統計モデル選択の概要

松嶋　敏泰

1. はじめに

モデル選択はデータから適切なモデルを決定する決定問題と考えられるが，モデルの適切さとは何であろうか．工学で用いられるモデルの適切さは，そのモデルが使われる目的から考えることが可能であろう．本稿ではモデルの利用目的から適切さの視点を整理することによって，モデル選択の問題を概観する．

2. モデル選択とは

統計モデルを構成し，利用していく工学的プロセスの一般的流れは概ね以下のよう考えられる．

S1) モデルの選択，S2) 検定推定，S3) モデルの利用．

S1) では，対象母集団の性質や現象が発生するメカニズム等を考慮に入れモデルの候補を選択する．多くの場合ここで用いられる統計モデルはパラメトリックなモデルであり，本稿でもパラメトリックモデルで議論を進める．パラメトリックモデルの確率密度関数を $f(x|\theta_m, m)$ で表現する．ここで m はモデル，$\theta_m \in \Theta_m \subseteq \mathcal{R}^k$ はモデル m の k 次のパラメータとする．例えば，モデル m として正規分布を考えれば，パラメータは $\theta_m = (\mu, \sigma)$ の2次のパラメータとなる．

S2) では，選択されたモデル m に関して，データから θ_m を推定しパラメータも含めてモデルを一つに決定する．検定を用いて帰無仮説 m_1 と対立仮説 m_2 からモデルを一つに決定することもある．

S3) では，このように決定された一つのモデルを用いて予測，制御，圧縮など問題の目的に応じた意思決定がなされる．

S2) 以降では選択されたモデルに対して恣意的，主観的判断をなるべく排除する方向で，モデルの決定や利用がなされる．しかし，S1) のモデル選択においては，データのみでなく固有技術や過去の経験からの主観が入った判断を行わざるを得ない．この部分の判断からなるべく主観を取り除いて，データの情報のみを使ってモデルを選択するための基準がモデル選択基準と考えられる．

データからのみモデルを決定するといっても，全く何もないところからモデルを創り出すわけではなく，選択するモデルの範囲はあらかじめ設定しておく必要がある．それをモデルのクラスと呼ぶ．本稿の扱うモデルのクラスはパラメトリックモデルの有限集合とする．モデルクラスを決定する時点ではやはり主観的判断が必要となるわけで，S1) から完全に主観を取り除けるわけではない．モデル選択はS2) のみしか扱えなかった古典的統計推測問題をS1) へと拡張していった問題ともとらえることができる．

このモデルのクラスのタイプは，まず分離型，非分離型に分類される．非分離型はさらにモデル間に順序関係があるクラスとないクラスに大別される．順序関係があるクラスは階層型と呼ばれ，特に全順序関係があるクラスは入れ子 (nested) 型と呼ばれる．

例えば，自己回帰モデルのクラスのモデル間には全順序関係があり，高次のモデルは必ず低次のモデルを含んでおり，入れ子型のモデルクラスとなる．重回帰モデルのクラスは変数組み合わせによりモデル間に半順序関係が成り立ち，入れ子型ではないが階層型のクラスとなる．

モデルクラスの中からモデルを選ぶというS1) の問題は，S2) の検定を繰り返し1つのモデルに絞り込むことで解決可能に思われる．しかし，その仮説の組み合わせは莫大になり，順番に行われるそれぞれの検定は一般に独立ではなく，危険率などを調整することは非常に難しいと考えられる．

また，入れ子型モデルのクラスの場合は真のモデルより高次のすべてのモデルは真のモデルを含んでしまうので，仮説検定の枠組みでは解決できない問題といえる．検定では，データが出てきた分布モデルは何かという視点中心であるが，この例のようなモデル選択の問題では，どのモデルが適切なのかという視点が

まつしま　としやす　早稲田大学理工学部
〒169 新宿区大久保 3-4-1

必要となってくる.

モデル選択の問題でもモデルのクラスは仮定しているわけで，その中に真のモデルが含まれる場合と，含まれない場合両方を考える必要があろう．さらには，真のモデルというものが存在するのかという根本的問題まで溯る必要があるが，そこまでは立ち入らないこととする．

以上より，本稿ではモデルの良さ，適切さという視点から議論を進めていきたい．これについても定性的にはいろいろの視点があると思えるが，モデルをどう利用するかの最終的な目的から，本稿では次の2つの視点から考えていくことにしたい．

(1) モデルを用いて新たなデータに対して行動(広義の予測)を行った場合の良さ．

(2) モデルを用いて与えられたデータに対して行動(広義の圧縮)を行った場合の良さ．

前者では，同じ統計モデルから得られる新たなデータに対して，選択したモデルを用いて何らかの行動をとった場合の良さで，そのモデルの良さを評価する．例えば，選択したモデルを用いて $t-1$ 時点までの時系列データ $X^{t-1}: X_1 \cdots X_{t-1}$ から t 時点のデータ X_t を予測した精度でモデルの良さを評価することである．

後者の場合，選択したモデルを用いることで，与えられたデータがいかにコンパクトに記録されるか，データの特徴をいかにうまく表現できるか等で評価を行う．ある意味で記述統計的な側面からの評価ともいえる．

3. モデルの良さの評価尺度

本節では，前節であげた2つの視点からの具体的評価尺度を，いくつかの損失関数で考えてみる．

3.1 新たなデータに対しての行動

選択したモデルを新たなデータの予測に用いた場合の良さの尺度として次の損失関数が考えられる．

$$L_1(f(Z|\hat{\theta}_m(x),\hat{m}(x)),g(Z)) = E_{g(Z)}[|\hat{Z}(\hat{\theta}_m(x),\hat{m}(x))-Z|^2]. \quad (1)$$

ここで，$\hat{m}(x)$ は与えられたデータ x からのモデルの推定値，$\hat{\theta}_m(x)$ は $\hat{m}(x)$ のもとでのパラメータの推定値，$\hat{Z}(\hat{\theta}_m(x),\hat{m}(x))$ は推測したモデルを用いて Z を予測した値，$E_{g(Z)}$ は真の分布 $g(Z)$ での期待値を表す．

上式は新たなデータに対する2乗誤差を損失関数としたものである．例えば先ほど述べた時系列データの予測の場合には，X^{t-1} を x に X_t を Z に対応させればよい．

自己回帰モデルに対してこの損失からモデルを選択(モデルの次数を決定)する評価基準として，FPE(Final Prediction Error)[4] が赤池により AIC 以前に提案されている．また，重回帰モデルにおいて被説明変数 Y の新たなデータに対する2乗誤差を損失としたモデル選択基準が Mallows の C_p [9] といえる．

上の2つの回帰モデルの予測問題では平均値について予測が目的であったが，行動の目的が違えば平均値だけでなく分散などを予測することの方が重要な場合もあり得る．新たなデータに対する行動をもっと広義の予測ととらえた場合，行動の良さは真のモデルと推測したモデルの平均のパラメータの違いのみではなく，分布全体の違いで測るべきと考えられる．そこで分布間の(擬)距離の一つである Kullback-Leibler 情報量によって損失関数を定義することを考える[1].

$$L_2(f(Z|\hat{\theta}_m(x),\hat{m}(x)),g(Z)) = E_{g(Z)}[\log \frac{g(Z)}{f(Z|\hat{\theta}_m(x),\hat{m}(x))}]. \quad (2)$$

L_1 をはじめとする予測の直接的な損失関数やその他の分布間の距離を用いた損失関数に対して，L_2 は漸近的に同等であったり，上界になることが知られている．そこで広義の予測の総合的損失関数として L_2 を考えて良さそうに思える．この損失から考えたモデル選択基準が AIC であり，先に挙げた FPE も C_p も漸近的に AIC と同等になることはよく知られている．

この損失関数 L_2 のデータ X に関する期待値である危険関数 R_2 を以下で定義しておく。

$$R_2(f(Z|\hat{\theta}_m(X),\hat{m}(X)),g(Z)) = E_{g(X)}E_{g(Z)}[\log \frac{g(Z)}{f(Z|\hat{\theta}_m(X),\hat{m}(X))}]. \quad (3)$$

3.2 与えられたデータに対しての行動

離散値 x_i の n 個の系列 $x^n: x_1 \cdots x_n$ を2値 $y_i \in \{0,1\}$ の系列 $C(x^n) = y^{l(x^n)}: y_1 \cdots y_{l(x^n)}$ に符号化

[1] ここでの対数 log の底は任意でかまわないが，3.2節以降では記述長としての解釈をあたえるため，特に断らない限り底は2とする．また，底が e の自然対数は ln で表記する．

して送ることを考える．ここで $l(x^n)$ は x^n の符号語 $C(x^n)$ の長さで x^n によって変わるものとする．符号語から元の系列 x^n が一意に復号される必要十分条件として，符号語長は $\sum_{x^n} 2^{-l(x^n)} \leq 1$, を満たさねばならないことが知られており，Kraftの不等式[7]と呼ばれている．

符号の良さを測る基本的尺度は平均符号長 $\bar{L}_C = E_{g(X^n)}[l(X^n)]$ である．この意味での最適符号の符号長の割り振りかたは $l(x^n) = -\log g(x^n)$ [2]とすればよいことが示されており[7]，最適符号はKraftの不等式を等式で満たしている．

最適な符号長は真の分布 $g(x^n)$ を用いて決定されているが，真の分布が未知の場合はどのように決定すればよいであろうか．Kraftの不等式を最適符号は等号で満たすことを考えると，ある確率 $P_C(x^n)$ を仮定することで，符号長 $l(x^n) = -\log P_C(x^n)$ とみなすことが可能である．

よって，この符号化のための確率 $P_C(x^n)$ を最適に決めることが最適な符号長の割り振りを決めることになる．以上より，符号の決定問題が確率モデルを選択する問題と同等であることが明らかになった．

上式で決めた符号長と最適な符号長の差を損失と考え，その期待値をとった次式の危険関数は冗長度と呼ばれている．

$$R_3(P_C(X^n), g(X^n)) = E_{g(X^n)}[\log \frac{g(X^n)}{P_C(X^n)}]. \quad (4)$$

この式もKullback-Leibler情報量を示すが，期待値をとっている確率変数が L_2 とは違っている．L_2 では X で推測したモデルに対し新たな確率変数 Z で期待値をとっているが，R_3 では X に対して期待値をとっていることになる．

この違いは x^t までのデータで x_{t+1} のデータの符号語を決める逐次的（予測）符号 $C(x_{t+1}|x^t)$ と，今まで述べた長さ n の系列を一括に符号化する符号とを比べればより明らかとなる．この予測符号化の X_{t+1} に関する冗長度は以下で定義され，X_{t+1} を Z に，X^t を X に書き換えれば R_2 と同等の損失になる．

$$R_4(P_C(X_{t+1}|X^t), g(X_{t+1}|X^t)) = E_{g(X^{t+1})}[\log \frac{g(X_{t+1}|X^t)}{P_C(X_{t+1}|X^t)}]. \quad (5)$$

[2]現実の符号長は自然数であるので正確には小数点以下を切り上げなければならないが，理論的考察において問題はないため実数の符号長を用いる．このような実数の符号長は理想符号長と呼ばれている．

符号化の確率が $P_C(X^n) = \prod_{t=0}^{n-1} P_C(X_{t+1}|X^t)$ を充たす場合，各 X_t の損失の n 回の累積が長さ n の系列 X^n 全体の損失となることから次の関係が成り立つ．

$$R_3(P_C(X^n), g(X^n)) = \sum_{t=0}^{n-1} R_4(P_C(X_{t+1}|X^t), g(X_{t+1}|X^t)). \quad (6)$$

以上より R_2 や R_4 の損失を累積したものが R_3 の損失となり，記述長の評価基準がある意味で累積予測誤差の評価基準となっていることが明らかとなった．

この節では，離散確率変数について議論したが，この記述長の基準は連続値にも拡張される．本来連続値の符号化には無限桁の精度が必要であるため，記述長の基準は現実的には意味を成さない．しかし，一つの意味付けとして累積予測誤差の基準と考えることができるであろう．

4. 予測を目的としたモデル選択

新たなデータに対する行動決定を目的とした危険関数 R_2 を最小化するモデル選択を考える．R_2 の $E_{g(Z)}[\ln g(Z)]$ の部分は比較対象の各モデルに共通なので，危険関数としては

$$R_2' = E_{g(X)}E_{g(Z)}[\ln f(Z|\hat{\theta}_m(X), \hat{m}(X))] \quad (7)$$

の最大化を考えればよいことになる．

しかし，これを直接求めることは困難であるため，漸近不偏推定量を求めることを考える．モデルを m に固定したときに，L_2 を最小とするパラメータを θ_m^* と定義する．$\hat{\theta}_m(x)$ として最尤推定量を用いることとし，$\hat{\theta}_m(x)$ と θ_m^* での漸近展開と，最尤推定量の分布が漸近的に正規分布に従うことを利用して次式が求まる．

$$R_2' = E_{g(X)}[\frac{1}{n}\ln f(X|\hat{\theta}_m(X), \hat{m}(X))] - trace I(\theta_m^*)^{-1} J(\theta_m^*), \quad (8)$$

ここで，$I(\theta_m^*) = -E_{g(X)}[\frac{\partial^2}{\partial\theta\partial\theta'}\ln f(X|\theta_m)|_{\theta=\theta_m^*}]$, $J(\theta_m^*) = E_{g(X)}[\frac{\partial}{\partial\theta}\ln f(X|\theta_m)\frac{\partial}{\partial\theta}\ln f(X|\theta_m)|_{\theta=\theta_m^*}]$.

よって，損失関数の不偏推定量として次式が求まる．

$$\ln f(Z|\hat{\theta}_m(x), \hat{m}(x)) - trace I(\theta_m^*)^{-1} J(\theta_m^*). \quad (9)$$

$I(\theta_m)$ はモデル m の Fisher 情報量であり，真の分布 $g(x)$ が仮定したモデルクラスに含まれる時は，$I(\theta_m) = J(\theta_m)$ が成り立ち，第2項はパラメータの

次元 k となる．この場合がAICであり，L_2 最小化の目的のために，このAICを最大化するモデルを選択することが提唱されている[1]．

第1項の対数尤度 $\ln f(x|\hat{\theta}_m(x),\hat{m}(x))$ でモデルの良さ R_2' を測ると過大評価してしまう点に問題があり，漸近的な偏りを第2項により補正していることになる．これは，$\hat{\theta}_m(x)$ と x は独立でないため，最尤推定量をパラメータとした分布関数では $\sum_x f(x|\hat{\theta}_m(x),\hat{m}(x))=1$ が成り立つとは限らず，確率とは見なせないことからもそれを窺い知ることができる．

AICの有用性は入れ子型モデルの選択を考えるとはっきりする．次数の高いモデルは低いモデルを含んでいるので，古典的な統計手法で用いられる対数尤度のみの比較では必ず高次のモデルが選択されてしまう．これに対してAICではパラメータの次数を第2項で引くことで高い次数のモデルを選択しにくくし，データへの当てはまりの良さとモデルの複雑性のトレードオフをうまく調整している．このトレードオフ関係を明確に式で表現したAICは，その後のモデル選択研究に大きな影響を与えた．

AICの直接の発展形として，真のモデルがモデルクラスに含まれない場合，式(9)の第2項は k とならないため，この項を何らかの漸近不偏推定量に置き換えたTIC[2]が提案されている．

AICは期待値が直接計算できないため漸近展開を用いたが，ブートストラップ法を基本とし，経験分布のリサンプリングを利用して期待値の近似計算をしてしまうことも考えられる．この考えをを発展させたものがEIC[3]で，この場合推定値として最尤推定量以外も用いることが可能である．

また，本節の目的である予測に適したモデルを選択する意味での直接的手法としてクロス・ヴァリデーションがである．クロス・ヴァリデーションはデータの一部からモデルのパラメータを推定し，それを用いて残りのデータを予測することによってモデルの良さを決定する手法である．この手法がある条件の下でAICと漸近的に同等であることが証明されている[14]．

5. 圧縮を目的としたモデル選択

データの圧縮に関する危険関数 R_3 の最小化を目的としたモデル選択について述べる．この目的に対して2段階符号化の考えを用いた，MDL(Minimum Description Length)基準がRissanenにより提案されている[11]．この節ではRissanenの考え方を簡単にまとめる．

2段階符号化は，まずどのモデルを仮定しているのかを受信者に知らせるために，モデル $\hat{m}(x^n)$ とパラメータ $\hat{\theta}_m(x^n)$ を送り，次にそのモデルを用いて求まるデータの確率 $P(x^n|\hat{\theta}_m(x^n),\hat{m}(x^n))$ を使ってデータを符号化し，データを送る方法である．Rissanenはこの符号化による x^n を送るための符号長を以下のように表している．

$$\begin{aligned} l(x^n) &= -\log P(x^n|\hat{\theta}_m(x^n),\hat{m}(x^n)) \\ &\quad +l(\hat{\theta}_m(x^n)|\hat{m}(x^n))+l(\hat{m}(x^n)), \quad (10)\end{aligned}$$

ここで，第2項と第3項はそれぞれパラメータとモデルのインデックスを送るための記述長である．

パラメータの推定量 $\hat{\theta}_m(x^n)$ として最尤推定量を用いることを前提とし，推定値を量子化して送信するとする．この仮定で量子化の幅 δ を大きくすると，推定値を上位の桁だけを有効数値として送ることとなり第2項の記述長は小さくなるが，第1項を最小化している最尤推定値とのずれが生じこの項は大きくなる．このトレードオフに対して，Rissanenは最適な量子化幅が $\delta=O(\frac{1}{\sqrt{n}})$ であること示している．

この量子化幅を用いて式(10)を漸近展開すると，以下の記述長が求まる(この式では $l(\hat{m}(x^n))$ は考慮していない)．

$$\begin{aligned} l_m(x^n) &= -\log P(x^n|\hat{\theta}_m(x^n),\hat{m}(x^n)) \\ &\quad +\frac{k}{2}\log n+O(1). \quad (11)\end{aligned}$$

この記述長を最小にするモデルが最適なモデルであるという基準が，初期のMDL基準である．

この基準の定性的主張は，モデルのデータに対する適合性である対数尤度とモデルの複雑度であるモデル自体の記述長を，総合的に符号長という尺度で判断している点にある．このMDLでもAICで主張されたトレードオフ関係が見て取れ，第2項に補正項が入っている．これは先にも述べたように，最尤推定量をパラメータとした分布関数で求まる符号化確率ではその和が1とならず，Kraftの不等式を満す一意復号可能な符号が構成できない．そのため補正項が必要になると考えることもできる．

MDLにはその後いろいろ発展形が提案されている．最近では量子化や2段階符号化にこだわらず，与えられたデータを符号化するための最小記述長とい

う概念をstochastic complexityと定義し，いろいろな応用が考えられている．

6. ベイズ理論からのモデル選択

ベイズ決定理論を用いて，最も確からしいモデルを選ぶという単純な発想からのモデル選択基準を考えると，以下の式を最大化する m を選ぶことになる．

$$\log \int P(x|\theta_m, m)\mu(\theta_m|m)P(m)d\theta, \tag{12}$$

ここで，μ はパラメータの事前分布，$P(m)$ はモデルの事前分布を表す．

上式は事後確率最大の m を選んでおり，m を推測したときの平均誤り確率最小という評価尺度でベイズ決定を行っていることになる．この式をSchwarzはモデルに指数分布族を仮定して漸近展開することにより，事前分布によらない(定数オーダでしか影響しなくなる)形のBIC(Bayes Information Criterion)[12]を求めた．

$$\log P(x^n|\hat{\theta}_m(x^n), \hat{m}(x^n)) - \frac{k}{2}\log n + O(1). \tag{13}$$

BICは形としては先に述べたMDLと同じになるが，なぜであろうか．モデルを m に固定して，危険関数 R_3 をベイズ的に最適にする符号化の確率は以下のように求まる．

$$P_C^B(x^n|m) = \int P(x|\theta_m, m)\mu(\theta_m|m)d\theta. \tag{14}$$

これは θ_m を事前分布で平均化して周辺分布をとったモデルが，R_3 の意味で最適であることを表している．

つまり式(12)は m をモデルとして選択した場合の最適記述長（累積予測損失）を評価していることにもなる．BICはMDLと同じ評価基準で最適化を行っているとも見なせ，類似の結果がでて当然ともいえる．

一般的に，ある正規性をみたす分布族に対して，さらに正確に次のような漸近展開式が求まっている[6]．

$$-\log P_C^B(x^n|m) = -\log P(x^n|\hat{\theta}_m(x^n), m) + \frac{k}{2}\log\frac{n}{2\pi} + \log\frac{\sqrt{\det I(\hat{\theta_m})}}{\mu(\hat{\theta}_m|m)} + o(1). \tag{15}$$

このベイス最適なモデル $P_C^B(x^n)$ は，パラメータに最尤推定量を代入した分布を使うMDLやAICとは異なり，周辺分布を用いることがベイズ的には最適であると主張している．

しかし事前分布が決まらないと式(14)は計算できない．そこで，ベイズ基準ではなくminimax基準から R_3 を最適化する符号化を考えることにする．

ベイズ危険を最大化する事前分布（最悪の事前分布）が存在すれば，その事前分布を用いたベイズ解はminimax解となる．この性質より，事前分布 $\mu(\theta_m|m) = \sqrt{\det I(\theta_m)}/\int\sqrt{\det I(\theta_m)}d\theta$ を用いて式(14)で符号化することが，危険関数 R_3 に関するminimax解であることが示されている．この符号化を用いた場合の符号長は以下のように示せる[5]．

$$-\log P_C^m(x^n|m) = -\log P(x^n|\hat{\theta}_m(x^n), m) + \frac{k}{2}\log\frac{n}{2\pi} + \log\int\sqrt{\det I(\theta_m)}d\theta + o(1). \tag{16}$$

このようにベイズ理論をおしすすめると，R_3 に対する完全なベイズ最適な決定は次式となり[10]，モデルを一つに選択せずに，考え得るすべてのモデルの重みづけ平均をとることが最適となる．

$$P_C^B(x^n) = \sum_m \int P(x^n|\theta_m, m)\mu(\theta_m|m)P(m)d\theta. \tag{17}$$

予測の最適化を図った予測符号 $P_C(X_{t+1}|x^t)$ も，同様にモデルを重み付けすることで求められる[10]．このように目的関数だけをベイズ最適から追い求めると，1つのモデルを選択するというモデル選択の一般的枠組みとは違う戦略が出てくる。

7. 各モデル選択基準の性質

この節では再び，1つのモデルを選択するという本来の話題に戻そう．幾つかの目的からモデル選択基準を眺めてきたが、得られた基準はAICと同様に対数尤度のみによる過大評価を何らかの補正項を用いて補正するものであった。

これらのモデル選択基準を一般的に表現すると以下の式となる。

$$\ln f(x|\hat{\theta}_m(x), \hat{m}(x)) - c(n)\cdot k. \tag{18}$$

このパラメータ次元に対する係数 $c(n)$ は例えばAICの場合 $c(n)=1$，BICの場合 $c(n)=\frac{1}{2}\log n$ である．

この節ではこの一般的表現を用いて，$c(n)$ の違いによるモデル選択基準の一致性と有効性について考えてみたい．

選択されたモデルが $n\to\infty$ で真のモデルに一致するかについては，次のような結果が得られている．真のモデルに確率収束するための $c(n)$ の必要十分条件は $c(n)\to\infty$ かつ $c(n)/n\to 0$ である．また，概収束するための必要十分条件は $\liminf c(n)/\ln\ln n > 1$ かつ $c(n)/n\to 0$ である[8]．

これより，AIC，TICのような$c(n)$が定数である選択基準には一致性がなく，MDL，BICのように$c(n)=O(\log n)$の基準には一致性があることがわかる．概収束を満足するぎりぎりをねらった選択基準がHQで$c(n)=\alpha \ln \ln n$，$\alpha>1$となっている[8]．

入れ子型のモデルクラスを考え，真のパラメータの次元をk_0とした場合，AICもBICもk_0未満の次数のモデルに収束する確率は0となるが，AICの場合，k_0より大きい次数のモデルに収束する確率が0にはならず，少し大き目の次数のモデルを選択する傾向があるといえる．

一致性は真のモデルが仮定したモデルクラスに含まれている場合は論じることができるが，含まれていない場合，モデルが無限個のパラメータを用いなければ表現できない場合は意味をなさない．そこで，無限次元のパラメータ$\boldsymbol{\beta}$で表される回帰モデル$\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}$について，線形モデルの平均値に対する平均2乗誤差$E_g[|\boldsymbol{X}_m\hat{\boldsymbol{\beta}}_m-\boldsymbol{X}\boldsymbol{\beta}|^2]$を考えてみよう．ここで$\hat{\boldsymbol{\beta}}_m$と$\boldsymbol{X}_m$はそれぞれモデル$m$に対応する$k$次だけをとりだしたパラメータベクトルと計画行列である．

選択対象のモデルクラスの次元としてはサンプル数n以下だけを考えればよい．漸近的にこの損失が下限と一致するのは$c(n)=1$のAICの場合だけであることが示されている[13]．つまりこのような条件の下ではAICは漸近有効性をもつことがわかる．

この節では一致性と有効性について，各モデル選択基準を横並べにして比較を行った．望ましい性質をすべて満足するような選択基準は残念ながら存在しない．しかし，これはある意味で当たり前の結論ともいえる．なぜならば，前節までで述べてきたように，それぞれのモデル選択基準はそれぞれの目的のもとで構成されているのであるから，それ以外の部分では必ずしも良い性質を持つとは限らないわけである．

AICやTICなどの$c(n)=O(1)$とした基準は，予測に対する良さをめざして構成された基準であるので，補正項が軽くなっている．このため，ある程度複雑なモデルも積極的取り入れることで，上の条件で有効性をもつ反面，一致性を持たなくなったといえる．

また，BICやMDLなどは$c(n)=O(\log n)$は記述長を短くすることをめざして構成された基準であるため，なるべくモデルを複雑にしないように制御する力が強いといえる．このため，一致性は充たすが，無限次元パラメータをもつ複雑な対象に対しては有効性を持たなくなる．

8. おわりに

本稿ではモデル選択の膨大な研究のほんの一部について，ある特定の視点からまとめたにすぎない．しかし，モデル選択を行う際，形式的な選択基準の適用ではなく，その対象問題をモデル化したい目的や背景，各モデル選択基準が導かれた条件や性質を考慮した上で，モデルクラスや選択基準を決めていただくための参考に少しでもなれば幸いである．

参考文献

[1] 坂元慶行, 石黒真木夫, 北川源四郎. 情報量統計学, 情報科学講座 *A-5-4*. 共立出版, 1983.

[2] 竹内啓. 情報統計量の分布とモデルの適切さの基準. 数理科学, (153):12–18, 1976.

[3] 北川源四郎, 石黒真木夫, 坂元慶行. 情報量基準AICとEIC. 電子情報通信学会 技術研究報告 *IT92*, 1993.

[4] H. Akaike. Fitting autoregressive model for prediction. *Ann. Inst. Statist. Math.*, 21:243–247, 1969.

[5] B. S. Clarke. Jeffreys' prior is asymptotically least favorable under entropy risk. *J. Stattistical Planning and Inference*, 41:37–60, 1994.

[6] B. S. Clarke and A. R. Barron. Information-theoretic asumptotics of Bayes methods. *IEEE Tranc. Inf. Theory*, 36(3):453–471, May 1990.

[7] R. G. Gallager. *Information theory and reliable communication*. Wiley, 1968.

[8] E. J. Hannan and B. G. Quinn. The determination of the order of an autoregression. *J. Roy. Statist. Soc.*, B 41:190–195, 1979.

[9] C. L. Mallows. Some comments on c_p. *Technometrics*, 15:661–675, 1973.

[10] T. Matsushima, H. Inazumi, and S. Hirasawa. A class of distortionless codes designed by Bayes decision theory. *IEEE Trans. Inf.Theory*, 37(5):1288–1293, Sep 1991.

[11] J. Rissanen. Universal coding, information, prediction and estimation. *IEEE Trans. Inf. Theory*, 30(4):629–636, July 1984.

[12] G. Schwarz. Estimating the dimension of a mode. *The Annals of Statistics*, 6(2):461–464, 1978.

[13] R. Shibata. An optimal selection of regression variables. *Biometrika*, 68:45–54, 1981.

[14] M. Stone. An asymptotic equivalence of choice of model by cross-validation and Akaike's Criterion. *J. Roy. Statist. Soc.*, B-39:44–47, 1977.