学習と階層 — ベイズ統計の立場から —

伊庭幸人

統計数理研究所 〒 106 東京都 港区 南麻布 4-6-7
Email: iba@ism.ac.jp

要旨

学習と階層の問題をベイズ統計の観点から論じた。前半では、平滑化を例として経験ベイズ法について解説した。 特に経験ベイズ法において場合の数 (エントロピー) の果たす役割を強調した。また、EM 法や学習方程式との関係、脳のモデルとの関係についても述べた。後半では、2 次元以上のマルコフ場に対して経験ベイズ法を適用する場合の原理的な困難について論じた。

本稿は 1995 年 4 月 26-28 日に開かれた " **RWC** 情報統合ワークショップ **'95**" に投稿した解説論文 (ISM Research Memo. No.558) である。限定配布の予稿集に載ったのみで、今後の使用は自由ということであるが、適当なレビュー誌が思い当たらないので "物性研究" の場を借りて発表させて頂くことにした。内容は統計数理研究所などで研究されていることを自分なりの理解の仕方でまとめたものに、自分の研究を通じて、あるいはいろいろな人との議論を通じて考えたことを加えたものである。

なお、関連する文献については、本号に掲載の "ベイズ統計と統計物理 (物性研究 1993 年 9 月号) への訂正と追加" も参照されたい。

目 次

1 はじめに **4**

2 ベイズ統計と階層の統合 **4**
2.1 informative prior 4
2.2 経験ベイズ法 5
2.3 経験ベイズ法における “場合の数” の役割 7
2.3.1 階層と事前分布の規格化 7
2.3.2 周辺化の役割 8
2.4 経験ベイズ法の微分形 9
2.5 脳の情報処理と経験ベイズ法 10

3 マルコフ場モデルとベイズの枠組 **10**
3.1 事前分布は生成モデルとしての意味を持つべきか: マルコフ場の場合 10
3.2 モデルを変更するという考え方 11
3.3 評価規準を変更するという考え方 12
3.4 その他の考え方 13

4 おわりに:モデルの階層性は情報処理にとって不可欠か **13**

KEYWORDS

学習、メタ知識、循環、ベイズ統計、 経験ベイズ法、平滑化、場合の数、内部モデル、マルコフ場、相転移、生成、予測

用語について

- “ 経験ベイズ法”
  “経験ベイズ法” というのは別の意味に多く使われる用語だから誤解を招く表現だという意見があ った。これに対して、本論文で説明した手法は “parametric” empirical Bayes というのが正しいが、“経験ベイズ” でも別に良いのでは、という意見もあ った。代案として、言葉の感じからは “ 階層ベイズ法” が良いような気がするが、これは (6) 式に対応するもので、(7) 式の最大化とは違うという説もある。今回はとりあえずもとのままにしておく。
- “ マルコフ連鎖モンテカルロ法”
  統計物理でいう普通のモンテカルロ法 (動的モンテカルロ法) のことを最近の統計学者はこのように呼ぶ。以前 “マルコフ鎖モンテカルロ法” と書いたのは Markov chain の誤訳 (!) であ った。Gibbs sampler はその一種 (熱浴法に相当?) のことで、その下位概念である。 ボルツマンマシンというのは — 何なのか誰に聞いてもはっきりしない。動的モンテカルロ法のことか、simulated annealing 法のことか、特定のモデルの意味か、学習法の意味か — これらをあまり区別しない人のための用語なのかもしれない。
- “ メタレベル”
  “メタレベル” というのは、形式論理での用法からいうと適切な表現ではないのではないか、というコメントがあ った。雑な言い方だ ったかもしれないと反省した。ただ、本論文では触れなか ったが、MDL 規準の立場で階層ベイズ法を解釈した場合、
  (符号化したもの)>
  (符号化の方法を符号化したもの)>
  (符号化の方法の符号化の方法を符号化したもの)
  のような階層に対応するから、その意味では、言語を記述する言語=メタ言語、といった用法と似ているような気もする。

IIW-95 予稿集との異同

- 文献の訂正 ・日本語版の追加 (下線部)

  Akaike,H. and Kitagawa,G.(編)(1994,**1995**)
  時系列解析の実際 I ,II 朝倉書店.
  Kitagawa,G.(1987)
  Journal of the American Statistical Association **82** 1032-1063.
  MacKay,D.J.C.(1992)
  Neural Computation 4 pp.415-447; pp.448-472; **pp.698-714**.
  Sakamoto,Y.(1991)
  Categorical Data Analysis by AIC, Kluwer Academic Publisher
  (坂元慶行 カテゴリカルデータのモデル分析 共立出版 (**1985**)).

- 変更
  Gray et al.(1994), Dubes and Jain(1989) → たとえば、Gray et al.(1994)...
  もう少し詳しい説明や文献引用が望ましいと思うが、細かくなるのでまたの機会に譲る。
- 訂正
  Hidden Markov Chain → hidden Markov chain
  略称としては HMM (Hidden Markov Model) が普通。

1　はじめに

情報処理システムを考える上で、学習はきわめて重要な問題である。オンラインでの動作をみる限り学習なしで動作しているようにみえるシステムでも、よく考えると学習に相当する部分がある。それはたとえば、ユーザに対する事前の調査に含まれていたり、システム構築者による “前処理” 部分のデリケートな調節であったりする[1]。

学習は、“外界から得られたデータに基づいて、規則を構成すること[2]” と表現できる。ここで、規則というのは、個別のデータを正確に記述するという意味ではなく、一般化を含んでいなければならない。そこには、有限のデータから無限を推測ないし構成するという難しい問題が含まれている。これは、たとえば閉区間の上の関数の有限データからの推定を考えてみてもわかる。与えられたデータから考えられる関数は無限にある[3]。

有限のデータから規則を推測ないし構成するためには、システムは内部にあらかじめ予備知識 (モデル) をもたねばならないだろう。しかし、完全な規則を内部モデルとして所有しているなら、規則の構成などははじめから問題にならない。このジレンマを解決する方法は、なんらかの “循環” を持ち込むことに求められる。すなわち、“種” になる知識がはじめにあって、それに基づいて、新たな知識を獲得し、前の知識を修正 ・発展させるわけである。ここで問題になるのは、どのような風に知識を与え、どのように修正していけば良いかということである。

こうした問題について具体的に論じるのが本論文の目的である。Sec.2 においては、informative prior を用いた smoothing を例として、階層的なベイズモデルと経験ベイズ法について解説する。ベイズモデルの設定及び経験ベイズ法による階層の統合において “場合の数” [4] が重要な役割を持つことを強調した。情報統合のための機構として “制約充足” が注目されているが、場合の数の役割は素朴な制約充足の枠組の盲点になっている。Sec.2 では脳における階層の統合にも触れる。Sec.3 では、マルコフ場について Sec.2 の枠組を適用する際の問題点について述べる。ここでは、マルコフ場の事前分布としての適合性が論じられるとともに、Sec.2 の枠組自体に関する疑問も提出される。

2　ベイズ統計と階層の統合

2.1　informative prior

例として、閉区間上の実数値関数を学習するという問題を考えてみる (たとえば、Tanabe and Tanaka (1983))。この問題は画像や時系列の平滑化のプロトタイプになっている。データ $y = \{y_j\}$(j番目の入力 $t(j)$ に対する出力が $y_j,\ \ j = 1, 2...K$) が与えられているとする。これから、関数 $y = f(t)$ を構成したい。予備知識としては、“関数が滑らか” ということを与えることにしよう。

この種の知識を表現するには、ベイズ統計の枠組が便利である。ベイズ統計では、予備知識を “ パラメータ”x の確率分布の形であらわす。これを 事前分布 (**prior distribution**) という。いまの問題では、関数の表現として、多数の点を結んだ折れ線を考え、各頂点をパラメータ $x = \{x_t\}\ \ (t = 1, 2...N)$ とする (簡単のために x の添字と位置 t – 先ほどは連続量として書いた – を同一視した)。すると、線の “滑らかさ”

[1] システム自身が、また、システムと相互作用する外部の対象が、小さな部分に分割して調整/調査できるなら, 学習は比較的容易であろう。一般に精密科学における実験条件の統制はそれを意図している。それが困難なのが **“Real World”** の定義ともいえる。

[2] あえて再構成とはいわない。

[3] 変数が連続量でなく、外界の有限な記号化がすでに与えられている場合でも、多変数の場合には入力として可能な記号列の種類が組合わせ爆発を起こすので、事実上の無限を扱う必要がある。このタイプの問題は、たとえばアンケート調査の解析では良く知られている (Sakamoto(1991) Chap.1-3)。

[4] 連続量の場合は、むしろ “体積効果” というのが良いかもしれない。“エントロピーの効果” と呼ぶのが最も正確だが、かえって誤解される場合もあるので、ここでは避けた。

は x の事前分布

$$\pi_\lambda(x) = \frac{1}{Z_\pi} \exp(-\lambda \sum_t (x_{t+1} - 2x_t + x_{t-1})^2) \tag{1}$$

によって表現することができる。Z_πは分布の規格化定数である[5]。この例のような、知識を積極的に表現するための事前分布を **informative prior** という (これに対して、公平さやランダム性の表現を意図とする事前分布を **ignorant prior** という)。

これ以外に、データ y とパラメータ x との関係についての知識が必要である。これは x を与えたときのデータ yの確率分布 $P(y|x)$ で表現される。これを x の関数と見たとき 尤度関数 (**likelihood function**) $L(y|x)$ と呼ぶ。ガウス的な雑音がデータに含まれていると考えると、尤度関数は普通の最小 2 乗法の場合と同様、

$$L_{\sigma^2}(y|x) = (\frac{1}{\sqrt{2\pi\sigma^2}})^K \exp(-\frac{1}{2\sigma^2} \sum_j (x_{t(j)} - y_j)^2) \tag{2}$$

のように書ける。データの背後にある正解を推定するのでなく、積極的に規則を構成するという立場からすると、雑音というのは不適切な考え方であるが、とりあえず、これはやむおえないものとする[6]。

これで必要な知識が表現できたわけである。ベイズの枠組での推論結果に対応するものは、 事後分布 (**posterior distribution**)

$$P_{pos}(x) = \frac{L_{\sigma^2}(y|x)\pi_\lambda(x)}{\sum_x L_{\sigma^2}(y|x)\pi_\lambda(x)} \tag{3}$$

として与えられる。$\sum_x$はあらゆる可能な x に関する和であり、x が連続量であれば積分に読みかえるとする。事後分布はパラメータの空間における確率分布で、各パラメータがどれだけ確からしいかを与えるものである。一般に事後分布は高次元空間における確率分布になるから、そこからさらに推定値を取り出す手続きが必要である。ひとつの方法は、事後分布の mode を x の推定値として利用することである。これを **MAP** 推定値 (Maximum A Posteriori estimate) という[7]。いまの場合は、

$$P_{pos}(x) = \frac{\exp(-E_{pos}(x))}{\sum_x \exp(-E_{pos}(x))} \tag{4}$$

$$E_{pos}(x) = \frac{1}{2\sigma^2} \sum_j (x_{t(j)} - y_j)^2 + \lambda \sum_t (x_{t+1} - 2x_t + x_{t-1})^2 \tag{5}$$

となり、MAP 推定値は (5) 式を最小にする $\{x_t\}$ によって与えられる。

事前分布を利用してパラメータの間の関係を制約することで、パラメータの数そのものは減らさずに実効的な自由度を減少させたことがポイントである。このおかげで、特定のモデル族の癖を不用意に導入することなしに、overfitting を回避してデータの規則性を抽出することができた。もし、滑らかさを表現する事前分布に由来する (5) 式の第 2 項がなければ、単にデータをつなぐ曲線が得られる。これではデータをそのまま記憶しているのと同じであり、1 章で考えた意味での “規則の構成” を行なったことにはならない。

2.2 経験ベイズ法

前節では informative prior を用いるベイズ統計について解説した。しかし、前節の範囲では重大な問題が解決されていない。それは、“ ハイパーパラメータ”λ,σ^2の決定法が与えられていないという問題である。これは、与えた予備知識について反省を加える手段がないことに対応している。先の例では、$\lambda\sigma^2$ の値の

[5] 実はこのままでは Z_πは無限大になる。これは重心の平行移動と平均傾きの変化の 2 つの変換に対して事前分布の密度が不変な結果であるが、いくつかの方法で除去することができる。あとの議論では $\log Z_\pi$のλ依存性だけが関係するので、ここではこれ以上詳しく議論しない。

[6] 雑音の分布形まで (適当な範囲の中で) 推定できれば、“規則の構成” にだいぶ近づくだろう。本論文ではそこまでは論じないが、雑音の強さの推定までは考える。

[7] 別の方法として、目的に応じた損失関数を最小にするような推定値をとることも考えられる。たとえば、必要な量の 2 乗誤差の事後分布のもとでの期待値を最小にするような推定値は、事後分布による期待値に相当する。

大小に応じて、得られる MAP 推定値は直線に近くもなれば、データの間をぐにゃぐにゃとつなぐ滑らかでない線にもなる。

(5) 式の最小化は、いわゆる “ 罰金つき最尤法”として知られているものと同じである。また、いわゆる “制約充足”形式の処理のうちで、評価関数の大域的最適化に計算が帰着されるものの典型的な例でもある。この意味で、informative prior を用いるベイズ統計は “罰金つき最尤法”や “制約充足”のひとつの解釈と考えられるが、そのような解釈をするメリットはここまでの範囲でははっきりしない。MAP 推定値以外の情報、たとえば、誤差の情報が事後分布に含まれている点が異なるなどと主張してみても、λやσ^2を決める客観的手段なしには、あまり説得力があるとは思えないのである。

以下では、ハイパーパラメータの決定問題についてベイズ統計の枠組の中で考えることにする。“ハイパーパラメータ” もパラメータであるから、それに関する事前分布を考えて、事後分布を作れば良いというのが基本的な考え方である。一般に、尤度関数に含まれるハイパーパラメータをα、事前分布に含まれるハイパーパラメータをγ とし、それらに対する事前分布を$\tilde{\pi}(\alpha, \gamma)$ とする (例ではα,γはそれぞれσ^2,λに相当)。すると、ハイパーパラメータを含めた事後分布$\tilde{P}_{pos}(x; \alpha, \gamma)$ は、

$$\tilde{P}_{pos}(x; \alpha, \gamma) = \frac{L_\alpha(y|x)\pi_\gamma(x)\tilde{\pi}(\alpha, \gamma)}{\sum_{x,\alpha,\gamma} L_\alpha(y|x)\pi_\gamma(x)\tilde{\pi}(\alpha, \gamma)} \tag{6}$$

と書ける。

パラメータ $\{x_t\}$ の数が十分多く、かつ、t が “十分離れた” 各部分の振舞いが互いに独立と見なせるとしよう。“十分離れた” というのは画像なら十分遠く離れた、時系列なら十分時刻の離れたという意味である。さらに、ハイパーパラメータα,γが大域的な性質を決めるもので、かつ、その性質が対象全体でほぼ一様だとする。さきの関数推定の例ではこれらの条件が実質的に満たされていることが多い。この場合、事後分布$\tilde{P}_{pos}(x; \alpha, \gamma)$ において、α,γの周辺分布が鋭い山を持ち、その位置が$\tilde{\pi}(\alpha, \gamma)$ にあまり依存しないことが期待できる[8]。

このような状況を踏まえて、赤池とその協力者たちはハイパーパラメータの決定法として、周辺尤度と呼ばれる量の対数

$$l(\alpha, \gamma) = \log \sum_x L_\alpha(y|x)\pi_\gamma(x) \tag{7}$$

を最大化するハイパーパラメータα,γを選択する方法を提案している (Akaike (1980), Akaike (1989))。このような考え方の歴史的な起源は古い (たとえば Good(1965)) が、種々の大規模問題に適用して実用性を明らかにした点で赤池学派の貢献は大きい。同様の方法は他の研究者たちによっても独立に採用されている (たとえば Geman and McClure(1987), MacKay(1992))。また、Baum らによって 1960 年代から研究されている hidden Markov chain に関する推定問題 (たとえば、Devijver and Dekesel(1987) 参照) でも、本質的には周辺尤度の最大化が扱われている。これらの方法はそれぞれいろいろな名前で呼ばれているが[9]、本論文では 経験ベイズ法と呼ぶことにする (これが統計学の世界では最も広く通用する名称のようであるが、別の意味で使われる場合もあるらしい)。

経験ベイズ法の有効性は文字通り経験的に確かめられている。一見すると、パラメータを推定するのと同じデータからハイパーパラメータを推定するのは無謀なように思うかもしれないが、実際に多くの例で reasonable な答が与えられる。これらの例の解説は本論文では行なわない (前節から論じている問題については Tanabe and Tanaka (1983) を参照。時系列に関しては Akaike and Kitagawa(1994,1995), Kitagawa(1987), Higuchi et al.(1988) などに、密度推定や空間パターンを含む問題については Sakamoto (1991) (Chap.4-6), Ogata and Katsura (1988) などに応用例が示されている)。 そのかわり、この方法の

[8] ここの議論は、統計物理における熱力学極限の存在及び self-averaging 性に関する議論と同じ種類のものである。

[9] たとえば、 **ABIC** 法、 タイプ **II** 最尤法、 階層ベイズ法、 **evidence framework** などの呼び名がある。また、単に (α,γに関する) 最尤法と呼ばれることもある。

背後にある仕組みについて少し考えてみる。これによって、素朴な意味での制約充足の考えに含まれていなくて、ベイズモデルに含まれているものは何かという問題に光を当てたいと思う。

2.3 経験ベイズ法における “場合の数” の役割

2.3.1 階層と事前分布の規格化

パラメータによって表現されるもの (例では関数形) も一種の知識である。これと比較した場合、ハイパーパラメータによって表現される知識 (例では滑らかさの程度) は、ひとつ階層が上の “メタ知識” といえる。この意味で、前節では、知識とメタ知識の間の階層の統合を考えたことになっている。しかし、見方をかえて、パラメータもハイパーパラメータもベイズの枠組では本来同格のはずだと考えるなら、前節のどこかではじめて階層が出現したというふうにも考えられる。この場合に、どこで階層が現れたかというと、ひとつの答えは、(6) 式と (7) 式の間ということになる。これは、パラメータ $x = \{x_t\}$ が局所的で多数あるのに対し、ハイパーパラメータα, γは大域的でかつ数が少ないということに (あるいは、そういう表現と馴染むようなデータの性質に) 階層の起源を求める考え方である。これはもちろん正しいのであるが、もうひとつ、(6) 式自体に、すでに階層性が作り込まれているという点を認識する必要がある。われわれはまず x に対する事前分布を与え、そのあとでα, γ に対する事前分布をつけ加えたのであるが、その手順そのものがすでに階層的なのである。

はっきりさせるために次のように考えてみる。いま、“制約” を 「“エネルギー” $E_\pi(x;\gamma)$ 及び$\tilde{E}_\pi(\gamma)$ をなるべく小さくする」という形で考えよう。これに “対応” する事前分布としては、$Z^{(1)}$,$Z^{(2)}$をそれぞれの規格化定数として、

$$\pi^{(1)}(x;\gamma) = \frac{\exp(-E_\pi(x;\gamma) - \tilde{E}_\pi(\gamma))}{\sum_{x,\gamma} \exp(-E_\pi(x;\gamma) - \tilde{E}_\pi(\gamma))} = \frac{1}{Z^{(1)}} \exp(-E_\pi(x;\gamma) - \tilde{E}_\pi(\gamma)) \tag{8}$$

と、

$$\begin{aligned} \pi^{(2)}(x;\gamma) &= \frac{\exp(-E_\pi(x;\gamma))}{\sum_x \exp(-E_\pi(x;\gamma))} \; \frac{\exp(-\tilde{E}_\pi(\gamma))}{\sum_\gamma \exp(-\tilde{E}_\pi(\gamma))} \\ &= \frac{1}{Z^{(2)}} \exp(-E_\pi(x;\gamma) - \tilde{E}_\pi(\gamma) - \log \textstyle\sum_x \exp(-E_\pi(x;\gamma)) \end{aligned} \tag{9}$$

の 2 つが考えられる。経験ベイズ法の枠組に導くのは、後者 (9) の方である。

この 2 つは、文章で書けば、

1. x, γを重み $\exp(-E_\pi - \tilde{E}_\pi)$ で同時に選択する。
2. まず、γを重み $\exp(-\tilde{E}_\pi)$ で選択し、次に、選んだγに対して x を重み $\exp(-E_\pi)$ で選択する。

にそれぞれ相当する。この差が重要になるのは、“最初の階層の” 規格化定数
$\sum_x \exp(-E_\pi(x;\gamma))$ がγによって大きく違う場合である。滑らかさに関する事前分布 (1) はその良い例であって、問題の規格化定数の対数は、

$$\log Z_\pi = -\frac{N-2}{2} \log \lambda + const. \tag{10}$$

のようになる[10]。この場合、規格化定数のλ依存性は分布 (1) が生成する可能性のある x の “場合の数” [11] がλによって大きく違うことに由来する。実際、λを小さくするほど、バラエティに富んだ線が生成されるようになる。

[10] 前に述べた無限大の除去の問題を考慮して扱うと係数は $N-2$ になる。
[11] 正確にいえば分布のエントロピー。

以上で説明したような事前分布の規格化の仕方に関わる問題、特に “場合の数” の効果は素朴な制約充足の枠組では論じることが難しい。ここでは、尤度関数に含まれるほうのハイパーパラメータαについては省略したが、こちらについては、尤度関数 $L(y|x)$ のデータ yについての規格化が重要な役割を演じる。これも素朴な形の制約充足の枠組には包含しにくいという点では同じである。

この種の効果はベイズ的なモデル選択の問題でも重要な役割を演じる。たとえば、100 次以下の多項式を “ランダム” に生成するような事前分布を考えるとする。この場合に、次の 2 つの定義は全然違う (ここで、“係数をランダムに選ぶ” というのは、たとえば、十分に分散の大きい正規分布から選ぶことを意味する)。

1. 100 次までの係数を互いに独立にランダムに選ぶ。
2. まず、最大次数 Nを 0 から 100 の間でランダムに選ぶ。
   しかるのち、N次までの係数を互いに独立にランダムに選ぶ。

前者の事前分布から出発すると、最大次数に近いモデルが選ばれる傾向があるのに対し、後者の選び方ではより reasonable な結果が期待できる。当たり前の例に見えるが、BIC 規準 (Schwarz(1978)) や MDL 規準 (Rissanen(1989)) によるモデル選択の議論の基礎には、これに類した階層的な事前分布の設定がある。ただし、これらの手法では、メタ知識の修正に相当する部分は考えないのが普通である。

2.3.2 周辺化の役割

経験ベイズ法に関するもうひとつの論点は、なぜパラメータ x については最大化をせずに和をとるかということである。(6) 式において、x とα, γについて同時に mode をとることにしても良いのではないだろうか。別の表現をすれば、(7) 式の代わりに MAP 推定値 x_{MAP} を代入した

$$l_0(\alpha, \gamma) = \log L_\alpha(y|x_{MAP})\pi_\gamma(x_{MAP}) \tag{11}$$

を最大化しては何故いけないのであろうか。

基本的には、これに対する答えは経験的なものである。少なくとも x が連続量の場合については、同時に mode をとる方法 ((11) を最大化する方法) ではうまくいかないことが知られている。一番簡単な説明は、数個、先の例でいえば 2 個のハイパーパラメータに対して最大化をする代わりに、$2 + K$(たとえば $K = 1000$) 個のパラメータに関して同時に最大化すれば、overfitting が再現されるのは当然ではないかという議論である。これは少々乱暴な議論ではあるが、おそらく正しいと思われる。

いずれにしても、この事実は、ハイパーパラメータの決定 (階層の統合) のためには、パラメータの MAP 推定値だけでなく、そのまわりの “そこそこ良い解” (MAP 推定値のまわりの揺らぎ) の情報が欠かせないことを示している。最適解の良さはいまいちでも、沢山の “そこそこ良い解” のある方が評価される場合もあるのである。ここでも、素朴な意味での制約充足の枠組に含まれない “場合の数” の効果が重要な役を演じている。

理解を深めるために、事前分布と事後分布をギブス分布の形に書き直してみる。x の事前分布を

$$\pi_\gamma(x) = \frac{\exp(-E_\pi(x)}{\sum_x \exp(-E_\pi(x))} \tag{12}$$

と書こう。また、尤度関数の方のハイパーパラメータはしばらく考えないことにして、
$L(y|x) = \exp(-E_L(y|x))$ と書く。このとき、周辺尤度の表式は、

$$l(\gamma) = \log \sum_x \exp(-E_{pos}(x)) - \log \sum_x \exp(-E_\pi(x)) \tag{13}$$

となる。但し、ここで、

$$E_{pos}(x) = E_L(y|x) + E_\pi(x) \tag{14}$$

である。(13) の第 2 項は事前分布の規格化から生じる項で、前節で論じたものである。これに対して、第 1 項での x に関する和は、事後分布に関して “そこそこ良い解” の効果を勘定にいれる役割をしている。両者の差が周辺尤度になっている。周辺尤度を使うことは、統計物理の言葉でいうなら “エネルギー” の代わりに “自由エネルギー” (あるいは “事後自由エネルギー” と “事前自由エネルギー” の差) を用いることに相当することも (13) からわかる (Iba (1989))。

2.4 経験ベイズ法の微分形

経験ベイズ法を微分形に書き直してみる。積分の形に書いた場合、パラメータ x の事後分布がハイパーパラメータの決定に果たす役割はいまひとつはっきりしなかった。それは和$\sum_x$の中に隠されているのであるが、微分形に書き直すことでよりわかりやすくなる。また、微分形に直すことは学習方程式や EM 法との関連を理解するためにも有効である。以下では、事前分布が指数分布族であることを仮定する。すなわち、ハイパーパラメータ$\gamma = \{\gamma_\mu\}$ に共役な十分統計量 $\{T_\mu(x)\}$ が存在して、

$$E_\pi(x) = -\sum_\mu \gamma_\mu T_\mu(x) \tag{15}$$

と書けるとする。

(13) と (14,15) から、周辺尤度のγ_μでの微分を求めるのは、対数関数の微分に注意すれば容易であって、

$$\frac{\partial l}{\partial \gamma_\mu} = \langle T_\mu(x)\rangle_{pos} - \langle T_\mu(x)\rangle_\pi \tag{16}$$

となる。但し、$\langle\quad\rangle_{pos}$ $\langle\quad\rangle_\pi$は、それぞれ、事後分布、事前分布での期待値を表わす。各μについて右辺を零とおくと、

$$\forall\mu \quad \langle T_\mu(x)\rangle_{pos} = \langle T_\mu(x)\rangle_\pi \tag{17}$$

となる。(17) 式は、“事前分布と事後分布 (= 事前分布によるデータの解釈) が、十分統計量 $\{T_\mu\}$ の目で見て同じになる” ということを意味している。経験ベイズ法に対するこの解釈は重要である。

(17) 式を逐次的に解く方法のひとつが、いわゆる **EM アルゴリズム** (Dempster et al.(1977), Devijver and Dekesel(1987)) である。そこでは、

1. ハイパーパラメータ $\{\gamma_\mu\}$ –メタ知識 – を与えて、事後分布での期待値 $\{\langle T_\mu(x)\rangle_{pos}\}$ – 知識 –を構成する。
2. 式 (17) の最初の項に $\{\langle T_\mu(x)\rangle_{pos}\}$ を代入した式をを満たすようにハイパーパラメータ $\{\gamma_\mu\}$ – メタ知識 – を修正する。

を交互に行なうという形で 1章で述べた “循環” が実現されている。この循環は、知識とメタ知識という階層をまたいでそれらを統合するような循環である。また、仮想的な時間τを定義して

$$\frac{d\gamma_\mu}{d\tau} = \langle T_\mu(x)\rangle_{pos} - \langle T_\mu(x)\rangle_\pi \tag{18}$$

とおけば、 ボルツマンマシンの学習方程式 (Hinton and Sejnowski(1986)) の形になる[12]。これは、(16) の第 2 項が複雑な場合に有効である。

学習方程式が (18) の形になるためには 2.3節で述べたことが効いている。2.3.2節で述べた “MAP 推定値のまわりの揺らぎの効果” が無視できれば、第 1 項で事後分布での期待値をとるかわりに MAP 推定値 x_{MAP}を十分統計量の表式に代入したもの $T_\mu(x_{MAP})$ を用いてよいことになる。また、2.3.1節で述べた規格化の問題は学習方程式の第 2 項の存在に関係している。

[12] いわゆるボルツマンマシン (一般の Hopfield Network に対するもの) では、$\{\gamma_\mu\}$ の個数と x の成分の数が (少なくとも) 同じくらい多いので、式は似ていても趣旨はかなり違ってくる。これに対して、マルコフ場についての学習方程式 (Geman and McClure(1987), Iba(1991), Ohtsuki and Kawato(1991)) は経験ベイズ法そのものといってよい。

2.5　脳の情報処理と経験ベイズ法

脳の情報処理、たとえば初期視覚における処理と経験ベイズ法を関係づける考えがある (Kawato and Inui(1990))。また、これに関連して、学習方程式 (18) の右辺の第 2 項と夢 (もしくは他の内部的な生理現象) の関連が論じられている。この観点に立てば、ベイズ統計における事前分布はわれわれの中にある世界のモデルに相当することになる。そこで起きてくる疑問は、事前分布にどの程度自律的な意味を与えうるかということである。この疑問は脳のなかで知識がどのようにコードされているかという問題にも関わってくる。この問題については、事前分布としてマルコフ場を使うことの是非という問題と絡めて、Sec.3で扱うことにする。

もうひとつの問題は、ハイパーパラメータに相当するものの大域的な一様性を保つことに関するものである。ハイパーパラメータが生得的に一様でそのまま不変に保たれているなら、もちろん問題はない。しかし、学習によって変化させるのであれば、なにか一様性を保つしくみが必要になる。より正確にいえば、長期に渡って眺めた多数の画像に対して学習がなされる場合 (神経系の各ユニット当たりのデータ数が十分多い場合) には、一様性を保つしくみは必ずしも必要ない。この場合は入力画像の統計的一様性がハイパーパラメータの一様性をある程度保証してくれる。これに対して、本論文で述べてきたような場合、つまり少数の (基本的には一枚の) 入力画像に対してハイパーパラメータが適応的に変化することを期待する場合は、積極的に一様性を保つ仕組みが必要になる (本来、経験ベイズ法はこちらの状況を想定している)。初期視覚における適応でどちらが求められているのかは不明であるが、一様性を保つための機構としては、長い軸索のニューロン、最近話題になっている脳内における動的な長距離相関、化学物質の拡散などが候補になるかもしれない。

実際のところ、ベイズの枠組における事前分布や事後分布が脳内にそのままの形で実在するとは考えにくい。脳はアクティブエレメントの集まった力学系にどちらかというと似ており、それを使ってベイズ的な計算をするのは無駄が多すぎる。しかし、ここであげた問題は、ある程度枠組を越えた一般性を持つかもしれない。

3　マルコフ場モデルとベイズの枠組

3.1　事前分布は生成モデルとしての意味を持つべきか: マルコフ場の場合

種々のマルコフ場モデル、たとえばイジング模型やその “多色版” であるポッツ模型、さらに線過程 (line-process) を含む模型などが、非ガウス的な画像処理モデルとして注目されている (Geman and Geman(1984), Besag(1986), Marroquin et al.(1987) Dube and Jain(1989), Besag et al.(1991), Possolo(1991), Chellappa and Jain(1993)) [13]。そこで、これらについて経験ベイズ法を適用してハイパーパラメータを決定しようという試みがなされた (たとえば、Geman and McClure(1987), Ogata(1990), Iba(1991) , Ohtsuki and Kawato(1991))。このような試みはそれなりにうまくいくようにもみえたが、いろいろ問題があることもわかってきた。これは、2 次元以上でかつ非ガウス性の強いモデルにおいては、いわゆる協同現象が顕著に見られ、一般次元のガウスモデルや 1 次元の非ガウスモデルとは質的な違いが生じてくることに関連している。

イジング模型 (正方格子上の positive な最隣接相互作用のみを有するイジング模型)

$$\pi(\{x_i\}) = \frac{1}{Z_\pi} \exp(J \sum_{i,j \in C(i)} x_i x_j), \quad x_i \in \{\pm 1\} \tag{19}$$

を事前分布とする場合を考えてみる。ここで、2 値の確率変数 (パラメータ) $\{x_i\}$ は正方格子上に配置さ

[13] マルコフ場による画像処理の文献では、本論文でいうパラメータを画素、ラベル、状態ベクトルなどと呼び、ハイパーパラメータをパラメータと呼んでいる場合がある。

れている。また、$C(i)$ は点 i の正方格子での 4 つの隣接点、Z_πは分布の規格化定数をそれぞれ示す。ハイパーパラメータである結合定数 Jは、隣同士に同じ色の来る条件付確率を制御する。Jが大きい (強結合) の場合には、隣同士に同じ色の来る条件付確率が大きくなる。

$x_i = \pm 1$ をそれぞれ黒と白と呼ぶことにすると、イジング模型を事前分布としてうまく再構成されると主張されている画像のほとんどは、黒と白の大きな塊からなるようなものである。実際、このようなパターンに雑音を入れたものを処理すると、MAP 推定値として黒と白の大きな塊が復元されてくる。ところが、問題なのは、事前分布であるイジング模型が自分自身でこのような大きな塊を生成することは、ほとんどのハイパーパラメータ Jの値についてありそうもないということである。結合定数の小さいイジング模型で生成される典型的なパターンは、白と黒の小さいスケールの塊の混合からなっている。これに対して、結合定数の大きい場合は、ほとんど一様な黒の “大海” の中に白い “小島” がぽつぽつとあるパターンか、その黒白を反転したもののどちらかになる。この 2 つの Jの領域の間にあるのがイジング模型の臨界点 (2 次の相転移点) である。

経験ベイズ法の観点から特にまずいのは、式 (17) に意味がなくなることである。この式は事前分布と事後分布が十分統計量に関して似ていることを主張している。いまの場合の十分統計量は白と黒が隣接している対の個数をあらわす。臨界点より強結合側のイジング模型を事前分布とした場合、これは “小島” のまわりの境界の長さの総和に対応する。これに対して白と黒の大きな塊からなるパターンでは、この量は白と黒の大きな塊の境界の長さに対応している。これらを等しくおくことに意味を見出すのは困難である。このような状況は、イジング模型に限らず、2 次元の非ガウス的マルコフ場ではしばしば見うけられるものである。式 (17) の右辺が事前分布の階層的な規格化に由来することに注意されたい。

以上のような問題については、いくつかの論文で明示的に問題にされている (たとえば、Gray et al. (1994), Dubes and Jain (1989))。しかしながら、この問題を認識していない研究者は現在でも多いように思われる。上記の文献でも指摘されているように、画像処理関係の論文で見かける “マルコフ場の典型的サンプルパターン” に “大きな塊” からなるものが多く見られるのはほとんどがサンプル生成法の誤りが原因である。マルコフ連鎖モンテカルロ法 (ボルツマンマシン, Gibbs sampler) の普通のタイプのもの (local な updation をするもの) を臨界点より強結合側で使用すると、緩和時間が極めて長くなるため、大きな塊からなるパターンが過渡的に出現するのである。

いわゆる線過程を含む問題では事情は異なる、かもしれない。“かもしれない” というのは、線過程を含む処理について事前分布の部分を正しくシミュレートしていると確信できるような研究を知らないからである。学習方程式 (18) をマルコフ連鎖モンテカルロ法を利用して解く場合には、事後分布のシミュレーションで生成された状態を、事前分布のシミュレーションの初期状態に使用することが多い (Ohtsuki and Kawato(1991))。この方法でもっともらしい答えが得られたとしても、定義通りの計算をしているのかどうかは不明である。また、事前分布の部分を独立に (たとえばオフラインで) 生成した研究においても、本当に典型的なサンプルパターンに対して統計量を計算しているかどうかは疑わしい。線過程を含む事前分布についても、事前分布は期待されるようなものではないかもしれない。

以下では、この問題に関して、いくつかの違った見方からのアプローチを提示する。重要な点は、枠組や方法論 (いまの場合、経験ベイズ法など) とそれを用いる対象 (いまの場合、マルコフ場モデルなど) は独立に論じられないということである。研究の発展段階においては、両者の良否や背後にある思想を同時に検討していかなければならない。

3.2 モデルを変更するという考え方

まず考えつくのは、よりもっともらしいパターンを生成できるようにモデルを変更することである。そのためのひとつの方法は、Gray et al.(1994) らが試みているように、相転移のないモデルを作ってそれに限定して考えることである。しかし、もっともらしいパターンを生成するモデルを得るのは簡単ではない

らしい。また、仮にそれができたとして相転移のあるモデルと同等の処理能力が得られる否か、という疑問もある[14]。

別の方向は、特徴点やテンプレートなどを含んだより高次のモデル (たとえば、Grenandar and Keenan (1989), Phillips and Smith (1994)) を考えて、それとマルコフ場モデルを統合して考えることである。つまり、顔なら顔の事前分布によって、顔の特徴点とか大域的な形状を生成し、それに肉づけしたり補間したりして完全な顔を生成するのにマルコフ場モデルを用いるわけである[15]。この方向は、工学としては正統的なように思われるが、高次のモデルの作り方、統合のしかた (特に、場合の数の処理)、高次のモデルの学習など多くの難題が予想される。

3.3 評価規準を変更するという考え方

別の方向の解決策は、経験ベイズ法をやめてしまうことである。経験ベイズ法では、(17) 式に表現されているように生成モデルとして良い事前分布を選ぶことが、良い事後分布を得るもとであると考えた。しかし、本来の目標は、事前分布の良さではなくて、事後分布もしくは MAP 推定値の良さである。したがって、こちらを直接評価しようとするのは合理的なことである。とはいっても、たとえば "MAP 推定値の良さ" をデータへの当てはまりの良さと解釈したのでは、 データを単につなぐのが一番良いことになってしまう。overfitting への逆戻りである。

これを防ぐ方法は、事後分布もしくは MAP 推定値の良さを現在のデータに対する当てはまりで評価するのではなくて、それらに基づいた将来のデータの "予測能力" で評価することである[16]。"予測能力" をうまく定義するのはもちろん困難である。規則がわからなければ将来もわからないはずであるが、逆にここで "予測" を考える目的は規則の構成である。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。

このような循環をうまく定式化しようとする試みとして、将来の予測の良さの解析的な評価による規準 (たとえば、Shibata(1989), Moody(1992), Kitagawa(1993; **PIC**)) や Bootstrap 法による評価による規準 (Kitagawa et al.(1993,1995); **EIC**)) がある。これらは、罰金つき最尤法の場合に AIC(赤池情報量規準) を拡張する試みとも考えられる。関連した古典的な手法としては **Cross Validation** (Wahba(1990)) がある。

マルコフ場についての問題は、マルコフ場を事前分布としてみることの困難に関わっていた。したがって、これらの手法のどれかを採用することにすれば、問題はなくなるように思われる。ところで、これらの規準を採用した場合にも、"ベイズモデル " という立場にこだわることに意味があるのだろうか。もはや、事前分布が単独で生成モデルとしての意味をもたないのであれば、"罰金つき最尤法" あるいは "制約充足" と呼んだ方が良いという考えもありうる。そう言い切ってしまって良いかは別として、経験ベイズ法を採用した場合に比べて、ベイズ的な定式化の役割が減少することは確かだろう[17]。

現段階では、上のどの規準もまだ完全に説得力を持つとはいえない。たとえば、EIC や CV は本来は独立同分布 (i.i.d.) の場合に有効な手法をもとにしているため、関数推定、画像や時系列などの相関の強いデータに対しては多少とも無理があり、そのためいろいろな疑問の余地を残している。また、経験ベイズ法をやめることで外見上なくなったかのように見える "場合の数" の効果が、これらの規準の世界でなんらかの形で残っているかどうかにも興味がもたれる。

[14] それでは、イジング模型やポッツ模型は本当にガウス場などとは別格に有効なのか、有効だとすればどういう状況でか、と問われると、筆者は比較研究をきちんとしていないので、自信を持って答えられない。こちらの方をまず検証する必要があるかも知れない。

[15] このあたりの考えは、東京大学教養学部における川人光男氏の講義 (1994) の際の川人氏との議論を参考にした。

[16] 赤池学派は一貫して予測を本質と見た議論を展開しており、経験ベイズ法の導入にあたってもそうした立場からの説明がなされている (Akaike (1989))。本論文での経験ベイズ法の説明はそれに比べると通常のベイズ統計寄りかもしれない。

[17] 北川の PIC 規準だけは別であって、この規準の導出は事前分布が単独で生成モデルとしての意味を持つことを前提としている。したがって、PIC はベイズ的な定式化なしには意味をもたない。このことは PIC を興味あるものにしているが、反面、いま論じているマルコフ場の問題の解決策にはならないことも示唆している。

3.4 その他の考え方

- 臨界点の利用

  ほとんどすべてのハイパーパラメータ Jの値について、2 次元イジング模型は、小さな塊のまじりあったパターンを作るか、一様な地の上に反対色の小さな塊が散らばったパターンを作るかのどちらかであると述べた。しかし、その境界である臨界点の近傍では – われわれが欲しいものとは違うかもしれないが – 大きなスケールの構造を含む特有のパターンを作る。この領域 (厳密には一点) では、空間相関は距離のべきで減衰する。パターンが特徴的な長さを持たないという点や不安定な領域と安定領域の境目であるという点は、画像処理のために好都合な特徴とも考えられる。したがって、臨界点直上にあるモデルを画像処理のための普遍的な事前分布として用いるという考えはひとつの可能性として考慮に値するかも知れない。

- pseudolikelihood 法の利用

  いくつかの研究 (Qian and Titterington(1989,1993), Besag et al.(1991)) では pseudolikelihood 法 (たとえば Besag(1986) を参照) の考えとマルコフ連鎖モンテカルロ法による事後分布からのサンプル生成を組み合わせて、ハイパーパラメータの推定を行なっている。pseudolikelihood 法を用いることで、事前分布からのサンプル生成はアルゴリズム的には不要になるが、事前分布が事後分布に似ていることを要求していることは経験ベイズ法と変わらない。そうすると、やはり経験ベイズ法と同じ問題を抱えることになりそうである。しかし、pseudolikelihood 法では、パターン全体についての確率分布に対して距離を考えるのではなくて、空間的に局所的な条件つき確率に対して距離 (分布が似ているかどうか) を考える。この点の違いがいまの状況でどういう意味を持つかは考えてみる必要がある。

4 おわりに:モデルの階層性は情報処理にとって不可欠か

階層、とくにあらかじめシステムに作り込まれた階層は、非線形動力学の陣営からは、旧弊な機構として厳しい批判をうけている (Kaneko and Tsuda(1994))。これに対して、いままでの議論は、モデルに階層を作り込むことが、規則を構成するために効果的な方法であることを示している。果たして、階層は情報処理にとって必須のものなのだろうか。

表現あるいはコーディングと確率は表裏一体のものであって、表現を固定することは暗黙の形でシステムに対して “自然” な雑音を決めることになる、という思想はかなり普及している。たとえば、遺伝子のコードを決めることは暗黙のうちに “自然” な突然変異のあり方を決める、というように。カオス力学系でさえ、初期条件あるいは丸め誤差という形で、表現からくる自然な確率というものをとりこんでいる。階層についてはどうであろう。たとえば、“表現とアルゴリズムが区別されていない動的な情報処理” のような理念のもとで研究を進めることで、階層の概念を脱却することが可能であろうか。それとも、(“確率” の場合と同じように)“階層” も、いつまでもどこかに潜んでいて、われわれがそこから脱却したと思った瞬間に再出現するような概念なのだろうか。

本論文でしてきた議論は、本論文の扱っている範囲に限っても、不十分なものである。“予測” の考え方についても、階層が 2 段以上になった場合のこと (これはしばしば計算上の著しい困難を招く可能性がある) についても十分な議論はされていない。また、本論文の範囲自体も、モデルの階層性が情報処理にとって不可欠かどうかを考えるためには狭すぎる。階層に関する議論をさらに深めていくことはやわらかい情報処理の未来を考える上で重要な課題であろう。

謝辞

田辺国士氏をはじめとする統計数理研究所の研究者の方々に直接もしくは文献を通してお教え頂いたことが本論文の骨子をなしています。また、津田一郎氏との長年にわたる議論は問題意識を深める上で大変重要な役割を果たしました。津田氏の “解釈学的循環” (Tsuda(1984)) と統計学における循環の関係については、本論文の執筆を通して考えさせられました。川人光男氏にはマルコフ場に関して、また、脳内のモデルに関しての議論を通していろいろお教えいただきました。2.5節及び 3 章には川人氏との議論を通じて得られた内容が多く含まれています。麻生英樹氏、岡隆一氏及びその他の RWCP PKA 部会の方々との討論も有益でした。これらの方々に深く感謝します。

参考文献

Akaike,H.(1980)
in Bayesian Statistics,
Eds. Bernardo,J.M, DeGroot,M.H., Lindley,D.V., and Smith,A.F.M.,
University press, Valencia.

Akaike,H.(1989)
in ベイズ統計学とその応用, (編) 鈴木雪夫 国友直人, 東大出版会.

Akaike,H. and Kitagawa,G.(編)(1994,1995)
時系列解析の実際 I ,II 朝倉書店.

Besag,J.(1986)
The Journal of the Royal Statistical Society B 48 pp.259-302 (with discussions).

Besag,J., York,J. and Mollié,A (1991)
Annals of Institute of Statistical Mathematics 43 pp.1-59 (with discussions).

Chellappa,R. and Jain,A.(eds.)(1993)
Markov Random Fields: Theory and Application, Academic Press, San Diego.

Dempster,A.P., Laird,N.M., Rubin,D.B. (1977)
The Journal of the Royal Statistical Society B 39 pp.1-38 (with discussions).

Devijver,P.A. and Dekesel,M.M.(1987)
in NATO ASI series F30 Pattern Recognition Theory and Applications,
Eds. P.A.Devijver and J.Kittler, Springer-Verlag.

Dubes,R.C. and Jain,A.K.(1989)
Journal of Applied Statistics 16 No.2 pp.131-164
(この巻全体が統計的画像処理の特集号).

Geman,S. and Geman,D.(1984)
IEEE Transactions on Pattern Analysis and Machine Intelligence 6 pp.721-741.

Geman,S. and McClure,D.E.(1987)
in Proc. of the 46th Session of the ISI, Bulletin of the ISI, Vol.52.

Good,I.J.(1965)
The Estimation of Probabilities, MIT Press, Cambridge, Mass.

Gray,A.J., Kay,J.W and Titterington,D.M.(1994)
IEEE Transactions on Pattern Analysis and Machine Intelligence 16 pp.507-513.

Grenandar,U. and Keenan,D.M. (1989)
Journal of Applied Statistics 16 No.2 pp.207-222.

Higuchi,T., Kita,K. and Ogawa,T.(1988)
Applied Optics 27 pp.4514-4519.

Hinton,G.E. and Sejnowski,T.J.(1986)
in Parallel Distributed Processing Vol.1,
Eds. Rumelhart,E. and McClelland,J.L., MIT press, Cambridge.

Iba,Y.(1989)
in Cooperative Dynamics in Complex Physical Systems,
Ed. Takayama,H., Springer-Verlag, Berlin.

Iba,Y.(1991)
統計数理 Vol.39 No.1 pp.1-21.

Kaneko,K. and Tsuda,I.(1994)
Physica D 75 pp.1-10.

Kawato,M. and Inui,T.(1990)
電子情報通信学会論文誌 J73-D-II, pp.1111-1121.

Kitagawa,G.(1987)
Journal of the American Statistical Association 82 1032-1063.

Kitagawa,G.(1993)
ISM Research Memo. No.481.

Kitagawa,G.,Ishiguro,M.,Sakamoto,Y.(1993)
信学技報 Vol.92, No.503, IT92-133, pp.49-62.

Kitagawa,G.,Ishiguro,M.,Sakamoto,Y.(1995)
ISM Research Memo. No.540.

MacKay,D.J.C.(1992)
Neural Computation 4 pp.415-447; pp.448-472; pp.698-714.

Marroquin,J., Mitter,S. and Poggio,T. (1987)
Journal of the American Statistical Association 82 pp.76-89.

Moody,J.E.(1992)
in Advances in Neural Information Processing Systems 4,
Eds. Moody,J.E., Hanson,S.J. and Lippmann,R.P., Morgan Kaufmann Publishers, San Mateo CA.

Ogata,Y.(1990)
Annals of Institute of Statistical Mathematics 42 pp.403-433.

Ogata,Y. and Katsura,K.(1988)
Annals of Institute of Statistical Mathematics 40 pp.29-39.

Ohtsuki,H. and Kawato,M. (1991)
in Proc. International Joint Conference on Neural Networks, Seattle.

Phillips,D.B. and Smith, A.F.M.(1994)
Journal of the American Statistical Association Vol.89 pp.1151-1163.

Possolo,A.(ed.)(1991)
Spatial Statistics and Imaging, Hayward, Institute of Mathematical Statistics.

Qian,W. and Titterington,D.M.(1989)
Journal of Applied Statistics 16 No.2 pp.267-282.

Qian,W. and Titterington,D.M.(1993)
IEEE Transactions on Pattern Analysis and Machine Intelligence 15 pp.748-752.

Rissanen,J.(1989)
Stochastic Complexity in Statistical Inquiry, World Scientific, Singapore.

Sakamoto,Y.(1991)
Categorical Data Analysis by AIC, Kluwer Academic Publisher
(**坂元慶行 カテゴリカルデータのモデル分析 共立出版** (1985)).

Schwarz,G.(1978)
Annals of Statistics 6 pp.461-464.

Shibata, R.(1989)
in From Data to Model, Ed. J. C. Willems, Springer-Verlag.

Tanabe,K and Tanaka,T.(1983)
月刊 地球 5 pp.179-186.

Tsuda,I.(1984)
Progress of Theretical Physics Supplment No.79 pp.241-259.

Wahba,G.(1990)
Spline Models for Observational Data, SIAM, Philadelphia Pennsylvania.