\cdot))\}=\mathrm{risk}(q, p(\cdot))+O(n^{-1.5})$$

がいえる．このように，TIC の期待値は，$O(n^{-1})$ の項まで risk に等しいが，TIC の値そのものは次式のようにあまり良い推定とはいえない．

$$\mathrm{TIC}(\hat{q}, p(\cdot))=\mathrm{risk}(q, p(\cdot))+O_p(n^{-0.5})$$

つまり，かなり分散の大きな推定量である．

3. 分散の評価

二つのモデル $p_1(x \mid \theta_1)$ と $p_2(x \mid \theta_2)$ があった時，$\triangle\mathrm{risk}=\mathrm{risk}(q, p_1(\cdot))-\mathrm{risk}(q, p_2(\cdot))>0$ なら，モデル2はモデル1より良いことになる．実際には $\triangle$risk は計算できないので，その推定量 $\triangle\mathrm{TIC}=\mathrm{TIC}(\hat{q}, p_1(\cdot))-\mathrm{TIC}(\hat{q}, p_2(\cdot))$ を用いてどちらのモデルが良いかを判断する．もし，$\triangle$TIC がその分散 $V\{\triangle\mathrm{TIC}\}$ に対して十分に大きくなければ，二つのモデルに有意な差はないといえる（図2). 本章では，$V\{\triangle\mathrm{TIC}\}$ とその推定量について，Kishino and Hasegawa (1989) や Shimodaira (1993a) で得られている結果について述べる．証明については付録を参照．

定理1. データ数 n が十分大きい時，漸近的に，

$$V\{\triangle\mathrm{TIC}\}=V\{e^*(x)\}/n+O(n^{-2})$$

である．ただし，$e^*(x)=d(x, p_1^*)-d(x, p_2^*)$ とおく．

系1. $\hat{V}\{\hat{e}(x)\}/n$ は，$V\{\triangle\mathrm{TIC}\}$ の推定量であり，漸近的に，

$$\hat{V}\{\hat{e}(x)\}/n=V\{\triangle\mathrm{TIC}\}+O_p(n^{-1.5})$$

である．ただし，$\hat{e}(x)=d(x, \hat{p}_1)-d(x, \hat{p}_2)$ とおく．

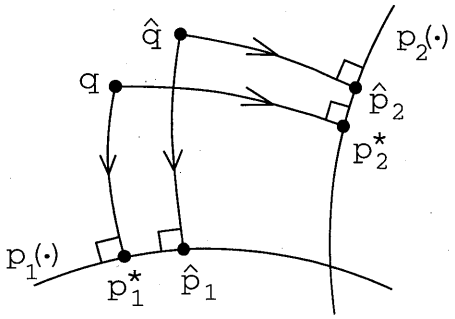

図2. 確率分布の空間の幾何学的関係. この図の場合, △risk<0 であり, モデル1の方がモデル2より良い. ところが, △TIC>0 であり, TIC 最小化法では, 実際と逆の結論が得られる. この場合, △TICの分散は比較的大きいと予想される. 本稿の方法では, このような場合, モデル1とモデル2の良さに有意な差はないと結論される.

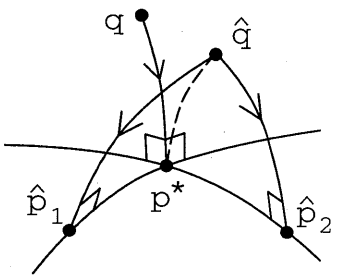

図3. 確率分布の空間の幾何学的関係. この図の場合, $p_1^*=p_2^*=p^*$ である. $O_p(n^{-1.5})$ を無視すると, $D(\hat{q}, \hat{p}_1)=D(\hat{q}, p^*)-D(\hat{p}_1, p^*)$ と書ける. 従って $O(n^{-2.5})$ を無視すると, $V\{\triangle \mathrm{TIC}\}=V\{D(\hat{q}, \hat{p}_1)-D(\hat{q}, \hat{p}_2)\}=V\{D(\hat{p}_1, p^*)-D(\hat{p}_2, p^*)\}$ とみなせて, これは, $V\{D(\hat{p}_1, p^*)\}+V\{D(\hat{p}_2, p^*)\}-2\mathrm{Cov}\{D(\hat{p}_1, p^*),\ D(\hat{p}_2, p^*)\}$ となる. この3つの項は, それぞれ, 定理3の右辺の3つの項に対応している.

従って $\hat{V}\{\hat{e}(x)\}/n$ は $V\{\triangle \mathrm{TIC}\}$ の推定量として使える. 6章では, 真の分布の正規性の仮定の下に, $V\{e^*(x)\}/n$ を推定してこれを用いている.

ところで, $V\{e^*\}$ は各々のモデルについての最適な分布 p_1^* と p_2^* の間の一種の距離とみなせる. 実際, $d_{ij}=\sqrt{V\{d(x, p_i^*)-d(x, p_j^*)\}}$ とおくと, 分散の性質から, d_{ij} は距離の公理をみたす. さらに, 確率分布の空間での K-L 情報量と次のような関係がある.

定理2. $D(q, p)$ は K-L 情報量とする. $D(q, p_1^*)$ と $D(q, p_2^*)$ が十分に小さい時,

$$V\{e^*\}=J+O(\| D(q, p_1^*)^{1/2}, D(q, p_2^*)^{1/2} \|^3)$$

ただし, $J=D(p_1^*, p_2^*)+D(p_2^*, p_1^*)$ とする.

このように, $D(q, p)$ として K-L 情報量を用いていて, 3つの分布 p_1^*, p_2^*, q が十分に近いと仮定できる時は, $V\{e^*\}\approx D(p_1^*, p_2^*)+D(p_2^*, p_1^*)$ と近似できる. この解釈に基づいて, 5章では, モデル地図を構成する.

さて, 二つのモデルの最適な分布 p_1^* と p_2^* が一致してしまう場合, $V\{e^*\}=0$ であり, $V\{\triangle \mathrm{TIC}\}=O(n^{-2})$ となる. この場合の結果を証明なしで述べる (図3).

定理3. $p_1^*=p_2^*$ とする. このとき,

$$V\{\triangle \mathrm{TIC}\}=\frac{1}{2n^2}(\mathrm{tr}\ G_1^* H_1^{*-1} G_1^* H_1^{*-1}+\mathrm{tr}\ G_2^* H_2^{*-1} G_2^* H_2^{*-1} -2\mathrm{tr}\ G_{12}^* H_2^{*-1} G_{21}^* H_1^{*-1})+O(n^{-2.5})$$

ただし, G_{12}^* の各成分は, $g_{12ij}^*=E\{\partial_i d_1(x, \theta_1^*)\partial_j d_2(x, \theta_2^*)\}$ とする.

定理 4. $p_1^*=p_2^*$ とする．このとき，

$$E\{\hat{V}\{\hat{e}\}/n\}=2V\{\triangle\mathrm{TIC}\}+O(n^{-2.5})$$
$$V\{\hat{V}\{\hat{e}\}/n\}=O(n^{-4})$$

である．

$p_1^*=p_2^*$ の場合について考える．この時，$\hat{V}\{\hat{e}\}/n$ によって $V\{\triangle\mathrm{TIC}\}$ を推定した場合，期待値としては 2 倍になる程度でそれほど悪いわけではないが，分散が非常に大きく，標準誤差は真値と同じオーダである．したがって，多重比較に用いる統計量 $\triangle\mathrm{TIC}/\sqrt{\hat{V}\{\hat{e}\}/n}$ は，かなり不安定で，分母が非常に小さくなった時とても大きな値をとり得る．これを安定化するために，$V\{\triangle\mathrm{TIC}\}$ の推定量として，

$$\frac{1}{n}\hat{V}\{\hat{e}\}+\frac{1}{2n^2}(\mathrm{tr}\,\hat{G}_1\hat{H}_1^{-1}\hat{G}_1\hat{H}_1^{-1}+\mathrm{tr}\,\hat{G}_2\hat{H}_2^{-1}\hat{G}_2\hat{H}_2^{-1}-2\mathrm{tr}\,\hat{G}_{12}\hat{H}_2^{-1}\hat{G}_{21}\hat{H}_1^{-1}) \tag{3.1}$$

を本稿では用いた．追加したオーダ $O(n^{-2})$ の項は，$V\{\triangle\mathrm{TIC}\}$ を consistent に推定する量である．また $\hat{V}\{\hat{e}\}/n>0$ なので，式 (3.1) は $V\{\triangle\mathrm{TIC}\}$ を大きめに見積もる推定量になる．したがって，これに基づいて構成するモデル選択の信頼集合は，保守的になる．

一方，$p_1^*\neq p_2^*$ の時は，式 (3.1) は $V\{\triangle\mathrm{TIC}\}$ を $O(n^{-1})$ まで consistent に推定する．一般に $p_1^*\neq p_2^*$ の時，式 (3.1) の第 2 項は，$V\{\triangle\mathrm{TIC}\}$ の $O(n^{-2})$ の項の consistent な推定量になるわけではない．実際にはこの他の項がある．しかしいずれにしても，系 1 で与えられているように，$\hat{V}\{\hat{e}(x)\}/n$ は $O_p(n^{-1.5})$ のゆらぎがあるので，$O(n^{-2})$ の項を consistent に評価しようとするのは漸近的には意味がない．実際には n は有限なので，p_1^* と p_2^* が非常に近い場合にモデル選択の結果が不安定になるのを避けるために，本稿では式 (3.1) を用いた．

実際に (3.1) の第 2 項（これを $V_{2\mathrm{nd}}$ とおく）を計算するのは手間がかかる．そこで，$G_1^*=H_1^*$, $G_2^*=H_2^*$ と仮定できる場合について，この項を大雑把に見積もる方法について述べる．$m_1=\dim\theta_1$, $m_2=\dim\theta_2$ とおくと，ある $m_c\leq\min(m_1, m_2)$ が存在して，

$$|m_1-m_2|\leq 2n^2V_{2\mathrm{nd}}\leq(m_1-m_c)+(m_2-m_c)\leq m_1+m_2$$

がいえる．$p_1^*=p_2^*$ の時には，m_c は二つのモデルが共通に含むモデルの次元である．特に，モデル 1 がモデル 2 を含む場合には，$2n^2V_{2\mathrm{nd}}=m_1-m_2$ となる．実際に $V_{2\mathrm{nd}}$ を計算するのが面倒な時には，このように上限を見積もって使えば十分だろう．

4. モデル信頼集合

モデルの候補を $N=\{1, 2, \ldots, l\}$ とする．この l 個のモデルの TIC と risk を $\mathrm{TIC}_1, \mathrm{TIC}_2, \ldots, \mathrm{TIC}_l$ と $\mathrm{risk}_1, \mathrm{risk}_2, \ldots, \mathrm{risk}_l$ とする．risk を小さい順に並べたものを $\mathrm{risk}_{[1]}\leq\mathrm{risk}_{[2]}\leq\cdots\leq\mathrm{risk}_{[l]}$ と書く．risk を一番小さくするモデル [1] を見つけ出すことがここでの目的である．本章は，Shimodaira (1993b) にそって述べる．

risk 最小のモデルが複数あることを考えて，$\mathcal{T}^*=\{i\in N: \mathrm{risk}_i=\mathrm{risk}_{[1]}\}$ とおく．観測データから構成するモデルの信頼集合を $\mathcal{T}\subset N$ と書く．一番良いモデルを与えられた有意水準で含み，かつなるべく小さくなるように，$\mathcal{T}$ を決めたい．Tukey 流の多重比較の方法に従えば，有意水準 α は次のように定義される．$\Pr\{\mathcal{T}^*\subset\mathcal{T}\}\geq 1-\alpha$ (Hochberg and Tamhane (1987))．これとは別に，Gupta 流の subset selection の方法に従えば，有意水準は次のように定義される．各 $k\in\mathcal{T}^*$ について $\Pr\{k\in\mathcal{T}\}\geq 1-\alpha$ (Gupta and Panchapakesan (1979))．本稿ではこの二通

りの多重比較について検討した．実際の応用では $\text{risk}_{[1]}<\text{risk}_{[2]}$ が一般的にいえる．この時 $|\mathcal{T}^*|=1$ となり，上記の二通りの有意水準の定義は同じになる．その場合でも，Tukey の信頼集合の方が Gupta のものより大きくなり保守的なので，その意味で，Gupta の方が優れているといえる．しかし，risk を小さくするモデルのいくつかが，ほとんど$\text{risk}_{[1]}$と同じ値を持っていることは現実に考えられ，このような場合を考慮すると，比較的良いモデルをもれなく選び出したい時には Tukey のやり方にも意味がある．

信頼集合を構成するために，本稿では次のような近似計算を行なった．まず，$\text{TIC}_1, \text{TIC}_2, \ldots, \text{TIC}_l$ は多変量正規分布に従うとした．すべてのモデルの最適な分布が互いに異なる時，漸近的にこの仮定は正当化されるが，もし，$p_1^*=p_2^*$ なら，$\text{TIC}_1-\text{TIC}_2$ は漸近的に χ^2 の線形結合になる．現実の応用では，$p_1^*=p_2^*$ のようなことは一般的におこらないが，p_1^* と p_2^* がデータ数 n に対して十分近い場合，やはり問題が残る．さらに，信頼集合の計算で $\triangle\text{TIC}$ の分散の推定量を用いるのだが，それをその真値とみなした．これは一種の多変量 t-分布を正規分布で近似してしまうことに相当する．このような近似により，得られる信頼集合は厳密なものではない．とくにデータ数 n が小さい場合には注意が必要である．

信頼集合の計算法について述べる．導出は付録参照．まず，統計量 T_{ij} と S_i を次のように定義する．

$$T_{ij}=\frac{\text{TIC}_i-\text{TIC}_j}{\sqrt{V\{\text{TIC}_i-\text{TIC}_j\}}}, \qquad S_i=\max_{j\neq i, j\in N} T_{ij}$$

c_i を適当な定数として，信頼集合を $\mathcal{T}=\{i\in N: S_i\leq c_i\}$ と書くことにする．有意水準 α から定数 c_i は次のように計算される．まず，確率変数 Z_{ij} を定義する．

$$Z_{ij}=\frac{(\text{TIC}_i-\text{risk}_i)-(\text{TIC}_j-\text{risk}_j)}{\sqrt{V\{\text{TIC}_i-\text{TIC}_j\}}}$$

Tukey 流の場合は，すべての $i\in N$ について $c_i=c$ とおき，

$$\Pr\left\{\max_{i\in N}\max_{j\neq i, j\in N} Z_{ij}\leq c\right\}=1-\alpha \tag{4.1}$$

Gupta 流の場合は，各 $i\in N$ について，

$$\Pr\left\{\max_{j\neq i, j\in N} Z_{ij}\leq c_i\right\}=1-\alpha \tag{4.2}$$

より定数 c_i を定義する．実際の計算では，$(\text{TIC}_1-\text{risk}_1, \ldots, \text{TIC}_l-\text{risk}_l)$ を平均ベクトル 0 の多変量正規分布にしたがう乱数として生成し，モンテカルロ計算を行なった．共分散行列は，$V\{\triangle\text{TIC}\}$ の推定量を使って計算する．各モデルの P-値は，$S_i=c_i$ となるような α の値として計算した．

5. モデル地図

データ数に対してモデルの候補の数が多い場合，モデル信頼集合は大きくなる傾向がある．先験知識にしたがって，モデルの候補をなるべく絞っておくべきなのはいうまでもないのだが，実際の応用では，候補の数が非常に多くなる場合も少なくない．このようにして信頼集合が大きくなってしまった場合，どういう傾向のモデルが選ばれているのかを調べることが重要になる．これにより，新たにデータをとるための実験の計画を考えたり，モデルの候補のクラスを考え直すこともできるだろう．このためのひとつの手段として，本章では「モデル地図」について

考える．石黒（1994）もいっているように，各モデルを「モデル空間」の点として表現した地図は，モデル探索の過程での有力な方法になり得る．

モデル間の幾何学的な関係を議論するために，モデル間の「距離」を定義しなければならない．これには様々な可能性があるが，ここではモデル選択のための地図を描くことが目的なので，TIC を使って次のように定義する．モデル i とモデル j の距離を d_{ij} と書く時，

$$d_{ij}=\sqrt{V\{\mathrm{TIC}_i-\mathrm{TIC}_j\}}.$$

これが距離の公理を満たすのは分散の性質より明らかである．したがって，モデルが l 個あれば，$l-1$ 次元ユークリッド空間に l 個の点をとってそれらの距離が d_{ij} になるようにできる．この l 個の点のひとつひとつがモデルを表すものと考えることができる．

このように定義したモデル地図は，真の分布が変われば違ったものになる．したがって，この地図を，「真の分布から見た」モデル地図と呼ぶことにする．実際には真の分布は知らず，d_{ij} もデータから推定した $\hat{d}_{ij}$ を使うので，描くことができるのは，経験分布から見た地図になる．また，l 次元空間は実際に紙に描けないので，6 章での例では，多次元尺度法を用いて 2-3 次元におとして描いた．

定理 1 と定理 2 より，大雑把にいって，$d_{ij}\approx\sqrt{J(p_i^*, p_j^*)/n}$ である．従って，d_{ij} は「確率分布の空間」での各モデルの最適分布間の距離とも解釈できる．実際には，経験分布から見た地図を描くので，$\hat{d}_{ij}$ は各モデルの推定分布間の距離とみなせる．

モデルの「高さ」を TIC で定義すると，多重比較で用いた統計量 T_{ij} は，モデル i からモデル j を見た傾きになる．従って，図 4 からわかるように，各モデルの TIC の順序と P-値の順序は必ずしも一致しない．しかし経験的にはこれらの順序はほぼ同じになることが多いようである．

モデル i とモデル j の良さの判別は $|T_{ij}|$ が大きいほど容易なので，$|T_{ij}|<c$ は良さの判別ができなくなる条件といえる．実は，式 (4.1) の Tukey の多重比較の定数 c を用いると，すべてのモデルの対比較を同時に検定していることになる．特に，モデル i がモデル j を含む場合について調べてみる．この時，確率分布の空間で K-L 情報量のピタゴラスの定理が近似的に成り立つことを考えると，$\mathrm{TIC}_i-\mathrm{TIC}_j\approx -D(\hat{p}_i, \hat{p}_j)+\triangle m/n$ と書ける．ただし $\triangle m$ はモデルの自由度の差である．$d_{ij}\approx\sqrt{2D(p_i^*, p_j^*)/n}$ を考えると，$T_{ij}\approx -n\hat{d}_{ij}/2+\triangle m/n\hat{d}_{ij}$ と近似できる．したがって，$|T_{ij}|<c$ は $|n\hat{d}_{ij}-\sqrt{2\triangle m+c^2}|<c$ で近似できる．c^2 が $2\triangle m$ に比べて大きい時は，さらに $n\hat{d}_{ij}<2c$ と近似できる．これより，$n\hat{d}_{ij}$ が c に対してある程度大きければ，二つ

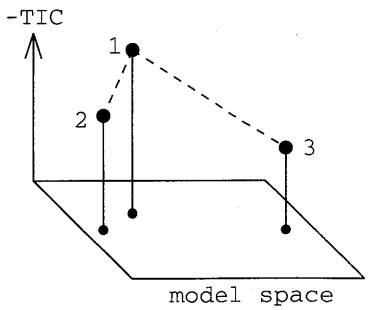

図 4．モデル空間を平面で表し TIC を高さで表した．この図の場合，3 つのモデルがあり，$\mathrm{TIC}_1<\mathrm{TIC}_2<\mathrm{TIC}_3$ である．しかし，モデル 2 はモデル 1 に非常に近いので，$T_{21}>T_{31}$ となる．したがって，Tukey の P-値は，モデル 3 の方がモデル 2 より大きくなる．

のモデルの良さの判別がつくことがわかる.

6. 数値例

重回帰分析の変数選択問題についていくつかの数値例をあげる. $x=(x_i)$ を説明変数ベクトル, y を従属変数とする. x と y の同時分布を $q(x, y)=q(y\mid x)q(x)$ と書く. (x, y) の n 個の独立な観測値 $((x_{i1}), y_1), \ldots, ((x_{in}), y_n)$ が与えられているとする. $q(x, y)$ でなく $q(y\mid x)$ の良い推定を得ることが目的なので, $d(x, y, p)=-\log p(x, y\mid\theta)$ の代わりに $d(x, y, p)=-\log p(y\mid x, \theta)$ を用いる.

真の分布 $q(x, y)$ の正規性を仮定すると, $G^*=H^*$ となり, TIC は AIC と同じになる. すなわち, m 個の説明変数を使ったモデルの TIC は,

$$\mathrm{TIC}=\frac{1}{2}(1+\log 2\pi+\log\hat{\sigma}^2)+\frac{m+2}{n}$$

となる. また, $\varepsilon^*=y-\beta_0^*-\beta_1^*x_1-\cdots-\beta_m^*x_m$ とおくと,

$$V\{e^*\}=1-\left(\frac{E\{\varepsilon_1^*\varepsilon_2^*\}}{\sigma_1^*\sigma_2^*}\right)^2$$

となる. とくに, モデル1がモデル2を含む時は $E\{\varepsilon_1^*\varepsilon_2^*\}=E\{\varepsilon_1^{*2}\}$ なので,

$$V\{e^*\}=1-(\sigma_1^*/\sigma_2^*)^2$$

となる.

以下の節で述べる数値例では, (3.1) によって $V\{\triangle\mathrm{TIC}\}$ を推定した. また, Tukey の P-値は, Holm の sequentially rejective procedure を適用して検出力をあげた (Hochberg and Tamhane (1987), p. 56). 各モデルは推定に用いた説明変数の番号を {1, 2} のように表示した. また, モデル地図における距離は, $n\hat{d}_{ij}$ を用いた.

6.1 新生児の体重のデータ

データセット BABY は佐和 ((1979), p. 57) よりとった. データ数は $n=15$ で, 4つの説明変数がある. モデルの候補は4つの説明変数を使う使わないのすべての組合せ $2^4=16$ 個を考えた.

図5には Tukey と Gupta の方法で計算した各モデルの P-値を示した. 有意水準 20% では, Tukey でも Gupta でも8つのモデルが選ばれる. 選ばれたモデルは, 説明変数 X_1 を含むモデルである. 図6に各モデルの AIC と Gupta の P-値を示した. 説明変数 X_1 を含む8個のモデルが, AIC を十分に小さくしていることがこれからもすぐに読みとれる. 実は, 回帰係数 $\beta_1, \ldots, \beta_4$ の t-統計量は, 順に, 8.1, −2.3, 3.1, −1.6 であり, これからも X_1 が選ばれやすいことがわかる.

図7にモデル地図を示す. この地図では, モデル信頼集合に含まれる8つのモデルのみを Torgerson (1958) の多次元尺度構成法を用いて3次元におとして描いた. 計算は S 言語の cmdscale 関数を用いた (渋谷・柴田 (1991)). これを模式的に表すと, 図8のようになる. また, これを階層的クラスタ分析した結果を図9に示す. 計算は S 言語の hclust 関数を用いた. モデルの横の数値は Gupta の P-値である.

どのモデル地図の表示からも読みとれるように, 信頼集合の8つのモデルは, 直方体の頂点にならんでいる. モデル {1, 2, 3, 4} は経験分布を代表していると考えられ, また各モデルの推

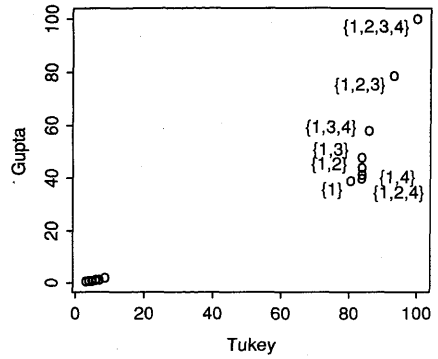

図5. データセット BABY における Gupta と Tukey の P-値.

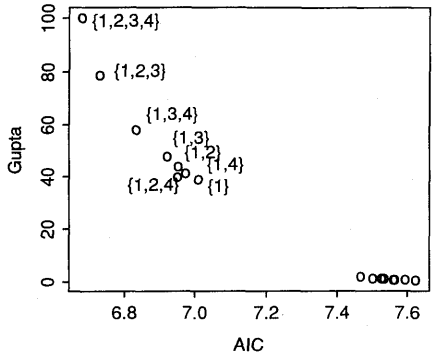

図6. データセット BABY における Gupta の P-値 と AIC の値.

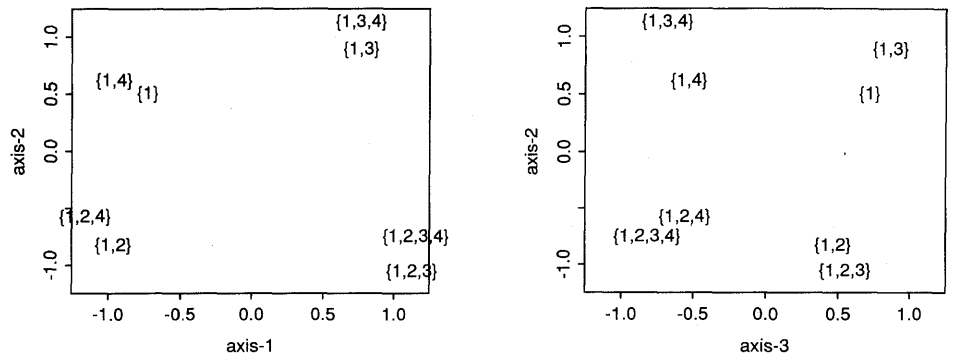

図7. データセット BABY のモデル地図.

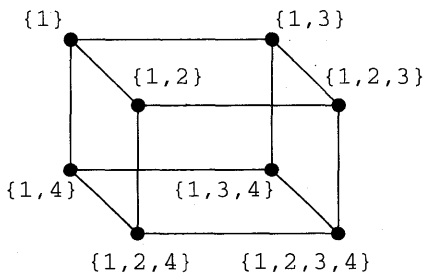

図8. データセット BABY のモデル地図 (模式図).

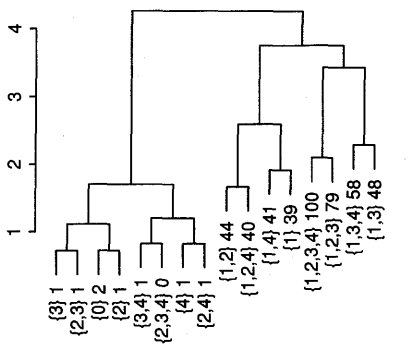

図9. データセット BABY のクラスタ分析.

定量は図2のように射影で表現できるので，このように8つのモデルがならんでいるのは当然のことといえる．ただし，直方体がゆがんで菱形のようになっていないことより，3つの説明変数 X_2, X_3, X_4 の従属変数を説明する力がほぼ，独立に作用していることがわかる．つまり確率分布の空間において，モデル {2} の軸，モデル {3} の軸，モデル {4} の軸が，ほぼ直交していることを表す．また，直方体の3辺の長さがほぼ等しいので，経験分布はこれらモデルの3軸より，ほぼ等距離にあることもわかる．これはまさに図2のような状況になっているので，モデル {1, 2}, モデル {1, 3}, モデル {1, 4} の間で良さの判別がつきにくいことが理解できる．また，変数 X_1 を含む8つのモデルの P-値が大きいことより，モデル {1, 2, 3, 4} が，他の7つのモデルから，モデルの良さを判別できるほどには十分に離れていないことがわかる．

モデル {1, 2, 3, 4} の最適な分布 $p^*_{\{1,2,3,4\}}(y \mid x)$ と，モデル {1} の最適な分布 $p^*_{\{1\}}(y \mid x)$ が等しい時，変数 X_1 を含むすべてのモデルの最適な分布は等しくなる．例えば，真の分布 $q(y \mid x)$ がモデル {1} に含まれるときがこれに当たる．このとき，risk を一番小さくするのは，一番小さいモデル {1} であるが，データ数 $n \to \infty$ の極限をとっても，TIC の "inconsistency" より，変数 X_1 を含むモデルが選ばれる確率は，どれも0より大きい値に収束する．このことを考慮に入れると，この例の場合には，モデル {1} だけを選んでも良いと思う．ただし，この例では $n=15$ であって，とても n が十分に大きいとはいえないので，この結論が正しいということを数理的に保証することはできない．

6.2 ハルドのセメント発熱量のデータ

データセット HALD は Draper and Smith ((1981), p. 629) よりとった．データ数は $n=13$ で，4つの説明変数がある．モデルの候補は $2^4=16$ 個を考えた．

図10には Tukey と Gupta の P-値を示した．有意水準20%では，Tukey でも Gupta でも同じ7個のモデルが選ばれる．前節で説明したように TIC の inconsistency を考慮に入れると，{1, 2} と {1, 4} と {2, 3, 4} の3つだけを選んでも良いと思う．すなわち，信頼集合の7つのモデルは，この3つのうちどれかを必ず含んでいる．

図11を見ると，モデル {1, 2, 3, 4} の AIC の方がモデル {1, 4} より小さいのに，P-値の順序は逆になっていることがわかる．これは，{1, 2, 3, 4} が {1, 2, 3} や {1, 2, 4} に近いためであり，ちょうど図4の実例になっている．

図12にモデル地図を示す．この地図では，モデル信頼集合に含まれる7つのモデルのみを多次元尺度構成法を用いて3次元におとして描いた．また，これを階層的クラスタ分析した結果を図13に示す．

図12を見ると，TIC の inconsistency を考慮して選ばれる3つのモデル {1, 2}, {1, 4}, {2, 3, 4} が，信頼集合の7つのモデルを取り囲んでいるのがわかる．観測された経験分布がこの3つのモデルに近く，したがって，これら3つのモデルのうちひとつでも含むような7つのモデルについて，どれが良いかを判定できないことが読みとれる．

変数 X_1 を含む8つのモデルについてのモデル地図を図14に示す．実は，これら8つのモデルは，ほぼこの2次元平面にのっている．本来なら前節での例のように，これら8つのモデルは3次元的に配置しているべきである．このように2次元におちてしまうのは，{1, 2}, {1, 3}, {1, 4} の3軸が，独立ではないことを示している．HALD のデータは，4つの説明変数の和がほぼ $X_1+X_2+X_3+X_4 \approx 100$ になるデータであり，これは当然の結果といえる．これより {1, 2, 3}, {1, 2, 4}, {1, 3, 4} の3つのモデルが {1, 2, 3, 4} に近いことも理解できる．また，図14を見ると，{1, 3} が {1} に非常に近い．これは，{1} から {1, 3} への説明力の増加が乏しいことを示す．

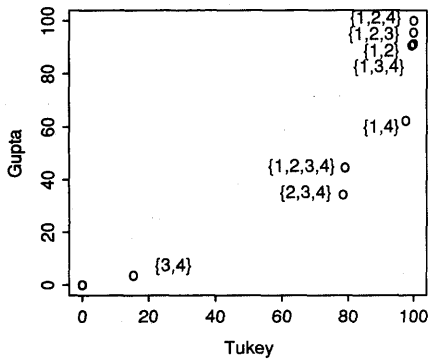

図10. データセット HALD における Gupta と Tukey の P-値.

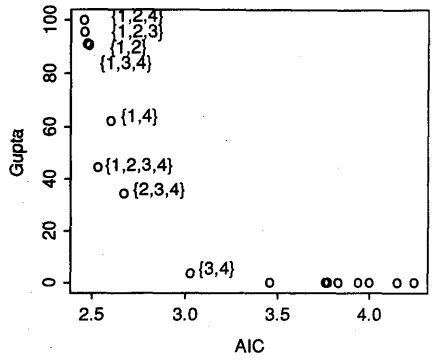

図11. データセット HALD における Gupta の P-値と AIC の値.

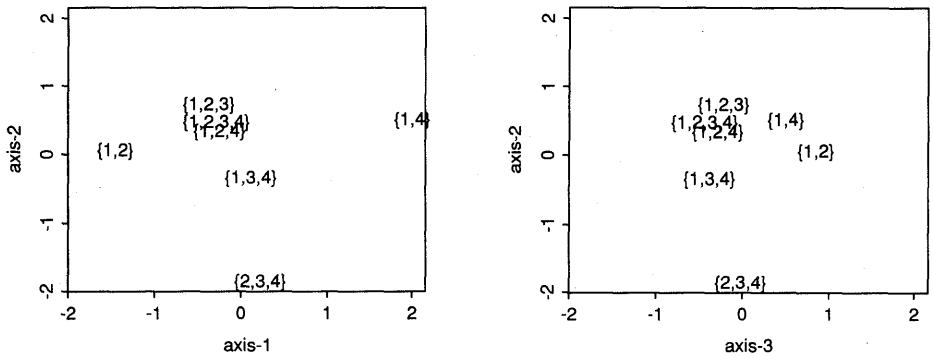

図12. データセット HALD のモデル地図.

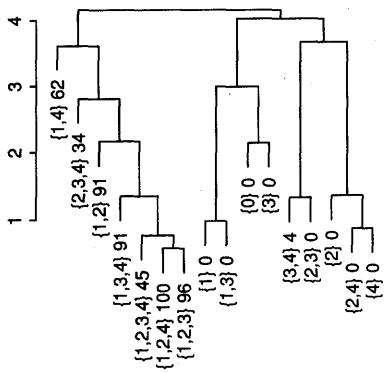

図13. データセット HALD のクラスタ分析.

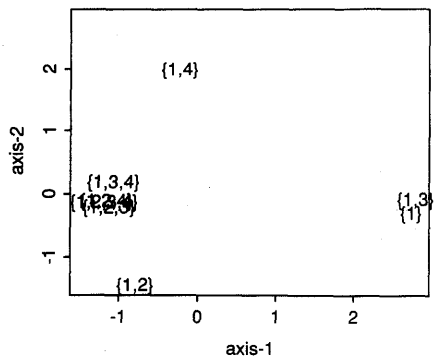

図14. データセット HALD のモデル地図（変数 X_1 を含む 8 つのモデル）.

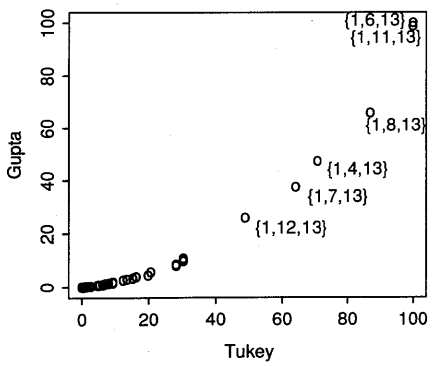

図15. データセット BOSTON における Gupta と Tukey の P-値.

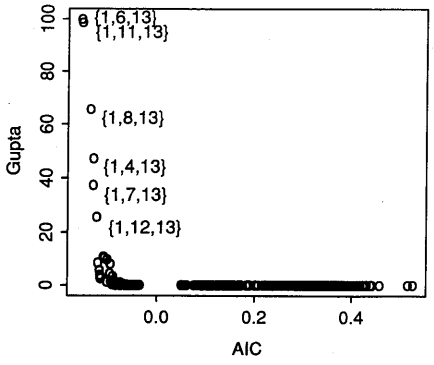

図16. データセット BOSTON における Gupta の P-値と AIC の値.

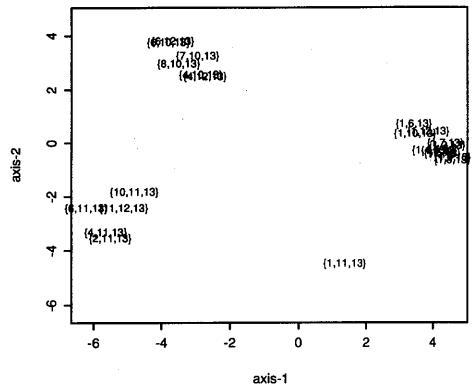

図17. データセット BOSTON のモデル地図.

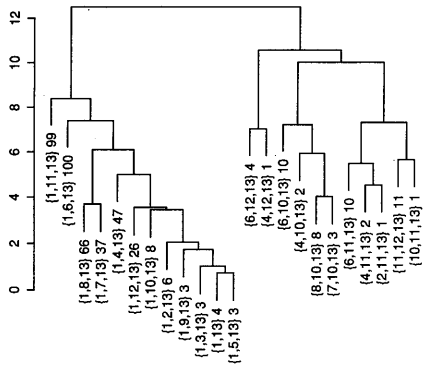

図18. データセット BOSTON のクラスタ分析.

6.3 ボストン各地域の住宅価格のデータ

データセット BOSTON は Belsley et al. ((1980), p. 244) よりとった. データ数は $n=506$ で, 13 個の説明変数がある. まず最初に, 13 個の説明変数のうち 3 個以下を選び出す 378 通りのモデルを候補とした. 図 15 より, 信頼集合の上位 6 個のモデルは, Gupta では有意水準 20%, Tukey では有意水準 40% に相当する. 図 16 を見ると, この 6 つのモデルは AIC を小さくする上位 6 つのものであることがわかる.

Gupta の P-値が 1% 以上の 23 個のモデルについて, モデル地図を図 17 に, そのクラスタ分析を図 18 に示す. モデル地図より, この 23 個のモデルは 4 つのグループに分けられることが読みとれる. これらはすべて {13} を含み, {1} と {11} を含むか含まないかによって分けられていることがわかる. 特に, Gupta の P-値が 20% 以上の 6 つのモデルは, {1, 13} を含む. 従って, 観測された経験分布は {1, 13} に近く, これを含むようなモデルの中では, 良さの判定は難しい. そのようなモデルの中では {1, 6, 13} と {1, 11, 13} が特に P-値を大きくしているが, 他のモデルの方が risk を小さくしている可能性も棄却できない.

次に, 13 個の説明変数から 4 個以下を選び出す 1,093 通りのモデルの候補を考えた. 結果は,

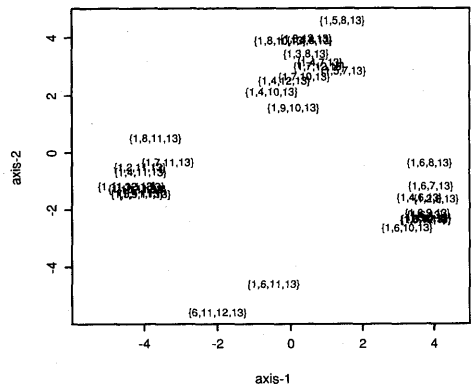

図19. データセット BOSTON のモデル地図 (4個までの説明変数).

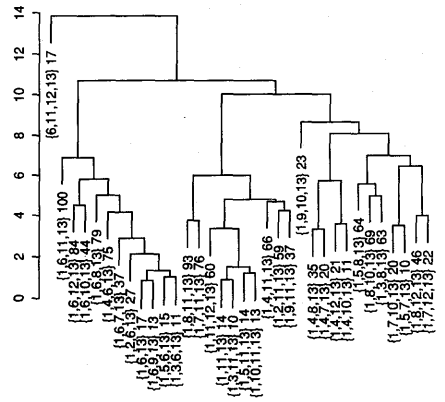

図20. データセット BOSTON のクラスタ分析 (4個までの説明変数).

有意水準 20% では，Tukey で 62 個が，Gupta で 22 個のモデルが選ばれた．これでは非常に多くのモデルが選ばれてしまい，結果の解釈は難しい．Gupta の P-値が 10% 以上の 34 個のモデルについて，モデル地図を図 19 に，クラスタ分析を図 20 に示す．モデル地図より，この 34 個のモデルが 4 つのグループに分けられることが読みとれる．これらはひとつのモデルを除きすべて {1, 13} を含み，{11} と {6} を含むか含まないかによって分けられていることがわかる．P-値を最大にするのは {1, 6, 11, 13} であり，観測された経験分布はこの付近にあるといえるが，その経験分布は同時に {1, 13} を含む多くのモデルにも近く，それらの差が有意とはいえない．

データセット BOSTON について，上記のように二通りのモデル候補のクラスを考えたが，いずれの結果も，{1, 13} を含むモデル間では良さの判定は難しいといえる．TIC の inconsistency を考慮すると，モデル {1, 13} のみを選ぶのもひとつの方法といえる．

7. むすび

AIC 最小化法は一般に広く使われているが，多重比較の立場で見ると，本稿で示されたように，その信頼性に問題がある．つまり，AIC を最小にしたモデルは確かに良いモデルかも知れないが，候補の中の他のモデルがすべて悪いと言い切れるわけではない．モデル選択を，探索的に用いる場合も，仮説検証的に用いる場合も，この点に注意を払う必要がある．

本稿では，与えられた有意水準の下で，モデルの信頼集合を構成した．これは，モデルを一つだけ選ぶことによって，他の可能性を無視してしまうよりも，あり得る可能性を積極的に拾い上げることに重点をおいた方法といえる．従って，実際の応用の場面では，これで問題が解決するのではなく，本稿での方法は，続く統計処理への示唆を与える為のものである．

その意味で，得られたモデル信頼集合やモデルの P-値の情報をまとめ上げることが重要になる．モデル地図を利用して比較的良いモデルに共通する性質を見つけ出すなどの方法について，さらに考える必要がある．

信頼集合を構成する為の有意水準をどう決めたら良いかという問題がある．本稿では 20% を主に用いたが，この値に積極的な根拠はない．モデルの P-値を大きい順に並べて，適当に上位のものを選ぶのでは，モデルの AIC を小さい順に並べて，適当に上位のものを選ぶのと，同じようでもある．しかし，モデルの P-値は，モデルの良さのひとつの規格化された物差しとみな

せる点で，AIC とは異なる．すなわち，あるモデル選択における AIC の差を，別のモデル選択での AIC の差と直接比較することは，あまり意味がない．

本稿では，Tukey や Gupta の多重比較の方法を利用することにより，モデル信頼集合を構成した．これら以外にも様々な方法が考えられる．例えば，ブートストラップ確率に基づき信頼集合を構成する方法などがあげられる (Felsenstein and Kishino (1993)). 何をもって error とするかによって，色々な有意水準の定義が考えられ，それに応じて，信頼集合の構成法が与えられる．モデル選択の目的や実用上の計算の容易さなどに応じて，それらの方法を検討する必要があろう．

一方，Akaike (1979) によれば，$\exp(-\mathrm{AIC}/2)$ は「モデルの尤度」とみなせる．この考えに従えば，AIC 最小化法は，モデルの適当な prior のもとでの，posterior mode を選ぶことと表現できる．本稿のモデルの P-値は，$\exp(-\mathrm{AIC}/2)$ と同様にモデルの尤度といえるかもしれないが，両者は計ろうとしている量が違うので，直接の比較はできない．また，各モデルによって推定された分布を，モデルの事後分布によって平均化した分布を推定量とする方法も Akaike (1979) で提案されている．これは，Bayes 流の枠組の中で，良いモデルに共通する性質をまとめ上げる一つの方法になっている．

本稿では P-値を計算するのに，4 章で述べたような近似をおこなった．データ数 n が比較的小さい場合，これが実用上の問題になる可能性があり，さらに検討が必要である．また，モデルの候補が多い場合，P-値を計算するためのモンテカルロ計算は手間のかかるものとなる．P-値を計算するのに，すべてのモデルの組合せについて Z_{ij} を生成するのではなく，実用上は，AIC を比較的小さくするものだけについて計算すれば十分であろう．

謝　辞

有益な御助言を頂いた東京大学計数工学科の甘利俊一先生，廣津千尋先生に感謝します．慶應義塾大学理工学部の柴田里程先生には，分散の高次項の問題について指摘されました．モデル地図を描くというアイデアは，統計数理研究所の石黒真木夫先生によるものであり，また本稿を書くに当たり多くの御助言を頂きました．東京大学教養学部の岸野洋久先生には，信頼集合の傾向を調べるべきだと指摘されました．査読者の方からは貴重なコメントをいただきました．研究環境についての援助をしてくださった指導教官の中野馨先生に感謝します．

付　録

定理 1 の証明. $\triangle\hat{\theta}=\hat{\theta}-\theta^*$ とおき，$\mathrm{TIC}(\hat{q}, p(\cdot))$ を $\hat{\theta}$ の周りで Taylor 展開すると，

$$\mathrm{TIC}(\hat{q}, p(\cdot))=D(\hat{q}, p(\theta^*))-(1/2)\mathrm{tr}(\hat{H}\triangle\hat{\theta}\triangle\hat{\theta}')+(1/n)\mathrm{tr}G^*H^{*-1}+O(\|\triangle\hat{\theta}\|^3)$$

これより二つのモデルの TIC の差は，

$$\begin{aligned}\triangle\mathrm{TIC}&=\mathrm{TIC}(\hat{q}, p_1(\cdot))-\mathrm{TIC}(\hat{q}, p_2(\cdot))\\&=\hat{E}\{e^*(x)\}-(1/2)(\mathrm{tr}\,H_1^*\triangle\hat{\theta}_1\triangle\hat{\theta}_1'-\mathrm{tr}H_2^*\triangle\hat{\theta}_2\triangle\hat{\theta}_2')\\&\quad+(1/n)(\mathrm{tr}G_1^*H_1^{*-1}-\mathrm{tr}G_2^*H_2^{*-1})+O(\|\triangle\hat{\theta}_1, \triangle\hat{\theta}_2\|^3)\end{aligned}$$

ただし，$O(\|\triangle\hat{\theta}_1, \triangle\hat{\theta}_2\|^3)$ は $\triangle\hat{\theta}_1$ と $\triangle\hat{\theta}_2$ に関して 3 次以上のオーダの項とする．これより，

$$\triangle\mathrm{TIC}-E\{\triangle\mathrm{TIC}\}=\hat{E}\{e^*(x)\}-E\{e^*(x)\}+O(\|\triangle\hat{\theta}_1, \triangle\hat{\theta}_2\|^2)+O(n^{-1})$$

従って，$E\{O(\|\triangle\hat{\theta}_1, \triangle\hat{\theta}_2\|^3)\}=O(n^{-2})$ より，

$$V\{\triangle \mathrm{TIC}\}=V\{\hat{E}\{e^*(x)\}\}+O(n^{-2})$$
$$=V\{e^*(x)\}/n+O(n^{-2})$$

となる．系1は，$\hat{V}\{\hat{e}(x)\}=V\{e^*(x)\}+O_p(n^{-0.5})$ より明らか．

定理2の証明． 次のように $\alpha=(\alpha_1, \alpha_2)$ でパラメトライズして，$p(x\mid\alpha)$ をつくる．

$$\log p(x\mid\alpha)=(1-\alpha_1-\alpha_2)\log q(x)+\alpha_1\log p_1^*(x)+\alpha_2\log p_2^*(x)-c(\alpha)$$

この時，$p_1^*=p(1, 0)$, $p_2^*=p(0, 1)$, $q=p(0, 0)$ である．さて，$\log p$ を l で表し，$p\partial_i l=\partial_i p$ に注意すると，

$$l_\alpha=l_0+\sum(\partial_i l_0)\alpha_i+O(\|\alpha\|^2),\qquad p_\alpha=p_0(1+\sum(\partial_i l_0)\alpha_i)+O(\|\alpha\|^2)$$

とかける．これより直ちに，

$$J(p_\alpha, p_\beta)=D(p_\alpha, p_\beta)+D(p_\beta, p_\alpha)$$
$$=\int(p_\alpha-p_\beta)(l_\alpha-l_\beta)dx$$
$$=\int p_0(\sum\partial_i l_0(\alpha_i-\beta_i))^2dx+O(\|\alpha, \beta\|^3)$$
$$=\sum\sum g_{ij}(\alpha_i-\beta_i)(\alpha_j-\beta_j)+O(\|\alpha, \beta\|^3)$$

ただし，$g_{ij}=\int p_0(\partial_i l_0)(\partial_j l_0)dx$ とおく．ところで，$E\{\partial_i l_0\}=\int p_0\partial_i l_0 dx=0$ に注意すると，

$$V\{l_\alpha-l_\beta\}=\int p_0(l_\alpha-l_\beta)^2dx+O(\|\alpha, \beta\|^4)$$
$$=\int p_0(\sum\partial_i l_0(\alpha_i-\beta_i))^2dx+O(\|\alpha, \beta\|^3)$$
$$=J(p_\alpha, p_\beta)+O(\|\alpha, \beta\|^3).$$

最後に，$\alpha=(1, 0)$, $\beta=(0, 1)$ とおけば良い．

多重比較の計算の導出

確率変数 U_{ij} と事象 E_{ij} を次のように定義する．

$$U_{ij}=\mathrm{TIC}_i-\mathrm{TIC}_j-c_i\sqrt{V\{\mathrm{TIC}_i-\mathrm{TIC}_j\}},\qquad E_{ij}=\{U_{ij}-(\mathrm{risk}_i-\mathrm{risk}_j)\leq 0\}$$

これより確率変数 U_i と事象 E_i を

$$U_i=\max_{j\neq i}U_{ij},\qquad E_i=\bigcap_{j\neq i}E_{ij}$$

で定義する．この時すぐわかるように，

$$U_i\leq 0\ \Leftrightarrow\ S_i\leq c_i,\qquad E_i\ \Leftrightarrow\ \max_{j\neq i}Z_{ij}\leq c_i$$

である．これより信頼集合は $\mathcal{T}=\{i\in N:\ U_i\leq 0\}$ と書け，また (4.1) は $\Pr\{\bigcap_{i\in N}E_i\}=1-\alpha$, (4.2) は $\Pr\{E_i\}=1-\alpha$ と書けることがわかる．このとき，この α が実際に有意水準になっていることを示す．

まず $k\in\mathcal{T}^*$ とおく．すると任意の $j\in N$ について $\mathrm{risk}_k-\mathrm{risk}_j\leq 0$ がいえて，$\{k\in\mathcal{T}\}$ という事象が E_k を含むことがわかる．従って，各 $k\in\mathcal{T}^*$ について $\Pr\{E_k\}\leq\Pr\{k\in\mathcal{T}\}$ がいえる．Gupta の方法では，すべての $i\in N$ について $\Pr\{E_i\}=1-\alpha$ とおくことにより，たしかに α が

有意水準になっていることがわかる．また，Tukeyの方法では $\Pr\{\mathcal{F}^* \subset \mathcal{F}\} \geq \Pr\{\bigcap_{k \in \mathcal{F}} E_k\} \geq \Pr\{\bigcap_{i \in N} E_i\} = 1 - \alpha$ に注意すれば良い．

参 考 文 献

Akaike, H. (1974). A new look at the statistical model identification, *IEEE Trans. Automat. Control,* **AC-19**, 716-723.

Akaike, H. (1979). A Bayesian extension of the minimum AIC procedure of autoregressive model fitting, *Biometrika,* **66**, 237-242.

Belsley, D.A., Kuh, E. and Welsch, R.E. (1980). *Regression Diagnostics,* Wiley, New York.

Draper, N. and Smith, H. (1981). *Applied Regression Analysis,* 2nd ed., Wiley, New York.

Felsenstein, J. and Kishino, H. (1993). Is there something wrong with the bootstrap on phylogenies? A reply to Hillis and Bull, *Systematic Biology,* **42**, 193-200.

Gupta, S.S. and Panchapakesan, S. (1979). *Multiple Decision Procedures,* Wiley, New York.

Hochberg, Y. and Tamhane, A.C. (1987). *Multiple Comparison Procedures,* Wiley, New York.

石黒真木夫 (1994). AICはなぜ役にたつのか?，応用数理，**4**(2)，19-33.

Kishino, H. and Hasegawa, M. (1989). Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea, *Journal of Molecular Evolution,* **29**, 170-179.

佐和隆光 (1979).『回帰分析』，朝倉書店，東京.

渋谷政昭，柴田里程 (1991).『S言語』，共立出版，東京.

Shimodaira, H. (1993a). On the variance of the information criterion for model selection (submitted for publication).

Shimodaira, H. (1993b). An application of multiple comparison techniques to model selection (submitted for publication).

竹内 啓 (1976). 情報統計量の分布とモデルの適切さの規準，数理科学，**153**，12-18.

竹内 啓 (1983). AIC基準による統計的モデル選択をめぐって，計測と制御，**22**，445-453.

Torgerson, W.S. (1958). *Theory and Methods of Scaling,* Wiley, New York.

Proceedings of the Institute of Statistical Mathematics Vol. 41, No. 2, 131-147 (1993) 147

A Model Search Technique Based on Confidence Set and Map of Models

Hidetoshi Shimodaira

(Department of Mathematical Engineering and Information Physics, University of Tokyo)

This paper describes a procedure for choosing a set of "good" models from competing candidates using multiple comparison techniques. Furthermore, a map of models is introduced by defining a distance of models in order to examine the selected "good" models.

Since the log-likelihood, which is the first term of Akaike's information criterion (AIC), has large variance, we choose several models which have relatively small values of AIC, instead of choosing only one model which minimizes AIC. First, we construct an estimator of the variance of the difference between the AIC values of any two models. Then, using the variance estimator, we choose a confidence set of models which includes the best model, at a given confidence coefficient. The result of the procedure will be shown in nominal "P-value" for each model, where the P-value is the largest significance level for which that model is included in the confidence set of models.

The confidence set will be very large when the number of candidate models is large compared to that of the data samples. In such a case, it is important to find a pattern in the models in the confidence set. A geometrical interpretation of the selected models seems a likely means for obtaining such a problem. Here the distance of two models is defined as the square root of variance of the difference between their AIC values. This distance is approximately proportional to the square root of the Kullback-Leibler divergence of the estimated distributions of the two models. The map of models is drawn using the classical multi-dimensional scaling method with the distance defined.

The model map gives us a lot of information concerning the model selection. It will show the location of the true distribution relative to candidate models. Since similar models are close each other on the map, a pattern of model structures in the confidence set can be found. A diagnosis on some problems, such as multi-collinearity, can be made.

The variable selection problem in multiple regression analysis is an example of model selection. Some numerical examples will be shown to illustrate our procedure.

Key words: AIC, model selection, model search, confidence set, multiple comparison, subset selection, model map, multiple regression, variable selection.