1998 年情報論的学習理論ワークショップ
1998 Workshop on Information-Based Induction Sciences (IBIS'98)
Hakone, Japan, July 11-12, 1998.

モデル選択とその周辺

Model Selection and Related Topics in Statistical Sciences

伊庭 幸人*
Yukito Iba

Abstract: There are a number of statistical methodologies, each of which has its own philosophy and mathematical background. A way to understand them is to regard them as tools for defining "*possible* samples from *possible* worlds" from a set of *real* samples from the *real* world. That is, different methods correspond to different ways to generate *imaginary* data from *real* data. In this note, we review the issues on model selection from this point of view. We also give a few remarks on philosophi-mathematical subjects around model selection. In the appendix, we discuss psychopathology of schizophrenia from the viewpoint of statistical sciences.

1 はじめに

「統計学とは何か」と聞かれたら何と答えるか。ひとつの答えとして、「データと事前の知識から、データを解釈するのにふさわしいアンサンブルを構成して、その中でデータを解釈する技術」ということが可能ではないだろうか。いいかえれば、「現実の世界から可能な世界(の集合)を構成する技術」ともいえる。「雑音を除去して本来あったものを復元する」「測定値から理論式のパラメータの正しい値を推定する」という考えを哲学でいう「模倣説」の統計学とするなら、この答えの背後にある考えは「構成説」の統計学といえる。

これは、情報理論も同じではないかと思う。Shannon の情報圧縮の基本定理はアンサンブル (ある確率分布から発生するメッセージの全体) についての言明である。1個の孤立したメッセージを符号化するというのは意味がない。さらに、実際の画像やテキストの圧縮を考えれば明らかなように、そこで使われる「アンサンブル」は、一般には既知ではなくて、それ自身データを分析して構成しなければならない。この点を強調したのが Rissanen であって、これによって、統計学と情報理論はその基盤の共通性を確認したことになる。本稿では、この観点、「現実の世界から可能な世界を構成する技術」という観点から、多様な統計手法、特にモデル選択に関する手法を見なおしてみることにする。

本稿の内容は主に解説とコメントである[1]。筆者は、実世界のデータ解析に携わった経験も少ないし、理論の面でもこの分野の専門家ではない。本稿で述べることも、いわば素人の感想であるから、「本物」の専門家からみれば多くの奇異な点があると思うが、ご容赦願いたい。モデル選択についてのもっときちんとした解説は、たとえば統計学辞典 [23] や柴田 [18] にある。最新のレビューとしては下平 [20] をすすめる。金谷の解説 [5] も本稿に関連した問題に触れている。

なお、筆者は、階層ベイズ法についてのレビュー [2] の中で、「解釈モデル」と「生成モデル」の関係について EM アルゴリズムや学習方程式、川人らの順逆モデルによる認知の理論に絡めて述べた。この方向の話はその後も進展しているようであるが、本稿の話はそれとは別である[2]。

2 なぜ確率モデルを使うか

モデル選択の話に入る前に、「統計学はなぜ確率モデルを使うか」について考えてみる。統計学や情報理論の専門家があらためてこれを問うことは少ないかもしれないが、実際のデータをいじったり、人工知能の研究者やチューリングマシンに基づく複雑さの定義を好む人達 – 「複雑系」の人などに多い – と議論していると、しばしばこの問題を考えさせられる。

*統計数理研究所 , 〒 1068569 港区 南麻布 4-6-7 tel. 03-5421-8759, e-mail iba@ism.ac.jp,
The Institute of Statistical Mathematics, 4-6-7, Minami-Azabu, Minato-ku, Tokyo 1068569, Japan

[1] 付録として、中井久夫の分裂病論を統計科学の視点から論じた小文を付した。

[2] むしろ [2] の 3.3 節で短く「ここには前に述べたのとはまた別の種類の "循環" が存在する」と述べた部分を膨らましたものとも考えられる。

われわれの観点からすれば、この答えは「可能な世界の集合を表現するためにもっとも便利な数学が確率論だから」ということになる。確率モデルのもっとも重要な特徴は、そこから (適当な乱数発生器があれば) サンプルを発生させることが可能な点である。平均値というのは単なる数値にすぎないが、正規分布をあてはめた場合、その正規分布からあらためてデータを発生させることができる。このことから生じる「自分で自分を評価できる能力」がいろいろな統計手法の基礎になっている。

一部の人工知能の研究者は、アナログ的な「確信度」を考えることが重要で、特に「確率」である必要はないと主張するが、これは、「可能な世界」をアンサンブルとして考えることの重要性を軽視している。外延的な「可能性の集合」を扱うとすれば、仮に確率構造そのものではなくても、それに相当近い数学を必然的に扱うことになるように思われる。

確率モデルの使用と並んで重要なのが、モデルに内部構造としての「未知パラメータ」を考えることである。さきに、正規分布を「あてはめる」といったが、あてはめるという操作が意味を持つためにはなんらかの意味でパラメータに相当するものが必要である。

たとえば、ベイズ統計の枠組みでは、世界の可能性についての事前の知識をパラメータ x の事前分布 $\pi(x)$ とパラメータ x を定めたときのデータ y の分布 $p(y|x)$ によって表現する。データ $\tilde{y}$ が得られると、ベイズの定理から事後分布 $p(x|\tilde{y}) = p(\tilde{y}|x)\pi(x)/p(x)$(ここで $p(x) = \sum_x p(\tilde{y}|x)\pi(x)$) が定義され、さらに、「生成される可能性のあるデータの分布」である予測分布 $p(y|\tilde{y}) = \sum_x p(y|x)p(x|\tilde{y})$ が計算される。

この枠組はとても簡単ではあるが、単に各データ y に主観確率を割り当てる場合と比較すると、「パラメータ」x が含まれている分、複雑になっている。「パラメータ」を導入して、階層的な表現をとったおかげで、予測分布のようなものの定義が可能になったのである。Rissanen は情報圧縮の立場から、パラメトリックモデルを「2段階符号化」と呼んだが、この表現はベイズの枠組に内在する階層性をうまく表している[3,4]。Kolmogorov の複雑さのような階層性のない表現では、あるデータの複雑さは与えられても、「そのほかにどういう可能性があったか」という問題にはデータに依存しない答しかでてこない。これは「事前確率」しかない世界である。逆に、チューリングマシンに基づいて、予測分布のようなものを定義しようとしたら、何らかの階層的な仕掛けが必要になると思われる。

3 モデル選択とアンサンブルの表現

確率モデルを考えて、パラメータを推定すれば、そこから「ありうるデータ」を作り出すことができる。これを使って、将来の予測や残差の有意さなどについての情報が得られる。では、複数のモデルのうちどれが良いと判断されるのかを尋ねられたら、どうすればいいのだろう。これがいわゆるモデル選択の問題である。ここでは、統計学がこの問題についてどういう風に答えるのかをわれわれの立場からみてみよう。

3.1 ベイズ統計, MDL の場合

ベイズ統計の枠組でもっとも簡単にモデル選択を考えるには、モデルの事後確率を考えればよい。データの生成確率とパラメータの事前分布の組からなるモデル $\{p_i(y|x), \pi_i(x)\}$ に対して、もう一段レベルの高い “モデルの集合の上の事前分布” $\tilde{\pi}(i)$ を考える。するとデータ $\tilde{y}$ を得たときのモデル i の事後確率は、

$$p(i|\tilde{y}) = \frac{\sum_x p_i(\tilde{y}|x)\pi_i(x)\tilde{\pi}(i)}{\sum_i \sum_x p_i(\tilde{y}|x)\pi_i(x)\tilde{\pi}(i)}$$

となる。これが一番高いモデルが良いモデルだということになる。これは「ありうるデータ」を介さずに「ありうるモデル」の集まりをダイレクトに指定したかたちになっている。

$\tilde{\pi}(i)$ の影響が定量的にあまり重要でなく、$\pi_i(x)$ が i に依存しなければ、上の規準は $-\log \sum_x p_i(\tilde{y}|x)\pi(x)$ を最小にするモデル i を選べというのに近い。これは、メッセージ y を $p_i(y|x)$ と $\pi(x)$ を用いて符号化したときの符号長とみなすことができる。さらに、x が実数でデータ数 N がパラメータ x の個数 k に比べて多いときは、$\sum_x$(というか $\int dx$) を鞍点近似することができる。その結果が、いわゆる BIC 規準 [14] とか MDL 規準 [12] である[5]。

MDL や BIC は、本来はモデルの予測能力を論じるものではない。次に論じる AIC との関係は微妙である。AIC のように予測能力を情報量損失ではかった場合、その表式は符号長の式と似たものになる。また、「模倣説」

[3] ここでいっているのは、普通のベイズモデルに内在する階層性のことである。後述の、いわゆる「階層ベイズモデル」はこの意味では2重 (あるいは3重以上) の階層を持つモデルということになる。

[4] 統計学の目的を直接測定できないパラメータの推定にあると考える立場 (問題によってはそのような観点が適している場合もあるだろう) からみれば、設定に「パラメータ」が含まれるのは当然である。これに対して、「予測」(赤池) や「情報圧縮」(Rissanen) のような目的のための手段として統計手法を考えるならば、パラメータの道具としての役割が浮かび上がってくる。あとでみるように、事前分布なしでも「ありうるデータ」の生成はできるが、パラメータに相当する内部構造なしにはできない。

[5] MDL の場合は、特定の $\pi(x)$ の選び方がいかに普遍的であるかについての議論がこれに加わる。しかし、筆者には「MDL は並のベイズとは全く別物」とはあまり思えない。むしろ、Jeffreys 流の検定 [4] や Lindley paradox [15] など過去のベイズ理論とのつながりの方を強く感じる。

的な考え方になるが、 MDL や BIC を使って「真のモデル」に一致するモデルが選べれば、それは予測の意味でも良いはずだという議論もできる。「真のモデル」とデータ数無限大の極限については後でまた触れる。

3.2 AIC の場合

AIC(赤池情報量規準)[13, 1] は、統計学の流儀をベイズと非ベイズ (いわゆる正統派、あるいは「頻度主義」) に分類した場合には後者に属する考えである。したがって、事前分布は使わず、モデルによるデータの生成を通じてモデルの評価を行う。モデルの評価規準としてその「予測能力」を考えるのが AIC の特徴であるが、「予測能力」を測るには、実際には存在しない「ありうるデータ」の集合が必要となる。この部分を AIC では次のような論理で処理している。

まず、データ $\tilde{y}$ に対して "対数尤度" $\log p(\tilde{y}|x)$ が最大になるようなパラメータ $\hat{x}$(最尤推定値) を求める。これは、たとえば、データの "真の分布" と $p(y|x)$ の間の情報量損失が最小になるような近似として正当化される[6]。

つぎに、与えられたモデル族 $\{p_i(y|x)\}$ の中で最良と考えられるものを求めるのであるが、この場合に、各モデルの良さを評価するのに、「もし $p(y|\hat{x})$ が本当の分布であったとしてそこからデータが発生しているとしたら、$p(y|\hat{x})$ の良さは "見かけの良さ" $\log p(\tilde{y}|\hat{x})$ からどれだけずれるか」という風に考えて、補正項を導出する。$\log p(\tilde{y}|\hat{x})$ は「自分自身の良さを同じデータから自分で評価している」ために明らかに甘い評価になっているが、それがどの程度であるかを自分自身で確かめるわけである。この段階では、$p(y|x)$(に $\hat{x}$ を入れたもの) は、評価される側のモデルであると同時に、「ありうるデータ」の生成を通じてモデルを評価するためのモデルとしても使われている。このように同じモデルを 2 重に使用することで、「モデルに自分自身を評価させる」というのが AIC のひとつのポイントであると考えられる。

具体的には、$p(y|\hat{x})$ から実際のデータと同数 (N 個) のデータ w を発生させ、w から求めた最尤推定値 $\hat{x}(w)$ を入れた分布 $p(y|\hat{x}(w))$ の「見かけの良さと真の良さの差の期待値」B を、$p(y|\hat{x})$ が真の分布だという仮定のもとに計算して補正項とする[7]。補正項 B を式で書くと、

$$\sum_w p(w|\hat{x}) \cdot \left\{ \log p(w|\hat{x}(w)) - \sum_z p(z|\hat{x}) \log p(z|\hat{x}(w)) \right\}$$

となる。N が大きいとして展開を行って計算すると、周知の簡潔な結果 $B \sim k$ (k はパラメータ数, ここでは情報量を測るのに「N で割らない」、主要項が N のオーダーになるような定義をしている) が得られる [13]。補正項への寄与がパラメータの種類に依存しないのは、推定しにくいパラメータほど情報量損失への寄与もすくないためである (調和振動子の統計力学の「等分配の法則」にちょっと似ている)。

3.3 TIC の場合

モデルを評価するためのデータを生成するモデルとして $p(y|\hat{x})$ を取ることが必然的であるとは必ずしもいえない。いわゆる TIC(竹内情報量規準) [24, 25] は、生成モデルに相当する部分を未知の "真の分布" $g(y)$ として、AIC に対応する結果を与えたものである。B に相当する補正項の形は AIC のように簡単にならず、評価されるモデル $p(y|\hat{x})$ によって複雑な形になる。一般に補正項は $g(x)$ に依存するが、(少なくとも) 簡単な場合には、データから推定することができる。たとえば、評価されるモデルが正規分布の場合は、$g(x)$ の 2 次と 4 次のモーメントを用いて補正項を書きあらわすことができる [24, 25] ので、これをデータから計算した値に置き換えて、補正項を求めることができる。

TIC はデータを生成するモデルとしてかなり一般のものを仮定しても、AIC に似た議論が可能なことを示している。k/N が大きい漸近極限の力をフルに利用しているともいえるかもしれない。問題点としては、まず、評価されるモデルがより複雑な場合にうまく補正項の評価ができるかどうかということがある。また、可能であってもその揺らぎが大きければ、生成モデルについてより強い仮定をした AIC よりかえって不安定になる可能性もある。

3.4 尤度比検定の場合

通常の尤度比検定の場合、分布 $p_1(y|x)$ で示される「帰無仮説」と $p_2(y|x)$ で示される「対立仮説」を比較するのに、データ $\tilde{y}$ についての尤度比 $p_1(\tilde{y}|\hat{x}_1)/p_2(\tilde{y}|\hat{x}_2)$ を使う。ここで $\hat{x}_1$, $\hat{x}_2$ はそれぞれモデル p_1, p_2 での最尤推定値である。ここで「小さいモデル (パラメータに関する拘束の強いモデル)」である「帰無仮説」$p_1(y|x)$ の方からデータを生成させたと仮定して、尤度比の対数の振舞いがどうなるかを調べ、それと実際の尤度比の比較で、p_1, p_2 のいずれかを採用するかを決めるのが尤度比検定の手順である。尤度比検定は有意水準を決めての「検定」であって、予測能力を評価するという論理は使わないが、数学的には AIC に近い。逆に AIC は尤度比検定と最尤推定を「予測」という観点から融合したもの

[6] 以下では、i.i.d. を仮定する。また、複数のデータ $y = \{y_1, ..., y_N\}$ に対して、$p(y|x) = \Pi_j p(y_j|x)$ のように書く。

[7] ここで、バイアスの期待値 B を評価するのがポイントである。単に $p(y|\hat{x})$ からデータを発生させて自分自身を直接評価しても、分布のエントロピー $-\sum_y p(y|\hat{x}) \log p(y|\hat{x})$ を得るだけで、効果はない。

ともいえる。

われわれの見地からすると、この枠組みでは、尤度比の構成に使われるモデル p_1, p_2 がデータの解釈に使われるモデルであるのに対し、小さいモデル (拘束の強いモデル)p_1 がこれらを評価するための「ありうるデータ」を生成するモデルである。より複雑な逐次検定の枠組み [25] に対しては、検定の手順の中で、「ありうるデータ」を作るモデルが系統的に差し替えられていくという見方ができる。

3.5 bootstrap の場合

統計学でいう bootstrap 法では、データのリサンプリングを通じて誤差などを計算する。N 個の標本からなり、各標本が i.i.d.(独立同分布) とみなせるデータがあったとき、乱数によって標本の番号を復元抽出して、各 N 個からなる「ありうるデータ」の集合を生成する。もとのデータと「ありうるデータ」の違いは、復元抽出なので、重複して選ばれる標本がある一方で、使われない標本もでてくることである。

bootstrap 法のモデル選択に対する応用として、bootstrap 法を使った情報量規準 (EIC) がある [3, 19]。この方法では、AIC で $p(y|\hat{x})$ を使って「ありうるデータ」を生成するところで、かわりに bootstrap 法を使って B に相当する罰金項を評価する。これによってモデルの予測能力を見積もるわけである。

i.i.d. の仮定のもとでは復元抽出による「ありうるデータの生成」という考えは広い適用範囲を持っている。しかし、赤池が強調しているように、世の中のデータの多く、たとえば時系列や画像などは決してそのまま i.i.d. と見なせるものではない。こうした場合に bootstrap の考えを拡張しようとすると、いろいろな手続きを持ちこむことになる。たとえば、画像や時系列では、各測定点の標本そのものではなくて、ある塊 (ブロック) を復元抽出することが考えられる。また、別の考え方としては、データにモデルをあてはめた (データをモデルで解釈した) 残りの「残差」の部分を復元抽出して擬似データを生成することもできる。この方法は EIC の平滑化や時系列への応用では常用されている[8]。

前者のような考え方では、データを解釈するモデルとは別に、「ありうるデータの生成」のためのモデルを暗黙のうちに導入していることになる。また、後者 (残差の bootstrap) では、残差があてはめたモデルによって違うので、それを通して、データを解釈するモデルが「ありうるデータの生成」の過程に絡んでくる。

[8] 単純な bootstrap とこれらの違いは、ある種の場合、たとえば回帰分析では、観測点の分布をどうモデルに組み入れるかという問題に関係する。

3.6 cross-validation の場合

cross-validation(交差確認法) と呼ばれるのは、要するに、データを何らかの形で、推定・学習用のデータとテスト用のデータに分けて、前者でモデルを訓練した結果を、後者で評価するという方法である。誤差の推定や予測能力の評価を通じてのモデル選択などがこれによって可能になる。素朴な cross-validation ではデータを適当に 2 分してしまうが、より洗練された方法として、データのうち 1 個をテスト用に、残りの $N-1$ 個を学習用に使い、N 通りのデータの分け方についての結果の平均をとるという方法も使われている[9]。

cross-validation の発想は「実力テスト」のそれと同じといえる。教科書から問題を出すと、答えを丸暗記してくる生徒がいるといけないから、問題の一部をテスト用に隠しておくというわけである。

i.i.d. の仮定をして理論的に議論をする場合、推定・学習用のデータとテスト用のデータの分け方は、純粋に技術的・数学的な問題であるように思える。しかし、実際には、むしろその分割の方法を通じて、何を i.i.d. と見なすのか、どのような「未来」の範囲での予測の最良性を望むのかが表現 — モデル化 — されていると考えるべきである。この暗黙のモデルを用いて、「ありうるデータの生成」を行うのが cross-validation であるといえる。

3.7 応用問題: 階層ベイズ法と罰金付き最尤法

データの性質を、多項式のようなパラメトリックモデルであらわすかわりに、もっと柔軟な表現、たとえば「滑らかな曲線でよく近似される」のような表現でモデル化できれば便利である。こうした考えは古くからあるが、1980 年代からは時系列や画像などの分野を中心に盛んになった ([1, 8, 11] 及び伊庭 [2] の文献参照)。

このようなモデルを表現するには大きくわけて 2 つの方法がある。ひとつの方法は、ベイズの枠組によるもので、いくつかの階層をもったベイズモデルを利用する。この方法は、データ数 N に匹敵する多数のパラメータ x を考える点と、パラメータ x の事前分布 $\pi(x|\alpha)$ を「パラメータに関する無知の表現」と考えずに、滑らかさなどの情報を積極的に表現する手段 (informative prior) とする点に特徴がある。この際、ひとつ階層が上の “ハイパーパラメータ”α は滑らかさの程度などの、一段上のレベルの情報をあらわすのに用いる。

[9] この方式の cross-validation は漸近的に TIC に等価であることが示されている [22]。ということは一致性をもたないわけである。一致性を持つための条件は [16] にある。

もうひとつの見方は、罰金付き最尤法と呼ばれる考え方である。この見方では、事前分布は考えず、データ y が生成される確率分布 $p(y|x)$ のみを考えるが、「パラメータ x の推定の手続き」として、普通の最尤法の $\log p(\tilde{y}|x)$ の最大化ではなく、$\log p(\tilde{y}|x) + U_\alpha(x)$ の最大化を考える。ここで、「罰金」$U_\alpha(x)$ を $\log \pi(x|\alpha)$ に等しくとれば、x の罰金付き最尤推定はベイズモデルで x を事後分布

$$p(x|\tilde{y},\alpha) = \frac{p(\tilde{y}|x)\pi(x|\alpha)}{\sum_x p(\tilde{y}|x)\pi(x|\alpha)}$$

を最大化するように選ぶのと同じである。α の決定方法や事後分布・予測分布の利用に踏み込まないレベルでは、informative prior を用いたベイズモデルと罰金付き最尤法は同じ式の別の解釈ということになる[10]。

こうしたモデルについて、ハイパーパラメータ α の最適な値を決めること、および、informative prior あるいは罰金項を含めたモデル全体を選択するにはどうしたらよいかということは、近年の統計学の重要な問題であった。ここでは、その全体をレビューする余裕はないが、もっとも重要と思われる問題にだけ触れておく

ベイズ的なモデル選択の場合にモデル i を選択したのと形式的に同じ筋道でモデルの番号 i のかわりに α の推定を考えると、$l(\alpha) = -\log \sum_x p_(\tilde{y}|x)\pi(x|\alpha)$ のような量 (赤池の ABIC, MacKay の evidence に符号・定数因子を除き対応) を最小化すれば良いように思われる。このような考え方は古くからある (たとえば Good の TYPE II 最尤法)。より新しくは、赤池による “赤池ベイズ”(ABIC 法)[1] や Gull や MacKay による evidence framework[11] などが良く知られている。また、隠れマルコフモデルや時系列の状態空間モデルの最尤推定 [8] も同じ構造をもっている[11]。

この方法でのハイパーパラメータ α の選択が、平滑化などの場合にかなりうまく機能することは、多くの実例で示されているが、その意味はどうだろうか。ベイズの枠組の近似として、あるいは、最短符号長という立場からみれば、$l(\alpha)$ の最小化でよいように思われる。問題は予測という立場をとったときである。この場合も、ベイズモデル $p(y|x)$, $\pi(x|\alpha)$ で指定される構造が「実在」していて、良い α を選ぶことが「真の事前分布」を選ぶことに、ひいては良い予測に繋がるのなら良いだろう。しかし、必ずしもそこまでの「実在性」がない場合には問題がある。

$l(\alpha)$ を最小にする手続きは、大雑把にいえばハイパーパラメータ α だけを与えたときの混合分布 $p(y|\alpha) = \sum_x p(y|x)\pi(x|\alpha)$ の予測力を最大にしようとしていることになる。しかし、実際にこのようなモデルを使用するときに、α のみを与えた混合分布 $p(y|\alpha)$ を用いて予測を行うということはあまりない。むしろ、データからの罰金付き最尤推定値 x^* を代入した分布 $p(y|x^*)$ の予測力、あるいは、ベイズ的に定義された予測分布 $p(y|\tilde{y},\alpha) = \sum_x p(y|x)p(\tilde{y}|x)\pi(x|\alpha)/p(\tilde{y}|\alpha)$ の良さを最大にしたい場合が多いと考えられる。主として罰金付き最尤法に近い立場で、これを評価する規準を作ろうとしたのが、EIC[3] や GIC[10] などであり[12]、よりベイズ寄りの立場で考えたのが PIC[9] である。

4 いろいろな観点から

「ありうるデータ」あるいは「ありうる世界」を生成するモデルという観点からの分析は表にまとめた通りである。以下では他の論点について簡単に紹介する。

4.1 部分か全体か？

手法を分類するための別の軸として、どれだけ「世界の全体」をみているか、ということがある。この意味で 1 対比較の検定はもっとも問題を局地化して考えているといえる。AIC はそれに比べると、より問題の全体を扱う枠組をめざしていると思われる。しかし、AIC もパラメータ数 k について可能なモデルの数が指数的に多くなるようなケースではうまくいかない[13]。これに対してベイズ系の手法は、考えている範囲で、すべての可能なデータ、パラメータ、モデルについて矛盾なく確率を割り当てるので、もっとも「全体の記述」が可能な手法といえる。

[10] 同じなら、ベイズモデルのほうがかっこいいようにも思える。しかし、ベイズ的な解釈は事前分布 $\pi(x|\alpha)$ が単独でも意味を持つ、つまりデータなしで $\pi(x|\alpha)$ をシミュレートしたときにある程度意味を持つことを含意している。これは相転移点以下のイジング模型や線過程を含む 2 次元画像モデルなどでは疑問である。一方、罰金項という表現はデータおよび分布 $p(y|x)$ と一緒になってはじめて有効に働くということを示唆している。詳しくは [2]。

[11] これらの分野では、(罰金を含んだ形の)ABIC に相当するものを AIC と呼ぶことがある [8]。これはすぐあとでいう「混合分布」の AIC の意味である。これらの場合、観測されない状態がパラメータ x に、パラメータがハイパーパラメータ α に対応する。

[12] 本ワークショップでの村田昇氏の講演もこれに関係したものではないかと思う。

[13] たとえば、重回帰分析ですべての submodel(可能な説明変数の数が K 個とすると 2^K 個) の比較を行う場合には、AIC は本来は使えない (関連した理論的な結果については [17])。ただし、bootstrap 法による規準 (EIC) をうまく利用することでこの問題が解決できるという報告もある [7]。この問題に関連しては、統計物理的手法による重回帰分析の解析 (本予稿集の樺島祥介氏の論文参照) も興味深い。

まとめの表

手法	評価されるモデル	評価のためのモデル	備考
ベイズ	$p_i(y\|x)$, $\pi_i(x)$	$\tilde{\pi}(i)$	
AIC	$p(y\|x)$	$p(y\|\hat{x})$	$\hat{x}$ は最尤推定値
TIC	$p(y\|x)$	$g(y)$	$g(y)$ は “真の分布”
尤度比検定	$p_1(y\|x)$, $p_2(y\|x)$	$p_1(y\|x)$	p_1: 帰無仮説、 p_2: 対立仮説
bootstrap	$p(y\|x)$ or 任意の手続き	「標本」の復元抽出	「標本」の定義がモデル
cross-validation	$p(y\|x)$ or 任意の手続き	評価用のデータ	評価用のデータを分ける仕方がモデル

「評価のためのモデル」は直接使用する場合とバイアス評価に使用する場合がある (本文参照)

「全体の記述」が可能ということは、帰納推論のための「自動機械」として使いやすいという意味でもある。実際に、「自動機械」の開発が本来の目的である人工知能の研究者にはベイズ系の手法が容易に受け入れられるようである。これに対して、検定はもっとずっと「道具」的な側面が強い。仮説を作るのは人間であって、機械ではない。AIC、EIC は、その中間にあって、使いやすい統計学の姿を模索する試みの中から生まれて来たとも考えられる[14]。

4.2 現実は無限に複雑か ?

データ数 N 無限大の漸近極限を考える上で重要なのは、有限の複雑さの「真のモデル」を想定するのかどうかということである。このことは、モデルの「一致性」(真のモデルが考えている範囲に含まれる場合にデータ数無限大の極限で正しいモデルが選ばれること) の意義を考える上で重要である。赤池や AIC を支持する統計学者の著作 [21] をみると「現実は無限に複雑である。しかし、われわれはデータの有限性ゆえにその一部しか知り得ないし、可能な以上に知ろうとすればかえって不利益をこうむる。」という世界観が伺われる。これは、たとえば、セメント工場の制御などを想定すれば理解できる考えである。また、こうした考えは、検定の漸近効率の定義などを通じて、古典的な検定論に潜在的に含まれているとも考えられる。このような状況では、データ数の増大はパラメータの精度の増大よりも、モデルの複雑さ (有効なパラメータ数) の増加をもたらす。したがって、BIC や MDL の $\log\sqrt{N}$ というパラメータ精度を示す因子は現れないことになる。

一般に理論を作ろうとすれば、思想的に「構成説」であろうと「模倣説」であろうと、なんらかの極限を考察することは避けられないように思われる。しかし、与えられた極限での結果を決定的と考えるよりは、極限と現実の関係を絶えず見なおしていくように努力する方が、理論と応用の双方にとって有益なのではないかと思う。その意味で、赤池や柴田 [17] が試みているような、「真のモデルが無限に複雑な場合」についての数学的考察は有意義であろう。

4.3 相互評価の非対称性

情報量損失 (KL-divergence)

$$D(P||Q) = \sum_x P(x)\log\frac{P(x)}{Q(x)}$$

はいわゆる「擬距離」であって、分布 P と Q について非対称 $D(P||Q) \neq D(Q||P)$ である。この効果は分布の間の距離がある程度以上離れている場合にのみ問題になる。また、P,Q が正規分布の場合、それぞれの分散が違う場合のみ、この非対称性がみられる。

このような非対称性は、日常の感覚からいっても、もっともな感じがする。A 氏が B 氏のことを自分に似ていると思っても、B 氏は A 氏のことを自分に似ていないと思うかもしれない。これは A 氏と B 氏とで、「似ている」ということを定義するのに、どのような面をどれだけ重視するかが異なっているためである。もし、A 氏が B 氏にあらゆる面で似ているのであれば、価値観もまた似ているであろうから、「1 次の無限小」のオーダーでは 2 人の違いについての 2 人の意見は一致することになる。

モデル選択という文脈でいえば、AIC の考え方は、分散 (雑音) も平均値や回帰係数もすべて差別せずに「パラメータ」としてみることに特徴がある[15]。「雑音」の部分もモデルの一部として、他のパラメータと全く同様に推定されると考えることは、赤池の「構成説」的な統計観の成立にとって重要なポイントであるが、これは同時にここでいう意味の非対称性を重視しないことにもなっているように思われる。

現在の筆者からみると、非対称性の問題、あるいは、「分散のような変数」と「平均のような変数」の違いは、やはりそれなりに重要なのではないかと感じるが、不勉強のために十分理解するに至っていない。統計学や情報

[14] どの種類の統計手法に肩入れするかという問題は、今世紀の科学哲学の展開にも関係している。いわゆる「論理実証主義」に対して、Popper が展開した「反証主義」は、仮説の設定の非合理性 (自動化不能性) と反証の客観性 (自動化可能性) を主張して、非ベイズ的な検定論を積極的に擁護しているともみられる。

[15] これに対して、雑音の部分を区別する、あるいは、区別した方が良いという主張も多い。たとえば、柴田 [18], 金谷 [5, 6] など (後者はもともとの AIC とはかなり違う状況を想定しているが)。

幾何の専門家が、このあたりの問題について明快な研究成果あるいは解説を与えてくれることを期待したい。

謝辞

専門的な助言を頂き、多数の文献を教えて下さった下平英寿氏に感謝します。石黒真木夫氏には原稿へのコメントを頂きました。その他の情報量統計学派の方々にも、田辺国士氏をはじめとして、日頃いろいろお教えを頂いています。柳本武美氏には、非対称性の意義、反証主義と検定、Lindley paradox などについてお教え頂きました。

参考文献

[1] Akaike,H. (1998), *Selected Papers of Hirotsugu Akaike*, Eds. E. Parzen, K. Tanabe, G. Kitagawa, Springer, New York.

[2] 伊庭幸人 (1996) , 学習と階層 — ベイズ統計の立場から —,「物性研究」65-5 (1996 年 2 月号) 657-677. (www.ism.ac.jp/~iba/)

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

[4] Jeffreys (1961), *Theory of Probability*, Oxford, 3rd ed., (1st ed. 1939).

[5] Kanatani, K. , What is the Geometric AIC — Reply to My Reviewers, (www.ail.cs.gunma-u.ac.jp/~kanatani/)

[6] 金谷健一 (1996) 情報量基準による幾何学的モデルの選択, 情報処理学会論文誌, Vol.37, No.6, 1073-1080.

[7] 北川源四郎, 石黒真木夫, 坂元慶行 (1994), 情報量規準 AIC と EIC, 電子情報通信技術研究報告 (信学技報), Vol.92, No.503, IT92-133, 49-62.

[8] Kitagawa,G and Gersch,W.(1996), *Smoothness Priors Analysis of Time Series*, Lecture Notes in Statistics, No.116, Springer-Verlag, New York.

[9] Kitagawa,G (1997), Information criteria for the predictive evaluation of Bayesian models, *Communications in Statistics, Theory and Methods*, Vol. 26, No. 9, 2223-2246.

[10] Konishi,S. and Kitagawa, G. (1996), Generalized information criteria in model selection, *Biometrika*, Vol.83, No.4, 875-890.

[11] MacKay, D.J.C. (1992), A practical Bayesian framework for backprop networks, *Neural Computation*, 4, 448-472.

[12] Rissanen (1983), A universal prior for integers and estimation by minimum description length, *The annals of statistics*, Vol.11, No.2, 416-431. Rissanen,J.(1989) , *Stochastic Complexity in Statistical Inquiry* World Scientific, Singapore.

[13] 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.

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

[15] Shafer, G. (1982), Lindley's paradox, *Jounal of the American Statistical Association*, 77, 325-334.

[16] Shao,J. (1993), Linear model selection by cross-validation, *Jounal of the American Statistical Association*, 88, 486-494.

[17] Shibata,R. (1981), An optimal selection of regression variables, *Biometrika*, 68, 1, 45-54.

[18] 柴田里程 (1988), 変数選択理論の現状, 数学, 36, 344-352.

[19] Shibata,R. (1997), Bootstrap estimate of Kullback-Leibler Information for model selection, *Statistica Sinica*,7, 375-394 .

[20] 下平英寿 (1997), モデル選択理論の新展開, (www.ism.ac.jp/~shimo/papers_ftp-j.html)

[21] Stone,M. (1979), Comments on model selection criteria of Akaike and Schwarz, *Journal of Royal Statistical Society, Ser.B*, 41,276-278.

[22] Stone,M. (1977), An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion, *Journal of Royal Statistical Society, Ser.B*, 39 44-47.

[23] 竹内啓ほか編 (1989), 統計学辞典, 東洋経済新報, 第 III 部, 13 章, モデル選択, 459-465.

[24] 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, No.153, 1976 年 3 月号, 12-18.

[25] 竹内啓 (1983), AIC 基準による統計的モデルの選択をめぐって, 計測と制御, Vol.22, No.5, 445-453.

(以上で、**URL** の頭の http://は略したので注意)

付録： 統計学的な病
— 中井久夫の分裂病論をめぐって —

われわれの日常は、多くの物事を無意味な偶然、「雑音」とみなすことで成立している。すべてのことに意味のある世界に住むとしたらどうであろう。たとえば、自分のまわりで異常に「数字が揃う」と主張する人達がいる。自分が時計を見るといつも「ぞろ目」である、見掛けた車のナンバープレートが毎回「ぞろ目」であった云々。多くの場合、これは偶然で説明できる度数で起こっているにすぎないことを、異常に気にしているということで説明できるだろう。この種の非合理的な解釈は誰でもやることだが、規則的な数列の出現に恐怖や神秘を感じ、自分が何らかの意味で選ばれている証拠と信じるに至るならば、そこには狂気の匂いが感じられる。

中井久夫 [1] は現代の日本でもっともすぐれた精神医学者の一人であるが、中井の分裂病論は、分裂病に親和性のある人間を理解するキーワードとして、「徴候優位性」「微分回路的な認知の優位性」をあげている点で、統計科学の観点から興味深く感じられる。徴候優位性とは、ささいな徴候もそれを雑音とみずに反応することであり、モデル選択の言葉でいうなら overfitting の状況に対応すると考えられる。また微分認知の優位性とは、時系列において現在の微小なトレンドに左右されることを意味し、平滑化の言葉でいうならば undersmoothing に対応すると考えられる。これに対して、中井は、過去の重みから抜け出しにくい、うつ病親和的な性格を、「積分回路的な認知の優位」と特徴づけている。

「徴候優位性」は状況によっては有利な性質であるが、いったんバランスを崩して悪循環に陥れば、分裂病的な状態に導く罠ともなる。発病直前の不安定な状態については、たとえば、以下のように説明されている [2]。

> 徴候的空間は、図式的空間とは異なり、自明的に三次元であるとはいえない。それは無限に多義性の湧く、いわば発酵性の空間である。そして、不安がより精密な予測を無限に追求させる時、それに対する歯止めがない。これを徴候性の相において追求するとき、覚醒度は天井知らずに上がる。(中略) 天井を突きぬければ、超覚醒はもはや単なる量の問題でなく、質的変化である。そこで見えてくるものは、ふだん滑らかに見える意識の肌が、ちょうど肉体の肌を顕微鏡にかけた時のように、あらい木目として映ってくる (どうして普段の意識の肌は滑らかなのであるか? — この疑問のほうが重要かも知れない)。(中略) この裂隙から見えるものは絶対の空無であるが、徴候を求める意識の指向性はその中に何らかの徴候を見ようとする。この試みは、「なにもないもの」に対する解読の試みであって、正解はありえない。ここでさらに恐怖が倍加し、錯誤はとめどない彷徨への路に足を踏み出させる。(太字筆者)

これらの記述を読んで統計的検定やハイパーパラメータの選択 (の失敗) を連想するのは、あまりに「徴候優位」な感じ方であろうか。筆者にはこの記述は統計的検定や情報量統計学の解説のホラー版のように読める。また、この後にくる混乱の描写は、計算機科学的にいえば「とめどない彷徨」による計算資源の枯渇 (「一般フレーム問題」における計算量爆発) を思わせる。中井によれば、破局の直前に、人は「思路の、努力感を全く伴わないところの無限延長、無限分岐」を体験することがあるという。

分裂病の背景に外界からの情報のフィルタリングの障害をみること、また、分裂病の発病前後の状態を外界からの刺激に過敏な「超覚醒」状態と考えることは中井が最初でも最後でもないと思われるが、中井の一連の仕事は、(1 次的に出現する) 妄想の基盤、発病直前の状態、回復後の一過性の症状 (「知覚潰乱発作」) などを一貫して徴候優位という立場から捉えている点で特に興味深いものがある。

「心のランダムさ」といったものは、これらの問題にどうかかわってくるのだろう。中井は、分裂病的な状態というものは内的なランダムさが不足した状態であり、同時に外的な「偶然」を生かす能力が落ちた状態であるととらえているようである。極度の「徴候優位」の状態、われわれの見方からすれば、最大のモデルが常に選択されるような状態では、「雑音」(残差) はゼロになってしまう。これが「偶然のない状態」のひとつの解釈である。別の可能性は、「ランダムさの不足」から過剰な「徴候優位性」が生じるということである。本文でみた通り、モデル選択・hyperparameter tuning のための有力な方法には乱数発生を必要とするものが多くある。もし、このような方法が脳内で使われていたとすると[16]、乱数発生能力の減少はバイアスの評価を不能にさせ、overfitting をもたらすかもしれない。これに関係して、分裂病者では乱数列を言ったり書いたりする能力が低下しているという報告がある (もっとも、これは疾患特異的ではなさそうであるし、乱数テスト [3] が、脳内の過程をそのまま反映しているのかどうかも疑問であるが。)。

分裂病論では、木村敏の考えが、筆者のまわりではよく知られている。木村の考え方が「オートポイエーシス」や力学系理論に親近性があるのに対して、統計科学の精神により近い中井の見方もあることを多くの人に知って欲しかったので、やや場違いかもしれないが、ここで 1 節をさいて論じた[17]。

参考文献

[1] 中井久夫の精神医学関係の著作の大部分は、岩崎学術出版社から出ている「中井久夫著作集」(第 1 期 3 巻、第 2 期 3 巻 + 別巻 2) に再録されている。比較的手に入りやすいものとしては、「分裂病と人類」(東大出版会) があるが、この本はベストではないと思う。最近「最終講義」がみすず書房より出版された。

[16] 幼稚な想像であるが、cross-validation のために入力にランダムにマスクをかけるような機構が脳内にある、といったことが考えられる。

[17] ついでであるが、読者が精神医学者と共同研究される機会があった場合、実験によって患者に負担をかけることのないようお願いしておく。分裂病圏の患者には、一見無害と見えるテストが重大な打撃となる可能性がある。

[2] 中井久夫, 上田宣子, 分裂病発病前後の「不連続的移行現象」— 特に一回的短期間現象とその関連における超覚醒現象について —, 分裂病の精神病理 16, 東大出版会.

[3] Wagenaar,W.A. (1972), Generation of random sequences by human subjects: A critical survey of literature, *Psychological Bulletin*, 77, 65-72. 村上公克・乱数テスト研究会 (1973), 人間乱数, 自然, 28, 8, pp.49-57(1973 年 8 月号). 伊庭幸人, 田中美栄子 (1996), 人間乱数, 「複雑系 4」研究会報告,「物性研究」66-5, 1996. Iba, Y. and Tanaka-Yamawaki, M. (1996), Statistical Analysis of Human Random Number Generators, *Methodologies for the Conception, Design, and Application of Intelligent Systems*, Vol.2, pp.467-472. Eds. Yamakawa,T. and Matsumoto,G., World Scientific, 1996.