

メトロポリス的モンテカルロ法の 巨視的パラメータ推定への応用

—— 2次元イジング模型の場合 ——

統計数理研究所 伊 庭 幸 人

(1991年4月 受付)

1. 本論文の内容

最近、通常の方法では ABIC (type II likelihood, 大規模混合モデルの周辺尤度)の計算が容易でない場合に、メトロポリス的モンテカルロ法 (Metropolis et al. (1953), Binder (1986)) を利用してその最適化を行い、巨視的パラメータ (ハイパーパラメータ (注 1)) を決定することが試みられている (Geman and McClure (1987), Ogata (1990), Ohtsuki and Kawato (1991)).

本論文では、2次元イジング模型の場合について、ABIC とメトロポリス的モンテカルロ法を利用した巨視的パラメータ推定の実験を報告する。結果がわかっている問題を扱い、推定結果の分布などを調べることによって、所与の方法の有効性を確かめるのが目的である。

2次元イジング模型は2値画像のマルコフ場モデルとして典型的なものである。格子上的マルコフ場モデルは、生態学等での空間パターンの特徴付け (Besag (1974)), texture の表現 (Cross and Jain (1983), Geman et al. (1987)), 画像復元 (Geman, S. and Geman, D. (1984), Derin et al. (1984), Besag (1986), Derin and Elliott (1987), Devijver and Dekesel (1987)) など広い範囲に応用されている。

磁場なしの2次元イジング模型にメトロポリス的モンテカルロ法を適用した場合のマルコフ鎖のダイナミクスに関しては、Miyashita and Takano (1985) に詳しい (ただし狭義のメトロポリス法でなく、熱浴法 (Gibbs sampler) が用いられているので注意)。

本稿に関連した2次元イジング模型に関する研究は、尤度関数が多少違っているが、Besag et al. (1991) にも含まれている。また、これに続く Besag の rejoinder (discussion に対する回答) では Qian and Titterton の研究 (投稿中とのこと) を引用してさらに議論がなされている。Besag らの論文には、簡単な実例 1 つに関する結果が示されているだけで、答のなかった場合の詳しいテストはなされておらず、もちろん推定結果の分布に関して触れられていない。また、方法的にも pseudo-likelihood を用いるなど、本稿とは異なっている (8.1 節参照)。いわゆる線過程 (line process, Geman, S. and Geman, D. (1984)) を含む場合に不完全データから巨視的パラメータを推定する問題は、ATR 視聴覚研究所のグループによって研究されている (Ohtsuki and Kawato (1991))。

2. 一般論

以下で述べる方法 (Geman, D. and Geman, S. (1986), Geman and McClure (1987)) は, メトロポリスのモンテカルロ法によって, 事前分布及び事後分布での期待値が計算できることに基づくものである. この方法は, 観測雑音の混入していない場合にメトロポリスのモンテカルロ法を利用して尤度を最大化するアルゴリズム (Penttinen (1984), Ogata and Tanemura (1981, 1984, 1985, 1989), 本郷 他 (1991)) を観測雑音がある場合 (一般に “不完全データ” の場合) に拡張したものととも考えられる.

微視的パラメータ (画素またはスピン (注1)) を $\{x\}$ とする. 事前分布 π と尤度関数 L が

$$(2.1) \quad \pi(\{x\}) = \frac{\exp(-E_\pi(\{x\}))}{\sum_{\{x\}} \exp(-E_\pi(\{x\}))}$$

$$(2.2) \quad L(\{y\} | \{x\}) = \frac{1}{Z_L} \exp(-E_L(\{y\}, \{x\}))$$

のようにギブス分布の形に書かれているとする. ここで, Z_L は尤度関数の $\{y\}$ に関する規格化定数である. 以下では, Z_L が $\{x\}$ に依存しないことを仮定する.

このとき, 事後分布は

$$(2.3) \quad E_{\text{pos}}(\{x\}) = E_\pi(\{x\}) + E_L(\{y\}, \{x\})$$

を用いて,

$$(2.4) \quad P_{\text{pos}}(\{x\}) = \frac{\exp(-E_{\text{pos}}(\{x\}))}{\sum_{\{x\}} \exp(-E_{\text{pos}}(\{x\}))}$$

となる. この場合,

$$(2.5) \quad -\frac{1}{2} ABIC = \log \sum_{\{x\}} L(\{y\} | \{x\}) \pi(\{x\})$$

は,

$$(2.6) \quad -\frac{1}{2} ABIC = \log \sum_{\{x\}} \exp(-E_{\text{pos}}) - \log \sum_{\{x\}} \exp(-E_\pi) - \log Z_L$$

と書ける. ここで, $\sum_{\{x\}}$ はあらゆる可能な $\{x\}$ に関する和である.

E_π が変数 λ を含んでいるとすると, λ による ABIC の微分は,

$$(2.7) \quad -\frac{1}{2} \frac{\partial (ABIC)}{\partial \lambda} = \frac{\sum_{\{x\}} \left(-\frac{dE_\pi}{d\lambda} \right) \exp(-E_{\text{pos}})}{\sum_{\{x\}} \exp(-E_{\text{pos}})} - \frac{\sum_{\{x\}} \left(-\frac{dE_\pi}{d\lambda} \right) \exp(-E_\pi)}{\sum_{\{x\}} \exp(-E_\pi)}$$

のように $-\frac{dE_\pi}{d\lambda}$ の事後分布での期待値と事前分布での期待値の差の形になる. λ はもともと問題に含まれている巨視的パラメータでもよいし, 人為的に加えたもの (「温度」など) でもよい. E_L に含まれている変数による微分についても同様に期待値による表現が可能である. これらを利用すると, メトロポリスのモンテカルロ法を用いた simulation で事後分布での期待値及び事前分布での期待値を求め, 巨視的パラメータを修正して, 再び simulation を行うという操作を反復することにより, ABIC の最小化を行うことができる. 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か, 後者を off-line で計算して表にしておく必要がある.

筆者は、ABIC法について学んだ際に上記の方法で巨視的パラメータの推定が可能なることに気づいたが、同様の原理に基づく方法がGemanらによってすでに提案・実験されていることがわかった (Geman, D. and Geman, S. (1986), Geman and McClure (1987)). 統計数理研究所のOgataも独立に類似の表式に基づく方法を提案した。Ogataは繰返し法による最小化を行う代わりに λ についての1次元積分を計算してABICの値自体を計算している (Ogata (1990)). イジング型神経回路網におけるシナプス荷重の決定に同じようなアルゴリズムを用いた仕事として、Hinton and Sejnowski (1986)がある。発表の時期はこれが最も早い。Hintonらの問題意識とGemanやOgataの研究の問題意識には大きな違いがあるように思われる (注2)。Hintonらの用語では、(2.7) から導かれる微分方程式 (後出の (5.5), (5.6) 式に相当) をボルツマンマシンの学習方程式という。

3. 実験の内容

本論文では模擬データとして、真に2次元イジング模型に従うものを用いる。実際には、メトロポリスのモンテカルロ法で生成したマルコフ鎖から十分な間隔でサンプルしたパターン (snapshot) に雑音を加えたものをデータとする。以下、真の結合定数を J^{true} 、雑音の強さを g^{true} で示す。一般にこのような問題では境界条件が重要な要素となるが、ここでは簡単のため周期境界条件 (トーラス上の格子) で実験を行った。乱数は第6章ではM系列の疑似乱数 (伏見・手塚 (1981)) と統計数理研究所の物理乱数を併用した。第7章では物理乱数のみを使用した。

巨視的パラメータを知ること自体が目的の問題 (たとえば texture の特徴付けの問題) を念頭においた場合は、この方法は有効なテスト方法である。しかし、画像修復のための巨視的パラメータの推定という文脈を見た場合、画像修復実験に利用できる模擬データは、非常に狭い範囲の J^{true} でしか得られないという問題がある。模擬データを作るためのイジング模型の結合定数 J^{true} が臨界点より大きいと、パターンは (境界条件にもよるが) ほとんど白か黒にそろってしまう。また、 J^{true} が臨界点より小さくても、それに近い場合は、かなり大きな系で実験しないとやはり同じ問題が起こる。逆に J^{true} が臨界点よりある程度以上小さい場合には、画像修復は非常に困難になる。

本稿では、結合定数 J^{true} が中程度までの場合を中心に扱う (第6章) が、結合定数がより大きい (臨界点に近い) 場合に関する結果も述べる (第7章)。前者では巨視的パラメータの推定自体が主たる興味の対象であるが、後者では画像修復的な要素も含まれてくる。臨界点以下の場合については8.2節で簡単に考察する。

4. 問題の説明

正方格子上的各格子点 i 上の2値のパターン $\{x_i\}$ ($x_i = \pm 1$) を考える。“真の事前分布” が2次元イジング模型

$$(4.1) \quad \pi(\{x_i\}) = \frac{1}{Z_\pi} \exp\left(\frac{1}{2} J \sum_{j \in N(i), i} x_i x_j\right)$$

であることを知っているとする。ただし、 J は未知で、これをデータ $\{y_i\}$ から推定するのが以下の目的である。 $N(i)$ は i と直接の相互作用のある画素の集合で、今の場合は正方格子上的4つの隣接格子点とする。 Z_π は規格化定数 (分配関数)。

観測雑音の大きさを p とする。すなわち、各画素（微視的パラメータ）が独立に確率 p で符号反転したものがデータ $\{y_i\}$ として与えられるとする。このとき、

$$(4.2) \quad h = \frac{1}{2} \log \frac{(1-p)}{p}$$

とおくと、尤度関数は、

$$(4.3) \quad L(\{y_i\} | \{x_i\}) = \frac{\exp(h \sum_i y_i x_i)}{(2 \cosh(h))^M}$$

と書ける。 M は画素の総個数、 \sum_i は各格子点ごとの和である。

これらに基づいて ABIC に当たるものを書き下すと、

$$(4.4) \quad -\frac{1}{2} ABIC(J, h) = \log \sum_{\text{config.}} \exp(JA_\pi + hA_L) - \log \sum_{\text{config.}} \exp(JA_\pi) - M \log(2 \cosh(h))$$

となる。ここで、 $\sum_{\text{config.}}$ は $\{x_i\}$ のすべての組合せ (2^M 通り) に関する和である。また、

$$(4.5) \quad A_\pi = -E_\pi/J = \frac{1}{2} \sum_{j \in N(i), i} x_i x_j$$

$$(4.6) \quad A_L = -E_L/h = \sum_i y_i x_i$$

とおいた。統計物理の立場からは、(4.4) 式の第1項はデータ $\{y_i\}$ 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項は磁場なしのイジング模型の自由エネルギーと解釈できる。

5. メトロポリス的モンテカルロ法の適用

(4.4) 式を J と h でそれぞれ微分して、一般論の (2.7) 式に当たるものを書き下すと、

$$(5.1) \quad -\frac{1}{2} \frac{\partial(ABIC)}{\partial J} = \frac{\sum_{\text{config.}} A_\pi \exp(JA_\pi + hA_L)}{\sum_{\text{config.}} \exp(JA_\pi + hA_L)} - \frac{\sum_{\text{config.}} A_\pi \exp(JA_\pi)}{\sum_{\text{config.}} \exp(JA_\pi)}$$

$$(5.2) \quad -\frac{1}{2} \frac{\partial(ABIC)}{\partial h} = \frac{\sum_{\text{config.}} A_L \exp(JA_\pi + hA_L)}{\sum_{\text{config.}} \exp(JA_\pi + hA_L)} - M \tanh(h)$$

となる。

事前分布での期待値を $\langle \rangle_\pi$ 、事後分布での期待値を $\langle \rangle_{\text{pos}}$ のように書くことにすると、これらはそれぞれ、

$$(5.3) \quad -\frac{1}{2} \frac{\partial(ABIC)}{\partial J} = \langle A_\pi \rangle_{\text{pos}} - \langle A_\pi \rangle_\pi$$

$$(5.4) \quad -\frac{1}{2} \frac{\partial(ABIC)}{\partial h} = \langle A_L \rangle_{\text{pos}} - M \tanh(h)$$

と書ける。

したがって、ABIC 最小化のためには微分方程式

$$(5.5) \quad c_J \frac{dJ}{dt} = -\frac{1}{2} \frac{1}{M} \frac{\partial(ABIC)}{\partial J} = \frac{1}{M} (\langle A_\pi \rangle_{\text{pos}} - \langle A_\pi \rangle_\pi)$$

$$(5.6) \quad c_h \frac{dh}{dt} = -\frac{1}{2} \frac{1}{M} \frac{\partial(ABIC)}{\partial h} = \frac{1}{M} \langle A_L \rangle_{\text{pos}} - \tanh(h)$$

を解けばよいことになる。ここで t は仮想的な時間(メトロポリスのなモンテカルロ法における仮想的時間 (MCS) とは別物), c_J, c_h は時定数で以下では $c_J=1$ とする。右辺は画素の総個数 M で割った形にしてある。

この微分方程式を適当に差分化してオイラー法で解けばよいのだが、(5.4) の右辺を零とおいたものは h について

$$(5.7) \quad h = \operatorname{arctanh} \left(\frac{1}{M} \langle A_L \rangle_{\text{pos}} \right) = \frac{1}{2} \log \frac{1 + \frac{1}{M} \langle A_L \rangle_{\text{pos}}}{1 - \frac{1}{M} \langle A_L \rangle_{\text{pos}}}$$

のように解けるので、こちらに関しては毎回それを用いて h を更新することにした。はじめは(5.6)を解いていたが、時間の刻みを(5.5)と同じにしたためか、うまくいかなかった。

Geman らは(5.5)の右辺第2項に当たる部分を off-line で計算することにより、計算の効率化を図っているが、ここではその方法は採用しなかった。off-line で表を作っておくと、(5.5)も“解く”ことができ、いわゆる EM 法のようなアルゴリズムになる。しかし、この方法は巨視的パラメータが多くなると使えない。

これらの式は、事後分布での期待値及び事前分布での期待値を含んでいる。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メトロポリスのモンテカルロ法を長く走らせなければならないが、実際には有限で打ち切る必要がある。また、(5.5)を差分化する際の時間刻み dt も重要である。打ち切り回数が少なく、時間刻みが大きいと、偏りがなくなったあとでも、 $1/J$ と「時間」 t の曲線はメトロポリスのモンテカルロ法の誤差(「熱揺らぎ」)の影響をうけて凸凹する。この場合も $1/J$ の「時間平均」をとれば、ほぼ正しい結果が得られる可能性がある。逆に刻みが小さく、打ち切り回数も大きいときは、凸凹が少ない代わりに、偏りがなくなるのに多くの MCS を要する。しだいに打ち切り回数を増やす方法なども考えられるが、ここでは使わなかった。

打ち切り回数(第6章では50 MCS)ごとに $(1/J, h)$ を更新するが、これを1 iteration と定義する。各 iteration での $\{x_i\}$ の初期条件としては、前の iteration の最後の $\{x_i\}$ を用いた。事後分布を生成するメトロポリスのモンテカルロ法と事前分布を生成するメトロポリスのモンテカルロ法は完全に独立に行い、初期条件も事後分布の方は事後分布の方から、事前分布は事前分布からそれぞれ引き継ぐこととした。一番最初の $\{x_i\}$ としては、完全にランダムなパターン(各点が独立に確率 $1/2$ で 1 または -1 をとるパターン)を用いた。

以下の例では、 J の初期値として臨界点 ($1/J \sim 2.27$) より強結合となる値 ($1/J = 2.0$ など) をしばしば用いたが、最終的に得られた J は少数の例外を除いて弱結合側の値となった。一般に臨界点より J が大きい場合には、すべて 1 (またはすべて -1) のパターンからはじめた場合とランダムなパターンからはじめた場合で実質的に答が違う可能性があり、注意深い検討が必要である。いまの場合は、 J の真の値が臨界点よりある程度小さく、しかもそのことを事前に知っている場合に限定して考えているので、このような場合もランダムなパターンからはじめている。

計算は以下の順番で行った。メトロポリスのモンテカルロ法のアルゴリズムとしては狭義のメトロポリス法を用いた。

1. 最初の10 iteration ほど $(1/J, h)$ を初期値に固定したまま“空走り”させる。これは、

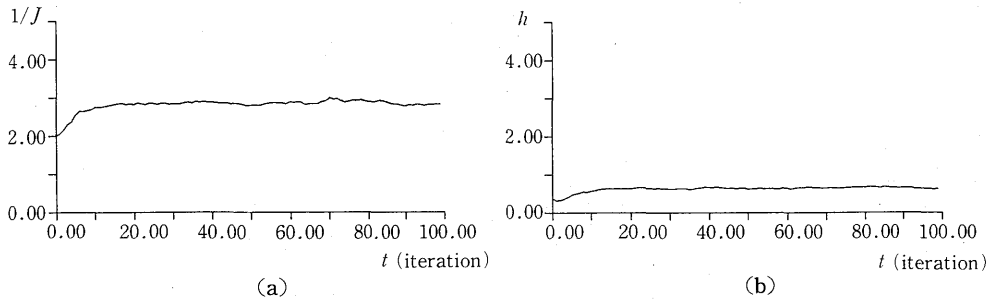


図1. (a) $(1/J, h)$ の初期値が $(2.0, 0.5)$ の場合. 横軸が「時間」 t , 縦軸が $1/J$. (b) $(1/J, h)$ の初期値が $(2.0, 0.5)$ の場合. 横軸が「時間」 t , 縦軸が h .

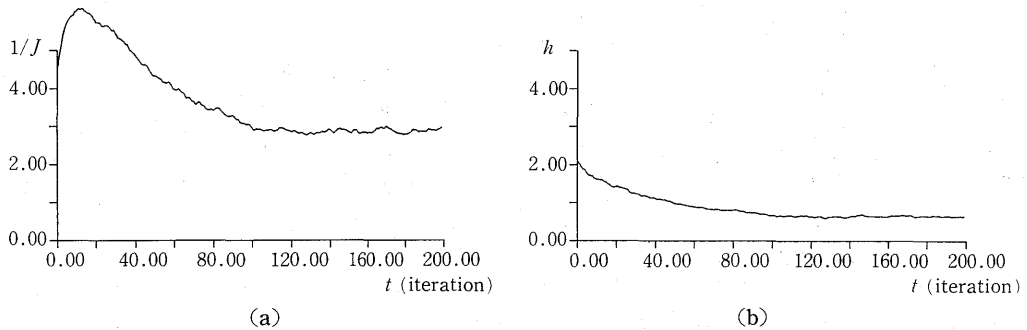


図2. (a) $(1/J, h)$ の初期値が $(4.0, 2.5)$ の場合. 横軸が「時間」 t , 縦軸が $1/J$. (b) $(1/J, h)$ の初期値が $(4.0, 2.5)$ の場合. 横軸が「時間」 t , 縦軸が h .

$\{x_i\}$ の初期条件が緩和するのを待つためである.

2. 次に方程式を解きはじめる. 200 から 400 iteration 程度で偏りがなくなったとみなす.
3. その後の 100 から 200 iteration で, 毎回方程式を解きつつ, $1/J$ 及び h の「時間平均」を計算する.

計算結果の例を図1(a-b), 図2(a-b)に示した. これらは大きさ 20×20 , 真の値が $(1/J^{\text{true}}, h^{\text{true}}) = (2.8, 0.2)$ ($(1/J^{\text{true}}, h^{\text{true}}) = (2.8, 0.693)$) の場合である. なお, 図1, 図2でははじめの 10 iteration とばすことは行っていない.

6. 実験結果(結合定数 J があまり強くない場合)

6.1 推定値の分布

実験に使用した模擬データの典型例(雑音を入れる前と入れた後)を図3(a-d), 図4(a-c)に示した. 大きさ 40×40 の場合, $1/J^{\text{true}} = 3.3$ では 200 MCS おきに, $1/J^{\text{true}} = 2.8$ では 500 MCS おきに snapshot を取り出し, それに観測雑音を加えて模擬データとした.

推定結果のヒストグラムを図5-図10に示した. 実験条件と図の対応は表1を参照されたい. 図7は10個, 図8, 図9は40個, それ以外は20個の模擬データに対する結果のヒストグラムである.

$(1/J, h)$ の初期値はいずれも $(1/J, h) = (2.0, 0.5)$ である. 時間刻み dt は 0.5 (大きさ 20×20 の場合は 0.1), 打ち切り回数は 50 MCS とした. 計算量は $(300 \sim 600 \text{ iteration}) \times 50 \text{ MCS/itera-}$

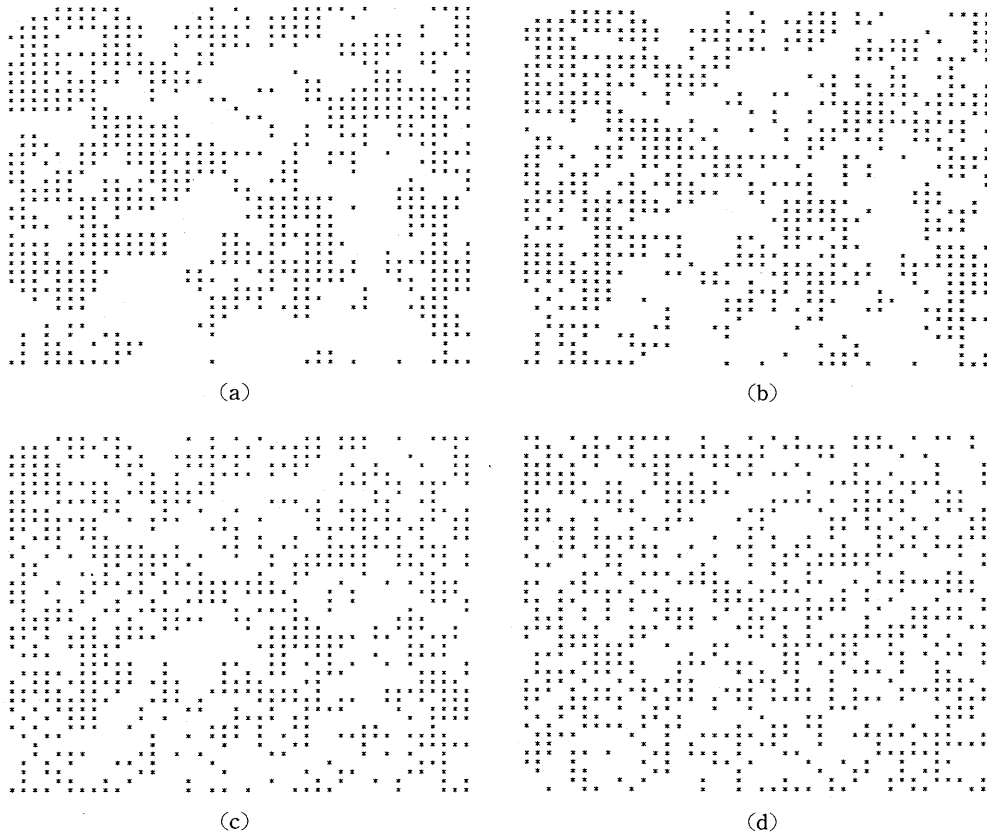


図3. $1/J^{\text{true}}=2.8$ の場合の snapshot と模擬データ. 大きさ 40×40 . (a) 観測雑音なし (正解). (b) 観測雑音 0.1. (c) 観測雑音 0.2. (d) 観測雑音 0.4.

tion=15000~30000 MCS 程度である. これを2つ並行して走らせるから全計算量は30000~60000 MCS となる. この計算には, 40×40 の場合, 統計数理研究所の M-682H で5分から10分の計算時間を要した. これは, 乱数の数の節約やベクトル化などを全く考慮しない素朴なプログラムを用いた場合なので, 工夫すれば改善が期待できる.

大きさが 40×40 で $1/J^{\text{true}}=2.8$, $p^{\text{true}}=0.1 \sim 0.2$ の場合には, 提案したアルゴリズムはよく機能しているといえる(図5, 図6). 同じ条件で $1/J^{\text{true}}=3.3$ とすると, かなり分散が大きくなるが, $1/J^{\text{true}}=2.8$ の場合との相違は明瞭である. また $p^{\text{true}}=0.1$ の場合と $p^{\text{true}}=0.2$ の場合の区別もできる(図8, 図9). $1/J^{\text{true}}=2.8$ の場合, $p^{\text{true}}=0.4$ になると推定はほとんど不可能である(図7, 左の図の横軸の目盛に注意). また $p^{\text{true}}=0.2$ でも, 大きさが 20×20 では分散が大きい(図10).

ヒストグラムは $1/J$ の大きい側に尾を引く傾向がある. これは, $1/J$ が小さい方にずれた場合の方がパターンが急激に変化することから見て, 直観的に納得できる. 模擬データに関する平均値も, $1/J$ と h がともに大きい側(弱結合・弱雑音の側)に偏るように見える.

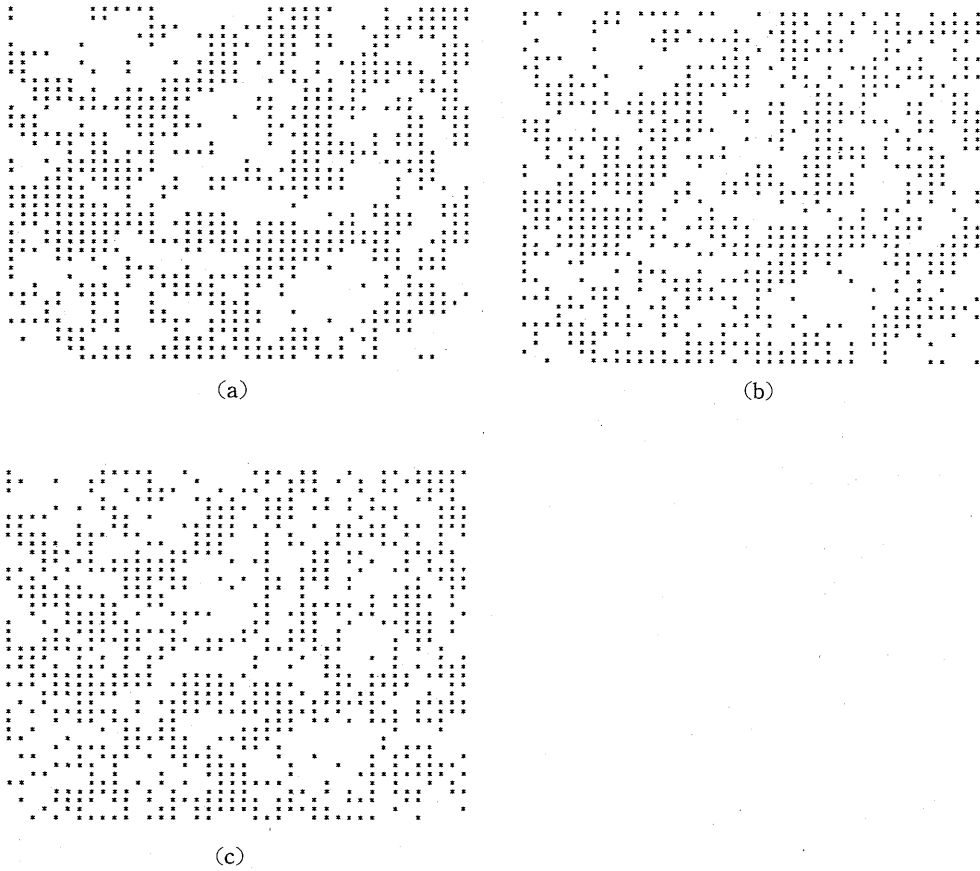


図4. $1/J^{\text{true}}=3.3$ の場合の snapshot と模擬データ, 大きさ 40×40 . (a) 観測雑音なし (正解), (b) 観測雑音 0.1, (c) 観測雑音 0.2.

表1.

大きさ	$1/J^{\text{true}}$	d^{true}	ヒストグラム (推定結果)	模擬データの例 (雑音混入前)	模擬データの例 (雑音混入後)
40×40	2.8	0.1	図5	図3(a)	図3(b)
40×40	2.8	0.2	図6	図3(a)	図3(c)
40×40	2.8	0.4	図7	図3(a)	図3(d)
40×40	3.3	0.1	図8	図4(a)	図4(b)
40×40	3.3	0.2	図9	図4(a)	図4(c)
20×20	2.8	0.2	図10		

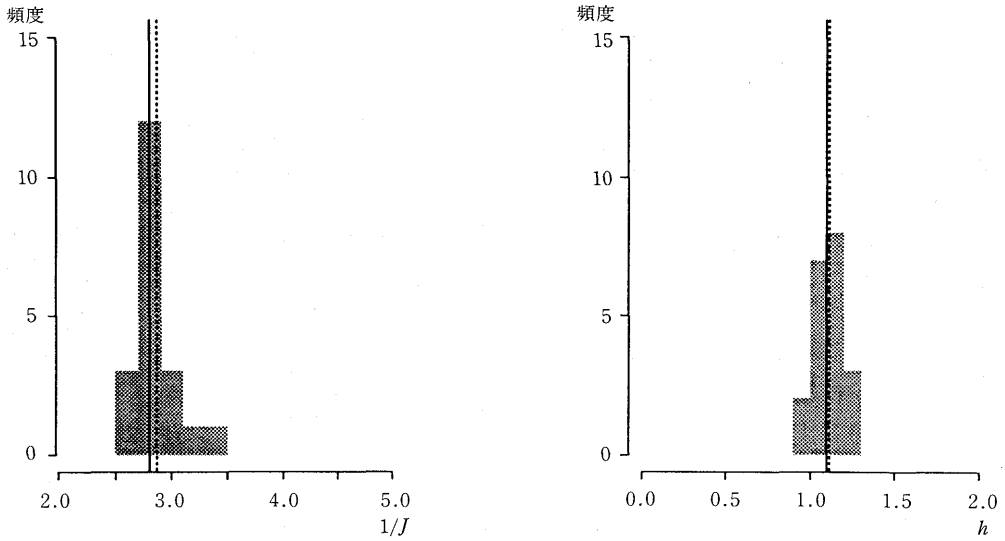


図5. 大きさ 40×40 , $1/J=2.8$, $p=0.1$. 実験に使用した模擬データの数は20. 左が $1/J$, 右が h の推定値のヒストグラム. 縦の点線が推定値の模擬データに関する平均, 実線が真の値.

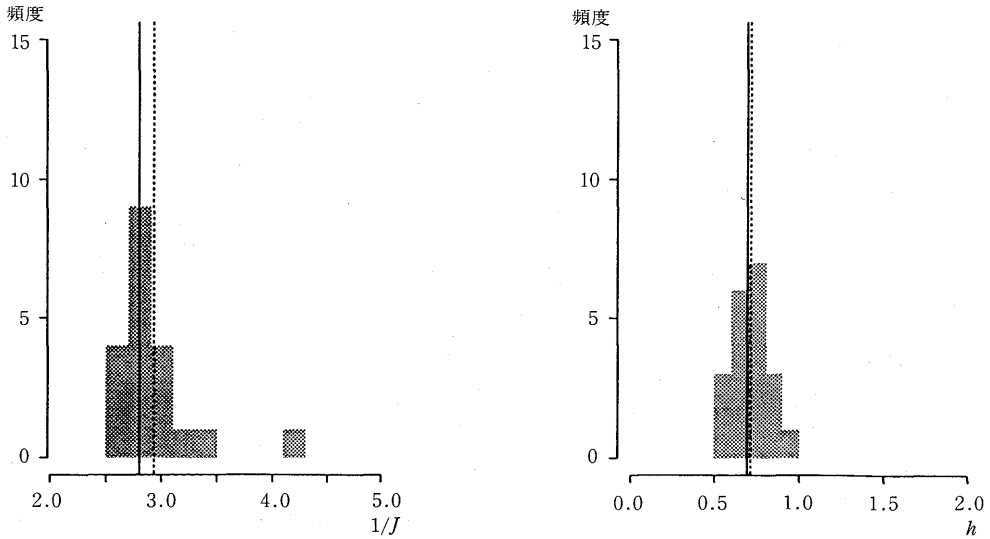


図6. 大きさ 40×40 , $1/J=2.8$, $p=0.2$. 実験に使用した模擬データの数は20. あとは図5の説明参照.

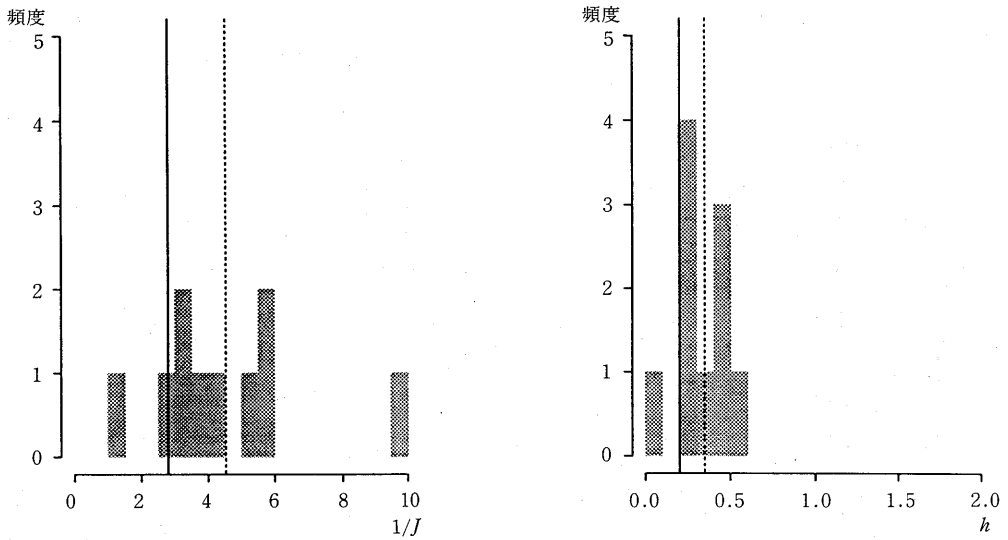


図7. 大きさ 40×40 , $1/J=2.8$, $p=0.4$. 実験に使用した模擬データの数は10. あとは図5の説明参照.

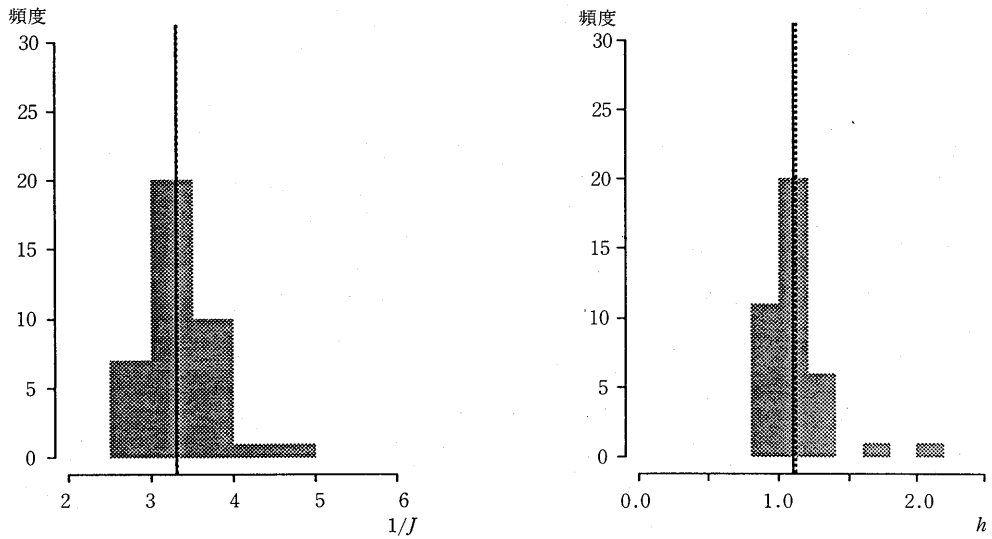


図8. 大きさ 40×40 , $1/J=3.3$, $p=0.1$. 実験に使用した模擬データの数は40. h が大きくなりすぎて計算を中止したものが40個中1個あり,これはヒストグラム及び平均値に含まれていない. あとは図5の説明参照.

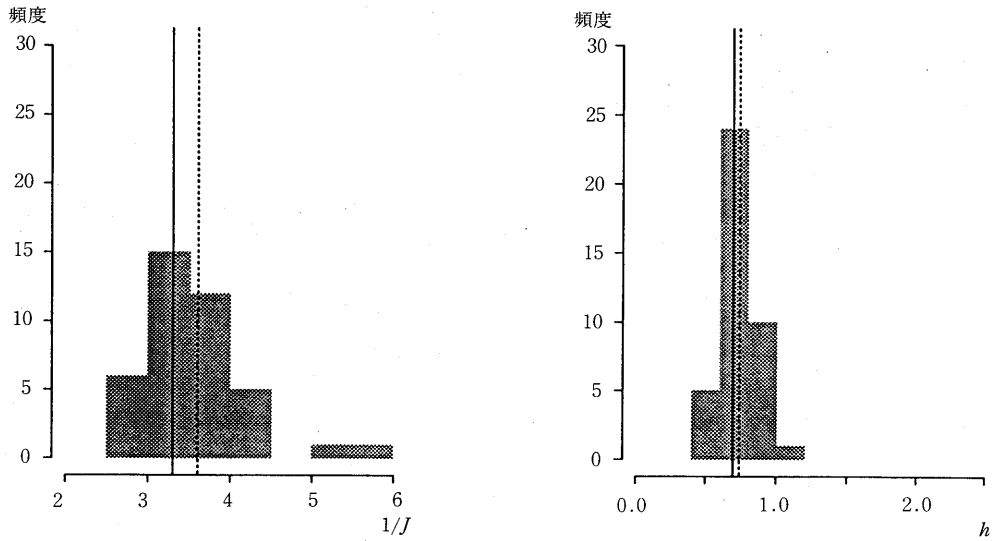


図9. 大きさ 40×40 , $1/J=3.3$, $p=0.2$. 実験に使用した模擬データの数は40. あとは図5の説明参照.

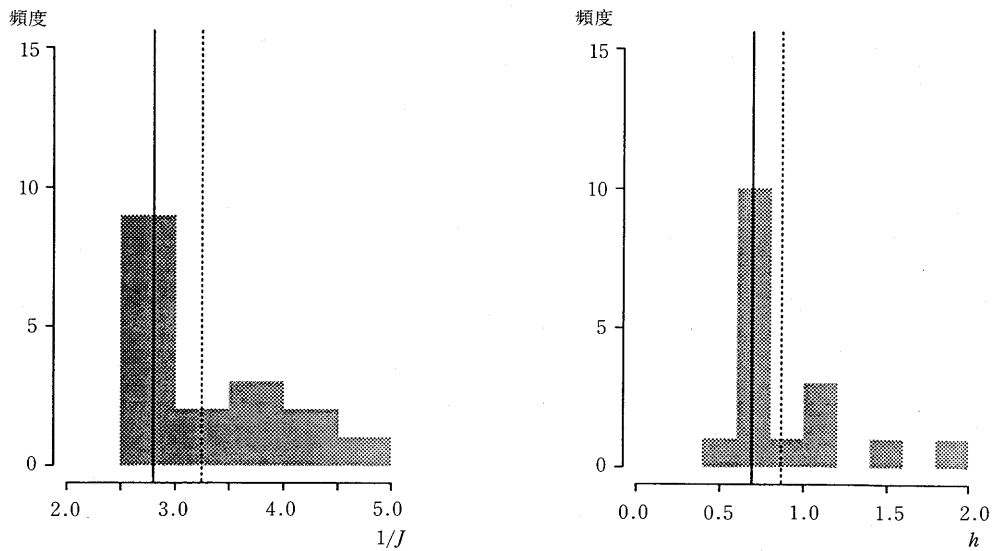


図10. 大きさ 20×20 , $1/J=2.8$, $p=0.2$. 実験に使用した模擬データの数は20. この図では h が大きくなりすぎて計算を中止したものが20個中3個あり, これらはヒストグラム及び平均値に含まれていない. あとは図5の説明参照.

6.2 収束と初期値依存性のチェック

狭義の最尤法による完全データからの推定(教師ありの学習)では、イジング模型を極限として含む指数型分布一般について巨視的パラメータに関する convexity が示せる(たとえば Smolensky (1986)). これに対して、ABIC を用いる不完全データからの推定(教師なしの学習)では、巨視的パラメータの空間で局所的な極値が出現する可能性がある。ここでは、その点を含めて、収束と初期値依存性について調べる。

図6と同じ条件の模擬データについて、初期値 $(1/J, h) = (2.0, 0.5)$ から出発した場合と $(4.0, 2.5)$ から出発した場合の $1/J$ の推定値を比較したものが図11である。結果はほぼ一致している。

図9と同じ条件の模擬データについて同じ実験をした結果が図12である。推定結果の $1/J$ の大きいものについては収束が悪くなる。この原因の1つは、 $(1/J, h) = (4.0, 2.5)$ から出発すると、一度 $1/J$ が大きくなってからまた小さくなり収束する、という挙動(図2(a)のような挙動)が見られることである。 $h=2.5$ という値は非常に大きいため、初期状態で(5.7)式の $(1/M) \langle A_L \rangle_{\text{pos}}$ がほとんど1になってしまい、そこに統計的揺らぎが加わるため、 h はなかなか減少しない。 $1/J$ の方は、もともと h を初期値2.5に固定した場合は増加方向に動くので、(この領域にしては) dt の大きさが大きすぎるためもあって、急激に増加してしまうわけである。このような場合でも、ある程度 h が小さくなれば、あとは急速に収束するようである。図12で特に大きく離れている1つについては、 $1/J$ の値が小さくなりはじめたところで計算時間が終わってしまっている。図11の場合も、推定結果の $1/J$ が最も大きい模擬データに関しては、使

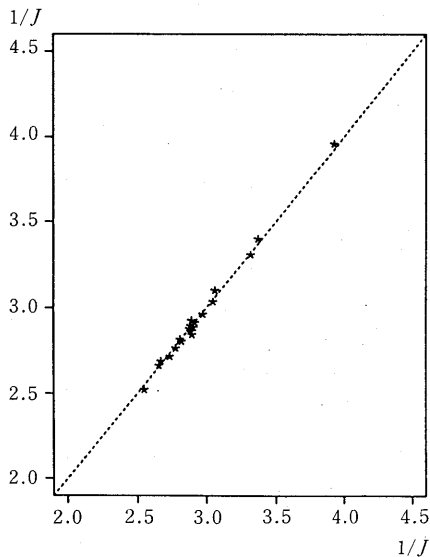


図11. $1/J^{\text{true}}=2.8$, $p^{\text{true}}=0.2$ の場合の初期値依存性。横軸は初期値 $(1/J, h) = (2.0, 0.5)$ から出発して、200 iteration (200 MCS ではない)後の100 iterationの平均。縦軸は初期値 $(1/J, h) = (4.0, 2.5)$ から出発して、400 iteration後の200 iterationの平均。20個の模擬データに関する比較。点線は横軸=縦軸の直線を表わす。

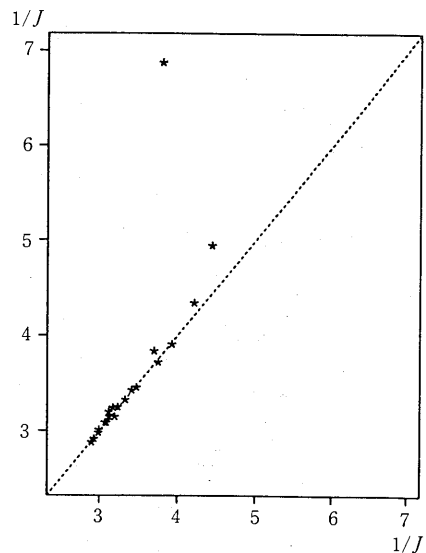


図12. $1/J^{\text{true}}=3.3$, $p^{\text{true}}=0.2$ の場合の初期値依存性。横軸は初期値 $(1/J, h) = (2.0, 0.5)$ から出発して、200 iteration後の200 iterationの平均。縦軸は初期値 $(1/J, h) = (4.0, 2.5)$ から出発して、400 iteration後の200 iterationの平均。20個の模擬データに関する比較。点線は横軸=縦軸の直線を表わす。

用する乱数列によって、ここで論じたような不安定性が見られる。

図7の場合（非常に推定結果の悪い場合）について、独立な乱数による2つの結果を比較したものが図13である(ともに $(1/J, h) = (2.0, 0.5)$ から出発した場合)。+印の点では一方の結果で $1/J$ が零に近付きすぎたため途中で計算を打ち切った。このほかに一方の結果で $1/J$ が大きくなりすぎ、図に入らなかった点が1つある。図13を見ると、大部分の点に関して推定値の大小の傾向は合うものの、良い精度で一致するとはいい難い。この不一致が局所的極値によるものかどうかは、もっと計算時間をかけないとわからないが、単に収束の速さなどの問題である可能性も大きい。なお、真の値 $(1/J, h) = (2.8, 0.2027)$ から出発した場合と $(1/J, h) = (2.0, 0.5)$ から出発した場合の比較でも図13と似た結果になった。

以上の結果から、非常に条件の悪い場合を除いて、ある程度の範囲の初期値に対して同一の推定結果が得られることが期待できる。収束の速さは必ずしも速くないが、図11や図12を見る限り、局所的極値が多数存在するようには見えない。

真の値が不明である場合にどのような初期値から出発するのが収束が速かつ安定か、というのは別に検討すべき問題である。 $(1/J, h) = (4.0, 2.5)$ では h が明らかに大きすぎる。 $(1/J, h) = (2.0, 0.5)$ では、ランダムパターンから出発したこともあって、比較的速く収束している。しかし、 $1/J$ の真の値が臨界点より弱結合側であることがわかっているなら、わざわざ初期値を強結合側にとる必然性はないと思われる。

6.3 打ち切り回数の影響

前に述べたように、本来 t を $t+dt$ に動かすたびに期待値が収束するまでメトロポリスのモンテカルロ法を長く走らせなければならないが、実際には有限で打ち切っている。本節ではこの打ち切り回数を50 MCSとしているが、これが少ないことが結果に影響しているのではないかという疑問がレフェリーにより提出された。50 MCSは緩和時間と同程度か、たかだかその数倍と思われるので、「時間平均」を推定値とすることを考慮しても、これはもっともな疑問である。

これを調べるために、打ち切り回数が50 MCSの場合と100 MCSの場合を比較したのが図14、図15である。それぞれ図6、図9と同じ条件の模擬データ10個ずつについての比較であって、 $(1/J, h)$ の初期値は $(1/J, h) = (2.0, 0.5)$ である。図14、図15によると、大部分の模擬データについて、両者の差は小さい。しかし、100 MCSでも揺らぎがかなりあり、打ち切り回数をさらに増やした場合との比較が必要かもしれない。打ち切り時間と時間刻み dt を適応的に変える方法の開発とともに、この点は今後の課題としたい。微分方程式の差分化にこだわらず、2

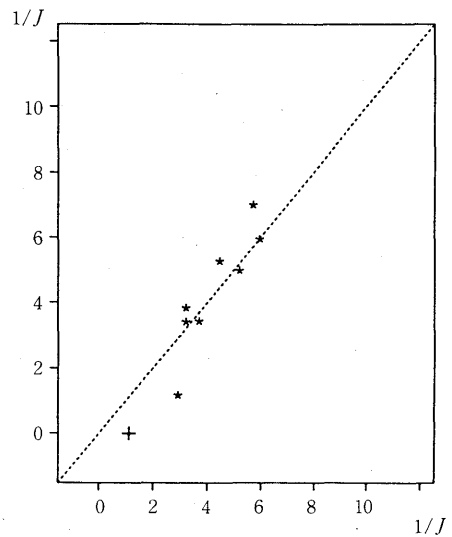


図13. $1/J^{\text{true}} = 2.8, p^{\text{true}} = 0.4$ の場合の収束性・乱数依存性。横軸は初期値 $(1/J, h) = (2.0, 0.5)$ から出発して、400 iteration後の200 iterationの平均。縦軸は同じ初期値から出発して、別の乱数列を使用し、550 iteration後の250 iterationの平均。10個の模擬データに関する比較(本文参照)。点線は横軸=縦軸の直線を表わす。

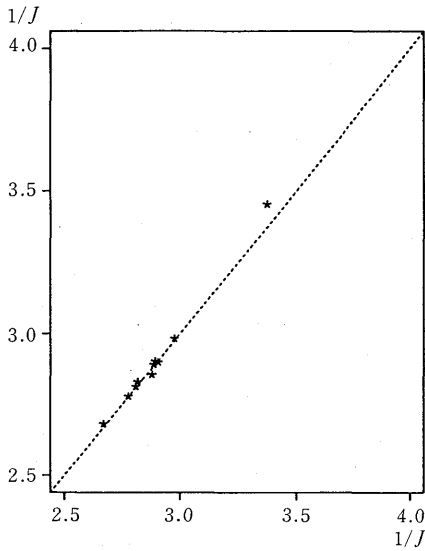


図 14. $1/J^{\text{true}}=2.8, p^{\text{true}}=0.2$ の場合の打ち切り回数依存性. 横軸は打ち切り回数 50 MCS, 縦軸は 100 MCS. ともに初期値 $(1/J, h)=(2.0, 0.5)$ から出発して, 200 iteration 後の 100 iteration の平均, 10 個の模擬データに関する比較. 点線は横軸=縦軸の直線を表わす.

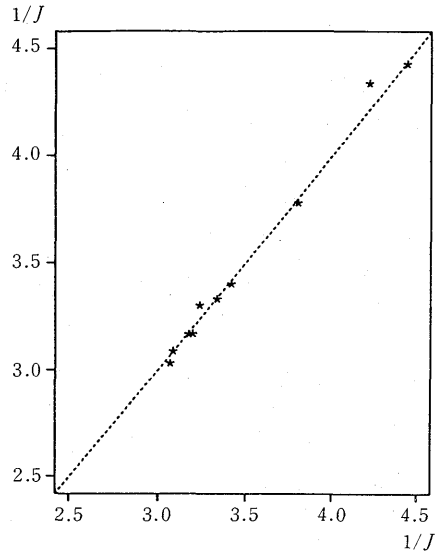


図 15. $1/J^{\text{true}}=3.3, p^{\text{true}}=0.2$ の場合の打ち切り回数依存性. 横軸は打ち切り回数 50 MCS, 縦軸は 100 MCS. ともに初期値 $(1/J, h)=(2.0, 0.5)$ から出発して, 200 iteration 後の 200 iteration の平均, 10 個の模擬データに関する比較. 点線は横軸=縦軸の直線を表わす.

分割の繰返しによる 1 次元探索を取り入れることも考えられる.

6.4 画像復元問題との関連

推定した J, h を利用して画素 $\{x_i\}$ を推定した場合, どの程度復元されるかも興味のあるところである. 巨視的パラメータとして, データを作成するのに用いた $J^{\text{true}}, h^{\text{true}}$ を与えた場合と本稿の方法で推定した J, h を用いた場合の画像復元効果の比較を表 2 に示した. 系の大きさはすべて 40×40 で, 実験結果は模擬データ 20 組について行ったものの平均である ($p^{\text{true}}=0.4$ の場合は 10 組). 復元はベイズ的な意味で重なりを最大にする解 (MPM 解(注 3)) に基づいて行った.

本節で考えている状況は画像復元に好都合ではないが, $1/J^{\text{true}}=2.8, p^{\text{true}}=0.2$ の場合には, (J, h) の真の値が既知の場合に近い結果が得られている. 処理に利用する巨視的パラメータによっては処理前より悪い結果になる可能性もあることに注意されたい. この方向でのより興味ある結果は第 7 章で述べる.

表 2.

$1/J^{\text{true}}$	p^{true}	正解との重なり (J, p) 既知	正解との重なり (J, p) 推定	正解との重なり 処理前
2.8	0.1	0.906	0.905	0.898
2.8	0.2	0.828	0.825	0.799
2.8	0.4	0.630	0.612	0.601

7. 実験結果（より臨界点に近い場合）

さらに臨界点の近くでどうなるか調べるために、大きさ 60×60 、結合定数 $1/J^{\text{true}}=2.45$ 、観測雑音 p^{true} が 0.25 及び 0.35 の場合を調べた（臨界点は $1/J \sim 2.27$ である）。模擬データは、10000 MCS おきに取り出した snapshot に観測雑音を加えて作成した。

初期条件は $(1/J, h) = (3.5, 1.5)$ とした。緩和時間が延びることを考慮して、各 iteration での打ち切り回数は、事後分布、事前分布について、それぞれ 100 MCS 及び 200 MCS とした。時間刻み dt は 0.5 で第 6 章と同じである。推定値の計算のためには、 $p^{\text{true}}=0.25$ の場合は 300 iteration 後の 100 iteration の、 $p^{\text{true}}=0.35$ の場合は 400 iteration 後の 100 iteration の、それぞれ平均をとった。模擬データ 1 つ当たりの全計算量は 120000~150000 MCS、統計数理研究所の M-682H で 40~50 分であった。

結果は表 3、表 4 に示した通りである。 $1/J, h$ の大きい側への偏りがあるように見えるが、かなり良い推定値が得られている。表 3、表 4 では、対応するデータ番号の結果は独立でない（同じ snapshot から作成した模擬データが使用されている）。

表 4 の場合、計算はほぼ収束していると思われるが、（これ以外の模擬データに対する結果で）400 iteration では収束しないものが少なくとも 1 つあった。この場合も 900 iteration では良い値に収束した。6.2 節と似た事情があり、初期条件 $(1/J, h) = (3.5, 1.5)$ からはじめると、いったん $1/J$ が大きい方に動き、そこから戻るのに時間がかかるのが収束を悪くしている。また、打ち切り回数ももっと増やすべきである。ここでは、事前分布側の打ち切り回数を事後分布側の 2 倍にとっているが、むしろ事後分布側を多くした方がよいかもかもしれない。

これらの場合、特に表 4 の場合には、画像修復能力がかなり認められる。画像修復能力は

表 3. $1/J^{\text{true}}=2.45, p^{\text{true}}=0.25$.

サンプル No.	$1/J$ の推定値 ($1/J^{\text{true}}=2.45$)	h の推定値 ($h^{\text{true}}=0.549$)	正解との重なり (J, p) 既知	正解との重なり (J, p) 推定	正解との重なり 処理前
No.1	2.48	0.551	0.831	0.833	0.750
No.2	2.60	0.553	0.813	0.813	0.744
No.3	2.43	0.544	0.829	0.829	0.749
No.4	2.49	0.610	0.829	0.827	0.757
平均	2.50	0.565	0.826	0.826	0.750

表 4. $1/J^{\text{true}}=2.45, p^{\text{true}}=0.35$.

サンプル No.	$1/J$ の推定値 ($1/J^{\text{true}}=2.45$)	h の推定値 ($h^{\text{true}}=0.310$)	正解との重なり (J, p) 既知	正解との重なり (J, p) 推定	正解との重なり 処理前
No.1	2.55	0.374	0.753	0.753	0.661
No.2	2.67	0.383	0.736	0.738	0.651
No.3	2.51	0.325	0.767	0.768	0.653
No.4	2.45	0.327	0.747	0.749	0.657
No.5	2.52	0.377	0.767	0.766	0.654
No.6	2.44	0.306	0.769	0.769	0.649
平均	2.52	0.349	0.757	0.757	0.654

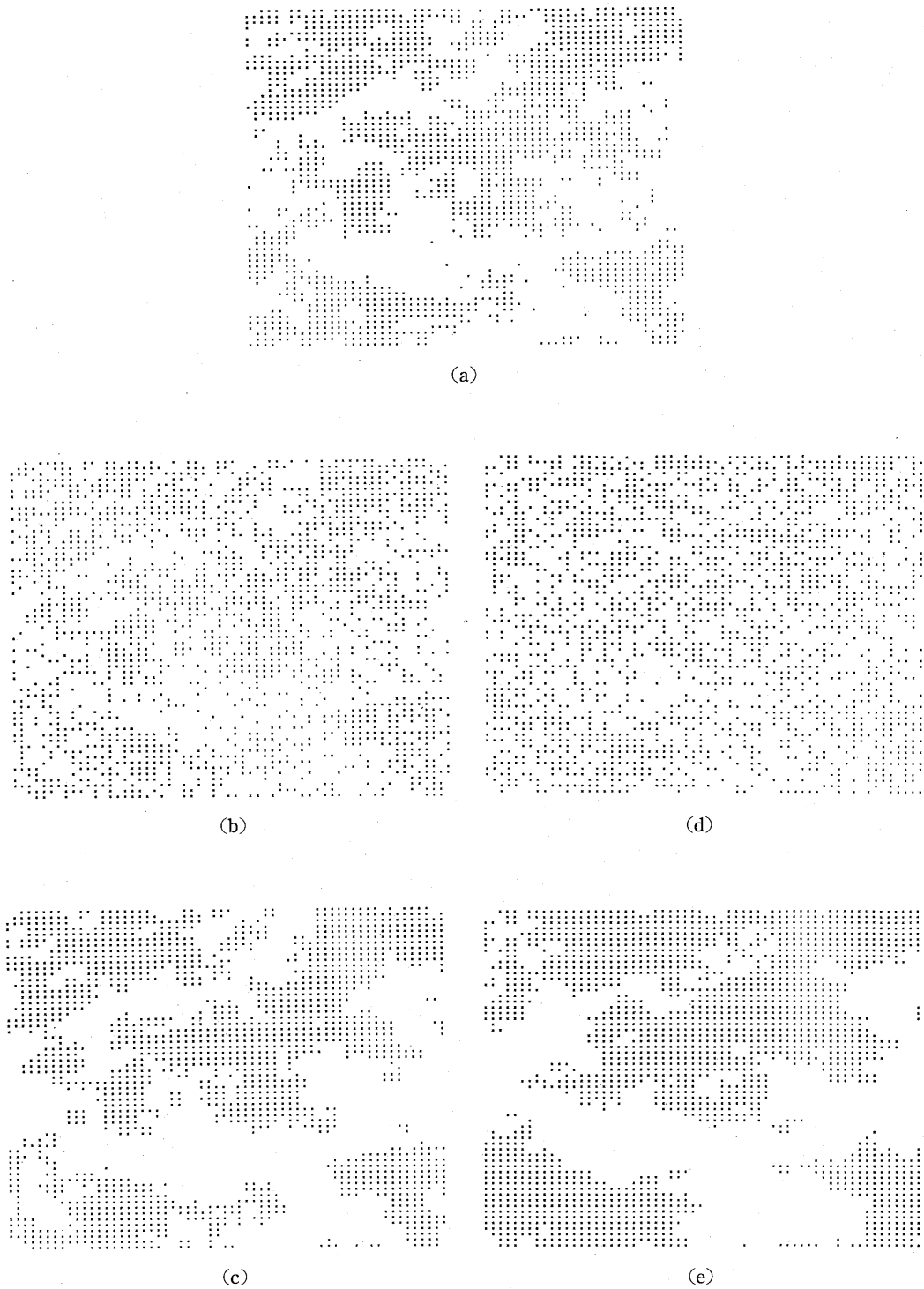


図 16. 大きさ 60×60 , $1/J^{\text{true}} = 2.45$ の場合の snapshot, 模擬データ, 処理結果. 画像修復は大域的な MAP 解でなく, MPM 解 (注 3) による. (a) 観測雑音なし (正解). (b) 観測雑音 0.25. (c) (b) を処理した結果. (d) 観測雑音 0.35. (e) (d) を処理した結果.

J, h の真の値を与えた場合と推定した場合とでほとんど変わらない。ただし、この領域での画像修復能力は用いる J, h についてある程度頑健と思われる。模擬データ（もとの画像及び雑音を加えたもの）と修復結果の例を図 16(a-e) に示した。

さらに臨界点に近い領域での研究を行うには、非常に多くの計算時間が必要である。そのような領域での推定は、フラクタルパターンに関する統計的推論の例としても興味深い。

8. 関連事項

8.1 他の方法との比較

離散的なモデルや非ガウスのモデルに関して ABIC 法による巨視的パラメータ推定を行う方法としては、メトロポリスのモンテカルロ法以外に、系を小さな部分に分けて近似的な ABIC を求める方法、系を帯状の部分に分割してそれぞれに（非ガウス）フィルターを適用する方法などが考えられる。これらの方法の共通の問題点は、微視的パラメータの状態数が増えたり相互作用の範囲が増えた場合に、モデルに忠実に計算を行おうとすると、急激に計算量が増えることである。メトロポリスのモンテカルロ法はそうした状況に強いと考えられる。フィルターによって離散画像モデルを扱った例としては、Derin et al. (1984), Devijver and Dekesel (1987) などがあるが、前者は観測雑音を含むデータからの巨視的パラメータ推定を行っておらず、後者は非常に特殊なマルコフ場に限定して扱っている。

観測雑音がない場合に巨視的パラメータを推定する方法に pseudo-likelihood 法 (Besag (1975, 1986), Geman and Graffigne (1986)) がある。Besag et al. (1991) では、pseudo-likelihood 法の考え方をメトロポリスのモンテカルロ法と組合せて、観測雑音がある場合に利用することが論じられている。これは、ABIC 法にこだわらなければ、興味深い方法である。この際、微視的パラメータの推定値(復元結果)に pseudo-likelihood 法を適用するのではなく、事後分布からサンプル (snapshot) を多数抽出して適用すべきであろう。筆者の誤解かもしれないが、Besag らは前者を行っているのではないと思われる点がある。

8.2 秩序相での推定について

イジング模型を事前分布として利用した際に、画像修復能力がより顕著になるのは、図 16 よりさらにはっきりとした塊からなるような画像に対してである。この場合、最適の J の値は臨界点よりかなり強結合側になる可能性がある。このような場合にいまの方法で J を推定するのがよいかどうかは、緩和の問題等を別にしても疑問の余地がある。臨界点よりある程度強結合側の場合、事前分布から抽出した典型的パターンは白または黒の地に小さな島状に反対の色(符号)のものが点在しているようなものである。これに対して、事後分布から抽出した典型的パターンは、画像がはっきりとした塊からなる場合、大きく白の領域と黒の領域で塗り分けられたものになる。このように事前分布と事後分布の性質が全く違う場合には、ABIC 法でなく他の方法(たとえば cross validation (交差確認法)) で J を推定した方が、実態に即しているのではないだろうか。

9. まとめ

2次元イジング模型の無秩序相(常磁性相)の問題では、ある程度条件が良ければ、ABIC 法とメトロポリスのモンテカルロ法を組合せることにより、結合定数と観測雑音の同時推定ができることが確かめられた。巨視的パラメータの空間で局所的極値にトラップされることは、この

例ではあまり問題にならない。結合定数が臨界点にある程度近い場合についても、この方法は有効と思われる。

今後の課題は、各 iteration での打ち切り回数及び時間刻みを適応的に制御する、1次元探索を効率的に行うなどの方法で、最終的な打ち切り回数を増やし、より正確に ABIC を最小化する方法を開発することである。

10. 注

注 1. Geman らは統計数理研究所の研究者の多くがパラメータと呼ぶものを画素、ラベルなどと呼び、ハイパーパラメータと呼ぶものをパラメータとしている。また、非線形フィルターの用語では、パラメータを状態ベクトルと呼び、ハイパーパラメータに当たるものをパラメータと呼ぶ。さらに、観測雑音を含まない場合には、ハイパーパラメータに当たるものが単なるパラメータになる。

混乱の原因としては、“ハイパー”という用語が階層性を示すのみでパラメータの特性を表わすものでないこと、接頭辞なしの“パラメータ”がいろいろの意味に使われていることが挙げられる。

本稿では、試みに“微視的パラメータ”、“巨視的パラメータ”という名称を使用した。個々の画素を“微視的パラメータ”、後出の J, c のようなパラメータ（ハイパーパラメータ）を“巨視的パラメータ”と呼ぶわけである。両者の本質的違いは、画像の大きさが増大しても前者の精度は増さないのに対し、後者は一般に精度が高くなっていくという点にある。統計物理においても、個々のスピンの向きや分子の位置のように系の大きさを大きくしても揺らぎが小さくならない量（微視的な変数）と全エネルギーや圧力のように揺らぎが小さくなる量（巨視的な変数）がある。“微視的”、“巨視的”という名称はこれとのアナロジーのつもりである。

他の代案としては、“局所的パラメータ”、“大域的パラメータ”というのがある（田辺（私信））。

注 2. 画像処理へのベイズ的アプローチは、巨視的パラメータの導入により overfit の問題（不適切性の問題）を解決することがもとの動機であった。これに対して、Hinton らの（狭義の）ボルツマンマシンでは、推定すべきパラメータ自体が沢山ある場合を主として考えており、overfit の問題は重視されていない。また、Hinton らの流れをくむ研究では、ボルツマンマシンは、問題に応じて統計モデルを考えるという立場からではなく、多機能ネットワークとして研究される場合が多い。

注 3. 2値画像の推定の場合、正しいパターンとの重なるの（ベイズ的な）期待値を最大にするには、各画素ごとに（1体の）周辺事後確率を考えて、“1である確率が1/2より大きければ1、小さければ-1”という基準で、パターンを決めればよい。Marroquin はこれを MPM 解（Maximum Posterior Marginal estimate）と呼んでいる（Marroquin (1985), Marroquin et al. (1987)）。

これに対して、パターン全体が完全に一致する確率を最大にする解が MAP 解（Maximum A Posteriori estimate）である。MPM 解のことを“ラベルごとの MAP 解”と呼ぶような用語法も可能なので、“大域的な MAP 解”と呼ぶのが紛れがないだろう。

場合によっては、大域的な MAP 解を用いるとパターンの復元能力が落ちることがある。MPM 解と大域的な MAP 解で差が生じるのはモデルの非ガウス性のためであるが、詳しい考

察は別の機会に譲る。

謝 辞

統計数理研究所の田辺國士教授と尾形良彦助教授, ATR 視聴覚研究所の川人光男氏, 岡本俊明氏, 本郷節之氏, 乾俊郎氏, 東京大学工学部の甘利俊一教授には有益な助言と励ましを頂きました。これらの方々に感謝します。

参 考 文 献

- Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems, *J. Roy. Statist. Soc. Ser. B*, **36**, 192-236.
- Besag, J. (1975). Statistical analysis of non-lattice data, *The Statistician*, **24**, 179-195.
- Besag, J. (1986). On the statistical analysis of dirty pictures, *J. Roy. Statist. Soc. Ser. B*, **48**, 259-302.
- Besag, J., York, J. and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics, *Ann. Inst. Statist. Math.*, **43**, 1-59.
- Binder, K. (1986). Introduction: theory and "technical" aspects of Monte Carlo simulations, *Monte Carlo Methods in Statistical Physics* (ed. K. Binder), Topics in current physics, Vol. 7, 1-45, Springer, Berlin.
- Cross, G.R. and Jain, A.K. (1983). Markov random field texture models, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **5**, 25-39.
- Derin, H. and Elliott, H. (1987). Modeling and segmentation of noisy and textured images using Gibbs random fields, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **9**, 39-55.
- Derin, H., Elliott, H., Cristi, R. and Geman, D. (1984). Bayes smoothing algorithms for segmentation of binary images modeled by Markov random fields, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **6**, 707-720.
- Devijver, P.A. and Dekesel, M.M. (1987). Learning the parameters of a hidden Markov random field image model: a simple example, *Pattern Recognition Theory and Applications* (eds. P.A. Devijver and J. Kittler), NATO ASI series F30, Springer, Berlin.
- 伏見正則, 手塚 集 (1981). 多次元分布が一様な疑似乱数列の生成法, *応用統計学*, **10**, 151-163.
- Geman, D. and Geman, S. (1986). Bayesian image analysis, *Disordered Systems and Biological Organization* (eds. E. Bienenstock et al.), NATO ASI series F20, Springer, Berlin.
- Geman, D., Geman, S. and Graffigne, C. (1987). Locating texture and object boundaries, *Pattern Recognition Theory and Applications* (eds. P.A. Devijver and J. Kittler), NATO ASI series F30, Springer, Berlin.
- Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **6**, 721-741.
- Geman, S. and Graffigne, C. (1986). Markov random field image models and their applications to computer vision, *Proceedings of International Congress of Mathematicians*, Berkeley, California (ed. A.M. Gleason), AMS, Providence, Rhode Island.
- Geman, S. and McClure, D.E. (1987). Statistical methods for tomographic image reconstruction, *Bull. Internat. Statist. Inst.*, **52**, Book 4, 5-21.
- Hinton, G.E and Sejnowski, T.J. (1986). Learning in Boltzmann machines, *Parallel Distributed Processing*, Vol. 1 (eds. E. Rumelhart and J.L. McClelland), MIT Press, Cambridge, Massachusetts.
- 本郷節之, 川人光男, 乾 俊郎, 三宅 誠 (1991). エネルギー学習機能をもつ局所並列確率アルゴリズムによる輪郭線抽出, *電子情報通信学会論文誌 D-II*, **J74**, 348-356.
- Marroquin, J. (1985). Optimal Bayesian estimators for image segmentation and surface reconstruction, MIT A.I. Memo 839, MIT, Cambridge, Massachusetts.
- Marroquin, J., Mitter, S. and Poggio, T. (1987). Probabilistic solution of ill-posed problems in computational vision, *J. Amer. Statist. Assoc.*, **82**, 76-89.

- Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953). Equation of state calculation by fast computing machines, *J. Chem. Phys.*, **21**, 1087-1092.
- Miyashita, S. and Takano, H. (1985). Dynamical nature of the phase transition of the two-dimensional kinetic Ising model, *Progr. Theoret. Phys.*, **73**, 1122-1140.
- Ogata, Y. (1990). A Monte Carlo method for objective Bayesian procedure, *Ann. Inst. Statist. Math.*, **42**, 403-433.
- Ogata, Y. and Tanemura, M. (1981). Estimation of interaction potentials of spatial point patterns through the maximum likelihood procedure, *Ann. Inst. Statist. Math.*, **33**, 315-338.
- Ogata, Y. and Tanemura, M. (1984). Likelihood analysis of spatial point patterns, *J. Roy. Statist. Soc. Ser. B*, **46**, 496-518.
- Ogata, Y. and Tanemura, M. (1985). Estimation of interaction potentials of marked spatial point patterns through maximum likelihood method, *Biometrics*, **41**, 421-433.
- Ogata, Y. and Tanemura, M. (1989). Likelihood estimation of soft-core interaction potentials for Gibbsian point patterns, *Ann. Inst. Statist. Math.*, **41**, 583-600.
- Ohtsuki, H. and Kawato, M. (1991). Training a hierarchical image model: estimation of MRF line-process energy provided with image intensity data without a high-level teacher, *Proceedings of the International Joint Conference on Neural Networks, Seattle*.
- Penttinen, A. (1984). Modelling interactions in spatial point patterns: parameter estimation by the maximum likelihood method, Jyväskylä studies in computer science, economics and statistics, Vol. 7, Ph.D Thesis, University of Jyväskylä, Finland.
- Smolensky, P. (1986). Harmony theory, *Parallel Distributed Processing*, Vol. 1 (eds. E. Rumelhart and J.L. McClelland), MIT Press, Cambridge, Massachusetts.

Macroscopic Parameter Estimation from Incomplete Data with Metropolis-type Monte Carlo Algorithm

Yukito Iba

(The Institute of Statistical Mathematics)

Estimation of hyper-parameters (macroscopic parameters) in Markov field models from incomplete data is important subject in spatial pattern analysis. The combination of type II likelihood method (ABIC method) and Metropolis-type Monte Carlo algorithm is the most general approach to this problem.

In this paper, this approach is tested in 2-dimensional Ising model, which is the most simple discrete Markov field model. The following numerical experiments are performed at various conditions :

1. Patterns generated by Metropolis algorithm at $J^{\text{true}} < J_c$ are sampled, where J_c is the critical coupling of 2-dim Ising model on square lattice ($1/J_c = 2.27$).
2. Observational noise is added to these patterns; i.e. the color (= sign) of each pixel (= spin) in these patterns is randomly changed with probability p^{true} .
3. The estimation algorithm is applied to these artificial data and estimated J and p are compared to J^{true} and p^{true} . Pattern recovery rate is also studied.

There are several types of Metropolis-type algorithm (Metropolis algorithm, Gibbs sampler, etc.). Among them Metropolis algorithm is used here. Boundary conditions are always periodic in the experiments.

At $1/J^{\text{true}} = 2.8$ good results are obtained when observational noise level is moderate ($p^{\text{true}} = 0.1$ or 0.2) and sample size is 40×40 . At $1/J^{\text{true}} = 3.3$ the algorithm also converges in most cases. Sample dependence is, however, rather high.

Pattern recovery rate is low in these cases. Local minima in hyper-parameter (J, p) space are not many (or not exist).

Experiments near to the critical point are also performed (sample size 60×60 , $1/J^{\text{true}} = 2.45$, $p^{\text{true}} = 0.25, 0.35$). Reasonable estimates of J and p are obtained. Pattern recovery rate is considerably high in this region (5~10%).