マルコフ連鎖モンテカルロ法とその統計学への応用

統計数理研究所　伊庭幸人

ver.3.4

本稿は、マルコフ連鎖モンテカルロ法の一般論を中心とした第 1 部と特定の事例についてやや詳しく解説した第 2 部からなっている。第 1 部は統計科学合同研究集会 (1994.4) の予稿を元にして、大幅に加筆変更し、統計への応用に関する文献とその解説を追加したものである。マルコフ連鎖の基礎に関連する事項と実装上の注意事項をそれぞれ付録として加えた。第 2 部では、分類の問題に関する未発表の研究を材料として、実際の統計の問題でマルコフ連鎖モンテカルロ法がどう使われるかを解説した。

第 1 章

第 1 部: 概説

マルコフ連鎖であらわされるダイナミクスを利用して、与えられた確率分布からのサンプルを得るアルゴリズムをマルコフ連鎖モンテカルロ法と呼ぶ。これは統計物理で 40 年にわたって重要な役割を果たしてきたが、近年になって統計学への応用が注目されている。統計学への応用に関しては統計数理研究所の尾形・種村らの先駆的な仕事があり、工学関係では組合せ最適化の手法 (simulated annealing 法) との関連でしばらく前から話題になっていたが、ベイズ統計を中心に広い範囲の統計の問題に応用されるようになったのはここ数年である。

マルコフ連鎖モンテカルロ法という名称についてであるが、物性物理の分野では、単にモンテカルロ法というとほとんどマルコフ連鎖を用いる方法を意味する。とくに区別する場合は、動的モンテカルロ法 というようだが、それほど一般的な呼称ではない。統計学や工学では他の種類のモンテカルロ法が広く使われているので、区別するための名称が必要である。最近、統計学の分野ではマルコフ連鎖モンテカルロ法 (**Markov chain Monte Carlo algorithms** , $(\mathbf{MC})^2$) という名称が定着してきているので、本稿ではそれを用いる。空間・画像の モデル としてのマルコフ場モデルと混同しないよう注意して欲しい。神経回路網関係では、ボルツマンマシンという用語があるが、この言葉は、モデルの名前なのか、学習法 (推定法) の名前なのか、マルコフ連鎖モンテカルロ法の工学的応用全体を指すのか、混乱しているようである。できれば使わない方が良いだろう。

1.1 マルコフ連鎖の作り方 (一般論)

以下では、確率分布からのサンプルを得るための手段として、定常分布が与えられた分布になるマルコフ連鎖をシミュレートすることを考える。このマルコフ連鎖の “時間” は本来仮想的なものであるが、統計物理では、実際の時間と対応させて考えることもある。この場合、マルコフ連鎖による状態遷移は物理系のダイナミクスに相当するわけである。統計学への応用ではこのような対応は意味を持たないが、 “ダイナミクス” というイメージ自体は有用と思われるので、本稿でも “マルコフ連鎖のダイナミクス” のような表現を用いることにする。

まず、 このようなマルコフ連鎖を作る一般的な方法を与えよう。定常分布が $\{P_\alpha^\infty\}$ になるようなマルコフ連鎖を作るには、状態 α から γ への遷移確率 $\{W(\alpha \to \gamma)\}$ を次の条件 (詳細釣合条件) をみたすように定めればよい。

$$P_\alpha^\infty W(\alpha \to \gamma) = P_\gamma^\infty W(\gamma \to \alpha) \tag{1.1}$$

有限系の場合、 (1.1) のほかに、 $W(\alpha \to \gamma) \neq 0$ であるような (α, γ) を結んだグラフが連結であるなど、いくつかの条件がみたされれば、定常分布が唯一に定まり、 $\{P_\alpha^\infty\}$ に一致することが示せる。

(1.1) 式は、

$$\frac{P_\alpha^\infty}{P_\gamma^\infty} = \frac{W(\gamma \to \alpha)}{W(\alpha \to \gamma)} \tag{1.2}$$

とも書けるが、これだけからは $\{W(\alpha \to \gamma)\}$ は一意には定まらない。そこで、可能な W の中から自然なものを選ぶわけであるが、状態を定義する変数のうち同時に変化するのが 1 つもしくは少数のみと仮定するのが普通である。 W でいえば、 $W(\alpha \to \gamma)$ のうちほとんどを零とし、状態 α と γ が “近い” ごく少数のもののみを非零とするわけである。具体的な例はあとに示す。このようなマルコフ連鎖を本稿では “局所的なダイナミクス” と呼ぶことにする。

(1.2) 式の左辺が比の形になっているために、遷移確率の表式を得るためには分布が規格化されている必要がないことを注意しておく。統計物理やベイズ統計に現れる分布では、規格化定数が解析的に計算できない場合も多いので、この点は非常に重要である。

1.1節及び 1.2節では、正確な数学的取り扱いよりも、概観を与えることを目標とする。詳細は付録 Aで補うが、厳密かつ一般的な結果については Geyer(1992), Tierney(1994), Winkler(1995) などを参照されたい。

以下では、連続時間のマルコフ過程を考え、 W を単位時間の遷移確率とする。 行列 L を

$$L_{\alpha\gamma} = \quad W(\gamma \to \alpha) \qquad (\alpha \neq \gamma) \tag{1.3}$$

$$\quad -\sum\nolimits_{\tilde{\gamma}\neq\alpha} W(\alpha \to \tilde{\gamma}) \quad (\alpha = \gamma) \tag{1.4}$$

で定義すると、適当な初期分布 $P_\alpha(0)$ から出発して、 $t > 0$ 時間後に状態 α にいる確率 $P(t) = \{P_\alpha(t)\}$ は方程式

$$\frac{dP(t)}{dt} = LP(t) \tag{1.5}$$

を満たす。この解は L の固有値 $\{\lambda_i\}$ と対応する固有ベクトル $\{V_i\}$ によって定まり、

$$P_\alpha(t) = P_\alpha^\infty + \sum_{i=1,2..} a_i V_i^\alpha \exp(\lambda_i t) \tag{1.6}$$

となる。 $\{a_i\}$ は初期分布 $P_\alpha(0)$ に依存する定数、 V_i^α は i 番目の固有ベクトル V_i の第 α 成分を意味する (V_i の次元 (成分数) は系のとりうる状態の総数に等しいことに注意)。

$\{\lambda_i\}$ のなかで最大のものは分布 $P^\infty = \{P_\alpha^\infty\}$ を固有ベクトルとして持つ $\lambda_0 = 0$ である。また、詳細釣合条件を考慮すると、 λ_0 以外の $\{\lambda_i\}$ はすべて負の実数であることがわかる (付録 Aの A.1節)。そこで、緩和時間スペクトル $\{\tau_i\} = \{-1/\lambda_i\}$ を定義すると、 $\{\tau_i\}$ はすべて正の実数で、そのうち最大のもの τ_1 が系が初期状態の記憶を失う (緩和する) までの時間を決める。以下では統計物理での用語法に従って、 τ_1 (場合によっては各 τ_i) を緩和時間 (relaxation time) と呼ぶ。また、 τ_i に対応する固有ベクトル V_i を i 番目の緩和モードと呼ぶことにする。実際の計算は離散時間のマルコフ連鎖に相当するため、上記とはいくつか違う点があるが、ことの本質はあまり変わらない (付録 Aの A.2節)。

初期状態の記憶を失うという意味は、多数の系を用意して、各々を独立に、与えられたダイナミクスに従って発展させたときに、 $\{\tau_1\} = \{-1/\lambda_1\}$ より十分長い時間がたてば、多数の系の分布が P_α^∞ に近付くということである。このことは、系の状態がマルコフ連鎖に従って変化するとき、初期状態から τ_1 以上時間がたったのちに、 τ_1 より十分大きい時間間隔でサンプルをぬきだせば、これらは定常分布からランダムに選んだものと見なせることを示唆している。

(1.6) 式は、 1 個の系を追いかけた場合に、出発した状態から τ_1 以上の時間がたてば、そこで 1 回だけ観測した統計量がその期待値の良い近似になるという意味では ない。これが成り立つのは、系が互いに独立とみなせる多くの部分からなりたっており、測定する量が各部分に対応する量の平均値であるような場合であるが、統計物理の場合はともかく、われわれの場合にはこの条件は満たされないと考えた方が良い。 “緩和” を最適化の場合の “収束” と混同してはならない。

普通、確率変数の数が大きくなると、系のとりうる状態の数はその指数関数で増えるが、局所的ダイナミクスのもとでの緩和時間は指数関数的に増えるとは限らない。たとえば、格子に確率変数が配置されたモデルの場合、ある程度以上はなれた部分の相関は指数関数的に失われることが多い。このような場合、格子の 1 辺の大きさ L がある距離以上になると、 τ_1 は一定値に収束すると考えられる。構成要素同士が非局所的に結合されているモデルの場合も、要素間の高次の相関が小さければ、緩和時間は系のとりうる状態の総数よりはるかに短くなりうる。

問題なのは、緩和時間の推定が (事前にも、事後的にも) 難しいことである。場合によっては、局所的で自然なダイナミクスのもとで緩和時間はきわめて長くなりうる。そのような場合には、全く間違った結果に導かれることもある。

従来から統計学などで使われてきたモンテカルロ法では、分布を大域的に近似するような重み関数を与えてサンプル生成を行ない、それを利用して必要な値を計算する。この方法では、問題が大規模になると、効果的な重み関数を与えるのがきわめて難しくなる。たとえば、確率変数の数が 1 個のとき $(1-\varepsilon)$ 位の近似度の場合、 1000 個の場合にそのまま拡張すれば $(1-\varepsilon)^{1000}$ の程度の近似度になる恐れがある。これが統計物理の分野でマルコフ連鎖モンテカルロ法が愛用されてきた主な理由である。確率変数の数が少なく、目的の分布が適当な変数変換によってガウス分布などで良く近似されるなら、従来の方法を使うべきである。

なお、マルコフ連鎖モンテカルロ法における仮想的時間について考える際には、 2 通りの単位 — マルコフ連鎖のステップ数を 確率変数の個数 で割ったものと割らないもの — があることに注意しなくてはならない。統計物理では、前者をモンテカルロステップ数 (MCS 数) と呼び、後者をフリップ数 (スピン・フリップ数) として区別している。上で τ_1 が一定値に収束すると述べたのは前者の単位ではかった場合である。

1.2 マルコフ連鎖を利用した統計量の計算

具体的なマルコフ連鎖の作り方 (メトロポリス法・ Gibbs sampler) については次の節で述べることにして、まず、ここでは、作ったマルコフ連鎖の使い方を述べる。

緩和時間より十分長い間隔でマルコフ連鎖から抜き出したサンプルが、与えられた分布からのランダムサンプルと見なせることを利用すると、任意の統計量 $A(\alpha)$ に対して、与えられた分布 P_α^∞ での $A(\alpha)$ の期待値

$$\langle A(\alpha)\rangle = \sum_\alpha A(\alpha) P_\alpha^\infty \tag{1.7}$$

あるいは

$$\langle A(\alpha)\rangle = \int A(\alpha) P_\alpha^\infty d\alpha \tag{1.8}$$

は、次のように書ける。

$$\langle A(\alpha)\rangle = \lim_{M\to\infty} \frac{1}{M} \sum_{m=1}^{M} A(\alpha_m) \tag{1.9}$$

ここで、 $\{\alpha_m\}$ はマルコフ連鎖から間隔 δ で抜き出したサンプル列である。

抜き出したサンプルの間隔 δ がマルコフ連鎖の緩和時間より長いという条件は 不必要 であることに注意しよう。これは次のように考えるとわかる。サンプル間隔が緩和時間より短い長さ M の列を考える。この列は、緩和時間より長い間隔 (R とする) ごとにとびとびにサンプルしなおすことで、長さ M/R の R 個の列に分割することができる。M/R が十分大きければ、この部分列のそれぞれについて統計量の平均値が求める期待値に収束するので、全体の平均もまた正しい値に収束する。

実は、サンプル間隔 δ を長くとることは 有害 である。間隔 δ を大きくすると (1.9) の右辺の分散は増大する (付録 Aの A.5節)。したがって、統計量の計算にかかる手間が過大にならない範囲で、なるべく頻繁にサンプルをとるのが良い。サンプル間隔が短い場合、統計量 (たとえば分布の分散) の不偏推定値を計算するためには、サンプル間の相関についての配慮が必要である (Kikuchi et al.(1994): ただし、これが問題になるのはかなり精密な解析をする場合だろう)。

平均 (1.9) の計算のためのシミュレーションの長さ $M\delta$ は最長の緩和時間 τ_1 に対して十分長いこと — たとえば 100 倍とか 10 万倍とか — が望ましい。正確には、計算したい統計量 A の相関が無視できるようになる時間 $\tilde{\tau}_A$ より十分長ければ良いわけだが、最長の緩和時間 τ_1 を目安にするのが一般には安全であろう (τ_1 と $\tilde{\tau}_A$ の関係は付録 Aの A.3節を参照)。そこで、まず行なうべきことは、 τ_1 ないし $\tilde{\tau}_A$ の上限を評価すること ("診断") であるが、これは一筋縄ではいかない場合もある。診断のために知っていると役立つことは 1.4節にまとめたが、基本は初期状態・乱数・シミュレーションの長さを変えて何回も試してみることに尽きる。

(1.9) の右辺の分散 σ_{MC}^2 を見積るにもこの方針で臨むのが簡単である。長さ M の計算を $\widehat{M}$ 回行なって、

$$\sigma_{MC}^2 \simeq \frac{1}{\widehat{M}-1} \sum_{\widehat{m}=1}^{\widehat{M}} \Big(\frac{1}{M}\sum_{m=1}^{M} A(\alpha_m^{\widehat{m}}) - \langle A(\alpha)\rangle\Big)^2 \tag{1.10}$$

とするわけである。ここで $\alpha_m^{\widehat{m}}$ は $\widehat{m}$ 回目の計算の m 番目のサンプルの意味である。また、 $\langle A(\alpha)\rangle$ は $M\widehat{M}$ 個のサンプルを全部を使った平均で置き換えるものとする。 (1.10) が成り立つには $M \gg \tilde{\tau}_A$ である必要がある。これが満たされない場合、 M を大きくするにつれて右辺は大きくなる (逆にこれを $\tilde{\tau}_A$ の見積りに使うこともできる)。 σ_{MC}^2 の推定についてのより洗練された議論は Geyer (1992) にある。分散 σ_{MC}^2 は各緩和モードからの寄与の和で表すことができる (付録 Aの A.4節) が、このことは定量的な見積りのためには直接役立たない。

実際の計算にあたっては、初期状態の特異性によるバイアスを除くために生成されるサンプル列の最初の部分を十分 "捨てる"(統計量の測定に使用しない) ことが重要である。これは、 (1.10) 式による誤差推定の場合も同じである。捨てる列の長さは系の最長の緩和時間 τ_1 より十分長いことが望ましいが、具体的な例でどの程度捨てるのがよいのかは難しい問題である。筆者は全体の $1/2$ とか $1/4$ とか捨てることが多いが、はっきりした根拠があってそうしているわけではない。理論的にいえば、最初の部分を多く捨てるのは、初期状態が P^∞ からの "典型的なサンプル" ではないために式 (1.6) の a_1 や a_2 が大きい値をとる可能性があるためである (付録 Aの A.6節)。サンプル間隔 δ を大きくすると分散が増大するというのは、 A.3 – A.5節の証明を見ればわかる通り、最初の部分を十分長く捨てた後の話である。

マルコフ連鎖モンテカルロ法は、基本的に期待値・周辺分布の計算法であって、積分の計算法ではない。したがって、たとえば、分布の規格化定数そのものを直接計算することはできない。積分そのものがどうしても必要な場合

は、求める積分の適当な外部変数 λ による対数微分を計算し、その情報を利用して 1 次元積分に帰着させるのが最も簡単である。すなわち、規格化されていない分布関数を $\widetilde{P}(\alpha)$ とするとき、

$$\frac{\partial}{\partial\lambda}\log\int\widetilde{P}(\alpha)d\alpha=\frac{\int(\frac{1}{\widetilde{P}(\alpha)}\frac{\partial\widetilde{P}(\alpha)}{\partial\lambda})\widetilde{P}(\alpha)d\alpha}{\int\widetilde{P}(\alpha)d\alpha}=\langle\frac{1}{\widetilde{P}(\alpha)}\frac{\partial\widetilde{P}(\alpha)}{\partial\lambda}\rangle \tag{1.11}$$

と期待値の形に書いて、これを積分することで、もとの積分 (の対数) を計算するわけである。 λ としてはいろいろな選び方が考えられるが、たとえば後出の “温度”T でも良い (Binder and Stauffer(1984))。このタイプの方法は Ogata (1990,1991) によって統計の問題に応用された。より進んだ積分の計算法としては、拡張されたアンサンブルに基づく諸手法の適用が考えられる (文献は 1.4.1節を参照されたい)。

ちなみに、$\langle\widetilde{P}(\alpha)^{-1}\rangle=\int d\alpha/\int\widetilde{P}(\alpha)d\alpha$ を利用して積分 $\int\widetilde{P}(\alpha)d\alpha$ を計算しようとしてもうまくいかない。大規模問題では、統計量 $\widetilde{P}(\alpha)^{-1}$ の母分散が非常に大きくなるので、左辺をマルコフ連鎖モンテカルロ法で計算しようとすると極度に収束が悪くなる。この場合、緩和時間の長短に関係なく駄目なことに注意されたい。

1.3 マルコフ連鎖の作り方 (通常使われる方法)

ここではマルコフ連鎖の作り方について簡単な方法を 2 種類説明し、具体的な確率分布に対する適用例を示す。また、 2 種の方法の両方を含む一般形も与える。

1.3.1 メトロポリス法

一般に、与えられた分布を $P(\alpha)$ とする (これからは肩の ∞ は省略する)。以下では、 $\alpha=\{\alpha_i\}(i=1..I)$ というように状態がベクトル的にあらわせる場合を考える。 $\{\alpha_i\}$ が離散変数の場合、 (狭義の) メトロポリス法の基本的なアルゴリズムは、任意の初期状態から出発して以下の手順を繰り返すことで得られる。

1. “動かすことを試みる変数”α_j を決める。

   最も基本的な方法は、乱数を用いて j を選ぶことである。任意の i について選ばれる確率が零でなければ、選び方が等確率である必要はない。但し、この確率を計算の途中で変更してはならない。

2. 変数 α_j をどう変更するか、 “動く先の候補”γ_j を決める。

   候補の選び方は等確率にする。すなわち、 α_j のとりうる値を $\{g_j^k,k=1..K_j\}$ とするとき、乱数を用いて確率 $1/(K_j-1)$ で $\gamma_j=g_j^k\,(g_j^k\neq\alpha_j)$ とする。

3. もとの状態 $\alpha=\{\alpha_1..\alpha_j..\alpha_I\}$ と動く先 $\gamma=\{\alpha_1..\gamma_j..\alpha_I\}$ について、

   $$r=P(\gamma)/P(\alpha) \tag{1.12}$$

   を計算する。

4. (0,1) の一様乱数 RND を発生し、 $RND<r$ なら、 α_j を γ_j で置き換える。

このアルゴリズムが詳細釣合条件を満たすことは、容易に示せる。なお、ステップ (1) で乱数を用いずに一定の順番で j を選んでも詳細釣合条件は満たされる。このことは、複数のマルコフ連鎖の合成という視点からみると理解しやすい。各 $i\,(i=1..I)$ について、 α_i のみを動かす遷移を考え、全体のマルコフ連鎖をそれら I 個の合成と見るわけである。同様に考えると、正方格子上のイジング模型 (1.3.3節) の場合、格子を 2 色の市松模様に塗り分けて、ひとつの色の格子点上の変数を一通り動かしたのち、別の色の格子点にうつる方法 (checkerboard updation) が可能である。このような手法でアルゴリズムの並列化が可能になる。

$\{\alpha_i\}$ が実数値をとる変数の場合は、次のようにすればよい。以下では、与えられた確率分布を $P(\alpha)d\alpha$、 $\alpha\in(-\infty,\infty)$ とする。

1. 離散変数の場合と同様に、適当な方法で j を選ぶ。

2. $\triangle\alpha$ を $(-\epsilon_j,\epsilon_j)$ の一様乱数として、 $\gamma_j=\alpha_j+\triangle\alpha$ とする。

3. もとの状態 $\alpha = \{\alpha_1..\alpha_j..\alpha_I\}$ と動く先 $\gamma = \{\alpha_1..\gamma_j..\alpha_I\}$ について、

$$r = P(\gamma)/P(\alpha) \tag{1.13}$$

を計算する。

4. (0,1) の一様乱数 RND を発生し、 $RND < r$ なら、 α_j を γ_j で置き換える。

$\{\epsilon_i\}$ の大きさは、緩和時間を決める重要な要素である。経験的には、ステップ (4) で試行が受理される ("動く") 割合 (acceptance ratio) が５０％前後となるようにするのが良いといわれている。これは経験則であり、条件付き分布の裾が長い場合など、特別な場合に当てはまるかどうかは疑問である。ただ、 acceptance ratio が大きすぎても緩和が遅くなることは覚えておく必要がある。

上の説明では、各成分 α_i が少数個の状態をとる離散変数あるいは 1 個の実数の場合を考えた。このような変数を同時に沢山変更するような拡張は形式的には容易に考えられるが、 acceptance ratio が減少するので、一般には必ずしも有利とは限らない。

メトロポリス法で変数の動く範囲が不等号で制約されている場合は誤りやすい。もっとも簡単な方法は、制約のない場合のマルコフ連鎖を構成した上で、単純に "はみ出す場合は動かさない" とすることである。制約のない場合に詳細釣合条件がみたされていれば、こうして作ったマルコフ連鎖も詳細釣合条件をみたす。

1.3.2 Gibbs sampler(熱浴法)

Gibbs sampler というのは統計学者の用語である。物理でいう熱浴法 (heat bath algorithm) とほぼ対応するように思われる。 Gibbs sampler のアルゴリズムは、 $\{\alpha_i\}$ が離散変数の場合、任意の初期状態から出発して以下の手順を繰り返すことで得られる。

1. 適当な方法で、 "動かす変数"α_j を決める。

2. 変数 α_j の "とりうる値" にそれぞれ確率を割り当てる。

   α_j のとりうる K_j 個の値を $\{g_j^k, k = 1..K_j\}$ とする。着目している以外の成分 $\alpha_i (i \neq j)$ はそのままで、 j 番目の成分 α_j が値 g_j^k をとった状態を、 $\gamma^k = \{\alpha_1..g_j^k..\alpha_I\}$ とするとき、確率 q^k を

$$q^k = \frac{P(\gamma^k)}{\sum_{\widetilde{k}} P(\gamma^{\widetilde{k}})} \qquad (k = 1..K_j) \tag{1.14}$$

   で計算する。

3. 乱数を用いて、 α_j を確率 q^k で状態 g_j^k に遷移させる。

   アルゴリズム的には、

   - $Q_k = \sum_{\widetilde{k}=1}^{k} q^{\widetilde{k}}$ $(k = 1..K_j)$、 $Q_0 = 0$ とする。
   - (0,1) の一様乱数 RND を発生する。
   - $Q_{k-1} \leq RND < Q_k$ なら、 α_j を g_j^k で置き換える。

   とすれば良い。

ステップ (2),(3) は、着目している成分以外を固定した条件つき分布 $P(\gamma_j \mid \alpha_1..\alpha_{j-1}, \alpha_{j+1}..\alpha_I)$ から γ_j をサンプルする ("取り直す") ことを意味する。メトロポリス法の場合と異なって、新しい α_j がどうなるかは、以前の α_j の状態に依存しない。今度も、詳細釣合条件が満たされることは簡単に確かめられるが、求める分布が不変なことを直接理解した方が早いかもしれない。ステップ (1) での変数の選び方は、メトロポリス法の場合よりも一般化できる (Geman,S. and Geman,D. (1984), 定理 A) が、通常はメトロポリス法と同じ範囲で考えて良いだろう。

同時に変更する変数の数を多くすると緩和は速くなるが、一般には条件つき確率の計算が急激に困難になるので、能率は必ずしも良くならない。 Gibbs Sampler を連続変数のモデルに適用した場合、一般にはステップ (2) の条件つき確率の計算が面倒であるが、分布の形が特別なために簡単になることもある。離散変数と連続変数の両方を含んでいて、離散変数を固定したときに連続変数の部分が多次元ガウス分布になるモデルはその例である。また、ある分布

の分散や平均値などがまた確率的に分布しているモデル (いわゆる混合分布モデル) のなかにも Gibbs sampler が簡単にインプリメントできるものがある (Gelfand et al.(1990), Gelfand and Smith (1990))。統計学への応用ではこのような例がしばしば見られることが、統計学で Gibbs sampler がメトロポリス法より先に注意を集めたひとつの理由であろう。

1.3.3 例: イジング模型

多数の 2 状態系 $\{S_i\}$ が大きさ $L \times L$ の正方格子上に配列している系を考える (図 1.1)。 2 つの状態のどちらにいるかを 2 値変数 $S = \pm 1$ であらわすことにする。このような 2 状態系のことをイジングスピン (Ising spin) と呼ぶことがある。 i でラベルされた格子点の上に、 $S_i = +1$ ならば黒い碁石が、 $S_i = -1$ の時白い碁石が、それぞれのっていると考えればよい。 2 状態系の間に相互作用が無ければ、この系は単に多数のイジングスピンをよせ集めたものである。そこで、スピン間に相互作用のあるモデルのうち最も簡単なものとして、各スピンが隣接する 4 個とのみ相互作用し、相互作用の形が格子上の位置によらないモデルを考える。

図 1.1: 正方格子上のイジングスピン

確率分布を式で書くと、

$$P(\{S_i\}) = \frac{1}{Z} \exp(J \sum_{nn} S_i S_j) \tag{1.15}$$

$$Z = \sum_{config.} \exp(J \sum_{nn} S_i S_j) \tag{1.16}$$

となる。ここで、 $\sum_{config.}$ は可能な 2^{L^2} 個の $\{S_i\}$ に関する和を意味するものとする。 $\sum_{nn}$ は正方格子上で隣接する (i, j) 対に関する和を表わす。 J は結合の強さを表す定数である。このモデルをイジング模型、正確には 2 次元正方格子上の最隣接相互作用のイジング模型と呼ぶ。モデルを完全に定義するためには、格子の境界での扱いを決めなくてはならないが、以下では第 1 部を通して、格子は周期境界条件もしくは自由境界条件になっていると仮定する。

イジング模型は、統計学でいうマルコフ場モデル (Besag(1974)) の最も単純なものである。イジング模型に関する文献は極めて多数あるが、ここでは、磁性体理論に関する教科書 (小口 (1970)) とマルコフ連鎖モンテカルロ法の適用に関する一論文 (Miyashita and Takano (1985)) を参照するにとどめる。 また、このモデルにはいろいろなバリエーションがあって、どこまでをイジング模型と呼ぶのかは必ずしもはっきりしないが、広義のイジング模型と統計学でいう auto-logistic model はほぼ同義ではないかと思われる。

イジング模型の分布関数は、

$$E(\{S_i\}) = -J \sum_{nn} S_i S_j \tag{1.17}$$

$$P(\{S_i\}) = \frac{1}{Z} \exp(-E(\{S_i\})) \tag{1.18}$$

$$Z = \sum_{config.} \exp(-E(\{S_i\})) \tag{1.19}$$

とも書ける。統計物理では上のように書いた場合 (あるいは、これに “温度”T(1.5節) を入れた場合) に “分布 (1.18) は ‘エネルギー’(1.17) に対する ‘ギブス分布’ である” という。 “エネルギー” という特別な関数や “ギブス分布” という特別な確率分布があるわけではないので注意されたい。規格化定数 Z は分配関数と呼ばれる。

- イジング模型に対するメトロポリス法

  まず $\{S_i\}$ の初期状態をセットする。

  1. 乱数を用いてスピン S_j を選ぶ。
  2. 選んだスピン S_j が符号を変えたとき減るエネルギー

  $$\triangle E = -2JS_j^{\alpha} \sum_{i \in nn(j)} S_i^{\alpha} \tag{1.20}$$

  を計算する。ただし、 $nn(j)$ は j の 4 個の隣接格子点を表わす。

3. (0,1) の一様乱数 RND を発生し、

(1.21) $$RND < \exp(\triangle E)$$

なら、注目しているスピンの符号を反転する。

1.- 3. を繰り返す。

- イジング模型に対する Gibbs sampler

まず $\{S_i\}$ の初期状態をセットする。

1. 乱数を用いてスピン S_j を選ぶ。
2. 選んだスピン S_j が $+1$ のときと -1 のときの系のエネルギー E_+,E_- をそれぞれ計算する。
3. (0,1) の一様乱数 RND を発生し、

(1.22) $$RND < \frac{\exp(-E_+)}{\exp(-E_+)+\exp(-E_-)}$$

なら、注目しているスピンを $+1$ で、そうでなければ -1 で置き換える。

1.- 3. を繰り返す。

実際の計算では、

(1.23) $$\frac{\exp(-E_+)}{\exp(-E_+)+\exp(-E_-)} = \frac{1}{1+\exp(\widetilde{\triangle E})} = \frac{1}{2}(1-\tanh(\widetilde{\triangle E}/2))$$

(1.24) $$\widetilde{\triangle E} = E_+ - E_- = -2J\sum_{i\in nn(j)} S_i^\alpha$$

を利用すればメトロポリス法と同じ程度の計算量で済む。

今のモデルでは $\exp(\triangle E)$,$\exp(\widetilde{\triangle E})$ として可能な値は限られているから、これを表にして記憶しておけば、計算量を大幅に節約できることを注意しておく。その他の実装上の問題については付録 Bを参照されたい。

1.3.4 一般形: Metropolis-Hastings 法

Gibbs sampler で条件つき確率の計算が大変な場合のひとつの方策は、近似的な条件つき確率と棄却手続きを導入することである。これをもう少し一般化すると、メトロポリス法と Gibbs sampler の両方を含む形のアルゴリズムを書くことができる。これは Metropolis-Hastings 法と呼ばれることがある。

アルゴリズムは、任意の初期状態から出発して以下の手順を繰り返すことで得られる。 γ_j に関する確率分布 $h_j^\alpha(\gamma_j)$ は、成分の番号 j と状態 $\alpha = \{\alpha_1..\alpha_j..\alpha_I\}$ に依存してよい。これは、最初に決めたら、あとは変えないものとする。

1. “動かすことを試みる変数”α_j を決める。

   最も基本的な方法は、乱数を用いて j を選ぶことである。任意の i について選ばれる確率が零でなければ、選び方が等確率である必要はない。但し、この確率を計算の途中で変更してはならない。

2. 変数 α_j をどう変更するか、 “動く先の候補”γ_j を決める。

   確率 $h_j^\alpha(\gamma_j)$ に従って、乱数を用いて候補 γ_j を選ぶ。

3. もとの状態 $\alpha = \{\alpha_1..\alpha_j..\alpha_I\}$ と動く先 $\gamma = \{\alpha_1..\gamma_j..\alpha_I\}$ について、

(1.25) $$r = \frac{P(\gamma)h_j^\gamma(\alpha_j)}{P(\alpha)h_j^\alpha(\gamma_j)}$$

を計算する。

4. (0,1) の一様乱数 RND を発生し、 $RND < r$ なら、 α_j を γ_j で置き換える。

メトロポリス法は任意の j,α,γ について、

$$h_j^{\gamma}(\alpha_j) = h_j^{\alpha}(\gamma_j) \tag{1.26}$$

が成立する場合に対応する。また、

$$\begin{aligned} h_j^{\alpha}(\gamma_j) = &\quad h_j^{\{\alpha_1..\alpha_j..\alpha_I\}}(\gamma_j) = \quad P(\gamma_j|\alpha_1..\alpha_{j-1},\alpha_{j+1}..\alpha_I) \\ h_j^{\gamma}(\alpha_j) = &\quad h_j^{\{\alpha_1..\gamma_j..\alpha_I\}}(\alpha_j) = \quad P(\alpha_j|\alpha_1..\alpha_{j-1},\alpha_{j+1}..\alpha_I) \end{aligned} \tag{1.27}$$

とした場合が Gibbs sampler に相当する。この場合、 (1.25) 式の分母・分子に $P(\alpha_1..\alpha_{j-1},\alpha_{j+1}..\alpha_I)$ を乗じることで、常に $r = 1$ となることが示せる。

マルコフ連鎖モンテカルロ法のアルゴリズムについての最近の発展については、 1.4.1節及び 1.6.1節の文献を参照されたい。

1.4 緩和に関する問題

1.4.1 緩和時間の長くなる状況とその診断

緩和に関する問題は、

1. 緩和が遅い場合にそれをを見ぬくにはどうしたら良いか。
2. 緩和を速くするにはどうしたら良いか。

の 2 つに分けられる。

後者については、 Metropolis-coupled MCMC 法 (MCMCMC) (温度並列シミュレーテッド・アニーリング法、交換法 : Geyer (1991), 木村, 瀧 (1990), 石川, 金久 (1993), 小西ほか (1995), 高畠 (1992), 伊庭 (1993b), 福島 (1994), Gilks et al. (1996))、 simulated tempering 法 (Marinari and Parisi (1992), Geyer and Thompson (1995))、マルチカノニカル法 (Berg and Neuhaus (1992), Berg and Celik (1992a, 1992b)) などの拡張されたアンサンブルを使う方法が注目されている。また、非局所的なダイナミクスも種々の観点から研究されている (Noguchi and Go (1985), Swendsen and Wang (1986,1987,1988), Liang (1992), Qian and Titterington (1989), Besag and Green (1993), Goodman and Sokal (1989), Neal (1995))。これらの方法は今後重要になるものと思われるが、解説は別の機会に譲りたい。

前者も一般的には困難な問題である。精密に緩和時間を知るためには統計量の期待値の計算よりはるかに手間がかかることが多い。緩和時間のおおまかな上限を知るだけなら簡単なように見えるが、必ずしもそうではない。特に緩和が多少遅い場合よりも、きわめて遅い場合の方がかえって気づきにくいので注意する必要がある。もちろん、有限系なら有限時間で緩和する証明があるので大丈夫といった考えは危険である。宇宙が出来てから現在までの時間よりも緩和時間が長い有限系の例は簡単に作ることができる。緩和がきわめて遅い場合にそれをチェックできる確実な診断法はないと思われるが、起こり得る事態についての知識があれば助けになる。緩和の遅くなる典型的状況と診断法のヒントをいくつか述べよう。

- 分布の峰が複数ある場合には緩和時間が長くなりがちである。分布のモードを求めるための数値的最適化が局所的極値にとらえられてしまう場合でも、マルコフ連鎖モンテカルロ法は確率的要素を含むため、そこから抜けだせる可能性がある。しかし、それにも限度があって、複数の峰のひとつに長時間とらえられてしまうことも起こる。複数の峰のひとつだけにとらえられていないことを確かめるには、初期状態を変えて何回もシミュレーションを行ない、結果が一致するかどうか調べるしかない。ランダムな初期状態だけでなく、極端な初期状態も試みるとよい。
- 水を静かに冷却していくと、零度より温度が下がっても凍らないことがある。これと同じことがマルコフ連鎖モンテカルロ法一般でも起こりうる。なんらかの状態の近くに分布が集中すべきなのに、その状態が “うまく見つからない” ために、いろいろな状態をまわりつづけるような場合である。この場合、確実に誤りを避ける方法はないが、分布のモードがあらかじめわかっていれば、それを初期状態として計算を行なったものと比較することで診断できる場合もある。

- 別のタイプの緩和が遅くなる場合として、ガウス分布 (線形系) での悪条件 (ill-condition) に相当する場合、すなわち、状態空間に細長い分布の “尾根” ができる場合がある。メトロポリス法で $T \to 0$ とすると一種の最急降下法になることからわかるように、局所的なダイナミクスを持つマルコフ連鎖モンテカルロ法は悪条件に弱い。

- その他、緩和が遅くなるので有名なケースとしては、いわゆる “critical slowing down” や空間パターンが違う性質の領域に分かれてその境界が乱歩する場合などがある。すぐ後で述べる臨界点より強結合側のマルコフ場での異常な緩和は後者の原因による。これらは一種の悪条件ともいえるが、ガウス分布での悪条件の場合とは本質的に違う面を含んでいる。

なお、 “エネルギー” (分布の対数 $\log P(\alpha)$) の振舞い だけ から緩和の判定をするのは好ましくない。 “エネルギー” は悪条件に対して鈍感であり、時にはモデルの対称性の影響 (付録 A の A.4節) を受けることがある。一般に、複数の統計量をモニターすること、また、サンプル列自体を可視化して観察することが重要である。

1.4.2 緩和の遅い例 — マルコフ場の場合

イジング模型 (1.15) 式において結合定数 J の値が小さいときは、生成されるパターンは白と黒の小さなクラスターからなり、ほぼ乱雑に見える (図 1.2a)。 J の値が大きくなるとクラスターの大きさは大きくなり (図 1.2b)、 “臨界点” に対応する値 $J = 0.4407..$ に近付くと白や黒の大きなクラスターが生成消滅を繰り返すようになる。これに伴って、メトロポリス法及び Gibbs sampler の緩和時間は急激に長くなる。これは、白や黒の大きなクラスターが生成消滅を繰り返すためであり、 critical slowing down と呼ばれる。結合定数 J の値を更に大きくすると白が優勢の状態か黒が優勢の状態 (図 1.2c) のいずれかが卓越するようになり、その間の遷移に要する時間は極めて長い。

図 1.2: イジング模型の生成する空間パタン。系の大きさは 40x40, 上から (a)J=1/5.0, (b)J=1/2.5, (c)J=1/2.0 に対応。 “すべて黒” の状態からはじめて 10000MCS 後。

また、 J が臨界点より大きい側で完全に乱雑な初期値から計算を始めた場合、白と黒の大きな領域に塗り分けられた状態が過渡的に出現して、なかなか消えないことがある (Miyashita and Takano (1985))。 S_i のとりうる状態数 (“色” の数) を 3 色以上に拡張した模型 (ポッツ模型,Wu(1982)) においても同様の現象が見られる。この現象は実際の統計の問題への応用でも問題になり、 “マルコフ連鎖モンテカルロ法によるマルコフ場のサンプリングによって生成された” とされる模擬データが典型的なサンプルとはいえない場合が少なからずあることが (たとえば、 Gray et al. (1994) に) 指摘されている。また、ポッツ模型で色数が多い場合には、先に述べた “過冷却” に相当する現象も起きる。本来はいずれかの “色” にほぼ揃った状態が高い頻度で出現するはずのときに、乱雑な状態からはじめると、種々の乱雑な状態を経めぐるばかりでなかなか揃わないのである (Kikuchi and Okabe(1992))。

以上はベイズ的な画像処理でいえば、事前分布からのサンプル生成に相当する場合であるが、応用上重要なのは事後分布からのサンプル生成である。 (1.15) を事前分布として、事後分布を考えると、たとえば、

$$P(\{S_i\}) = \frac{1}{Z}\exp(J\sum_{nn} S_i S_j + \sum_i h_i S_i) \tag{1.28}$$

のように非一様な “外場” h_i を含む分布になる。この 場合、 h_i の分布によって緩和は速くもなりうるし、極めて遅くもなりうる。

ここでは離散変数の場合を主に述べたが、連続変数で裾の重い分布を事前分布として用いた場合も緩和は遅くなることが多い。たとえば、 Kitagawa(1987) で扱われている非ガウス平滑化の問題で事後分布からのサンプル生成に通常のマルコフ連鎖モンテカルロ法を適用すると、緩和時間は極めて長くなる。

1.5 simulated annealing

いま、与えられた確率分布 $P(\alpha)$ に対して変形した分布、

$$P_T(\alpha) = \frac{P(\alpha)^{1/T}}{\sum_\alpha P(\alpha)^{1/T}} \tag{1.29}$$

を考える。 T を温度という。温度 T が小さいときには、分布 P_T は最大確率の状態の上に集中する (実際、 $T \to 0$ では、メトロポリス法のアルゴリズムは局所的な逐次改良法と見なせる)。これを利用して、マルコフ連鎖モンテカ

ルロ法において、T の大きいところから出発して、ゆっくりと $T \to 0$ とすることで、組合せ的最適化の問題を解くことができる。ゆっくりと $T \to 0$ とする意味は、はじめから $T = 0$ とするのに比べて、温度が有限のときに、局所的な最適解のまわりから抜け出す機会が多いということにある。この方法を、材料が歪んだ状態 (局所的極値) にあるのを加熱と除冷 (anneal) によって除く方法にたとえて、 simulated annealing 法 (模擬やきなまし法) と呼ぶ。マルコフ連鎖モンテカルロ法の一種として、 simulated annealing 法があるのではなく、マルコフ連鎖モンテカルロ法を組合せ的最適化に使うという考えを simulated annealing と呼ぶのである。 simulated annealing の考えは Kirkpatrick et al.(1983) で提唱されて以来、広く用いられるようになった。

simulated annealing 法がうまくいくためには、有限温度で出現しやすい状態の集合に大域的な最適解が含まれていることが望ましい。また、厳密に最適解が必要とされる場合より、近似的な解で良い場合に向いた方法である。実際の計算は、確率的に繰り返しバックトラックを行なっているのに相当するので、時間がかかる。人間が複雑なアルゴリズムを考える手間の分も計算機が努力していると考えれば、ある程度は容認できるだろう。数学的には、極めてゆっくりと温度 T を下げた場合に確率 1 で大域的な最適解に到達することが証明できる (Geman,S. and Geman,D.(1984), Winkler(1995)) が、そのような証明に実際的な意味があるかどうかは疑わしい。

simulated annealing 法の意義は、ランダムな揺らぎが工学的な応用で肯定的な意味を持つことを示したことである。逆に良くない影響は、マルコフ連鎖モンテカルロ法の本来の用途は分布からのサンプル生成であるのに、最適化という文脈でのみ考える傾向が一部で生じたことである。もちろん、多くの工学の問題では、 (近似的) 最適解のみが意味を持つ。しかし、少なくとも 2 つの分野 — 統計物理と統計学 — では多変数の確率分布からのサンプリングそれ自体が重要な課題なのである (Iba(1989), 伊庭 (1993a))。

1.6 参考文献: とくに統計への応用について

1.6.1 総論及び方法論

統計物理でのマルコフ連鎖モンテカルロ法について、網羅的に成果を集めた本として、 Binder (1986 (1st ed. 1979), 1987 (1st ed. 1984), 1992) がある。これらは網羅的すぎて読みやすくはないが、 2 冊目の冒頭の Binder and Stauffer (1984) は比較的わかりやすく入門向きである。 Binder と Heermann による教科書 (Binder and Heermann(1992)) が同じ Springer-Verlag から出ているが、これは臨界現象の解析が中心になっていて、統計学者や工学者にはあまり向かないと思う。

統計学者によるマルコフ連鎖モンテカルロ法の本としては、 Gilks et al.(1996) がある。内容はトピック別の分担執筆であるが、最新の内容を含んでおり興味深い。また、画像関係への応用を中心とした新しい教科書に Winkler(1995) がある。総合報告としては、少し古いが、 J.Roy.Stat.Soc. の同じ号に特集として 3 つの論文と討論が掲載されているのが目につく (Smith and Roberts (1993), Besag and Green (1993), Gilks et al. (1993))。 Besag et al. (1995) は具体例に基づいた報告と討論である。そのほかの報告としては、 Geyer (1992), Tierney (1994) などがあり、 Tanner の本 (1993) の後半にも解説がある。このうち Tierney (1994) は収束性などの理論的な面が中心になっている。

メトロポリス法は Metropolis et al.(1953) が、 simulated annealing は Kirkpatrick et al.(1983) がそれぞれ “原典” である。後者は読みやすい。統計の文献では Hastings(1970) が Metropolis et al.(1953) と共に引用されるが、メトロポリス法を独立に発見したという意味ではなく、一般化された形 (1.3.4節) を示したという意味である。 simulated annealing 法に関連した先駆的な仕事としてはわが国の Tsuda and Kiyono (1964) によるものがある。 Geman et al. (1990) には、 simulated annealing 法の同時発見者として Cêrný の名があげられている。

本論文では触れなかったが、計算手段としてのランジュバン方程式や Molecular Dynamics (とモンテカルロ法のハイブリッド法) も統計学に導入されている。これらの方法はメトロポリス法での比 $r = P(\gamma)/P(\alpha)$ が 1 成分のみの変化 $\gamma_j = \alpha_j + \triangle\alpha$ に対しても計算し難いような場合に特に効果を発揮する。たとえば、 $P(\alpha)$ が α の成分からなる行列 A の行列式 $\det A$ を含む式で表される場合である。ランジュバン方程式やハイブリッド法の統計に対する応用については、 Amit et al.(1991), Grenandar and Miller(1994) 及び Neal(1995) を、素粒子物理で開発された技法については Kronfeld(1993), Neal(1995) 及びその文献を参照されたい。

1.6.2 各論 (応用分野別)

ここまでは統計への具体的な応用についてほとんど述べなかった。第 2 部では特定の例題についてやや詳しい研究を紹介するが、ここでは種々の応用についての簡単な解説と文献をあげることにする。いかなる意味でも網羅的とはいえないことをお断わりしておく。 Gilks et al.(1996) の各章からも以下の各分野についての有力な情報が得られ

るので、ぜひ参照されたい。また、 “Bayesians Worldwide” のリストを使って海外の若手研究者の home page にアクセスすることもできる (http://bayes.stat.washington.edu/bayes_people.html)。

マルコフ連鎖モンテカルロ法の統計学への応用として早くから手がつけられていたのは、尤度関数の規格化定数が解析的に計算できない場合への応用である。そのほかの重要な応用はベイズ統計の諸問題に対するものである。必ずしもベイズ的な用語が用いられないような例 — 種々のミクスチャの問題、 missing data を含む問題、潜在変数を含む問題 — を含めれば、マルコフ連鎖モンテカルロ法の統計学への応用例の大部分は後者の範囲に属する。

まず、尤度関数 $L(y|x)$ の y に関する規格化定数が解析的に計算できない場合への応用であるが、次のようなものがあげられる。

- 空間点過程への応用

  尾形・種村は空間の点配置を相互作用する液体分子のつくるパターンとのアナロジーでとらえ、最尤法によって相互作用のパラメータを推定することを試みた。このモデルでは、規格化定数 (液体の統計力学では分配関数に相当する) が計算上の困難をもたらす。これを解決するために、尾形・種村はマルコフ連鎖モンテカルロ法を導入した (Ogata and Tanemura (1981,1984,1985,1989)、背景については長谷川, 種村 (1986))。マルコフ連鎖モンテカルロ法は、初期の仕事ではサンプルパタンの生成及び最尤法における近似のチェックに使用されており、のちには最尤法の計算そのものに利用されている。これらは世界的に見ても早い時期の仕事といえる。同じ方向の研究としては Penttinen (1984) がある。 Strauss (1986) も参照されたい。

- マルコフグラフ・マルコフ場への応用

  Frank and Strauss(1986), Strauss(1986) は social network のの特徴 (たとえば友人・知人関係が推移律を満たす度合) を表すモデルとしてマルコフグラフモデルを導入し、マルコフ連鎖モンテカルロ法の適用を論じている。また、 Cross and Jain(1983) はテクスチャ (肌理) のモデルとしてマルコフ場を論じ、サンプルパタンの生成法としてのマルコフ連鎖モンテカルロ法を導入した (ただし、緩和の問題に関する注意が不足していることが指摘されている)。マルコフ場に関するその後の発展については、すぐ後の “ベイズ統計とその周辺への応用” の “画像モデル・空間モデルへの応用” の項を参照されたい。

- DNA finger printing の解析への応用

  Geyer and Thompson(1992) では DNA finger printing の解析の問題を例として auto-logistic model (イジングモデル) に対するマルコフ連鎖モンテカルロ法の応用が論じられている。この文献は総説ではないが、討論つきで、歴史的側面にも触れている。

ベイズ統計とその周辺の問題に対する応用にはさまざまなものがある。

- 画像モデル・空間モデルへの応用

  画像再構成へのマルコフ連鎖モンテカルロ法の応用は Geman 兄弟の仕事の影響力のもとで普及した (Geman,S. and Geman,D.(1984), Geman,D. and Geman,S.(1986), Geman et al.(1987), Geman and McClure(1987), Geman et al.(1990))。 simulated annealing による画像再構成の仕事は同時期に他にも存在する (たとえば Camevali et al.(1985)) が、「ベイズの枠組の明示」、「simulated annealing 法の収束証明」、「マルコフ場モデルの確立」、「線過程の導入」などをひとつの総体として提示したところにこの仕事の意義がある。

  Geman,S. and Geman,D.(1984) では simulated annealing の形でのマルコフ連鎖モンテカルロ法の応用が強調されている (それ以外の応用についても示唆されているが、注意深く読まないと気づかないかもしれない)。これに対して、その後の文献では事後分布からのサンプル生成やそれを用いた期待値の計算も重視されている (たとえば、 Marroquin et al.(1987), Geman and McClure(1987), Dubes and Jain(1989), Ogata (1990,1991), Besag et al.(1991), 伊庭 (1991a), Rue(1995), Besag(1986) の 2.4 節などを参照)。

  画像再構成のための超パラメータ推定の問題は、ある意味では尾形・種村の仕事に代表されるような最尤法によるパラメータ推定の問題の “不完全データ” への拡張とも見ることができる。これに関連した研究としては、たとえば、 Geman and McClure(1987), Qian and Titterington(1989,1991,1993), Gray et al. (1992), Ogata (1990), Ohtsuki and Kawato (1991), Besag et al.(1991), Veijanen(1991), Dinten et al.(1991), 伊庭 (1991a) などがある。

  マルコフ場へのマルコフ連鎖モンテカルロ法の応用についての教科書としては、先ほども触れた Winkler(1995) がある。また、マルコフ場に関係した総合報告や論文集としては、 Besag (1986,1989), Journal of Applied

Statistics 16 No.2(1989, 統計的画像処理の特集号), Chellappa and Jain(1993), Possolo(1991) などがあげられる。視覚情報処理などの認知的側面とのかかわりについては Marroquin et al.(1987) を参照されたい。

物体の輪郭線や連続変形に対するに対するベイズ的なモデリングに興味が持たれている (active contour model, deformable template model など)。また、これらから出発して、より "高次" の物体の特性をモデル化する試みもある。これらの問題に対するマルコフ連鎖モンテカルロ法の応用に関しては、 Phillips and Smith(1994), Stovik(1994), Amit et al.(1991), Grenandar and Miller(1995) などを参照されたい。 Journal of Applied Statistics 21 No.1 (1994) は先にあげた特集の続編であるが、目次によればこの話題に関係する論文を複数含んでいるようである (未入手)。

- 時系列モデルへの応用

  この分野でもマルコフ連鎖モンテカルロ法は流行である。非線形・非ガウスの状態空間モデル、変化点の検出問題などが研究されている。筆者はあまり良く文献を知らないが、 Gordon and Smith (1990), 伊庭 (1991b), Carlin et al. (1992), Berry and Hartigan (1993), West(1995), Chan and Ledolter (1995) などをあげておく。

- 多層パーセプトロンへの応用

  多層パーセプトロン (Rumelhart et al. (1986), Cheng and Titterington (1994)) はいわゆる Neural Network のうちもっとも普及しているモデルである。このモデルでの overfitting を防ぐ方法のひとつとして、パラメータ ("シナプス荷重") に罰金項ないし事前分布を仮定する方法が提案されている (石川 (1991), Mackay (1992), Williams (1995))。これは、重回帰分析におけるリッジ回帰及びその一般化 (石黒 (1985)) に対応する。この方向へのマルコフ連鎖モンテカルロ法の応用については Neal(1995) がある。

- 動的デザイン (query sampling, active learning) への応用

  いままで獲得したデータをもとに、次のデータの採取計画を作るという研究 (動的デザイン, query sampling, active learning) が最近注目されている。この方面でも、マルコフ連鎖モンテカルロ法は有効である。単純な応用例として伊庭 (1994) を、より本格的な例として Paass and Kindermann (1995) をあげておく。後者では多層パーセプトロンモデルにおける query sampling へのマルコフ連鎖モンテカルロ法の応用が論じられている。

- 離散パラメータを持つベイズモデルへの応用

  マルコフ連鎖モンテカルロ法を離散パラメータを持つ統計モデルに応用するのは非常に興味深い。画像モデルや時系列におけるラベルも一種の離散パラメータであるが、それ以外の例もいろいろある。計算上の困難 (いわゆる組合せ爆発) のために多数の離散パラメータを持つ統計モデルはいままであまり正面から扱われなかったが、マルコフ連鎖モンテカルロ法がひとつの突破口をひらくことが期待される。

  いわゆる Neural Network の分野では、 Hinton と Sejnowski による "ボルツマンマシン" の研究がある (Hinton and Sejnowski(1986), Sejnowski et al.(1986))。これは結合定数が場所によって異なるイジング模型によってさまざまな機能 (たとえば図形の対称性の検出) を学習させる研究である。 Hinton らの研究では、 EM アルゴリズムの E-step の計算のためにマルコフ連鎖モンテカルロ法が用いられている。 Hinton らの "ボルツマンマシン" を統計モデルとして考えると、多数の (離散) 潜在変数のある対数線形モデルに相当する。類似のモデル (群) として Bayesian Network と呼ばれるものがあり、最近さかんに研究されているようである (たとえば、 Hrycej (1990))。

  重要な分野としてデータから遺伝子の家系図上の伝搬や染色体上の位置関係を推定する問題 (pedigree analysis, linkage analysis) がある。この問題については、そのままではアルゴリズムの緩和に問題があるために工夫が必要なこともあり、多くの論文が書かれつつある (Thompson(1991,1994), Kong(1991), Geyer and Thompson (1995))。

  別の興味深い応用は DNA 配列の local もしくは global なマルチプルアラインメントに関するものである (Lawrence et al. (1993), Liu et al. (1994), Liu et al. (1995))。石川, 金久 (1993) はマルチプルアラインメントの問題を拡張アンサンブルの一種 (本論文でいう交換法) を用いて取り扱っているが、最適化手法としての側面しか見ていないように思われる (注 1)。

  ベイズの枠組ではモデル選択を離散パラメータの推定問題として扱うこともできる。この方向へのマルコフ連鎖モンテカルロ法の応用としては、 Carlin and Chib(1995) 及びその文献, 伊庭 (1991b) などを参照されたい。

- 離散パラメータを含む分類の問題 — 画像以外 — への応用

  対象がいくつかのグループに分かれていることを仮定したモデルで、どのグループに属するのかを表すラベルとして離散パラメータが用いられることがある。ラベルはしばしば missing data として解釈される。

  分類に対するマルコフ連鎖モンテカルロ法の応用として分かりやすいのはクラスター分析での群内平方和の最小化への simulated annealing 法の適用である (佐藤 (1992))。この場合はベイズモデルが設定されているわけではないので、最適化手法としての適用に限定される。

  背後の統計モデルをより意識した分類法としては finite mixture model(Titterington at al. (1985)) に基づく方法がある。 finite mixture model にマルコフ連鎖モンテカルロ法を応用した研究としては Celeux and Diebolt (1985), Celeux (1986), Celeux and Govaert (1992), Celeux and Govaert (1993) 及び Diebolt and Robert (1994) があげられる。前 4 者は SEM アルゴリズムと称する方法などを論じている。最後の論文では、 各成分分布のパラメータ (normal mixture なら各成分の平均や分散) と所属グループを表す離散ラベルの同時事後分布をマルコフ連鎖モンテカルロ法でサンプルしている (通常の EM アルゴリズムによる扱いでは前者は点推定で置き換えられる)。

  より興味深いのは、分類の対象それ自身でなく、その pair に関してデータが与えられる場合である。このような問題は social network のデータなどから生じる。こうした場合へのマルコフ連鎖モンテカルロ法の適用については続く第 2 部で扱う。 Iba(1992) も参照されたい。

- ベイズ統計一般への応用、 “data augmentation”

  Gelfand, Smith たちは Gibbs sampler のさまざまな問題に対する応用を与えている (Gelfand and Smith(1990), Gelfand et al.(1990))。簡単なものをひとつあげると、正規分布の分散がさらに逆ガンマ分布しているようなモデルである。これらの例の多くでは、 Gibbs sampler が安価にインプリメントできる。また、なかには多くの変数を同時に取り直すようなダイナミクスが可能なものもある。 Tanner らによって統計学の内部で独立に考案された “data augmentation” (Tanner and Wong (1987), Tanner (1993)) のテクニックもこの文脈で考えられる (Gelfand and Smith (1990))。

第 2 章

第 2 部: 事例研究

第 1 部の最後で触れたように、離散パラメータを含むモデルの重要なグループとして、対象の分類のための離散ラベルを含んだ問題がある。ここでは、そのうち “相互関係データにもとづく分類の問題” を取り上げる。分類に関係した統計モデルとして良く知られている finite mixture model の場合、分類すべき対象それぞれに対してデータが与えられる。これに対して、ここで取り上げるタイプの問題では、分類すべき対象の 各対 についてデータが与えられる点が異なっている。以下では、このタイプの問題のひとつを取り上げ、マルコフ連鎖モンテカルロ法を適用して解析する。同じタイプの別の問題 (対戦データに基づく分類、とくに 3 すくみのある場合) の詳しい研究は Iba(1992) を参照されたい。

2.1 問題の説明

表 1 は米国のある小学校 (5,6 年生) の “友人関係” の調査の一部である (Hansell (1984) のデータを Wang and Wong (1987) より引用)。

表 1

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
1		1	1	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0
2	1		0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
3	1	1		0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	1	0	0	1	0	0	0
4	1	1	1		1	1	1	1	1	1	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0
5	1	1	0	0		0	0	0	1	1	0	1	0	1	1	0	1	1	0	0	0	1	1	1	1	0	0
6	0	1	0	0	1		0	0	0	1	0	0	1	1	0	1	1	1	0	0	0	0	1	0	1	0	0
7	0	0	0	0	1	1		0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	1	0	0	0	1	1	0		1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	1	0	0	1	0		0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0		0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	1	0	0	0	1	0		0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	1	0	0	0	1	0	0	0	0	0	0
14	0	0	0	0	1	0	0	0	0	0	0	0	0		1	0	1	1	0	0	1	0	0	1	1	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	1		0	0	1	0	0	1	1	0	0	1	0	0
16	1	1	0	1	1	1	0	1	0	1	1	0	0	1	1		1	1	1	0	1	1	1	1	1	0	1
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		1	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	1	1	0	0	0	0	0
19	0	0	0	1	0	0	0	0	0	0	0	1	0	1	1	1	1	1		0	1	1	0	1	1	0	0
20	1	0	1	0	1	1	1	0	0	0	1	1	0	0	1	0	1	1	0		0	0	0	1	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0		0	0	0	1	1	0
22	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0		1	1	1	0	1
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	1	1		1	1	0	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0		1	0	0
25	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0	0	1	0	1	1		0	0
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0
27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

表 1 のデータを $\widetilde{n}_{ij}$ とする。表の縦、横がそれぞれ i, j に対応する。 i が j の友人であれば、表の (i, j) 欄は 1、そうでなければ 0 とした。人数 N は 27 人。表 1 は非対称であるが、以下の解析では関係の有無のみに着目して、対称に直す。すなわち、 (i, j) 成分か (j, i) 成分のいずれかが 1 であれば、両者を 1 とする。こうしてできたデータが図 2.1である。以下では、これを $\{n_{ij}\}$ で示す。 $n_{ij} = \max(\widetilde{n}_{ij}, \widetilde{n}_{ji})$ である。

図 2.1: 行列 n_{ij}。縦軸が i、横軸が j。 $n_{ij} = 1$ のとき丸、 $n_{ij} = 0$ のとき空白を記入して表したもの。対角線にも丸を記入した。

なお、周辺度数を見ればわかるように、表 1 の非対称の程度はかなり顕著である。たとえば、 $\sum_j \widetilde{n}_{20,j} = 11$ に対し、 $\sum_i \widetilde{n}_{i,20} = 0$ となる。従って、このデータに限っていえば、われわれの見方はいくらか不自然な点がある。非

対称性が著しいのは、データを得るに当たって “友人” というより “好きな相手” あるいは “お気に入りの子” を尋ねるのに近い質問の仕方 (Wang and Wong(1987) 参照) がされたためかもしれない。

ここで、データ $\{n_{ij}\}$ をもとに生徒達をいくつかのグループに分けることを考える。グループの数の判定はもちろん興味ある問題であるが、ここでは簡単のために 2 つのグループ I 、 II に分けるとする。そして、 “同じグループ内の 2 人が互いに友人である確率を p_1、異なるグループ内の 2 人が互いに友人である確率を p_2 とし、任意の 2 人が友人であるか否かはすべて独立と考える” というモデルを考えよう。 p_1、 p_2 は組わけとともにデータから推定する。 $p_1 > p_2$ とする。

生徒の数 N だけの離散パラメータ $\{S_i\}$ を導入して、人物 i がグループ I に属するとき $S_i = +1$、グループ II に属するとき $S_i = -1$ とすると、このモデルの尤度関数 $L(\{S_i\}; p_1, p_2)$ は、

$$L(\{S_i\}; p_1, p_2) = \Pi_{i>j} L_{ij}(S_i, S_j; p_1, p_2) \tag{2.1}$$

$$L_{ij}(S_i, S_j; p_1, p_2) = p_1^{n_{ij}}(1-p_1)^{1-n_{ij}} \qquad if\ S_i = S_j \tag{2.2}$$

$$L_{ij}(S_i, S_j; p_1, p_2) = p_2^{n_{ij}}(1-p_2)^{1-n_{ij}} \qquad if\ S_i \neq S_j \tag{2.3}$$

となる。これを見やすく書きなおすと、

$$\log L(\{S_i\}; p_1, p_2) = \sum_{i>j} J_{ij}(p_1, p_2) S_i S_j + \sum_{i>j} C_{ij}(p_1, p_2) \tag{2.4}$$

$$J_{ij}(p_1, p_2) = \frac{n_{ij}}{2} \log \frac{p_1}{p_2} + \frac{1-n_{ij}}{2} \log \frac{1-p_1}{1-p_2} \tag{2.5}$$

$$C_{ij}(p_1, p_2) = \frac{n_{ij}}{2} \log(p_1 p_2) + \frac{1-n_{ij}}{2} \log((1-p_1)(1-p_2)) \tag{2.6}$$

とも書ける。 2 次の項 $S_i S_j$ が現れるところが、 finite mixture model の場合と異なっていることに注意されたい。この項の存在が問題を難しくしている。

統計物理でいえば、このモデルはイジング型の “スピングラス模型” (高山 (1991), Binder and Young (1986), Mezard et al.(1987)) に類似している。実際、スピングラス理論の教科書である Mezard et al.(1987) は社会関係の問題を冒頭で説明のために用いている。本論文での試みはこの “説明のためのたとえ話” を真面目な問題として取り上げたものとも考えられる。

p_1,p_2 をハイパーパラメータとするようなベイズ的な取り扱いをすることにして、 $\{S_i\}$ の事前分布を

$$\pi(\{S_i\}) = \text{一様分布} \tag{2.7}$$

と定める。これは、各 S_i が独立に確率 $1/2$ でグループ I またはグループ II に属するという事前知識に相当する。この事前分布の是非についてはあとで論じる。

2.2 関連研究・関連問題

われわれのモデルは social network の用語でいえば、 *a posteriori* block model の一種である。ここで、 *a posteriori* というのはブロック構造 (組わけ) をデータから自動的に推定することを意味する (これに対して、 *a priori* block model では解析者がブロック構造の候補を与える)。 a priori あるいは a posteriori な block model に関連した研究については Carrington et al.(1979-80), Panning(1982), Holland et al. (1983), Noma and Smith (1985), Wang and Wong (1987), Wasserman and Anderson (1987), Strauss and Ikeda (1990) 及びその引用文献を参照されたい。調べた範囲では a posteriori block model にマルコフ連鎖モンテカルロ法を適用した研究はないが、(特に 1987 年以後の) 文献の調査は不十分であって、先行研究については完全に把握できていない。 Strauss(1986) や Strauss and Ikeda (1990) には social network に対するマルコフ連鎖モンテカルロ法の応用が述べられているが、a posteriori な組わけへの応用は論じられていないようである。

Wasserman and Anderson(1987) も言及しているように、表 1 のような関係データからグループを検出するためには数量化理論や双対尺度法の適用も考えられる。本来はその種の方法との比較も必要であろう。最近、統計物理で話題となった畳み込み符号の問題 (Soulas (1989,1994), Ruján (1993), Nishimori (1993), 伊庭 (1993a)) とわれわれの問題の関係は注 2 を参照されたい。

2.3 計算したい量

p_1,p_2 をハイパーパラメータとして固定すると、$\{S_i\}$ の事後分布は、

$$P_{pos}(\{S_i\}) = \frac{L(\{S_i\};p_1,p_2)\pi(\{S_i\})}{\sum_{\{S_i\}} L(\{S_i\};p_1,p_2)\pi(\{S_i\})} \tag{2.8}$$

のようになる。ただし、$\sum_{\{S_i\}}$ はあらゆる可能な $\{S_i\}$ の組合せについての和を意味する。この事後分布からのサンプル生成をマルコフ連鎖モンテカルロ法で行ないたいわけである。

次に、p_1,p_2 の推定であるが、ここでは、$\{S_i\}$ に関して周辺化した (和をとった) 量

$$l(p_1,p_2) = \log \sum_{\{S_i\}} L(\{S_i\};p_1,p_2) \tag{2.9}$$

を最大化することで行なう。これは Good が TYPE II 最尤法、赤池が ABIC 法と呼んだ方法に相当する (たとえば伊庭 (1996a))。p_1,p_2 の推定法としては、これ以外に、

1. ハイパーパラメータ p_1,p_2 とパラメータ $\{S_i\}$ について同時に事後確率を最大化する方法
2. p_1,p_2 についても点推定で置き換えずに分布を考える方法

が考えられる。ここでは、これらの違いについては深入りしないが、少なくともパラメータが連続変数の場合には 1. は良くないという証拠がある。この問題についての一般的な解説は、伊庭 (1996a,1993a,1996b) とそこにあげた文献を参照されたい。また、2. に相当することをマルコフ連鎖モンテカルロ法を利用して行なう方法については Iba(1992) で少し議論した。

(2.9) 式を p_1,p_2 で微分したものはそれぞれ、

$$\frac{\partial}{\partial p_1} l(p_1,p_2) = \frac{1}{p_1(1-p_1)} \sum_{i>j} (\frac{1+\langle S_i S_j \rangle}{2})(n_{ij} - p_1) \tag{2.10}$$

$$\frac{\partial}{\partial p_2} l(p_1,p_2) = \frac{1}{p_2(1-p_2)} \sum_{i>j} (\frac{1-\langle S_i S_j \rangle}{2})(n_{ij} - p_2) \tag{2.11}$$

となる。ここで、$\langle \quad \rangle$ は事後分布での期待値をあらわす。これらの右辺を零とおいて p_1,p_2 について解くと

$$p_1 = \frac{\sum_{i>j} n_{ij} (\frac{1+\langle S_i S_j \rangle}{2})}{\sum_{i>j} \frac{1+\langle S_i S_j \rangle}{2}} \tag{2.12}$$

$$p_2 = \frac{\sum_{i>j} n_{ij} (\frac{1-\langle S_i S_j \rangle}{2})}{\sum_{i>j} \frac{1-\langle S_i S_j \rangle}{2}} \tag{2.13}$$

となる。(2.12),(2.13) の右辺を計算するためには、事後分布での期待値

$$\langle S_i S_j \rangle = \frac{\sum_{\{S_i\}} S_i S_j L(\{S_i\};p_1,p_2)\pi(\{S_i\})}{\sum_{\{S_i\}} L(\{S_i\};p_1,p_2)\pi(\{S_i\})} \tag{2.14}$$

を求めれば良いが、これはマルコフ連鎖モンテカルロ法によって事後分布 (2.8) からのサンプル生成を行なうことで計算できる。これを用いると、p_1,p_2 は、

- p_1,p_2 に適当な初期値を与え、マルコフ連鎖モンテカルロ法によって、(2.12),(2.13) の右辺を計算する。
- 計算した値で p_1,p_2 を置き換え、再びマルコフ連鎖モンテカルロ法を走らせる。

を繰り返すこと (いわゆる EM アルゴリズム (Dempster et al. (1977)) によって推定することができる。

2.4 アルゴリズムの実装

基本的には正方格子上の強磁性イジング模型と同じだが、結合の強さ J_{ij} が i,j に依存する点と任意の $i \neq j$ に対して $J_{ij} \neq 0$ である点が異なる。以下では、 S_i をとり直したあとでもその値が不変な場合に計算の一部を省くようにしたものを示す。なお、乱数の生成には伏見, 手塚 (1981) に示されたプログラムを用いた。

われわれのモデルに関する Gibbs sampler

まず $\{S_i\}$ の初期状態をセットする。各 i について $f_i = \sum_{j \neq i} J_{ij} S_j$ を計算してストアしておく。

1. 乱数を用いて S_j を選ぶ。
2. $\widetilde{\triangle E}$ を計算する。
   (2.15) $$\widetilde{\triangle E} = -2 f_j$$
3. (0,1) の一様乱数 RND を発生し、
   (2.16) $$RND < \frac{1}{1 + \exp(\widetilde{\triangle E})}$$
   なら注目している S_j を $+1$ で、そうでなければ -1 で置き換える。
4. もし、 S_j の値が 3. の操作で変化したならば、 f_i の値を修正する。すなわち、すべての i について、
   (2.17) $$f_i = f_i + 2 J_{ij} S_j$$
   とする。

1.- 4. を繰り返す。

2.5 実験結果

2.5.1 事後分布からのサンプル生成

まず生成されるサンプル列を見よう。以下では、 p_1, p_2 を $(p_1, p_2) = (0.549, 0.213)$ と取った場合を考える。あとで述べるようにこの値は EM アルゴリズムで求められたものである。生徒の総数 N は 27 である。

図 2.2: マルコフ連鎖モンテカルロ法で生成された列。ランダムな初期状態から出発して、 5000MCS 捨てたあとの 1000MCS を 5MCS おきに描いたもの。横軸は MCS 数、縦軸は生徒の番号 i。上段の右端から下段の左端に続く。

かなり綺麗に 2 つのグループに分かれることがわかる。種明かしをすると、 1 番目から 13 番目までは男子生徒、 14 番目から 27 番目までは女子生徒なのである。ほぼそのあたりで 2 つに分かれてみえるのは不思議ではないだろう。詳細に見ると、 5,6,12,20,26,27 番の生徒は 2 つのグループの一方に属するときと他方に属するときがある。それらは独立ではない。たとえば、 5 番が “白” のとき、 6,12,20 番は白に、 26 番は黒に分類されやすいことが読みとれる。

いまの場合、いろいろな初期値から出発して分布の局所的極大を探すと、次の 2 つが見つかる。

図 2.3: (a) が MAP 推定値 (事後分布のモード)。 (b) はそれに近い事後確率を持つ局所的な極大値。

図 2.3と図 2.2を比較すると、マルコフ連鎖モンテカルロ法で生成された列は図 2.3の 2 つの極大の近くを行ったり来たりしていると考えて良いことがわかる。この行き来には 50-200MCS くらいの時間がかかり、それが緩和時間のオーダーを決めている。

図 2.3の (a) と “完全に同じ” 状態が列のなかに現れる回数は (ある測定では)50000 回中 7396 回 (14.8%)、 (b) と完全に同じ状態が現れる回数は 5467 回 (10.9%) であった。比は $14.8/10.9 \simeq 1.35$ である。一方、 (a) と (b) の対数尤度の差は 0.332 なので、これから計算した事後確率の比は $\exp(0.332) \simeq 1.39$ となり、辻褄はあっている。この種の方法を診断に使うという提案 (Zellner and Min (1995)) があるが、より大きな系、また、より多くの状態が列に出現するような場合にはこのチェックはできないことを注意しておく。

2.5.2 EM アルゴリズムによる p_1,p_2 の推定

EM アルゴリズムの iteration は $p_1 > p_2$ を満たす各種の初期値に対して、 4-12 回程度で先に述べた値 (より正確には $(p_1, p_2) = (0.548, 0.213)$) に収束した。ただし、 $p_1 \simeq p_2$ であるような初期値 (J_{ij} の絶対値が小さくなるような初期値) については、与えられた計算時間 (各点 20000MCS) と収束判定条件のもとで誤った値に収束する例も見られた。ここでは、 p_1 と p_2 が共に前回の値と $\epsilon = 5 \times 10^{-4}$ 以下の差の場合に収束と判定している (サンプル平均の確率的変動を考慮して比較的ゆるい条件とした)。結果はまとめて図 2.4に示した。

なお、 $p_1 < p_2$ の初期値からは (今の場合には意味が無いと思われる) 値 $(p_1, p_2) \simeq (0.28, 0.56)$ に収束するのが見られたが、詳しくは調べていない。

図 2.4: EM アルゴリズムの収束。横軸が p_1、縦軸が p_2。 2 重丸が初期値、縦と横の破線の交点が収束先 $(0.548, 0.213)$ を示す。斜めの点線は $p_1 = p_2$ を表す。ここに示した範囲では、 1 つの初期値を除いて望ましい値に収束している。EM アルゴリズムの収束までの iteration 数の最大値は 12 回 (収束判定条件 $\epsilon < 5 \times 10^{-4}$)。各点ごとに、ランダムな初期状態から出発して 5000MCS 捨てた後の 20000MCS で期待値を計算。期待値の計算のためのサンプルは 1MCS ごとに 1 個とった。

2.5.3 統計量の期待値の計算

各 i について $S_i = 1$ である事後確率を示したのが図 2.5である。

図 2.5: 縦軸は $S_i = 1$ である周辺確率。横軸は i。ランダムな初期状態から出発して 5000MCS 捨てた後の 50000MCS で測定。初期状態と乱数のことなる 2 回の結果を縦棒の上端と $-$ 印でそれぞれ示した (ほとんど一致している)。期待値の計算のためのサンプルは 1MCS ごとに 1 個とった。

それ以外の興味ある量としては、生徒 i と j が同じグループに属する事後確率がある。これを計算して図示したのが、図 2.6である。事後確率の大きさが 2 重丸、丸、点、空白で示されている。図 2.1と比較されたい。

図 2.6: i と j が同じグループに属する事後確率。縦軸:i、横軸:j。シンボルは、確率 0.95 以上:2 重丸　0.95-0.80: 丸 0.80-0.50: 点　0.50 未満: 空白。対角線にも 2 重丸を記入した。ランダムな初期状態から出発して 5000MCS 捨てた後の 50000MCS で測定。期待値の計算のためのサンプルは 1MCS ごとに 1 個とった。

2.6 結果の検討

以上をみると非常にうまく話が運んだようにみえるが、実は多くの問題点がある。特に 2番目の問題が深刻である。

1. 黒と白の対称性をめぐる問題。

   事後確率の式からすぐわかるように、すべての S_i の符号を反転しても事後確率の値は変わらない。したがって、「すべての S_i の符号を反転した状態」(=「黒と白を逆転した状態」) の対は図 2.2の列の中に同じ確率で存在すべきであるが、そうはなっていない。初期状態を変えて実験すると、黒と白を逆転した状態を含む列も生じるが、今度は前の列に含まれた状態が出てこない。これでわかることは、この例でのマルコフ連鎖モンテカルロ法は実は緩和していないということである。一番長い緩和時間 τ_1 は極めて長くて、観測時間内で生成された状態はすべて 2 分割された状態空間の一方に “閉じ込められている” のである。

   そうすると、われわれの計算はすべて間違っているのだろうか。そうではない。積 S_iS_j はすべての S_i と S_j の符号を同時に反転しても不変だから、 EM アルゴリズムで利用した統計量及び図 2.6で示した統計量は符号反転に対して不変であり、一番長い緩和時間 τ_1 に対して “感受性” がない (τ_1 に対応する固有ベクトル V_1 への射影が零である)。したがってこれらの量の計算については問題はないのである。

   ただし、計算した量のうち、 S_i の期待値 (図 2.5) だけは別である。正しい計算をすれば、 S_i の期待値はすべて零になってしまう。実際、初期状態と乱数によっては符号が反対の結果も出る。 S_i の期待値が一見良く収束するように見えるので、「とりあえず良いことにする」のは正しくない。分類の不確定さがより大きい場合

には符号反転状態の間を行き来する緩和時間 τ_1 と次の τ_2 の大きさはそれほど違わないかもしれない。そのような場合に S_i の期待値を計算すれば不安定な結果が予想される。図 2.5のような図に意味を与えたいのであれば、 MAP 推定値 (図 2.3a) との重なりが 50% より大きい "半空間" に限定して期待値を計算するのがひとつの方法である (注 3)。

2. 尤度関数の妥当性に関する問題

   図 2.2の結果をよく見ると、別のおかしなところが目につく。たとえば、 26 番の生徒が友人関係を持っているのは 21 番の生徒のみである。にもかかわらず、図 2.2では S_{26} は S_5,S_6,S_{20} などと強く相関した振舞いをしている。

   この結果は、 "友人の総数の少ない生徒を人数の小さい方のグループに入れた方が全体の事後確率が高くなる"(逆も真) という (必ずしも意図しない) 性質がモデルにあるためと思われる。この効果は結果全体に影響を及ぼしている可能性がある。この点からすると、図 2.1のデータのように各生徒の友人の総数 (表の周辺度数) が著しく不均衡なデータを分析するには、われわれのモデルはあまり適切でないことになる。これは、関係の非対称性を扱うのが難しい点とならんで重要な欠点である。より実際のグループ構造を反映するようにするための改良、関係の非対称性を扱うための改良はいくつか考えられる (注 4)。

3. 事前分布の妥当性に関する問題

   事前分布 (2.7) は 2 つのグループの大きさが同じ程度であることを暗に含んでいるため、グループの大きさが不均等な場合には適切でない。グループの大きさの比をデータから推定するか、または、

   $$\pi(\{S_i\}) = \frac{1}{N+1}\frac{1}{{}_N C_l} \tag{2.18}$$

   とした方が良いだろう (たとえば、 Kitagawa and Akaike (1982))。ここで、 N は生徒の総数、 l は一方のグループに属する生徒の数 $l = \sum_i (S_i+1)/2$ である。この 2 つの方法の比較を含む議論と文献は伊庭 (1991b) を参照されたい。ただし、伊庭 (1991b) の問題とは違って、今の問題では黒と白の対称性及び黒白反転状態の間の遅い緩和がある点に注意する必要があるだろう。

以上をマルコフ連鎖モンテカルロ法の応用という観点からみるとどうなるであろうか。まず、考えるべき問題の大部分は計算法よりも、統計モデルや統計の方法論の方にあることがわかる。「計算統計学」というが、本当に問題になる部分はやはり「統計学」の部分であることを痛感する。しかし、ある意味では新しい計算法によって新たな問題点が見えてくるという面もある。たとえば、 MAP 推定値しか計算できないのであれば、最初の項 (1) で触れた問題には気づかなかったかもしれない。

次に、計算法の見地からみて、一番問題になるのは何かというと、やはり緩和の問題だと思われる。モデルに手を入れたり、より多様なデータに適用して行けば、うまく緩和しない場合に出会う可能性が高い。たとえば、事前分布を (2.18) 式に変えたとする。この事前分布のもとでは、データがはっきり 2 つのグループに分かれない場合、 1 つのグループがすべてを含むようになり、一方のグループが消滅する傾向がある。消滅するかどうかぎりぎりの条件で計算を行なった場合、緩和時間はおそらく長くなるだろう。

緩和時間の診断は重要ではあるが、もし、観測時間内に "緩和していない" という結論になった場合に、計算時間を長くする以外に対策が無いのであれば、大変困ったことになる。慎重に診断しようとすればするほど精神衛生に悪いことになってしまう。そこで、緩和を早くするようなできるだけ一般的な方策を開発することが重要になってくる。この方向の研究については第 1 部で文献を紹介した。

上記の各問題点の追求及び他の方法との比較については、別の機会に論じることとしたい。

注と謝辞

注

(注 1)

Thorne らは動的プログラミング (あるいは、非線形フィルタリング、 Baum アルゴリズム) の手法を拡張して、DNA 配列のアラインメントと分岐年代の推定を統合して扱った (Thorne et al.(1991,1992))。彼らはその中で、最適なアラインメントのみではなく、可能なアラインメントの全体 (事後分布) を考えることの重要性を指摘している (最適なアラインメントにのみに基づいて分岐年代を推定すると (短い方に) バイアスが生じる可能性がある)。マルコフ連鎖モンテカルロ法のマルチプルアラインメントへの応用においても、分布からのサンプル生成の側面が重要であるように思われる。

(注 2)

あるビットが 0 か 1 かということが、特定の人の属する集団が I か II かに相当し、符号化されたデータが $\{n_{ij}\}$ に対応する。畳み込み符号では通信路の対称性に相当する条件 $p_2 = 1 - p_1$ が自然な拘束になるという点が本論文の問題とは違う。また、通信路では p_1, p_2 が予め知られているとしても不自然でないだろう。マルコフ連鎖モンテカルロ法の応用という側面からいえば、迅速さが要求される符号解読は良い例とはいえない。

(注 3)

この問題の背景には意外に深いものがある。それは “統計的推定” というものの目的をどうとらえるかに関わってくる。もし、事後分布でなく、事後分布から作られた予測分布のみを考えるなら、こういう問題は一切起きないであろう。これは分類や変化点の検出それ自体は統計学の目的ではなく、これらは予測のための手段であると割り切ることに相当する。しかし、一方では分類や変化点の検出それ自体への興味も否定できない。そこにいろいろな悩みが生じる余地がある。

(注 4)

Wang and Wong (1987) は “友人を作る力の強さ” を各生徒に割り当てるようなモデル (実際にはその非対称版である “p_1 モデル”) にさらにブロック構造の影響を組み込んだモデルを考えている。彼らの研究は a priori なブロックモデルについてのものであるが、これを拡張して、グループ分けも推定するようにすることができるかもしれない。もしこれが可能なら、友人の総数が生徒によって大きく違う場合や関係が非対称な場合のグループ分けを扱うひとつの方法になりうる。ただし、モデルはかなり複雑になるであろう。

別の方向として、グループ数をもっと増やしたり、 “孤独な人” に対応するようなグループを設けたりすることも可能である。これによって、友人の総数が生徒によって大きく違う場合に対処できるようになることが期待される。上の案と比較すると、こちらの方が社会関係に対する見方としては一貫性があるように思われるが、モデル選択の規準を与えるのが難しいだろう。また、この方法では非対称性はうまくモデル化することができない。

謝辞

日頃から種々の助言を頂き、また文献を教えて下さった尾形良彦氏、田辺國士氏、樋口知之氏、菊池誠氏、福島孝治氏、中村永友氏、坂野鋭氏に感謝します。伊藤伸泰氏はこの原稿の “遠い祖先” に当たるものに詳しくコメントして下さいました。田村義保氏は本論文を投稿するよう勧めて下さいました。これらの方々にも感謝します。

付録 A

モンテカルロ法のためのマルコフ連鎖の基礎

A.1 L の固有値の性質の導き方

L の固有値 $\{\lambda_i\}$ の性質は次のようにして分かる。まず $\exp(tL)$ が確率行列であることと遷移のグラフの連結性から、$\exp(tL)$ の固有値は絶対値が 1 以下の複素数であり、とくに定常状態に対応する $\exp(\lambda_0) = 1$ は単純根である。そこで、L の固有値 $\{\lambda_i\}$ は、$\lambda_0 = 0$ を除いてすべて実部が負になる。

詳細釣合の式 (1.1) と L の定義 (1.3) から、

$$P^{\infty}_{\alpha} L_{\gamma\alpha} = P^{\infty}_{\gamma} L_{\alpha\gamma} \tag{A.1}$$

となる。これは、確率ベクトルの空間に定義された内積

$$(Q^1, Q^2)_G = \sum_{\alpha} Q^1_{\alpha} Q^2_{\alpha} (P^{\infty}_{\alpha})^{-1} \tag{A.2}$$

のもとで、行列 L がエルミートであることを示している。そこで、詳細釣合が満たされる場合は、$\{\lambda_i\}$ はすべて実数であることがわかる。これは詳細釣合条件をみたす連続時間のマルコフ過程の緩和は “振動的” ではありえないことを意味する。

この議論からわかるように、V_i は内積 $(\quad)_G$ について正規直交系をなし、(1.6) 式の展開係数 a_i は、

$$a_i = (P(0), V_i)_G = \sum_{\alpha} P_{\alpha}(0) V^{\alpha}_i (P^{\infty}_{\alpha})^{-1} \tag{A.3}$$

となる。また、任意の統計量 A の分散 σ^2_0 は、A をこの正規直交系で展開することにより、

$$\sigma^2_0 = \langle A^2 \rangle - \langle A \rangle^2 = \sum_{i=1,2..} (\sum_{\alpha} A(\alpha) V^{\alpha}_i)^2 \tag{A.4}$$

と書ける。

A.2 離散時間の場合に関する注意

離散時間の場合も連続時間の場合と類似の結果がなりたつ。行列 $\exp(tL)$ の代わりに遷移確率 W の行列を使って論じればよくて、導出はむしろ簡単である。得られる結果は連続時間の場合の $0 < \exp(\lambda_i) < 1 \ (i = 1, 2..)$ を $-1 < \Lambda_i < 1 (i = 1, 2..)$ に置き換えたものになる。これは、付録 Aのこれ以降の結果の大部分について成り立つ。連続時間の場合と異なる点は偶数周期の振動的な緩和が可能なことである。また、離散時間の場合には定常状態の一意性をいうために付加的な条件が必要である。たとえば、最低 1 つの α について $W(\alpha \to \alpha) \neq 0$ となれば十分である。

ほとんどの “自然な” 例では上記の付加条件が満されているが、例外的な場合には詳細釣合条件のもとでも非減衰振動がおこりうる。たとえば、イジング模型 $(1.3.3$節$)$ で結合定数 J が完全に零 [ほとんど零] の場合に、動かすスピン S_i を規則的に選ぶタイプのメトロポリス法を適用すると周期 2 の振動が起きて緩和しない [緩和が遅い] 。この種の原因による遅い緩和は滅多に起きないが、起きたときには驚かされることになる。

A.3　任意の統計量の緩和 (平衡緩和)

本節では、任意の統計量の時間相関を各緩和モード V_i の寄与に分解する式を示し、導出する。以下では、サンプル α_0 を採取する時刻 $t = 0$ 以前に十分多くのサンプルを捨てているものと仮定して、その後の時刻 $t = m\delta\,(m = 1, 2..M)$ に生成されるサンプルの列 $\{\alpha_m\}$ に関する期待値を考えることにする。これは、初期状態 α_0 を P^∞ からのランダムサンプルとして、生成される列 $\{\alpha_m\}$ に関する期待値を考えるのと同等である。(標準的記法ではないが) この意味の期待値を 2 重括弧 $\langle\langle\quad\rangle\rangle$ で記すことにする。1 重括弧 $\langle\quad\rangle$ は従来通り P^∞ に関する期待値 (母平均) である。この記号を用いるとわれわれの求めるものは
$\langle\langle A(\alpha_0)A(\alpha_m)\rangle\rangle$ と書ける。

すると、求める表式は、

$$\langle\langle A(\alpha_0)A(\alpha_m)\rangle\rangle = \langle A\rangle^2 + \sum_{i=1,2..}(\sum_\alpha A(\alpha)V_i^\alpha)^2\exp(-m\delta/\tau_i) \tag{A.5}$$

となる。A.1節の (A.4) 式を使うと上式は、

$$\frac{\langle\langle A(\alpha_0)A(\alpha_m)\rangle\rangle - \langle A\rangle^2}{\langle A^2\rangle - \langle A\rangle^2} = \frac{\sum_{i=1,2..}(\sum_\alpha A(\alpha)V_i^\alpha)^2\exp(-m\delta/\tau_i)}{\sum_{i=1,2..}(\sum_\alpha A(\alpha)V_i^\alpha)^2} \le \exp(-m\delta/\tau_1) \tag{A.6}$$

とも書ける。$\sigma_0^2 = \langle A^2\rangle - \langle A\rangle^2$ は分布 P^∞ のもとでの A の分散 (母分散) である。(A.6) 式から、はじめの方を十分捨てたあとでは、統計量 A の "規格化された" 時間相関は最長の緩和時間できまる指数関数 $\exp(-t/\tau_1)$ よりも速く減衰することがわかる。従って、統計量 A の時間相関が無視できるようになる時間 $\widetilde{\tau}_A$ は τ_1 の数倍より小さいことになる。(A.5) 式の右辺が複数の指数関数の和になっているので、統計量 A の時間相関が $1/e$ になる時間 τ_A と $\widetilde{\tau}_A$ の関係は単純でないことを注意しておく。

(A.5),(A.6) の導出は以下の通りである。まず、初期状態 $P(0)$ として α_0 の上のデルタ分布をとった場合の時間発展 $P_\alpha(t)$ を本文の (1.6) 式を用いて書く。A.1節の (A.3) 式から係数 a_i を決めると

$$a_i = V_i^{\alpha_0}(P_{\alpha_0}^\infty)^{-1} \tag{A.7}$$

となるから、(1.6) 式を用いて、$\lambda_i = -1/\tau_i$ に注意すると、

$$P_\alpha(t) = P_\alpha^\infty + \sum_{i=1,2..} V_i^{\alpha_0}V_i^\alpha(P_{\alpha_0}^\infty)^{-1}\exp(-t/\tau_i) \tag{A.8}$$

と求まる。ここで定義から、

$$\langle\langle A(\alpha_0)A(\alpha_m)\rangle\rangle = \sum_{\alpha_0}\sum_\alpha A(\alpha_0)A(\alpha)P_\alpha(m\delta)P_{\alpha_0}^\infty \tag{A.9}$$

と書けるから、これに (A.8) 式で $t = m\delta$ としたものを代入して整理すると (A.5) 式を得る。

A.4　サンプル平均の分散の表式

本節では、サンプル平均の分散 σ_{MC}^2 を各緩和モード V_i からの寄与に分解する式を示し、導出する。この式の重要性は緩和時間スペクトル $\{\tau_i\}$ と σ_{MC}^2 を概念的に関係づけている点にある。実際の問題では緩和時間スペクトル $\{\tau_i\}$ も、問題の統計量の各緩和モードへの射影 $\sum_\alpha V_i^\alpha A(\alpha)$ も分からないので、この公式から定量的な情報を得ることは できない ことを注意しておく。

まず、求める分散 σ_{MC}^2 をきちんと定義する必要がある。ここでは、十分多くのサンプルを捨てたあとのサンプル列を用いてサンプル平均を計算するとして、前節の記号を使って

$$\sigma_{MC}^2 = \langle\langle(\frac{1}{M}\sum_{m=1}^M A(\alpha_m))^2\rangle\rangle - \langle\langle\frac{1}{M}\sum_{m=1}^M A(\alpha_m)\rangle\rangle^2 \tag{A.10}$$

と定義することにする。この定義のもとで、求める表式は、シミュレーションの長さ $M\delta$ が十分長いとき、

$$\sigma_{MC}^2 \simeq \frac{1}{M}\sum_{i=1,2..}(\sum_\alpha A(\alpha)V_i^\alpha)^2\frac{1+\exp(-\delta/\tau_i)}{1-\exp(-\delta/\tau_i)} \tag{A.11}$$

あるいは、

$$\sigma_{MC}^2 \simeq \frac{1}{M}(\langle A^2\rangle - \langle A\rangle^2) + \frac{1}{M}\sum_{i=1,2..}(\sum_{\alpha} A(\alpha)V_i^{\alpha})^2 \frac{2\exp(-\delta/\tau_i)}{1-\exp(-\delta/\tau_i)} \tag{A.12}$$

となる。

因子 $(1+\exp(-\delta/\tau_i))/(1-\exp(-\delta/\tau_i))$ は複雑に見えるが、 $\tau_i \gg \delta$ ならば、 $\sim 2\tau_i/\delta$ となる。この場合、 "i 番目の緩和モードの寄与は、 V_i に対しての統計量 $A(\alpha)$ の射影の 2 乗と緩和時間 τ_i の積に比例する" という解釈ができる (ここでいう "射影" は内積の定義 (A.2) の最後の因子を含まない)。特別の場合として、最長の緩和時間に対応する緩和モード V_1 への射影が対称性から零になる場合がある。このとき、問題の統計量 A のサンプル平均の分散に対しては、最も遅い緩和モードの寄与が完全になくなる。この実例は第 2 部の 2.6節の 1項を参照されたい。

以下、式 (A.11),(A.12) の導出を与える。時間並進に対する不変性

$$\langle\langle A(\alpha_{m_1})A(\alpha_{m_2})\rangle\rangle = \langle\langle A(\alpha_0)A(\alpha_{(m_2-m_1)})\rangle\rangle \tag{A.13}$$

及び $\langle\langle A(\alpha_m)\rangle\rangle = \langle A\rangle$ を利用すると、

$$\sigma_{MC}^2 = \frac{1}{M^2}\{M\langle A^2\rangle + 2\sum_{m=1}^{M-1}(M-m)\langle\langle A(\alpha_0)A(\alpha_m)\rangle\rangle\} - \langle A\rangle^2 \tag{A.14}$$

となる。この式に (A.5) 式を代入し、 $M\delta \gg \tau_1$ と仮定して等比級数の和をとると、求める式 (A.12) が得られる。

A.5 サンプル間隔とサンプル平均の分散

シミュレーションの長さ (MCS 数) を同じにしたとき、サンプルをとる間隔が長いほどサンプル平均の分散は増大する (ただし、はじめの方のサンプルは十分捨て、サンプルは等しい間隔でとると仮定する)。この事実は、 (A.11) 式を使うと簡単に示すことができる (以下の議論は Geyer (1992) に基づく)。

シミュレーションの長さが同じであるから、サンプルをとる間隔が δ_1 のときのデータの個数 M_1 を M_0/δ_1 とすると、間隔が δ_2 のときのデータの個数 M_2 は M_0/δ_2 となる。 (A.11) 式より、それぞれの場合の分散は、

$$\sigma_1^2 = \frac{\delta_1}{M_0}\sum_i(\sum_{\alpha} V_i^{\alpha}A(\alpha))^2 \frac{1+\exp(-\delta_1/\tau_i)}{1-\exp(-\delta_1/\tau_i)} \tag{A.15}$$

$$\sigma_2^2 = \frac{\delta_2}{M_0}\sum_i(\sum_{\alpha} V_i^{\alpha}A(\alpha))^2 \frac{1+\exp(-\delta_2/\tau_i)}{1-\exp(-\delta_2/\tau_i)} \tag{A.16}$$

となる。ところが、 $\delta_1 > \delta_2$ のとき、各 i について

$$\delta_1 \frac{1+\exp(-\delta_1/\tau_i)}{1-\exp(-\delta_1/\tau_i)} > \delta_2 \frac{1+\exp(-\delta_2/\tau_i)}{1-\exp(-\delta_2/\tau_i)} \tag{A.17}$$

が示せるので、 $\sigma_1^2 > \sigma_2^2$ となる。どんな短い時間間隔で取ったサンプル対にも、僅かなりとも互いに独立な情報が含まれていることを考えれば、これは不思議ではないだろう。

A.6 非平衡緩和

いままで、初期状態 α_0 が P^{∞} からのランダムサンプルの場合を主に扱ってきたが、最後に α_0 についての平均をとらない場合を考えよう。これは統計物理でいえば非平衡状態からの緩和を考えることに相当する。

以下では、 α_0 を固定して、 α_0 から出発する間隔 δ の列 $\{\alpha_m\}(m=1..M)$ に関する平均を $\langle\langle\quad\rangle\rangle_{neq}$ で表そう。すると、 $\langle\langle A(\alpha_m)\rangle\rangle_{neq} - \langle A\rangle$ は時刻 $m\delta$ での統計量 A の バイアス を示すことになる。これは (A.8) 式から計算できて、

$$\langle\langle A(\alpha_m)\rangle\rangle_{neq} - \langle A\rangle = \sum_{i=1,2..}\frac{V_i^{\alpha_0}}{P_{\alpha_0}}(\sum_{\alpha} V_i^{\alpha}A(\alpha))\exp(-m\delta/\tau_i) \tag{A.18}$$

となる。この場合も右辺が $\exp(-t/\tau_1)$ の定数倍で押えられることには変わりないが、初期状態によっては定数因子が非常に大きくなる可能性がある。

付録 B

実装上の基礎知識

- 場合分け

  メトロポリス法の場合、確率の比 $P(\gamma)/P(\alpha)$ を直接に計算するより、その対数である “エネルギーの差”$\triangle E_{\alpha\gamma} = \log(P(\gamma)/P(\alpha))$ を計算するのが普通である。この場合、 $\triangle E_{\alpha\gamma} > 0$ の場合を分けることが多い。条件文がひとつ増えるかわり、乱数と exp の計算が回避できる。また、これによって、無用の overflow が exp の計算で生じるのを防げる場合がある。計算機によっては、 underflow を防止するために、 exp の中身が小さ過ぎるときの場合わけも必要になる。

- 乱数

  乱数を複数の手続きに使用するので、乱数に系列相関があるとおかしな結果になることがある。実際には、乱数関係でおこるひどい失敗は、初歩的なプログラムミスによるものが多い。

  よく聞く間違いの例。以下では、 IR を $\{1..M\}$ の整数乱数、 RND を $(0,1)$ の実数乱数とする。

  - IR を下位ビットがランダムでない整数乱数 IRND から IR=MOD(IRND,M)+1 として作ってしまった。
  - IR を RND から IR=INT(RND*M)+1 として作ったところ、 IR=M+1 となる場合が生じた (これはプログラムの移植のさいに問題になることがある)。
  - 単精度実数を供給する乱数ルーチンに、間違って倍精度の配列を渡した。

  乱数ルーチンは最も内側のループで呼ばれるので、呼びだしのオーバーヘッドが大きい。乱数をまとめてとるか、 inline 展開すると早くなる。

- バグ

  マルコフ連鎖モンテカルロ法は、プログラムが簡単な代わり、バグがわかりにくいという特性がある。これは乱数を使うためもあるが、ひとつひとつは簡単なステップがより集まって複雑な効果を生ずるというアルゴリズムの特性によるところが大きい。もし、小さい系で厳密な計算との比較が可能なら、はじめにやっておいた方が良い。

- プログラムの移植

  移植して結果が変わる場合には本当に間違っている場合とそうでない場合がある。本当に間違っている場合は、疑似乱数ルーチンの移植性がないためであることが多い。本当は間違っていない場合も多い。これは、マルコフ連鎖モンテカルロ法のアルゴリズムが実数比較を含むことから生ずる。とくに、場合分けによって乱数を節約した場合は、乱数列がひとつ分ずれることで、それ以降まったく違う状態列を生じて驚かされることがある。