有限混合分布モデルによる有限温度の準安定状態の表現

統計数理研究所　　伊庭 幸人[1]
東京大学 物性研究所　　福島 孝治[2]

1　はじめに

本研究の目的は，有限温度のシミュレーションデータから “準安定状態” を検出する，あるいは，構成することである．最適化 (温度 $\to 0$) における local optima を定義することは難しくないが，有限温度となると話は別で，有限系の準安定状態をシミュレーションから構成するための確立された手段はないと思う．スピングラス理論では，平均場近似を基盤として考えることが多いので，このことはあまり問題にされないのかもしれないが，有限系のシミュレーションにリアリティを感じるシミュレーション屋としては考えてみたい問題である．

ここでは，この問題に対して有限混合分布モデル **Finite Mixture Model** によるアプローチを試みる．この方法では「準安定状態」は確率分布 (ギブス分布) からのサンプルの集合に対して定義され，サンプルの順序にはよらない．この意味で，これは静的な定義である「準安定状態」はもっとダイナミカルに定義されるべきであるという考えもあると思うが，ここでは静的な範囲でどのようなことができるかを考えたい．

筆者の一人 (伊庭) は従来から統計物理と統計学の境界領域に興味を持ってきた．その文脈では，本研究は，(1) 統計物理の手法や考え方 $\Rightarrow$ 統計学，(2) 統計学の手法や考え方 $\Rightarrow$ 統計物理，の 2 方向のうち，後者 (2) の例とみなせる．

2　Finite Mixture Model(有限混合分布モデル)

有限混合分布モデルとは P_c $(c = 1...c_{max})$ を簡単な分布，λ_c をそれぞれのパラメータとするときに，その混合

$$P(y) = \sum_c w_c P_c(y|\lambda_c), \qquad \sum_c w_c = 1 \tag{1}$$

のことである．これらは，グループ分けやクラスター分析の道具として広く使われている [1]．また，教師なし学習のモデルの一つとして，artificial neural network の分野での研究対象にもなっている [2]．ここで，P_c は一般に「クラスター」「テンプレート (を中心としたかたまり)」「プロトタイプ (を中心としたかたまり)」と解釈されるが，それを「準安定状態」と考えようというのがわれわれの提案である．ここで，一般に，

$$p_i(c) = \frac{w_c P_c(y|\lambda_c)}{\sum_c w_c P_c(y|\lambda_c)} \tag{2}$$

という量は，クラスター c (準安定状態 c) へのサンプル i の「所属度」をあらわすと考えられる．$\sum_c p_i(c) = 1$ となることに注意されたい．

[1] E-mail:iba@ism.ac.jp
[2] E-mail:hukusima@chiral.issp.u-tokyo.ac.jp

有限温度のシミュレーションデータに有限混合分布をあてはめる素朴な方法は，準安定状態の中心 (テンプレート) を local optima のひとつにとり (あるいは物理的な考察からあらかじめ選んでおき)，その周辺のサンプルから各成分 P_c のパラメータ λ を推定するというものである．ここでは，もう一歩すすめてこれらをすべて同時に，データとテンプレートの個数以外の予備知識なしに，自己無矛盾に決定することを試みる．そのための方法については，あとで述べる．

有限混合分布の応用で，もっともよく使われるのは P_c たちが (多変量) ガウス分布である場合 (Gaussian Mixture) [3] であるが，あとの例では，離散変数の分布 (イジング模型) などを考えるので，それ用のモデルを開発した．次のように記号を定める．

定数
i_{max}: サンプル数 (自然数)
j_{max}: 各データの成分数 (= スピンの数)(自然数)
c_{max}: テンプレート数 (自然数)
パラメータ
x_c^j: クラスター c のテンプレートの j 成分 (実数)
w_c: クラスター c の大きさ (正実数)
データ
d_i^j: i 番目のサンプルの j 成分 (成分が ± 1 のベクトル)

すると，ここで解析に使うモデルは，

$$P(\{\{d_i^j\}\}\,|\,\{\{x_c^j\}\},\{w_c\}) = \prod_i \left\{ \sum_c w_c \frac{\exp(\sum_j x_c^j\, d_i^j)}{Z_c} \right\} \tag{3}$$

$$Z_c = \Pi_j \{\exp(+x_c^j) + \exp(-x_c^j)\} \tag{4}$$

と書ける．ここで，c 番目のテンプレート $\{x_c^j, j = 1..j_{max}\}$ は実数値をとるので，Gaussian Mixture でいえば各成分の分散共分散行列に相当する部分も含んでおり，物理的には「準安定状態 c での平均場」のようなものである．但し，われわれの目的はシミュレーションの output の解析であって「平均場近似」ではないことに注意されたい．w_c はテンプレート c に所属するサンプルの割合 (所属度の重みつきで計算したもの) である．あるいは，準安定状態 c のサイズといってもよい．

上のモデルは，テンプレート c でのスピン j に作用する「平均場」x_c^j のかわりに「平均磁化」$m_c^j = \tanh x_c^j$ を用いて書くこともできる．

$$P(\{\{d_i^j\}\}\,|\,\{\{m_c^j\}\},\{w_c\}) \;=\; \prod_i \left\{ \sum_c w_c\, \Pi_j \frac{1 + d_i^j m_c^j}{2} \right\} \tag{5}$$

このほうが，発散が起らず，数値計算上有利なことが多い，

3　EM アルゴリズムによるあてはめ

データ $\{\{d_i^j\}\}$ とテンプレートの個数 j_{max} のみ与えて，テンプレート (平均場の方向と強さの両方)，各テンプレートに所属するサンプルの割合を同時に推定するには，(3) で定義される $P(\{\{d_i^j\}\}\,|\,\{\{x_c^j\}\},\{w_c\})$ を，パラメータ $\{\{x_c^j\}\},\{w_c\}$ について最大化すればよい (最尤推定)．具体的には次の式を解くことになる．

$$\frac{\partial \log P}{\partial x_c^j} = 0 \tag{6}$$

$$\frac{\partial}{\partial w_c}(\log P - \lambda \sum_b w_b) = 0 \tag{7}$$

ここで，λ は条件 $\sum_c w_c = 1$ に対応する未定乗数，P は $P(\{\{d_i^j\}\} \,|\, \{\{x_c^j\}\}, \{w_c\})$ の略である．上の微分を計算することにより，上式を解くためのアルゴリズム (有限混合分布に対する EM アルゴリズム) が導かれる．ここで $p_i(c)$ は計算のための中間変数であるが，先に述べたように，サンプル i のテンプレート c に対する所属度とも解釈することができる．

EM アルゴリズム：適当な初期条件からはじめて以下を収束するまで反復する．

- **E-step**

$$p_i(c) := \frac{w_c \Pi_j \frac{1+d_i^j m_c^j}{2}}{\sum_b w_b \Pi_j \frac{1+d_i^j m_b^j}{2}} \tag{8}$$

- **M-step**

$$m_c^k := \frac{\sum_i d_i^k \cdot p_i(c)}{\sum_i p_i(c)} \tag{9}$$

$$w_c := \frac{\sum_i p_i(c)}{d_{max}} \tag{10}$$

EM アルゴリズムは最尤推定値を求めるための最適化の一手法として導かれるものであるが，ある種の「自己組織的」分類手法としても解釈できる．すなわち，E-step においては，テンプレート m_c^j と各テンプレートの重み w_c を与えて，各サンプルのそれぞれについて，各テンプレートへの所属度 $p_i(c)$ を計算する．M-step においてはサンプルをそれぞれの所属度を重みとして平均することでテンプレートを更新し，また所属度の和に応じてテンプレートの重みを更新する．これをくりかえすことで，テンプレートと分類が自然に組織化されて，テンプレートおよびその分類の規準 (重み) と各サンプルの分類 (所属度) が同時に求まることを期待するわけである．

4 $+-$ 対称性を仮定したモデル

有限混合分布モデルに基づく方法の利点として，問題の対称性を比較的容易にとりこむことができるという点がある．たとえば，シミュレートされるモデル (データを供給するモデル) が $d_i^k \to -d_i^k$ という $+-$ 対称性 (スピン反転対称性) を持っているなら，準安定状態も $+-$ 対称性で結ばれた対で出現するはずである．これは，データを解析する有限混合分布モデルにおいては，あるテンプレートに対して，これを符号反転したものが存在することに対応している．解析が狙いどおりにいっているならば，わざわざ仮定しなくても，符号反転で結ばれたテンプレートが対で出現することが期待されるが，逆に対称性を仮定したモデル化を行えば，より安定な解析が可能になる．いまの場合には，$+-$ 対称性を仮定したモデルは，

$$P(\{\{d_i^j\}\} \,|\, \{\{x_c^j\}\}, \{w_c\}) = \prod_i \left\{ \sum_c w_c^+ \frac{\exp(+\sum_j x_c^j d_i^j)}{Z_c} + \sum_c w_c^- \frac{\exp(-\sum_j x_c^j d_i^j)}{Z_c} \right\} \tag{11}$$

$$Z_c = \Pi_j \{\exp(+x_c^j) + \exp(-x_c^j)\} \tag{12}$$

と表現できる (ここでは w_c については対称性を仮定していないが，さらに $w_c^- = w_c^+$ と置くこともできる．)．このモデルも，前と同様に，EM アルゴリズムによってあてはめることができる．たとえば，

$$m_c^j = \tanh x_c^j \tag{13}$$

についての更新の式 (M-step) は

$$m_c^k := \frac{\sum_i d_i^k \cdot (p_i^+(c) - p_i^-(c))}{\sum_i (p_i^+(c) + p_i^-(c))} \tag{14}$$

と修正される．ここで，$p_i^+(c)$, $p_i^-(c)$ は，それぞれ，テンプレート c とそれを $+-$ 反転したテンプレートに対するサンプル i の所属度であって，E-step で，

$$p_i^+(c) := \frac{w_c^+ \Pi_j \frac{1+d_i^j m_c^j}{2}}{\sum_b w_b^+ \Pi_j \frac{1+d_i^j m_b^j}{2}} \tag{15}$$

$$p_i^-(c) := \frac{w_c^- \Pi_j \frac{1-d_i^j m_c^j}{2}}{\sum_b w_b^- \Pi_j \frac{1-d_i^j m_b^j}{2}} \tag{16}$$

と計算される．

off-lattice のヘテロポリマーなどを扱うためには，3 次元回転と重心移動についての不変性を取り入れたモデルを作ることが望ましいが，あてはめの計算が困難になると思われる．このような場合に適用できる実用的な手法については，さらに研究が必要である．

5 数値実験

まず，スピングラスの SK 模型

$$P(\{x_i\}) = \frac{\exp(-\beta E(\{x_i\}))}{Z},\ x_i \in \pm 1,\ Z \text{ は規格化定数 (分配関数)} \tag{17}$$

$$E(\{x_i\}) = \sum_{(i,j)} J_{ij} x_i x_j \tag{18}$$

$$J_{ij} \leftarrow \textbf{ i.i.d. } N(0, \frac{\sigma^2}{N}) \tag{19}$$

を解析の対象として考える．

このモデル (解析対象) に対してモンテカルロ・シュミレーション (交換モンテカルロ法を使用) で分布からのサンプリングを行い，そのサンプル (snapshot) をデータとした．逆温度 β は σ^2 を単位として 2，サンプル数は全部で 2000 個である．用いたサンプルの一部を図 1 に示した．交換モンテカルロ法を使用しており，サンプル間隔が 100MCS なので，図 1 で隣接するサンプル間にはすぐ見て取れるような相関はない．このこともあって，図 1 を一瞥して準安定状態を見出すのはかなり困難なようにも思うが，どうであろうか．

このデータを，テンプレート数 4 の $+-$ 対称性を仮定した有限混合分布モデル (11) で解析した結果が，図 2 と図 3 である．図 2 には得られたテンプレート 4 枚を示した．また，ここでは示さないが，$+-$ 対称性を仮定しないモデル (3) を用いて解析しても，ほぼ対応するテンプレートが得られることも確かめられた．ただし，図 2 の上で，対となるテンプレートの一方が他に比べて全体に一様に縮んでいる (m_c^j の絶対値が小さくなっている) のが気になるところである．

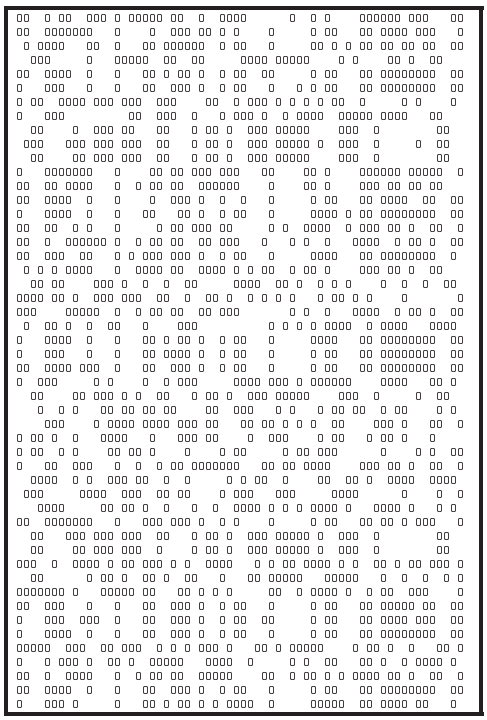


図 1: モンテカルロ法によって得られたサンプルの例。横方向に 64 サイトのスピン変数の”+1”を O で、“-1” を空白で表し、縦方向に各サンプリングを示した。

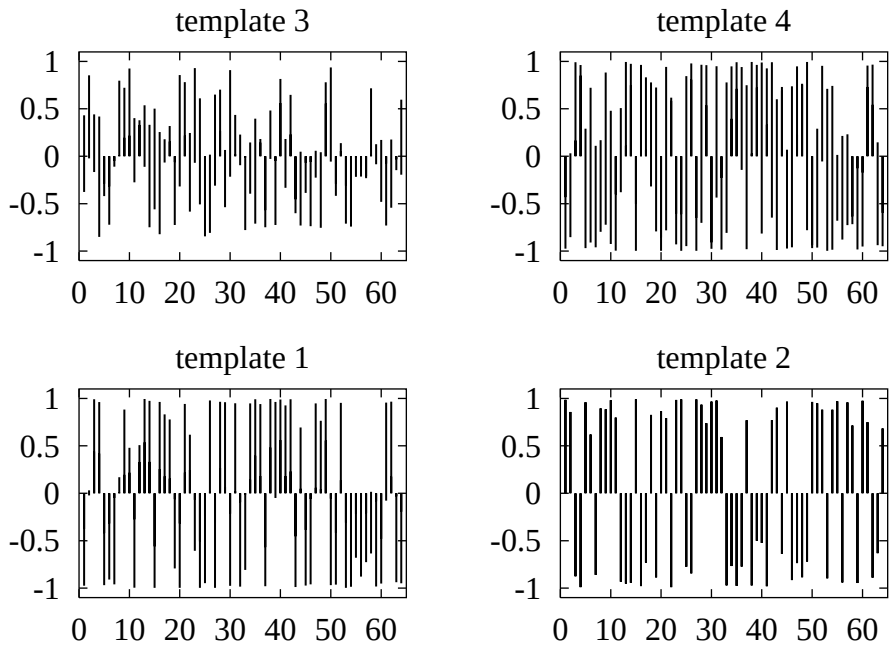


図 2: SK 模型における各テンプレートでのサイトごとの平均磁化.

各サンプル i を，所属度 $p_i(c)$ が最も大きい準安定状態 c に帰属させるという方法でサンプルを分類することができる．図 3 には各テンプレートごとにそれに帰属されたサンプルの一部を示した「縮んだ」テンプレートに対応する (c) を除いて，(a)(b)(d) では，もっともらしい結果が得られていることがわかる．

SK 模型の代わりに「答」がほぼわかっている問題として，Hopfield 模型 (対称結合連想記憶), すなわち，(17),(18) で

$$J_{ij} = \frac{1}{N}\sum_k \xi_i^k \xi_j^k \tag{20}$$

としたものを考える．スピンの数は 64 個で，記憶させたパターンは 4 個，記憶させる 2 値パターン $\{\xi_i\}^k$, $k = 1..k_{max}$ は

$$\xi_i^k \leftarrow \textbf{ i.i.d } \frac{1}{2}\delta(\xi - 1) + \frac{1}{2}\delta(\xi + 1) \tag{21}$$

のように，ランダムに生成させたものを用いた．

このモデル (逆温度 β=2) に対して，モンテカルロシュミレーション (交換モンテカルロ法，100MCS ごとにサンプル採取) を行って得たサンプル 1000 個に対して，テンプレート数 4 の有限混合分布モデル (11) を EM アルゴリズムであてはめた．得られたテンプレートが図 4 である．記憶させたパターンとその $+-$ 反転したものががほぼテンプレートとして再現されているが，左上のテンプレートでやはり「縮み」が見られる．

6 問題点

- テンプレートの “縮み”
  例で見られた，テンプレートが全体に一様に縮む現象は問題である．レプリカ間の重なりの分布を，データにあてはめた有限混合分布モデルを利用して評価すると，もとのデータから直接計算したものと合わないが，これにも「縮み」の現象が関わっているように思われる．

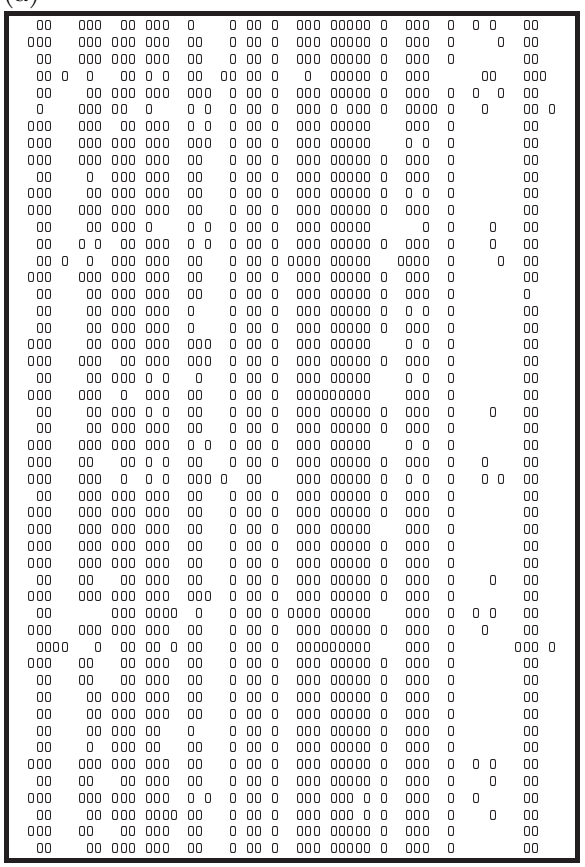
  現象的には，どのテンプレートにも本来属さないサンプルが特定のテンプレートに集まってくることが「縮み」の原因であるように思われる．したがって「ごみ箱」的な役割を果たす，すべての成分 k について $m_0^k = 0$ であるようなテンプレートをあらかじめ入れることで，問題を解決することが考えられる．

  実際に試してみると，Hopfield 模型では「ごみ箱」法で縮みが解消される例があったが，一般にはうまくいかないようである．より深い原因．おそらく，ひとつのテンプレートで表現される準安定状態の中でのスピン相関が無視できないということが関係しているのではないかと考えている．

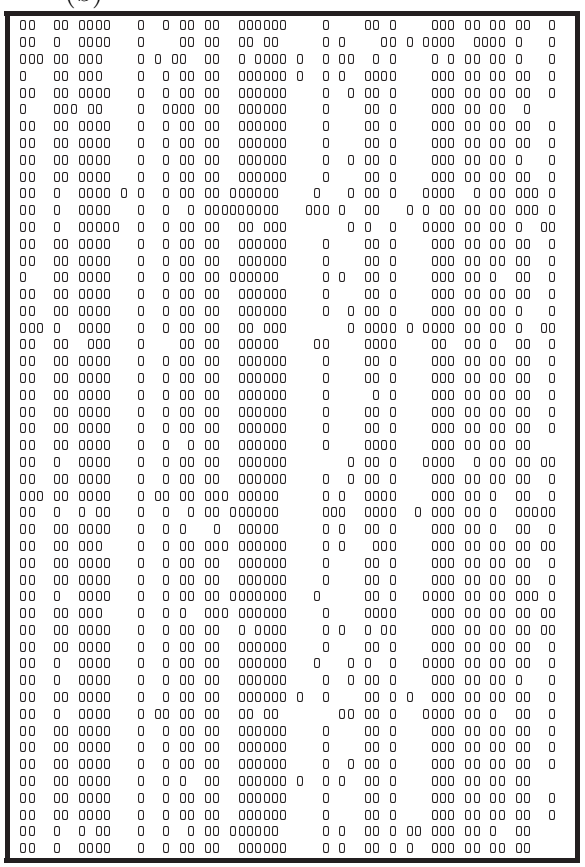
- 準安定状態の数の決定
  当然問題になるのが，与えられた有限個のデータから「見える」準安定状態の妥当な数を決定する問題である．有限系の問題では，あらかじめ定義された準安定状態というものがあるわけではなく，むしろ，有限混合分布のあてはめが，それを定義する試みであるともいえる．しかし，同数のデータを別の乱数を利用したシミュレーションで生成した場合に，毎回別のテンプレートが検出されるとしたら，明らかにテンプレートの数が多すぎると考えてよいだろう (必要条件としての安定性)．

  これと似た原理に基づく考えが，統計学やニューラルネットの学習で最適パラメータ数の決定に使われている Cross Validation(交差確認法) である．この方法では，データに当ては

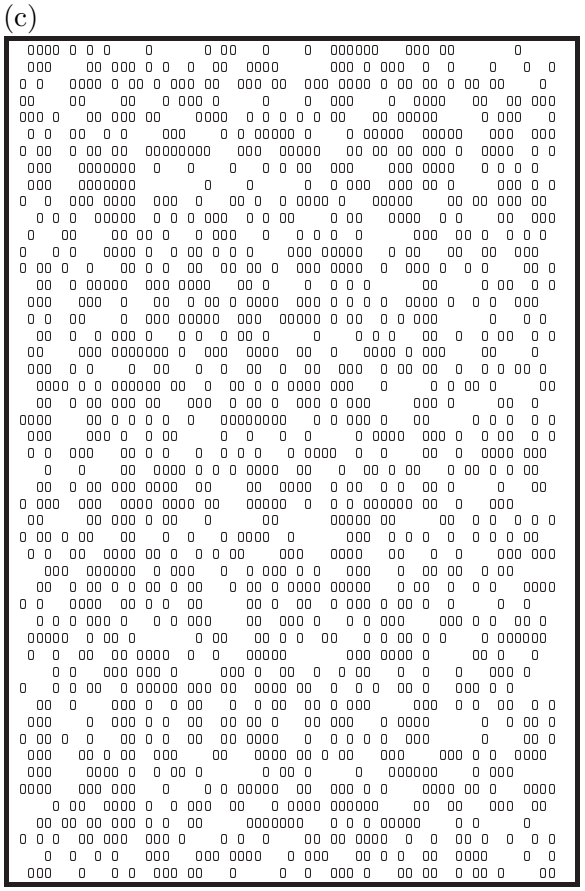
(a)



(b)



(c)



(d)

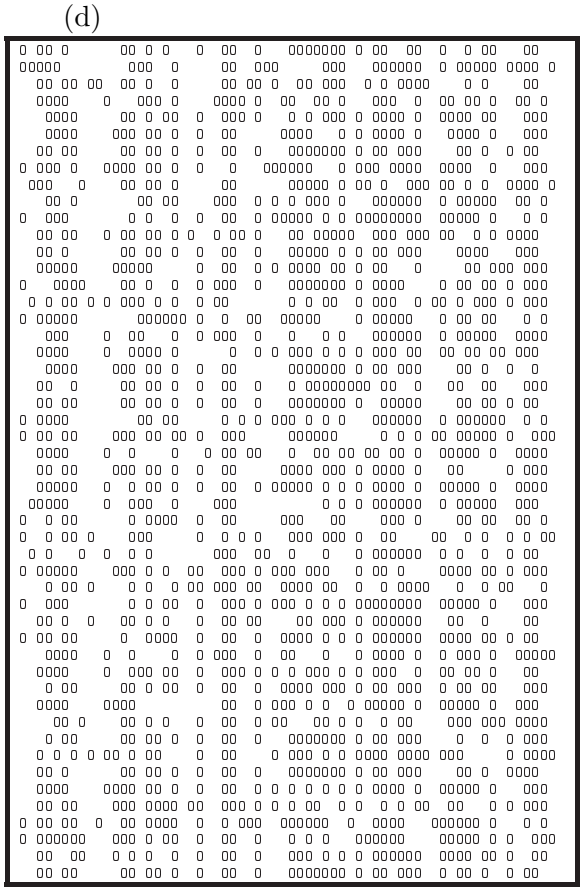


図 3: 各テンプレートに帰属されたサンプルの一部

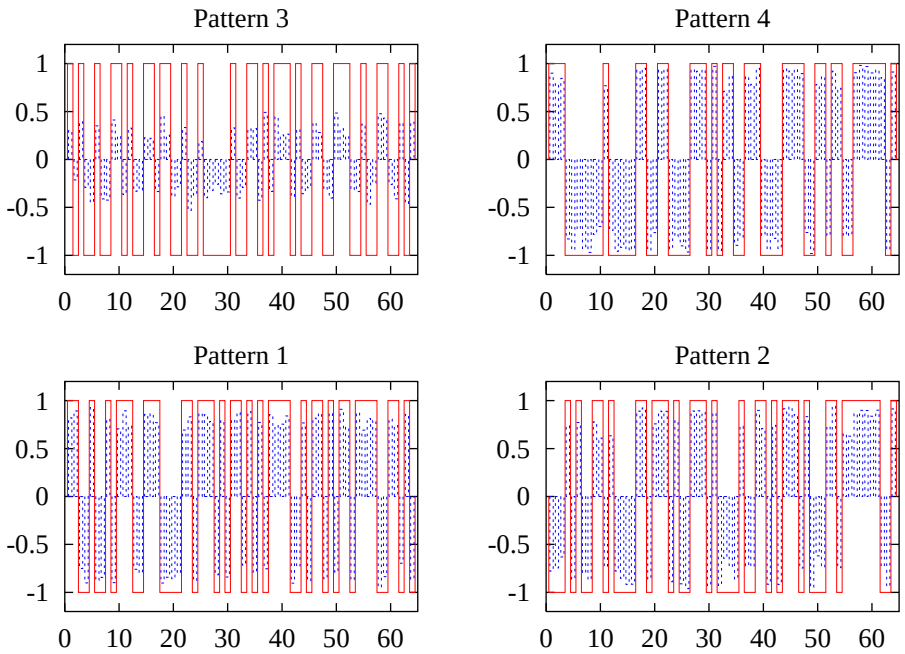


図 4: Hopfield 模型における各テンプレートでのサイトごとの平均磁化 (破線)．実線は記憶させたパターン ．

めたモデルを，それとは独立なデータを用いて検証し，そのデータに対してもっとも高い尤度 (など) を与えるパラメータ数のモデルを選択する．これはモデルの予測能力を測っていることになる．普通は，このために，データをいくつかの組 (学習用データと検証用データ) に分けたりするが，いまの場合はデータがシミュレーションによっていくらでも製造できるので簡単である．

ところが，奇妙なことに，交差確認法を今の例に適用するとテンプレートの数が相当に増加しても，モデルの予測能力がどんどん上がっていくという結果が得られる．生成されたテンプレートを調べてみると，ベクトルとしての向きがほぼ同じで，大きさだけが異なるものが複数含まれているようである．先の表現でいえば「縮んだ」テンプレートがもとのものに重複して出現する傾向があるということになる．これも，おそらく準安定状態内の相関の効果ではないかと思われるが，まだよくわかっていない．

参考文献

[1] 有限混合分布モデルについてのテキストは，★ D. M. Titterington, A. F. M Smith, and U. E. Makov: *Statistical Analysis of Finite Mixture Distributions* (John Wiley and Sons, Chichester, 1985). 本稿には直接関係ないが，ついでに，動的モンテカルロ法を用いた最新の研究をあげると．★ S. Richardson and P. J. Green: Bayesian analysis of mixtures with an unkown number of components, Journal of Royal Statistical Society B **59** (1997) 731 [with discussion].

[2] 統計物理の手法による有限混合分布モデル「の」研究 (~ を用いた研究，ではなくて) には，たとえば以下のようなものがある．最近の発展についてはよく知らない．★ N. Barkai and

H. Sompolinsky: Statistical mechanics of the maximum-likelihood density estimation, Phys. Rev. E **50** (1994) 1766. ★ T. L. H. Watkin and J-P. Nadal: Optimal unsupervised learning, J. Phys. A **27** (1994) 1899. ★ M. Biehl and A. Mietzner: Statistical mechanics of unsupervised structure recognition, J. Phys. A **27** (1994) 1885.

[3] 研究会が終わってから知ったが，以下の論文でタンパク質の準安定状態を記述するために使われているモデルは，Gaussian Mixture と本質的に同じだと思われる．ただし，EM アルゴリズムのようなテンプレートとそのパラメータ，サンプルの帰属を同時に行う手法は使われていないようである．論文を教えてくださった郷先生に感謝します．★ A. Kitao, S. Hayward and N. Go, PROTEINS, 33:496 (1998)