統計数理 (1999)
第47巻 第2号 273-275

「特集 情報量規準」について

統計数理研究所　石　黒　真木夫 (オーガナイザー)

(受付 1999年9月8日)

広範な情報量規準の世界を十分に紹介するためには，おそらく，よく企画された双書を要する．このひとつの特集がそれに代わり得るはずがないが，まがりなりにも問題の現場から情報量規準の理論までと技術的側面の一端をカバーしたものになった．論文を寄せられた方々に心からお礼申しあげたい．この特集がきっかけになってどこかで新しい研究が芽吹けば本望である．

1. 統計的モデル・データ・情報量規準

情報量規準の世界を理解するには統計的モデルという概念を理解しておく必要がある．ここでは端的に，統計的モデルとはコンピュータの中で実際のデータの振舞いを再現するものであると定義する．コンピュータの中でという限定がきつすぎると思われるかもしれないが，こう理解しておく方がイメージがはっきりする．コンピュータは単に数値計算だけでなく論理的な数式処理もこなす機械であるから過度に制約したことにはならない．

データの振舞いという，よく使われる，あたかもデータが主体的に行動するものであるかのような言い方は，人間によって集められるのを待っている本来受動的な存在であるデータが，しばしば予想をうらぎり，しかも測定のたびに異なる姿を示すことを的確に言い留める表現である．測定が繰り返し得るものであることが暗に仮定されていることに注意されたい．

データの振舞いでもっとも顕著なのは観測を繰り返した時に値が変動する事である．我々の統計的モデルはこの点を「再現」しなければならない．警句的に言えば再現性のとぼしさを再現するのが統計的モデルという事になる．この一見矛盾を解くかぎが確率論である．データを確率変数の実現値とみなし，確率モデルの形で統計モデルを構成するのである．

このような統計的モデルの使い道として

- その再現性を利用したシミュレーション・予測
- データの振舞いを再現するためにモデルに与えた構造を手がかりとするデータ生成機構 (＝現象) の理解

の2通りある．

データを確率変数の実現値とみなすということは，その確率変数になんらかの分布を想定するということである．このデータを生成しているはずの分布を「真のモデル」と呼ぶ．

情報量規準はデータにあてはめられた統計モデルの真のモデルからの離れ具合を測る量である．現在，この範疇に属する規準がいろいろ提案されているが，その嚆矢を放った赤池 (1973) の仕事は，

1．多くの統計的問題が統計モデルの真のモデルからの離れ具合を測る問題として定式化で

きること，
2．測る規準としてKullback-Leibler情報量をとるべきこと，
3．最尤法であてはめられた統計モデルのKullback-Leibler情報量（の差）が推定できること

を示すものであった．Kullback-Leibler情報量はモデルを用いたシミュレーションの結果の分布が真の分布に一致する確率の対数をとったものとみなすことができる．通常の条件のもとではこの量の真値は分らない．1，2番の項目は，さまざまな問題が統一的に扱える理論的枠組が用意されたことを意味している．3番目の項目が指しているのはよく知られたAICであり，赤池が用意した枠組が理論的に有用なだけでなく実用的なものであることを意味している．

ここに拓かれた情報量統計学の分野での研究には以下のようなものがある．

・AICの利用．さまざまなモデルの評価．特にモデル構築過程の制御．
・AICの前提条件をゆるめた場合のKullback-Leibler情報量推定法の研究．
・AICを精密化してKullback-Leibler情報量推定の精度を上げる研究．
・AIC（あるいはこれを拡張したもの）の確率変数としての性質の研究．

2. 各論文の性格

この特集の論文では，一回の測定で得られた時系列データの解析の事例が多い．本来繰り返しがない事象の記録の場合にも定常性を仮定することによって統計的モデルによる解析が可能なのである．

松宮論文：水産資源学の課題とその課題を解決するために収集されるデータおよびそこで用いられている具体的モデルの数々をあげている．現象を支配している要因の追及にAICを利用した説明変数選択を利用した事例が紹介されている．

樋口論文：オーロラとも密接に関係する地球の磁気圏で起こる電流現象の分類に情報量規準によるモデル選択を応用した事例である．データを見て分類するという従来人間の判断を要する作業を自動化する，エキスパートシステムの一つの部品としての使い方といっていいかもしれない．モデル選択を変数選択型と次数選択型に分類すれば次数選択型の事例である．

近藤論文：牛乳の価格と販売量のPOSデータが解析されている．ブランド間の競合の存在下での販売量の変動の研究である．これは自動化することがおそらく不可能な，能動的なモデル構築を要する発見の過程における情報量規準の利用のよい例である．

飯野・尾崎論文：金融資産価格の不連続な変動を表現する「ジャンプ拡散モデル」をデータにあてはめ，AICを計算する方法を提案し，外国為替レート時系列データの解析例を示している．

ここでいう「ジャンプ」は研究者によってはシステムノイズのアウトライア（外れ値）と呼ぶものである．アウトライアの概念はモデルと密接な関係にある．あるモデルで説明できない値が出現した場合にアウトライアと呼ぶからである．データにアウトライアが含まれる場合，いわゆる裾の重い分布を利用してアウトライアが原理的に現れないようにする方法がある．アウトライアを含むデータからモデルのパラメータを推定するロバスト推定法と呼ばれる一群の手法もある．また，その両者の中間に位置すると見なされる方法がある．この論文でも利用されている混合正規分布を利用する方法である．

矢船・石黒・北川論文：薬物動態解析にあたってデータ測定点の選択にKullback-Leibler情報量を利用する研究である．真のモデルが分っていれば問題は簡単であるが意味がない．真の

分布のパラメータの分布が個体差に由来し既知であるという前提のもとで議論が展開されている．

井元・小西論文：真の分布とモデルを固定した上でいくつか異なるKullback-Leibler情報量の推定量を含む種々のモデル選択規準の挙動を比較している．真の分布を知っている超越者の観点からの比較である．

北川・小西論文：AICの拡張を論じた論文である．ロバスト推定量又は罰則付き最尤法でパラメータ推定を行った場合のモデル評価，ベイズ型予測分布モデルのモデル評価などが論じられている．

石黒論文：情報量規準を利用する研究におけるモデル構築，あてはめ，公表の過程で有用な技術を紹介したものである．

論文中で言及されているデータの提供が可能かどうか問い合わせたところ，松宮氏から水産資源データ，近藤氏からPOSデータに関してデータ・関連情報の提供が可能である旨の回答をいただいた．関心のある方は著者に相談されたい．樋口氏からはデータの提供は著者の権限の範囲にないが，関連情報の提供には応じるとの申し出をいただいている．

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 (eds. S. Kotz and N. L. Johnson), Springer, New York, (1992).)