統計数理(2019)
第 67 巻 第 2 号 193–214

創立 75 周年記念号
[総合報告]

情報量規準 AIC の統計科学に果たしてきた役割

小西 貞則†

(受付 2019 年 1 月 21 日；改訂 4 月 12 日；採択 4 月 25 日)

要　　旨

情報量規準 AIC は，導出の基本概念を尤度原理と Kullback-Leibler 情報量に置き，これを予測という視点から理論を展開したことが本質的であった．モデリングの過程におけるモデルの評価と選択は，多様なモデルとその推定法が提唱される度に問題が提起され，AIC の基本的考え方を理論的・実際的側面から研究することによって，新たなモデル評価基準の提唱へと繋がっていった．本論文では，AIC の果たしてきた役割を概観し，一般に情報量規準と呼ばれるモデル評価基準がどのように提唱されてきたかを述べる．また，ベイズアプローチに基づく予測分布モデル等の評価を目的として提唱された，AIC に基礎を置く情報量規準についてもふれる．

キーワード：AIC，ABIC，BIC，DIC，GIC，PIC，TIC，WAIC.

1. はじめに

データの背後にある現象の解明と予測・制御，そして新たな知識発見のための基礎的な役割を担う現象のモデリングに，本質的な役割を果たしてきたのが情報量規準 AIC(Akaike Information Criterion)である．Akaike (1973, 1974)の提唱した AIC は，最尤法によって推定したモデルを確率分布で表現し，その良さを Kullback-Leibler 情報量(Kullback and Leibler, 1951)によって予測の視点から評価したことで，極めて適用範囲の広い柔軟な手法となり，諸分野の現象解明に大きく寄与してきた．赤池・北川 編 (1994, 1995)や Bozdogan ed. (1994), Parzen, Tanabe and Kitagawa (1998)には，自然科学はもとより社会科学の様々な分野で AIC が情報抽出や予測・制御にどのように寄与したかを紹介している．また，赤池弘次博士の第 22 回京都賞受賞を記念して 2007 年に出版された「赤池情報量規準 AIC」(室田・土谷 編, 2007)には，自らの言葉で情報量規準 AIC 導入に至る経緯とその効果について述べている．

情報量規準 AIC は，候補として挙げたモデル集合の中で，近似モデルの良さを相対比較することを目的とし，導出の基本概念を尤度原理と Kullback-Leibler 情報量に置き，これを予測という視点から理論を展開したことが本質的であった．これは，統計科学の尤度原理と情報科学の情報理論を融合することによって，モデルの評価と選択に新たな方向性を提起したといえる．

蓄積されたデータに内包される有用な情報を抽出，活用するため，これまでに様々なモデルとモデルの推定法が提唱されてきた．モデルの推定法という観点からみると，確率分布で表現されたモデルを，最尤法，正則化法，L_1 ノルム型正則化法，ベイズアプローチなど，それぞれの手法の特徴を考慮して推定する．さらに，モデリングの過程において重要な役割を果たすの

† 中央大学 理工学部：〒 112–8551 東京都文京区春日 1–13–27 (現 九州大学大学院 数理学研究院：〒 819–0395 福岡市西区元岡 744)

が，推定したモデルの評価と選択である．この問題に多くの研究者が取り組み，設定したモデルとその推定法に対応して AIC の基本理念を理論的・実際的側面から研究し，新たなモデル評価基準の提唱へと繋がって行った．

本稿では，AIC 導出の理論をもう一度振り返ってみることから始め，一連のモデリングのプロセスの中で，AIC の果たしてきた役割を概観し，一般に情報量規準と呼ばれるモデル評価基準がどのように提唱されてきたかを述べる．2 節で AIC 導出の過程を整理し，情報量規準と呼ばれるモデル評価基準を定式化する．3 節で，多種多様なモデルと推定法に対応して，AIC 導出の基本概念を展開して新たに提唱されたモデル評価基準について述べる．4 節では，ベイズアプローチによって構築されたモデルの評価を目的として，AIC の基本的な考え方に基づいて導出されたいくつかのモデル評価基準について述べる．5 節では，Akaike (1980b)の提唱した ABIC(Akaike Information Criterion)を紹介すると共に，AIC としばしば比較の対象として取り上げられる BIC (Schwarz, 1978)との相違点等についてふれる．6 節では，モデル選択の不確定性とそれに対処する一つの方法である Akaike ウェイト (Akaike, 1978b, 1979; Burnham and Anderson, 2002)について述べる．

2. 情報量規準

現象解明のためのモデリングは，当該分野の知識とデータをもとにモデル集合を想定し，この中から現象発生の確率的メカニズムを最もよく近似するモデルを評価し選択する．本節では，このモデルの評価・選択という問題に対して，情報量規準がどのように定式化されてきたかを，Akaike (1973, 1974)の基本的な考え方を踏襲して整理する．

2.1 AIC 導出の基本的考え方

いま，データ $\boldsymbol{y}=\{y_1, y_2, \ldots, y_n\}$ は，未知の密度関数 $g(y)$(確率分布関数 $G(y)$)に従って生成されたとする．データを発生した $g(y)$ は，真の分布，あるいは真のモデルと考える．観測された有限個のデータ $\boldsymbol{y}$ に内在する情報を抽出するために，確率分布によって表現されたモデル集合 $\{f(y|\boldsymbol{\theta}); \boldsymbol{\theta} \in \Theta \subset R^p\}$ を想定し，モデルに含まれる p 次元パラメータベクトル $\boldsymbol{\theta}$ を，推定量 $\hat{\boldsymbol{\theta}}$ で置き換えた $f(y|\hat{\boldsymbol{\theta}})$ で真のモデル $g(y)$ を近似する．推定したモデル $f(y|\hat{\boldsymbol{\theta}})$ は，データを発生した真のモデル $g(y)$ との近さを測ることによってその良さを評価する．Akaike (1973, 1974)は，分布間の距離を測る基準として Kullback-Leibler 情報量(K-L 情報量)を採用し，モデルの評価を予測の視点から捉えることによって AIC 導出に繋げた．これは，以下のように述べることができる．

推定したモデル $f(y|\hat{\boldsymbol{\theta}})$ とデータを発生した真のモデル $g(y)$ との距離は，予測の視点を入れて K-L 情報量で測るとき，次の式で与えられる．

$$I\{g(z), f(z|\hat{\boldsymbol{\theta}})\} = E_G\left[\log \frac{g(Z)}{f(Z|\hat{\boldsymbol{\theta}})}\right] = E_G[\log g(Z)] - E_G[\log f(Z|\hat{\boldsymbol{\theta}})]. \tag{2.1}$$

ここで，期待値は $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\boldsymbol{y})$ を固定して真の分布 G に関してとる．予測の視点とは，データ $\boldsymbol{y}$ とは独立に，真の分布からランダムに採られた将来のデータ $Z = z$ の従う分布 $g(z)$ を，モデル $f(z|\hat{\boldsymbol{\theta}})$ で近似したときの平均的な良さを測ることを意味する．

(2.1)式の K-L 情報量の右辺第一項 $E_G[\log g(Z)]$ は，個々のモデルに依存せず一定であることから，第 2 項の期待対数尤度と呼ばれる $E_G[\log f(Z|\hat{\boldsymbol{\theta}})]$ の値が大きいモデルほど真のモデルに近いといえる．期待対数尤度は，真のモデルに依存する未知の量である．そこで，n 個のデータを発生した未知の真の分布 G を，各データ点 y_i に確率 $1/n$ を付与した経験分布関数 $\hat{G}$ で推定する．これは，離散型確率分布の期待値をとることから

$$E_{\hat{G}}[\log f(Z|\hat{\boldsymbol{\theta}})] = \frac{1}{n}\sum_{i=1}^{n}\log f(y_i|\hat{\boldsymbol{\theta}}) = \frac{1}{n}\log f(\boldsymbol{y}|\hat{\boldsymbol{\theta}})$$

となり，モデル $f(y|\hat{\boldsymbol{\theta}})$ の対数尤度 $\ell_n(\hat{\boldsymbol{\theta}}) = \log f(\boldsymbol{y}|\hat{\boldsymbol{\theta}})$ が求まる．したがって，モデルの対数尤度は期待対数尤度 $(\times n)$ の一つの推定量である．

しかし，対数尤度は期待対数尤度$(\times n)$の推定量ではあるが，データ $\boldsymbol{y}$ とは独立に真のモデル g から発生した将来のデータ $\boldsymbol{z}$ に基づく対数尤度 $f(\boldsymbol{z}|\hat{\boldsymbol{\theta}}(\boldsymbol{y}))$ ではなく，モデルの推定に用いたデータ $\boldsymbol{y}$ を再び利用した $f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}(\boldsymbol{y}))$ で推定していることから，推定のバイアス

$$\log f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}(\boldsymbol{y})) - nE_G[\log f(Z|\hat{\boldsymbol{\theta}}(\boldsymbol{y}))] \tag{2.2}$$

を生じる原因となっている．これは，一般に $\log f(\boldsymbol{z}|\hat{\boldsymbol{\theta}}(\boldsymbol{y})) < \log f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}(\boldsymbol{y}))$ となることからも分かる．(2.2)式は，ある特定のデータ $\boldsymbol{y}$ に対するバイアスであるが，大きさ n のデータを g から繰り返し抽出したときの平均的なバイアスは

$$b(G) = E_{G(\boldsymbol{y})}[\log f(\boldsymbol{Y}|\hat{\boldsymbol{\theta}}(\boldsymbol{Y})) - nE_{G(z)}[\log f(Z|\hat{\boldsymbol{\theta}}(\boldsymbol{Y}))]] \tag{2.3}$$

で与えられる．ここで，期待値は $\boldsymbol{Y}$ の同時分布 $\prod_{i=1}^{n} g(y_i)$ に関してとる．したがって，このバイアスを何らかの方法で求めて，もし，バイアスがデータを生成した真の確率分布 G に依存していれば，$b(G)$ の一致推定量 $\hat{b}(G)$ で対数尤度のバイアスを補正した $\ell_n(\hat{\boldsymbol{\theta}}) - \hat{b}(G)$ が期待対数尤度$(\times n)$の推定量として求まる．一般に，-2 を掛けた

$$\mathrm{IC} = -2\log f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}) + 2\hat{b}(G) \tag{2.4}$$

を，K-L 情報量の推定量として導かれたモデル評価基準であることから情報量規準という．IC 値が小さいモデルほど K-L 情報量の値も小さく，真のモデルに近いといえる．

情報量規準 AIC は，最尤法によって推定したモデル $f(y|\hat{\boldsymbol{\theta}}_{ML})$ を評価するための基準で，期待対数尤度$(\times(-2n))$の近似推定量として導かれ，次の式で与えられた．

$$\mathrm{AIC} = -2\log f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_{ML}) + 2\ \text{（モデルの自由パラメータ数）} \tag{2.5}$$

ただし，$\hat{\boldsymbol{\theta}}_{ML}$ は $\boldsymbol{\theta}$ の最尤推定量とし，$\log f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_{ML})$ は n 次元データベクトル $\boldsymbol{y}$ に基づくモデルの最大対数尤度である．最大対数尤度で期待対数尤度を推定したとき，平均的にどの程度過大に推定しているかを表す(2.3)式のバイアスが，結果としてモデルの自由パラメータ数と一致することを示している．AIC の値を最小とするモデルを選択する方法は，AIC 最小化法と呼ばれている．

多数のパラメータで特徴付けられたモデルほど，観測したデータへのモデルの当てはまりはよい．しかし，複雑すぎるとモデルは将来の現象予測に有効に働かない．AIC は予測の観点から最適なモデルを選択するための評価基準で，モデルのデータへの適合度を最大対数尤度 $\log f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_{ML})$ で捉え，モデルの自由パラメータ数をモデルの複雑さに対するペナルティとして組み込んでいるといえる．

2.2 情報量規準の定式化

(2.3)式のバイアス補正項 $b(G)$ は，モデルを最尤法で推定するか，あるいは正則化法などで推定するかによって，また真のモデルと想定したモデルの関係をどう捉えるかによって異なる形をとる．いま，最尤法で推定したモデルを $f(y|\hat{\boldsymbol{\theta}}_{ML})$ とする．このとき，(2.3)式のバイアス $b(G)$ は，最尤推定量の漸近的性質(例えば，小西・北川, 2004, p.42)を用いると，データ数 n に対して漸近的に $b(G) = \mathrm{tr}\{J^{-1}(G)I(G)\}$ となる．ただし，$J(G)$, $I(G)$ は次式で定義される

$p \times p$ 行列とし，式中 $\partial/\partial\boldsymbol{\theta} = (\partial/\partial\theta_1, \ldots, \partial/\partial\theta_p)^T$ は転置ベクトルを表す．

$$J(G) = -E_G\left[\frac{\partial^2 \log f(Z|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right], \qquad I(G) = E_G\left[\frac{\partial \log f(Z|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial \log f(Z|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^T}\right]. \tag{2.6}$$

いま，$\hat{J}(G), \hat{I}(G)$ をそれぞれ $J(G)$, $I(G)$ の一致推定量とすると

$$\mathrm{TIC} = -2\sum_{i=1}^{n} \log f(y_i|\hat{\boldsymbol{\theta}}_{ML}) + 2\mathrm{tr}\{\hat{J}^{-1}(G)\hat{I}(G)\} \tag{2.7}$$

が求まる．これは，竹内 (1976)によって与えられ，情報量規準 TIC と呼ばれている．

ここで，想定したパラメトリックモデル $\{f(y|\boldsymbol{\theta}); \boldsymbol{\theta} \in \Theta\}$ の中に真のモデル $g(y)$ が含まれる，すなわち，ある $\boldsymbol{\theta}_0 \in \Theta$ に対して $g(y) = f(y|\boldsymbol{\theta}_0)(F(y|\boldsymbol{\theta}_0))$ となるものが存在すると仮定する．このとき，(2.6)式の期待値を $G = F$ でとると $J(F) = I(F)$ が成立し $\mathrm{tr}\{J^{-1}(F)I(F)\} = p$ となり，情報量規準 AIC が導かれる．$I(F)$ はフィッシャー情報行列である．AIC は，TIC の漸近バイアスをモデルの自由パラメータ数で近似した評価規準であるといえる．導出の詳細は，小西・北川 (2004, 3 章)，Konishi and Kitagawa (2008, Chapter 3)を参照されたい．

Akaike (1974)の論文では，想定したモデル集合の中に真のモデルは含まれていないという仮定のもとで AIC 最小化法を議論し，データを発生した真のモデルの近傍に適切にパラメトリックモデルを想定すれば，最尤法に基づくモデルの対数尤度のバイアスはモデルの自由パラメータ数で近似できると述べている．これによって，情報量規準 AIC は，個々のモデルに対して漸近バイアス $\mathrm{tr}\{\hat{J}^{-1}(G)\hat{I}(G)\}$ を解析的に導出する必要がなくなり，また，パラメータ数 p は当然未知の確率分布 G にも依存しないことから，バイアスの推定による変動も取り除かれ，適用上極めて柔軟な手法となったといえる．

情報量規準の構成においては，モデル $f(z|\hat{\boldsymbol{\theta}})$ の期待対数尤度 $E_G[\log f(Z|\hat{\boldsymbol{\theta}})]$ を予測の観点から推定することが本質的であった．これは，観測データ $\boldsymbol{y}$ に基づいて構築したモデルを，真のモデルからランダムに抽出した将来のデータ $\boldsymbol{z}$ でモデルを評価するという考え方を定式化したことで実現した．同様に，予測の観点から種々の予測誤差を捉えることを可能とした極めて汎用性の高い手法が，Stone (1974)によるクロス・バリデーション(Cross-Validation; 交差検証法)である．

クロス・バリデーションは，観測データ $\boldsymbol{y}$ のみに基づいて予測の観点からモデルを評価する方法で，モデルの推定に用いるデータとモデルの評価に用いるデータを分離して行う．クロス・バリデーションによると期待対数尤度は，

$$\sum_{i=1}^{n} \log f(y_i|\hat{\boldsymbol{\theta}}_{ML}^{(-i)}) \tag{2.8}$$

と推定される．ただし，$\hat{\boldsymbol{\theta}}_{ML}^{(-i)}$ は，n 個の観測データの中から i 番目のデータ y_i を取り除いた残りの $(n-1)$ 個のデータに基づく最尤推定値とする．Stone (1977)は，クロス・バリデーションによるモデル評価基準と AIC は漸近的に同等であることを示した．その証明は，候補モデル集合の中に真のモデルは含まれていないという仮定のもとで行っており，導出の過程で(2.7)式の TIC を与えている．

一般に，K-L 情報量に基づいて予測の視点から導かれる AIC タイプの情報量規準は，期待対数尤度の推定量を求める問題に帰着される．同様に，期待対数尤度を予測の視点からクロス・バリデーションによって推定したのが(2.8)式であった．このことから，Konishi and Kitagawa (2008, p.245)は，次節で述べる汎関数理論を用いることによって，AIC タイプの情報量規準はクロス・バリデーションと漸近的に同等であることを示した．

Sugiura (1978)は，ガウス型線形回帰モデルに対して，真のモデルが想定したモデルに含まれる場合に，(2.3)式の期待対数尤度のバイアスを精密に求め，修正情報量規準 $\mathrm{AIC_c}$ と呼ばれる評価基準を提唱した．Hurvich and Tsai (1989)，Fujikoshi and Satoh (1997)，藤越・杉山 (2012)，McQuarrie and Tsai (1998)等は，時系列モデル，多変量回帰モデルに対して正規性の仮定のもとでバイアス補正項を求めて，AIC を修正した形の情報量規準を提案している．修正情報量規準は，正規性等の条件下で求められているが，パラメータ数 p に比べてデータ数 n がそれほど多くないときは，実際上有効であることが数値的に検証されている(例えば，Burnham and Anderson, 2002 を参照されたい)．

3. 最尤法の枠組みを外した情報量規準

では，ロバスト推定，正則化最尤法など，最尤法を含むより広いクラスの推定法によって構築されたモデルの評価を可能とする情報量規準は，どのように構成すればよいであろうか．本節では，K-L 情報量の推定量として導かれたいくつかの情報量規準について述べる．

3.1 一般化情報量規準 GIC

最尤法を含むより広いクラスの推定法で構築したモデルの評価を，統計的汎関数に基づくアプローチによって可能にしたのが，一般化情報量規準 GIC(generalized information criterion: Konishi and Kitagawa, 1996)である．

データを発生した真のモデル $g(y)$ は，候補モデル集合 $\{f(y|\boldsymbol{\theta}); \boldsymbol{\theta} \in \Theta \subset R^p\}$ に含まれないとする．このとき，モデルのパラメータは $g(y)(G(y))$ に従って発生したデータによって推定される．そこで，一般にパラメータ θ_i の推定量 $\hat{\theta}_i$ は，確率分布 G の実数値関数，すなわちある統計的汎関数 $T_i(G)$ が存在して，n 個のデータそれぞれに等確率 $1/n$ をもつ経験分布関数 $\hat{G}$ に対して $\hat{\theta}_i = T_i(\hat{G})(i = 1, 2, \ldots, p)$ で与えられるとする．この $T_i(G)$ を第 i 要素とする p 次元汎関数ベクトルを $\boldsymbol{T}(G) = (T_1(G), \ldots, T_p(G))^T$ とすると，p 次元推定量は $\hat{\boldsymbol{\theta}} = \boldsymbol{T}(\hat{G})$ で与えられる．例えば，標本平均 $\overline{y}_n = n^{-1}\sum_{i=1}^n y_i$ を定義する汎関数は $T_\mu(G) = \int y dG(y)$ であり，この汎関数 T_μ によって $\overline{y}_n = T_\mu(\hat{G}) = \int y d\hat{G}(y)$ で与えられることが分かる．標本数 n を無限大とすると，経験分布関数 $\hat{G}$ は真の分布 G に法則収束することから，$\hat{\boldsymbol{\theta}} = \boldsymbol{T}(\hat{G})$ は $\boldsymbol{\theta} = \boldsymbol{T}(G)$ に対して一致性をもつ推定量である．

一般化情報量規準 GIC は，$f(y|\boldsymbol{\theta})$ のパラメータを汎関数で定義される推定量 $\hat{\boldsymbol{\theta}} = \boldsymbol{T}(\hat{G})$ で置き換えたモデル $f(y|\hat{\boldsymbol{\theta}})$ の評価基準で，次の式で与えられた．

$$\mathrm{GIC} = -2\sum_{i=1}^n \log f(y_i|\hat{\boldsymbol{\theta}}) + \frac{2}{n}\sum_{i=1}^n \mathrm{tr}\left\{ \boldsymbol{T}^{(1)}(y_i; \hat{G}) \frac{\partial \log f(y_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \right\}. \tag{3.1}$$

ただし，$\boldsymbol{T}^{(1)}(y; \hat{G})$ は，その第 i 要素 $T^{(1)}(y; \hat{G})$ が次の式で与えられる点 $\hat{G}$ での汎関数微分で，p 次元経験影響関数ベクトルと呼ばれる．

$$T_i^{(1)}(y; \hat{G}) = \lim_{\epsilon \to 0} \frac{T_i((1-\epsilon)\hat{G} + \epsilon\delta_y) - T_i(\hat{G})}{\epsilon}.$$

ここで，δ_y は点 y 上に確率 1 をもつ分布とする．影響関数は，ロバスト推定において，分布のわずかな変化に対して推定値がどれだけ変化するかを調べるために用いられた (Huber, 1981; Hampel et al., 1986)．

一般化情報量規準 GIC は，最尤法をはじめとしてロバスト推定法，様々な L_2 ノルム正則化項をもつ正則化最尤法などによって推定されたモデルの評価を可能とするモデル評価基準である．これらの推定量は，一般に標本空間とパラメータ空間の直積空間上で定義された実数値関

数 $\psi_i(y,\boldsymbol{\theta})$ に対して，次の同時方程式の解 $\hat{\boldsymbol{\theta}}$ として与えられる．

$$\sum_{i=1}^{n}\psi_j(y_i,\hat{\boldsymbol{\theta}})=0,\qquad j=1,2,\ldots,p. \tag{3.2}$$

ここで，$\boldsymbol{\psi}=(\psi_1,\psi_2,\ldots,\psi_p)^T$ とベクトル表示して，これを $\boldsymbol{\psi}$-関数と呼ぶ．最尤推定量 $\hat{\boldsymbol{\theta}}_{ML}$，正則化最尤推定量 $\hat{\boldsymbol{\theta}}_R$ は，それぞれ

$$\boldsymbol{\psi}_{ML}(y,\boldsymbol{\theta})=\frac{\partial\log f(y|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}},\qquad \boldsymbol{\psi}_R(y,\boldsymbol{\theta})=\frac{\partial\{\log f(y|\boldsymbol{\theta})-\lambda R(\boldsymbol{\theta})\}}{\partial\boldsymbol{\theta}} \tag{3.3}$$

としたときの解である．ただし，$R(\boldsymbol{\theta})$ は正則化項，$\lambda>0$ は正則化パラメータと呼ばれ，モデルのデータへの適合度と当てはめたモデルの滑らかさを連続的に調整する役割を果たす．

この $\boldsymbol{\psi}$-関数に対して，(3.1)式の GIC の影響関数は

$$\boldsymbol{T}^{(1)}(y,G)=J(\boldsymbol{\psi},G)^{-1}\boldsymbol{\psi}(y,G) \tag{3.4}$$

で与えられる．ただし，$J(\boldsymbol{\psi},G)$ は，次式で与えられる $p\times p$ 行列で，(2.6)式の行列 $J(G)$ に相当する．

$$J(\boldsymbol{\psi},G)=-E_G\left[\frac{\partial\boldsymbol{\psi}(Z,\boldsymbol{\theta})^T}{\partial\boldsymbol{\theta}}\right].$$

ここで，(3.4)式の影響関数を(3.1)式の GIC へ代入すると，(3.2)式の同時方程式の解として与えられる推定量 $\hat{\boldsymbol{\theta}}$ に基づくモデル $f(y|\hat{\boldsymbol{\theta}})$ の評価基準

$$\mathrm{GIC_R}=-2\sum_{i=1}^{n}\log f(y_i|\hat{\boldsymbol{\theta}})+2\mathrm{tr}\{J(\boldsymbol{\psi},\hat{G})^{-1}I(\boldsymbol{\psi},\hat{G})\} \tag{3.5}$$

が求まる．ただし，$I(\boldsymbol{\psi},G)$ は

$$I(\boldsymbol{\psi},G)=E_G\left[\boldsymbol{\psi}(Z,G)\frac{\partial\log f(Z|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^T}\right]$$

で与えられる $p\times p$ 行列で，これは(2.6)式の行列 $I(G)$ に対応する．

情報量規準 $\mathrm{GIC_R}$ のバイアス補正項の推定値は，一般に実数値関数 $h(z|\boldsymbol{\theta})$ の期待値 $E_G[h(Z|\boldsymbol{\theta})]$ $(\boldsymbol{\theta}=\boldsymbol{T}(G))$ を，経験分布関数 $\hat{G}$ に関する期待値 $E_{\hat{G}}[h(Z|\hat{\boldsymbol{\theta}})]=n^{-1}\sum_{i=1}^{n}h(y_i|\hat{\boldsymbol{\theta}})$ $(\hat{\boldsymbol{\theta}}=\boldsymbol{T}(\hat{G}))$ で推定した結果を用いている．

特別な場合として，(3.3)式の $\boldsymbol{\psi}_{ML}$ を(3.5)式へ代入すると最尤法に基づく TIC が求まる．さらに，Fisher 一致性の概念 $(\boldsymbol{T}(F_{\boldsymbol{\theta}})=\boldsymbol{\theta};F_{\boldsymbol{\theta}}=F(y|\boldsymbol{\theta}))$ を適用することによって，M 推定などのロバスト推定に対しても AIC のバイアス補正項であるモデルの自由パラメータ数に対応する結果が求まり，AIC は M 推定量に基づくモデルの評価基準へと自然に拡張される（小西・北川, 2004, p.77; Konishi and Kitagawa, 2008, p.131）．GIC の導出とその応用および精密化については，Konishi and Kitagawa（1996），Konishi（1999, 2002），Konishi and Kitagawa（2003），小西・北川（2004, 4 章），Konishi and Kitagwa（2008, Chapter 5）を，統計的汎関数については，von Mises（1947），Fernholz（1983）などを参照されたい．

確率過程に対する情報量規準は，Uchida and Yoshida（2001, 2004）によって与えられた．Lv and Liu（2014）は，モデル集合を一般化線形モデル（McCullagh and Nelder, 1989）として，候補モデル集合の中には真のモデルは含まれないという仮定のもとで，AIC タイプのモデル評価基準を求めた．結果は，(2.6)式の行列 J,I に対応するものを一般化線形モデルのもとで求めているが，GIC の特別な場合と考えられる．Shen and Ye（2002），Shen, Huang and Ye（2004）

は，それぞれガウス分布と指数型分布族に対して，期待対数尤度の近似的に不偏な推定量として導いた適応型モデル評価基準を提唱した．これらは，AIC 導出の基本概念から導かれたものであるが，汎用性という点では問題が残る．

3.2 正則化法と平滑化パラメータの選択

非線形回帰モデルの関数推定に対しては，最尤法は有効に機能しない場合が多く，このため対数尤度に曲線(曲面)の局所変動の程度を考慮に入れた正則化最尤法(罰則付き最尤法)が用いられる．その際，平滑化パラメータ(正則化パラメータ)がモデルの複雑さの程度を調整し，データへの過適合による汎化能力の低下を抑制する働きをする．本節では，非線形回帰モデリングの過程で本質的な平滑化パラメータの選択に用いられてきたモデル評価基準について述べる．

いま，目的変数 y と p 次元説明変数 $\boldsymbol{x}$ に関して観測された n 組のデータ集合に，回帰モデル $y = u(\boldsymbol{x};\boldsymbol{\beta}) + \varepsilon$ を当てはめるとする．現象の平均構造を捉える回帰関数 $u(\boldsymbol{x};\boldsymbol{\beta})$ に対して，スプライン，B-スプライン，動径関数などを仮定してモデル化する．これらのモデルを統一的に表すと，回帰関数を非線形関数 $b_j(\boldsymbol{x})$ の線形結合とした

$$y = \sum_{j=1}^{m} \beta_j b_j(\boldsymbol{x}) + \varepsilon, \qquad \varepsilon \sim F(\varepsilon) \tag{3.6}$$

で与えられ，基底展開法に基づく非線形回帰モデルと呼ばれる(例えば，Hastie, Tibshirani and Friedman, 2009, 5 章; 小西, 2010, 3 章)．

基底展開に基づく非線形回帰モデルは，対数尤度関数にペナルティ項(正則化項)を課した正則化最尤法，すなわち $\log f(\boldsymbol{y}|\boldsymbol{\beta}) - \lambda R_n(\boldsymbol{\beta})$ の最大化によって推定する．正則化項 $R_n(\boldsymbol{\beta})$ としては，関数の曲率を考慮した 2 階微分の積分の離散近似，パラメータ $\boldsymbol{\beta}$ の差分や 2 乗和等が説明変数の次元と分析目的に応じて用いられる(小西・北川, 2004, p.92)．正則化法は，Good and Gaskins (1971)によって密度推定の枠組みで提唱され，その後，縮小推定量や本稿 5 節で述べるように，ベイズモデルとの関係が明らかにされた (Akaike, 1980b; Kitagawa and Gersch, 1984, 1996; Shibata, 1989)．

正則化最尤法によって推定したモデルの複雑さの程度は，平滑化パラメータ λ に加えて基底関数の個数 m にも依存する．そのため，平滑化の程度を調整するこれらのパラメータの値を決める問題をモデル選択として捉え，AIC に基づく様々なモデル評価基準が提唱された．Hastie and Tibshirani (1990)は，AIC のバイアス補正項である自由パラメータ数を，基底関数の個数と平滑化パラメータを含む有効自由度 (effective degrees of freedom)で置き換えたモデル評価基準を提唱した．その後，ガウス型線形回帰モデルの枠組みで求められた修正情報量規準 $\mathrm{AIC_c}$ (Sugiura, 1978)に含まれる変数の個数を有効自由度で置き換えた評価基準も提唱された (Hurvich, Simonoff and Tsai, 1998, 等)．しかし，限られた設定のもとでの数値比較の有効性は認められるが，理論的整合性には課題が残る．

これに対して，汎関数の枠組みで導出した GIC の特別な場合として与えられた(3.5)式の $\mathrm{GIC_R}$ へ，正則化最尤推定量を与える(3.3)式の $\boldsymbol{\psi}_R$ を代入すると，平滑化パラメータ λ をもつ正則化最尤法に基づくモデルの評価基準が求まる．この結果を用いて，基底展開法に基づく非線形回帰モデルを正則化最尤法によって推定したときの平滑化パラメータの選択，基底関数の個数を決める評価基準を導出してモデリングに組み込んだ解析手法が提案された(Imoto and Konishi, 2003; Ando, Konishi and Imoto, 2008; Kawano and Konishi, 2011; Tateishi and Konishi, 2011; Kawano, Misumi and Konishi, 2012; Park and Konishi, 2017 等)．また，$\mathrm{GIC_R}$ は，関数データ解析 (Ramsay and Silverman, 2005)において，経時的に観測・測定されたデータの関数

化にも適用された（Araki et al., 2009a, 2009b; Kayano, Dozono and Konishi, 2010; Matsui and Konishi, 2011 等）．

AIC に基づくモデル評価基準は，一般化加法モデル GAM（generalized additive model; Hastie and Tibshirani, 1990）における Wood, Pya and Säfken（2016）や混合効果モデルに対する Liang, Wu and Zou（2008），Yu and Yau（2012），Misumi and Konishi（2016）など，様々な手法のモデリングの過程で用いられて，モデルの評価と選択に貢献してきた．

Shibata（1989）は正則化法によるモデルとその評価について議論し，Regularized Information Criterion（RIC）を提唱した．Murata, Yoshizawa and Amari（1994）は，ニューラルネットワークモデルの最適なパラメータ数，あるいは隠れ層の個数の決定を目的とした Network Information Criterion（NIC）を提唱した．さらに，正則化項を考慮した損失関数に基づくモデルの推定と評価を議論している．これらは，それぞれのモデリングの目的に合わせて，AIC 導出の基本的な考え方を踏襲して提唱されたモデル評価基準である．

3.3 スパースモデリング

データ数に比してモデルのパラメータ数が大幅に上回る大規模モデリングでは，モデルの推定とモデルの評価を分離して行うことの限界が指摘された．一つは，候補となるモデルが多数に上ることによる計算量の限界，一つはモデル選択の信頼性（Brieman, 1996）などが挙げられる．このような状況の中で回帰モデリング，特に，線形回帰モデルの推定と変数選択に新たな方向性を示したのが，lasso（least absolute shrinkage and selection operator; Tibshirani, 1996）であった．これは，損失関数に回帰係数の絶対値（L_1 ノルム）の和を正則化項として付与した推定法で，その特徴はモデルの推定と変数選択を同時に実行できる点にあった．このため，高次元線形回帰モデルに対する有効なモデリングとして注目を集め，様々な L_1 型正則化線形回帰モデリング（スパースモデリング）の研究が急速に進展した（川野 他, 2010; Konishi, 2014, Section 2.3; Hastie, Tibshirani and Wainwright, 2015; 廣瀬, 2016; 川野・松井・廣瀬, 2018 等）．

スパースモデリングでは，調整パラメータ λ の値の増加に伴って，回帰係数の推定値は 0 へと縮退する．基本的には，調整パラメータの値を与えたもとでモデルをスパース推定し，その結果 0 でない回帰係数の推定値に対応する説明変数の個数をモデルの自由パラメータ数として AIC や 5 節（5.2）式の BIC を用いて評価するプロセスを繰り返すことは可能である．この方法に対して，様々なスパース推定法の特徴，データ数とパラメータ数との関係やモデル選択の一致性等を考慮した理論研究が進展し，新たなモデル評価基準が提唱された．

Efron et al.（2004），Zou, Hastie and Tibshirani（2007）は，Stein のリスク不偏推定の枠組みで lasso に対してモデルの自由度を与え，AIC, BIC, Mallows（1973）の C_p に基づいた評価基準を検討した．Kato（2009）は微分幾何学的アプローチによって，より広い lasso タイプの自由度の不偏推定について議論した．モデルの自由度については，Ye（1998），Efron（2004）を併せて参照されたい．Zhang, Li and Tsai（2010），Fan and Tang（2013）は，AIC のバイアス項の 2 と対応する BIC の $\log n$ を，データ数 n に依存する正の実数列で置き換えてモデルの複雑さを制御することで，調整パラメータの選択を議論している．Hirose, Tateishi and Konishi（2013）は，様々なスパース回帰モデリングに対する自由度を数値的に計算するアルゴリズムを提唱し，AIC, 修正情報量規準 $\mathrm{AIC_c}$，BIC, Mallows' C_p などに基づくモデル評価基準による調整パラメータの選択法を与えた．Ninomiya and Kawano（2016）と Umezu et al.（2019）は，それぞれ lasso と bridge（Frank and Friedman, 1993），SCAD（smoothly clipped absolute deviation; Fan and Li, 2001）などの非凸正則化法に対して，一般化線形モデルの枠組みで AIC 導出の基本概念に基づいてモデル評価基準を提唱した．

BIC は，候補モデル集合に真のモデルは含まれているとしたとき一致性をもち，しかも AIC

よりはより単純なモデルを選択する傾向にある．このような理由により，スパースモデリングの調整パラメータの選択に，BIC をもとにしたモデル評価基準が提唱されている．Wang, Li and Tsai (2007)は，SCAD の調整パラメータの選択に対して(5.2)式の BIC の自由パラメータ数をモデルの自由度で置き換えた評価基準を提唱した．Wang, Li and Leng (2009)は，lasso，SCAD を含む L_1 正則化法に対して，BIC を基準とした調整パラメータ選択法に対して理論的整合性を議論している．

3.4 ブートストラップ情報量規準

前節までに述べた情報量規準は，データ発生の確率構造とモデル推定に関して，それぞれ異なる条件下で漸近理論に基づいて導出された．それに対して，ブートストラップ情報量規準は，個々のモデルの対数尤度のバイアスをブートストラップ法 (Efron, 1979)を適用して数値的に近似したものである (Ishiguro, Sakamoto and Kitagawa, 1997; Konishi and Kitagawa, 1996)．なお，本節ではデータ $\boldsymbol{y}$ とブートストラップ標本 $\boldsymbol{y}^*$ の違いをモデルの中で示すため，推定量 $\hat{\boldsymbol{\theta}}$ を $\hat{\boldsymbol{\theta}}(\boldsymbol{y})$ と表す．

情報量規準構成においては，推定したモデル $f(y|\hat{\boldsymbol{\theta}}(\boldsymbol{y}))$ の期待対数尤度 $nE_G[\log f(Z|\hat{\boldsymbol{\theta}}(\boldsymbol{y}))]$ を対数尤度 $\log f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}(\boldsymbol{y}))$ で推定したときのバイアスの補正が本質的であった．ブートストラップ法の基本的な考え方は，未知の確率分布 G からの標本 $\boldsymbol{y} = \{y_1, \ldots, y_n\}$ に基づく推測過程を，データから推定した既知の確率分布である経験分布関数 $\hat{G}$ からの標本であるブートストラップ標本 $\boldsymbol{y}^* = \{y_1^*, \ldots, y_n^*\}$ に置き換えて実行する点にある．このため，ブートストラップ標本 $\boldsymbol{y}^*$ に基づいて推定したモデルを $f(y|\hat{\boldsymbol{\theta}}(\boldsymbol{y}^*))$ とする．

次に経験分布関数を真の分布としたときの $f(y|\hat{\boldsymbol{\theta}}(\boldsymbol{y}^*))$ の期待対数尤度は，$\hat{G}$ が n 個の各データに等確率 $1/n$ をもつ離散型確率分布の確率分布関数であることから

$$E_{\hat{G}}[\log f(Z|\hat{\boldsymbol{\theta}}(\boldsymbol{y}^*))] = \int \log f(z|\hat{\boldsymbol{\theta}}(\boldsymbol{y}^*))d\hat{G}(z) = \frac{1}{n}\sum_{i=1}^{n} \log f(y_i|\hat{\boldsymbol{\theta}}(\boldsymbol{y}^*)) = \frac{1}{n}\log f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}(\boldsymbol{y}^*))$$

となる．一方，期待対数尤度の一つの推定量である対数尤度は，モデルをブートストラップ標本によって推定し，推定したモデル $f(y|\hat{\boldsymbol{\theta}}(\boldsymbol{y}^*))$ の評価を再びブートストラップ標本を用いて行うことから，$\log f(\boldsymbol{y}^*|\hat{\boldsymbol{\theta}}(\boldsymbol{y}^*))$ で与えられる．従って，ブートストラップ法によって期待対数尤度を対数尤度で推定したときのバイアスは，

$$E_{\hat{G}}[\log f(\boldsymbol{y}^*|\hat{\boldsymbol{\theta}}(\boldsymbol{y}^*)) - \log f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}(\boldsymbol{y}^*))]$$

と推定される．

この期待値は，$\hat{G}$ が既知の確率分布(経験分布関数)であることを利用して，モンテカルロ法によって数値的に近似できるところにブートストラップ法の最大の特徴がある．すなわち，経験分布関数からの大きさ n の標本の反復抽出とは，観測データからの大きさ n の標本の復元抽出の反復と同値(小西・越智・大森, 2008, p.9)であることを利用して

$$b(\hat{G}) \approx \frac{1}{B}\sum_{i=1}^{B}\{\log f(\boldsymbol{y}^*(i)|\hat{\boldsymbol{\theta}}^*(i)) - \log f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}^*(i))\} := b_B(\hat{G})$$

と近似する．ただし，$\boldsymbol{y}^*(i)$ は i 番目のブートストラップ標本，$\hat{\boldsymbol{\theta}}^*(i)$ は i 番目のブートストラップ標本に基づく推定値とする．このとき，対数尤度のバイアスを補正した情報量規準 EIC (extended information criterion)は，

$$\mathrm{EIC} = -2\sum_{i=1}^{n} \log f(y_i|\hat{\boldsymbol{\theta}}) + 2b_B(\hat{G}) \tag{3.7}$$

で与えられる．

ブートストラップ法は，実行プロセスの中で解析的アプローチを，観測データ自身を反復抽出(リサンプリング)するというモンテカルロ計算法で置き換えたことにより，極めて緩やかな仮定のもとで，より複雑な問題に適用できる柔軟な統計手法となった．しかし，バイアス推定の標本変動に加えて，ブートストラップリサンプリングによる変動が生じることから，バイアス項の差異でモデルの違いを見るときには十分注意を払う必要がある．このブートストラップバイアス推定の確率変動を減少させるための方法が，Konishi and Kitagawa (1996)，Kitagawa and Konishi (2010)によって提案された．また，Konishi and Kitagawa (1996)は，ブートストラップバイアス推定および変動減少法の理論的整合性を汎関数の枠組みで証明した．

4. ベイズモデルの評価基準

本節では，ベイズアプローチによって構築されたモデルの評価を目的として，AIC 導出の基本的な考え方を踏襲して提唱されたいくつかのモデル評価基準について述べる．

データ $\boldsymbol{y}$ を発生した真のモデル $g(y)$ に対して，想定したモデル集合を $\{f(y|\boldsymbol{\theta}); \boldsymbol{\theta} \in \Theta \subset R^p\}$ とし，パラメータベクトル $\boldsymbol{\theta}$ の事前分布を $\pi(\boldsymbol{\theta})$ とする．このとき，データ $\boldsymbol{y}$ に対する $\boldsymbol{\theta}$ の事後分布は，

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}} \tag{4.1}$$

である．さらに，真のモデル g からランダムに抽出された将来のデータ z に対して，データ $\boldsymbol{y}$ を与えたもとで，モデル $f(z|\boldsymbol{\theta})$ の事後分布に関する期待値

$$h(z|\boldsymbol{y}) = E_{\pi(\boldsymbol{\theta}|\boldsymbol{y})}[f(z|\boldsymbol{\theta})] = \int f(z|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta} \tag{4.2}$$

として与えられるのが予測分布である．

4.1 ベイズ予測分布の情報量規準

予測分布モデルの評価を K-L 情報量に基づいて行うとき，期待対数尤度 $E_G[\log h(Z|\boldsymbol{y})]$ の推定が本質的となる．AIC の場合と同様に未知の確率分布 G を経験分布関数 $\hat{G}$ で置き換えると，$E_{\hat{G}}[\log h(Z|\boldsymbol{y})] = n^{-1}\sum_{i=1}^{n} \log h(y_i|\boldsymbol{y}) = n^{-1}\log h(\boldsymbol{y}|\boldsymbol{y})$ が求まる．従って，予測分布モデルの期待対数尤度を対数尤度で推定したときのバイアスは

$$b_{\mathrm{pred}}(G) = E_{G(\boldsymbol{y})}[\log h(\boldsymbol{Y}|\boldsymbol{Y}) - nE_{G(z)}[\log h(Z|\boldsymbol{Y})]]$$

で与えられ，バイアスを補正した

$$\mathrm{IC}_{\mathrm{pred}} = -2\log h(\boldsymbol{y}|\boldsymbol{y}) + 2b_{\mathrm{pred}}(\hat{G}) \tag{4.3}$$

が予測分布に対する情報量規準である（Akaike, 1980a)．

Konishi and Kitagawa (1996, p.878)は，汎関数の枠組みでこのバイアスを求めて，予測分布モデルに対する情報量規準を導出した．さらに，積分のラプラス近似 (Tierney and Kadane, 1986; Davison, 1986)を用いて，最尤法によって推定したモデル $f(z|\hat{\boldsymbol{\theta}}_{ML})$ に対して，予測分布を $h(z|\boldsymbol{y}) = f(z|\hat{\boldsymbol{\theta}}_{ML}) + O_p(n^{-1})$ と近似して，TIC，AIC と同様の情報量規準が求まることを示した．Kitagawa (1997)は，モデルと事前分布がともに多変量正規分布と仮定した線形ガウス

型ベイズモデルの情報量規準を導出し，これを PIC（predictive information criterion）と呼んだ．

4.2 逸脱度情報量規準 DIC

Spiegelhalter et al.（2002）は，ベイズの観点から AIC と同タイプのモデル評価基準を提唱し，これを DIC（deviance information criterion; 逸脱度情報量規準）と呼んだ．AIC の評価の対象とするモデルは，最尤法によって推定したモデル $f(z|\hat{\boldsymbol{\theta}}_{ML})$ であった．これに対して，DIC は最尤推定量をパラメータのベイズ推定量である事後平均 $\hat{\boldsymbol{\theta}}_B = E_{\pi(\boldsymbol{\theta}|\boldsymbol{y})}[\boldsymbol{\theta}|\boldsymbol{y}]$ で置き換えたモデル $f(z|\hat{\boldsymbol{\theta}}_B)$ を評価の対象とした．モデル $f(z|\hat{\boldsymbol{\theta}}_B)$ と真のモデルとの近さを，K-L 情報量で予測の観点から測ったとき，AIC のモデルの自由パラメータ数に対応する有効自由度を次の式で与えた．

$$b_{\mathrm{DIC}} = 2\{\log f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_B) - E_{\pi(\boldsymbol{\theta}|\boldsymbol{y})}[\log f(\boldsymbol{y}|\boldsymbol{\theta})]\}$$

従って，バイアスを補正した DIC は

$$\mathrm{DIC} = -2\log f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_B) + 2b_{\mathrm{DIC}} \tag{4.4}$$

で与えられる．

一般に，最大対数尤度 $\log f(\boldsymbol{y}|\hat{\boldsymbol{\theta}})$ がモデルのデータへの当てはまりの良さを表すのに対して，逆に当てはまりの悪さを表す $-2\log f(\boldsymbol{y}|\hat{\boldsymbol{\theta}})$ を逸脱度という．DIC の $-2\log f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_B)$ は事後平均に対する逸脱度に基づいており，この意味で逸脱度情報量規準と呼ばれている．

DIC の有効自由度 b_{DIC} は，ベイズモデルの設定によっては負の値を採ることもあり，このため有効自由度を事後分布に関する $\log f(\boldsymbol{y}|\boldsymbol{\theta})$ の分散 $b_{\mathrm{DICa}} = 2\mathrm{Var}_{\pi(\boldsymbol{\theta}|\boldsymbol{y})}\{\log f(\boldsymbol{y}|\boldsymbol{\theta})\}$ とした DIC も提案されている（Gelman et al., 2013）．

4.3 情報量規準 WAIC

Spiegelhalter et al.（2014）では，DIC の果たしてきた役割を再考するとともに，いくつかのデメリットも議論されている．これに対して，Watanabe（2009, 2010）の提唱した WAIC（widely applicable information criterion）は，ベイズモデリングの過程の中にマルコフ連鎖モンテカルロ法による予測分布の積分計算を組み込んだ汎用性の高い情報量規準として用いられている．

WAIC は，(4.2)式の予測分布に対して期待対数尤度 $\sum_{i=1}^{n} E_G[\log h(z_i|\boldsymbol{y})]$ を

$$\sum_{i=1}^{n} \log h(y_i|\boldsymbol{y}) = \sum_{i=1}^{n} \log \int f(y_i|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}$$

で推定したときのバイアスを

$$b_{\mathrm{WAIC}} = 2\left\{\sum_{i=1}^{n} \log h(y_i|\boldsymbol{y}) - \sum_{i=1}^{n} E_{\pi(\boldsymbol{\theta}|\boldsymbol{y})}[\log f(y_i|\boldsymbol{\theta})]\right\}$$

で与えた．この結果，ベイズ型予測分布モデルの評価を可能とする WAIC は，

$$\mathrm{WAIC} = -2\sum_{i=1}^{n} \log h(y_i|\boldsymbol{y}) + 2b_{\mathrm{WAIC}} \tag{4.5}$$

で与えられた．その特徴は，事後分布 $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ から反復発生させた $\boldsymbol{\theta}_s$ によって

$$b_{\mathrm{WAIC}} = 2\sum_{i=1}^{n}\left\{\log\left(\frac{1}{S}\sum_{s=1}^{S} f(y_i|\boldsymbol{\theta}_s)\right) - \frac{1}{S}\sum_{s=1}^{S}\log f(y_i|\boldsymbol{\theta}_s)\right\}$$

とシミュレーションによって求めることができ，また特異モデルにも適用可能であることにある．さらに，DIC の事後分布に関する分散に基づく有効自由度に対応する

$$b_{\mathrm{WAICa}} = \sum_{i=1}^{n} \mathrm{Var}_{\pi(\boldsymbol{\theta}|\boldsymbol{y})}\{\log f(y_i|\boldsymbol{\theta})\} = \sum_{i=1}^{n} [E_{\pi(\boldsymbol{\theta}|\boldsymbol{y})}[\{\log f(y_i|\boldsymbol{\theta})\}^2] - \{E_{\pi(\boldsymbol{\theta}|\boldsymbol{y})}[\log f(y_i|\boldsymbol{\theta})]\}^2]$$

で補正した WAIC を提唱した．

5. 赤池ベイズ情報量規準 ABIC とベイズ型モデル評価基準 BIC

AIC としばしば比較の対象として取り上げられるモデル評価基準が，Schwarz (1978)によって提唱された BIC である．BIC は，候補モデル集合の中でベイズ理論による事後確率が最も高くなるようなモデルを最適なモデルとして選択するもので，K-L 情報量に基づいて予測の視点から導出された AIC とは考え方を異にしている．従って，BIC は，Bayesian information criterion というよりも，ベイズ型モデル評価基準と呼ぶ方が適切である．

一方，Akaike (1980b)の提唱した ABIC(Akaike Bayesian information criterion)は，AIC の考え方に基づいてベイズモデルの評価を目的としたもので，この点でベイズ情報量規準といえる評価基準である．本節では，AIC，ABIC と BIC の違いを見るために BIC 導出の概略を示した後，ABIC について述べる．また，モデル選択の一致性についてふれる．

5.1 モデルの事後確率と BIC

いま，r 個のモデルの候補を $M_1, M_2, \ldots, M_r$ とし，各モデル M_i は確率分布モデル $f_i(y|\boldsymbol{\theta}_i)$ $(\boldsymbol{\theta}_i \in \Theta_i \subset R^{p_i})$ とパラメータベクトル $\boldsymbol{\theta}_i$ の事前分布 $\pi_i(\boldsymbol{\theta}_i)$ によって特徴付けられているとする．ベイズアプローチでは，データ $\boldsymbol{y}$ が観測されたとき，パラメータベクトル $\boldsymbol{\theta}_i$ の事前分布 $\pi_i(\boldsymbol{\theta}_i)$ で積分した

$$p_i(\boldsymbol{y}) = \int f_i(\boldsymbol{y}|\boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i$$

を評価の対象とする．この $p_i(\boldsymbol{y})$ は，データ $\boldsymbol{y}$ がモデル M_i から観測される尤もらしさを表しており，周辺尤度あるいは周辺分布と呼ばれる．

次に，i 番目のモデルが生起する事前確率を $P(M_i)$ とすると，i 番目のモデルの事後確率は，ベイズの定理より

$$P(M_i|\boldsymbol{y}) = \frac{p_i(\boldsymbol{y})P(M_i)}{\sum_{j=1}^{r} p_j(\boldsymbol{y})P(M_j)}, \qquad i = 1, 2, \ldots, r \tag{5.1}$$

で与えられる．この事後確率は，データ $\boldsymbol{y}$ が観測されたとき，そのデータが i 番目のモデルから生起する確率であり，従って，r 個のモデルの中から一つのモデルを選択するとき，事後確率最大のモデルを採用するのが自然である．これは，(5.1)式の分母がすべてのモデルに共通であることから，分子の $p_i(\boldsymbol{y})P(M_i)$ を最大にするモデルを選択することと同等である．さらに，事前確率 $P(M_i)$ はすべて等しいとした場合には，データの周辺尤度 $p_i(\boldsymbol{y})$ を最大にするモデルを選択することになる．

Schwarz (1978)の提唱した BIC は，積分で表された周辺尤度を，積分のラプラス近似 (Barndorff-Nielsen and Cox, 1989, p.169; 宮田, 2018)を用いて近似した結果得られたもので，一般に，周辺尤度 $p(\boldsymbol{y})$ に対して次の式で与えられた．

$$-2\log p(\boldsymbol{y}) = -2\log\left\{\int f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}\right\} \approx -2\log f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_{ML}) + p\log n. \tag{5.2}$$

ただし，$\hat{\boldsymbol{\theta}}_{ML}$ はモデル $f(y|\boldsymbol{\theta})$ の p 次元パラメータベクトル $\boldsymbol{\theta}$ の最尤推定値である．BIC の値を最小とするモデルを最適なモデルとして選択する．導出の詳細は，小西・北川 (2004, p.155)，Konishi and Kitagawa (2008, p.215)を参照されたい．

導出方法からも分かるように，BIC は想定したモデル集合の中に真のモデルが含まれるという仮定を陽に置いているわけではない．事前確率は，想定した各モデルが真のモデルである確率を与えていると考えることができる．しかし，導出の過程ですべて等しいとしており，またモデルの複雑さを表すパラメータ数の項には，パラメータの事前分布の情報も現れていない．この問題に対して，積分のラプラス近似を精密化して導いたのが，一般化ベイズ型モデル評価基準 GBIC (Konishi, Ando and Imoto, 2004)である．

5.2 赤池ベイズ情報量規準 ABIC

ABIC は，季節調整法の大規模モデリングの過程で用いられた正則化法をベイズの観点から捉えて，ベイズモデリングに本質的な評価基準として提唱されたもので，以下のように述べることができる．

一般に正則化最尤法は，対数尤度関数に正則化項あるいはペナルティ項と呼ばれる制約 $R(\boldsymbol{\theta},\boldsymbol{\lambda})$ を課した正則化対数尤度関数

$$\log f(\boldsymbol{y}|\boldsymbol{\theta}) - \frac{1}{2}R(\boldsymbol{\theta},\boldsymbol{\lambda})$$

の最大化によってパラメータを推定する．ここで，$\boldsymbol{\lambda}$ $(\in \Lambda \subset R^q; q < p)$ は q 次元パラメータとし，正則化法の枠組みでは平滑化パラメータに相当する．この式は，

$$\log f(\boldsymbol{y}|\boldsymbol{\theta}) + \log\left\{\exp\left(-\frac{1}{2}R(\boldsymbol{\theta},\boldsymbol{\lambda})\right)\right\} = \log\left\{f(\boldsymbol{y}|\boldsymbol{\theta})\exp\left(-\frac{1}{2}R(\boldsymbol{\theta},\boldsymbol{\lambda})\right)\right\}.$$

と書き表すことができる．さらに，指数関数の項を規格化して確率分布 $\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})$ とすると，これはハイパーパラメータ $\boldsymbol{\lambda}$ によって規定される $\boldsymbol{\theta}$ の事前分布である．このとき，データ $\boldsymbol{y}$ の周辺分布あるいは周辺尤度は

$$p(\boldsymbol{y}|\boldsymbol{\lambda}) = \int f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})d\boldsymbol{\theta}$$

で与えられる．

ベイズモデルの周辺分布 $p(\boldsymbol{y}|\boldsymbol{\lambda})$ を，ハイパーパラメータ $\boldsymbol{\lambda}$ を持つパラメトリックモデルと考えると，モデルの評価は次の AIC の枠組みで捉えることができ，その評価基準は

$$\text{ABIC} = -2\left\{\max_{\boldsymbol{\lambda}} \log p(\boldsymbol{y}|\boldsymbol{\lambda})\right\} + 2q = -2\max_{\boldsymbol{\lambda}} \log\left\{\int f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})d\boldsymbol{\theta}\right\} + 2q \tag{5.3}$$

で与えられる．この評価基準は，Akaike (1980b)によって提唱され，赤池ベイズ情報量規準 (Akaike Bayesian information criterion)と呼ばれている．この方法によると，ベイズモデルのハイパーパラメータ $\boldsymbol{\lambda}$ は，対数周辺尤度 $\log p(\boldsymbol{y}|\boldsymbol{\lambda})$ の最大化，すなわち，最尤法によって推定しており，また，ハイパーパラメータによって特徴付けられた複数のベイズモデルの相対的な良さを比較するには，ABIC 最小化法によってモデルを選択すればよい．

このようにして推定したハイパーパラメータを $\hat{\boldsymbol{\lambda}}$ とすると，パラメータ $\boldsymbol{\theta}$ の事前分布 $\pi(\boldsymbol{\theta}|\hat{\boldsymbol{\lambda}})$ に対して，$\boldsymbol{\theta}$ の事後分布

$$\pi(\boldsymbol{\theta}|\boldsymbol{y};\hat{\boldsymbol{\lambda}}) = \frac{f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\hat{\boldsymbol{\lambda}})}{\int f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\hat{\boldsymbol{\lambda}})d\boldsymbol{\theta}}$$

が求まる．パラメータ $\boldsymbol{\theta}$ の推定値としては通常，事後分布のモード，すなわち $\pi(\boldsymbol{\theta}|\boldsymbol{y};\hat{\boldsymbol{\lambda}}) \propto f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\hat{\boldsymbol{\lambda}})$ を最大とする $\hat{\boldsymbol{\theta}}$ が用いられる．ABIC 最小化法によるモデリングは，周辺分布 $p(\boldsymbol{y}|\boldsymbol{\lambda})$ に対して最尤法によるハイパーパラメータの推定とモデル選択を行ったあと，パラメータ $\boldsymbol{\theta}$ の事後分布 $\pi(\boldsymbol{\theta}|\boldsymbol{y};\hat{\boldsymbol{\lambda}})$ の最大化によって $\boldsymbol{\theta}$ の推定値を求めていることが分かる．

ABIC 最小化法は，当初経済データの季節調整法の開発に用いられた（Akaike, 1980b, 1980c; Akaike and Ishiguro, 1980）．その後，コホート分析（中村, 1982），2 値回帰モデル（Ishiguro and Sakamoto, 1983），地震学（Ogata, Katsura and Tanemura, 2003; Ogata, 2004; 尾形, 2015），状態空間モデリング（Kitagawa, 1987, 1996, 1998）など様々な分野のモデリングに利用されている．

5.3 モデル選択の一致性

AIC と BIC は，それぞれ K-L 情報量とベイズ事後確率という異なる基準に基づいて導出されたものであり，適用上異なる性質をもつことは明らかである．一般に，理論的側面からは，BIC はモデル選択に対して一致性を有していることが強調され，これに対して，AIC は ミニマックス最適性を有することを両者の違いとして見ることが多い．

Shibata（1976）は，AIC 最小化法によって次数選択を行う場合，標本数 n を無限に大きくするとき，真の次数を選択する確率は 1 とならない，すなわち次数選択の一致性を有しないことを示している．また，Shibata（1983）は，想定したモデルの中に真のモデルが含まれない場合，AIC によって選択されたモデルは，平均 2 乗誤差の意味で漸近的に最適であることを示した．Nishii（1984）は，ガウス型線形回帰モデルに対して AIC, BIC を含むいくつかの評価基準の一致性をもつ条件等について検討した．Akaike（1978a）は，AIC 最小化法のベイズの観点からの捉え方を述べ，多変量ガウス分布モデルに対して，等事前確率の仮定の下で AIC は一致性は有しないが，ミニマックス解を与えることを示した．AIC タイプのモデル選択基準に対するミニマックス最適性と関連論文については，Yang（2005）を参照されたい．

AIC と BIC を比べるとき，モデル選択の一致性という観点からしばしば議論されてきた．一致性は，真のモデルが想定したモデル集合に含まれているという仮定のもとでの証明であることもあり，これに対して様々な指摘がなされた．Burnham and Anderson（2002），小西・北川（2004, p.66），甘利（2007, 第 II 編 1 章, p.66），北川（2007, 第 II 編 2 章, p.83），Konishi and Kitagawa（2008, p.73）は，一致性をめぐる議論に否定的で，実際問題に対峙し現象解明に向けてモデリングを行うという観点から，それぞれの見解を明確に述べている．

実際，モンテカルロ・シミュレーションによって，AIC，BIC 等のモデル評価基準を比較検証するとき，データを発生させるモデルを真のモデルとすることが多い．さらに，真のモデルは想定したモデルの中に含まれ，しかも単純なモデルからデータを発生させることから，BIC との比較において一致性を有しない AIC が劣るという検証結果がしばしば見受けられる．このような設定のもとでのモデル評価基準の比較は，現象をモデル化するという立場からみると必ずしも実際的ではないことが指摘されている．赤池（1995, p.199），Burnham and Anderson（2002, p.20）は，真のモデルとは，本来，無限次元であり，従ってそれを観測された有限個のデータから完全に再現することは現実的でない．そこでモデル集合を想定して，その中から最良の近似モデルを選択するという考えに基づいており，少なくとも想定したモデル集合の中には，データを発生した真のモデルは含まれていないとしている．このような点を踏まえて，Burnham and Anderson（2002）は実データの解析，シミュレーションを通して，AIC の有用性を様々な角度から検証している．

6. モデル選択の不確定性と Akaike ウェイト

情報量規準によるモデル選択では，その値によってデータから構築した複数の候補モデルを順位付けして，データ発生の確率構造を最も良く近似するモデルを一つ選択する．しかし，評価基準はデータに依存しており，それ自身確率変数であることからモデル選択に起因する不確定性を生じる．したがって，評価基準値の近い他の候補モデルとの差をどのように解釈するかという問題に繋がる．

Akaike (1978a, 1978b, 1979, 1983)は，(2.5)式の AIC に対して $(-1/2)$ AIC が期待対数尤度 $(\times n)$の漸近的不偏推定量であることから

$$\exp\left(-\frac{1}{2}\mathrm{AIC}\right) = \exp\{\log f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_{ML}) - p\} = f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_{ML})e^{-p}$$

は最尤法によって推定したモデルの尤度とみなせ，また近似的に事後分布を表しているとした．したがって，この式を利用して AIC 値にそれほど大きな差のない候補モデル集合に対しては，平均化によって融合したモデルの構築を示唆している（Akaike, 1978b, 1979; Bozdogan, 1987）．この考え方を発展させて Akaike (1978b)，Burnham and Anderson (2002)は，最小 AIC 値と各候補モデルの AIC 値との差に基づくモデルの相対的な確からしさを，尤度と関連づけて，以下のように定式化している．

いま，データ $\boldsymbol{y}$ に基づいて推定したモデル集合を $\{f_i(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_i); i = 1, \ldots, r\}$ とする．各モデルに対する AIC 値を AIC_i とし，最小 AIC 値を $\mathrm{AIC}_{\min}$ とおく．このとき，両者の差

$$\Delta\mathrm{AIC}_i = \mathrm{AIC}_i - \mathrm{AIC}_{\min}, \qquad i = 1, \ldots, r$$

は，最小 AIC 値を基にしたモデルの相対的な比較を表している．さらに，

$$\exp\left(-\frac{1}{2}\Delta\mathrm{AIC}_i\right)$$

は，データが与えられたとき，モデル i の近似的な尤度を表すことから，この AIC 値の差を基準化した次の式は Akaike ウェイトと呼ばれる（Burnham and Anderson, 2002, p.75）．

$$w_i = \frac{\exp\left(-\frac{1}{2}\Delta\mathrm{AIC}_i\right)}{\sum_{k=1}^{r}\exp\left(-\frac{1}{2}\Delta\mathrm{AIC}_k\right)}$$

ウェイトの総和は $\sum_{i=1}^{r} w_i = 1$ であることから，w_i はモデル集合に属する各モデルの相対的な確からしさの指標となっている．Burnham and Anderson (2002)の提唱したマルチモデル推測とは，Akaike ウェイトを複数のモデルの相対的な良さとして用いて，モデル集合に基づく推論を実行する方法であるといえる．AIC 最小化法によって選択された一つのモデルが，明らかに他のモデルより現象予測に優れたモデルであればよいが，そうでない場合には順序付けられた複数のモデルを融合してモデルの推測を行う方が有効であり，Akaike ウェイトはそのための基準指標を示すと考えることができる．

マルチモデル推測は，モデル選択の不確定性に対処するための一つのアプローチであるといえる．同様に，2 つのモデルの AIC の差の有意性の検定や信頼集合を構成して不確定性を検証する方法（Linhart, 1988; Shimodaira, 1997），さらに観測データからブートストラップ標本を反復抽出してモデル選択を繰り返し実行し，各モデルが選択される頻度を推定するブートストラップ選択確率に基づく方法などが提唱されている．モデル選択の不確定性については，Kishino and Hasegawa (1989)，Shimodaira and Hasegawa (1999)，Burnham and Anderson (2002, Chapter 6)，下平 (2004, 2007)を参照されたい．

7. おわりに

本稿では，AIC 導出の理論を振り返ってみることから始め，一連のモデリングのプロセスの中で AIC の果たしてきた役割を再考し，一般に情報量規準と呼ばれるモデル評価基準がどのように提唱されてきたかを述べた．

情報量規準 AIC 導出の基本的考え方は，最尤法によって推定したモデルを確率分布で表現し，それを Kullback-Leibler 情報量によって予測の視点から評価したことであった．この基本概念は，データからの情報を確率分布モデルで捉え，当該分野で蓄積された知識を事前分布としてベイズ理論によってモデルに同化させた予測分布モデルの評価・選択へと繋がった．1980 年に提唱された赤池情報量規準 ABIC は，5.2 節で述べたようにモデルのパラメータ数がデータ数を超えるような大規模モデリングの過程で用いられた正則化法をベイズの観点から捉えて，ベイズモデリングに本質的な評価基準として提唱されたものであった．本稿は，理論的側面から情報量規準について述べてきたが，AIC, ABIC は，自然科学はもとより社会科学の様々な分野で，現象解明のためのモデリングに重要な役割を果たしてきた．

近年，諸科学・産業界では，計測・測定技術，計算機関連技術の高度な発展によって，大規模・高次元データが獲得，蓄積され，次々とデータベース化されつつある．特に，少数かつ高次元データ，大量かつ超高次元データからの効率的な情報抽出技術の開発研究が強く希求され，国際的に研究が推進されている．このような状況の中で，正則化法，L_1 ノルム正則化法，ベイズモデルは，現象分析のための重要な解析手法として用いられ，適用上の問題点の克服と汎化能力の向上を目指して，AIC タイプのモデル評価基準が提唱されてきたといえる．

情報量規準 AIC の理論は，これからますます多様な形式で獲得される複雑な大規模データの背後に潜む有益な情報やパターンを高効率に抽出・処理するための新しいモデリングの中でも活かされるものと思われる．同時に，過去から学び未来を予測する技術を獲得するためには，新たな発想でモデリングに取り組む必要があることも確かである．

謝　辞

統計数理「創立 75 周年記念号」への発表の機会を与えて下さいました統計数理研究所特任教授の田村義保先生に感謝致します．東京大学特任教授の北川源四郎先生には，数々の貴重なご意見, ご指摘を賜りました．また，査読者には，貴重なご指摘をいただきました．ここに記して厚く御礼申し上げます．

参 考 文 献

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, *2nd International Symposium on Information Theory* (eds. B. N. Petrov and F. Csaki), 267-281, Akademiai Kiado, Budapest. (Reproduced in *Breakthroughs in Statistics, Vol.1, Foundations and Basic Theory* (eds. S. Kotz and N.L. Johnson), Springer-Verlag, New York, (1992).)

Akaike, H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, **AC-19**(6), 716-723.

Akaike, H. (1978a). A Bayesian analysis of the minimum AIC procedure, *Annals of the Institute of Statistical Mathematics*, **30**, 9-14.

Akaike, H. (1978b). On the likelihood of a time series model, *The Statistician*, **27**, 217-235.

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

Akaike, H. (1980a). On the use of predictive likelihood of a Gaussian model, *Annals of the Institute of Statistical Mathematics*, **32**, 311-324.

Akaike, H. (1980b). Likelihood and the Bayes procedure, *Bayesian Statistics* (eds. J. M. Bernardo, M. H. De Groot, D. V. Lindley and A. F. M. Smith), 143-166 (discussion 185-203), University Press, Valencia, Spain.

Akaike, H. (1980c). Seasonal adjustment by a Bayesian modeling, *Journal of Time Series Analysis*, **1**, 1-13.

Akaike, H. (1983). Information measures and model selection, *International Statistical Institute*, **44**, 277-291.

赤池弘次 (1995).『時系列解析の心構え，時系列解析の実際 II』(赤池弘次, 北川源四郎 編), 第 12 章, 朝倉書店, 東京.

Akaike, H. and Ishiguro, M. (1980). A Bayesian approach to the trading-day adjustment of monthly data, *Time Series Analysis* (eds. O. D. Anderson and M. R. Perryman), 213-226, North-Holland, Amsterdam.

赤池弘次, 北川源四郎 編 (1994).『時系列解析の実際 I』, 朝倉書店, 東京.

赤池弘次, 北川源四郎 編 (1995).『時系列解析の実際 II』, 朝倉書店, 東京.

甘利俊一 (2007).『赤池情報量規準—その思想と新展開, 赤池情報量規準 AIC—モデリング・予測・知識発見—(室田一雄, 土谷 隆 編)』, 第 II 編, 1 章, 52-78, 共立出版, 東京.

Ando, T., Konishi, S. and Imoto, S. (2008). Nonlinear regression modeling via regularized radial basis function networks, *Journal of Statistical Planning and Inference*, **138**, 3616-3633.

Araki, Y., Konishi, S., Kawano, S. and Matsui, H. (2009a). Functional regression modeling via regularized Gaussian basis expansions, *Annals of the Institute of Statistical Mathematics*, **61**, 811-833.

Araki, Y., Konishi, S., Kawano, S. and Matsui, H. (2009b). Functional logistic discrimination via regularized basis expansions, *Communications in Statistics — Theory & Methods*, 2944-2957.

Barndorff-Nielsen, O. E. and Cox, D. R. (1989). *Asymptotic Techniques for Use in Statistics*, Chapman and Hall, New York.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions, *Psychometrika*, **52**, 345-370.

Bozdogan, H. (ed.) (1994). *Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling*: *An Informational Approach*, Kluwer Academic Publishers, the Netherlands.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection, *Annals of Statistics*, **24**, 2350-2383.

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference*: *A Practical Information — Theoretical Approach*, Springer-Verlag, New York.

Davison, A. C. (1986). Approximate predictive likelihood, *Biometrika*, **73**, 323-332.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife, *Annals of Statistics*, **7**, 1-26.

Efron, B. (2004). The estimation of prediction error: Covariance penalties and cross-validation, *Journal of the American Statistical Association*, **99**, 619-632.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression, *Annals of Statistics*, **32**, 407-499.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalised likelihood and its oracle properties, *Journal of the American Statistical Association*, **96**, 1348-1360.

Fan, Y. and Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood, *Journal of the Royal Statistical Society*, **B75**, 531-552.

Fernholz, L. T. (1983). *von Mises Calculus for Statistical Functionals*, Lecture Notes in Statistics, **19**, Springer-Verlag, New York.

Frank, I. and Friedman, J. (1993). A statistical view of some chemometrics regression tools, *Technometrics*, **35**, 109-148.

Fujikoshi, Y. and Satoh, K. (1997). Modified AIC and C_p in multivariate linear regression, *Biometrika*, **84**, 707-716.

藤越康祝, 杉山高一 (2012).『多変量モデルの選択』, 朝倉書店, 東京.

Gelman, A., Carlin, J. C., Stern, H. and Dunson, D. B. (2013). *Bayesian Data Analysis*, 3rd ed., Chapman and Hall/CRC, New York.

Good, I. J. and Gaskins, R. A. (1971). Nonparametric roughness penalties for probability densities, *Biometrika*, **58**, 255-277.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics, The Approach Based on Influence Functions*, Wiley, New York.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*, Chapman and Hall, London.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning*, 2nd ed., Springer, New York.

Hastie, T., Tibshirani, R. and Wainwright, M. (2015). *Statistical Learning with Sparsity*, Chapman & Hall/CRC, New York.

廣瀬 慧 (2016). スパースモデリングとモデル選択, 電子情報通信学会誌, **99**, 392-399.

Hirose, K., Tateishi, S. and Konishi, S. (2013). Tuning parameter selection in sparse regression modeling, *Computational Statistics & Data Analysis*, **59**, 28-40.

Huber, P. J. (1981). *Robust Statistics*, Wiley, New York.

Hurvich, C. M. and Tsai, C. L. (1989). Regression and time series model selection in small samples, *Biometrika*, **76**, 297-307.

Hurvich, C. M., Simonoff, J. S. and Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion, *Journal of the Royal Statistical Society*, **B60**, 271-293.

Imoto, S. and Konishi, S. (2003). Selection of smoothing parameters in B-spline nonparametric regression models using information criteria, *Annals of the Institute of Statistical Mathematics*, **55**, 671-687.

Ishiguro, M. and Sakamoto, Y. (1983). A Bayesian approach to binary response curve estimation, *Annals of the Institute of Statistical Mathematics*, **35**, 115-137.

Ishiguro, M., Sakamoto, Y. and Kitagawa, G. (1997). Bootstrapping log likelihood and EIC, an extension of AIC, *Annals of the Institute of Statistical Mathematics*, **49**, 411-434.

Kato, K. (2009). On the degrees of freedom in shrinkage estimation, *Journal of Multivariate Analysis*, **100**, 1338-1352.

Kawano, S. and Konishi, S. (2011). Semi-supervised logistic discrimination via regularized Gaussian basis expansions, *Communications in Statistics — Theory and Methods*, **40**, 2412-2423.

川野秀一, 廣瀬 慧, 立石正平, 小西貞則 (2010). 回帰モデリングと L_1 型正則化法の最近の展開, 日本統計学会誌, **39**, 211-242.

Kawano, S., Misumi, T. and Konishi, S. (2012). Semi-supervised logistic discrimination via graph-based regularization, *Neural Processing Letters*, **36**, 203-216.

川野 秀一, 松井秀俊, 廣瀬 慧 (2018).『スパース推定法による統計モデリング』, 共立出版, 東京.

Kayano, M., Dozono, K. and Konishi, S. (2010). Functional cluster analysis via orthonormalized Gaussian basis expansions and its application, *Journal of Classification*, **27**, 211-230.

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

Kitagawa, G. (1987). Non-Gaussian state-space modeling of nonstationary time series (with discussion), *Journal of the American Statistical Association*, **82**, 1032-1063.

Kitagawa, G. (1996). Monte Carlo filter and smoother for non-Gaussian nonlinear state space models, *Journal of Computational and Graphical Statistics*, **5**, 1-25.

Kitagawa, G. (1997). Information criteria for the predictive evaluation of Bayesian models, *Communications in Statistics — Theory and Methods*, **26**, 2223-2246.

Kitagawa, G. (1998). Self-organizing state space model, *Journal of the American Statistical Association*, **93**, 1203-1215.

北川源四郎 (2007).『情報量規準と統計的モデリング，赤池情報量規準 AIC—モデリング・予測・知識発見—(室田一雄, 土谷 隆 編)』, 第 II 編, 2 章, 79-109, 共立出版, 東京.

Kitagawa, G. and Gersch, W. (1984). A smoothness priors-state space modeling of time series with trend and seasonality, *Journal of the American Statistical Association*, **79**, 378-389.

Kitagawa, G. and Gersch, W. (1996). *Smoothness Priors Analysis of Time Series*, Lecture Notes in Statistics, **116**, Springer-Verlag, New York.

Kitagawa, G. and Konishi, S. (2010). Bias and variance reduction techniques for bootstrap information criteria, *Annals of the Institute of Statistical Mathematics*, **62**, 209-234.

Konishi, S. (1999). Statistical model evaluation and information criteria, *Multivariate Analysis, Design of Experiments and Survey Sampling* (ed. S. Ghosh), 369-399, Marcel Dekker, New York.

Konishi, S. (2002). Theory for statistical modeling and information criteria — functional approach, *Sugaku Expositions*, **15**(1), 89-106, American Mathematical Society.

小西貞則 (2010).『多変量解析入門—線形から非線形へ—』, 岩波書店, 東京.

Konishi, S. (2014). *Introduction to Multivariate Analysis*: *Linear and Nonlinear Modeling*, Chapman & Hall/CRC, New York.

Konishi, S. and Kitagawa, G. (1996). Generalized information criteria in model selection, *Biometrika*, **83**, 875-890.

Konishi, S. and Kitagawa, G. (2003). Asymptotic theory for information criteria in model selection — functional approach, *Journal of Statistical Planning and Inference*, **114**, 45-61.

小西貞則, 北川源四郎 (2004).『情報量規準』, 朝倉書店, 東京.

Konishi, S. and Kitagawa, G. (2008). *Information Criteria and Statistical Modeling*, Springer, New York.

Konishi, S., Ando, T. and Imoto, S. (2004). Bayesian information criterion and smoothing parameter selection in radial basis function network, *Biometrika*, **91**, 27-43.

小西貞則, 越智義道, 大森裕浩 (2008).『計算統計学の方法—ブートストラップ，EM アルゴリズム，MCMC—』, 朝倉書店, 東京.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency, *Annals of Mathematical Statistics*, **22**, 79-86.

Liang, H., Wu, H. and Zou, G. (2008). A note on conditional AIC for linear mixed-effects models, *Biometrika*, **95**, 773-778.

Linhart, H. (1988). A test whether two AICs differ significantly, *South African Statistical Journal*, **22**, 153-161.

Lv, J. and Liu, J-S. (2014). Model selection principles in misspecified models, *Journal of the Royal Statistical Society*, **B76**, 141-167.

Mallows, C. L. (1973). Some comments on C_p, *Technometrics*, **15**, 661-675.

Matsui, H. and Konishi, S. (2011). Variable selection for functional regression models via the L_1 regularization, *Computational Statistics and Data Analysis*, **55**, 3304-3310.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed., Chapman & Hall/CR, London.

McQuarrie, A. D. R. and Tsai, C.-L. (1998). *Regression and Time Series Model Selection*, World Scientific, Singapore.

Misumi, T. and Konishi, S. (2016). Mixed effects historical varying coefficient model for evaluating dose-response in flexible-dose trials, *Journal of the Royal Statistical Society* (*Applied Statistics*), **C65**, 331-344.

宮田庸一 (2018). ラプラス近似のベイズ統計学への応用とその周辺, 数学, **70**, 275-295.

Murata, N., Yoshizawa, S. and Amari, S. (1994). Network information criterion-determining the number of hidden units for an artificial neural network model, *IEEE Transactions on Neural Networks*, **6**, 865-872.

室田一雄, 土谷 隆 編 (2007).『赤池情報量規準 AIC—モデリング・予測・知識発見—』, 共立出版, 東京.

中村 隆 (1982). ベイズ型コウホート・モデル—標準コウホート表への適用—, 統計数理研究所彙報, **29**, 77-97.

Ninomiya, Y. and Kawano, S. (2016). AIC for the Lasso in generalized linear models, *Electronic Journal of Statistics*, **10**, 2537-2560.

Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression, *Annals of Statistics*, **12**, 758-765.

Ogata, Y. (2004). Space-time model for regional seismicity and detection of crustal stress changes, *Journal of Geophysical Research*, **109**(B3), B03308, doi:10.1029/2003JB002621.

尾形良彦 (2015). 地震の確率予測の研究—その展望, 統計数理, **63**(1), 3-27.

Ogata, Y., Katsura, K. and Tanemura, M. (2003). Modelling heterogeneous space-time occurrences of earthquakes and its residual analysis, *Applied Statistics*, **52**, 499-509.

Park, H. and Konishi, S. (2017). Principal component selection via adaptive regularization method and generalized information criterion, *Statistical Papers*, **58**, 147-160.

Parzen, E., Tanabe, K. and Kitagawa, G. (1998). *Selected Papers of Hirotugu Akaike*, Springer, New York.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*, 2nd ed., Springer, New York.

Schwarz, G. (1978). Estimating the dimension of a model, *Annals of Statistics*, **6**, 461-464.

Shen, X. and Ye, J. (2002). Adaptive model selection, *Journal of the American Statistical Association*, **97**, 210-221.

Shen, X., Huang, H-C. and Ye, J. (2004). Adaptive model selection and assessment for exponential family distributions, *Technometrics*, **46**, 306-317.

Shibata, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion, *Biometrika*, **63**, 117-126.

Shibata, R. (1983). Asymptotic mean efficiency of a selection of regression variables, *Annals of the Institute of Statistical Mathematics*, **35**, 415-423.

Shibata, R. (1989). Statistical aspects of model selection, *From Data to Model* (ed. J. C. Willemsa), 215-240, Springer-Verlag, New York.

Shimodaira, H. (1997). Assessing the error probability of the model selection test, *Annals of the Institute of Statistical Mathematics*, **49**, 395-410.

下平英寿 (2004). 情報量規準によるモデル選択とその信頼性評価,『モデル選択—予測・検定・推定の交差点』(甘利俊一, 竹内 啓, 竹村彰通, 伊庭幸人 編), 統計科学のフロンティア 3, 第 I 部, 1-76, 岩波書店, 東京.

下平英寿 (2007). モデル選択とブートストラップ,『赤池情報量規準 AIC—モデリング・予測・知識発見—』, 第 II 編, 4 章, 133-156, 共立出版, 東京.

Shimodaira, H. and Hasegawa, M. (1999). Multiple comparisons of log-likelihoods with applications to phylogenetic inference, *Molecular Biology and Evolution*, **16**, 1114-1116.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion), *Journal of the Royal Statistical Society*, **B64**, 583-639.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2014). The deviance information

criterion: 12 years on (with discussion), *Journal of the Royal Statistical Society*, **B76**, 485-493.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion), *Journal of the Royal Statistical Society*, **B39**, 111-147.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion, *Journal of the Royal Statistical Society*, **B39**, 44-47.

Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections, *Communications in Statistics*, **A7**, 13-26.

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.

Tateishi, S. and Konishi, S. (2011). Nonlinear regression modeling and detecting change points via the relevance vector machine, *Computational Statistics*, **26**, 477-490.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society*, **B58**, 267-288.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities, *Journal of the American Statistical Association*, **81**, 82-86.

Uchida, M. and Yoshida, N. (2001). Information criteria in model selection for mixing processes, *Statistical Inference and Stochastic Processes*, **4**, 73-98.

Uchida, M. and Yoshida, N. (2004). Information criteria for small diffusions via the theory of Malliavin-Watanabe, *Statistical Inference and Stochastic Processes*, **7**, 35-67.

Umezu, Y., Shimizu, Y., Masuda, H. and Ninomiya, Y. (2019). AIC for the non-concave penalized likelihood method, *Annals of the Institute of Statistical Mathematics*, **71**(2), 247-274.

von Mises, R. (1947). On the asymptotic distribution of differentiable statistical functions, *Annals of Mathematical Statistics*, **18**, 309-348.

Wang, H., Li, R. and Tsai, C. L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method, *Biometrika*, **94**, 553-568.

Wang, H., Li, B. and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters, *Journal of the Royal Statistical Society*, **B71**, 671-683.

Watanabe, S. (2009). *Algebraic Geometry and Statistical Learning Theory*, Cambridge University Press, Cambridge, UK.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory, *Journal of Machine Learning Research*, **11**, 3571-3594.

Wood, S. N., Pya, N. and Säfken, B. (2016). Smoothing parameter and model selection for general smooth models, *Journal of the American Statistical Association*, **111**, 1548-1575.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation, *Biometrika*, **92**, 937-950.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection, *Journal of the American Statistical Association*, **93**, 120-131.

Yu, D. and Yau, K. K. (2012). Conditional Akaike information criterion for generalized linear mixed models, *Computational Statistics & Data Analysis*, **56**, 629-644.

Zhang, Y., Li, R. and Tsai, C.-L. (2010). Regularization parameter selections via generalized information criterion, *Journal of the American Statistical Association*, **105**, 312-323.

Zou, H., Hastie, T. and Tibshirani, R. (2007). On the "degrees of freedom" of the lasso, *Annals of Statistics*, **35**, 2173-2192.

The Role of Information Criterion AIC in Statistical Science

Sadanori Konishi

Department of Mathematics, Faculty of Science and Engineering, Chuo University;
Now at Faculty of Mathematics, Kyushu University

The Akaike information criterion (AIC) provides a useful tool for evaluating models estimated by the maximum likelihood method, and a number of successful applications of AIC have been reported in diverse fields of the natural and social sciences. AIC was essentially derived as an estimator of the Kullback-Leibler information from the predictive point of view, and it provided a new paradigm for model selection and evaluation problems in statistical science. The first objective of this paper is to provide a brief explanation of the concept and derivation of the AIC and related criteria.

With the development of modeling techniques such as regularization, sparse modeling, and Bayes modeling, it is necessary to present criteria that enable us to evaluate models constructed by various estimation procedures. The second objective of this paper is to review the AIC type of information criteria for evaluating models estimated by various techniques, with emphasis on the choice of the adjusted parameters, including a smoothing parameter. We review some advances in Bayesian information-theoretic criteria, where criteria were constructed, using the concept of the degrees of freedom as a bias-corrected adjustment. We also describe information-theoretic criteria for evaluating a Bayesian predictive distribution, derived from the fundamental principle behind AIC.

Key words: AIC, ABIC, BIC, DIC, GIC, PIC, TIC, WAIC.