日本生態学会誌 65：157 - 167（2015）

特集 2 生態学におけるモデル選択

モデル選択と予測：その考え方と方法

箱山　洋

水産総合研究センター

An introduction to model selection for prediction

Hiroshi Hakoyama

Fisheries Research Agency

要旨：データにモデルを当てはめるとは、観察した自然現象を確率変数で表現し、その確率分布を推定することである。頻度論的な立場からは、自然、もしくは、そのメカニズムを確率モデルとして正しく表現した「真のモデル」が、データを発生させたと考える。未知の真のモデルの確率分布をデータと近似モデルから精度よく推定できれば、結果としてよい予測につながる。本質的に、近似モデルのパラメータ数とデータに含まれる情報量が、確率分布の推定精度を決定する。また、一般に利用できるデータの量は限られている。したがって、与えられたデータに対してパラメータ数の異なる複数のモデルを用意し、最善のモデルを選択すること、すなわち、モデル選択が一つの統計学的な問題となる。ここでは、このようなモデル選択と予測に関する基本的な考え方を、ヒストグラム・モデル、線形回帰モデルを例としながら説明する。

キーワード：不一致、カルバック・ライブラー不一致、TIC、AIC

は　じ　め　に

生態学のなかでも生物資源の管理や絶滅危惧種の保全などの分野では、個体群や群集の動態を精度よく予測することが求められている。ここでは予測モデルの複雑さが重要な問題である。本特集で巌佐（2015）が解説したように、対象のモデル化において変数を束ねることは本質的なことであり、モデルに取り込む要素の程度によって、動態の予測に用いる確率モデルの複雑さは異なる。例えば、個体群動態のモデルを考えるとき、年齢構造を考慮するモデルと考慮しないモデルでは（例えば、Ludwig and Walters 1985；箱山 2015）、前者のほうが複雑なモデルである。また、群集においても、個々の種間の相互作用を考慮する複雑なモデルや、種と種間の相互作用を束ねた単純なギルド構造のモデルを考えることができる（Iwasa et al. 1987, 1989）。与えられたデータに対して、複雑さの異なる複数の確率モデルがあるとき、どの程度の複雑さのモデルが、対象の動態予測において優れているだろうか。

様々な要因があるため、この問いに答えるのは簡単ではないが、少なくともモデル選択の観点では、どのモデルの予測が優れているかは利用できるデータ量（もしくは、データの情報量）に依存すると言えるだろう。現実的で複雑なモデルが必ずしも予測に優れたモデルではない。定性的な知識も含めて対象に対する知識をできるだけ取り入れて作った複雑な確率モデルをオペレーティング・モデル（operating model）といい、経験的には、現実の限られたデータに対しては、それよりずっとパラメータ数の少ない近似モデルのほうが予測に優れている（Linhart and Zucchini 1986；Zucchini 2000）。

実際にデータと複数の確率モデルが与えられたとき、どのモデルが予測において優れているのかという問題に、モデル選択は定量的に答えることができる。モデル選択では、真の確率分布と近似モデルの確率分布のずれ（不一致，dicrepancy）を基礎として、候補となる近似モデルの中から最善の予測を与えると推定されるモデルを選択

2014 年 10 月 28 日受付、2015 年 2 月 24 日受理
e-mail: hako@affrc.go.jp

する。ここでは、この統計的な枠組みについて、簡単なモデルを例として説明する。まず、わかりやすい導入例であるヒストグラム・モデルを説明する。次に、Linhart and Zucchini（1986）の定式化に従って、不一致、特にカルバック・ライブラー不一致（Kullback-Leibler discrepancy；Kullback and Leibler 1951）について説明し、竹内（1976）によるカルバック・ライブラー不一致からの TIC（Takeuchi's information criterion）および AIC（Akaike's 'A' information criterion；Akaike 1973, 1974）の導出を説明する。最後に、線形重回帰モデルを例として、カルバック･ライブラー不一致、TIC、AIC、AICc（Hurvich and Tsai 1989）を計算することで、その違いを比較し、規準に基づいたモデル選択の性質を考察する。本稿では確率モデルは確率分布関数と同義とする。

モデル選択の考え方

ある確率変数を近似モデルで予測する場合、真の確率分布との不一致が小さい近似モデルを用いれば予測精度が高いが、その不一致の程度は、モデルの複雑さとサンプルサイズに大きく影響を受ける。このことを説明するために、Linhart and Zucchini（1986）に習い、単変量の確率分布をヒストグラムで近似することを考えてみよう。ヒストグラムは近似確率分布モデルであり、各分割の相対頻度がパラメータ、分割数がパラメータ数である。自然のメカニズムである真の確率分布を正確に知ることはできないが、ここでは説明のために、図 1 に示した混合正規分布を真の確率分布であるとして、その真の分布からモンテカルロ・データを発生させて、ヒストグラムを書くことにする。

まず、n=100 のデータを、分割数 I=10 のモデルと分割数 I=80 のモデルに当てはめて、それぞれヒストグラムを描いた（図 1a, b）。両者を比較すると、分割数 I=10 のヒストグラムは、分割が粗いながらも、真の分布をそれなりによく近似している。これに対して、分割数 I=80 のヒストグラムは、モデルの可塑性は高いがデータが不十分なために各分割の変動が激しく、真の分布をうまく近似できていない。このことを、オーバーフィッティング（overfitting）と言う。このデータでは、真の確率分布とヒストグラムのずれ（不一致）は、ガウス不一致（Gauss discrepancy, 後述）の基準で、分割数 I=80 よりも分割数 I=10 のほうが小さく、単純なモデルのほうが真の分布をよりよく近似している。

同じ n=100 のサンプルサイズでも、どちらのモデルの近似がよいかは別のデータでは変わりうる。しかしながら、十分多数の独立なデータから何度も二つのモデルのヒストグラムを描いて、不一致を計算し、平均的な近似の良さとして期待不一致（不一致の期待値）を計算すると、それはデータに依存しない関係であり、真の確率分布・近似モデル・サンプルサイズのみに依存して決定される定数である。この例では、n=100 のサンプルサイズのとき、期待不一致でも分割数 I=80 よりも分割数 I=10 のほうが小さく、平均的には単純なモデルのほうが、真の分布をよりよく近似している。

次に、n=10^5 のデータを、分割数 I=10 のモデルと分割数 I=80 のモデルに当てはめて、それぞれヒストグラムを描いた（図 1c, d）。サンプルサイズが大きいため、それぞれの近似モデルのパラメータ（各分割の相対頻度）の推定値は、そのモデルの中では最も不一致の小さいパラメータ値に近いものとなっている。両者を比較すると、分割数 I=10 のヒストグラムは、分割が粗いため可塑性が小さく、サンプルサイズが増えても、分布のずれが比較的大きい。一方、分割数 I=80 のヒストグラムは、モデルの可塑性が大きくデータが十分あるため、真の分布とのずれは極めて小さくなっている。このデータでは、分割数 I=10 よりも分割数 I=80 のほうが不一致が小さく、複雑なモデルのほうが真の分布をよりよく近似している。期待不一致も、分割数 I=10 よりも分割数 I=80 のほうが小さく、平均的にも複雑なモデルのほうが真の分布をよりよく近似している。

このように、サンプルサイズが小さいときには単純なモデルの不一致が小さく、サンプルサイズが大きいときには複雑なモデルの不一致が小さい傾向がある。サンプルサイズに応じて不一致の小さいモデルを選択して予測に用いれば、予測精度が高い。これが、予測のためのモデル選択の基本的な考え方である。

では、具体的には、どうやってデータからモデルを選択するのか。真の確率分布が未知であるから、通常はデータの実現値ごとの不一致を計算することはできないが、不一致の期待値のモデル間の差をデータから推定することはできる。すなわち、モデル選択とは、期待不一致のモデル間の差の推定量を用いて、期待不一致最小という意味で最善のモデルをデータから推定することである。比較する確率モデルのなかで期待不一致最小となる最善のモデルは、データには依存せず、真の確率分布、近似モデル、サンプルサイズによって決定される。この統計的枠組みについて、以下順を追って説明する。

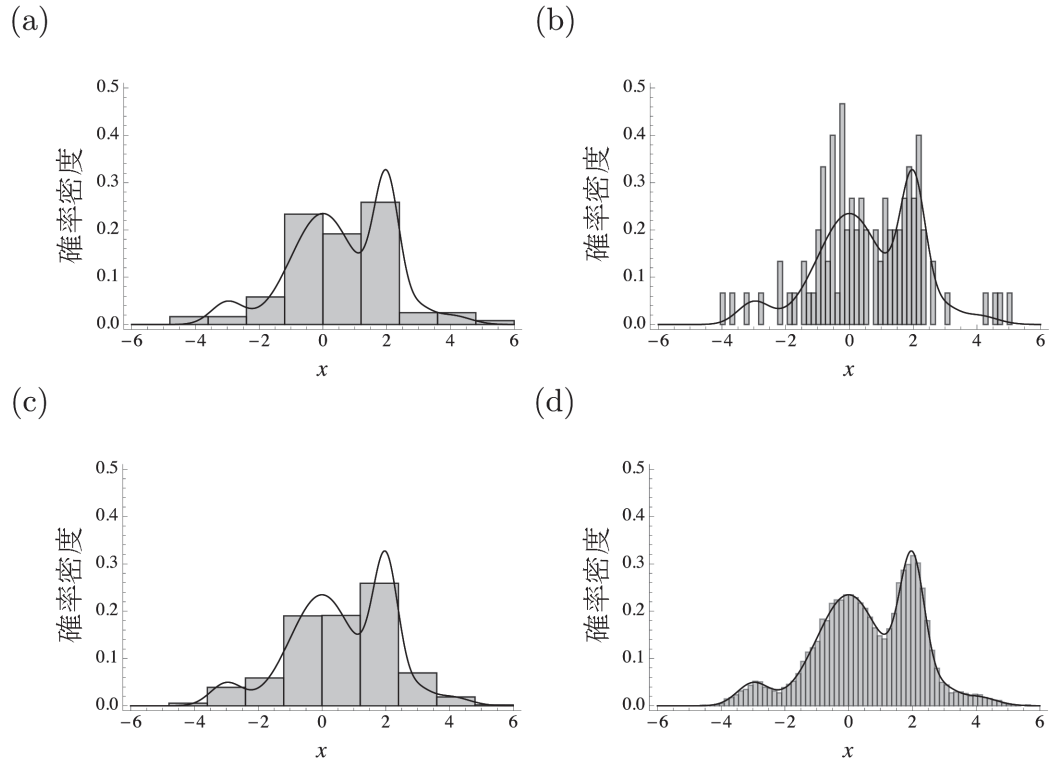

図 1. 真の確率分布（混合正規分布）と近似分布（ヒストグラム）のずれの比較。図の実線は真の確率分布を表す。真の確率分布は、正規分布 $\mathcal{N}(\mu=-3, \sigma=0.5)$, $\mathcal{N}(2, 0.4)$, $\mathcal{N}(3, 0.4)$, $\mathcal{N}(4, 0.6)$ を、重み {2, 1, 0.2, 0.1, 0.1} で混合した混合正規分布である。（a）n=100 のデータを分割数 I=10 のヒストグラム・モデルに当てはめ、真の分布と比較した図、（b）分割数 I=80 のモデルに（a）と同じデータを当てはめ、真の分布と比較した図、（c）n=10^5 のデータを分割数 I=10 のモデルに当てはめ、真の分布と比較した図、（d）分割数 I=80 のモデルに（c）と同じデータを当てはめ、真の分布と比較した図。

不一致：確率分布と確率分布のずれ

ある確率分布と別の確率分布のずれを計る量である不一致は、二つの確率分布関数の汎関数として定義されるが、目的や実用に応じた様々な不一致関数がある。連続型確率変数と離散型確率変数のどちらにも不一致は定義できるが、ここでは、連続型確率変数について、ガウス不一致とカルバック・ライブラー不一致について説明する。Linhart and Zucchini（1986）は二つの確率分布のずれを表す言葉として distance, loss function, discrepancy などを検討し、discrepancy という言葉が適切であると議論した。その訳語として、不一致という言葉を使う。

ガウス不一致

ある自然現象を表す連続な確率変数 X の真の確率分布を F と表し、その確率密度関数を $f(x)$ とする。真の確率分布 F は X を発生させる自然のメカニズムであり、通常は正確には知り得ない。これに対して、観察から得た知識を用いて作った近似確率モデルを $G_{\boldsymbol{\theta}}$ とし、その確率密度関数を $g_{\boldsymbol{\theta}}(x)$ とする。確率変数は大文字（例えば、X）、その実現値は小文字（例えば、x）で表記する。$\boldsymbol{\theta}=(\theta_1, \theta_2, \cdots, \theta_p)'$ はモデルのパラメータを表す。図 2 に概念図として、真の確率分布の確率密度関数 $f(x)$ と近似モデルの確率密度関数 $g_{\boldsymbol{\theta}}(x)$ を重ねて書いた。二つの分布が一致していない部分は灰色で示されているが、このずれを定量化するのに、$f(x)$ と $g_{\boldsymbol{\theta}}(x)$ の差の二乗を定義域で積分した量は自然であり、ガウス不一致と呼ばれている：

$$\tilde{\Delta}_{\mathrm{Gauss}}(G_{\boldsymbol{\theta}}, F) = \int (f(x) - g_{\boldsymbol{\theta}}(x))^2\, dx, \tag{1a}$$

$$= \int f(x)^2 dx - 2\int f(x) g_{\boldsymbol{\theta}}(x) dx + \int g_{\boldsymbol{\theta}}(x)^2 dx. \tag{1b}$$

ガウス不一致は常に非負であり（$\tilde{\Delta}_{\mathrm{Gauss}}(G_{\boldsymbol{\theta}}, F) \geq 0$）、等号が成り立つのは F=$G_{\boldsymbol{\theta}}$ のときだけであることに注意しよ

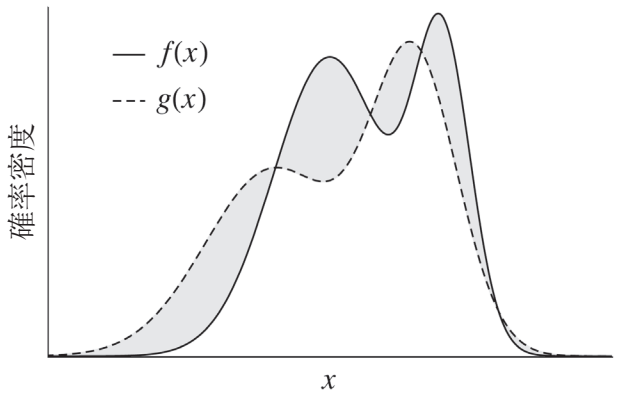

図 2. 真の確率分布 $f(x)$ と近似確率分布 $g(x)$ のずれを表す概念図。灰色の部分は二つの分布の不一致部分を表す。

う。不一致が小さいほど分布のずれが小さく、近似モデルが真のモデルに一致するときに不一致が最小となるようになっている。一般に不一致はこの非負である性質を満たしていなければならない。

ガウス不一致の積分式（式 1a）を展開すると式 1b となるが、式 1b の右辺第一項は真のモデルにのみ依存した項である。従って、二つの近似モデルの不一致の差を考える場合には必要なく、この項を除いた式もまたガウス不一致として用いられる：

$$\Delta_{\mathrm{Gauss}}(G_{\boldsymbol{\theta}}, F) = -2\int f(x) g_{\boldsymbol{\theta}}(x) dx + \int g_{\boldsymbol{\theta}}(x)^2 dx. \qquad (2)$$

式 1 と区別するために式 2 にはチルダをつけない。実際のモデル選択の基礎となるのは式 2 の不一致である。Linhart and Zucchini（1986）は、このガウス不一致を基礎として、前節のヒストグラム・モデルの期待不一致の不偏推定量を導出し、モデル選択を行った。

不一致は近似モデル $G_{\boldsymbol{\theta}}$ の関数型にも依存するが、単一の近似モデル $G_{\boldsymbol{\theta}}$ を扱っていて誤解のない場合、不一致がパラメータ $\boldsymbol{\theta}$ の関数であることを強調して、不一致 $\Delta(G_{\boldsymbol{\theta}}, F)$ を $\Delta(\boldsymbol{\theta})$ と略記する。

カルバック・ライブラー不一致

シャノンの情報量（Shannon and Weaver 1949）と関係のあるカルバック・ライブラー不一致（Kullback-Leibler 情報量）も二つの確率分布のずれを計る汎関数の一つであり、真の分布のもとでの対数確率分布比の期待値である。カルバック・ライブラー不一致は、二つの確率分布のずれを表す関数としてガウス不一致ほど直感的ではないが、モデル選択の汎用的な規準である AIC の基礎となる重要な不一致である。カルバック・ライブラー不一致は次のように定義される：

$$\tilde{\Delta}_{\mathrm{K\text{-}L}}(G_{\boldsymbol{\theta}}, F) = \int f(x) \ln \frac{f(x)}{g_{\boldsymbol{\theta}}(x)} dx, \qquad (3a)$$

$$= \int f(x) \ln f(x) dx - \int f(x) \ln g_{\boldsymbol{\theta}}(x) dx. \qquad (3b)$$

カルバック・ライブラー不一致も常に非負の値を取り、最小の 0 となるのは、近似モデルが真のモデルに一致するときだけである。$\ln x \le x - 1$ という不等式から、常に $\tilde{\Delta}_{\mathrm{K\text{-}L}}(\boldsymbol{\theta}) \ge 0$ であることを証明できる。

少し具体例を計算してみて、イメージをつかんでみよう。$f(x)$、$g_{\boldsymbol{\theta}}(x)$ が、それぞれ、$\mathcal{N}(0, 1)$、$\mathcal{N}(\mu, \sigma^2)$ に従う正規分布であるとする。このとき、$\tilde{\Delta}_{\mathrm{K\text{-}L}}(\mu, \sigma^2) = 0.5\,(\ln \sigma^2 - 1 + (1 + \mu^2)/\sigma^2)$ である。この式から、平均差 μ の二乗で不一致が大きくなることがわかる。また、近似モデルの分散 σ^2 が、真の値の 1 より大きくても、小さくても不一致が大きくなることもわかる。近似モデルの平均が真のモデルと異なる場合を、図 3a, b に示した。図 3a は、真のモデルと近似モデルの確率密度関数を示しており、平均差が大きい近似モデル（2）のほうが、平均差が小さい近似モデル（1）よりも、真のモデルとのずれが大きいことが見て取れる。これに対して、それぞれのモデルのカルバック・ライブラー不一致の被積分関数 $f(x)\ln(f(x)/g_{\boldsymbol{\theta}}(x))$ を描いたものが、図 3b である。カルバック・ライブラー不一致は、灰色の領域で表した符号付きの面積となるが、近似モデル（2）のほうが、近似モデル（1）より符号付きの面積が大きくなっており、不一致が大きいことが見て取れる。常に $\tilde{\Delta}_{\mathrm{K\text{-}L}}(\boldsymbol{\theta}) \ge 0$ であるから、正の面積は負の面積より大きいことに注意しよう。$f(x)$ と $g_{\boldsymbol{\theta}}(x)$ が正規分布の場合には $\ln(f(x)/g_{\boldsymbol{\theta}}(x))$ は直線関数であり、その値は分布の中心から外れると極端に大きく（小さく）なるが、分布の中心から外れたところでは $f(x)$ が十分に小さいので、$\tilde{\Delta}_{\mathrm{K\text{-}L}}(\mu, \sigma^2)$ の主要な被積分領域は分布の中心部だけにある。

カルバック・ライブラー不一致においても、式 3b の右辺第一項は真の分布 $f(x)$ のみの関数であるため、近似モデルの比較には必要がない。したがって、この項を除いた相対的な不一致をモデル選択の基礎として用いる：

$$\Delta_{\mathrm{K\text{-}L}}(G_{\boldsymbol{\theta}}, F) = -\int f(x) \ln g_{\boldsymbol{\theta}}(x) dx = -E_{F,X}\left[\ln g_{\boldsymbol{\theta}}(X)\right], \qquad (4)$$

ただし、$E_{F,X}[]$ は、真の分布 F の確率密度関数 $f(x)$ で確率

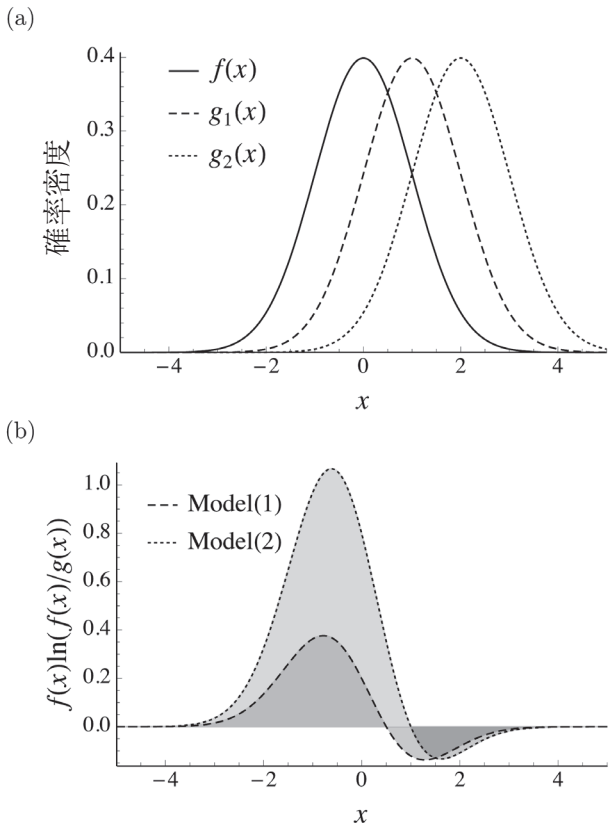

図 3. 真の分布と近似分布が正規分布の場合のカルバック・ライブラー不一致の例。(a) 真の確率分布$\mathcal{N}(0, 1)$と二つの近似確率分布$\mathcal{N}(1, 1)$、$\mathcal{N}(2, 1)$の確率密度関数。それぞれ、$f(x)$, $g_1(x)$, $g_2(x)$で表す。(b) カルバック・ライブラー不一致関数の被積分関数$f(x)\ln(f(x)/g(x))$の分布。灰色の符号付き面積を積分すると、カルバック・ライブラー不一致が得られる。

変数Xの関数の期待値をとることを表す。確率変数が自明な場合には、$E_F[]$と略記する。

確率変数としての不一致

近似確率モデル$G_{\boldsymbol{\theta}}$のパラメータをデータ$\boldsymbol{x}=(x_1, x_2, \cdots, x_n)'$から推定するのであれば、その汎関数である不一致$\Delta(\hat{\boldsymbol{\theta}})=\Delta(G_{\hat{\boldsymbol{\theta}}}, F)$は確率変数である。ただし、パラメータ$\boldsymbol{\theta}$の推定量はデータ$\boldsymbol{x}$の関数であり、ハットを付けて$\hat{\boldsymbol{\theta}}(\boldsymbol{x})$と表す。

この節では、第二節でヒストグラムを例に説明したモデルの複雑さ・サンプルサイズと不一致$\Delta(\hat{\boldsymbol{\theta}})$の大きさの関係を定式化して説明する。この説明のために、確率変数である不一致$\Delta(\hat{\boldsymbol{\theta}})$を二つの要素（近似による不一致、推定による不一致）に分解する。近似による不一致はモデルの複雑さに関係した定数であり、推定による不一致はサンプルサイズに影響を受ける確率変数である。

近似による不一致

ある近似モデル$G_{\boldsymbol{\theta}}$のなかで、もっとも不一致が小さいパラメータ値を$\boldsymbol{\theta}_0$とする：

$$\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}} \Delta(G_{\boldsymbol{\theta}}, F).$$

$G_{\boldsymbol{\theta}_0}$は近似モデル$G_{\boldsymbol{\theta}}$の中で最善の近似モデルであり、$G_{\boldsymbol{\theta}_0}$と真の分布$F$との不一致$\Delta(\boldsymbol{\theta}_0)=\Delta(G_{\boldsymbol{\theta}_0}, F)$を「近似による不一致」という。近似による不一致$\Delta(\boldsymbol{\theta}_0)$は、$g_{\boldsymbol{\theta}}(x)$と$f(x)$の関数型にのみ依存し、データに依存しない定数である。近似による不一致は単純なモデルで大きい傾向がある。

真の分布Fが未知であれば、$\boldsymbol{\theta}_0$は計算できない。しかしながら、$\hat{\boldsymbol{\theta}}$は$\boldsymbol{\theta}_0$の推定量であり、サンプルサイズが十分大きければ、精度良く$\boldsymbol{\theta}_0$を推定できる。例えば、先のヒストグラム・モデルの場合、十分大きなデータから推定したパラメータ値（各分割の相対頻度）は$\boldsymbol{\theta}_0$に近いものとなっている（図 1c と 1d のヒストグラム）。

推定による不一致

データから推定したパラメータ値$\hat{\boldsymbol{\theta}}$を持つ近似モデル$G_{\hat{\boldsymbol{\theta}}}$と、最善の近似モデル$G_{\boldsymbol{\theta}_0}$の間の不一致$\Delta(G_{\hat{\boldsymbol{\theta}}}, G_{\boldsymbol{\theta}_0})$を「推定による不一致」という。推定による不一致は確率変数である。例えば、先のヒストグラム・モデルの例では、分割数I=80 のモデルについて、図 1d のヒストグラムのパラメータは十分$\boldsymbol{\theta}_0$に近いから、図 1b と 1d のヒストグラムの間の不一致が、おおよそ、図 1 のn=100 のデータの推定による不一致となっている。

サンプルサイズが小さいときには、複雑なモデルの推定による不一致が大きい傾向がある。例えば、図 1 のn=100 のデータでは、分割数I=10 よりも分割数I=80 のモデルの推定による不一致のほうが大きい（おおよそ、I=10 のモデルの推定による不一致は図 1a と図 1c のヒストグラムの間の不一致であり、I=80 のそれは図 1b と 1d のヒストグラムの間の不一致である）。

全体の不一致

確率変数である全体の不一致$\Delta(G_{\hat{\boldsymbol{\theta}}}, F)$は、近似による不一致$\Delta(G_{\boldsymbol{\theta}_0}, F)$と推定による不一致$\Delta(G_{\hat{\boldsymbol{\theta}}}, G_{\boldsymbol{\theta}_0})$によって決定される。近似による不一致$\Delta(G_{\boldsymbol{\theta}_0}, F)$は、サンプルサイズによらず、単純なモデルで大きい傾向がある。一方、サンプルサイズが小さいときには、複雑なモデルの推定による不一致$\Delta(G_{\hat{\boldsymbol{\theta}}}, G_{\boldsymbol{\theta}_0})$は、近似による不一致$\Delta(G_{\boldsymbol{\theta}_0}, F)$

を上回る大きさとなる傾向がある。これらの効果が組合わさることで、サンプルサイズが小さいときには、複雑なモデルよりも単純なモデルの全体の不一致 $\Delta(G_{\hat{\theta}}, F)$ が小さい傾向（予測精度が高い傾向）が出てくる。

真の分布 F と最善の近似モデル G_{θ_0} が一致する場合（$f(x) \equiv g_{\theta_0}(x)$）、全体の不一致 $\Delta(G_{\hat{\theta}}, F)$ は、近似による不一致 $\Delta(G_{\theta_0}, F)$ と推定による不一致 $\Delta(G_{\hat{\theta}}, G_{\theta_0})$ の和である：

$$\Delta(G_{\hat{\theta}}, F) = \Delta(G_{\theta_0}, F) + \Delta(G_{\hat{\theta}}, G_{\theta_0}). \qquad (5)$$

この場合、近似による不一致と推定による不一致の効果は相加的にはたらく。一般には真の分布と最善の近似モデルは一致しないが、最善の近似モデルが十分真の分布に近ければ（$f(x) \approx g_{\theta_0}(x)$）、近似による不一致と推定による不一致が相加的に全体の不一致を決定すると考えてよい。

モデル選択の規準：期待不一致の推定量

ある確率変数 X の予測をするモデルを選ぶ場合、モデル選択の規準は期待不一致（不一致 $\Delta(\hat{\boldsymbol{\theta}})$ の期待値）の推定量である。ただし、ここでの不一致は、式 2 や式 4 のように真の分布のみに依存する計算できない項を除いた相対的な不一致である。モデル選択では、データから各モデルの期待不一致を推定し（モデル選択規準を計算し）、期待不一致が小さいと推定されるモデルを予測に優れたモデルとして選択する。

不一致の分布の期待値は近似モデルとサンプルサイズで決定される。即ち、あるサンプルサイズに対して、複数の近似モデルのなかに期待不一致最小となる近似モデルが存在する。平均的に小さな不一致を示すモデルは、問題とする確率変数 X の予測精度が高いと考えられるから、一つの考え方として、期待不一致の推定量をモデル選択の規準としてよいだろう。推定した期待不一致を比較することによって、真の分布が未知のまま予測精度の高いと推定される近似モデルを選ぶことができる。

例えば、カルバック・ライブラー不一致の場合であれば、$\Delta_{\text{K-L}}(\hat{\boldsymbol{\theta}}(\boldsymbol{X})) = -E_{F,Z}[\ln g_{\hat{\theta}(X)}(Z)]$ の期待値 $E_{F,X}[\Delta_{\text{K-L}}(\hat{\boldsymbol{\theta}}(\boldsymbol{X}))]$ の推定量を導出することができる。ここで、Z とデータ $\boldsymbol{X}=(X_1, X_2, \cdots, X_n)'$ の各 X_i は互いに独立で同一の確率密度関数 f（真の分布）に従う確率変数である。モデル選択規準の TIC や AIC はカルバック・ライブラー不一致の期待値（AIC は、その $2n$ 倍）の推定量に他ならない。

漸近近似規準

この節では、竹内（1976）に沿ってカルバック・ライブラー不一致と最尤推定量に基づいた近似規準である TIC と AIC の導出を説明する。

一般に多くの場合で、期待不一致 $E_{F,X}[\Delta(\hat{\boldsymbol{\theta}}(\boldsymbol{X}))]$ や、その良い推定量は複雑すぎて導くことができない。また、個々のモデルの規準をその都度導出するのは汎用性に欠ける。そこで、漸近理論に基づいた近似的な規準が考えられた（Akaike 1973, 1974；竹内 1976; Sugiura 1978；Hurvich and Tsai 1989）。サンプルサイズが十分に大きく、真の分布 F と最善の近似分布 G_{θ_0} が十分に近いという仮定のもとで、規準は比較的簡単な式となる。AIC は、そのような漸近近似規準のひとつである。

まず、次項で期待不一致のバイアスした推定量である最小経験不一致（minimum empirical discrepancy）を説明する。つぎに、次々項で説明する最尤推定量 $\hat{\boldsymbol{\theta}}$ の漸近分布を用いて、最後の項で最小経験不一致 $\Delta_n(\hat{\boldsymbol{\theta}})$ の期待不一致の推定量としてのバイアスを評価し、モデル選択の規準となるバイアスの小さい期待不一致の推定量である AIC と TIC を導出する。

以下、断りがない限り、不一致はカルバック・ライブラー不一致を意味し、$\Delta(\hat{\boldsymbol{\theta}})$ をカルバック・ライブラー不一致 $\Delta_{\text{K-L}}(\hat{\boldsymbol{\theta}})$ の略記とする。また、期待不一致 $E_{F,X}[\Delta(\hat{\boldsymbol{\theta}}(\boldsymbol{X}))]$ は $E_F[\Delta(\hat{\boldsymbol{\theta}})]$ と略記する。

最小経験不一致：期待不一致のバイアスした推定量

通常、最尤推定では、データを発生させた真のモデル F とデータに当てはめるモデルのなかの最善のモデル G_{θ_0} が等しいと仮定するが、ここではこれまでの議論の通り、一般には $F \neq G_{\theta_0}$ とする。Akaike（1973）は、最尤推定原理の拡張として、最尤推定は経験不一致（empirical discrepancy）$\Delta_n(\boldsymbol{\theta})$ の最小化として明快に理解できると考えた。経験不一致は、近似モデル $G_{\boldsymbol{\theta}}$ と経験分布（empirical distribution; 試行結果から推測した確率分布）F_n との不一致であり、次のように定義される：

$$\Delta_n(\boldsymbol{\theta}) = \Delta(G_{\boldsymbol{\theta}}, F_n) = -\frac{1}{n}\sum_{i=1}^{n} \ln g_{\boldsymbol{\theta}}(x_i), \qquad (6a)$$

$$= -\frac{1}{n} L(\boldsymbol{\theta}; \boldsymbol{x}), \qquad (6b)$$

ただし、$L(\boldsymbol{\theta}; \boldsymbol{x}) = \sum_{i=1}^{n} \ln g_{\boldsymbol{\theta}}(x_i)$ は対数尤度である。最尤推定量 $\hat{\boldsymbol{\theta}}$ は、この経験不一致の最小化（尤度の最大化）で得ることができる。最尤推定量 $\hat{\boldsymbol{\theta}}$ を経験不一致に代入した

最小経験不一致 $\Delta_n(\hat{\boldsymbol{\theta}})$ は、期待不一致 $E_F[\Delta(\hat{\boldsymbol{\theta}})]$ の一致推定量（サンプルサイズが十分に大きいとき不偏推定量となる推定量）であるが、サンプルサイズが小さいとき、最尤推定量 $\hat{\boldsymbol{\theta}}$ を代入したことによるバイアスがある。このため、最小経験不一致 $\Delta_n(\hat{\boldsymbol{\theta}})$（もしくは、最大対数尤度 $L(\hat{\boldsymbol{\theta}}; \boldsymbol{x})$）は、モデル選択の規準には適さない。

最尤推定量の漸近分布

近似確率モデル $G_{\boldsymbol{\theta}}$のパラメータ $\boldsymbol{\theta}=(\theta_1, \theta_2, \cdots, \theta_p)'$ に関する尤度方程式 $\partial L(\boldsymbol{\theta}; \boldsymbol{x})/\partial\boldsymbol{\theta}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}=0$ を $\boldsymbol{\theta}_0$ の周りでテイラー展開し、2 次の項までとる計算から、サンプルサイズ n が大きいとき、$\hat{\boldsymbol{\theta}}$ の漸近分布が多変量正規分布に従うことを計算できる：

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \sim \mathcal{N}(0, J(\boldsymbol{\theta}_0)^{-1} I(\boldsymbol{\theta}_0) J(\boldsymbol{\theta}_0)^{-1}), \tag{7}$$

ただし、$\boldsymbol{\theta}_0$ は不一致 $\Delta(\boldsymbol{\theta})$ を最小にする定数である。$J(\boldsymbol{\theta})$ と $I(\boldsymbol{\theta})$ は、次のような対称行列で定義される：

$$J(\boldsymbol{\theta}) = -E_F\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} \ln g_{\boldsymbol{\theta}}(\boldsymbol{x})\right],$$

$$I(\boldsymbol{\theta}) = E_F\left[\frac{\partial}{\partial\boldsymbol{\theta}} \ln g_{\boldsymbol{\theta}}(\boldsymbol{x}) \frac{\partial}{\partial\boldsymbol{\theta}'} \ln g_{\boldsymbol{\theta}}(\boldsymbol{x})\right].$$

ここで、$g_{\boldsymbol{\theta}_0}(x) \equiv f(x)$ であれば、$J(\boldsymbol{\theta}_0)=I(\boldsymbol{\theta}_0)$ となり、フィッシャー情報行列に一致する。フィッシャー情報量は、シャノンの情報量とは違った概念である。フィッシャー情報行列の逆行列は漸近的には最尤推定量の分散共分散行列であり、大きな情報を持つとは推定量の分散が小さいことである。

カルバック・ライブラー不一致の展開

カルバック・ライブラー不一致 $\Delta(\hat{\boldsymbol{\theta}})$ を $\boldsymbol{\theta}_0$ のまわりで展開し、2 次の項までとる：

$$\Delta(\hat{\boldsymbol{\theta}}) \approx \Delta(\boldsymbol{\theta}_0) + \frac{1}{2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' J(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0). \tag{8}$$

次に、多少天下り的であるが、$M=\Delta(\boldsymbol{\theta}_0)-\Delta_n(\boldsymbol{\theta}_0)$ がでるように、$\Delta(\boldsymbol{\theta}_0)$ を展開する。ここでは、$\Delta_n(\hat{\boldsymbol{\theta}})$ を $\boldsymbol{\theta}_0$ のまわりで展開して 2 次の項までとった式を使う。$n \to \infty$ で、観測情報行列の平均 $n^{-1}\sum_{i=1}^{n} \partial^2 \ln g_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x})/(\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}')$ が、$J(\boldsymbol{\theta}_0)$ に確率収束することから、

$$\Delta(\boldsymbol{\theta}_0) \approx \Delta_n(\hat{\boldsymbol{\theta}}) + \frac{1}{2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' J(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + M. \tag{9}$$

式 8, 9 から、

$$\Delta(\hat{\boldsymbol{\theta}}) \approx \Delta_n(\hat{\boldsymbol{\theta}}) + (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' J(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + M. \tag{10}$$

式 10 の右辺第二項と第三項が、不一致 $\Delta(\hat{\boldsymbol{\theta}})$ に対する最小経験不一致 $\Delta_n(\hat{\boldsymbol{\theta}})$ のバイアスである。データの実現値ごとに、このバイアスを計算することはできないが、その期待値はデータから推定できる。

先に式 7 で求めた推定量 $\hat{\boldsymbol{\theta}}$ の漸近分散共分散行列より、対称行列 J の二次形式とトレースの性質から、

$$E_F\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' J(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\right] \approx \frac{1}{n} \operatorname{trace}\left[J(\boldsymbol{\theta}_0)^{-1} I(\boldsymbol{\theta}_0)\right],$$

また、$E_F[M]=0$ である。式 10 の両辺の期待値を取り、これらを代入すると、カルバック・ライブラー不一致の期待値の展開式が得られる：

$$E_F\left[\Delta(\hat{\boldsymbol{\theta}})\right] \approx E_F\left[\Delta_n(\hat{\boldsymbol{\theta}})\right] + \frac{1}{n} \operatorname{trace}\left[J(\boldsymbol{\theta}_0)^{-1} I(\boldsymbol{\theta}_0)\right]. \tag{11}$$

$E_F[\Delta(\hat{\boldsymbol{\theta}})]$ の推定量（規準）のひとつは、式 11 の右辺第一項の期待値を外し、$\boldsymbol{\theta}_0$ を $\hat{\boldsymbol{\theta}}$ で置き換えたものになる：

$$\mathrm{TIC} = \Delta_n(\hat{\boldsymbol{\theta}}) + \frac{1}{n} \operatorname{trace}\left[J(\hat{\boldsymbol{\theta}})^{-1} I(\hat{\boldsymbol{\theta}})\right], \tag{12}$$

この規準が TIC である。最小経験不一致に補正項であるトレース項が加わっている。

式 11 の特別な場合として、前節で述べたように、最善のモデル $g_{\boldsymbol{\theta}_0}(x)$ が真のモデル $f(x)$ に十分近い場合は $J(\boldsymbol{\theta}_0) \approx I(\boldsymbol{\theta}_0)$ であり、トレース項はパラメータ数 p で近似できる:

$$\operatorname{trace}\left[J(\boldsymbol{\theta}_0)^{-1} I(\boldsymbol{\theta}_0)\right] \approx p,$$

このとき、式 12 の右辺第二項のトレース項を p で置き換え、両辺を $2n$ 倍すると、AIC に一致する：

$$\mathrm{AIC} = -2L(\hat{\boldsymbol{\theta}}; \boldsymbol{x}) + 2p. \tag{13}$$

AIC は、最尤推定から最大対数尤度さえ計算できれば、正確な期待不一致やその良い推定量を導けないモデルでも広く用いることができる。近似モデルが柔軟で真のモデルを十分に扱える場合（パラメータ数が多い場合）、近似による不一致が小さいので、$\operatorname{trace}[J(\boldsymbol{\theta}_0)^{-1} I(\boldsymbol{\theta}_0)]=p$ とする近似の精度が高くなる。

TIC も AIC も漸近近似に基づいて導出されたため、サ

ンプルサイズが小さいときには、バイアスが大きくなる。そこで、Sugiura（1978）は、正規分布や二項分布に従うモデルについて、有限サンプルサイズでの補正をしたc-AIC（the corrected AIC）を導き、Hurvich and Tsai（1989）は、Sugiura（1978）の成果を踏まえて、一般回帰モデルに対してAICcを導いた。AICcは次の式で与えられる：

$$\mathrm{AICc} = \mathrm{AIC} + \frac{2p(p+1)}{n-p-1}. \tag{14}$$

線形回帰モデルの例

モデル

ここまでの議論の具体例として、線形回帰モデルを調べる。具体例によって、問題を掘り下げて行く。真のモデルを既知として、モンテカルロ・データを用いて、近似モデルでのモデル選択を行ってみよう。

まず、真のモデル F として、説明変数が二つの重回帰モデルを考える：

$$Y = 0.1X_1 + 0.5X_2 + 1 + \epsilon, \tag{15a}$$

$$X_1 \sim \mathcal{N}(1,1),\ X_2 \sim \mathcal{N}(1,1),\ \epsilon \sim \mathcal{N}(0,1),\ X_1 \perp\!\!\!\perp X_2 \perp\!\!\!\perp \epsilon. \tag{15b}$$

ただし、$\perp\!\!\!\perp$は確率変数が独立であることを表す。このモデルでは、係数の大きさの違いから説明変数では X_2 のほうが X_1 よりも効果が大きいが、どちらも切片に比べれば効果は小さい。ここで、X_2 は観測するのが難しい未知の要因だとしよう。この場合、真のモデルはモデル選択の近似モデルの候補から外れる。また、Y と X_1 のデータだけを得ることになる。

次に、二つの近似確率モデルを考える。第一の近似モデル G_1 は X_1 を説明変数とした単回帰である：

$$Y = a_1 X_1 + a_2 + \epsilon, \tag{16a}$$

$$\epsilon \sim \mathcal{N}(0, \sigma_1^2). \tag{16b}$$

第二の近似モデル G_2 は、X_1 を含まない切片のみのモデルである：

$$Y = a_3 + \epsilon, \tag{17a}$$

$$\epsilon \sim \mathcal{N}(0, \sigma_2^2). \tag{17b}$$

モデル選択は、極めて複雑な自然＝真のモデルとして、それを単純な確率モデルで近似するというスキームであるから、この例はその小さな模型であると言えるだろう。

カルバック・ライブラー不一致

不一致には、カルバック・ライブラー不一致を用いる。F, G_1, G_2 は正規分布に従い、その x_1, x_2 の条件付き確率密度関数は、それぞれ次のようになる：

$$f(y \mid x_1, x_2) \sim \mathcal{N}(0.1x_1 + 0.5x_2 + 1,\ 1), \tag{18a}$$

$$g_1(y \mid x_1, x_2) \sim \mathcal{N}(a_1 x_1 + a_2,\ \sigma_1^2), \tag{18b}$$

$$g_2(y \mid x_1, x_2) \sim \mathcal{N}(a_2,\ \sigma_2^2). \tag{18c}$$

真のモデル F と近似モデル G が正規分布で、その確率分布がそれぞれ$\mathcal{N}_F(\mu, \sigma^2)$、$\mathcal{N}_G(m, s^2)$ であるとき、カルバック・ライブラー不一致は次のような簡単な式になる：

$$\Delta_{\text{K-L}}\left(m,\ s^2\right) = \frac{1}{2}\ln 2\pi + \frac{1}{2}\ln s^2 + \frac{(\mu - m)^2 + \sigma^2}{2s^2}. \tag{19}$$

式 18, 19 を用いれば、条件付き不一致 $\Delta_{\text{K-L}}(G_1,\ F \mid x_1,\ x_2)$、$\Delta_{\text{K-L}}(G_2, F \mid x_1, x_2)$ を計算することができる。

規準

不一致 $\Delta_{\text{K-L}}(\hat{\boldsymbol{\theta}} \mid x_1, x_2)$ が x_1, x_2 の条件付きであるから、その期待値 $E_F[\Delta_{\text{K-L}}(\hat{\boldsymbol{\theta}} \mid x_1, x_2)]$ も x_1, x_2 の条件付きとなる。この条件付き期待不一致を X_1, X_2 に関して期待値をとったものを期待不一致としてモデルを比較する：

$$E_X\left[E_F\left[\Delta_{\text{K-L}}\left(\hat{\boldsymbol{\theta}} \mid X_1, X_2\right)\right]\right]. \tag{20}$$

モンテカルロ・データから様々なサンプルサイズ n の不一致 $\Delta_{\text{K-L}}(\hat{\boldsymbol{\theta}} \mid X_1,\ X_2)$ の分布や式 20 の期待値を計算した。各サンプルサイズ n ごとに、モンテカルロ・データは 10^5 回発生させた。また、TIC, AIC, AICc の規準についてもモンテカルロ・データごとに計算し、それらの確率分布を得た。これらの規準は式 20 の推定量となっている。

AIC と AICc は簡単に計算できるが、TIC ではトレースの項を計算しなければいけない。計算してみると、近似

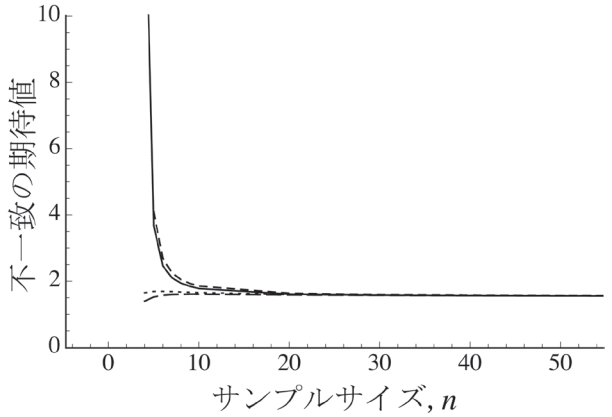

図 4. 線形回帰モデルにおけるサンプルサイズごとの期待不一致と規準の推定バイアス。実線はカルバック・ライブラー不一致 $\Delta_{\text{K-L}}(G_1, F)$ の期待値であり、真の値である。一番下の長い破線は TIC、点線は AIC、破線は AICc の期待値を表す。それぞれ、10^5 回のモンテカルロ・データから平均値を計算して描いた。$\Delta_{\text{K-L}}$ や TIC と比較するために、AIC と AICc は $2n$ で割った。

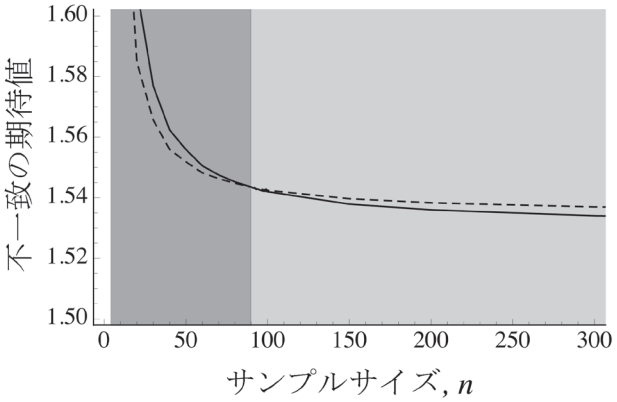

図 5. 線形回帰モデルにおけるサンプルサイズごとの期待不一致。実線は不一致 $\Delta_{\text{K-L}}(G_1, F)$ の期待値であり、破線は $\Delta_{\text{K-L}}(G_2, F)$ の期待値である。n=90 を境に両者の大小関係は入れ替わっている。期待不一致の小さいモデルが最善であるから、濃い灰色の領域では G_2 が最善のモデルである。薄い灰色の領域では G_1 が最善のモデルである。

モデル G_1 と G_2 において、$I(\hat{\boldsymbol{\theta}})$ は比較的簡単な式となるが、$J(\hat{\boldsymbol{\theta}})$ は煩雑となる。$I(\hat{\boldsymbol{\theta}})$ よりも観測情報行列 $\mathcal{I}(\hat{\boldsymbol{\theta}}) = \partial_2 \ln g_{\hat{\theta}}(\boldsymbol{x})/(\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}')$ のほうが、フィッシャー情報行列 $I(\boldsymbol{\theta}_0)$ の推定量としてよいことが知られているから（Efron and Hinkley 1978）、ここでは $I(\boldsymbol{\theta}_0)$ と $J(\boldsymbol{\theta}_0)$ の推定に、簡単に計算できる観測情報行列を用い、数値的にトレース項を計算する。ただし、G_1 と G_2 ともに $I(\hat{\boldsymbol{\theta}}) = \mathcal{I}(\hat{\boldsymbol{\theta}})$ であり、この部分には違いがない。

規準の推定バイアス、分散

図 4 は TIC, AIC, AICc の推定バイアスを示している。実線のカルバック・ライブラー不一致の期待値からのずれが推定バイアスである。サンプルサイズが小さいとき（$n < 20$）、漸近近似によって導かれた TIC と AIC のバイアスが大きい。一方、AICc はサンプルサイズが小さいときにもバイアスが少なく、有限サンプルサイズでの補正がうまくいっていることがわかる。サンプルサイズが大きいとき（$n > 20$）、どの規準もほとんどバイアスなく期待不一致を推定できている。

一般に、不偏推定量の分散は小さいほどよい。TIC, AIC, AICc の分散を考えてみると、AIC と AICc は定義上同一の分散を持つ。TIC は、トレースの項も確率変数である分、AIC や AICc よりも分散が大きい。しかし、この線形回帰モデルの例では、分散の大きさの差は無視できるほど小さかった。

最善のモデル

これまで議論してきたように、最善のモデル（期待不一致が最小のモデル）はデータに依存せず、真のモデル・近似モデル・サンプルサイズによって決定される。当然、規準にも依存しない。図 5 に、単回帰モデル G_1 および切片のみのモデル G_2 の期待不一致と、サンプルサイズの関係を示す。サンプルサイズ n=90 で二つのモデルの期待不一致の大小関係は入れ替わっている。すなわち、$n \leq 90$ では G_2 が最善のモデルであり、$n > 90$ では G_1 が最善のモデルである。

真のモデル F の構造に、単回帰モデル G_1 と切片のみのモデル G_2 のどちらが近いかといえば、それは単回帰モデル G_1 である。しかしながら、$n \leq 90$ では G_2 が最善のモデルであった。このことは、真のモデルの構造に近いモデルが、最善のモデルとなるとは限らないことを示している。今回の例では真のモデルは選択候補にないが、仮に真のモデルが選択候補にあったとしても、サンプルサイズが小さいときには最善のモデルにならない。どの程度サンプルがあれば、サンプルサイズが大きいと言えるのかには絶対的な指標はなく、対象のモデルによって違う。

また、この結果は、サンプルサイズが小さいとき、最善の予測（期待不一致が最小と言う意味で最善の予測）には手持ちのデータ X_1 は使わない方が良いことを意味している。利用できるデータをすべて利用することは、必ずしも最善の予測にならない。

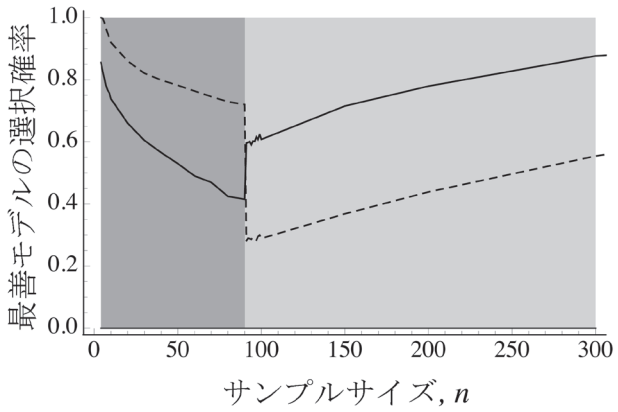

図 6. 線形回帰モデルにおけるサンプルサイズごとの最善のモデルの選択確率。実線はカルバック・ライブラー不一致で選択した場合。破線は AICc で選択した場合。二つの灰色の領域は、図 5 と同様に、最善のモデルの領域を示す。

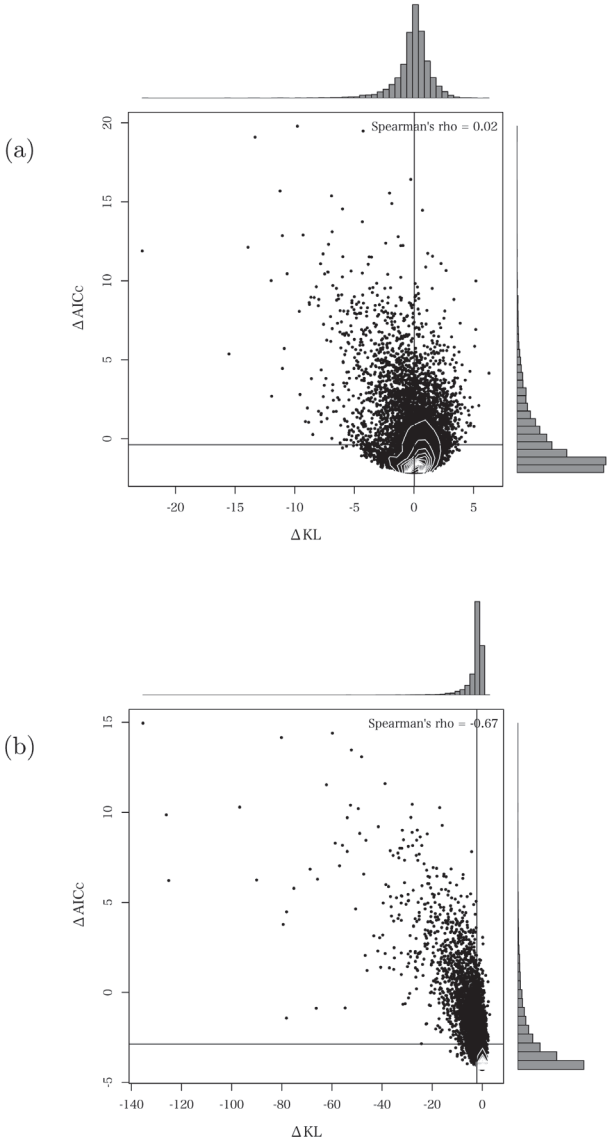

図 7. カルバック・ライブラー不一致の差 ΔKL の 2n 倍と ΔAICc の散布図（$n_{simulation}=10^5$）。それぞれの平均値を直線で表す。直線の交点は重心である。ヒストグラムはそれぞれの周辺分布を表す。(a) サンプルサイズ n=90、(b) サンプルサイズ n=10。

最善のモデルの選択確率のバイアス

AICc でモデル選択したとき、最善のモデルはどの程度の確率で選ばれるだろうか？図 5 のサンプルサイズに対する不一致の期待値の変化からすると、選択確率は、n=4 の最小サンプルサイズで高く、サンプルサイズの増加とともに次第に減少し、n=90 で 0.5 となり、そこから滑らかに増加に転じていくのが、モデル選択規準の持つ性質としては理想であろう。真の分布のずれであるカルバック・ライブラー不一致で選択する場合と AICc で選択する場合を比較しながら、サンプルサイズに対する選択確率の変化を図 6 に示した。実際には最善のモデルの選択確率は面白いパターンを示している。

実線で示したカルバック・ライブラー不一致での選択確率は、サンプルサイズ $4 \le n \le 90$ で、0.8 から 0.4 程度まで減少し、n=90 を境に急激に上昇し 0.6 程度になり、そこから n=300 までにゆっくりと 0.8 程度まで上昇している。

破線で示した AICc での選択確率も同様のパターンを示しているが、カルバック・ライブラー不一致での場合と比べるとバイアスしており、より極端である。AICc では、$n \le 90$ で 0.7 から 1 という非常に高い選択確率を持つが、$90 < n < 300$ では 0.3 から 0.5 程度の低い選択確率しか持たない。

最善のモデルが切り替わる n=90 で、選択確率が 0.5 にならずに急激な選択確率の変化が起きるのは、カルバック・ライブラー不一致の差（ΔKL）や AICc の差（ΔAICc）の分布が非対称だからである。最善のモデルの推定は、ΔKL もしくは ΔAICc の正負で判断する。最善のモデルが切り替わる n=90 では、ΔKL と ΔAICc の分布の期待値はおおよそ 0 となっている（図 7a）。しかしながら、非対称な分布では分布の中央値は平均から大きくずれている（図 7a, b）。このことが、選択確率が偏ること、最善のモデルが切り替わるとき急激な選択確率の変化が起こることの理由である。

不一致と規準の関係

図 7 に不一致の差 ΔKL と規準の差 ΔAICc の関係を示す。

期待不一致の推定量として導出されたAICcは、図4に示したように期待不一致の不偏推定量に近い推定量となっている。しかしながら、その分布の形は不一致の分布とは全く異なる。また、不一致と規準の間には複雑な関係があり、そのことが図7に示したΔKLとΔAICcの関係にも現れている。不一致と規準の関係は、漸近的には、式10に示した経験不一致のバイアス項の関数型によって決定される。図7bはn=10の場合であるが、両者に負の相関があり、ΔKLとΔAICcが正の相関をしているという直感とは異なる。

同等の期待不一致を持つモデルの比較

最善のモデルが切り替わるn=90の近辺では、単回帰モデルG_1も切片のみのモデルG_2もほぼ同じ期待不一致を持つ。この場合、どちらのモデルで予測しても同じ予測精度だろうか？これまでモデル選択の拠り所にしてきた期待不一致の観点では、確かに同じ予測精度である。しかしながら、分布の形は両者で異なる。単純なモデルG_2のほうが、不一致の分散が小さい。すなわち、G_2で予測すると、極端に予測が良くなることや、極端に予測が悪くなることが比較的少ない。このような安定した予測を良しとするならば、同等の期待不一致を持つモデルの比較では、不一致の分散の小さい単純なモデルを選ぶという考え方もあるかもしれない。

ま と め

予測のためのモデル選択について、基本的な考え方と方法を説明した。Akaike（1973）以降、モデル選択の研究は著しく発展してきている（坂元ほか1983；小西・北川2004；Konishi and Kitagawa 2007）。生態学の分野でも今後ますますモデル選択の重要性が増すと予想される。

謝 辞

生態学会シンポジウム「生態学におけるモデル選択」および、本特集で、ご講演・ご執筆頂いた岸野洋久先生、巖佐　庸先生、粕谷英一先生に感謝致します。原稿に有意義なコメントを頂いた岸野洋久先生、別所和博氏に感謝致します。また、お二人の査読者に感謝致します。この研究は、水産総合研究センターの運営費交付金の支援を受けた。

引 用 文 献

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Proceedings of the Second International Symposium on Information theory, 267-281, Akademinai Kiado

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

Efron B, Hinkley DV (1978) Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. Biometrika, 65:457-482

箱山 洋 (2015) 予測, 検証, モデル構築. 日本生態学会誌, 65:197-202

Hurvich CM, Tsai CL (1989) Regression and time series model selection in small samples. Biometrika, 76:297-307

巖佐 庸 (2015) 動態モデルにおける完全アグリゲーション：変数を束ねてもモデルの予測に誤りが生じないのはどのようなときか？ 日本生態学会誌, 65:169-177

Iwasa, Y, Levin SA, Andreasen V (1989) Aggregation in model ecosystems II. Approximate aggregation, Ima Journal of Mathematics Applied In Medicine and Biology, 6:1-23

Iwasa, Y, Andreasen V, Levin S (1987) Aggregation in model ecosystems. I. Perfect aggregation, Ecological Modelling, 37:287-302

Konishi S, Kitagawa G (2007) Information criteria and statistical modeling. Springer Verlag, New York

小西 貞則, 北川 源四郎 (2004) 情報量規準. 朝倉書店, 東京

Kullback S, Leibler RA (1951) On information and sufficiency. The Annals of Mathematical Statistics, 22:79-86

Linhart H, Zucchini W (1986) Model selection. Wiley, New York

Ludwig D, Walters CJ (1985) Are age-structured models appropriate for catch-effort data? Canadian Journal of Fisheries and Aquatic Sciences, 42:1066-1072

坂元 慶行, 石黒 真木夫, 北川 源四郎 (1983) 情報量統計学. 共立出版, 東京

Shannon CE, Weaver W (1949) The mathematical theory of communication. Urbana, University of Illinois Press

Sugiura N (1978) Further analysts of the data by Akaike's information criterion and the finite corrections. Communications in Statistics-Theory and Methods, 7:13-26

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

Zucchini W (2000) An introduction to model selection. Journal of Mathematical Psychology, 44:41-61