解説

EM アルゴリズムの幾何学

Geometry of the EM Algorithm

赤穂 昭太郎
Shotaro Akaho
電子技術総合研究所
情報科学部
情報数理研究室
Electrotechnical Laboratory
Information Science Division
Mathematical Informatics Section

非会員
連絡先: 〒305 茨城県 つくば市 梅園 1–1–4 電子技術総合研究所 情報科学部 情報数理研究室
電話: 0298–58–5549
FAX: 0298–58–5841
e-mail : akaho@etl.go.jp

1 はじめに

実世界から得られる複雑で不完全な情報を処理するために, 確率・統計的なモデル化が有効な手段として用いられている. ニューラルネットモデルや, 音声認識や画像処理において用いられているマルコフモデルなどはその代表的なものである. 従来, ニューラルネットの学習ではバックプロパゲーションに代表される最急降下法がよく用いられてきた. しかしながら, 一般に複雑な確率モデルに関する推論は, 困難な非線形最適化問題に帰着されがちであり, 学習に時間を要したり, 局所最適解にトラップされやすいという問題がある.

本稿では, 不完全データからの学習アルゴリズムである EM アルゴリズムの解説を行う. EM アルゴリズムは, 最急降下法と同様, 解を逐次改良していく繰り返し探索のアルゴリズムであるが, ある種の確率モデルの学習においては大域的最適解への良好な収束性や特に初期段階での速い収束性が知られている.

ここで, 不完全データというのは必ずしも欠測値を含むデータや部分データのことだけを指すのではなく, 便宜的に変数を付け足すことによって簡単な問題に帰着できる場合を含んでおり, 広い範囲の応用に用いることが出来る.

EM アルゴリズムは, 統計学では古くから知られている手法であり, Dempster らによって一般的に定式化された [6]. 音声認識でも隠れマルコフモデルの Baum-Welch アルゴリズム [5] として実用的に用いられている. 最近, 画像の隠れマルコフモデルに対する応用やニューラルネットモデルの一つであるボルツマンマシンの学習アルゴリズムとの関係 [3], および Jordan らによる “Hierarchical Mixtures of Experts (HME)”[7] と呼ばれるモデルの提案などがあり, 注目を浴びている. また, Amari は幾何学的な観点から EM アルゴリズムを特徴づけた [2].

以下ではまず EM アルゴリズムとはどんなものかについて説明し, その幾何学的な意味を述べ, なぜアルゴリズムがうまくいくかについての直観的イメージを与える. さらに, HME モデルおよび隠れマルコフモデルの解説を行う.

2 EM アルゴリズムとは

2.1 不完全データの最尤推定

EM アルゴリズムでは, 不完全性をもたない**完全データ** $\boldsymbol{x} \in X$ というものが背後にあって, 観測されるデータ $\boldsymbol{y}$ はその完全データの**不完全データ**であると考える. つまり, $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x})$ なる, 既知の多対 1 写像が存在しているとする. 実際の応用では, べ

クトル $\boldsymbol{x}$ が $\boldsymbol{x} = (\boldsymbol{y}, \boldsymbol{z})$ のように分離できて, $\boldsymbol{y}$ だけが観測され, $\boldsymbol{z}$ が隠れた変数となっている場合が多い. 完全データは実際に想定されるデータであることもあるし, 推定を容易にするために仮想的に導入される場合もある.

ここで, パラメータ $\boldsymbol{\xi}$ をもつ完全データ $\boldsymbol{x}$ の確率分布 $f(\boldsymbol{x} \mid \boldsymbol{\xi})$ を考えよう. $f(\boldsymbol{x} \mid \boldsymbol{\xi})$ に対応して $\boldsymbol{y}$ の確率分布 $g(\boldsymbol{y} \mid \boldsymbol{\xi})$ が次のように与えられる.

$$g(\boldsymbol{y} \mid \boldsymbol{\xi}) = \int_{X(\boldsymbol{y})} f(\boldsymbol{x} \mid \boldsymbol{\xi}) \, \mathrm{d}\boldsymbol{x}. \tag{1}$$

ただし $X(\boldsymbol{y}) \subset X$ は $\boldsymbol{y}$ の逆写像である.

観測データ $\boldsymbol{y}$ が与えられた時に $g(\boldsymbol{y} \mid \boldsymbol{\xi})$ を $\boldsymbol{\xi}$ のもっともらしさを表す関数 (尤度) として考え, その最大値をとる $\boldsymbol{\xi}$ を求める方法が**最尤推定法**であり, データ $\boldsymbol{y}$ と確率モデル $g(\boldsymbol{y} \mid \boldsymbol{\xi})$ が与えられた時に $\boldsymbol{\xi}$ を推論するために用いられる典型的な方法である. また, 尤度 g のかわりに対数尤度 $\log g$ を最大化しても同じであるのでそちらを用いることもある.

ここで, 不完全データ $\boldsymbol{y}$ の最尤推定を直接行うことは難しいが, 完全データ $\boldsymbol{x}$ がわかっていれば最尤推定が簡単になる場合を考えよう. EM アルゴリズムは完全データの最尤推定を間接的に利用して不完全データの最尤推定を行おうとするものである. 次節で EM アルゴリズムを一般的な形で定式化する.

例 1 正規混合分布モデル

K 個の正規分布 $N(\mu_k, \sigma_k^2), (k = 1, \ldots, K)$ があるとし, サンプル y はこのうちの一つから出て来るものとする. k 番目の正規分布が選ばれる確率は π_k であるとする. 問題はサンプルが与えられたときのパラメータ $\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}$ の推定である. y の分布は次のように書ける.

$$g(y \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}) = \sum_{k=1}^{K} \pi_k \phi(y \mid \mu_k, \sigma_k), \qquad \sum_{k=1}^{K} \pi_k = 1. \tag{2}$$

ここで ϕ は正規分布の密度関数

$$\phi(y \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x-\mu)^2}{2\sigma^2}) \tag{3}$$

である. もしサンプルが何番目の正規分布から生成されていたかがすべてわかっていれば問題は自明となる. EM アルゴリズムを適用するために, その番号 z を含めたものを完全データとし, y を不完全データとみなす. 完全データ (y, z) の分布は次のように書ける.

$$f(y, z \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}) = \pi_z \phi(y \mid \mu_z, \sigma_z). \tag{4}$$

また, 独立にこの分布に従う N 個の完全データ $(y_1, z_1), (y_2, z_2), \ldots, (y_N, z_N)$ が与えられたときの対数尤度は

$$\sum_{i=1}^{N} \log f(y_i, z_i \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}) = \sum_{i=1}^{N} \log(\pi_{z_i} \phi(y_i \mid \mu_{z_i}, \sigma_{z_i})) \tag{5}$$

となる.

次節の例 2で正規混合分布モデルに対する EM アルゴリズムを書き下す. Jordan らの HME モデルはある意味でこのモデルを発展させたものになっている.

2.2 E ステップと M ステップ

EM アルゴリズムはパラメータをある適当な初期値に設定し, **E ステップ** (Expectation step) と **M ステップ** (Maximization step) と呼ばれる二つの手続きを繰り返すことにより $\boldsymbol{\xi}$ の値を逐次更新する方法であり, 次のように定式化される.

1. パラメータの初期値を適当な点 $\boldsymbol{\xi} = \boldsymbol{\xi}^{(0)}$ にとる.
2. $p = 0, 1, 2, \ldots$ に対して次の二つのステップを繰り返す.
   (a) E ステップ: 完全データの対数尤度 $\log f(\boldsymbol{x} \mid \boldsymbol{\xi})$ の, データ $\boldsymbol{y}$ とパラメータ $\boldsymbol{\xi}^{(p)}$ に関する条件つき平均を求める. つまり,

$$\begin{aligned} Q(\boldsymbol{\xi}) &= \mathrm{E}[\log f(\boldsymbol{x} \mid \boldsymbol{\xi}) \mid \boldsymbol{y}, \boldsymbol{\xi}^{(p)}] \\ &= \int f(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\xi}^{(p)}) \log f(\boldsymbol{x} \mid \boldsymbol{\xi}) \, \mathrm{d}\boldsymbol{x}. \end{aligned} \tag{6}$$

   を計算する.
   (b) M ステップ: $Q(\boldsymbol{\xi})$ を最大化する $\boldsymbol{\xi}$ を $\boldsymbol{\xi}^{(p+1)}$ とおく.

なお, 不完全データ $\boldsymbol{y}$ が与えられたときの完全データ $\boldsymbol{x}$ の条件つき分布は, Bayes の公式から

$$f(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\xi}) = \begin{cases} f(\boldsymbol{x} \mid \boldsymbol{\xi}) / g(\boldsymbol{y} \mid \boldsymbol{\xi}), & \boldsymbol{x} \in X(\boldsymbol{y}) \\ 0 & \boldsymbol{x} \notin X(\boldsymbol{y}) \end{cases} \tag{7}$$

で与えられる.

例 2 正規混合分布モデルの場合 [14]

例 1 で与えた正規混合分布モデルの EM アルゴリズムを定義通りに計算する. 観測データ $y_1, y_2, \ldots, y_N$ が与えられたとき, まず E ステップで計算すべき Q は,

$$Q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \frac{\pi_k^{(p)} \phi(y_i \mid \mu_k^{(p)}, \sigma_k^{(p)})}{g(y_i \mid \boldsymbol{\pi}^{(p)}, \boldsymbol{\mu}^{(p)}, \boldsymbol{\sigma}^{(p)})} \log(\pi_k \phi(y_i \mid \mu_k, \sigma_k)) \tag{8}$$

となる (z_i は離散値しかとらないので積分は和になる).

M ステップでは $Q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma})$ の最大化問題を解くことになる. 基本的にはパラメータに関して微分した関数のゼロ点を求めればよいが, $\boldsymbol{\pi}$ に関しては $\sum_{k=1}^{K} \pi_k = 1$ という条件があるので, Lagrange の未定係数法を用いることになる. 式を簡単にするために,

$$\hat{\phi}_k^{(p)}(y) \equiv \frac{\phi(y \mid \mu_k^{(p)}, \sigma_k^{(p)})}{g(y \mid \boldsymbol{\pi}^{(p)}, \boldsymbol{\mu}^{(p)}, \boldsymbol{\sigma}^{(p)})}, \tag{9}$$

$$\Phi_k^{(p)} \equiv \sum_{i=1}^{N} \hat{\phi}_k^{(p)}(y_i) \tag{10}$$

とおくと, 最終的に

$$\pi_k^{(p+1)} = \frac{\pi_k^{(p)} \Phi_k^{(p)}}{\sum_{j=1}^{K} \pi_j^{(p)} \Phi_j^{(p)}} \tag{11}$$

が得られる. $\boldsymbol{\mu}, \boldsymbol{\sigma}$ に関しても同様に

$$\mu_k^{(p+1)} = \frac{1}{\Phi_k^{(p)}} \sum_{i=1}^{N} \hat{\phi}_k^{(p)}(y_i)\, y_i \tag{12}$$

$$(\mu_k^{(p+1)})^2 + (\sigma_k^{(p+1)})^2 = \frac{1}{\Phi_k^{(p)}} \sum_{i=1}^{N} \hat{\phi}_k^{(p)}(y_i)\, y_i^2 \tag{13}$$

という結果を得る. アルゴリズムを導くために E ステップと M ステップを計算したが, 最終的には (11), (12), (13) という 1 ステップの式に帰着できる.

2.3 EM アルゴリズムの特徴

EM アルゴリズムの各々の繰り返しによって, 尤度が単調に増加することが証明されている. 従って, 局所的には最適解に収束し, 少なくとも初期解よりはよい解が得られる. もちろん一般に大域的に収束する保証はないが, 多くの応用例で良好な大域的収束性が経験的に知られている.

ただし, 最初のうちは速い収束を示すが, 収束の後期では遅くなると言われており, E ステップや M ステップが必ずしも容易に実行できないという問題も存在する. これらの EM アルゴリズムの問題点および対処法については 4.1節で述べる.

3 EM アルゴリズムの幾何学的意味

3.1 確率分布の空間

パラメータ $\boldsymbol{\theta}$ をもつ確率分布の集合 $S = \{f(\boldsymbol{x} \mid \boldsymbol{\theta})\}$ は $\boldsymbol{\theta}$ を局所座標系とする空間 (多様体) とみなすことができる. この空間に確率分布特有の微分幾何学的構造を導入して, 確率分布に関わる問題を幾何学的に考察するのが情報幾何学である. そこでは互いに双対な構造が自然にあらわれ, 双対座標系や双対射影といったものが重要な役割を果たしている [1, 3, 4].

本節では情報幾何学における基礎的な概念をまじえながら, EM アルゴリズムを幾何学的に解釈し, なぜ EM アルゴリズムがうまく働くかについて直観的イメージを与える.

3.2 統計的推定の幾何学的なイメージ

ここではまず, 与えられたデータに対して確率モデルをあてはめるという一般的な問題に対して幾何学的な意味づけを与える.

まず, 基本となる空間 S としては, あてはめるべき**モデルの空間** M を含む空間を考える (M は S 自身ということもある). 特に, S として**指数型分布族**と呼ばれる確率分布の空間を考えると, そこには自然に e- 座標系 $\boldsymbol{\theta}$ と m- 座標系 $\boldsymbol{\eta}$ という**双対的な座標系**が導入される (3.4節参照). 以下では S として指数型分布族だけを考える.

観測によって得られたデータは $\boldsymbol{\eta}$ 座標の 1 点として表すことができて, 統計的推論はその点からモデルの空間 M への (直交) **射影**として得られる (3.5節参照). 射影は, 双対座標系に付随して e- 射影と m- 射影という二つの双対的な射影を考えることができる (3.6節参照). 特に m- 射影は最尤推定と等価になっている.

以上の操作が不完全データを含まない場合の統計的推論の幾何学的イメージである (図 1). 次の節では不完全データを含む場合 (EM アルゴリズム) の幾何学的なイメージを与える.

図 1:

3.3 EM アルゴリズムの幾何学的イメージ

EM アルゴリズムのように不完全データしか得られない場合には観測データを S の 1 点として表すことができない. 一般にデータが不完全な分だけ自由度をもつので, 1 つの点ではなく S の部分空間 D となる. D を**データ多様体**と呼ぶことにする.

ここで次のアルゴリズムを考える.

1. モデル M の中の点 $\boldsymbol{\theta}^{(0)}$ を適当な初期値としてとる.

2. $p = 0, 1, 2, \ldots$ に対して次の二つのステップを繰り返す.

   (a) e- ステップ: $\boldsymbol{\theta}^{(p)}$ から D への e- 射影 (の $\boldsymbol{\eta}$ 座標) を $\boldsymbol{\eta}^{(p)}$ とする.

   (b) m- ステップ: $\boldsymbol{\eta}^{(p)}$ から M への m- 射影 (の $\boldsymbol{\theta}$ 座標) を $\boldsymbol{\theta}^{(p+1)}$ とする.

実はこのアルゴリズムが, 比較的ゆるやかな条件のもとで EM アルゴリズムに一致することが示されている. 従って, EM アルゴリズムは幾何学的には二つの空間 D, M の間の射影の繰り返しとして解釈することができる (図 2).

図 2:

以下の節ではここで用いられた指数型分布族の双対的な構造, モデル多様体やデータ多様体, 射影などについてさらに説明し, 上記のアルゴリズムが幾何学的考察から自然に導かれることを示す.

ちなみに, e- 射影と m- 射影を用いたアルゴリズムは小文字を用いて em アルゴリズムと呼ばれるが, EM アルゴリズムと似たような名前になっているのは偶然の一致である. EM が Expectation and Maximization の頭文字であるのに対し, em は Exponential and Mixture の頭文字である.

3.4 指数型分布族と双対座標系

指数型分布族というのは

$$f(\boldsymbol{x} \mid \boldsymbol{\theta}) = \exp\{\sum_i \theta_i F_i(\boldsymbol{x}) - \psi(\boldsymbol{\theta})\} \tag{14}$$

という形に書ける分布族であって, 多くの基本的分布が属している.

$\boldsymbol{\theta}$ が e- 座標系になっており, これに双対な m- 座標系は期待値パラメータと呼ばれ,

$$\eta_i = \mathrm{E}[F_i(\boldsymbol{x}) \mid \boldsymbol{\theta}] \tag{15}$$

で与えられる. S は $\boldsymbol{\theta}$ と $\boldsymbol{\eta}$ のそれぞれの座標系に関して微分幾何学的に "**平坦な**" 空間になっており性質がよい空間である. $\boldsymbol{\theta}$ と $\boldsymbol{\eta}$ は互いに次の変換により座標変換される (**ルジャンドル変換**).

$$\frac{\partial}{\partial \theta_i}\psi(\boldsymbol{\theta}) = \eta_i, \quad \frac{\partial}{\partial \eta_i}\varphi(\boldsymbol{\eta}) = \theta_i; \quad \varphi(\boldsymbol{\eta}) \equiv \sum_i \theta_i \eta_i - \psi(\boldsymbol{\theta}) \tag{16}$$

例 3 正規分布

正規分布の密度関数 $\phi(y \mid \mu, \sigma)$ は次のように書き表せる.

$$\log \phi(y \mid \mu, \sigma) = \frac{\mu}{\sigma^2}x + \frac{-1}{2\sigma^2}x^2 - \{\frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2)\}. \tag{17}$$

ここで,

$$\theta_1 = \frac{\mu}{\sigma^2}, \qquad \theta_2 = \frac{-1}{2\sigma^2}, \tag{18}$$

$$F_1(x) = x, \qquad F_2(x) = x^2, \tag{19}$$

などとおけばこれは指数型分布族の形をしている. また双対座標は, $\eta_1 = \mu$, $\eta_2 = \mu^2 + \sigma^2$ となる.

図 3 に正規分布の双対座標系を示す. (a) はパラメータ μ, σ^2 をそのまま座標系としてとったもので, その上の格子が双対座標系ではそれぞれ (b), (c) のように変換される. 人間にとって自然な μ-σ^2 座標系は, 幾何学的に自然な $\boldsymbol{\theta}$ 座標系や $\boldsymbol{\eta}$ 座標系からみると非常に曲がった座標系であることがわかる.

図 3:

正規混合モデルの完全データの分布 (4) に対しても同様な変形を行えば指数型分布族であることがわかる.

3.5 データ点とモデルの空間

3.2 節で, 統計的推論は S の上に置かれるデータ点からモデルの空間 M への射影として解釈されると述べた. M は S の部分空間として陽に与える場合もあるが, 特に M を定めなくても, 複数サンプルの分布を考えると自然にモデル M が導入される. 本節ではデータ点が空間のどこに対応づけられるか, および複数サンプルから導かれるモデルの空間の例を示す.

まず, S の中の (未知の) 分布 $\boldsymbol{\theta}_0$ に従って 1 つのサンプル $\boldsymbol{x}_0$ が観測されたとして, これを S の中の点に対応づけることを考える. 最尤推定は (14) 式の $\boldsymbol{x}$ に $\boldsymbol{x}_0$ を入れた関数を最大にするパラメータである. $\log f(\boldsymbol{x}_0 \mid \boldsymbol{\theta})$ を θ_i に関して微分して 0 とおくと,

$$F_i(\boldsymbol{x}_0) - \eta_i = 0,$$

という式が得られる. ここで ψ の微分には (16) 式を用いた. 以上のことから, 観測データは m- 座標系の点

$$\hat{\eta}_i = F_i(\boldsymbol{x}_0) \tag{20}$$

に対応づけられることがわかる. ここでさらに $\hat{\boldsymbol{\eta}}$ を m- 射影によりモデル M に射影すれば M に対する最尤推定が得られることになる (M が S 自身の場合は $\hat{\boldsymbol{\eta}}$ 自身が推定値となる).

不完全データ $\boldsymbol{y}$ しか与えられない場合は $\hat{\boldsymbol{\eta}}$ が一意的には定まらないが, 次のように, $\boldsymbol{y}$ が与えられたときの $\hat{\boldsymbol{\eta}}$ のとり得る値全体の集合

$$D = \{\hat{\boldsymbol{\eta}} \mid \hat{\eta}_i = F_i(\boldsymbol{x}), \boldsymbol{x} \in X(\boldsymbol{y})\} \tag{21}$$

を考えることができる. これが 3.3節で述べたデータ多様体である.

次に, 指数型分布族の空間 S^* をとり, その上の分布 $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_N$ に従う独立な N 回の観測 $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$ を考えよう. すると観測全体の分布の空間 S は直積空間

$$S = S^* \times S^* \times \cdots \times S^* \tag{22}$$

となる. このとき観測データは S の点 $(\hat{\boldsymbol{\eta}}_1, \hat{\boldsymbol{\eta}}_2, \ldots, \hat{\boldsymbol{\eta}}_N)$ として S の m- 座標を用いて表すことができる. データ多様体も (21) 式と同様に定義できる. 以下の例では二つの典型的な場合について直積空間 S に自然に導入されるモデルの空間 M について述べる.

例 4 独立同分布のモデル

サンプルが独立で, $\boldsymbol{\theta}_i$ がすべて共通のモデル, つまり M が $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 = \cdots = \boldsymbol{\theta}_N \equiv \boldsymbol{\theta}$ を満たす S の部分多様体となっている場合である.

この場合, 完全データが観測できる場合には直積空間 S を考える必要はなく, もとの空間 S^* の問題に帰着される. すなわち, 観測データは S^* の点 $(1/N)\sum_{i=1}^{N} \hat{\boldsymbol{\eta}}_i$ に対応づけられる ((20) 式と同様に導かれる). ただし我々が扱う不完全データの推論の場合, データ多様体は一般に S^* の部分空間としては表現できず, 直積空間のまま扱う必要がある.

この例にあてはまる例としては, 例 1,2 で述べた正規混合モデル, ボルツマンマシンのモデル, 隠れマルコフモデルのモデルなどがある.

例 5 条件つき分布の学習

ニューラルネットの学習に代表されるようなシステムの入出力関係の学習の場合には入力 $\boldsymbol{u}$ に対する出力 $\boldsymbol{v}$ の条件つき分布 $f(\boldsymbol{v} \mid \boldsymbol{u}, \boldsymbol{\xi})$ が学習の対象となる ($\boldsymbol{\xi}$ はシステムのパラメータ). この分布の空間を S^* とすると, $\boldsymbol{u}$ と $\boldsymbol{\xi}$ がパラメータを定め S^* の e- 座標を用いて $\boldsymbol{\theta}(\boldsymbol{u}, \boldsymbol{\xi})$ のように書ける. この場合は $\boldsymbol{\xi}$ は共通だが, $\boldsymbol{u}$ のほうは入力の変化に応じて変化するので, N 個の入出力のペアごとにすべて異なってお

り, 直積空間上で $(\boldsymbol{\theta}(\boldsymbol{u}_1,\boldsymbol{\xi}),\boldsymbol{\theta}(\boldsymbol{u}_2,\boldsymbol{\xi}),\ldots,\boldsymbol{\theta}(\boldsymbol{u}_N,\boldsymbol{\xi}))$ という空間を張る. この空間が条件つき分布の学習モデルにおけるモデル多様体となる. この空間は入力に応じて変化する空間であり, 一般に前の例のものよりも複雑なものとなる.

これにあてはまる例としては, 確率的多層パーセプトロンモデルや, 後で述べる Jordan らの HME モデルなどが挙げられる.

3.6 射影とダイバージェンス

e- 射影と m- 射影はそれぞれ, e- 座標系と m- 座標系における部分多様体への垂線の足として定義される. ここで二つの確率分布の間の隔たりを調べるために Kullback-Leibler の**ダイバージェンス**を導入する.

$$K(p\|q) \equiv \int p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\, \mathrm{d}\boldsymbol{x} \tag{23}$$

ここで $K(p\|q)$ は e- 座標系と関係が深いので e- ダイバージェンスと呼び, それと双対的関係にある $K(q\|p)$ を m- ダイバージェンスと呼ぶことにする. ダイバージェンスは距離の公理は満たさないが統計的に重要な量である.

射影とダイバージェンスとの間に次の定理が成り立つ (図 4).

図 4:

定理 1 (射影定理 [1]) 指数型分布族の空間 S の点 p および S の m-(e-) 座標系に関して平坦な部分多様体 M が与えられたとする. このとき, M の点 q が e-(m-) ダイバージェンスの最小値をとる点となるための必要十分条件は, q が p から M への e-(m-) 射影であることである.

M が平坦でない場合も e-(m-) 射影は e-(m-) ダイバージェンスの停留点 (微係数が 0 の点) になる.

この定理から, e- 射影および m- 射影がダイバージェンスの極小値を与えるという意味で自然なものであることがわかる (特に m- 射影は最尤推定になっている). また, 幾何学的には射影する部分多様体は平坦な空間であることが望ましい. EM アルゴリズムではほとんどの場合, データ多様体 D は m- 座標系に関して平坦な空間になっている. また, モデルの空間としても e- 座標系に関して (できるだけ) 平坦な空間をとることが多い. 従って D に対して e- 射影を行い, M に対して m- 射影を行うという em アルゴリズムは上記定理から自然に導かれる.

EM アルゴリズムを設計する際にも, 完全データの空間におけるモデルができるだけ平坦な空間 (M 自身指数型分布族であるなど) になるようにすることが必要になると考えられる.

4 モデルの実例と問題点

EM アルゴリズムは最近になって画像処理やニューラルネットの学習などといった分野への応用に使われ始めるようになってきた. 以下ではその具体例として Jordan らによって提案された HME モデルと, 音声や画像に対する隠れマルコフモデルの学習アルゴリズムへの応用について簡単に紹介し, EM アルゴリズムを実際問題に適用する場合の問題点について述べる. 本稿では省略するが, ニューラルネットの代表的なモデルである多層パーセプトロンも中間層の出力を隠れた変数として EM アルゴリズムを適用することができる.

4.1 実際への適用への問題点

例 2 において正規混合モデルに対する EM アルゴリズムが簡単な形で書けることを示したが, 実際に応用しようとするといろいろな問題が生じる.

まず, M ステップにおいては前節で見たように一般にモデルの空間が曲がった空間であるほど射影を行うのが困難になる. 条件つき確率の形の学習は曲がった空間となることが多い. そこで関数 Q を最大化するのはあきらめて, 少なくとも現在の Q よりは大きな値になるアルゴリズムを用いることが考えられる. そのようなアルゴリズムは GEM (一般化 EM) アルゴリズムと呼ばれるが, そのための技法としては従来から用いられて来た最急降下法や (擬似) Newton 法などが挙げられる.

次に, E ステップでは条件つき期待値を計算するので, 一般には数値積分 (離散データの場合は総和計算) を行う必要があるが, 問題サイズによっては大きな計算量となることがある. これを避けるための普通の方法はモンテカルロシミュレーションによって推定された積分値で置き換える方法である [18].

また, EM アルゴリズムの欠点として挙げられる問題の一つに収束の遅さがある. EM アルゴリズムは一般的に, 繰り返し初期の収束は Newton 法と同程度の速さで収束し, 尤度が単調に増加する保証があるので Newton 法より安全なアルゴリズムと言える. しかしながら, ある程度収束するとそれ以降収束速度が遅くなる傾向がある. 低い精度でよければ十分実用となるが, 高い精度が必要な場合にはなんらかの加速法を用いる必要がある [9, 18].

4.2 HME モデル

ここでは Jordan らによって提案された HME (Hierarchical Mixtures of Experts: 階層型エキスパート混合) モデル [7, 8] の簡単な紹介を行う.

このモデルは例 1, 2 で述べた正規混合分布を発展させたものであるのでそれから出発しよう. その密度関数 (2) を一般化して書くと

$$g(\boldsymbol{y} \mid \boldsymbol{\pi}, \boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_K) = \sum_{k=1}^{K} \pi_k p_k(\boldsymbol{y} \mid \boldsymbol{\xi}_k), \qquad \sum_{k=1}^{K} \pi_k = 1 \tag{24}$$

という関数形をしている. ここで, p_k は正規分布に限らず任意の分布族に一般化する. また, HME モデルでは例 5 で扱った条件つき分布の学習を行い, p_k は入力 $\boldsymbol{u}$ に対する出力 $\boldsymbol{v}$ の条件つき分布の形で書く. すると,

$$g(\boldsymbol{v} \mid \boldsymbol{u}, \boldsymbol{\pi}, \boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_K) = \sum_{k=1}^{K} \pi_k p_k(\boldsymbol{v} \mid \boldsymbol{u}, \boldsymbol{\xi}_k), \qquad \sum_{k=1}^{K} \pi_k = 1 \tag{25}$$

となる. Jordan らは p_k として一般化線形モデルをとっている.

各 p_k はエキスパートと呼ばれ, それぞれ入力 $\boldsymbol{u}$ に対して処理を施して $\boldsymbol{v}$ を出力するが, どのエキスパートの出力を選ぶかを確率 π_k で制御している. エキスパートの出力を選ぶ $\boldsymbol{\pi}$ にあたるネットワークはゲートネットワークと呼ばれる.

HME モデルではこの部分も一般化し, π_k は定数ではなく, $\boldsymbol{u}$ のパラメトリックな関数であるとする. 具体的には一般化線形モデルの一つである次の関数族をとる.

$$\pi_k(\boldsymbol{u}) = \frac{\exp(\boldsymbol{\theta}_k^T \boldsymbol{u})}{\sum_{j=1}^{N} \exp(\boldsymbol{\theta}_j^T \boldsymbol{u})}. \tag{26}$$

以上によって図5のようなネットワークが定まるが, これをさらに図6のようにゲートネットワークを多層化したものが HME モデルである.

HME モデルでは M ステップは最急降下法あるいは (擬似) Newton 法によって行われる. Jordan らは効率のよい逐次アルゴリズムを提案しており, バックプロパゲーションなどよりも非常に速く良質の解が得られることが報告されている.

図 5:

図 6:

4.3 隠れマルコフモデル

隠れマルコフモデルは音声認識への応用で比較的古くから EM アルゴリズムが用いられてきた. ここでは画像処理への応用が可能なマルコフランダム場の上の隠れマルコフモデルに一般化して説明する.

図 7 のように, 点集合とその近傍系が与えられているとしよう. **マルコフランダム場**は, 各点 i はある確率にしたがって**状態** z_i をとるが, その確率は近傍の点の状態に対する条件つき確率

$$a(z_i \mid z_j, j \in N_i) \tag{27}$$

で表されているとするモデルである.

図 7:

図 7(a) は音声のような時系列に対するモデルを表しており, 各点が時刻に対応し, ある時刻に対する状態が一つ前の状態の条件つき確率で与えられるとするモデル (マルコフ連鎖) である. また, 図 7(b) は画像のように 2 次元空間の信号に対するモデルを表しており, 各点の状態はそれに隣接する点の状態の条件つき分布として表される. 点集合がある状態をとる確率は Gibbs 分布として知られる分布で表現できる.

ここで, 各状態 z に対してある確率分布 $b(y|z)$ で記号 (または信号) y が出力されるとしよう. **隠れマルコフモデル**は, マルコフランダム場において状態 z は観測できず, 出力記号 y (の系列) だけが観測されるというモデルである. EM アルゴリズムは z を隠れた変数とし, 状態の遷移確率 $a(z_i \mid z_j, j \in N_i)$ および出力記号の確率 $b(y \mid z)$ を学習するために用いられる (N_i は点 i の近傍).

このモデルでは E ステップにおいて多数の確率の積和計算を行う必要がある. ただし, 時系列など 1 次元のマルコフランダム場に対しては, 途中の計算結果を再利用することによってかなり計算量を節約でき, Baum-Welch アルゴリズムとして知られる高速なアルゴリズムがある [5, 12].

一般に次元や近傍の大きさが増えれば増える程計算時間は爆発的に増大する. このため, 通常はモンテカルロシミュレーションによって近似することが多い. Zhang [19] は統計力学でよく用いられる平均場近似を用いて計算量を減らす手法を提案した. 時系列データの識別や画像データのセグメンテーションといった例題に対して良好な結果を示している.

5 おわりに

EM アルゴリズムの幾何学的な解釈を与えるとともに, 主に工学的応用の見地からモデルを紹介した. EM アルゴリズムは従来の最急降下法などと比べてより安全に局所最適解に収束し, 多くの場合に大域的最適解に近いところに収束するという点で新たなアルゴリズム設計の指針となると思われる.

実際の応用例も徐々に増えつつあるが, これから更にいろいろなモデルの提案や新たな応用分野も増えていくことが期待される.

また, 理論的にも情報幾何学的な枠組に基づいて評価や解析がなされるであろう. この方面の最近の研究では Amari[2] のほか, 不完全データを含むモデル選択に関する評価基準の提案 [16] やインクリメンタルなアルゴリズムの理論づけ [13] などが注目される.

EM アルゴリズムは最近再び見直されつつある Bayes 決定法とも関係が深い. Bayes 決定法では変量だけではなくパラメータに対しても確率分布を考えるが, 分布をもつパラメータは観測されない変数とみなすこともでき, EM アルゴリズムを適用することができる (潜在変数モデル [15, 18]).

本稿をまとめるにあたり Amari[2] を大いに参考にした. EM と em が異なる場合などの詳細な議論に関してはそちらを参照されたい. なお, 本稿では, Riemann 計量やアファイン接続といった情報幾何学で重要な役割を果たす概念に関する解説を行わなかった. これらを使った幾何学的考察により確率分布に特有な興味深い構造がいろいろ見出されている. 情報幾何学に関しては [1, 4, 11] などを参照のこと.

EM アルゴリズムはもともと統計学の分野で発展してきたものである [17]. 著者は統計学の専門家ではないので, 統計学における発展の歴史や経緯, 現状などについては触れられなかった. それらに関しては [10] を参考にされたい.

謝辞

本稿を書くきっかけを与えて頂いた栗田多喜夫氏に感謝いたします. また, 原稿の不備を指摘して頂いた査読者の方々にも感謝いたします. 本解説は「リアルワールドコンピューティングプログラム (RWCP)」の一環として行われたものです. 研究の機会を与えて下さった諏訪基情報科学部長, 大津展之新情報計画室長に感謝いたします.

参考文献

[1] Amari, S.: *Differential Geometrical Methods in Statistics.* Lecture Notes in Statistics. Springer-Verlag, 1985.

[2] Amari, S.: Information geometry of the EM and em algorithms for neural networks. Technical Report METR 94–04, University of Tokyo, 1994. (to appear in Neural Networks).

[3] Amari, S., Kurata, K., and Nagaoka, H.: Information geometry of Boltzmann machines. *IEEE Trans. Neural Networks*, Vol. 3, No. 2,, 1992.

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

[5] Baum, L., Petrie, T., Soules, G., and Weiss, N.: A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, Vol. 41, No. 1, pp. 164–171, 1970.

[6] Dempster, A., Laird, N., and Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc.* B, Vol. 39, pp. 1–38, 1977.

[7] Jordan, M. I. and Jacobs, R. A.: Hierarchical mixtures of experts and the EM algorithm. In *Proc. of IJCNN'93*, pp. 1339–1344, Nagoya, 1993.

[8] Jordan, M. I. and Xu, L.: Convergence results for the EM approach to mixtures of experts architectures. MIT A.I. Memo No. 1458, 1993.

[9] Louis, T.: Finding observed informatin using the EM algorithm. *Journal of the Royal Statistical Society* B, Vol. 44, pp. 98–130, 1982.

[10] 宮川雅巳: EM アルゴリズムとその周辺. 応用統計学, Vol. 16, No. 1, pp. 1–19, 1987.

[11] Murray, M. K. and Rice, J. W.: *Differential Geometry and Statistics.* Chapman & Hall, 1993.

[12] 中川聖一: 確率モデルによる音声認識. 電子情報通信学会, 1988.

[13] Neal, R. M. and Hinton, G. E.: A new view of the EM algorithm that justifies incremental and other variants. submitted to Biometrica, 1993.

[14] Redner, R. and Walker, H.: Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, Vol. 26, pp. 195–239, 1984.

[15] 島内 他 (編): アルゴリズム辞典. 共立出版, 1994.

[16] Shimodaira, H.: A new criterion for selecting models from partially observed data. In P.Cheeseman and R.W.Oldford (eds.), *Selecting Models from Data: AI and Statistics IV*, chapter 3, pp. 21–29. Springer-Verlag, 1994.

[17] 竹内啓 (編): 統計学辞典. 東洋経済新報社, 1989.

[18] Tanner, M. A.: *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions.* Springer-Verlag, 1993.

[19] Zhang, J.: The mean field theory in EM procedures for Markov random fields. *IEEE Trans. on Signal Processing*, Vol. 40, No. 10, pp. 2570–2583, 1992.