統計物理と統計的情報処理

大規模・非ガウスモデルをめぐる話題

Version 1.00

1996 年 4 月 18 日

統計数理研究所

伊庭幸人

目次

1 大規模・非ガウスモデルを用いた統計的情報処理 **4**

1.1 ベイズ統計の基礎 . . . 4

1.1.1 事前分布・尤度関数・事後分布 . . . 4

1.1.2 Ignorance Prior と Informative Prior . . . 5

1.1.3 “解” の定義と効用関数 . . . 5

1.2 Informative Prior . . . 7

1.2.1 Informative Prior と大規模モデル . . . 7

1.2.2 Non Gaussian Prior . . . 8

1.2.3 Infomative Prior と視覚研究 . . . 10

1.3 巨視的パラメータの推定: 学習の問題 . . . 12

1.3.1 事前知識の推定，学習 . . . 12

1.3.2 ABIC 法 . . . 12

1.4 大規模統計モデルをめぐる他の話題 . . . 16

1.4.1 多数の離散パラメータを含むモデル . . . 16

1.4.2 コネクショニズム・ Neural Network . . . 17

2 統計物理と統計的情報処理 **19**

2.1 非ガウス・大規模モデルと統計物理 . . . 19

2.2 メトロポリス的モンテカルロ法と微視的パラメータの推定 . . . 21

2.2.1 Simulation without Annealing . . . 21

2.2.2 Simulated Annealing . . . 22

2.2.3 Anneal する場合としない場合の比較 . . . 23

2.3 メトロポリス的モンテカルロ法と巨視的パラメータの推定 . . . 25

2.3.1 不完全データ (教師なし) の場合 . . . 25

2.3.2 完全データ (教師あり) の場合 . . . 26

2.4 他の統計物理の手法の適用 . . . 27

3 付録 A 情報量規準と統計学 (未完) **30**

3.1 モデル選択の規準 . . . 30

3.2 AIC の擬ベイズ的解釈 . . . 30

4 付録 B イジング型神経回路網と対数線型模型 (未完) **30**

この総説は，筆者のオリジナルな研究の部分 (以下本文と呼ぶ) のための準備であると同時に，いままで関係があきらかでなかったり，多くの人に別々のものと考えられてきた各分野の関連を明らかにするという目的をもっている．

本文を理解するための予備知識としては，視覚研究に関する節 (1.2.3節) は不要である．また，コネクショニズムや神経回路網について述べた節 (1.4.2節) もほとんど関係ないので，その目的のためには飛ばして読まれることを希望する．

本文中で統計物理とのアナロジーで用いた用語は，「　」で囲った．ただし，何回も表われる場合は省略した場合もある．

1　大規模・非ガウスモデルを用いた統計的情報処理

第 1 部では，大規模統計モデル，視覚研究，神経回路網などをめぐる最近の発展について広い視野から述べる．

1.1　ベイズ統計の基礎

上記の発展の多くは，ベイズ統計の枠組みで理解するのがもっとも一般的でかつ明快と思われる．そこで，本節ではまずベイズ統計の基礎概念を説明する．

1.1.1　事前分布・尤度関数・事後分布

もともとのベイズの立場では，統計的情報処理とは，あらかじめ持っている知識と新たに得たデータを結合して，どちらか一方では得られない効用を得ることと定義づけられる．さらに立場によっては，事前知識をどう調節したらよいか，あるいはその優劣をどうして比較したらよいかという問題がこれにつけ加わる．事前知識の調節や比較の問題は 1.3節で扱うことにして，ここでは事前知識とデータとの結合の問題およびその前提としての事前知識の表現の問題を考えよう．

ベイズ統計では，すべての知識を推定すべきパラメータに関する確率分布の形に表わす．この点が “統計” である由縁であって，一般の人工知能などとことなる点である．これは「熱平衡系」と一般の力学系・確率過程の違いに対応するとも考えられる．具体的には，以下のように考える．

まず，データを知らないときのパラメータ $\{x\}$ の確率分布 $\pi(\{x\})$ を定める ($\{\ \ \}$ を付けたのは複数，ときには多数のパラメータがあることを示すためである)．$\pi(\{x\})$ を事前分布 (**prior distribution**) という．この背後には，モデルのアンサンブルがあって，各モデルがパラメータ $\{x\}$ で標識されており，われわれが相手にしている系はそこから抽出されたという発想がある．$\{x\}$ はこの意味で確率変数とみなされるわけである．

つぎに，データとパラメータの関係を記述しなくてはならない．これは，パラメータが決まったときにさまざまなデータを得る確率 (順問題の確率) を与えることで決まる．これを $\mathrm{L}(\{y\} \mid \{x\})$ と書き，尤度関数 (**likelihood function**) と呼ぶ．これは $\{y\}$ に関しては確率分布であるが，$\{x\}$ についてはそうではない (規格化されていない)．

以上で，問題に対する事前の知識を事前分布と尤度関数という形で表現することができたので，こんどはデータと事前知識の結合について考えよう．これにはベイズの公式 (**Bayes formula**) を用いる．ベイズの公式とは，$\{y\}$ が与えられたときのパラメータ $\{x\}$ の条件付分布 $p\{x\} \mid \{y\})$ が，

$$p(\{x\} \mid \{y\}) = \frac{\mathrm{L}(\{y\} \mid \{x\})\pi(\{x\})}{\sum_x \mathrm{L}(\{y\} \mid \{x\})\pi(\{x\})} \tag{1}$$

で与えられるというものである．この式の導出は簡単であるが，重要なことなのできちんと説明する．

$\{x\}$ と $\{y\}$ がともに起こる確率を $f(\{x\},\{y\})$ とすると，

$$f(\{x\},\{y\}) = \mathrm{L}(\{y\} \mid \{x\})\pi(\{x\}) = p(\{x\} \mid \{y\})\Lambda(\{y\}) \tag{2}$$

と 2 通りに書ける．ここで，$\Lambda(\{y\})$ は ($\{x\}$ について何もわからない場合に) y が起きると期待される確率である．

$p(\{x\} \mid \{y\})$ は $\{x\}$ に関する確率分布であるから，$\sum_x p(\{x\} \mid \{y\}) = 1$ である．これを使うと，(2) 式から，

$$\sum_x \mathrm{L}(\{y\} \mid \{x\})\pi(\{x\}) = \Lambda(\{y\}) \tag{3}$$

となる．(2) 式の両辺を $\Lambda(\{y\})$ で割ったものにに (3) 式の左辺を代入すると求める式 (1) が得られる．

この状況を図解したのが図 1 である．上の図で黒丸の $\{y\}$ が得られたとして関係のある部分のみを取り出したのが下の図である．ベイズの公式は，特定の $\{x\}$ を経て $\{y\}$ に行く経路の相対確率が $\mathrm{L}(\{y\} \mid \{x\})\pi(\{x\})$ であることから導かれる．

さて，ベイズの公式を用いると具体的なデータ $\{y\} = \{y^d\}$ が得られたときの $\{x\}$ の分布は，

$$P(\{x\}) = \frac{\mathrm{L}(\{y^d\} \mid \{x\})\pi(\{x\})}{\sum_x \mathrm{L}(\{y^d\} \mid \{x\})\pi(\{x\})} \tag{4}$$

と求められる．$P\{x\}$ を事後分布 (**posterior distribution**) と呼ぶ．ベイズ統計の立場では得られる情報はすべて事後分布のなかに含まれることになるわけである．

なお，以下の記述ではしばしば y に当たるものと y^d に当たるものを記号として区別しないで使う (特に例の中)．

1.1.2 Ignorance Prior と Informative Prior

実際の例をベイズ的な形式で論じてみよう．ここでは，“最小 2 乗法” による直線 $y = at + b$ の当てはめというもっとも簡単な例を考える．この場合，尤度関数はデータ $\{y_i, (i = 1..m)\}$ についてのガウス分布

$$\mathrm{L}(\{y_i\} \mid \{a, b\}) = \frac{1}{(2\pi\sigma^2)^{m/2}} \exp(-\sum_i \frac{(y_i - at_i - b)^2}{2\sigma^2}) \tag{5}$$

となる (ここでは，データを取った点 $\{t_i\}$ と分散 σ は既知の定数として扱った)．

パラメータ $\{a, b\}$ に関する事前分布としては何をとったらよいだろうか．ひとつの考え方は，まったく何も知らないのだから，$(-\infty, +\infty)$ のどの点も同じぐらい確からしいと仮定することである．このような決め方を **ignorance prior** (無知を表わす事前分布) という．

このもとで，ベイズの公式から事後分布を計算すると，

$$P(\{a, b\}) = \frac{1}{Z_{pos}} \exp(-\sum_i \frac{(y_i - at_i - b)^2}{2\sigma^2}) \tag{6}$$

となる．Z_{pos} は規格化因子で，今の場合 (もしそうしたければ) 解析的に計算できる．

上記のような “無知” の表現には問題がないわけではなく，

1. 上の “事前分布” は規格化不能であり，実は $\{a, b\}$ についての確率分布になっていない
2. 見かけが一様なら本当に無知を表わしているといえるのか．とくに座標変換不変性をどう考えるか．

などの問題が生じうる．

規格化不能な事前分布 (**improper prior**) をどう考えるかということは議論の分かれる問題であるが，事後分布といっしょになって意味をもてば十分であるという見方もある．上の例でいえば，はじめから $(-\infty, +\infty)$ とせずに，$[-L, L]$ の一様分布から出発して，事後分布を作ってから $L \to \infty$ としたと思うわけである．

座標変換不変性の問題も非常に重要な問題であるが，本論文の趣旨とはあまり関係ないので，ここではこれ以上論じないことにする．

事前分布として ignorance prior をとる代わりに，積極的な事前知識を導入するという考えもありうる．これを一般に **informative prior** と呼ぶ．直線を当てはめる場合でいえば，あらかじめ $b \leq 0$ であることが知られている場合には，b の事前分布にそれを取り入れる (たとえば “$[0, +\infty)$ に一様に分布する” とする) のが自然であるが，これは最も簡単な informative prior の例になっている．

最近，画像解析，時系列解析等において，informative prior を組織的に利用することによって大きな成果があがっているが，これについては 1.2節でくわしく論ずることにする．

1.1.3 “解” の定義と効用関数

いままで，問題のベイズ的な意味の解とは事後分布そのものだと考えてきた．前節であげた例の場合などは，パラメータが 2 個しかないのでこれでもかまわないが，一般にはパラメータの数の非常に多い場合もあり，その場合，事後分布自体を観察することは不可能である．そこで，なんらかの方法で事後分布から必要な情報 (“解”) を取り出すことが必要になってくる．

いちばん簡単なのは事後分布を最大にするパラメータ (最頻値，**mode**) を解とすることで，この定義による解を **MAP** 解 (最大事後確率解，**Maximum A Posteriori estimate**) という．さきの例に適用すると，式 (6) の左辺を最大にする $\{a, b\}$，すなわち

$$l_2 = \sum_i (y_i - at_i - b)^2 \tag{7}$$

を最小にする $\{a, b\}$ が MAP 解ということになるが，これはいわゆる最小 2 乗法にほかならない．

別の意味の解も考えられる．たとえば，a, b それぞれの事後分布による期待値

$$\langle a \rangle_{pos} = \int a P(\{a, b\}) da db \tag{8}$$

$$\langle b \rangle_{pos} = \int b P(\{a, b\}) da db \tag{9}$$

を解と定義することもできる．実は，式 (6) で定義される事後分布の場合，期待値としての解は MAP 解に一致してしまう．しかし，これは事後分布がガウス分布の場合の特殊性であって，一般にはなりたたない．たとえば，事前分布に $b \leq 0$ を付け加えただけで，すでに両者は一致しなくなる (図 2 参照)．

いろいろな “解” のうちどれを選ぶかは何が目的であるかによる．これは，求められた事後分布のもとで，指定した量を最大 (最小) にするという形に定式化できる．指定する量のことを効用関数 (utility function) あるいは損失関数 (loss function) と呼ぶ．たとえば，期待値は正解からの誤差の 2 乗の事後分布による期待値を最小にする解である．また，MAP 解は正解のまわりの微小領域に入る可能性がもっとも大きい解である．

MAP 解の効用関数の説明がやや回りくどい言い方になっているのは連続パラメータを念頭に置いているからで，離散パラメータの場合は端的に “正解と一致する確率が最大になる解” となる．これは，パラメータが沢山ある場合は “すべてのパラメータの一致を要求し，ひとつでも外れた場合は “0 点” にする” ことに相当している．

また，これらの解の 誤差や的中確率 なども事後分布から抽出しうる情報である．

1.2 Informative Prior

本節では informative prior の利用について述べる．まずはじめに informative prior の導入によって柔軟なモデル構成が可能になることを説明し，次に非ガウス的なマルコフ場モデルによる画像処理について詳しく論じる．視覚研究との関連についても触れる．

1.2.1 Informative Prior と大規模モデル

常識的にいえば，統計モデルにおいてパラメータの数はデータの数よりはるかに少なくなければならない．パラメータの数が多ければ多いほどモデルの当てはまりは良くなるが，雑音を除いてデータに共通の特性を引き出すという統計的情報処理本来の意義が失われてしまう (**overfit**)．この事情を，予測の最良性という観点から定式化したのが，**AIC(Akaike's Information Criterion)** などの情報量規準であり，データの最短記述という見方から定式化したのが **MDL(Minimum Description Length)** である．これらが出現する以前はいわゆる有意性検定がそれらの代役を果していた．

infomative prior を導入すると，上の事情は違ってくる．適当な infomative prior のもとでは，パラメータの数がデータ数と同程度かそれ以上あっても，意味のある情報処理を行うことができる．これは，パラメータ同士が独立でなくなるからである．

例として，いわゆる平滑化 (**smoothing**) のうち最も簡単な場合を論じよう (図 3,[K.Tanabe, T.Tanaka (1983)])．直線上の M 個の点 $\{t_k, k = 1..M\}$ でデータ $\{y_k\}$ が与えられているとする．このとき，“もとの曲線” をどう推定するかというのが問題である．データの誤差の分布はガウス分布 $N(0, \sigma^2)$ に従うと仮定しよう (ここでは，σ^2 は既知とする)．

ここで考えなければならないのは，“曲線” とはなにか，ということである．ひとつの考え方は，“曲線とは多項式で表わされるものだ” というふうに，少ないパラメータで特徴付けられる範囲に対象を限定することである．この場合，データ数 M が 100 のときに 100 次式を当てはめるのはあきらかなナンセンスなので，AIC 最小化規準などによって次数を決定しなくてはならない．多項式と 3 角級数のどちらがよいかなどということも AIC で決められるから，思い付いた “曲線” の族をすべてためして見て，最もよいものを選べばよいことになる．

上の考え方で問題なのは，どうしても曲線族の “くせ” が結果に反映されてしまうことである．重要なのは，“曲線” の定義の是非ではなく，結果としての予測能力 (AIC ではかられるもの) だというのはもっともであるが，“曲線” という直観をもっと自然な形で表現できれば，予測能力の意味でもよいのではないかと思われる．

そこで，曲線の微小部分を表わす沢山のパラメータ $\{x_i, i = 1..N\}$ を用意して，それらに関する informative prior の形で “曲線らしさ” を表現する．

まず考えられるのは，曲線の微小部分を「原子」と見たときにそれらが「ばね」でつながっているというモデルである (図 4)．

$$\pi_\lambda(\{x_i\}) = \frac{1}{Z_\lambda} \exp(-\lambda \sum_i (x_{i+1} - x_i)^2) \tag{10}$$

(Z_λ は λ のみに依存する定数) この事前分布は，対象が “急に変動しない” ということをいっているので，それなりに有用だが，“曲線らしさ” をあらわすとはいえない．たとえば，直線でも傾斜が急なものは良くない (起こりにくい) ことになってしまう．

曲線をあらわす事前分布として，よりふさわしいのは，もうひとつ高階の差分を用いた

$$\pi_\lambda(\{x_i\}) = \frac{1}{Z_\lambda} \exp(-\lambda \sum_i (x_{i+1} - 2x_i + x_{i-1})^2) \tag{11}$$

である．ただし Z_λ は λ のみに依存する定数．この事前分布は対象が “なめらか” であることを表わしており，3 次スプラインのベイズ統計版ともいえるものである．物理的には (10) の弦にたいして，(11) は剛体の細い棒をあらわしている．

x_i の総数 N は一般にデータの数 M と同じか，大きくとってよい．データをとった点 t_k の “上” にある曲線の微小部分が $x_{\eta(k)}$ であるとすると，尤度関数はガウス雑音の仮定のもとで，

$$Ł_\sigma(\{y_k\} \mid \{x_i\}) = \frac{1}{Z_\sigma} \exp(-\sum_k \frac{(y_k - x_{\eta(k)})^2}{\sigma^2}) \tag{12}$$

となる．これから，MAP 解は

$$\frac{1}{\sigma^2} \sum_k (y_k - x_{\eta(k)})^2 + \lambda \sum_i (x_{i+1} - 2x_i + x_{i-1})^2 \tag{13}$$

を最小にする x_i として求められる．これは，非ベイズ的な用語では罰金付き最尤法と呼ばれるものである．

以上では境界条件については省略したが，実際にはちゃんと考えなければいけない．境界条件によっては π_λ は improper になるが， proper な事前分布の近似と考えられるので問題ない．このあたりが面倒なので，規格化定数を Z_λ, Z_σ などと書いたが，実際にはこれらは簡単に計算できる．

いまのような方法は非常に広い範囲の問題に使用できる．拡張の方向としては，

1. ほかの事前知識，たとえば “１年周期の変動がある”，“正値である”，“凸である” などを取り入れる．
2. ２次元以上の問題 (とくに画像問題) を考える．
3. 非ガウス的な事前分布を用いる場合を考える．

などが考えられるが，本稿の趣旨からすると，３番目 + ２番目がとくに重要である．これについては，画像問題における非ガウス的事前分布の利用を中心にして，次節 (1.2.2節) で詳しく扱うことにする．

以上において，λ や σ のような事前の知識にかかわる量の推定 (学習) をどうするかということが当然問題になるが，これは 1.3節で考える．なお，λ や σ のような量もある意味でパラメータであるが，本稿ではこれらを巨視的パラメータ (マクロなパラメータ， **macroscopic parameters**) と呼び，いままでパラメータと呼んでいたものを (区別する必要のある場合)，微視的パラメータ (ミクロなパラメータ， **microscopic parameters**) と呼ぶ．いろいろな研究グループの用語については，まだ説明していないものも含めて表 1 に示した．

1.2.2 Non Gaussian Prior

時系列や画像の問題で非ガウス的な事前分布を導入する最大の動機は，変化点や境界の検出である．また， 外れ値 (**outlier**) に対しての頑健性という要求もある．

いま格子上の微視的パラメータ $\{x_i\}$ からなる画像モデルの場合を考えると，このような事前分布の例としては，

$$\pi(\{x_i\}) = \frac{1}{Z_\pi} \exp(-\frac{1}{2} \sum_{j \in N(i),i} E_{loc}(x_i, x_j)) \tag{14}$$

とおいたとき，

- 分布の対数 (「エネルギー」) が Caucy 型

$$E_{loc}(x_i, x_j) = -\frac{\lambda}{(x_i - x_j)^2 + \xi^2} \tag{15}$$

- 分布が Caucy 型 ($\lambda = 1$) または Pearson system($0.5 < \lambda \leq \infty$)

$$E_{loc}(x_i, x_j) = \lambda \log \{(x_i - x_j)^2 + \xi^2\} \tag{16}$$

- しきい値型

$$E_{loc}(x_i, x_j) = \lambda (x_i - x_j)^2 \quad \text{if} \mid x_i - x_j \mid < \xi \tag{17}$$

$$E_{loc}(x_i, x_j) = \lambda \xi^2 \qquad \text{else} \tag{18}$$

などが考えられる (図 5)．ここで，$N(i)$ は格子上の点 i の近傍 (正方格子なら４個ないし８個) をあらわし，ξ, λ は巨視的パラメータ (ここでは定数扱い) である．Z_π は規格化定数であるが，こんどは簡単に求められない．上記は１階の差分が E_{loc} に含まれている場合であるが，前節に述べたように問題によっては２階の差分を用いたほうがよい場合もある．

これらの事前分布の狙いは，ξ 程度より大きい変位を与えるとばねが切れたり，力が出なくなるとすることによって，はっきりした境界だけを検出し，外れ値のまわりに及ぼす影響を軽減するということにある．(１階の差分の代わりに２階の差分を用いた場合は，１次元では棒，２次元では薄板が曲がりすぎると折れる・割れるというイメージになる)．時系列モデル (１次元) の場合に，(16) 式に対応する事前分布を使用した場合の結果 (変化点の検出) を図 6 に示した [G.Kitagawa (1987)]．

非ガウス的な事前分布へのもうひとつの道は，微視的パラメータが離散的な値をとる問題である．もとから画像が２値画像 (一般に q 値画像) である場合を想定してもよいし，なんらかのラベルを画像にはりつける場合を考えてもよい．

2 値画像の場合，2 値について対称な尤度関数は

$$\mathrm{L}(\{y_i\} \mid \{x_i\}) = \frac{\exp(-h \sum_i y_i x_i)}{(2\cosh(h))^M} \tag{19}$$

とかける．ここで，$y_i = \pm 1$，$x_i = \pm 1$ とした．M は微視的パラメータの数であり，h は雑音 ($+1 \leftrightarrow -1$ が起こる確率)p と

$$h = \frac{1}{2}\log\frac{1-p}{p} \tag{20}$$

で結ばれている．q 値画像の場合やラベル貼りの場合は，ありうる誤りの性質と確率によっていろいろな尤度関数が可能である．

このような場合の $\{x_i\}$ の事前分布としてもっとも簡単なものは，“似たものが集まりやすい” とした場合で，

$$\pi(\{x_i\}) = \frac{1}{Z_\pi}\exp(-\frac{1}{2}K \sum_{j\in N(i),i} \delta(x_i - x_j)) \tag{21}$$

とかける (δ はクロネッカーのデルタ)．これは統計物理における q 状態ポッツ模型になる．とくに 2 値 ($q = 2$) のときは有名なイジング模型である．これらのモデルは離散パラメータのモデルとしては最も簡単なものであるが，強い非ガウス性をもっている (このことは，連続パラメータのモデルの極限と考えれば容易に理解できる)．

ここで取り上げた事前分布は連続パラメータのものも離散パラメータのものも “各格子点の変数がとる値の確率分布がその点の近傍の変数を固定すると決ってしまう” という特徴を持っている．このような確率場をマルコフ場という．

非ガウス的なマルコフ場を利用した画像処理で大きな注目を集めたのは Geman 兄弟の研究である．Geman & Geman の研究は次の 2 点に特色がある．

- 補助的な微視的パラメータ (「ダミーのスピン」) を導入することによって複雑な事前情報の表現を可能にした．
- これらの複雑なモデルの MAP 解を **simulated annealing** 法によって求めた．

2 番目の点は本文と重要な関係があり，第 2 部でくわしく論じられる．1 番目の方は本文には直接関係ないが，実際のモデル作りをする上では重要なので，ここで簡単に説明する．

微視的パラメータの間に複雑な相互作用を入れれば，理論上はいくらでも複雑な事前知識が表わせるはずである．しかし実際には，“物体の縁は直線に近い” といった比較的簡単なものでも式の形で表わすのは面倒である．このような場合，補助パラメータを利用すると便利である．“物体の縁は直線に近い” の場合，被覆格子 (図 7) 上の補助パラメータ $\{l_{ij}\}$ を考え，事前分布を次のように書き下す (線過程 (line process) の導入)．

$$\pi(\{x_i\}) = \frac{1}{Z_\pi}\exp(-R(\{l_{ij}\}) - \frac{1}{2}K \sum_{j\in N(i),i} E_x(x_i, x_j, l_{ij})) \tag{22}$$

l_{ij} は i と j を結ぶ線上にある補助パラメータで，$\{0, 1\}$ の 2 値をとり，これが 1 のときは，i と j の間は “切れている” と見なされる．すなわち，

$$E_x(x_i, x_j, l_{ij}) = \delta(x_i - x_j)\delta(l_{ij} - 1) \tag{23}$$

このようにしておいて，補助パラメータ $\{l_{ij}\}$ の間に R で表わされる相互作用を導入する．R はやはり局所的な関数の和であって，たとえば図 8 のように定める．単純なポッツ模型 (21) による復元結果及び線過程を含む事前分布 (22,23) による復元結果を図 9 に示す．

Geman & Geman らは，さらに複雑な事前分布を用いて，境界の検出，テクスチャー (肌理，**texture**) の判別などを試みている．いちばん見栄えのする例のひとつを図 10 に示す．この例には複雑な事前分布が用いられている．

本節の内容に関する文献として，以下のものをあげておく．画像処理にマルコフ場を用いるという考えは，２０年近い歴史を持つらしいが，古い参考文献については [S.Geman, D.Geman (1984)] を参照されたい．

- 2 次元画像の処理
  - S.Geman, D.Geman (1984)
  - D.Geman, S.Geman (1986)
  - D.Geman, S.Geman, C.Graffigne (1987)
  -

- H.Derin, H.Elliott, R.Cristi, D.Geman (1984)
- P.Devijiver, M.Dekesel (1986)
- J.Besag (1986)
- D.Terzopoulos (1986)
- C.Koch, J.Marroquin, A.Yuille (1986)
- J.Marroquin, S.Mitter, T.Poggio (1987)
- B.Gidas (1989)
- Y.Ogata (preprint)

• SPECT(トモグラフィの一種) の画像処理

トモグラフィの問題では，尤度関数によって表わされる微視的パラメータとデータの関係が非局所的になることに特徴がある．

- S.Geman, D.E.McClure (1985)
- S.Geman, D.E.McClure (1987)

• １次元的な問題・時系列

微視的パラメータの配置が１次元的な問題，とくに時系列の場合 infomative prior を用いるベイズ的な扱いは，カルマンフィルターの非線形版に近くなる．この場合，“状態空間” が微視的パラメータの空間に相当するが，１次元性を生かして，統計物理の転送積分法 (転送行列法) に対応する “端から順に” 計算するアルゴリズムが使える．
(17,18) のような連続型の事前分布に関する研究としては，統計数理研究所の G.Kitagawa の研究がある．
離散モデルに対応するものは，隠れマルコフモデル (hidden Markov model) とよばれ，L.E.Baum らによって 1970 年代に導入された．これについては，[P.Devijiver, M.Dekesel (1986)] を参照されたい．

- G.Kitagawa (1987)
- P.Devijiver, M.Dekesel (1986)

以上はあくまで大体の傾向を示すもので，完全な文献リストではない．これらの文献の内容については，この総説および本文の該当箇所で必要に応じて論じることにする．

1.2.3 Infomative Prior と視覚研究

この主題は直接本文に関係ないが，非常に興味深い話題なので，簡単に触れることにする．

D.Marr の **‘Vision’**(1982) は近年の視覚研究において，もっとも有名な著作である．この著作の重要性は，それまでの視覚研究の結果を統一的な原理のもとで一連の処理の流れとしてまとめたことにある．D.Marr は，視覚研究とは人間の内部構造の研究である以上に外部世界の研究であり，われわれの普通見る世界の像がどのような拘束条件に従うかの研究である，ということを強調している．(これは，ある意味で哲学者カントの指摘と同じであり，ただ方向が反対なだけである)．

D.Marr の死後，その仕事を引き継いだ MIT のグループは，Geman & Geman の研究に示唆されて，D.Marr の “拘束条件” を infomative prior とよみかえることにより，初期視覚のさまざまな問題がベイズ統計の言葉で定式化できることを示した ([J.Marroquin, S.Mitter, T.Poggio (1987)])．このことは，しばしば “初期視覚の問題は不適切 (ill-posed) 逆問題である” という標語で示される．不適切 (**ill-posed**) という言葉は，データに対して微視的パラメータが多すぎるために informative prior を導入しなければ推定が困難になるような状況を表わしている．

初期視覚の問題の簡単な例として，明度 (表面反射率) の問題をとりあげよう．これは，D.Marr の本にとりあげられている題材のひとつである (オリジナルは [E.H.Land, J.J.McCann (1971), Horn(1977)])．

図 11 のような図形 (モンドリアン図形) を考える．２つの矢印のさす部分の濃さを比較すると，一見右が濃いように思われるが，実は両者の明度は同じである．この “錯覚” はわれわれがなめらかな明度変化 (“トレンド”) を照明の効果として分離し，差し引いていることに由来する．その一方で，急激な明度の変化は縁として見える．この実験は，われわれが，(15),(16),(17,18) のような境界を検出しつつ平滑化を行うモデルを持っていて，それを用いた計算を意識せずに行っているとすれば説明できる．

色つきの図形の場合は同じことを３原色に分けて適用すればよい．物体の表面の認識，立体視，運動の検出といった問題についても informative prior を用いた定式化ができる．たとえば，物体の表面を認識することは，点の集まりを不連続や角を保持しつつなめらかな曲面で補間する問題に帰着される．

この話題に関連して，いわゆる “Marr の３水準” に関する筆者の意見を述べておく．D.Marr は複雑な情報処理系を理解するための３つの水準として，

- 計算理論
- 表現とアルゴリズム
- ハードウェアによる表現

の 3 つをあげ，各水準のどれを考察しているのかをつねに意識することの重要性を説いている．

統計的情報処理の本質はすべての知識を確率分布でかきあらわすこと (統計力学との類似でいえば熱平衡理論の範囲に留まること) にあると考えられるが，これは，アルゴリズムとモデルが区別できるということと本質的な関係がある．これから考えると D.Marr による 3 水準の分離は複雑な情報処理系を理解するための枠組みとして統計的情報処理の枠組みを採用することを示唆しているともみられる．

実際，上の 3 つの水準は，統計的情報処理における，

- 統計モデルの作成
- モデルを当てはめるための数値計算法
- 計算機の選択とコーディング

にほぼ対応する (厳密にいえば，統計モデルの作成には狭義の計算理論だけでなく表現の選択の問題も含まれるが)．

また， D.Marr が計算理論を重視したのは， H.Akaike や J.Rissanen が統計モデルを明示することを重視したことに対応するとも考えられる．

実際には， D.Marr の主張は，はっきりとした物理的拘束条件に基づくモデルが望ましい，といったモデルを立てるための指針にまで及んでおり，上の要約はもちろんその一面を示したにすぎないが，その主張の利点とおそらくは限界を理解するのに役立つと思う．

1.3 巨視的パラメータの推定: 学習の問題

ここでは 1.1節， 1.2節で先送りにした問題，すなわち事前分布に含まれる巨視的パラメータを経験から決定するにはどうしたらよいかという問題を考察する．ここでは一般論を述べるが，具体例については第 2 部の 2.3節及び本文の例を参照されたい．

1.3.1 事前知識の推定，学習

まず，何組かの似たようなデータが得られている場合を考える．この場合，その一部について “正解” がわかっていれば，問題は簡単である．たとえば，

$$\pi_\lambda(\{x_i\}) = \frac{1}{Z_\lambda}\exp(-\lambda\sum_i (x_{i+1}-2x_i+x_{i-1})^2) \tag{24}$$

$$\text{Ł}_\sigma(\{y_k\}\mid\{x_i\}) = \frac{1}{Z_\sigma}\exp(-\sum_k \frac{(y_k-x_{\eta(k)})^2}{\sigma^2}) \tag{25}$$

の場合に，$\{y_i\}$ でなく $\{x_i\}$ が知られている例 (“教師” ，“完全データ”) がいくつかあれば，(25) 式は不要であり，(24) 式から λ を推定するのは単なる最尤法 (ベイズ的にいえば “π を λ についての尤度関数と思い，λ に ignorant prior を仮定して MAP 解を求める” こと) になる．今の場合，曲線としてどの程度くねくねしているものを許すのかを例によって教えたことになる．これは，“完全データの場合”あるいは “教師ありの学習”とよばれる．この場合も事前分布 π のが非ガウスの場合は，規格化定数 Z_λ を求めるのは困難であり，メトロポリス的モンテカルロ法や統計物理の手法が役立ちうる (第 2 部の 2.3節参照)．

教師ありの学習の場合，教師として使う “似たもの” の範囲をどう決めるかという問題がある．さらに世の中には正解の手に入れようのない問題も多い．そこで，雑音を含んだデータから巨視的パラメータ (上記の場合 λ と σ) を推定することが問題になる．これを “不完全データの場合”あるいは “教師なしの学習”と呼ぶ．これを行うためには，単純な最尤法とは別の手法 (原理) が必要である．次節で述べる ABIC 法はこのような手法のひとつである．

教師なしの場合も，学習ということの基本からすれば，似たようなパターンがいくつかある場合を考えるべきである．しかし実際には，十分大きい画像や時系列は互いにほとんど相関のない多くの部分からなると考えられるので，ひとつのデータから巨視的パラメータを推定できることも多い (以下ではむしろそちらを主体に考える)．この場合，同じデータから巨視的パラメータと微視的パラメータの両者を推定することになる．

なお，ベイズ推定と学習についての別の考え方として，ベイズの公式自体が学習をああらわしているというものがある．データがない場合は，事前の先入観にもとづいて判断をするしかないが，しだいにデータが増えていくと事前分布の役割が小さくなる．この過程を学習と考えるわけである．これは，1 つの場所を写した写真が何枚もあるとき，枚数が増えるにしたがって，知識が正確になっていく場合にあたる．

これに対して，われわれのいう学習 (巨視的パラメータの学習) は，同じ地方の写真がたくさんあるとき，そこからその地方の風景の特徴を掴んで，個々の写真の解釈に役立てるというような場合である．この場合，写真が 1 枚しかなくてもある程度大きければ，風景の特徴の学習は可能と思われる．

1.3.2 ABIC 法

巨視的パラメータの学習には，いくつかの方法が提案されているが，ここではそのうちの最も有力で一般性のある方法である **ABIC** 法を紹介する．この方法は第 2 部で述べる統計物理との類似という点からも興味深い．

いま，事前分布 $\pi(\{x\})$ が巨視的パラメータ λ を含み，尤度関数 Ł が巨視的パラメータ μ を含んでいるとする (ともに複数個あってよい)．

データ $\{y^d\}$ が与えられたとき， ABIC 法では， λ,μ を

$$-\frac{1}{2}ABIC(\lambda,\mu) = \int \text{Ł}_\mu(\{y^d\}\mid\{x\})\pi_\lambda(\{x\})d\{x\} \tag{26}$$

が最大 (ABIC が最小) になるように決める．ここで $\int d\{x\}$ は微視的パラメータ全部に関する多重積分 (多重和) を表わす．

この意味は，事前分布 $\pi_\lambda(\{x\})$ にしたがって微視的パラメータを生成したときの尤度 $\text{Ł}_\mu(\{y^d\}\mid\{x\})$ の期待値を最大にするということである．

事前分布 π が非ガウスの場合には，規格化定数が簡単にはもとまらないことが多い．そこでこれを明示し，統計物理との類似を見やすくするように書きかえると

$$\pi_\lambda(\{x\}) = \frac{\exp(E_\pi^\lambda(\{x\}))}{\int \exp(E_\pi^\lambda(\{x\}))d\{x\}} \tag{27}$$

となる．また，尤度関数 L を同様に書きかえると，

$$\mathrm{L}_\mu(\{y^d\} \mid \{x\}) = \frac{1}{Z_L^\mu(\{x\})} \exp(E_L^\mu(\{y^d\}, \{x\})) \tag{28}$$

と書ける．

このとき， ABIC は

$$-\frac{1}{2}ABIC(\lambda,\mu) = \log \int \exp(E_L^\mu(\{y^d\},\{x\}) + E_\pi^\lambda(\{x\}) - \log Z_L^\mu(\{x\}))d\{x\} - \log \int \exp(E_\pi^\lambda(\{x\}))d\{x\} \tag{29}$$

となる．

さらに，$\log Z_L^\mu\{x\}$ が $\{x\}$ に依存しなければこの部分を外に出して，

$$-\frac{1}{2}ABIC(\lambda,\mu) = \log \int \exp(E_L^\mu(\{y^d\},\{x\}) + E_\pi^\lambda(\{x\}))d\{x\} - \log \int \exp(E_\pi^\lambda(\{x\}))d\{x\} - \log Z_L^\mu \tag{30}$$

としてよい．

この形でみると，統計力学でいう「自由エネルギー」(正確にいうと自由エネルギーの差) との類似は明らかである．

ABIC の微分を 0 とおくことで， ABIC 最小解の満たすべき方程式を求めることができる．

$$\frac{dE_\pi^\lambda(\{x\}))}{d\lambda} = A^\lambda(\{x\}) \tag{31}$$

とするとき， ABIC を λ で微分したものを 0 とおいて得られる式は，

$$\frac{\int A^\lambda \exp(E_L^\mu + E_\pi^\lambda)d\{x\}}{\int \exp(E_L^\mu + E_\pi^\lambda)d\{x\}} - \frac{\int A^\lambda \exp(E_\pi^\lambda)d\{x\}}{\int \exp(E_\pi^\lambda)d\{x\}} = 0 \tag{32}$$

となる．ただし，記号の簡略化のため，$\{y^d\},\{x\}$ 依存性を明記するのをやめた．この式は簡単に解釈できて，“A の事後分布での期待値と事前分布での期待値が等しくなるような λ を選べ” ということになる．

これは，$\langle\ \rangle_{pos}$ で事後分布での期待値をあらわし，$\langle\ \rangle_\pi$ で事前分布での期待値をあらわすと

$$\langle A^\lambda \rangle_{pos} = \langle A^\lambda \rangle_\pi \tag{33}$$

と書ける．μ に関する対応する式も同様にして導けるが，詳しいことは第 2 部の 2.3節及び本文の実例にゆずることにする．

微分した形の式に明快な意味があるので，巨視的パラメータの推定に関しては，それを出発点としてもよいようにも思われる．しかし， ABIC を情報量規準 AIC(付録参照) の拡張として見る立場からは，モデルの選択のために最小化によって得られた ABIC の値自体が意味を持つことになるので，積分形の方が本質的である． AIC の拡張という立場からは (26) で定義した $-\frac{1}{2}ABIC$ から，巨視的パラメータの数を引いたものを規準とすべきだと考えられるが，この違いは問題にならないことが多いようである．

$\{x_i\}$ について和をとる代わりに，$\{x_i\}$ と $\{\lambda,\mu\}$ について同時に事後確率を最大化したらどうなるだろうか．これは巨視的パラメータと微視的パラメータを区別せず，両者について MAP 解をとることに相当する．

(30) 式でいえば第一項の積分を $\{x_i\}$ に関する鞍点 (「古典解」) で置き換えたことになる (この場合，鞍点のまわりの揺らぎは考慮していない．揺らぎを第 1 近似まで取り入れたもの (「半古典近似」) は，後で述べるように ABIC の近似値として有効である)．

(33) 式に当たるものは，定義より

$$\frac{\partial(E_L^\mu + E_\pi^\lambda)}{\partial x_j} \mid_{\{x_i\}=\{x_i^{MAP}(\lambda)\}} = 0 \tag{34}$$

であることに注意すると，

$$A^\lambda(\{x_i^{MAP}(\lambda)\}) = \langle A^\lambda \rangle_\pi \tag{35}$$

となる．

このやり方ではなぜいけないのだろうか．少なくとも，微視的パラメータが連続変数の場合は，この方法では over-fit の問題が再び起こると考えられる (AIC の補正項にあたる尤度の “不正利得” が無視できなくなる)．

微視的パラメータが離散的な場合はもっと微妙である．しかし，たとえば，本文で論じる texture の問題 (無秩序相のイジング模型のパラメータを雑音を含んだ snapshot から推定する問題) では，(33) の左辺にのみ MAP 解を代入するのは明らかに不自然である．

ABIC 法では高次元の積分 (または多重和) の計算が必要になる．事後分布がガウス分布の場合は，ABIC の計算はガウス積分であり，数値的には微視的パラメータの数に比例する大きさの行列 (しばしば疎行列) の QR 分解 (Gram-Schmidt 直交化) の計算に帰着される [K.Tanabe, T.Tanaka(1983)]．

統計数理研究所の研究者による初期の仕事の多くはガウス的な事前分布の場合を扱っている．このうち 密度推定 (**density estimation**, ヒストグラムの推定) の場合は，尤度関数が非ガウス的なため，事前分布がガウス的でも事後分布は非ガウスになる．しかし，この場合も事後分布は **convex** になるので，その意味では非ガウス性はそれほど強いとはいえない (convex の定義は 2.3.2節を参照)．この問題に対しては，事後分布を MAP 解のまわりに 2 次近似する手法 [M.Ishiguro, Y.Sakamoto (1984)] 及びメトロポリス的でないモンテカルロ法 [K.Tanabe, M.Sagae, S.Ueda (preprint)] によって ABIC の計算がなされている．

微視的パラメータが 1 次元的な配列をしている場合 (時系列など) には巨視的パラメータの推定に対しても非ガウスフィルター [G.Kitagawa (1987)] の手法が有力である．

非ガウス性が強い 2 次元以上の問題については，統計物理との類似に基づいてメトロポリス的モンテカルロ法を導入することが考えられるが，これは 2.3節及び本文の主題である．

以下では，ABIC 法に関係があると思われる手法をいくつかとりあげて論じる．用語の違いのために見のがされやすい類似点を強調するように心がけた．

1. **ABIC 法**
   ABIC 法という名前は，[H.Akaike (1980)] による (“a Baysian information criterion”，AIC のベイズ版)．巨視的パラメータの決定手段として ABIC 最小化を用いることはかなり古くから提案されているようであり，H.Akaike の上述論文では [Good (1965)] の **“タイプ 2 最尤法 (Type 2 Maximum Likelyhood Estimation)”**が引用されている．ほかにも **empirical Bayes** などの名称で類似の提案がなされているようである．H.Akaike の仕事の特徴は，ABIC の値自体が意味を持つことを強調したことと，時系列などの大規模モデルへの大胆な適用にある．その後，ABIC 法は統計数理研究所の研究者達によってさまざまな問題への応用がなされている (下記の文献参照)．

   - 時系列 (季節調整，地球潮汐の問題を含む)
     - H.Akaike (1980)
     - M.Ishiguro (1981)
     - Y.H.Tamura (1988)
     - G.Kitagawa (1981,1987) → 非線形フィルター
   - 離散スプライン
     - K.Tanabe, T.Tanaka (1983)
   - 2 値反応曲線・曲面
     - M.Ishiguro, Y.Sakamoto (1983)
     - M.Ishiguro, Y.Sakamoto (1985)
   - 密度推定
     - M.Ishiguro, Y.Sakamoto (1984)
     - K.Tanabe, M.Sagae, S.Ueda (preprint)
   - マーク付き点過程 (2 次元)
     - Y.Ogata, K.Katsura (1988)
   - コウホート分析
     - T.Nakamura (1986)
   - その他 (論文未発表のものを含む)
     - 電波干渉計のデータ解析 (M.Ishiguro)
     - 生存時間分布関数のモデル (T.Kamakura)
     - 白血病のモデル (Y.H.Tamura)

- 呼吸器のモデル (K.Tanabe)
- 地震の頻度と大きさの関係の推定 (M.Ishiguro)
- リッジ回帰の一般化 (M.Ishiguro)

- Y.Sakamoto, M.Ishiguro の仕事については，単行本 'カテゴリカルデータのモデル分析 (坂本慶行 1985，共立出版)' に詳しい．

2. 非ガウスフィルター・隠れマルコフモデル

   R.E.Kalman による時系列の状態空間表現 (**state space reprisentation**) では，状態ベクトル (**state vector**) であらわされる内部空間を考えるわけであるが，これは実質的に微視的パラメータの空間とみなしてよい．この場合，(巨視的) パラメータの推定は状態ベクトルについて和をとった "尤度" で行われるが，これは ABIC 法と同じである．

   いわゆるカルマンフィルターではモデルがガウス的なので，これらの点は表面から隠されているが，G.Kitagawa の最近の研究では，非ガウス・非線型の場合が論じられており，ABIC 法との関連も指摘されている．

   前出の隠れマルコフ鎖も本質的には同じもので，微視的パラメータ (状態ベクトル) が離散的な場合に当たる．

   微視的パラメータの配列が 1 次元的であることは，アルゴリズム的には大きな利点がある．これらの場合，ABIC に当たる量はフィルターのアルゴリズム (「転送積分法」) の副産物として簡単に得られる．

3. 経路積分によるパターン判別

   W.Bialek と A.Zee は [W.Bialek, A.Zee (1987), A.Zee (preprint)] などで，経路積分 (path integral) とパターン判別の関連を論じている．最良の経路 (MAP 解) のみでなく，そのまわりの揺らぎも考えた方が良いという趣旨で，ABIC 法の思想に近い主張と思われる．ただし，彼らの研究は推定型の問題についてではなく，判別・検定型の問題についてである．かれらの主張は第 2 部で述べる simulation without annealing にも関係があると思われる．

4. **EM 法**

   EM 法 (Expectation-Maximization algorithm) というのは統計的方法の名称というよりはアルゴリズムの名称であり，欠測値のある場合，あるいは "潜在構造" のある場合に，(32) 式に類似の式を iteration で解くことにより "最尤解" を求める方法である [A.P.Dempster, N.M.Laird, D.B.Rubin (1977)]．

   EM 法は必ずしもベイズモデルに対するものではないが，A.P.Dempster らは，タイプ 2 最尤法による hyper parameter estimation への応用も述べている．

   いずれにしても，欠測値のある場合，あるいは "潜在構造" のあるモデルにおける "最尤法" は実質的には ABIC 法に近いように思われる．

5. ボルツマンマシンの学習方程式・「エネルギー」学習法

   G.E.Hinton らはイジング型神経回路網 (狭義のボルツマンマシン) の "シナプス荷重"(結合定数) の学習規則として，(32) 式に類似の式を提案した ([G.E.Hinton, T.J.Sejnowski(1986)])．

   このモデルでは，結合定数が場所によってすべて異なっており，それ自体沢山あるので，巨視的パラメータの推定を行っているとはいえない．しかし，隠れたパラメータに関して和を取った "尤度" を最大にするという点では ABIC 法と同じである (彼らは EM 法及び隠れマルコフ鎖に関する文献を引用しているが，発想は独立と思われる)．この研究の特徴は，(32) 式にあたる式の各辺を計算するのに，メトロポリス的モンテカルロ法を用いたことである．

   なお，[G.E.Hinton, T.J.Sejnowski(1986)] は先ほど論じた "なぜ (32) 式の第 1 項 (あるいは第 1 項と第 2 項の両方) で鞍点をとらずに，期待値 (積分) を計算するのか" という問題にも 1 節をさいているが，あまり明快ではない

   イジング型神経回路網 (狭義のボルツマンマシン) についての詳しいことは，付録 B を参照されたい．

   また，Geman & Geman らは非ガウス的な事前分布を用いたトモグラフィの画像処理問題で (32) 式をメトロポリス的モンテカルロ法の助けで解くことを提案し，実行している ([D.Geman, S.Geman (1986), S.Geman, D.E.McClure (1987)])．これは，ABIC 法そのものであるが，Geman & Geman らは単に最尤法と呼んでいる．アルゴリズムについては，EM 法の変形として説明されているが，G.E.Hinton らの研究が引用されているので，おそらくそれに示唆されたものであろう．この研究については第 2 部 (2.3節) で ABIC 法とメトロポリス的モンテカルロ法を組合せた他の研究とともに詳しく説明する．

1.4 大規模統計モデルをめぐる他の話題

1.4.1 多数の離散パラメータを含むモデル

離散的なパラメータを多数含んだモデルは画像モデルに限らない．

いわゆる **mixture** の問題はその 1 例である．これは，各標本が g 個の分布 $\{P_k, k = 1..g\}$ のいずれから抽出されたか不明であるという条件のもとで，分布のパラメータと各標本がどの分布に由来するかを同時に推定するという問題である．この種の問題は，各標本にそれがどこから抽出されたかを示す離散パラメータ $\{S_k\}(S_i \in \{1, 2..g\})$ が付属していると考えることによって，多数の離散パラメータを含んだ問題として解釈しなおすことができる．離散パラメータは数が多くても実質的な自由度は少ないため，なかなか overfit にはならず，informative prior なしでも相当に数が多くなりうる．

いろいろな問題，たとえば，外れ値 (outlier) の検出問題 [G.Kitagawa, H.Akaike (1982)]，クラスター分析 ([N.E.Day (1969), J.H.Wolfe (1970), A.J.Scott, M.J.Symons (1971),D.A.Binder (1978)] が mixture の問題の形に表わせる．ここでは，クラスター分析の場合について簡単に説明しよう．

クラスター分析は，多次元空間の点をいくつかの群に分けるための雑多な手法の混合体であるが，その中のある種のものは統計モデルにもとずく解析法として定式化することが可能である (たとえば，[A.D.Gordon (1981)] を参照)．古典的な例でいうと，階層的な方法や最小展張木 (Minimum Spanning Tree) の方法によるクラスター分析は統計モデルでは書けないが，群内分散を最小化する方法は (本稿でいう意味の) 統計的情報処理に親近性がある．

A.J.Scott, M.J.Symons (1971) は後者のタイプのクラスター分析に対して，ガウス分布の mixture problem としての定式化を与えた．もとの形式は，必ずしも “丸く” ないクラスターを含む非常に一般的なもので，そこに価値があるともいえるが，ここではうんと簡単な場合のみ考えることにする．

データを $(\{y_i\}, i = 1..n)$ とする．これらはすべて p 次元のベクトルとする．すなわち，p 次元空間において “かたまり”(クラスター) を探すわけである．いま，簡単のために “丸い” クラスターのみを考え，クラスターの個数は既知とする．このとき，各クラスターは平均 m_k，分散 σ_k^2 の p 次元正規分布で示される．各データが g 個のクラスター (ガウス分布) のどれにに属するかを定める離散パラメータを $\{S_i\}(S_i \in \{1, 2..g\})$ とすると，尤度関数は，

$$\mathrm{L}(\{y_i\} \mid \{m_k, \sigma_k^2\}_{k=1}^g, \{S_i\}_{i=1}^n) = \frac{1}{Z_L} \exp(-E_L) \tag{36}$$

$$E_L = \sum_{k=1}^{g} \sum_i \delta(S_i - k)\left(\frac{\mid y_i - m_k \mid^2}{2\sigma_k^2} + \frac{\log \sigma_k^2}{2}\right) \tag{37}$$

とかける (Z_L は $\{m_k, \sigma_k^2\}, \{S_i\}$ に依存しない規格化定数)．

ベイズ的に考えるなら，このほかにパラメータの事前分布をあたえる必要があるが，ここでは適当な ignorant prior を用いることとする．

この場合，$\{m_k, \sigma_k^2\}$ の扱いとしては，次の 3 つが考えられる．

1. $\{m_k, \sigma_k^2\}$ を巨視的パラメータとみて ABIC 法 (EM 法) を用いる．

   この場合，微視的パラメータ $\{S_i\}$ について和をとった尤度を用いて $\{m_k, \sigma_k^2\}$ を先に推定し，次に $\{S_i\}$ の事後分布を考えることになる．いまの場合，尤度が $\delta(S_i - k)$ について 1 次なので，和はすぐ計算できるが，それを用いて $\{m_k, \sigma_k^2\}$ を計算するのはかなり面倒な非線形最適化になる．その後，微視的パラメータ $\{S_i\}$ を推定する．これは複数の既知のガウス分布から得られたデータを判別する問題であり，超平面のどちら側にあるかを調べることに帰着される．これは [N.E.Day (1969), J.H.Wolfe (1970)] などによって研究された方法である． EM 法との関係及び他の文献については，[A.P.Dempster, N.M.Laird, D.B.Rubin (1977)] を参照されたい．

2. $\{m_k, \sigma_k^2\}, \{S_i\}$ について同時に尤度 (事後分布) を最大化する

   まず，$\{S_i\}$ を固定しておいて，$\{m_k, \sigma_k^2\}$ を尤度を最大にするように決める (これは解析的にできる)．$\{m_k, \sigma_k^2\}$ を消去したことにより，S_i の間には複雑な相互作用が生じるが，この系に対して逐次改良法 (ICM) を用いることにより，$\{S_i\}$ を推定する．この方法は [A.J.Scott, M.J.Symons (1971)] によるが，群内分散を最小化する方法の一般化になっている．

3. $\{m_k, \sigma_k^2\}, \{S_i\}$ の事後分布を考え，必要に応じて周辺分布を求める．

   この方法では，(2) の場合と同様，巨視的パラメータと微視的パラメータを区別しない．

なお，この問題とガウス的な多次元判別問題との相違をわれわれの言葉で表現すれば，判別問題が “教師あり” なのに対し，クラスター分析は “教師なし” であるといえる．

一般に mixture の問題はパラメータ $\{S_i\}$ の間に直接の相互作用がないため，画像問題に比べると簡単である．しかし，問題を少し変えて，データそのものが何種類かの分布から引き出されるとする代わりに，データを作り出す要素が何種類かある場合を考えると，パラメータ $\{S_i\}$ の間に直接の相互作用を持つモデルが生ずる．

本文で論じる “対比較に基づく分類” はこのような問題のもっとも簡単な例である．この課題では，リーグ戦の対戦の勝敗表がデータとして与えられているときに，各個体がいくつかのグループのどれかから抽出されているとして，各個体の所属と各グループ間の強弱が問われる．

このような種類の問題に対して，第 2 部で述べるような統計物理との類似に基づいた手法を適用することはまだ殆ど行われていない (註参照)．本文の筆者の研究はその意味では新しい試みである．

> 群内分散を最小化する方法によるクラスター分析に simulated annealing 法を適用した研究があるらしいが，統計モデルという意識があるのかどうかは分からない．
>
> 統計数理研究所のシンポジウム (1989 Dec.) で A.F.J.Smith と議論したところによると，A.F.J.Smith もメトロポリス的なモンテカルロ法 (Gibbs Sampler) を “通常の統計の問題” に適用することに興味を持っているとのことであった．上記に関係のある話題としては，抗体 (antibody) の分析に関してあらわれる大規模な mixture problem を扱っている由である (未発表)．
>
> 通常の統計モデルと新しい見方の関係という意味では，カテゴリカルデータの分析に用いられる対数線型模型 (**log linear model**) がイジング型神経回路網 (狭義のボルツマンマシン) として知られているものに形式的に非常に近いことが注目される．これは両者を知っていればすぐに気付くことであるが，両者の関連を明確に指摘した文献は見たことがない．両者の類似点と相違点を分析することは付録 B に譲る．

1.4.2 コネクショニズム・Neural Network

最近，コネクショニズムの名のもとに，多数の要素をもった非線型のネットワーク (“神経回路網”,**“Neural Network”**) を利用した情報処理が流行している．このグループは心理学者，工学者，物理学者，生理学者等を含んでおり，その “教科書” **Parallel Distributed Processing** (1986) にちなんで，**“PDP”**グループとも呼ばれている．

コネクショニズムは単一の思想というより，ある傾向をもった思想の集合体と見られるが，そこで使われているモデル (具体的には **feed-forward back-propagation** 型ネットワークとイジング型神経回路網 (狭義のボルツマンマシン)) についていえば，統計モデルの一種として見るのが最も良い見方だと思われる．

普通，これらのネットワークは‘神経回路網” として考えられており，歴史的にはパーセプトロンの系譜を引き継ぐものとされている．これはもちろん誤りではないが，実は 3 層パーセプトロン自体が (ガウス的な) 統計モデルなのである．この見地から見れば，パーセプトロンは忘れられていたのではなく，多次元の判別分析あるいは数量化 2 類として広く実用に使われていたという考え方さえ可能である (すべての “神経回路網” が統計モデルと見なせるわけではないことに注意．たとえば，非対称結合の連想記憶のようなものは通常の仕方では統計モデルとは解釈できない．ここで重要なのは，結合が対称か否かではなく，機能が確率分布に基いて記述できるか否かである)．

また，新しい神経回路網の特徴は，“隠れユニット”(**hidden units**) を持つにもかかわらず，系統的な学習が可能であることとされているが，これについても統計学的な対応物が存在する．実際，前に註のなかで説明したように，イジング型神経回路網の場合，隠れユニットのある場合の学習則は ABIC 法ないし EM 法とほとんど同じである．これは，第 2 部の 2.3節の例と付録 B を比較すれば一目瞭然である．この場合，“隠れユニット” は微視的パラメータ，あるいは非線型フィルターにおける内部空間に対応するわけである．

Geman&Geman は PDP グループと研究上のつながりがあり，その仕事とイジング型神経回路網との関係は比較的よく知られている．このとき，Geman&Geman における補助的な要素 (線過程) を PDP の隠れユニットに同定することが多いようであるが，不完全データの場合は補助的な要素がなくても画素自体がすでに “隠れている” ことを注意しておく．これは，特定の画素が 100% 分かっており他が全く不明という場合を考えるとよくわかる．この場合はもちろん不明の画素が隠れユニットであるが，それと比較した場合，どこに雑音が入っているか分からないときはすべての画素が少しずつ隠れていることになる．

コネクショニストのネットワークが統計モデルの一種とみなせるといっても，伝統的な統計モデルとの違いはもちろん存在する．これは，

1. データの種類によらず処理できる万能の非線型モデルをめざしている (統計学でも頑健性ということはいわれるが，本質的にはモデルは “オーダーメイド” と考えられている)．
2. overfit に対す考え方の違い．
3. “分散表現” を標傍している．

4. データの与え方が違う．

などにまとめられる．

付録 B において，イジング型神経回路網と古典的な統計モデルである対数線型模型 (**log linear model**) の比較を行う．上記の相違点についてもそこで詳しく論じる．

> **Harmony Theory** について：
> P.Smolensky は，コネクショニズムの理論として統計的情報処理の枠組みの上に立って，**‘Harmony Theory’** を唱えている．これは，枠組みとしては統計的情報処理の枠組みに非常に近いものと考えられる．PDP の Harmony Theory の章では，統計物理とのアナロジー，特にメトロポリス的モンテカルロ法の応用が強調され，巨視的変数と微視的変数の区別，比熱の意味なども説かれている．これは，最後の点をのぞいてはオリジナルな理論というよりは総説と考えたほうがいいのかもしれない．総説としては興味深いものであり，参考にした点もあるが，視野の広さという点ではこの総説のほうがはるかに広い．

2　統計物理と統計的情報処理

第 2 部では，第 1 部で述べた展開の結果として，統計モデルが大規模化，非ガウス化し，統計物理 (とくにスピン系) の研究者が興味を持ってきたモデルに似てきたこと，その結果としてメトロポリス的モンテカルロ法に代表される統計物理の手法の適用が有効性をもつようになったことを説明する．

2.1　非ガウス・大規模モデルと統計物理

熱平衡統計力学と統計的情報処理は，どちらも確率分布を扱うので，その意味では似ていて当り前である．最大の違いは，統計物理は沢山の変数を扱うため，そのための技術や概念が発達しているのに対し，いままでの統計学ではそのようなことが問題にならなかった (あるいは難しいから回避されてきた) ということである．

ところが，計算機の発達にともなって統計モデルは大規模化し，上記のギャップは埋まりつつある．とくに，informative prior を使用した場合， overfit の問題はなくなるので，非常に大規模なモデルが出現しうる．たとえば，画像処理では微視的パラメータが数万個あっても不思議ではない．また， 1.4.1節で述べたように，パラメータが離散的な場合も大規模なモデルが出現しうる．

この結果としてあらわれた統計モデルの多くは統計物理ですでに研究されているモデルに次の 2 つの意味でかなり似ている．

まずひとつは informative prior として導入される 事前分布 が統計物理のモデルに似ているということである．とくに画像のマルコフ場モデルについては，イジング模型やポッツ模型はもちろん，もっと実用的なモデルも統計物理の言葉でいえば 2 次元のスピンモデルの 1 種といってよい型をしている．これらの場合，類似の対照になる統計物理のモデルは “純粋系” のモデルである．

次にいえるのは，大規模モデルによって生成される 事後分布 もまた統計物理のモデルに似ているということである．事後分布には “データ” が含まれているので，類比の対照になる統計物理のモデルも対応する要素 (たとえば非一様な外場) を含まなくてはならない．このような系についての統計物理での知見は多くないが，例外として “データ” が完全にランダムな場合 (雑音ばかりの場合) は “ランダム系の統計物理” としてかなり研究されている．

多くの場合，画像の大規模ベイズモデルから生ずる事後分布は，データを **quenched randomness** と見たとき，「ランダム磁場型」になる．たとえば，(19,21) から生成される事後分布は，2 状態 ($x_i \in \{-1,+1\}$) モデルの場合，

$$P(\{x_i\}) = \frac{1}{Z_{pos}} \exp(\frac{1}{2}J \sum_{j\in N(i),i} x_i x_j + h \sum_i y_i x_i) \tag{38}$$

となる (ただし $J = K/2$ とした)．これは場所によってことなる磁場 $\{hy_i\}$ が加わっているイジング模型にほかならない．$\{y_i\}$ が雑音を含んだデータの場合，分布 (38) の性質は，磁場が一様な場合と完全にランダムな場合の中間になると思われる．後者に対応する統計物理の問題は，局所的なダイナミクスのもとで沢山の準安定状態を持つ場合があるが，これは画像処理の場合にも同様の状況が (程度の差こそあれ) 存在することを示唆する．なお， “沢山の準安定状態を持つ” というかわりに， “組合せ論的な複雑さ (**combinatorial difficulty**) がある” といってもよいが，ここでは最適化 (「T=0」) だけではなく分布の生成についても考えているので，統計物理の言葉を用いることにする．

事後分布が「ランダム結合型」ないし「スピングラス型」になる問題も存在する．付録 B で述べるイジング型神経回路網の “想起状態” はその 1 例である．もっと統計モデルらしい例としては，筆者による “対比較にもとずく分類” の問題がある．このような問題の場合も，対応するランダム系の統計物理の性質から，沢山の準安定状態を持つ可能性があることがわかる．

このような類似性を理論的な研究に生かした例はいまのところ非常に少ない．現在話題になっているのは，メトロポリス的モンテカルロ法などの統計物理で常用されているアルゴリズムを対応する統計の問題に役立てる試みであり，本文及び総説の以下の節も主にこの問題を扱う．

しかし，このような場合でも，統計物理の諸概念や知見は，背景的に意味を持つはずである．これについては，次節及び本文の各所で指摘する．

> 統計物理と大規模モデルを用いる統計学の類似性を認識し，応用することは，いろいろなグループによって独立になされている．
>
> G.E.Hinton と T.J.Sejnowski のボルツマンマシン， P.Smolensky の Harmony Theory, Geman&Geman らのメトロポリス的モンテカルロ法による画像処理の研究は最もよく知られた例である．このうち， Geman&Geman の研究に関しては， U.Grenander の示唆によるところが大きいらしい [U.Grenander (1983)]．画像の texture の生成をメトロポリス的モンテカルロ法で行った研究としては， [G.C.Cross, A.K.Jain (1983)] がよく知られている．マルコフ場と統計物理の関係を最初に指摘したのは， [J.Moussoris (1974)] だといわれている．

統計数理研究所の Y.Ogata と M.Tanemura は点配置の問題と液体の統計力学の類似 [Y.Ogata,M.Tanemura (1981)] から出発して独自の路線をすすんでおり，マルコフ場を用いた画像処理の研究 [Y.Ogata(preprint)] はその路線上にあるといえる．

統計物理学者による最近の研究 (理論) としては，[W.Bialek, A Zee (1987,1988)] がある．

筆者は 1986 年から 1987 年にかけて統計数理研究所のグループの仕事 (Y.Sakamoto, M.Ishiguro, K.Tanabe などの研究) を学んだ際に Geman&Geman と似たようなことを考え，それ以来この方向の研究をしている．正確にいうと，その時すでに Geman&Geman(1986) を見ていたが，単なる simulated annealing の応用例だとしか思っておらず，読み直してはじめて意味が分かった．いわゆる PDP に接したのはもっとあとである．

2.2 メトロポリス的モンテカルロ法と微視的パラメータの推定

2.2.1 Simulation without Annealing

ベイズ統計でなんらかの効用関数を最大にする解を求めたいときやその解の誤差や確からしさを求めたいときに必要となるのは，多くの場合事後分布そのものではなくて，大部分の微視的パラメータに関して和をとった分布すなわち周辺分布 (marginal distribution) である (この用語は度数分布表の周辺和に由来する)．和をとること (周辺化) は，統計物理でいう粗視化の操作に対応する．

たとえば，微視的パラメータ x_i の期待値を求めるには，規格化されていない事後分布を $\tilde{P}\{x_i\}$ とすると，周辺分布

$$P_i^m(x_i) = \frac{\int \tilde{P}(\{x_j\})\Pi_{j\neq i}dx_j}{\int \tilde{P}(\{x_j\})\Pi_j dx_j} \tag{39}$$

を用いて，

$$\langle x_i \rangle = \int x_i P_i^m(x_i) dx_i \tag{40}$$

を求めればよい．

そこで，大規模ベイズ模型を扱うためには，規格化定数のわからない高次元・非ガウスの分布について，周辺分布を生成することが重要な問題になってくる．これは，まさに統計物理で問題になっていることであり，そこでの手法，特にメトロポリス的なモンテカルロ法が役立ちうる (他の統計物理の手法については 2.4節で述べる)．

ここでいうメトロポリス的なモンテカルロ法は狭義のメトロポリス法や熱浴法の総称であって，定常分布が欲しい分布になるようなマルコフ鎖を構成して，乱数を引いてシュミレートする方法である (**Gibbs Sampler** と呼ばれることもある)．この結果，実質的な緩和時間を $\tilde{\tau}$ とすると，$\tilde{\tau}$ 以上の間隔でマルコフ鎖から取り出されたサンプルは事後分布からのランダムサンプルと見なせ，これを用いて周辺分布を計算することができる (詳しいことは，統計物理の解説参照)．

たとえば，(38) の事後分布の場合，エネルギーが，

$$E_{pos}(\{x_i\}) = -\frac{1}{2}J \sum_{j\in N(i),i} x_i x_j - h\sum_i y_i x_i \tag{41}$$

で表わされる統計物理のモデルだと思ってモンテカルロシミュレーションを行えばよい．また，連続パラメータで事前分布が (15) なら，

$$E_{pos}(\{x_i\}) = -\frac{1}{2}\sum_{j\in N(i),i} \frac{\lambda}{(x_i - x_j)^2 + \xi^2} - \sum_k \frac{(y_k - x_{\eta(k)})^2}{\sigma^2} \tag{42}$$

とすればよい ($x_{\eta(k)}$ はデータ y_k の上にある面要素)．

なお，anneal しない ことを明示するうまい用語がないので，本稿ではこの手法を **simulation without annealing** と呼ぶことにする．これは，後出の **simulated annealing** を念頭に置いた筆者の造語である．

メトロポリス的なモンテカルロ法を用いた計算が速く収束するためには，一種の局所性あるいは独立性が必要である．さもなければ，最悪の場合には，すべての場合を調べあげるまで収束しないことになる．

2 次元の画像問題では明らかにある程度の局所性がある．すなわち，ある画素を推定するために大きな情報を与えるのはその近くの画素の状態で，うんと遠い画素の状態はあまり影響しない (もちろん例外もある)．

どの要素も全部の要素の何割かとつながっているような “ネットワーク型の問題” では，この種の独立性は明白ではなくなるが，これは必ずしもすべての場合を調べあげる必要があることを意味しない．実際，本文で扱った “対比較に基づく分類” の問題では，条件にもよるが，かなりすみやかな収束が得られる．統計物理での経験も，全部の場合を数え上げなくても結果が得られる場合が多くあることを示している．

どのような場合に局所性，独立性が成り立つかは重要な問題であるが，統計物理はこの問題に対していくらかの示唆をあたえうる．たとえば，相転移という現象は (何らかの意味で) 局所性の喪失と深く結び付いている．2 次元強磁性イジング模型の場合は転移点はそれより強結合 (低温) 側では，相関距離が無限大になる点として特徴付けられる．“ランダム系” であるスピングラスの場合も，相転移は局所性の喪失と結びついている (“スピングラス相” では局所の変化が全体に波及するが，“パラ相” では局所的にとどまる)．“ネットワーク型の問題” である SK 型のスピングラスでもやはり転移があり，転移点は 1 個のスピンを動かしたときの feedback loop の寄与が発散する点として特徴付けられる．

これらの例の第 1 の教訓は，モデルの相互作用のみかけの局所性と作られた分布の局所性が一致するとはかぎらないということである．しかし，それ以上の応用はそう簡単ではない．たとえば，相転移がなくてもメトロポリス的

モンテカルロ法の収束が遅くなることもある (2 次元のランダム磁場イジング模型の場合には，相転移がないのにもかかわらず低温での収束は極めて遅い) . また，統計的情報処理であらわれる系は有限系で，多くの場合非一様であるから，統計物理の結果がただちに使えるわけではない . これらを考えると，当面は統計物理の知見を念頭において，さまざまな実例に関して経験を積むことが重要と思われる .

以上では全部数え上げる場合との比較をしたが，最初から局所性を仮定した近似法と比較した場合の simulation without annealing の特徴は，逆に，局所性がそれほどなくても使える，ということである . 統計物理で臨界現象の研究にさかんに使用されているのはこのためである . また，“硬い” アルゴリズムと違って局所性の程度をあらかじめ仮定しないでも適用できるというのも大きな利点である .

もちろん，どんな方法でも同じであるが，万能視するのは危険である . 特に，収束が “遅い” 場合には気がついても，“極めて遅い” 場合には収束したと勘違いすることが多い . “有限系のエルゴード性の証明” などは実際の状況では何の意味ももたないことに注意して，慎重に計算しなくてはならない .

MAP 解は周辺分布を用いてかけないので simulation without annealing を用いてもとめることはできない . MAP 解とメトロポリス的モンテカルロ法の関係 (simulated annealing 法) については次節 (2.2.2) で述べる .

本節で述べた方法が画像処理に適用可能なことは，すでに [S.Geman, D.Geman (1984)] で示唆されている (U.Grenander([U.Grenand
が最も早いともいわれている) . しかし，実際には，Geman & Geman の初期の仕事はもっぱら simulated annealing 法によるものである . これに対して，J.Marroquin([J.Marroquin (1985)]) は anneal しない方法のすぐれた点を強調した . この問題については，次節 (2.2.2) で simulated annealing 法の説明をしてから，あらためて 2.2.3節で論じることにする . 筆者の貢献についてもそこで紹介する .

2.2.2 Simulated Annealing

MAP 解を求めるひとつの方法は，局所的な逐次改良法である . J.Besag はこれを **ICM(Iterated Conditional Mode)** と名付けて各種の画像問題に適用している [J.Besag (1986), S.Geman, D.E.McClure (1985,1987)] . この方法は簡単で計算も速いが，局所的な極値 (局所的なダイナミクスのもとでの準安定状態) に収束する可能性がある . 2.1節で述べたように，多くの局所的極値をもつ問題は少なくないので，逐次改良法で対応できる範囲には限界がある .

ここで，**simulated annealing** 法の適用が考えられる . 事後分布が，

$$P(\{x_i\}) = \frac{1}{Z_{pos}} \exp(-E(\{x_i\})) \tag{43}$$

とかけるとしたとき，これに「温度」T を導入して変形し，

$$P_T(\{x_i\}) = \frac{1}{Z_{pos}} \exp(-\frac{E(\{x_i\})}{T}) \tag{44}$$

とする . (38) の例でいえば，

$$P_T(\{x_i\}) = \frac{1}{Z_{pos}} \exp(\frac{1}{2}\frac{J}{T} \sum_{j \in N(i),i} x_i x_j + \frac{h}{T} \sum_i y_i x_i) \tag{45}$$

とするわけである . ここで，はじめは T を大きくして (「高温で」) メトロポリス的モンテカルロ法による simulation を実行し，ゆっくり $T \to 0$ とする . この場合，$T = 1$ でちょうどもとの事後分布になるわけだが，一般にはもっと高い T からはじめて差し支えない . $T = 0$ では正しいギブス分布 (44) は $-E(\{x_i\})$ を最大とする $\{x_i\}$ の上の δ 分布，すなわち MAP 解を与えるから，途中で準安定状態につかまらなければ $T \to 0$ で MAP 解が得られるはずである . はじめから $T = 0$ でメトロポリス的モンテカルロ法を行えば逐次改良法と同じであるが，ゆっくり冷やす (**anneal**，焼きなまし) 過程を経ている分だけ，準安定状態につかまる率が低く，また，仮につかまっても真の解に近いものが得られることを期待するわけである .

この方法の直観的根拠は文字通り，物理における「焼きなまし」にあるが，もう少しよく考えると，$E(\{x_i\})$ の曲面が図 12 の右のようなものでなく，左のようなものであることを前提にしていることがわかる (これについてのの議論は [E.B.Baum (1986), M.A.Moore (1987), J.Bernasconi (1987)] を参照) .

“完全にランダム” な系，すなわちランダム磁場イジング模型，スピングラス模型などについて，アニール速度を遅くしていったとき，得られる解のエネルギーの漸近形を予想する試みが統計物理の研究者によって行われている ([D.A.Huse, D.S.Fisher (1986)] 参照]) . もっと理論的な研究としては，非常に時間をかけて焼きなましを行った場合に必ず正解に収束するという証明 (たとえば [S.Geman, D.Geman (1984)]) があるが，このような証明が実際的な意味をもつのかどうかはよくわからない .

simulated annealing 法は [J.Kirkpatrick, C.D.Gelatt Jr, M.P.Vecchi (1982)] によって導入されたが，この際に念頭におかれたのは，巡回セールスマン問題， LSI の設計問題などの NP 完全な組合せ的最適化問題であって，特にベイズ統計とのつながりはなかった．これらの問題への応用では，コスト関数をエネルギーとみなして，形式的にギブス分布 $\exp(-E_\alpha/T)$ を考え，ゆっくり $T \to 0$ とすることによって (近似的な) 最適解を求めるわけである．この場合のギブス分布は全く人工的につくられたもので，ベイズ統計に応用した場合と違って，($T=1$ の場合の) 分布自体がとくに意味をもつわけではない．

simulated annealing 法による画像処理に対して，“NP 完全でない問題では，通常の組合せ的最適化のアルゴリズムを適用して厳密な MAP 解を求めたほうがよい” という批判がある．たとえば分布 (38) の場合， MAP 解を求めることはグラフの最小カットを求めることに帰着され，微視的パラメータの数の多項式オーダーで可能である (統計物理の解説参照)．実際，分布 (38) の場合はグラフ理論のアルゴリズムを用いたほうが simulated annealing 法より速いようである．また，この方法で得た正確な MAP 解と比較すると， simulated annealing 法による MAP 解は必ずしもよい近似解とはいえない (ICM で得た解より多少ましな程度に過ぎない) という主張もなされている (図 13, [B.T.Porteous, D.M.Greig, A.H.Scheult (1989) , J.Besag (1986) の discussion])．統計物理でも，ランダム磁場イジング模型の基底状態を求めるには，メトロポリス的なモンテカルロ法よりもグラフ理論のアルゴリズムを用いた方が速いという研究がある ([A.T.Ogelski(1986)])．

これに対する反論として，

- 上記の例は simulated annealing 法が特に苦手とする例なのではないか．
- 少しモデルを複雑にすると，多項式オーダーの解法はなくなるか，少なくとも非常に面倒になるのではないか．
- MAP 解を求めるための解法は，周辺分布が欲しい場合には役立たない．いいかえれば， simulated annealing を代替することはできても， simulation without annealing のかわりにはならない．

といったことが考えられる．

1 番目の説によると，画像問題はすべて苦手な問題に入りかねず，実際そういう主張をする人もいるが，もっと調べてみないとよく分からない．なお，筆者の扱った “対比較による分類” の問題では， simulated annealing 法は非常に有効であった (本文参照)．3 番目の論点は重要なので，次節 (2.2.3) 以降で詳しく論じる．

> [V.Cêrný(1982)] も J.Kirkpatrick らと独立に同様のことを考えたらしい．さらに古い時代のものとしては， S.Watanabe がすでにクラスター分析にランジュバン方程式を応用する可能性を論じているのが注目される [認識とパターン (岩波新書) に言及があるが，オリジナルの論文は不明．また，アニールするのかしないのかも不明．]．また， [T.Tsuda, T.Kiyono(1964)] はわずか 2 変数の問題についてではあるが， simulated annealing 法に近い方法で方程式の大域解を求めることを論じている．
>
> simulated annealing 法をはじめて画像処理の問題に適用したのは，すでに述べたように， [S.Geman D.Geman(1984)] であるが，同時期の独立の研究がいくつかあるようである．筆者が読んだものとしては， [P.Camevali, L.Coletti, S.Patarnello(1985)] がある (ただし，モデルも簡単で，ベイズ統計との関係も言及されていない)．

2.2.3 Anneal する場合としない場合の比較

J.Marroquin は [J.Marroquin (1985)] において， MAP 解のかわりに MPM 解 (Maximum Posterior Marginal estimate) を simulation without annealing で計算することを提案している．これは各格子点をばらばらに考えて，それぞれで最も確率の大きい状態を選んだ解である．2 値画像 ($x_i \in \{-1,+1\}$) の場合で説明すると，各格子点 i ごとに微視的パラメータが $+1$ をとる確率を p_i としたとき，

$$x_i^{MEM} = +1 \quad \text{if} \quad p_i > 0.5 \tag{46}$$

$$x_i^{MEM} = -1 \quad \text{if} \quad p_i < 0.5 \tag{47}$$

で定まる $\{x_i^{MEM}\}$ が MEM 解になる． MEM 解を効用関数で特徴付けると， “もとのパターンとの重なりの期待値を最大にする解” ということになる．これに対して MAP 解は， “もとのパターンとすべての微視的パラメータが厳密に一致する可能性を最大にする解” である．

J.Marroquin の主張は，

- 雑音が多い場合には， “もとのパターンとすべての微視的パラメータが厳密に一致する可能性” はほとんどないのに，それを最大にする意義は疑わしい．
- MEM 解の方がずっと良い結果を与える例がある．
- simulated annealing より simulation without anealing のほうが速く収束するし， anneal の速度を調節する手間がいらない．

といったところにある．

J.Marroquin の指摘以外に， simulation without annealing の重要な利点として，

- 解のベイズ的意味での誤差や確からしさが計算できる．

という点がある．これは，[S.Geman, D.Geman (1984)] や [J.Besag (1986) の discussion] ですでに示唆されているが，筆者や Y.Ogata の主張するところでもある．

また， simulation without annealing は技術的な面だけでなく，思想的な面からも興味深い．統計物理の立場からいうと，「温度」T が有限 ($T = 1$) の状態が意味を持ち， “動いている状態” で情報処理が行われるというのは目新しく感じられる．これに比べると，最近 simulated annealing 法に関連して論じられている「有限温度の最適化問題」(たとえば「有限温度の巡回セールスマン問題」) などは，確率分布と関係ない問題に技巧的に確率を入れて作った問題であり， T=0 以外ではもともとの工学的意味がない点で面白味にかける．

simulation without annealing には最適化という概念の絶対視をのりこえるという意味もあるかもしれない． simulated annealing は大ざっぱな最適化という主張に基づいているが，これは図 14 でいえば， “(*) が求められなくても (**) が求められれば十分ではないか” というふうに解釈できる．しかし，本当は (*) も (**) もいらないのであって必要なのはそのあたりに雲のように広がった分布なのかもしれない (標語にすれば “点から分布へ”)．

筆者は， simulation without annealing に非常に関心があり，以下のような研究を試みた (詳細は本文参照)．

- “対比較に基づく分類” の問題において，求める量によっては simulation without annealing のほうが結果がよいことを示した．
- Marroquin の示した例には疑問があることを指摘し，別の条件で同じ問題を調べた．結果は simulation without annealing を支持するものであった．
- 周辺分布に対する近似として平均場近似を提案し，その効用と限界を具体例でしめした (2.4節参照)．
- 周辺分布そのものに関心のある場合を具体例で示し，そのような問題に simulation without annealing を適用して見せた．また，どのような場合に周辺分布が要求されるのか考察した．

[J.Marroquin(1985)] には，重なり最大解としての MEM 解を提唱すること自体が新しいかのような書き方がしてあるが，これは書きすぎで，このことはこの分野の専門家には以前から知られていたようである (たとえば [H.Derin, H.Elliott, R.Cristi, D.Geman (1984)])．また，前述のように simulation without annealing という方法が可能なことも早くから知られていた．興味深い点は， MEM 解の利点をはっきり指摘したことと， (必ずしも良い例とはいえないが)MEM 解の方がずっと良い例があると主張し，それを simulation without annealing で実際に求めてみせた点にある．

simulation without annealing に関する J.Marroquin(1985) 以後の文献を以下にあげる．

[S.Geman, D.E.McClure (1987)] は SPECT の画像処理問題に simulation without annealing を適用している．この問題はほぼ連続変数 (256 段階) であって，各点での期待値が解として用いられた (MMSE 解， Minimum Mean Squared Error estimate)． (15) 型の事前分布が用いられている． MMSE 解は他の方法で得た近似的な MAP 解とわずかに違う程度と述べられているが，いまの場合に近似的な MAP 解がどの程度正しい MAP 解になっているのかは不明である (いうまでもなく，正しい MAP 解のほうが近似的な MAP 解より好ましい振舞いをするとは限らない)．

[Y.Ogata(preprint)] では，連続変数の 2 次元画像処理問題が扱われている．事前分布は (16) 型その他であり，やはり期待値 (MMSE 解) が解として求められている． MAP 解との比較はなされていないが，誤差が同時に求められているところに特色がある．

この原稿を執筆中に行われた統計数理研究所のシンポジウム (1989 Dec.) での発表では，招待講演者の J.York (J.Besag との共著)， A.F.J.Smith の両者とも， simulation without annealing の立場であった． J.York は離散および連続変数の画像処理を扱い，誤差が求められることを強調していた． A.F.J.Smith の方は一般論であったが，具体的な問題について，すでに数編の公表論文があるらしい (未入手)．

2.3 メトロポリス的モンテカルロ法と巨視的パラメータの推定

メトロポリス的なモンテカルロ法の巨視的パラメータの推定に対する応用を論じる．はじめに不完全データ (教師なし) の場合を取り上げる．この場合， ABIC 法とメトロポリス的モンテカルロ法を組合せた方法が使えるが，これは simulation without annealing の応用例になっている．このとき解くべき式は， neural network の用語で学習方程式と呼ばれるものと本質的に同じである．次に，上記の特殊な場合として，完全データ (教師あり) の場合を簡単に論じる．教師ありの場合はいろいろな意味で簡単であり，推定する巨視的パラメータの種類によっては，巨視的パラメータの空間での convexity が示せる．

2.3.1 不完全データ (教師なし) の場合

ABIC 法とメトロポリス的モンテカルロ法を組合せた方法による巨視的パラメータの推定法は [S.Geman, D.Geman (1986), S.Geman, D.E.McClure (1987)] によって，トモグラフィの画像処理の問題で導入された．

筆者は同様な方法で， 2 次元イジング模型の雑音を含むパターンから巨視的パラメータを推定することを試みた．この問題はトモグラフィの問題より易しい点が多いが，雑音の大きさと結合の強さの同時推定を試みた点やさまざまな条件で正解の知られているデータによるテストを行った点には意義があると思う．

以下では，筆者の例を用いて ABIC 法とメトロポリス的モンテカルロ法を組合せた方法による巨視的パラメータの推定法を解説する (実験の結果については本文参照)．

この例では，事前分布と尤度関数はそれぞれ，

$$\pi(\{x_i\}) = \frac{1}{Z_\pi}\exp(\frac{1}{2}J\sum_{j\in N(i),i} x_i x_j) \tag{48}$$

$$\mathrm{L}(\{y_i\}\mid\{x_i\}) = \frac{\exp(h\sum_i y_i x_i)}{(2\cosh(h))^M} \tag{49}$$

となるから，

$$A_\pi = -E_\pi/J = \frac{1}{2}\sum_{j\in N(i),i} x_i x_j \tag{50}$$

$$A_L = -E_L/h = \sum_i y_i x_i \tag{51}$$

とおくと，(26) 式より， ABIC は

$$-\frac{1}{2}ABIC(J,h) = \log\sum_{config.}\exp(JA_\pi + hA_L) - \log\sum_{config.}\exp(JA_\pi) - M\log(2\cosh(h)) \tag{52}$$

とかける．

ここで， M は微視的パラメータ (= 画素)$\{x_i\}$ の数であり， $\sum_{config.}$ は $\{x_i\}$ のすべての組合せ (2^M通り) に関する和である．

(52) 式を見ると，第 1 項がデータ $\{y_i\}$ に由来する非一様な磁場中のイジング模型の自由エネルギー，第 2 項が磁場なしのイジング模型の自由エネルギーにそれぞれ対応していることが分かる．

ABIC 最小解の満たす方程式は， (52) 式を J と h でそれぞれ微分して 0 とおくことにより，

$$-\frac{1}{2}\frac{\partial(ABIC)}{\partial J} = \frac{\sum_{config.} A_\pi\exp(JA_\pi + hA_L)}{\sum_{config.}\exp(JA_\pi + hA_L)} - \frac{\sum_{config.} A_\pi\exp(JA_\pi)}{\sum_{config.}\exp(JA_\pi)} = 0 \tag{53}$$

$$-\frac{1}{2}\frac{\partial(ABIC)}{\partial h} = \frac{\sum_{config.} A_L\exp(JA_\pi + hA_L)}{\sum_{config.}\exp(JA_\pi + hA_L)} - M\tanh(h) = 0 \tag{54}$$

となる．

これらはそれぞれ，

$$\langle A_\pi\rangle_{pos} = \langle A_\pi\rangle_\pi \tag{55}$$

$$\langle A_L\rangle_{pos} = M\tanh(h) \tag{56}$$

と書ける．この場合，尤度関数の方に入っている巨視的パラメータ h(一般論の μ) で微分した式の方も分かりやすい形をしているが，これは尤度関数がデータと微視的パラメータに関して対称的な形をしているおかげである．

ここで，simulated without annealing を用いて，事後分布での期待値 $\langle A_\pi \rangle_{pos}$，$\langle A_L \rangle_{pos}$ 及び事前分布での期待値 $\langle A_\pi \rangle_\pi$ を求め，巨視的パラメータを修正して，再び simulation をするという操作を繰り返すことで (55),(56) 式を解くのが，Geman & Geman の方法の骨子である．

微分方程式で書くと，

$$c_J \frac{dJ}{dt} = \frac{1}{2}\frac{\partial(ABIC)}{\partial J} = \frac{\sum_{config.} A_\pi \exp(JA_\pi + hA_L)}{\sum_{config.} \exp(JA_\pi + hA_L)} - \frac{\sum_{config.} A_\pi \exp(JA_\pi)}{\sum_{config.} \exp(JA_\pi)} \tag{57}$$

$$c_h \frac{dh}{dt} = -\frac{1}{2}\frac{\partial(ABIC)}{\partial h} = \frac{\sum_{config.} A_L \exp(JA_\pi + hA_L)}{\sum_{config.} \exp(JA_\pi + hA_L)} - M\tanh(h) \tag{58}$$

となる (t は仮想的な時間，c_J, c_h は時定数)．これらはイジング型神経回路網における “学習方程式” に相当する．また，「エネルギー」の期待値に関連する式なので，“エネルギー学習法” の式と呼ぶこともある．実際にこれを解くときには，仮想的な時間 t に関して差分化するので，最急上昇法あるいは EM 法とでもいうべきものになる．

ここで，t はメトロポリス的なモンテカルロ法における仮想的時間 (**MCS**) とは別物である．理想的には，t を $t+dt$ に動かすたびに期待値が収束するまでモンテカルロ法を十分長く走らせなければならない (詳しくは本文で論ずる)．

なお，事後分布での期待値の計算と事前分布での期待値の計算は別の simulation を要するので，2 つの simulation を並行して走らせることになるが，[S.Geman, D.E.McClure (1987)] では後者を off-line で計算して表にしておく方法をとっている．

[Y.Ogata (preprint)] も独立に類似の方法に到達している．Y.Ogata の方法の特徴は (55),(56) を iteration で解く代わりに，積分を計算してしまうところにある．上の例では，ABIC の J に関する微分が simulation without annealing を利用して求められるわけであるが，$J = 0$ のときの ABIC はすぐわかるので，1 次元の積分を行えば ABIC が求められる．

より一般には，(45) の「温度」T のような量をを導入して，それに関する微分量を simulation without annealing で計算し，1 次元積分を行うことになる．この際，$T = \infty$ での ABIC の値は簡単に求められるのが普通であるから，それを基準にとればよい．準安定状態が多くあるモデルでも「高温側」ではメトロポリス的モンテカルロ法の収束は良いので，その意味でも高温側に接続するのが望ましい．この方法は統計物理で **thermodynamic integration** といわれるものに対応している．Y.Ogata は温度の代わりに巨視的パラメータの共通尺度因子を使う方法も提案している (上記の例では温度と同じ)．

この方法の利点は，

- ABIC の値自体を計算することによって，異なるモデル族の優劣を判定できる可能性がある．
- 巨視的パラメータの空間で局所的極値につかまらないですむ．

などである．ただし，巨視的パラメータが多い場合には，探索すべき空間が大きくなりすぎるので適用できない．

Y.Ogata はこの方法で事前分布 (15),(16) を用いた画像処理の問題を扱っている．

> M.Kawato, T.Okamoto ら ATR 視聴覚研究所のグループは線過程を含んだ問題について同様の学習法を実験し，有望な結果を得ている由である (私信)．この場合，画像としては雑音を含んでいないものを与えるが，線過程についての教師は与えないのでやはり不完全データによる学習ということになる．
>
> J.York & J.Besag は統計数理研究所のシンポジウム (1989 Dec.) での発表で，事後分布による期待値のみを simulation without annealing で計算し，事前分布による期待値を近似的方法 (peudo likelihood 法) で計算する方法を提案した．例題は筆者のものと同じであったが，例を 1 つ示しただけで，答の分かっている場合のテストはやっていないようであった．
>
> これら以外にも，トモグラフィの問題などでは，似た方法で巨視的パラメータを推定した例があるらしい (未入手)．

2.3.2 完全データ (教師あり) の場合

完全データ (教師) として雑音を含まない微視的パラメータ $\{x_i^{true}\}$ が与えられている場合を考える．この場合，前節で扱った問題は，イジング模型 (48) で生成されたパターン (雑音なし) から巨視的パラメータ J を推定する問題になる．J についての尤度は (48) 式より，

$$L(J) = \pi_J(\{x_i^{true}\}) = \frac{\exp(\frac{1}{2}J\sum_{j\in N(i),i} x_i^{true}x_j^{true})}{\sum_{config.}\exp(\frac{1}{2}J\sum_{j\in N(i),i} x_i x_j)} \tag{59}$$

となり，(53) に対応する式は

$$\frac{1}{2}\sum_{j\in N(i),i} x_i^{true}x_j^{true} = \frac{\sum_{config.}(\frac{1}{2}\sum_{j\in N(i),i} x_i x_j)\exp(J\frac{1}{2}\sum_{j\in N(i),i} x_i x_j)}{\sum_{config.}\exp(J\frac{1}{2}\sum_{j\in N(i),i} x_i x_j)} \tag{60}$$

となる．これは 1 個の完全データしかない場合であるが，独立なデータが何個もある場合は，尤度は積 (相乗平均)，「エネルギー」は和 (相加平均) になることに注意して，それらをまとめて推定に使用してもよい (積分型の式を使えば，どれとどれをまとめるとよいかを AIC を用いて決められる)．

完全データの問題は，以下の 2 つの意味で不完全データの問題より簡単である．

まず，計算の大変な部分 (60の右辺) に，データが含まれていないということがある．従って，計算は「純粋系」の問題であって，必ずしもメトロポリス的モンテカルロ法によらずとも，いろいろな近似法が有効に使える．J.Besag による **peudo likelihood** 法はそのひとつである．

次に巨視的パラメータの空間での **convexity** の問題がある．ある関数が (弱い意味で)**convex** (凸) であるとは，定義域のすべての点でヘッセ行列の固有値が (零を含む) 定符号であることを意味する (図 15)．この場合，局所的極値はなく，ニュートン法や最急降下法のような方法で容易に大域的な極値に到達できる．

convexity は完全データ (教師あり) の場合すべてに成り立つわけではない．成り立つのは，たとえば $\{t_k\}$ は任意の関数とするとき，

$$L(\lambda) = \pi_\lambda(\{x_i^{true}\}) = \frac{1}{Z_\pi(\{\lambda_k\})}\exp(\sum_k \lambda_k \cdot t_k(\{x_i^{true}\}) - f(\{x_i^{true}\})) \tag{61}$$

と書ける場合である．ここで，$\{x_i^{true}\}$ は連続型もしくは離散型の完全データ，$\lambda = \{\lambda_k\}$ は巨視的パラメータである．

このような型の分布族を指数型分布族 (**exponential family**) という (普通は連続パラメータの場合をいうが，離散パラメータの場合はその極限と考えられる)．最隣接相互作用のイジング模型 (60) は指数型分布族に属する．また，最隣接以外の相互作用や 4 体の相互作用が入ったイジング模型も同様である．指数型分布族に 属さない 分布の例としては，コーシー分布やしきい値モデルがある (裾の広さやしきい値をパラメータと考えた場合)．

convexity の証明は簡単である (たとえば [P.Smolensky (1986)])．実際に $-\log L(\lambda)$ の 2 階微分を計算すると，

$$-\frac{\partial^2 \log L(\lambda)}{\partial\lambda_i\partial\lambda_j} = \langle (t_i - \langle t_i\rangle_\pi)(t_j - \langle t_j\rangle_\pi)\rangle_\pi \tag{62}$$

となり，この “分散共分散型” の行列が非負対称であることはすぐに分かる．ここで，$\langle\ \rangle_\pi$ は分布 π による期待値を意味し，たとえば，

$$\langle t_i\rangle_\pi = \int t_i(\{x_k\})\pi_\lambda(\{x_k\})d\{x_k\} \tag{63}$$

等である．

> ATR のグループは，線過程を含む問題 (顔の輪郭の抽出) の教師ありの場合を研究した [ ]．
>
> Y.Ogata と M.Tanemura は [Y.Ogata, M.Tanemura (1984,1986)] において，メトロポリス的モンテカルロ法を用いて，完全データからの巨視的パラメータの推定を行っている．対象となっている問題はマルコフ場でなく空間点過程の問題で，計算の内容は「液体」の simulation に似ている． A.Penttinen の学位論文 [A.Penttinen (1984)] もこれと類似の内容を含んでいるが，メトロポリス的モンテカルロ法で Hessian まで計算し，ニュートン法で巨視的パラメータを求めるという大胆な試みが行われている．

2.4 他の統計物理の手法の適用

本節ではメトロポリス的モンテカルロ法以外の統計物理の手法を統計的情報処理に導入することについて論じる．この場合，たとえば事前分布の規格化定数を求める問題や完全データ (教師あり) の問題のような「純粋系」の問題は割合簡単であるが，事後分布での期待値の計算や ABIC の計算は非一様な系を扱うため面倒である．本節では主に後者を考える．

統計物理で有効性が知られている手法が，工学でも独立に発明されていることはしばしばある．たとえば，非線型フィルターとして使われているものの発想は，前にも触れたが，転送積分法 (転送行列法) と同じである．

- マルチスケールのアルゴリズムとくりこみ群

  広い意味でフラクタル的な考え，あるいは，さまざまな尺度でものを見ることは，画像問題ではとくに重要と思われる．

統計物理では，臨界点の性質をフラクタル的な観点から近似的に調べる方法としてくりこみ群の方法が使われている．これをマルコフ場による画像処理に応用した研究として，[B.Gidas (1989)] がある．実空間くりこみ群の手法が用いられている．

なお，統計物理に由来するものではないが，最適化手法にこのような考えを取り入れたものとして，**multi-grid** 法がある [J.Walsh(), D.Terzopoulos ()]．これは元来，線型計算の方法として発達した方法であるが，最近では非線型版やメトロポリス的モンテカルロ法版 ([J.Goodman, A.D.Sokal (1986), D.Kandal, E.Domany, D.Ron, A.Brandt, E.Loh (1988)]) もある．また，統計物理の問題にも応用されている ([R.G.Edwards, J.Goodman, A.D.Sokal (1988)])．

最適化の困難を増す要因としては，

- 目的関数の方向による曲率の違い (“渓谷” の存在)．
- 局所的な極値の存在

の 2 つがある．ニュートン法は前者を simulated annealing 法は後者を重視した解法である．両者に同時に対応するのは難しいが，multi-grid 法は画像問題の特徴を生かしているので，ある程度両方に対応できるのではないかと期待されている．

また，心理実験によれば，人間の視覚処理もいくつかの異なった尺度 (粗視化レベル) で行われているといわれており，それとの比較も興味深い．

- 平均場近似

平均場近似 (分子場近似，**mean field approximation**) はスピン系の統計物理でもっとも基本的な近似法であり，非一様な系にも容易に適用できる．平均場近似の説明は統計物理の解説及び本文に譲るが，簡単にいえば，1 体の周辺分布 (たとえば “ある画素が黒い確率”) についての自己無撞着な式を解く方法ということになる．筆者はこれをイジング的な事前分布の場合に適用し，詳しく調べた．この方法の発展としては，クラスター近似的な拡張や非線型フィルター (転送行列法) と組み合わせた方法なども考えられる．

筆者は 1987 年ごろから，物理学会等で平均場近似を画像処理に適用するという考えを発表してきたが，最近になって以下の研究を知った．

1. イジング型神経回路の平均場近似

   C.Peterson らはイジング型神経回路網に平均場近似を応用して，いろいろな課題の学習に成功している [C.Peterson, J.R.Anderson (1987)，C.Peterson, E.Hartman (1989)]．平均場近似の結果，シグモイド的な応答をする要素 (“tanh 要素”) からなるネットワークが生ずるが，これは recurrent になっている点で通常の backpropagation-feedfoward 型の神経回路網と異なっている．
   MIT の人 (Geisel?) が，上記の仕事を画像問題に応用する研究をしているらしい (未発表?)．平均場法によって，C.Koch らの研究 ((17,18) のような事前分布を用いた画像処理問題を Hopfield-Tank 法によって扱ったもの．[C.Koch, J.Marroquin, A.Yuille (1986)]) が再解釈できることが示唆されている由である．

2. Relaxation Labelling 法の “統計学的解釈”

   relaxation labelling 法などの名で呼ばれる協調的な画像解釈のアルゴリズムが 1970 年代から存在する [A.Rosenfeld, R.A.Hummel, S.W.Zucker (1976), A.Lev, S.W.Zucker, A.Rosenfeld (1977), R.A.Hummel, S.W.Zucker (1983)]．
   この方法は，局所的な確率が自己無矛盾になるように繰り返し法で解くもので，その意味では平均場法と類似しているが，データが初期条件として与えられるだけで，それ以後は参照されないという点で大きな違いがある．(なお，[J.Marroquin (1985)] の周辺分布に対する “deterministic approximation” もこの点では，relaxation labelling 法に似ている)．
   これを改良して統計モデルの近似という形に書き直すという経路で平均場近似に到達することも可能なはずで，そういう仕事がないかどうか注意していたのであるが，この原稿の執筆中に，[J.M.Kay D.M.Titterington (1986)] という論文を見つけた．これには，まさに上記のことがなされており，平均場法と等価な形式が与えられている．
   これにはまた，このような “データを保持する” 型の relaxation labelling 法についての研究として，[N.L.Hjort, E.Mohn (1985)] が引用してある．この論文は未入手だが，Hjort らの方法自体は平均場法と少し違うようである．

3. Iterated Conditional Expectation

統計数理研究所のシンポジウム (1989 Dec.) で J.York の予稿から得た情報によれば，A.Owen [A.Owen (1989?)] が上記の研究をしているそうである．題名から推察すると平均場近似に関係がありそうだが，論文未入手のため詳細は不明である．

3　付録 A　情報量規準と統計学　(未完)

3.1　モデル選択の規準

3.2　AIC の擬ベイズ的解釈

4　付録 B　イジング型神経回路網と対数線型模型　(未完)

表 1: 用語の対照表

this paper	Akaike	Geman&Geman	non Gaussian filter
microscopic parameter	parameter	pixel variable label variable etc.	state vector
macroscopic parameter	hyper parameter	parameter	parameter
MABICE	MABICE	MLE(modified EM algorithm)	MLE

MLE = Maximum Likelihood Estimate

MABICE = Minimum ABIC Estimate