1999 年情報論的学習理論ワークショップ
1999 Workshop on Information-Based Induction Sciences (IBIS'99)
Izu, Japan, August 26-27, 1999.

EM アルゴリズムの関連話題
– Wake-Sleep アルゴリズムと再帰的 EM –
Two related subjects of the EM algorithm
– The Wake-Sleep algorithm and the recursive EM algorithm–

池田 思朗*
Shiro Ikeda

Abstract: In this article, we show two topics of the EM (expectation-maximization) algorithm. One is the Wake-Sleep (W-S) algorithm which is proposed by P. Dayan and G. Hinton. The algorithm was believed to converge by the similarity between W-S and EM algorithms. But we have clarified that they are different and the convergence is not clear in general cases. The other topic is the recursive EM algorithm. This algorithm approximate the Fisher's scoring method by using the EM algorithm recursively and tries to accelerate the EM algorithm. We show the background of the algorithm and some numerical simulations.

1 はじめに

EM(Expectation Maximization) アルゴリズムは，直接観測できない確率変数をもつ確率モデルの最尤推定のために，Dempster ら [4] によって提案された．EM アルゴリズムは様々なモデルに広く用いられており，音声認識で広く使われている HMM (Hidden Markov Model)[10] などでは大きな成功を納めている．具体的な計算はモデル毎に異なるが，EM アルゴリズムは繰り返し演算で最尤推定を求める手法である．各繰り返しで行う演算は通常簡単であり，その導出も容易であることが EM の特徴である．

今回は IBIS に招待されたことから，以前から行なってきた EM アルゴリズムに関する話題を提共したい．ここでは特に 2 つの話題について述べる．

1 つめの話題は Wake-Sleep アルゴリズムについてである．このアルゴリズムは，Helmholtz マシン [3, 5] に対する学習則として提案された．当初 Helmholtz マシンは $\{0,1\}$ を確率的に出力する細胞が集まったネットワークとして提案された．Helmholtz マシンは 2 つの層から成り，その層の間で双方向に 2 のモデルを別のパラメータによって定義する．このモデルに対し提案された学習則 Wake-Sleep アルゴリズムはその単純さから注目を浴びたが収束性については明らかではなかった．この収束性について説明した結果を紹介する．

2 つめは再帰的 EM アルゴリズムである．

EM アルゴリズムは一般に収束が遅い．一方，最尤推定法の 1 つである Fisher のスコアリング法 (スコアリング法) [8] は EM アルゴリズムと同様に繰り返し演算で最尤推定を行い，収束は EM アルゴリズムよりも速い．ただし，各繰り返し演算の計算では Fisher の情報行列の逆行列を求めなければいけない．これは神経回路網モデルなどでは計算が難しい．そこで，再帰的に EM アルゴリズムを用いることで EM アルゴリズムとスコアリング法を結びつけることができることを示す．

このアルゴリズムは 2 つの段階からなる．まず，与えられたデータを用いて通常の EM アルゴリズムを行う．次の段階では，与えられたデータではなく，モデル自身がデータを作り出し，そのデータを用いて EM アルゴリズムを行う．この 2 つの段階を通じて得たパラメータを用いると，単に EM アルゴリズムを行うよりも良いパラメータを作りだせる．アルゴリズムの理論的導出と計算機実験の結果を示す．

同様にスコアリング法を近似する手法は EM アルゴリズムを加速する方法としていくつか提案されている．それらの手法と比べれば，再帰的な EM アルゴリズムの計算量は等しいか少なくてすみ，安定している．また，

*科技団，さきがけ研究 21 〒 351-0198 埼玉県和光市広沢 2-1 理研 脳総研 情報創成 tel. 048-467-9663, e-mail shiro@brain.riken.go.jp, Presto, JST, Information Synthesis lab. BSI, RIKEN, Wako, Saitama, Japan

再帰的な EM も含め，これらのスコアリング法の近似手法が，計算量も含めて EM アルゴリズムを加速しているかどうかは真の分布とモデルの分布との関係によっており，一般には明らかではない．この点についてもシミュレーションを通じて示す．

2 EM アルゴリズムと Wake-Sleep アルゴリズム

2.1 EM アルゴリズム

確率変数を $X = (Y, Z)$ とし，Y は観測できる確率変数，Z は観測できない隠れた確率変数と定義する．$p(x;\boldsymbol{\theta}) = p(y, z;\boldsymbol{\theta})$ である．ここでは，$p(x;\boldsymbol{\theta})$ は指数型分布族であるとして扱う．

$$p(x;\boldsymbol{\theta}) = \exp\left(\sum_{i=1}^{n}\theta^i r_i(x) - k(\boldsymbol{r}(x)) - \psi(\boldsymbol{\theta})\right). \quad (1)$$

$\boldsymbol{\theta} = (\theta^1, \cdots, \theta^n)^T$ は自然母数と呼ばれ $\psi(\boldsymbol{\theta})$ はその関数である．また $\boldsymbol{r}(x) = (r_1(x), \cdots, r_n(x))^T$ であり $k(\boldsymbol{r}(x))$ はその関数である．$p(x;\boldsymbol{\theta})$ の y についての周辺分布は，

$$p(y;\boldsymbol{\theta}) = E_{p(z;\theta)}\left[p(x;\boldsymbol{\theta})\right] = \int p(x;\boldsymbol{\theta})d\mu(z)$$

と表され $p(x;\boldsymbol{\theta})$ が指数型分布族であっても $p(y;\boldsymbol{\theta})$ は必ずしも指数型分布族には属さない．

サンプルとして得られるデータは観測できる確率変数 y についての経験分布 $\hat{q}(y) = \sum_{s=1}^{N}\delta(y_s)/N$ $\{y_1, \cdots, y_N\}$ のみである．$\hat{q}(y)$ から $\boldsymbol{\theta}$ を推定したい．

$l(y;\boldsymbol{\theta}) = \log p(y;\boldsymbol{\theta})$ とすると対数尤度は，

$$L(Y^N;\boldsymbol{\theta}) \stackrel{\text{def}}{=} \frac{1}{N}\sum_{s=1}^{N} l(y_s;\boldsymbol{\theta}) = E_{\hat{q}(y)}\left[l(y;\boldsymbol{\theta})\right],$$

となる．この最尤推定を行う場合に EM アルゴリズムを適用できる．

EM アルゴリズムは繰り返し演算で最尤推定を求めるアルゴリズムであり，ある初期パラメータ $\boldsymbol{\theta}_0$ からパラメータを更新していく．新しいパラメータ $\{\boldsymbol{\theta}_t\}$ $(t = 1, 2, 3, \cdots)$ を求める際には，次の 2 つの手続きを行う．

- Expectation-ステップ: $Q(\boldsymbol{\theta}, \boldsymbol{\theta}_t)$ を計算する

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}_t) = E_{\hat{q}(y)p(z|y;\theta_t)}\left[l(y, z;\boldsymbol{\theta})\right]$$

- Maximization-ステップ: $Q(\boldsymbol{\theta}, \boldsymbol{\theta}_t)$ を最大にするパラメータを求める．

$$\boldsymbol{\theta}_{t+1} = \operatorname*{argmax}_{\theta} Q(\boldsymbol{\theta}, \boldsymbol{\theta}_t)$$

EM ステップにより $\boldsymbol{\theta}_t$ から $\boldsymbol{\theta}_{t+1}$ を得るが，この新たなパラメータに関して尤度の増加 $L(Y^N;\boldsymbol{\theta}_{t+1}) \geq L(Y^N;\boldsymbol{\theta}_t)$ が示せる [4]．EM ステップを繰り返すとパラメータは収束し，これが最尤推定であると考えられる．

EM アルゴリズムの情報幾何的な解釈としては甘利 [2] が行なった結果がある．確率変数 $\boldsymbol{X}$ を考え，確率密度関数 $p(\boldsymbol{x})$ の空間 S 考える．パラメタ $\boldsymbol{\theta}$ で表現される確率密度関数 $p(\boldsymbol{x};\boldsymbol{\theta})$ を確率モデルとすると，集合 $\{p(\boldsymbol{x};\boldsymbol{\theta})\}$ は空間 S の中で部分多様体を成す．これをモデル多様体 M と呼ぶことにする．一方，周辺分布が観測データの経験分布に一致する分布全体を観測多様体 D と呼ぶ．この 2 つの多様体の間を e 射影と m 射影 [1] を繰り返すのが EM アルゴリズムと考えられる (図 1)．

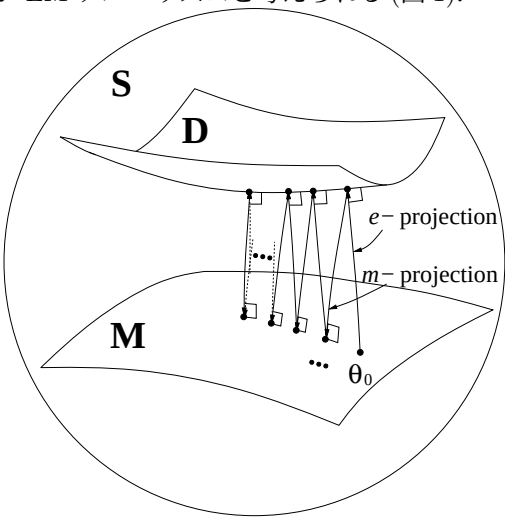


図 1: EM アルゴリズムの情報幾何的な解釈

2.2 Helmholtz マシンと W-S アルゴリズム

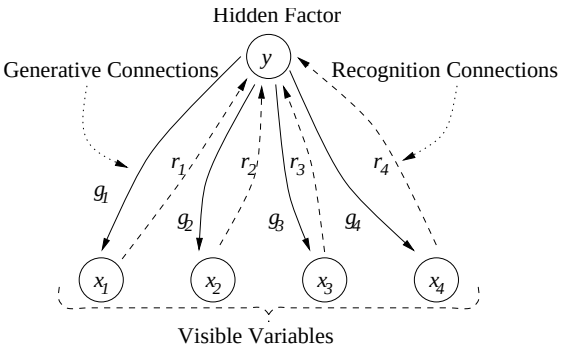


図 2: Helmholtz マシン

ここで Helmholtz マシン [3] とその学習法である Wake-Sleep (W-S) アルゴリズムについて述べる．Helmholtz マシンは，神経回路網モデルの 1 つとして提案された．このモデルには外部からの信号を受ける下

位の細胞の層と外からは直接観測できない高次の細胞の層がある．ここでは簡単のために高次の細胞は 1 つとする．この 2 つの層からなるネットワークに 2 組のパラメータを与える．1 つは入力を外から受け，入力層から高次への結合を持つ認識モデル (recognition model) であり，もう 1 つは高次から低次への結合を持つ生成モデル (generative model) である．前章の内容に合わせて書くと認識モデルはモデル多様体 M を表わし，生成モデルはデータ多様体 D を表わしている．この 2 つの多様体に別々のパラメータを与えたのが Helmholtz マシンである．

最も単純な例としては，線型なモデルに正規分布のノイズを加えたものがあげられる．これは因子を 1 つもつ因子分析のモデルと同値である [9]．認識モデルと生成モデルを以下に示す．

認識モデル $z = \boldsymbol{r}^T \boldsymbol{y} + \delta,$

$\boldsymbol{r}$ は n 次元の実ベクトルで $\delta \sim \mathcal{N}(0, s^2)$ は雑音である．これは z を観測できない確率変数として，$\boldsymbol{y}$ が与えられたときの z の条件付き確率を記述したモデルである．これによってデータ多様体 D を構成できる．

生成モデル $\boldsymbol{y} = z\boldsymbol{g} + \boldsymbol{\epsilon},$

$\boldsymbol{y} = (y_1, \cdots, y_n)^T$ は n 次元の実ベクトルであり，$z \sim \mathcal{N}(0, 1)$ は隠れ変数である．$\boldsymbol{g}$ は “因子負荷” と呼ばれ，$\boldsymbol{\epsilon} \sim \mathcal{N}(0, \Sigma)$ は正規分布に従う雑音である．ただしその共分散行列は対角行列である．$\Sigma = \mathrm{diag}(\sigma_i^2)$. 生成モデルは $\boldsymbol{y}$ と z の確率分布を与えるものであり，モデル多様体 M を構成する．

データ $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_N$ が与えられたとき，モデルのパラメータを最尤推定したい．W-S アルゴリズムもこの目的で提案された． W-S アルゴリズムでは，次の手続きによって学習を行なう [9].

Wake-phase: データ $\{\boldsymbol{y}_s\}$ から $\boldsymbol{y}$ をランダムに選び，認識モデルを用いて z を生成する，$z = \boldsymbol{r}_t^T \boldsymbol{y} + \delta, \delta \sim \mathcal{N}(0, s_t^2)$. 生成モデルの $\boldsymbol{g}$ と Σ を次のように更新する．α は小さな正の定数で β は 1 から少しだけ小さい定数，$\overline{\ }$ はこのようにして得られた $\boldsymbol{y}$ と z についての平均である．

$$\boldsymbol{g}_{t+1} = \boldsymbol{g}_t + \alpha\overline{(\boldsymbol{x} - \boldsymbol{g}_t y)y} \quad (2)$$

$$\sigma_{i,t+1}^2 = \beta\sigma_{i,t}^2 + (1-\beta)\overline{(x_i - g_{i,t}y)^2}, \quad (3)$$

Sleep-phase: 生成モデルに従って $\boldsymbol{y} = z\boldsymbol{g}_{t+1} + \boldsymbol{\epsilon}, z \sim \mathcal{N}(0, 1), \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathrm{diag}(\sigma_{t+1}^2))$, $\boldsymbol{y}$ と z を生成する．$\boldsymbol{r}$ s^2 を次のように更新する．

$$\boldsymbol{r}_{t+1} = \boldsymbol{r}_t + \alpha\overline{(y - \boldsymbol{r}_t^T \boldsymbol{x})\boldsymbol{x}} \quad (4)$$

$$s_{t+1}^2 = q(\boldsymbol{\eta})\beta s_t^2 + (1-\beta)\overline{(y - \boldsymbol{r}_t^T \boldsymbol{x})^2}. \quad (5)$$

W-S アルゴリズムは実現が容易であること，local な情報のみから学習できることから注目を集めた．当初，EM との類似性からその収束性が信じられていたが，パラメータの更新則を考えると必ずしもその収束性は明らかではない．$\boldsymbol{y}$ についての経験分布を $\hat{q}(\boldsymbol{y})$ として，認識モデルを $\hat{q}(\boldsymbol{y})q(z|\boldsymbol{y}; \boldsymbol{r}, s^2) = q(\boldsymbol{\eta})$ 生成モデルを $p(\boldsymbol{y}, z; \boldsymbol{g}, \Sigma) = p(\boldsymbol{\theta})$ と書くと，$KL(,)$ を Kullback-Leibler のダイバージェンス，γ を小さな定数として，

W-phase:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \gamma \frac{\partial KL(q(\boldsymbol{\eta}_t), p(\boldsymbol{\theta}_t))}{\partial \boldsymbol{\theta}}$$

S-phase:

$$\boldsymbol{\eta}_{t+1} = \boldsymbol{\eta}_t + \gamma \frac{\partial KL(p(\boldsymbol{\theta}_{t+1}), q(\boldsymbol{\eta}_t))}{\partial \boldsymbol{\eta}}$$

と書ける．KL ダイバージェンスは対称ではないので，W-phase と S-phase では異なる量を最小化していることが分る．線型のモデルの場合には常に $p(z|\boldsymbol{y}; \boldsymbol{g}, \Sigma) = q(z|\boldsymbol{y}; \boldsymbol{r}, s^2)$ となる $\{\boldsymbol{r}, s^2\}$ が存在し，Sleep phase を十分長く取ることで $KL(p(\boldsymbol{\theta}_{t+1}), q(\boldsymbol{\eta}))$ を最小にする $\boldsymbol{\eta}$ は $KL(q(\boldsymbol{\eta}), p(\boldsymbol{\theta}_{t+1}))$ も最小にすることが分り，収束性が確認できる [6].

ここでは因子が 1 つのモデルについてのみ述べた．一般の因子分析のモデルでは z はベクトルとなり $\boldsymbol{z}$ に関して認識モデルと生成モデルを定義でき，W-S アルゴリズムも作れる．しかし，認識モデルのノイズの共分散行列が対角行列とならないことから Wake-phase がローカルなデータの生成モデルでなくなり，神経回路網モデルとしてはあまり面白味がない．

Helmholtz マシンの特徴は M 多様体と D 多様体に別のパラメータを用いることである．これに対し EM を行なうことはできる．一方 W-S アルゴリズムを行なう場合にはそれぞれの phase で異なる関数を最小化しているため，一般的には収束性が明らかでない．

3 再帰的 EM アルゴリズム

3.1 EM アルゴリズムとスコアリング法との関係

ここでは再帰的 EM アルゴリズムについて説明する．

まず，EM アルゴリズムの一回の EM ステップを通じて得られるパラメータの性質について述べる．一回の EM ステップで $\boldsymbol{\theta}_t$ から $\boldsymbol{\theta}_{t+1}$ が得られたとする．このとき次の近似が成り立つ (証明に関しては [8](3.76),[12] を参照のこと)．なお，この近似が成り立つのは，指数型分布族の場合のみで，曲指数型分布族では成り立たない．

$$\boldsymbol{\theta}_{t+1} \simeq \boldsymbol{\theta}_t + {G_X}^{-1}(\boldsymbol{\theta}_t)\partial L(Y^N;\boldsymbol{\theta}_t). \tag{6}$$

ここで $\partial = (\partial_1, \cdots, \partial_n)^T = (\partial/\partial\theta^1, \cdots, \partial/\partial\theta^n)^T$ であり，$G_X(\boldsymbol{\theta}) = (g_{X\,ij}(\boldsymbol{\theta}))$ は確率分布 $p(x;\boldsymbol{\theta})$ の Fisher 情報行列である．定義は，次の通りである．

$$\begin{aligned} g_{X\,ij}(\boldsymbol{\theta}) &= E_{p(x;\theta)}\left[\partial_i l(x;\boldsymbol{\theta})\partial_j l(x;\boldsymbol{\theta})\right] \\ &= -E_{p(x;\theta)}\left[\partial_i\partial_j l(x;\boldsymbol{\theta})\right]. \end{aligned}$$

次に Fisher のスコアリング法について述べる．スコアリング法も繰り返し演算によってパラメータを更新するが，その更新ルールは，

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + {G_Y}^{-1}(\boldsymbol{\theta}_t)\partial L(Y^N;\boldsymbol{\theta}_t), \tag{7}$$

と表される．スコアリング法は EM アルゴリズムよりも収束が速いことが知られている．これは (6) 式と (7) 式の係数行列 $G_X(\boldsymbol{\theta})^{-1}$ と $G_Y(\boldsymbol{\theta})^{-1}$ の差によって生じる．$G_Y(\boldsymbol{\theta}) = (g_{Y\,ij}(\boldsymbol{\theta}))$ も $G_X(\boldsymbol{\theta})$ と同様に Fisher 情報量行列であるが周辺分布 $p(y;\boldsymbol{\theta})$ の情報量行列である．

$$\begin{aligned} g_{Y\,ij}(\boldsymbol{\theta}) &= E_{p(y;\theta)}\left[\partial_i l(y;\boldsymbol{\theta})\partial_j l(y;\boldsymbol{\theta})\right] \\ &= -E_{p(y;\theta)}\left[\partial_i\partial_j l(y;\boldsymbol{\theta})\right]. \end{aligned}$$

$G_X(\boldsymbol{\theta})$ と $G_Y(\boldsymbol{\theta})$ との間には次の関係式が成り立つ．

$$G_Y(\boldsymbol{\theta}) = G_X(\boldsymbol{\theta}) - G_{Z|Y}(\boldsymbol{\theta}) \tag{8}$$

$G_{Z|Y} = (g_{Z|Y\,ij}(\boldsymbol{\theta}))$ は次のように定まる条件付き Fisher 情報量行列である．

$$\begin{aligned} g_{Z|Y\,ij}(\boldsymbol{\theta}) &= -E_{p(y;\theta)}\left[E_{p(z|y;\theta)}\left[\partial_i\partial_j l(z|y;\boldsymbol{\theta})\right]\right] \\ &= E_{p(y;\theta)}\left[g_{Z|y\,ij}(\boldsymbol{\theta})\right]. \end{aligned}$$

G_Y，G_X，$G_{Z|Y}$ は一般に正定値対称行列である．

スコアリング法で用いる Fisher 情報量行列 ${G_Y}^{-1}$ は EM アルゴリズムの対象となる確率分布では直接求めることが難しい．そこで，EM アルゴリズムを用いてスコアリング法を近似する手法を提案した．理論的導出には次の定理が重要となる．

定理 1. ${G_Y}^{-1}$ は次のように G_X, $G_{Z|Y}$ によって展開できる．

$${G_Y}^{-1} = \left(I + \sum_{i=1}^{\infty}({G_X}^{-1}G_{Z|Y})^i\right){G_X}^{-1} \tag{9}$$

証明　(9) 式は，G_Y，G_X，$G_{Z|Y}$ の同時対角化により簡単に導かれる [8]．

この結果を用いると (7) 式は，

$$\begin{aligned} \boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t + {G_Y}^{-1}\partial L(Y^N;\boldsymbol{\theta}_t) \\ &= \boldsymbol{\theta}_t + {G_X}^{-1}\partial L(Y^N;\boldsymbol{\theta}_t) \\ &\quad + {G_X}^{-1}G_{Z|Y}{G_X}^{-1}\partial L(Y^N;\boldsymbol{\theta}_t) \\ &\quad + ({G_X}^{-1}G_{Z|Y})^2{G_X}^{-1}\partial L(Y^N;\boldsymbol{\theta}_t) \\ &\quad + \cdots \end{aligned} \tag{10}$$

と書き直せる．(6) 式と (10) 式を比べると，EM アルゴリズムはスコアリング法を G_X で展開したときの 1 次近似だとみなせる．

3.2　再帰的 EM アルゴリズム

スコアリング法はパラメータ $\boldsymbol{\theta}$ を計量 G_Y に基づいて最急降下の方向に更新していく．これは通常 EM アルゴリズムよりも収束が速い．しかし，${G_Y}^{-1}$ の計算は簡単でない場合も多い．EM アルゴリズムを再帰的に用いてスコアリング法を近似する手法について説明する．

ある $\boldsymbol{\theta}_t$ から一度 EM ステップを行い，パラメータを一度更新したとする．このとき得られた $\boldsymbol{\theta}_{t+1}$ は，一つの確率分布 $p(y;\boldsymbol{\theta}_{t+1})$ を与える．そこで，経験分布の $\hat{q}(y)$ の代わりに $p(y;\boldsymbol{\theta}_{t+1})$ を真の分布としてパラメータ $\boldsymbol{\theta}_t$ を EM ステップで更新する．もし $p(y;\boldsymbol{\theta}_{t+1})$ が連続分布の場合には $p(y;\boldsymbol{\theta}_{t+1})$ にしたがってデータを生成して，そのデータを用いて学習を行う．離散分布の場合には $p(y;\boldsymbol{\theta}_{t+1})$ そのものを真の分布として学習を行えば良い．EM ステップを 1 回行ったあとで得られたパラメータを $\bar{\boldsymbol{\theta}}_{t+1}$ とすると，この新たに得られたパラメータは $\boldsymbol{\theta}_t$ とも $\boldsymbol{\theta}_{t+1}$ とも異なる．$\boldsymbol{\theta}_t$，$\boldsymbol{\theta}_{t+1}$，$\bar{\boldsymbol{\theta}}_{t+1}$ の 3 つのパラメータから，より良い推定量を作り出す (図 3)．まず $\bar{\boldsymbol{\theta}}_{t+1}$ の持つ性質を示す．

定理 2. $p(y;\boldsymbol{\theta}_{t+1})$ を真の分布とし，$\boldsymbol{\theta}_t$ から一度 *EM* ステップを行い，得られたパラメータを $\bar{\boldsymbol{\theta}}_{t+1}$ とする．このとき，$\bar{\boldsymbol{\theta}}_{t+1}$ には次の性質がある．

$$\bar{\boldsymbol{\theta}}_{t+1} - \boldsymbol{\theta}_t \simeq {G_X}^{-1}G_Y{G_X}^{-1}\partial L(Y^N;\boldsymbol{\theta}_t). \tag{11}$$

証明　付録 A を参照のこと．

(6) 式，(8) 式と (11) 式から，

$$\begin{aligned} & \bar{\boldsymbol{\theta}}_{t+1} - \boldsymbol{\theta}_t \\ \simeq\ & {G_X}^{-1}(G_X - G_{Z|Y}){G_X}^{-1}\partial L(Y^N;\boldsymbol{\theta}_t) \\ \simeq\ & (\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t) - {G_X}^{-1}G_{Z|Y}{G_X}^{-1}\partial L(Y^N;\boldsymbol{\theta}_t) \end{aligned} \tag{12}$$

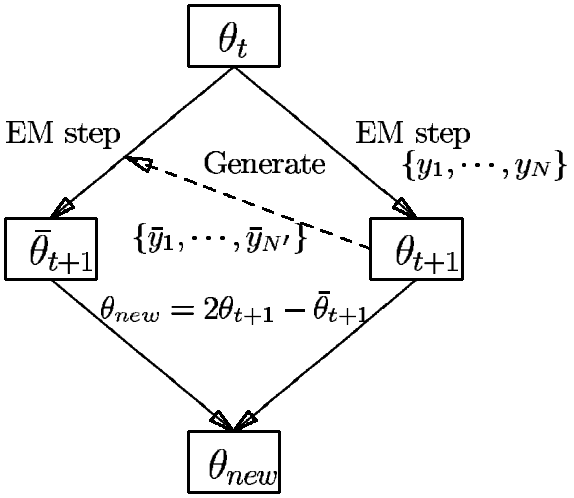


図 3: アルゴリズムの概要

が得られる．スコアリング法の 2 次の項の近似は.

$$\begin{aligned} & {G_X}^{-1} G_{Z|Y} {G_X}^{-1} \partial L(Y^N; \boldsymbol{\theta}_t) \\ \simeq \;& (\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t) - (\bar{\boldsymbol{\theta}}_{t+1} - \boldsymbol{\theta}_t) = \boldsymbol{\theta}_{t+1} - \bar{\boldsymbol{\theta}}_{t+1}, \end{aligned}$$

となる．スコアリング法の 2 次までの近似は,

$$\begin{aligned} \boldsymbol{\theta}' &= 2\boldsymbol{\theta}_{t+1} - \bar{\boldsymbol{\theta}}_{t+1} \\ &= \boldsymbol{\theta}_t + (\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t) + (\boldsymbol{\theta}_{t+1} - \bar{\boldsymbol{\theta}}_{t+1}) \\ &\simeq \boldsymbol{\theta}_t + {G_X}^{-1}(I + G_{Z|Y}{G_X}^{-1}) \partial L(Y^N; \boldsymbol{\theta}_t), \end{aligned}$$

とすればよい．また，同様の手法を用いて更に高次までスコアリング法を近似できる．

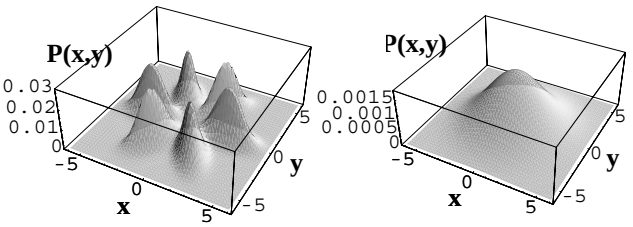
系 1. $p(y; \bar{\boldsymbol{\theta}}_{t+i-1})$ を真の分布 *(*教師*)* として $\boldsymbol{\theta}_t$ から *EM* ステップを一回行い，得られたパラメータを $\bar{\boldsymbol{\theta}}_{t+i}$ とする *(*$i = 1, 2, \cdots,$ であり，$\bar{\boldsymbol{\theta}}_t = \boldsymbol{\theta}_{t+1}$ と定める*)*. $\bar{\boldsymbol{\theta}}_{t+i}$ は次の性質を持つ.

$$\begin{aligned} \bar{\boldsymbol{\theta}}_{t+i} - \boldsymbol{\theta}_t &\simeq ({G_X}^{-1} G_Y)^i {G_X}^{-1} \partial L(Y^N; \boldsymbol{\theta}_t) \\ &= (I - {G_X}^{-1} G_{Z|Y})^i {G_X}^{-1} \partial L(Y^N; \boldsymbol{\theta}_t) \end{aligned}$$

証明　定理 2 の証明と同じ方法で行えば良い (付録 A).

この結果を用いると，$\bar{\boldsymbol{\theta}}_t$, $\cdots$, $\bar{\boldsymbol{\theta}}_{t+i}$, と $\boldsymbol{\theta}_t$ から，$({G_X}^{-1} G_{Z|Y})^i {G_X}^{-1} \partial L(Y^N; \boldsymbol{\theta}_t)$ が近似でき，スコアリング法を i 次まで近似できる．ただし，対象とするモデルが連続分布の場合，次章のシミュレーションのように Monte Carlo 的な手法を用いる必要があるため，2 次以上の近似は誤差が大きくなり安定しない.

また $i \geq n$ であれば，それ以上は線型従属となるから，実際に EM ステップを行うまでもなく，線形演算でより高次の近似を順次求めることができる.



モデル 1　　モデル 2

図 4: 学習に使った 2 つのモデル

3.3　シミュレーション

実際にアルゴリズムがどのように働くかを知るために，2 次元の正規混合分布 [13] を用いて実験を行った．図 4 に 2 つのモデルの密度関数を示す．両方とも 6 つの正規分布の重ね合わせで定義されている．ただし，モデル 2 では分散が大きいので，個々の正規分布は全体の分布からは確認できない.

混合正規分布に対する EM アルゴリズムは簡単で，その計算量は少ない．アルゴリズムを使って次のように実験を行った.

1. 教師分布から y について 1000 個のサンプルを作る．モデルの初期分布のパラメータ $\boldsymbol{\theta}_0$ を定める.
2. 教師分布から得られたデータを用いて，EM ステップを一回行い，$\boldsymbol{\theta}_t$ から $\boldsymbol{\theta}_{t+1}$ を得る.
3. 1000 個の新しいデータを $p(y; \boldsymbol{\theta}_{t+1})$ から生成する.
4. 新しく作られたデータを用いて，EM ステップを一回行い，$\boldsymbol{\theta}_t$ から $\bar{\boldsymbol{\theta}}_{t+1}$ を求める.
5. 新しいパラメータを $\boldsymbol{\theta}_{new} = 2\boldsymbol{\theta}_{t+1} - \bar{\boldsymbol{\theta}}_{t+1}$ とし，$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{new}$ と定めて，2 へもどる.

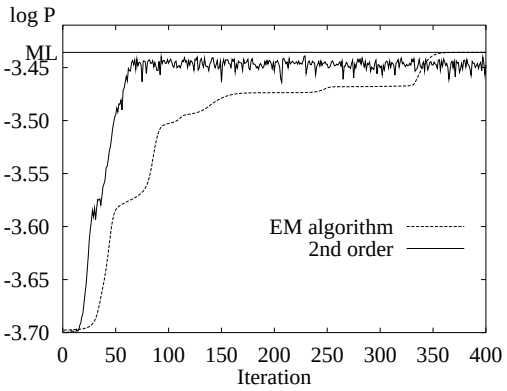


図 5: 対数尤度の変化

学習の際に尤度がどのように変化していくかを図 5 に示す．提案するアルゴリズムでは完全に収束すせず，絶えずふらついてる.

ここで，計算量の点から提案するアルゴリズムの 1 ステップを見直してみる．提案するアルゴリズムの 1 ステップは実際には 2 ステップの EM アルゴリズムを含んでいる．もし，計算量も含めて EM アルゴリズムと速さを比べるのであれば，提案するアルゴリズムの横軸を変えて比べる方が適切であろう．図 6 に結果を示す．

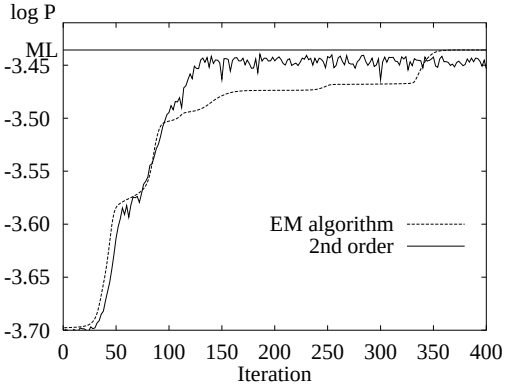


図 6: 計算量を考慮した対数尤度の変化

後半のふらつきをなくし，途中から通常の EM アルゴリズムに変えることを考える．ここでは，切替えるタイミングを決めるため，

$$\begin{aligned} \lambda(t) &= \eta\lambda(t-1) + (1-\eta)L(Y^N;\boldsymbol{\theta}_t), t = 1, \cdots, \\ \lambda(0) &= L(Y^N;\boldsymbol{\theta}_0) \end{aligned} \tag{13}$$

という関数を用いて $\lambda(t)$ の値が下がったら，通常の EM アルゴリズムに切替えることにした．なお，η は 0.7 とした．結果を図 7 に示す．この結果をみると，ほぼ 3 倍程度収束が速いことがわかる．提案したアルゴリズムと EM アルゴリズムを組み合わせることで，速く収束するアルゴリズムを構成できる．

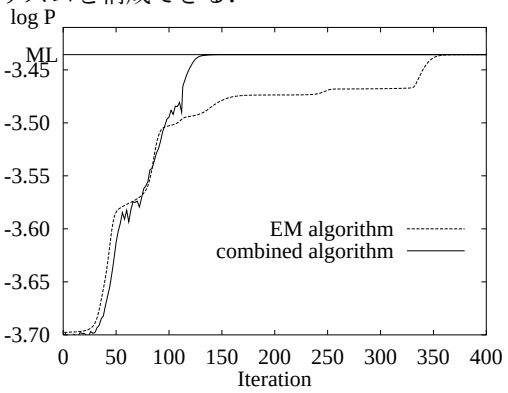


図 7: EM アルゴリズムと組合せた場合

3.4 考察

情報幾何的にはこのアルゴリズムは図 8 のように理解できる．図中の D と M は図 1 と同様である．再帰的に

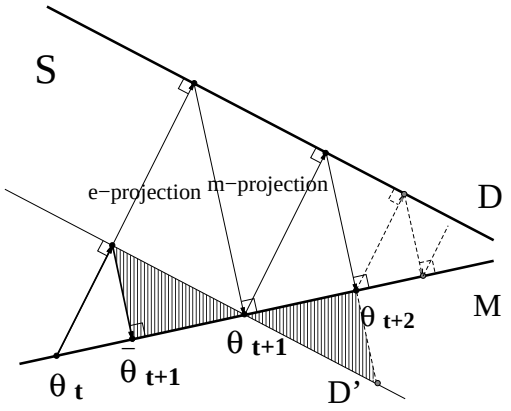


図 8: 再帰的 EM アルゴリズムの幾何

EM を用いるときには，一度 EM を行なった後，その $p(x;\boldsymbol{\theta}_{t+1})$ からデータを作り，一度 EM ステップを行なう．これは $p(x;\boldsymbol{\theta}_{t+1})$ を通る新たな多様体 D' を作り，この M と D' との間で e-projection と m-projection を行なっていることと等しい．この D' と M の間で一度 EM ステップを行なったものが $\bar{\boldsymbol{\theta}}_{t+1}$ である．これに対し，単に 2 度 EM ステップを行なった結果得られた $\boldsymbol{\theta}_{t+2}$ を図示してある．図中に 2 つの三角形が示してあるが，直感的には，全てが高次元の線型空間であれば，2 つの三角形は合同であり，$\boldsymbol{\theta}_{new} = \boldsymbol{\theta}_{t+2}$ であることが確かめられる．もちろん D も M も多様体であるので一般的には $\boldsymbol{\theta}_{new} \neq \boldsymbol{\theta}_{t+2}$ であり，そのずれによって $L(Y^N;\boldsymbol{\theta}_{new})$ と $L(Y^N;\boldsymbol{\theta}_{t+2})$ のどちらの尤度が大きいかは変わってくる．

ここに 1 つの例を示す．前章のシミュレーションでは図 4 のモデル 1 を真の分布とし，EM, 再帰的 EM の初期モデルはモデル 2 とした．これに対し，逆の状況を考える．モデル 2 を真の分布として，モデル 1 を初期モデルとして学習を始める．尤度の変化の様子を図 9 に示す．ここでは計算量は考えずに表示してある．

この結果から 2 次の近似をしてもほとんど収束の速さは変わらず，高次の近似をする意味が無い．計算量まで考えるのであれば，ほぼ 2 倍の計算量で得るものはほとんどないことが分る．このように，対象の分布とモデルの初期分布によって，用いる効果があるかどうかは変わってくる．

ここで紹介した再帰的な手法と同様に EM アルゴリズムの加速に関しては様々な方法が提案されている．そのほとんどは再帰的 EM アルゴリズムと同様にスコアリング法の近似を用いている．したがって，上でしめした

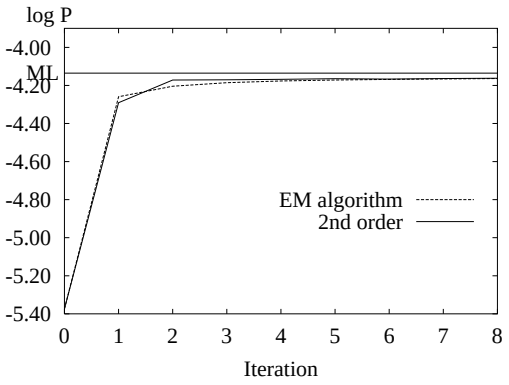


図 9: 真の分布と初期分布を変えた結果

問題はどの加速アルゴリズムも持っている問題である．

多くの加速アルゴリズムでは EM ステップを $\boldsymbol{\theta}_{t+1} = EM(\boldsymbol{\theta}_t)$ という関数であると定義し，このヤコビアン J と $(\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t)$ を用いて，スコアリング法の最急降下の方向を展開し，EM アルゴリズムの加速としている．この J は，ここでは ${G_X}^{-1}G_{Z|Y}$ と示したものと近似的に等しい．Aitken 加速では関数 $\boldsymbol{\theta}_{t+1} = EM(\boldsymbol{\theta}_t)$ から直接そのヤコビアンとして J を求める [8]．ヤコビアンを求めるのに必要な計算量は EM アルゴリズムの 1 ステップと同じ程度であるとすれば，2 次の近似を行うのに対し，用いる計算量は同じ程度である．ただし，このようにして求めた J の固有値は必ずしも 0 と 1 の間に存在しない．

Louis Turbo では [7] J の具体的な計算手順は与えられていない．これに対し Meng と Rubin は EM アルゴリズムを使って J を計算する方法を提案している [11]．彼らの方法では J を求めるために，EM ステップをパラメータの数だけ行う．一度 J を求めてしまえば，スコアリング法を何次までででも近似できるが，2 次の近似を求めるためにもパラメータ数分の EM ステップを行う必要がある．一方，本論文で提案した手法では，2 次の近似を求めるためには，EM アルゴリズムを 2 回行えば良く，高次の場合でも，それがパラメータ数以下ならば次数と同じ回数の EM を行えば良いだけである．それ以上の場合には単なる線形演算を行えば良く，Meng と Rubin の手法と比べ，パラメータ数よりも低い次数の近似を得たいのならば計算量は少く，大きい次数の近似には同じ計算量が必要となる．したがって Meng と Rubin の手法と比べても本手法の方が計算量が少なくてすむのである．

4 結び

本稿では EM アルゴリズムに関する 2 つの話題を提共した．Wake-Sleep アルゴリズムに関しては，KL ダイバージェンスの向きの違いから一般のモデルに対しては収束性が明らかでないことを示した．ただし，認識モデルが生成モデルに含まれるような場合，Sleep phase を十分長く取ることで必ず収束はする．この条件の元でも Wake-Sleep アルゴリズム が有用な場合があるのかを調べる必要があるだろう．

また再帰的 EM アルゴリズムを通じて EM アルゴリズムとスコアリング法を結びつけ，一種の加速法となることを示した．ただし，計算量も含めて常に EM アルゴリズムを加速するかは一般に明らかではない．今後の課題として，加速となるのかどうかを，モデル多様体とデータ多様体との関係に基づいて明らかにすることを考えている．

参考文献

[1] 甘利俊一, 長岡浩司. 情報幾何の方法. 岩波講座 応用数学 [対象 12]. 岩波書店, 1993.

[2] Shun-ichi Amari. Information geometry of the EM and em algorithm for neural networks. Technical report, Department of Mathematical Engineering and Information Physics, University of Tokyo, 1994. Technical report of Mathematical Engineering.

[3] Peter Dayan, Geoffrey E. Hinton, and Radford M. Neal. The Helmholtz machine. *Neural Computation*, Vol. 7, No. 5, pp. 889–904, 1995.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statistical Society, Series B*, Vol. 39, pp. 1–38, 1977.

[5] G. E. Hinton, P. Dayan, B. J. Frey, and R. M. Neal. The “wake-sleep” algorithm for unsupervised neural networks. *Science*, Vol. 268, pp. 1158–1160, 1995.

[6] Shiro Ikeda, Shun-ichi Amari, and Hiroyuki Nakahara. Convergence of the wake-sleep algorithm. to appear in Advances in Neural Information Processing Systems 11, 1999.

[7] Thomas A. Louis. Finding the observed information matrix when using the EM algorithm. *J. R.*

Statistical Society, Series B, Vol. 44, No. 2, pp. 226–233, 1982.

[8] Geoffrey J. McLachlan and Thriyambakam Krishnan. *The EM Algorithm and Extensions*. Wiley series in probability and statistics. John Wiley & Sons, Inc., 1997.

[9] Radford M. Neal and Peter Dayan. Factor analysis using delta-rule wake-sleep learning. *Neural Computation*, Vol. 9, No. 8, pp. 1781–1803, 1997.

[10] L.R. Rabiner, S.E. Levinson, and M.M. Sondhi. On the application of vector quantization and hidden Markov models to speaker-independent, isolated word recognition. *The Bell System Technical Journal*, Vol. 62, No. 4, pp. 1075–1105, April 1983.

[11] Martin A. Tanner. *Tools for Statistical Inference – Observed Data and Data Augmentation Methods*, Vol. 67 of *Lecture Notes in Statistics*. Springer-Verlag, 1991.

[12] D.M. Titterington. Recursive parameter estimation using incomplete data. *J. R. Statistical Society, Series B*, Vol. 46, No. 2, pp. 257–267, 1984.

[13] Lei Xu and Michael I. Jordan. On convergence properties of the EM algorithm for Gaussian mixture. A.I.Memo No.1520, C.B.C.L. Paper No.111, 1995.

A 定理 2 の証明

(6) 式より,

$$\begin{aligned}\boldsymbol{\theta}_{t+1}-\boldsymbol{\theta}_t &\simeq {G_X}^{-1}\partial L(Y^N;\boldsymbol{\theta}_t)\\ &= {G_X}^{-1}\partial\left(E_{\hat{q}(y)}\left[l(y;\boldsymbol{\theta})\right]\right)\Big|_{\theta=\theta_t}\end{aligned}$$

と書ける．$\hat{q}(y)$ を $p(y;\boldsymbol{\theta}_{t+1})$ と置き換え，(6) 式の導出と同様の手続きを行うと,

$$\begin{aligned}&\bar{\boldsymbol{\theta}}_{t+1}-\boldsymbol{\theta}_t \simeq {G_X}^{-1}\partial\left(E_{p(y;\theta_{t+1})}\left[l(y;\boldsymbol{\theta})\right]\right)\Big|_{\theta=\theta_t}\\ =\ & {G_X}^{-1}\int p(y;\boldsymbol{\theta}_{t+1})\partial l(y;\boldsymbol{\theta})\Big|_{\theta=\theta_t}d\mu(y).\end{aligned}\tag{14}$$

ここで $p(y;\boldsymbol{\theta}_{t+1})$ を次のように展開する,

$$\begin{aligned}p(y;\boldsymbol{\theta}_{t+1}) &\simeq p(y;\boldsymbol{\theta}_t)\\ &\quad +p(y;\boldsymbol{\theta}_t)\left(\partial l(y;\boldsymbol{\theta}_t)\right)^T(\boldsymbol{\theta}_{t+1}-\boldsymbol{\theta}_t)\end{aligned}$$

この結果を用いると (14) 式は次のように近似できる．

$$\begin{aligned}&\bar{\boldsymbol{\theta}}_{t+1}-\boldsymbol{\theta}_t\\ \simeq\ & {G_X}^{-1}\int\Big(p(y;\boldsymbol{\theta}_t)\partial l(y;\boldsymbol{\theta}_t)\\ &+p(y;\boldsymbol{\theta}_t)\partial l(y;\boldsymbol{\theta}_t)\partial l(y;\boldsymbol{\theta}_t)^T(\boldsymbol{\theta}_{t+1}-\boldsymbol{\theta}_t)\Big)d\mu(y)\\ =\ & {G_X}^{-1}\left(\int p(y;\boldsymbol{\theta}_t)\partial l(y;\boldsymbol{\theta}_t)\partial l(y;\boldsymbol{\theta}_t)^T d\mu(y)\right)\\ &\cdot(\boldsymbol{\theta}_{t+1}-\boldsymbol{\theta}_t)\\ =\ & {G_X}^{-1}G_Y(\boldsymbol{\theta}_{t+1}-\boldsymbol{\theta}_t)\\ \simeq\ & {G_X}^{-1}G_Y{G_X}^{-1}\partial L(Y^N;\boldsymbol{\theta}_t).\end{aligned}$$

ゆえに (11) 式を得る．ここでは次の結果を用いた,

$$\int p(y;\boldsymbol{\theta}_t)\partial l(y;\boldsymbol{\theta}_t)d\mu(y)=0.$$

また，連続の分布の場合，提案するアルゴリズムでは Monte Carlo 法を用いたが $\bar{\boldsymbol{\theta}}_{t+1}$ が Monte Carlo 法の影響で一点に定まらない．漸近的な $\bar{\boldsymbol{\theta}}_{t+1}$ の分布を示しておく．今 $p(y;\boldsymbol{\theta}_{t+1})$ にしたがって，サンプルを N' 個生成したとする $\{\bar{y}_1,\cdots,\bar{y}_{N'}\}$．$\hat{p}(y;\boldsymbol{\theta}_{t+1})$ を次のように定める．

$$\hat{p}(y;\boldsymbol{\theta}_{t+1})=\frac{1}{N'}\sum_{i=1}^{N'}\delta(y-\bar{y}_i)$$

また $\boldsymbol{\theta}^*_{t+1}$ を $\hat{p}(y;\boldsymbol{\theta}_{t+1})$ に対する最尤推定点とする．これらを用いて 2 次まで (14) 式を展開する．

$$\begin{aligned}&\int\hat{p}(y;\boldsymbol{\theta}_{t+1})\partial l(y;\boldsymbol{\theta})\Big|_{\theta=\theta_t}d\mu(y).\\ =\ & E_{\hat{p}(y;\theta_{t+1})}\left[\partial l(y;\boldsymbol{\theta}^*_{t+1})\right] &(15)\\ & -E_{\hat{p}(y;\theta_{t+1})}\left[\partial^2 l(y;\boldsymbol{\theta}^*_{t+1})\right](\boldsymbol{\theta}^*_{t+1}-\boldsymbol{\theta}_t) &(16)\end{aligned}$$

(15) 式は 0 であり $E_{\hat{p}(y;\theta_{t+1})}\left[\partial^2 l(y;\boldsymbol{\theta}^*_{t+1})\right]$ は漸近的に $-G_Y(\boldsymbol{\theta}_{t+1})$ と等しく $\boldsymbol{\theta}^*_{t+1}$ は $\boldsymbol{\theta}_{t+1}$ を中心に，分散行列が $G_Y(\boldsymbol{\theta}_{t+1})^{-1}/N'$ の正規分布に従う．したがって，$\bar{\boldsymbol{\theta}}_{t+1}$ の分散行列は，${G_X}^{-1}G_Y(\boldsymbol{\theta}_{t+1}){G_X}^{-1}/N'$ 程度である．