EM アルゴリズムの動的性質
— 確率推論におけるミクロとマクロの絡み合い —

井上 純一 †

Abstract : 不完全データから確率モデルのパラメータを推定するための常套手段として広く知られている EM アルゴリズムを統計物理の視点から解説する. データ間に強い相関がある大自由度確率モデルへの適用に際し, EM アルゴリズムの処理過程はミクロ/マクロな変数が相互に絡み合ったダイナミックスとなるが, その様相がアルゴリズムの収束性/精度に及ぼす影響をレプリカ法を用いずに, 画像修復を例にとった簡単な計算機実験を通して詳しく見ていきたい.

Keywords : EM アルゴリズム, 最尤推定, 画像修復, マルコフ確率場, マルコフ連鎖モンテカルロ法, 情報統計力学

1 はじめに

情報化社会が複雑になるにつれて, 大規模なデータを扱う情報システムを構築して望みの処理をさせたいときには適切な確率モデル/グラフィカル・モデルを選び出し, そこでの確率伝播法を用いた推論/予測を実行せよ, というのがスタンダード・アプローチになるような情勢へと世の中は確実に向かっているようだ. このような確率モデルはたいていの場合, 個々のデータを表現する確率変数の他, システムを特徴づけるパラメータをも含む. 例えば, 1 次元正規分布を複数張り合わせた混合分布では各々の正規分布の平均値と分散, そして混合比がそのパラメータとなる. こうした場合, パラメータの最尤推定には EM アルゴリズム [1, 2, 3, 4] を用いるのが常套手段であるが, つなぎ合わせる分布の個数が増えるほど, 計算に手間がかかって難しくなることは明らかであろう. しかし, パラメータがたとえ数個しかない場合であっても, データを表現する確率変数が高次元で高い相関を持つときには別種の問題が生じ, これは確率伝播法が解決してきた困難と同根のものである. 本稿では, そのような場合に EM アルゴリズムが直面する問題点と情報処理過程を [ミクロ] な確率変数と [マクロ] なパラメータが相互に絡み合ったダイナミックスという観点から解説する.

2 統計的手法に基づく画像の復元

典型的な大自由度/強相関を持つデータを扱う確率推論の例として, 画像修復の問題 [5] を取り上げる. 画像修復とは劣化された画像から原画像を復元する問題であり, 統計的な復元手法を用いる場合には劣化過程, 及び, 原画像の各々を確率モデルで表現する. ここでは 2 値画像に対する劣化過程として各画素が確率 $\mathrm{e}^{-h}/2\cosh h$ で独立に反転するものとし, 次の条件付き確率で表す.

$$P_h(\{\tau\}|\{\sigma\}) \quad = \quad Z_l^{-1}\,\mathrm{e}^{h\sum_{i=1}^{N}\tau_i\sigma_i},\ \ Z_l = \mathrm{tr}_{\{\tau\}}\mathrm{e}^{h\sum_{i=1}^{N}\tau_i\sigma_i} = (2\cosh h)^N \tag{1}$$

ここに N は全画素数であり, 劣化画像 $\{\tau\} = (\tau_1, \cdots, \tau_N)$, 推定画像 $\{\sigma\} = (\sigma_1, \cdots, \sigma_N)$ において位置 i の画素値をそれぞれ $\tau_i, \sigma_i \in \{-1, 1\}$ で表した. 特に平面画像を考える際には 2 次元格子点 (x, y) と i を一貫した規

† 〒 060-0814 札幌市北区北 14 条西 9 丁目 北海道大学大学院情報科学研究科

則で $i \mapsto (x,y)$ と関連つければよい. また, 画素に関するトレースを $\mathrm{tr}_{\{\tau\}}(\cdots) = \sum_{\tau_1=\pm 1}\cdots\sum_{\tau_N=\pm 1}(\cdots)$ で定義した. 本稿では一貫してこの表記を用いることに注意されたい. 一方, 原画像の確率モデルとしては, 同じ値をとる隣接画素対が高頻度で現れるように

$$P_J(\{\sigma\}) \quad = \quad Z_m^{-1}\mathrm{e}^{J\sum_{<ij>}\sigma_i\sigma_j},\ \ Z_m = \mathrm{tr}_{\{\sigma\}}\mathrm{e}^{J\sum_{<ij>}\sigma_i\sigma_j} \tag{2}$$

と選ぶ. ここで $<ij>$ は隣接する画素対を表し, 例えば 2 次元正方格子を考えるならば, (x,y) の位置にある画素の最隣接とは $(x+1,y),(x-1,y),(x,y+1),(x,y-1)$ の 4 点が対応する. 統計学では (1) 式を**尤度**, (2) 式を**事前確率**と呼ぶ. さて, 統計的な画像の復元を行う際にはこれら 2 つの確率を用いて**事後確率**: $P_{J,h}(\{\sigma\}|\{\tau\})$ を計算する. 具体的には**ベイズ公式**により直ちに

$$P_{J,h}(\{\sigma\}|\{\tau\}) \quad = \quad \frac{P_h(\{\tau\}|\{\sigma\})P_J(\{\sigma\})}{\mathrm{tr}_{\{\sigma\}}P_h(\{\tau\}|\{\sigma\})P_J(\{\sigma\})} = \frac{\mathrm{e}^{J\sum_{<ij>}\sigma_i\sigma_j+h\sum_i\tau_i\sigma_i}}{\mathrm{tr}_{\{\sigma\}}\mathrm{e}^{J\sum_{<ij>}\sigma_i\sigma_j+h\sum_i\tau_i\sigma_i}} \tag{3}$$

が得られる. 従って, 劣化画像 $\{\tau\}$ に対して事後確率を計算し, それを推定画像 $\{\sigma\}$ の関数としてみた場合, この事後確率を最大にする配列 $\{\sigma\}$ を推定値に選ぶ方策をとることができる. これを**事後確率最大化法**と呼ぶ. 上式より, 事後確率の最大化はエネルギー関数: $H(\{\sigma\}|\{\tau\}) = -J\sum_{<ij>}\sigma_i\sigma_j - h\sum_i\tau_i\sigma_i$ の最小化に等しい. このようにミクロな確率変数である各画素の推定値を求めるためには上記の最適化問題を解けばよい. しかし, このためには確率モデルを特徴づけるパラメータ J,h 自体も手持ちの観測データ $\{\tau\}$ から推定しなければならない. そこで, 次節では EM アルゴリズムを用いて J,h の最尤推定値を決定する手続きを見ていく. このアルゴリズムの処理過程を計算機実験により具体的に追ってみることにより, 本稿の副題に掲げた「ミクロとマクロの絡み合い」の意味が明らかになる.

3　EM アルゴリズムによるパラメータの最尤推定

確率モデルのパラメータ J,h を決定するには劣化画像 $\{\tau\}$ に対し, 次に定義される**周辺尤度**をコスト関数として導入し, パラメータに関する最適化問題を解けばよい (**周辺尤度最大化法**) [6].

$$\begin{aligned}\mathcal{L}(J,h:\{\tau\}) \quad &= \quad \log \mathrm{tr}_{\{\sigma\}}P_h(\{\tau\}|\{\sigma\})P_J(\{\sigma\}) \\ &= \quad \log \mathrm{tr}_{\{\sigma\}}\mathrm{e}^{J\sum_{<ij>}\sigma_i\sigma_j+h\sum_i\tau_i\sigma_i} - \log \mathrm{tr}_{\{\sigma\}}\mathrm{e}^{J\sum_{<ij>}\sigma_i\sigma_j} - \log 2\cosh h\end{aligned} \tag{4}$$

ここに, 上式右辺の第 2,3 項はそれぞれ, 事前分布, 尤度の規格化因子 Z_m, Z_l の対数をとったものであることに注意しょう. 劣化画像に関する平均操作 $[\cdots]_{\{\tau\}}$ を施した場合の周辺尤度が確率モデルのパラメータの真値 (J_*,h_*) で最大値をとること, つまり, 不等式: $[\mathcal{L}(J_*,h_*:\{\tau\})]_{\{\tau\}} \geq [\mathcal{L}(J,h:\{\tau\})]_{\{\tau\}}$ が成り立つことは確率分布 $P_{h_*}(\{\tau\}|\{\sigma\})P_{J_*}(\{\sigma\})$ と $P_h(\{\tau\}|\{\sigma\})P_J(\{\sigma\})$ 間のカルバック距離の非負性を用いることにより, あるいは可解モデルのレプリカ解析 [7] によって簡単に示すことができる. この問題において我々は原画像に関する情報を何も持たず, その部分のデータが欠落しているので, 推定値 $\{\sigma\}$ の自由度に関して和 $\mathrm{tr}_{\{\sigma\}}(\cdots) = \sum_{\sigma_1=\pm 1}\cdots\sum_{\sigma_N=\pm 1}(\cdots)$ をとり, 観測データ $\{\tau\}$ を残して確率分布 $P_h(\{\tau\}|\{\sigma\})P_J(\{\sigma\})$ を周辺化したものの対数をもってパラメータ決定のコスト関数としているわけである. 従って, パラメータの最尤推定値を求めるには周辺尤度 $\mathcal{L}$ をパラメータに関して最大化すればよい. その一手法として **EM アルゴリズム (Expectation-Maximization)** が知られている. この方法では周辺尤度を直接最大化せず, 尤度関数のステップ依存した事後確率での平均: $\mathrm{tr}_{\{\sigma\}}P_{J_t,h_t}(\{\sigma\}|\{\tau\})\log P_h(\{\tau\}|\{\sigma\})P_J(\{\sigma\})$ で定義される **Q 関数**を最大化する手続きにより周辺尤度を間接的に最大化する.

$$\begin{aligned}Q(J,h|J_t,h_t) \quad &= \quad -NJ\frac{\mathrm{tr}_{\{\sigma\}}\varepsilon_B\,\mathrm{e}^{J_t\sum_{<ij>}\sigma_i\sigma_j+h_t\sum_i\tau_i\sigma_i}}{\mathrm{tr}_{\{\sigma\}}\mathrm{e}^{J_t\sum_{<ij>}\sigma_i\sigma_j+h_t\sum_i\tau_i\sigma_i}} - Nh\frac{\mathrm{tr}_{\{\sigma\}}\varepsilon_C\,\mathrm{e}^{J_t\sum_{<ij>}\sigma_i\sigma_j+h_t\sum_i\tau_i\sigma_i}}{\mathrm{tr}_{\{\sigma\}}\mathrm{e}^{J_t\sum_{<ij>}\sigma_i\sigma_j+h_t\sum_i\tau_i\sigma_i}} \\ &- \quad \log \mathrm{tr}_{\{\sigma\}}\mathrm{e}^{J\sum_{<ij>}\sigma_i\sigma_j} - N\log 2\cosh h\end{aligned} \tag{5}$$

ここは 2 ページ目

ここに $\varepsilon_B = -(1/2N)\sum_{<ij>}\sigma_i\sigma_j, \varepsilon_C = -(1/N)\sum_i \tau_i\sigma_i$ である．EM アルゴリズムは Q 関数の計算 (Expectation) とその最大化 (Maximization)：

$$J_{t+1} = \arg\max_J Q(J,h|J_t,h_t),\ \ h_{t+1} = \arg\max_h Q(J,h|J_t,h_t) \tag{6}$$

の繰り返しで構成される．ここではまず劣化過程の反転確率 p が既知であるとし，$h = h_* = (1/2)\log((1-p)/p)$ とおいて J の更新のみを考えよう．このとき更新式の『方程式』は具体的に

$$u_m(J_{t+1}) \equiv \frac{\mathrm{tr}_{\{\sigma\}}\varepsilon_B \mathrm{e}^{J_{t+1}\sum_{<ij>}\sigma_i\sigma_j}}{\mathrm{tr}_{\{\sigma\}}\mathrm{e}^{J_{t+1}\sum_{<ij>}\sigma_i\sigma_j}} = \frac{\mathrm{tr}_{\{\sigma\}}\varepsilon_B \mathrm{e}^{J_t\sum_{<ij>}\sigma_i\sigma_j + h_*\sum_i\tau_i\sigma_i}}{\mathrm{tr}_{\{\sigma\}}\mathrm{e}^{J_t\sum_{<ij>}\sigma_i\sigma_j + h_*\sum_i\tau_i\sigma_i}} \equiv u_p(J_t,h_*) \tag{7}$$

と書ける．幸いなことに上式左辺 u_m はデータ $\{\tau\}$ を含まない確率変数 $\{\sigma\}$ のみからなる系における統計量 ε_B の期待値であり，十分に多数個 $(N\to\infty)$ の画素が 2 次元正方格子上に配置されている場合には以下に示す厳密解 $u_m(J)$ が知られており (Onsager (1944))，更新式が

$$u_m(J_{t+1}) = u_p(J_t,h_*),\ \ u_m(J) = -\coth 2J\left\{1+\frac{2}{\pi}(2\tanh^2 2J-1)K(k)\right\} \tag{8}$$

と書けることになる．ただし，$K(k)$ はその母数が $k = 2\tanh 2J/\cosh 2J$ で与えられる完全楕円関数：

$$K(k) = \int_0^{\pi/2}\frac{d\phi}{\sqrt{1-k^2\sin^2\phi}} \tag{9}$$

である．従って，具体的なアルゴリズムの処理過程は，まず初期値 J_0 を適当に設定し，$u_p(J_0,h_*)$ を計算する．ついで，その値 u_p と $u_m(J)$ の値が等しくなるような J を J_1 とし，その J_1 に対して再度 $u_p(J_1,h_*)$ を計算し，その値 u_p と $u_m(J)$ の値が等しくなるような J を J_2 とする $\cdots\cdot$ というように逐次的に進んで行くことになる．そしてパラメータ J は $J_0\to J_1\to J_2\to\cdots$ のように更新される．このダイナミックスの固定点が EM アルゴリズムの与える解である．このように書くだけならば容易だが，しかし問題はそう簡単ではない．$u_p(J,h_*)$ は観測データ $\{\tau\}$ を含む非一様な系での期待値であり，u_m のような厳密解は無い．従って，これを正確に評価するには 2^N の和 $\mathrm{tr}_{\{\sigma\}}(\cdots)$ を計算する必要があるのだが，N が大きな場合には，この計算が現実的ではなくなるのである．

この種の平均値 u_p の計算技法として最近では統計物理においてベーテ近似として知られる確率伝播法が脚光を浴び，確率推論/予測に関する多くの問題に適用され，成功していることは周知のとおりである (例えば，本特集の田中和之氏の記事を参照)．実際，上記の問題を切り抜ける場合でも確率伝播法の適用が可能なのだが，ここではあえて見方を変え，ミクロ変数である各画素 $\sigma_i(i=1,\cdots,N)$ に次の確率過程を課し，十分な時間が経過した後に系 $\{\sigma\}$ が平衡分布 (事後分布)：$\mathrm{e}^{-H(\{\sigma\}|\{\tau\})}/\mathrm{tr}_{\{\sigma\}}\mathrm{e}^{-H(\{\sigma\}|\{\tau\})}$ に収束することに着目する．このとき事後分布でのアンサンブル平均が上述確率過程からのサンプリング平均 (時間平均) で置き換えられることを利用し，式 (7) で与えられる ε_B の重み付き平均を $(1/N_{\mathrm{MCS}})\sum_{k=1}^{N_{\mathrm{MCS}}}\varepsilon_B(k)$ で近似計算することを考えよう．この方法を**マルコフ連鎖モンテカルロ法 (MCMC：Markov Chain Monte Calro 法)** と呼ぶ．

ミクロ変数の従う確率過程 (シングルスピンフリップ・メトロポリス法)

(i) $i\in\{1,2,\cdots,N\}$ をランダムに選び，その位置の画素を反転させる．$\sigma_i\to-\sigma_i$.
(ii) $\Delta E = H(F_i\{\sigma\}|\{\tau\}) - H(\{\sigma\}|\{\tau\})$ を計算し (F_i は位置 i の画素を反転させる演算子)，$\Delta E\le 0$ であれば，(i) の反転を受け入れ，$\Delta E>0$ であっても，確率：$\mathrm{e}^{-\Delta E}$ で反転を受け入れる．
(iii) (i)(ii) を $N\times N_{\mathrm{MCS}}$ 回繰り返す．

なお, この MCMC 法は u_m の厳密解を知らない場合, $H = -J\sum_{<ij>}\sigma_i\sigma_j$ と置くことで u_m の計算にも適用できる. つまり, EM アルゴリズムを適用する前に予め MCMC 法で J と u_m の対応関係をテーブルにしておく. そして J の更新毎に $u_p = u_m$ を与える J をそのテーブルから拾い出し, 該当値を次ステップの J 値とするように更新を進めればよい. 図 1 (左) に比較のため u_m の厳密解と MCMC 法による結果 ($N = 50\times 50, N_{\rm MCS} = 2\times 10^5$) を載せる. ところで, MCMC 法では $N_{\rm MCS}$ の値をいくらに設定するか, 言い換えれば, パラメータを更新する際, 直前のパラメータ値で特徴づけられる系がどのくらい平衡状態から遠いのか, によって結果が異なるので注意が必要である. 事実, 図 1 (左) からわかるように, u_m の厳密解と MCMC 法から求めた u_p の交点が EM アルゴリズムの決定する J 値を与えるのだが, この交点の値は $N_{\rm MCS} = 100$ と $N_{\rm MCS} = 2\times 10^5$ とでは有意な差があり, $N_{\rm MCS}$ を大きくとり, 系を十分に緩和させた場合の方が真値 $J_* = 0.465$ に近いことが見てとれる. 前に述べた確率伝播法は基本的に平均場近似の考え方に基づいており, ミクロ変数 (の期待値) の更新は EM アルゴリズムにおけるマクロ変数の更新式と同様に確定的な式で与えられ, 逐次的に画素間の相関を取り込むことにより精度が向上する. 一方, MCMC 法ではミクロ変数に確率過程を課し, そこから生成される系の時系列の長時間平均で期待値を近似計算するため, 系の平衡状態への緩和時間がミクロ変数だけでなく, パラメータの推定精度にとっても決定的である. これは MCMC 法を用いた EM アルゴリズムではミクロ変数の確率過程 (ダイナミックス) がマクロ変数 J, h に共役な統計量 $\varepsilon_B, \varepsilon_C$ の期待値を決定し, それが Maximization ステップを介して J, h に影響し, さらに, そうして更新された J, h がミクロ変数の確率過程にフィードバックされる ⋯ といった具合にミクロ変数とマクロ変数が相互に絡み合いながら情報処理が進んで行くことになるからである. 特に確率的情報処理の課題では事後確率が必ず何らかの観測データを含み, 空間的に非一様なものとなる. そこで統計物理の知見を借りるならば, 空間的に非一様な系の平衡状態への緩和は一般的に言って非常に遅い, という事実が知られている. そうであるならば Expectation ステップ, つまり u_p の計算では系の緩和に関して慎重にならざるを得ず, どのタイミングで処理を Maximiation ステップへと切り替えてマクロ変数を更新したならば, どの程度の精度が得られるのか, を調べることはアルゴリズムの性能を議論する上ではとても重要になる. それはここで示した簡単な数値実験からも明らかであろう.

さて, 最後に劣化度 h をも未知とし, これと J を EM アルゴリズムで同時推定しよう. h の更新式は $\hat{u}_p(J,h)$ を ε_C の事後確率での平均値として, $h_{t+1} = \tanh^{-1}(\hat{u}_p(J_t, h_t))$ と書けるので, これと h も未知であるとした場合の J の更新式 : $u_m(J_{t+1}) = u_p(J_t, h_t)$ とを組んで反復計算を行う. 結果を図 1 (右) に示す. この図より, 原画像として $J_* = 0.465$ と選んだ事前確率 (2) からのスナップショットに選び, 劣化率を $p = 0.1$ ($h_* = 1.1$) に定め, $N_{\rm MCS} = 100$, 及び, $N_{\rm MCS} = 2\times 10^5$ としてアルゴリズムを動作させた場合, それらの収束値には小さいが有意な差があることが見て取れる. なお, 当然のことながら図 2 (b1) のような印鑑等の自然画像に対しても MCMC 法に基づく EM アルゴリズムを適用することができる. そのパラメータの時間発展と最終的に得られる修復画像とをそれぞれ図 1 (右) と図 2 (b3) に載せておく.

4　おわりに

本稿では欠落データを含む確率モデルの最尤推定値を求めるために広く用いられている EM アルゴリズムを統計物理の切り口から説明してきた. 統計物理からみて EM アルゴリズムが興味深いのは, 観測データとして空間 (確率場) に乱れが入り込み, 結果として緩和が非常に遅い系のダイナミックスが処理過程の背後に現れるためである. 実際に本稿で示したように, そのダイナミックスの詳しい解析が EM アルゴリズムの性能を評価する上での一つのキーポイントとなっている. 一方, 統計物理では J や h などのパラメータは温度, 磁場など, 物質が置かれている環境を調節するための変数であり, 通常はこれらの値を固定した上での緩和過程が実験/理論的に詳しく調べるられている. しかし, EM アルゴリズムでは周辺尤度最大化原理に基づきそれらのパラメータ自体も異なる時間スケールで動いて行くわけで, そうした状況を統計物理の問題として考えて直してみることは意味が無いわけではない. 例えば事後確率がパラメータの選び方

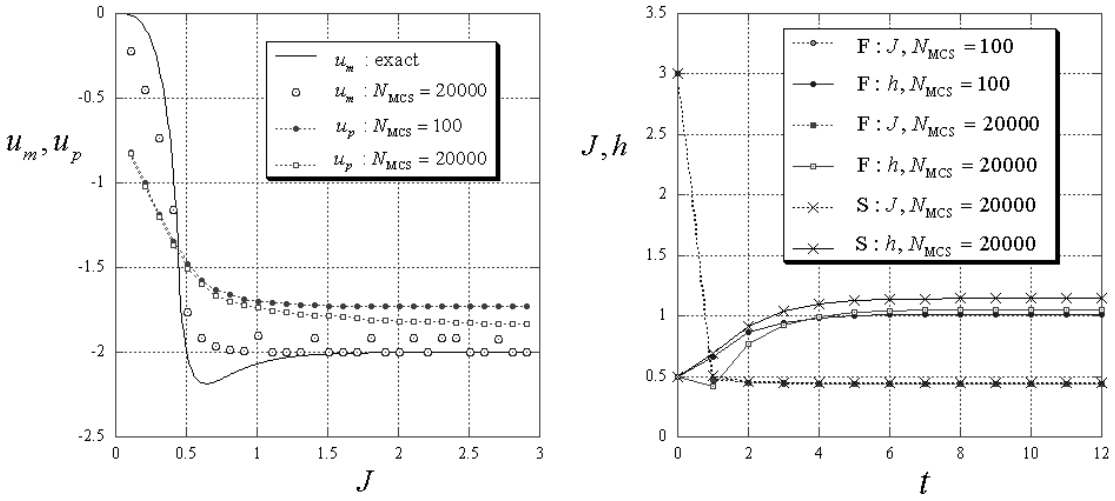

図 1: 左図は $u_m(J)$ の厳密解と $N_{\mathrm{MCS}} = 2 \times 10^5$ での MCMC 法による近似解, 及び, $u_p(J, h_*)$ の $N_{\mathrm{MCS}} = 100, 2 \times 10^5$ での MCMC 法による計算結果. u_m, u_p の交点が J の推定値を与える (真値は $J_* = 0.465$.) 右図は EM アルゴリズムの処理過程. プロット・キャプション F は $J_* = 0.465$ で与えられる事前確率 (2) からのスナップショットを原画像に選んだ場合であり, プロット・キャプション S は図 2 (b1) の印鑑画像を原画像としたものである. 劣化率はともに $p = 0.1$ ($h_* = 1.1$).

で相転移などの臨界現象を引き起すようなクラスの分布に対し, かつ, 真のパラメータ値がその臨界点に一致するような場合に EM アルゴリズムを動作させれば, 系は周辺尤度最大化原理に従って自己組織的にその臨界点へと向かうことになるであろう. そのような現象/緩和過程を詳しく解析してみることは興味深いのではなかろうか.

参考文献

[1] A. P. Dempster *et.al.*, *Journal of the Royal Statistics*, Series B (methodological), **39**, pp.1-38 (1977).

[2] 渡辺美智子, 山口和範 編著, 「EM アルゴリズムと不完全データの諸問題」, 多賀出版 (2000).

[3] 赤穂昭太郎, 情報処理, Vol. 37, No. 1, pp. 43-51 (1996).

[4] 宮川雅巳, 応用統計学, Vol. 16, No. 1, pp.1-19 (1987).

[5] K. Tanaka, *Journal of Physics A : Mathematical and General*, **35**, pp. R81-R150 (2002).

[6] 伊庭幸人, 統計数理 第 39 巻 第 1 号 pp. 1-21 (1991).

[7] J. Inoue and K. Tanaka, *Physical Review E*, **65**, pp. 016125-1 - 016125-11 (2002).

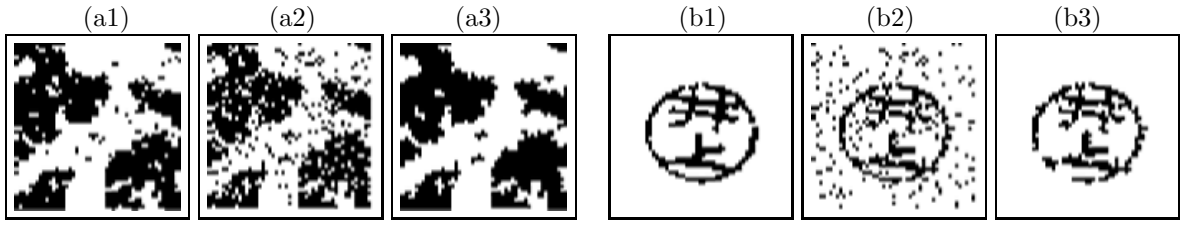

図 2: 原画像 (a1)(b1) ((a1) は $J_* = 0.465$ の事前分布 (2) からのスナップショット), 劣化画像 (a2)(b2) (劣化率 : $p = 0.1$) $N_{\rm MCS} = 2 \times 10^2$ での EM アルゴリズムで求められたパラメータ (図 1 (右) の収束点) で温度制御 $T = 3/\sqrt{t}$ でのシミュレーテッド・アニーリング法による最大事後確率推定により復元された画像 (a3)(b3).