多様体

Amari, Shun-ichi. "Information geometry of the EM and em algorithms for neural networks." Neural networks 8.9 (1995): 1379-1408.

交通・都市・国土学研究室
月田 光

概要

前提

・ノイズの入った教師データから入出力のモデル化を行う際に，ニューラルネットワーク(NN)が用いられる.
・データが観測不能な値や欠損値(隠れ変数)を含む場合，隠れ変数を推定する.

→隠れ変数の推定に，EMアルゴリズムとemアルゴリズムが提案されている

EMアルゴリズム：条件付き期待値を用いた反復的な統計手法
emアルゴリズム：多様体においてKLダイバージェンスを反復的に最小化

目的

・EMアルゴリズムとemアルゴリズムがほとんどの場合等価であることを示す.
・EMとemに統一的な情報幾何学的枠組みを与え，その等価性を保証する条件を証明.
・例として以下を考える.
(1) 確率的多層パーセプトロン，(2)混合エキスパートモデル，(3)混合正規分布

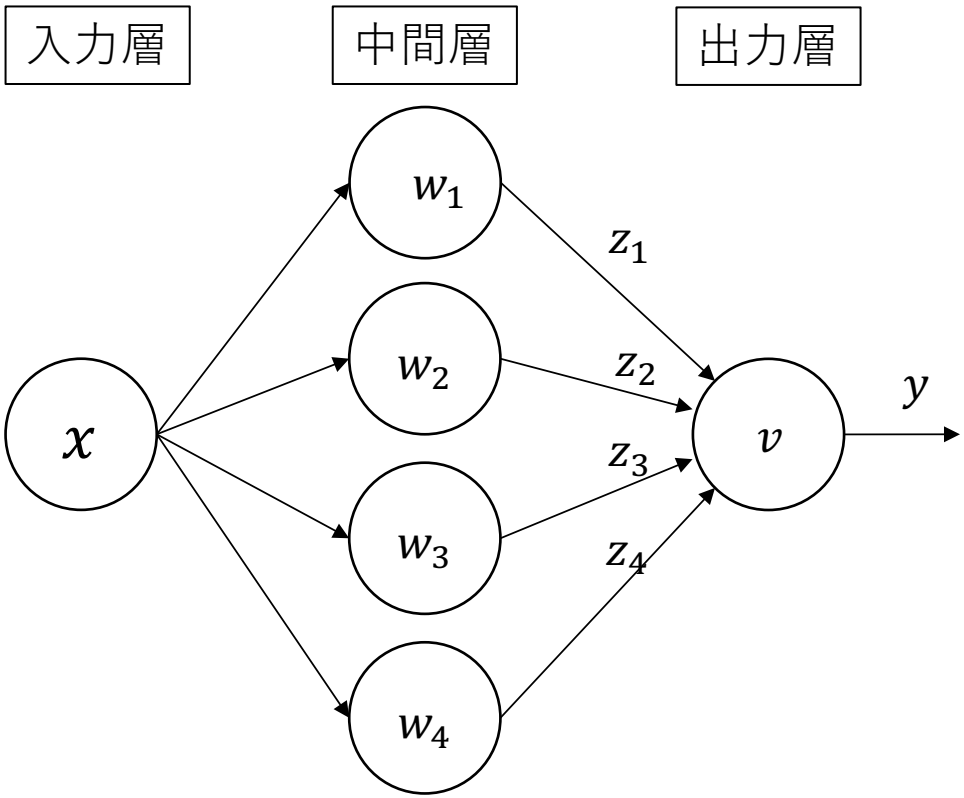
論文構成，目次

第1章はじめに　一部抜粋・和訳

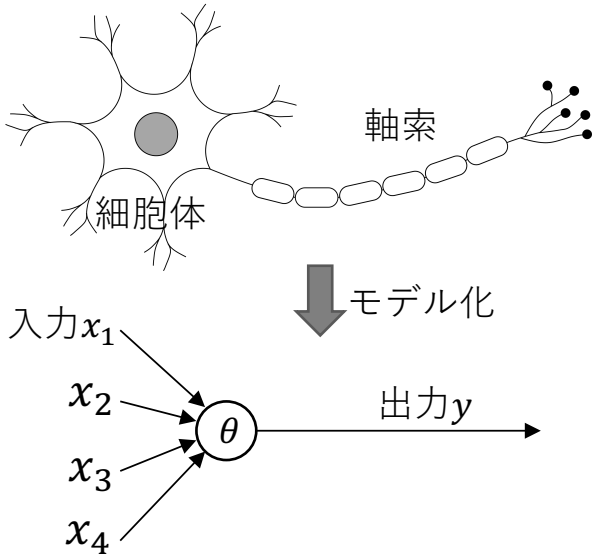
第2章は、基本的な統計モデルである指数族と曲線指数族についての紹介である。第3章は統計的予備知識で、最尤推定を幾何学的に説明する。ニューラルネットワークや統計学に慣れた読者はこれらのセクションを飛ばしてもよい。第4章では、本理論の主要な枠組みを紹介する。ニューラルネットワークの集合Mは、関連する確率分布の多様体Sに部分多様体として埋め込まれることが示される。問題は、DとMの間のKullback-Leibler divergenceを最小にする、MのネットワークとDのフィルドインデータを見つけることである。本論文の主要部分である第6章では、情報幾何学的な観点から両者の分析が行われる。第7章では、この2つのアルゴリズムがほとんどの実用的なケースで等価であることが示される。第8章では学習方法を扱い、第9章ではMとDにおける双測地的勾配流を研究する。 第10章では正規混合とエキスパートネットの混合を研究する。付録として、情報幾何学の直観的な入門を簡単に述べる。

ニューラルネットワーク(NN)

・入力を線形変換する計算処理が神経ネットワーク状に結合した数理モデル
・系全体で見ると，教師データから学習可能な非線形関数の普遍的な近似となる.
・入出力関係は x を入力したときの出力 y の条件付き確率 p(y|x) で確率的に記述.
・機械学習のモデルとして用いられる.



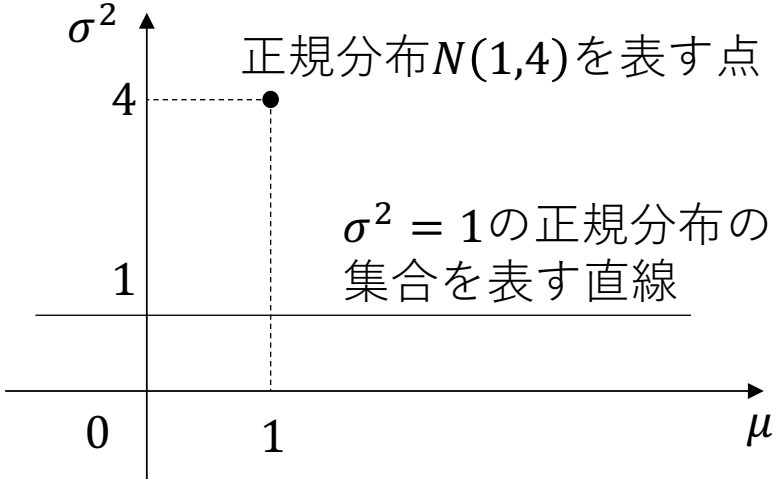
ベクトル$\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)$で表されたパラメータを持つNNにおける確率分布$P(y;\boldsymbol{\theta})$や条件付き確率分布$P(x|y;\boldsymbol{\theta})$を考える.



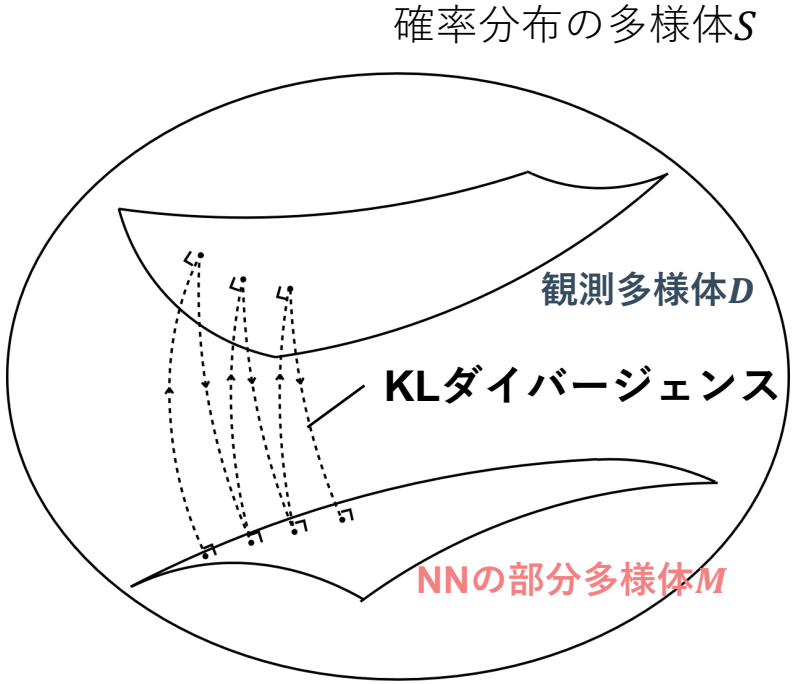
情報幾何学・確率分布の多様体

・情報幾何は、確率分布の多様体の情報構造に由来し、新しい微分幾何学的概念を持つ新しい数学的枠組みとして発展してきた.

例：正規分布$N(\mu, \sigma^2)$を表す座標平面



パラメータの次元が$n = 2$のため平面で，パラメータを一つ固定した部分集合は直線で表されているが，パラメータ数nを一般化すると，これらは多様体，部分多様体となる.

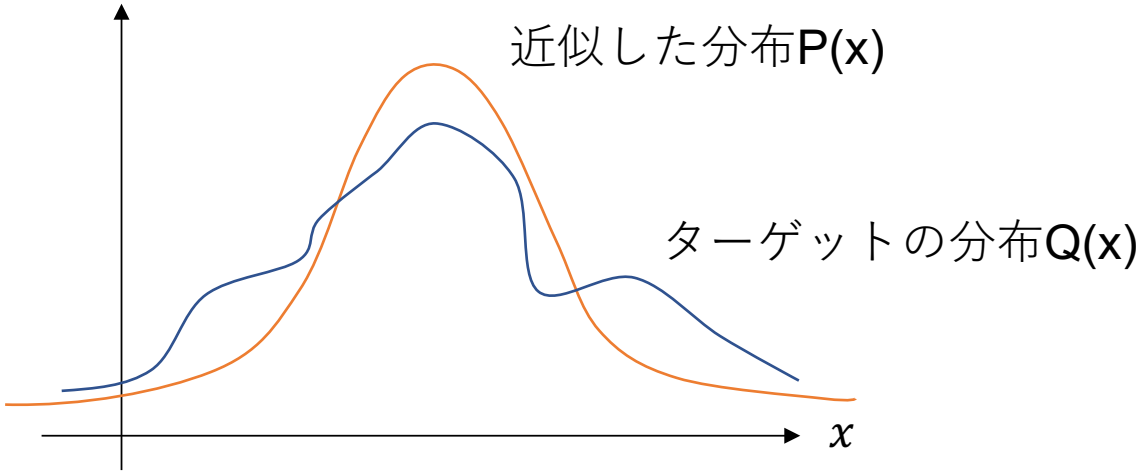


距離の尺度としてKLダイバージェンスを用い，これを最小化するM上の点を求める.

KLダイバージェンス(Kullback–Leibler divergence)

・二つの確率分布の違いを数値化したもの.
・モデル化したある確率分布P(x)が，観測された確率分布Q(x)をどれほど良く近似しているかを数値として表す.
・非負の値をとり，ぴったり一致する場合に0となる.
・計算方法は交差エントロピーから情報エントロピーを引く.

$$
\begin{aligned}
K(Q||P) &= H(Q,P) - H(Q) \\
&= E_{x\sim Q}[-\log P(x)] - E_{x\sim Q}[-\log Q(x)] \\
&= \int Q(x) \log \frac{Q(x)}{P(x)} dx
\end{aligned}
$$



指数型分布族の定義

・ $r_i(x), i = 0, \ldots, n$ をxの関数，ベクトル$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$をパラメータとし，確率変数$x$の確率密度関数が以下の式で表されるとき，その族$S = \{p(x; \boldsymbol{\theta})\}$を指数型分布族という．ただし，$x$はベクトルでも良い．

$$p(x;\theta) = \exp\left\{\sum_{i=1}^{n} \theta_i r_i(x) + k(x) - \psi(\theta)\right\}$$

・ xを$\boldsymbol{r}$に変換し，$\boldsymbol{r} = (r_1, \ldots, r_n) \in R^n$を$R^n$上の適当な測度$\mu(\boldsymbol{r})$に関して以下の分布を持つベクトル確率変数として扱うことができる．ただし，項$k(x)$とヤコビアンは$\mu(\boldsymbol{r})$に吸収される．

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

例1：正規分布

$$r_1 = x, r_2 = x^2, \theta_1 = \frac{\mu}{\rho^2}, \theta_2 = \frac{1}{2\rho^2}, \psi(\theta) = \frac{\mu^2}{2\rho^2} + \log\left(\sqrt{2\pi}\rho\right)$$

例2：離散分布 $p_i = Prob(x = i), i = 0, \ldots, n$とすると

$$r_i = \log\frac{p_i}{p_0}, \theta_i = \delta_i(x), \psi(\theta) = -\log p_0 = \log\{1 + \sum \exp(\theta_i)\}$$

確率的多層パーセプトロン

右下の図のような出力が一つの隠れ層パーセプトロンを考える.

・入力をx, m個隠れユニットの出力を$\boldsymbol{z} = (z_i), i = 1, \ldots, m, z_i = 0\ or\ 1$とする.

・z_iの確率を次のように定義

$$p(z_i|x) = \varphi(z_i, \boldsymbol{w_i} \cdot \boldsymbol{x})$$

ただし, ・φはシグモイド関数

$$\varphi(z, u) = \frac{\exp(zu)}{1 + \exp(u)}$$

・$\boldsymbol{w}_i$はi番目の隠れユニットのシナプスの重みベクトル

・$\boldsymbol{x}$に$x_0 \equiv 1$を追加し, $\boldsymbol{w}_i \cdot \boldsymbol{x} = \sum_{j=0}^{n} w_{ij} x_j$となる.

・$p(z_i = 0|x) + p(z_i = 1|x) = 1$になる.

次に出力ユニットを考える.

・隠れ層から信号$\boldsymbol{z}$を受け取り, 2値出力yを返す.

・ yの確率は$\boldsymbol{z}$に依存し, その確率を次のようにおく.

$$p(y|\boldsymbol{z}) = \varphi(y, \boldsymbol{v} \cdot \boldsymbol{z})$$

ただし, $\boldsymbol{v}$は出力ユニットのシナプスの重みベクトル

$\boldsymbol{z}$に$z_0 \equiv 1$を追加し, 定数項も考慮.

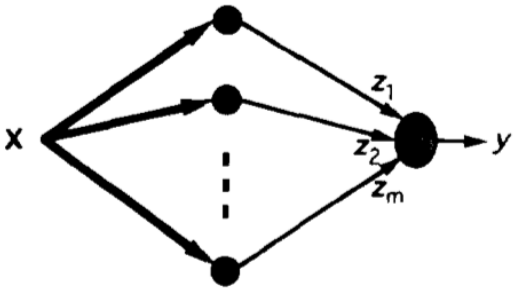


FIGURE 2. Stochastic multilayer perceptron.

確率的多層パーセプトロン

・入力xが与えられた時，変数(y, z)の条件付き確率は以下の式となる．

$$p(y, \boldsymbol{z}|x; \boldsymbol{u}) = p(y|\boldsymbol{z}; \boldsymbol{u}) \prod_{i=1}^{m} p(z_i|x; \boldsymbol{u})$$

・ただし， uは全パラメータを要約した変数である．

$$\boldsymbol{u} = (\boldsymbol{w_1}, \dots, \boldsymbol{w_m}, \boldsymbol{v})$$

・2値出力のパーセプトロンの期待値や， $[0,1]$の出力をする決定論的パーセプトロンでは以下の値となる．

$$z_i = \varphi(1, \boldsymbol{w_i} \cdot x), y = (1, \boldsymbol{v} \cdot \boldsymbol{z})$$

・ xが確率分布$\mathrm{q}(x) > 0$に従って生成されるとき， (y, z, x)の全確率分布は次の式．

$$p(y, \boldsymbol{z}, x; \boldsymbol{u}) = q(x) p(y, \boldsymbol{z}|x; \boldsymbol{u})$$

・対数を取ると次の式になる．

$$l(y, \boldsymbol{z}, x; \boldsymbol{u}) = \log p(y, \boldsymbol{z}|x; \boldsymbol{u}) + \log q(x)$$

・xの分布$\mathrm{q}(x)$が未知でも，条件付き確率の対数$\log p(y, \boldsymbol{z}|x; \boldsymbol{u})$を最大化することが$l(y, \boldsymbol{z}, x; \boldsymbol{u})$の最大化である．

確率的多層パーセプトロンと指数型分布族

・対数をさらに変形する.

$$l(y, \boldsymbol{z}, x|\boldsymbol{u}) = y\boldsymbol{v}\cdot\boldsymbol{z} + \sum_i z_i \boldsymbol{w_i}\cdot\boldsymbol{x} - \log\{1+\exp \boldsymbol{v}\cdot\boldsymbol{z}\} - \sum \log\{1+\exp(\boldsymbol{w}_i\cdot x)\}$$

・$\boldsymbol{z}$の値として，各成分が$\{0,1\}$をとるm次元ベクトル$\boldsymbol{k}$を導入．またデルタ関数を導入．$\delta_{\boldsymbol{k}}(\boldsymbol{z})$は$\boldsymbol{k}$でインデックスされた$2^m$次元のベクトル確率変数

$$\delta_{\boldsymbol{k}}(\boldsymbol{z}) = \begin{cases} 1 & \boldsymbol{z}=\boldsymbol{k} \\ 0 & \boldsymbol{z}\neq\boldsymbol{k} \end{cases}$$

・また $w_{\boldsymbol{k}} = \sum_i \boldsymbol{k_i}\boldsymbol{w_i}$を置き，$\sum_i \boldsymbol{z_i}\boldsymbol{w_i}\cdot\boldsymbol{x} = \sum_{\boldsymbol{k}} \delta_{\boldsymbol{k}}(\boldsymbol{z})\boldsymbol{w_k}\cdot\boldsymbol{x}$とすることで，次式となる.

$$l(y, \boldsymbol{z}|\boldsymbol{x}, \boldsymbol{u}) = \sum_{\boldsymbol{k}} \boldsymbol{k}\cdot\boldsymbol{v} y \delta_{\boldsymbol{k}}(\boldsymbol{z}) - \sum_{\boldsymbol{k}} \delta_{\boldsymbol{k}}(\boldsymbol{z}) \log\{1+\exp(\boldsymbol{k}\cdot\boldsymbol{v})\}$$
$$+ \sum_{\boldsymbol{k}} \delta_{\boldsymbol{k}}(\boldsymbol{z})\boldsymbol{w_k}\cdot\boldsymbol{x} - \sum_i \log\{1+\exp(\boldsymbol{w_i}\cdot\boldsymbol{x})\}$$

確率的多層パーセプトロンと指数型分布族

・以下のようにおくと，$p(y, \boldsymbol{z}|\boldsymbol{x}, \boldsymbol{u})$が指数型分布族であるとわかる．

$$r_{1,\boldsymbol{k}} = y\delta_{\boldsymbol{k}}(\boldsymbol{z}), r_{2,\boldsymbol{k}} = \delta_{\boldsymbol{k}}(\boldsymbol{z}), \theta_{1,\boldsymbol{k}} = \boldsymbol{k}\cdot\boldsymbol{v}, \theta_{2,\boldsymbol{k}} = \boldsymbol{w}_{\boldsymbol{k}}\cdot\boldsymbol{x} - \log\{1+\exp(\boldsymbol{k}\cdot\boldsymbol{v})\}$$

$$\psi(\boldsymbol{\theta}) = \sum_i \log\{1 + (\boldsymbol{w}_i \cdot \boldsymbol{x})\}$$

$$p(\boldsymbol{r}|\boldsymbol{x};\boldsymbol{u}) = \exp\{\boldsymbol{r}\cdot\boldsymbol{\theta} - \psi(\boldsymbol{\theta})\}$$

繰り返し観測と積空間

・$\boldsymbol{r_1}, \ldots, \boldsymbol{r_T}$を同じ分布$p(\boldsymbol{r}; \boldsymbol{\theta})$からの$T$個の独立な観測値とする.

$$p(\boldsymbol{r_1}, \boldsymbol{r_2}, \ldots, \boldsymbol{r_T}; \boldsymbol{\theta}) = \prod_{t=1}^{T} p(\boldsymbol{r_t}; \boldsymbol{\theta})$$

・ $p(r; \boldsymbol{\theta})$が指数型の場合

$$p(\boldsymbol{r_1}, \boldsymbol{r_2}, \ldots, \boldsymbol{r_T}; \boldsymbol{\theta}) = \exp\{(\sum r_t) \cdot \boldsymbol{\theta} - T\psi(\boldsymbol{\theta})\}$$

・よって全体としても指数型分布族になる.

・次に条件付き分布$p(y, \boldsymbol{z}|\boldsymbol{x})$を考える.
・$(\boldsymbol{x_t}, y_t, \boldsymbol{z_t}), t = 1, \ldots, T$を指数型分布族からの独立した観測値とする.
・ $p\{r_t; \boldsymbol{\theta}(\boldsymbol{x_t})\} = \exp\{\theta(x_t) \cdot r_t - \psi_t\}$とすると，結合条件分布は次の式.

$$p\{(y_1, \boldsymbol{z_1}), \ldots, (y_T, \boldsymbol{z_T})|\boldsymbol{x_1}, \ldots, \boldsymbol{x_T}\} = \exp\left\{\sum \boldsymbol{\theta}(\boldsymbol{x_t}) \cdot \boldsymbol{r_t} - \sum \psi_t\right\}$$

・ ただし， $\boldsymbol{r} = \boldsymbol{r}(y, \boldsymbol{z})$

繰り返し観測と積空間

・ $S_t = \{p(r_t; \theta(x_t))\}$は$t$番目の入力に対する条件付き確率分布の多様体であり，
全部の結合分布は以下の積空間に与えられる．

$$S_T^* = S_1 \times S_2 \times \cdots \times S_T$$

・S_T^*の観測値$\boldsymbol{r}_T^*$を次のようにおく．

$$\boldsymbol{r}_T^* = (\boldsymbol{r_1}, \ldots, \boldsymbol{r_T})$$

・パラメータ$\boldsymbol{\theta}_T^*$を次のようにおく．

$$\boldsymbol{\theta}_T^* = \{\boldsymbol{\theta}(\boldsymbol{x_1}), \ldots, \boldsymbol{\theta}(\boldsymbol{x_T})\}$$

・ $\boldsymbol{\theta}_T^* \cdot \boldsymbol{r}_T^* = \sum_{t=1}^{T} \boldsymbol{\theta_t} \cdot \boldsymbol{r_t}, \psi(\boldsymbol{\theta}_T^*) = \sum_{t=1}^{T} \psi(\boldsymbol{\theta}_t)$とおくと次式となり，指数型分布族．

$$p(\boldsymbol{r}_T^*; \boldsymbol{\theta}_T^*) = \exp\{\boldsymbol{\theta}_T^* \cdot \boldsymbol{r}_T^* - \psi(\boldsymbol{\theta}_T^*)\}$$

・ただし， $\boldsymbol{\theta}_T^*$は$\theta_{1,\boldsymbol{k}} = \boldsymbol{k} \cdot \boldsymbol{v}, \theta_{2,\boldsymbol{k}} = \boldsymbol{w_k} \cdot \boldsymbol{x} - \log\{1 + \exp(\boldsymbol{k} \cdot \boldsymbol{v})\}$より$\boldsymbol{w_i}$と$\boldsymbol{v}$に制限され，
任意の値をとることができず， S_T^*の部分領域に制限される

曲線指数型分布族

- $S = \{p(\boldsymbol{r};\theta)\}$を$n$次元指数型分布族とし，$M$をその$m$次元部分多様体とすると，$S$の一部を占める部分集合$M$は曲線指数型分布族(Curved Exponential Families)という．**MはSの部分多様体である**.
- $\boldsymbol{u} = (u_1, \ldots, u_m), m \leq n$を$M$の座標系とすると，$M$は$\boldsymbol{u}$の関数$\boldsymbol{\theta}(\boldsymbol{u})$を用いて次の分布により構成される

$$p\big(\boldsymbol{r};\boldsymbol{\theta}(\boldsymbol{u})\big) = \exp\big\{\boldsymbol{\theta}(\boldsymbol{u})\cdot\boldsymbol{r} - \psi\big(\boldsymbol{\theta}(\boldsymbol{u})\big)\big\}$$

- $\boldsymbol{\theta}_T^* = \{\boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta_T}\}$は積空間$\boldsymbol{S}_T^*$の座標であり，パラメータ$\boldsymbol{u}$を用いて$\boldsymbol{\theta_t} = \boldsymbol{\theta}(\boldsymbol{x_t}, \boldsymbol{u})$.
- $\boldsymbol{x_t}$が与えられた時，$\boldsymbol{\theta}_T^*$における自由なパラメータは$\boldsymbol{u}$であり，Tが増加しても次元数は固定.

- よって，M_T^*はS_T^*の部分多様体を形成し，M上の分布$p\big(\boldsymbol{r};\boldsymbol{\theta}(\boldsymbol{u})\big)$は$S$に属しており，$\boldsymbol{\theta}(\boldsymbol{u})$は$S$上での座標である(右図).

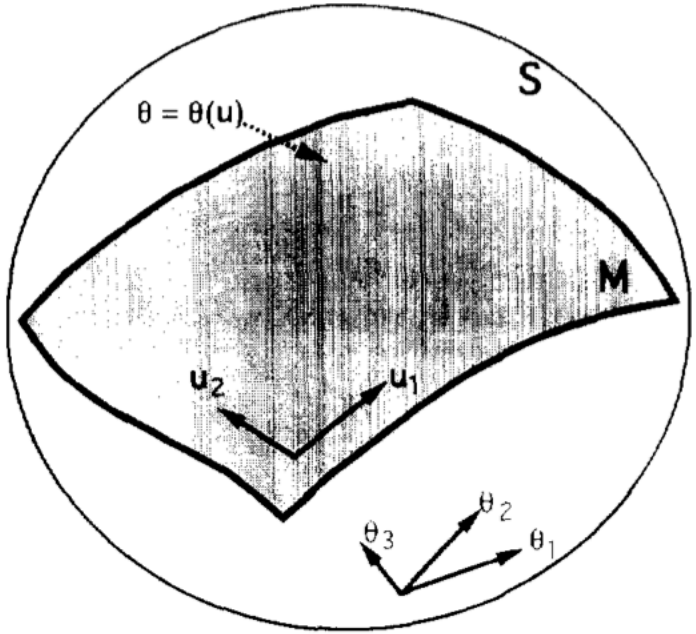


FIGURE 3. Curved exponential family.

曲線指数型分布族の例：正規分布の乗算モデル

・ εを正規分布$N(0,1)$に従う確率変数とする.
・ノイズεが入った大きさ1の信号を入力， u倍に増幅された出力xがされるとする.

$$x = u(1+\varepsilon)$$

・ xは$N(u, u^2)$に従う.
・この集合Mは正規分布$S = \{N(\mu, \rho^2)\}$の曲線指数型分布族である． 平均μと分散ρ^2は独立ではなく，共通のuによって規定される.
・よって， MはS上で曲線(放物線)となる

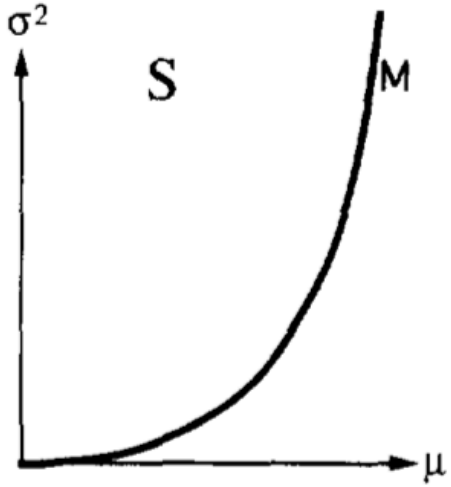


FIGURE 4. Example of curved exponential family.

期待値パラメータ

・すべての確率変数$\boldsymbol{r}$が可視であると仮定する.
・パラメータ$\boldsymbol{\theta}$で表される分布$p(\boldsymbol{r};\boldsymbol{\theta})$の期待値を$\boldsymbol{\eta}$とする.

$$\boldsymbol{\eta}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}[\boldsymbol{r}] = \int \boldsymbol{r} p(\boldsymbol{r};\boldsymbol{\theta}) d\boldsymbol{\mu}(\boldsymbol{r})$$

・ただし，確率の積分 $\int p(\boldsymbol{r};\boldsymbol{\theta}) d\boldsymbol{\mu}(\boldsymbol{r}) = \int \exp\{\boldsymbol{\theta}\cdot\boldsymbol{r} - \psi(\boldsymbol{\theta})\} d\boldsymbol{\mu}(\boldsymbol{r})$ は1であることより，これを$\boldsymbol{\theta}$で微分することで，以下の式が導ける.

$$\boldsymbol{\eta}(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \psi(\boldsymbol{\theta})$$

・さらに，$\boldsymbol{\theta}$と$\boldsymbol{\eta}$の変換は1対1であるため，$\boldsymbol{\eta}$はSのもう一つのペラメータとして使用できる．ここで，$\boldsymbol{\eta}$を期待値パラメータと呼ぶ.
・よって，$\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\eta})$と逆変換できる.

・$\boldsymbol{\eta}$で定義される確率分布$p^*\{\boldsymbol{r};\boldsymbol{\eta}(\boldsymbol{\theta})\} = p(\boldsymbol{r};\boldsymbol{\theta})$を導入.
・$\varphi(\boldsymbol{\eta})$を分布$p^*\{\boldsymbol{r};\boldsymbol{\eta}(\boldsymbol{\theta})\}$のエントロピーの負とする.

$$\varphi(\boldsymbol{\eta}) = \int p^*(\boldsymbol{r};\boldsymbol{\eta}) \log p^*(\boldsymbol{r};\boldsymbol{\eta})\, d\mu(\boldsymbol{r})$$

・すると，$\boldsymbol{\theta}(\boldsymbol{\eta})$は次のように与えられる(Amari, 1985).

$$\boldsymbol{\theta}(\boldsymbol{\eta}) = \frac{\partial}{\partial \boldsymbol{\eta}} \varphi(\boldsymbol{\eta})$$

期待値パラメータの例

・正規分布の場合

$$\eta_1 = E[x] = \mu = \frac{\theta_1}{2\theta_2}, \eta_2 = E[x^2] = \mu^2 + \rho^2 = \left(\frac{\theta_1}{2\theta_2}\right)^2 - \frac{1}{2\theta_2}$$

・離散分布の場合

$$\eta_i = E[\delta_i(x)] = p_i = \frac{\exp(\theta_i)}{1 + \sum \exp(\theta_i)}$$

・確率的パーセプトロン$S = \{p(y, \boldsymbol{z}; \boldsymbol{\theta})\}$の場合

$$\boldsymbol{\eta} = (\eta_{1,k}, \eta_{2,k})$$

$$\eta_{1,k} = E[y\delta_k(\boldsymbol{z})] = \varphi(1, \boldsymbol{v} \cdot \boldsymbol{k}) \prod \varphi(k_i, \boldsymbol{w_i} \cdot \boldsymbol{x})$$

$$\eta_{2,k} = E[\delta_k(\boldsymbol{z})] = \prod \varphi(k_i, \boldsymbol{w_i} \cdot \boldsymbol{x})$$

期待値パラメータと最尤推定量

・期待値パラメータは最尤推定量の研究に有用である.

・指数型分布族$S = \{p(\boldsymbol{r};\boldsymbol{\theta})\}$の尤度関数の対数は以下の式.

$$l(\check{\boldsymbol{r}};\boldsymbol{\theta}) = \boldsymbol{\theta}\cdot\check{\boldsymbol{r}} - \psi(\boldsymbol{\theta})$$

$\check{\boldsymbol{r}}$ ：$\boldsymbol{r}$の観測値
$\widehat{\boldsymbol{\theta}}$：$\boldsymbol{\theta}$の最尤推定量

・これを微分することにより最尤推定量が求められる.

$$\frac{\partial l}{\partial \theta} = \check{\boldsymbol{r}} - \boldsymbol{\eta}\left(\widehat{\boldsymbol{\theta}}\right) = \mathbf{0}$$

・よって期待値$\boldsymbol{\eta}$は観測値そのもの．$\widehat{\boldsymbol{\theta}}$も$\widehat{\boldsymbol{\theta}} = \frac{\partial}{\partial \boldsymbol{\eta}}\varphi(\widehat{\boldsymbol{\eta}})$から求められる.

・$\widehat{\boldsymbol{\eta}}$または$\widehat{\boldsymbol{\theta}}$は$S$における点を決定するものであり，これを観測点と呼ぶ.

繰り返し観測と観測点

・$\boldsymbol{r_1}, \dots, \boldsymbol{r_T}$を$T$個の独立な確率変数とし，$\boldsymbol{r_t}$は分布$p(\boldsymbol{r_t}; \boldsymbol{\theta_t})$に従うとする．

$$p(\boldsymbol{r_1}, \dots, \boldsymbol{r_T}; \boldsymbol{\theta_1}, \dots, \boldsymbol{\theta_T}) = \prod_{t=1}^{T} p(\boldsymbol{r_t}; \boldsymbol{\theta_t})$$

$$p(\boldsymbol{r_T^*}; \boldsymbol{\theta_T^*}) = \exp\{\boldsymbol{\theta_T^*} \cdot \boldsymbol{r_T^*} - \psi\}$$

・期待座標$\boldsymbol{\eta_T^*} = (\boldsymbol{\eta_1}, \dots, \boldsymbol{\eta_T})$は$\boldsymbol{\eta_t} = E_{\theta_t}[\boldsymbol{r_t}]$で与えられ，観測値$\check{\boldsymbol{r}}_\boldsymbol{t}$ により以下の通り．

$$\boldsymbol{\eta_T^*} = (\check{\boldsymbol{r}}_\boldsymbol{1}, \dots, \check{\boldsymbol{r}}_\boldsymbol{T})$$

・対応する最尤推定量$\widehat{\boldsymbol{\theta}}_\boldsymbol{t}$は次のように与えられる．

$$\widehat{\boldsymbol{\theta}}_\boldsymbol{t} = \frac{\partial \varphi(\widehat{\boldsymbol{\eta}}_t)}{\partial \boldsymbol{\eta}}$$

・NNの場合，$\boldsymbol{\theta}_t, \boldsymbol{\eta}_t$は入力$\boldsymbol{x_t}$とパラメータ$\boldsymbol{u}$により$\boldsymbol{\theta}_t = \boldsymbol{\theta}(\boldsymbol{x_t}, \boldsymbol{u}), \boldsymbol{\eta}_t = \boldsymbol{\eta}(\boldsymbol{x_t}, \boldsymbol{u})$と書かれ，これが曲線指数型分布族$M^*$を定義する．しかし観測点$\widehat{\boldsymbol{\eta}}_\boldsymbol{T}^*, \widehat{\boldsymbol{\theta}}_\boldsymbol{T}^*$がこれを満たすとは限らず，必ずしも$M$に含まれるとは限らない．
・よって最尤推定量$\widehat{\boldsymbol{u}}$を得るためには，Mに射影する必要がある．

繰り返し観測と観測点

・すべての$\boldsymbol{r_t}$が同じ分布$p(\boldsymbol{r_t};\boldsymbol{\theta})$に従うとする．$\boldsymbol{\theta}=\boldsymbol{\theta_1}=\cdots=\boldsymbol{\theta_T}$

$$p(\boldsymbol{r_1},\ldots,\boldsymbol{r_T};\boldsymbol{\theta})=\exp\left\{\sum_{t=1}^{T}\boldsymbol{r_t}\cdot\boldsymbol{\theta}-T\psi(\theta)\right\}$$

・観測値の平均値$\bar{\boldsymbol{r}}=\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{r}_t$を導入し，$S_T^*$を単一の$S$に還元できる．

$$p(\bar{\boldsymbol{r}};\boldsymbol{\theta})=\exp\{T(\bar{\boldsymbol{r}}\cdot\boldsymbol{\theta})+Tk(\bar{\boldsymbol{r}})-T\psi(\boldsymbol{\theta})\}$$

・これは$\bar{\boldsymbol{r}}$が$\boldsymbol{\theta}$や$\boldsymbol{u}$を推定するのに十分な統計量であることを示している．

・ただし，$k(\bar{\boldsymbol{r}})$は$\boldsymbol{r}$の変換によって生じ，$\exp\{Tk(\bar{\boldsymbol{r}})\}=\int_{\bar{\boldsymbol{r}}}\prod d\mu(\boldsymbol{r_t})$で表され，$d\bar{\mu}(\overline{\boldsymbol{r}})=\exp\{Tk(\overline{\boldsymbol{r}})\}d\mu$で除去できる．

・i.i.d.である場合は，S_T^*を参照せず指数型関数族Sで$\bar{\boldsymbol{r}}\cdot\boldsymbol{\theta}-\psi(\boldsymbol{\theta})$を最大化することで最尤推定量$\widehat{\boldsymbol{\theta}}$が与えられる．

・以上より，$\bar{\boldsymbol{r}}=\frac{\partial}{\partial\boldsymbol{\theta}}\psi(\boldsymbol{\theta}),\widehat{\boldsymbol{\eta}}=\bar{\boldsymbol{r}}$．観測されたデータ$\check{\boldsymbol{r}}_t$はFisherの意味で情報損失無しに$S$における$\eta$座標系の一点$\widehat{\boldsymbol{\eta}}$に要約される．

曲線指数型分布族での推定

・ S上のMにおける最尤推定を考える.

・観測点$\hat{\boldsymbol{\eta}} = \bar{\boldsymbol{r}}$が与えられた時，最尤推定量$\hat{\boldsymbol{u}}$は次式の対数尤度最大化で求める.

$$l(\bar{\boldsymbol{r}}; \boldsymbol{u}) = \bar{\boldsymbol{r}} \cdot \boldsymbol{\theta}(\boldsymbol{u}) - \psi\{\boldsymbol{\theta}(\boldsymbol{u})\}$$

・これは，$p(\bar{\boldsymbol{r}}; \hat{\boldsymbol{\theta}})$から$p(\bar{\boldsymbol{r}}; \boldsymbol{\theta}(\boldsymbol{u}))$へのKLダイバージェンスを最小化することと等価.

$$K(\hat{\boldsymbol{\theta}} || \boldsymbol{\theta}(\boldsymbol{u})) = \int p(\bar{\boldsymbol{r}}; \hat{\boldsymbol{\theta}}) \log \frac{p(\bar{\boldsymbol{r}}; \hat{\boldsymbol{\theta}})}{p(\bar{\boldsymbol{r}}; \boldsymbol{\theta}(\boldsymbol{u}))} d\mu(\bar{\boldsymbol{r}})$$
$$= -H(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\eta}} \cdot \boldsymbol{\theta}(\boldsymbol{u}) + \psi(\boldsymbol{\theta}(\boldsymbol{u}))$$

・幾何学的には$\hat{\boldsymbol{\theta}}$を$M$に射影することで与えられる.

・尤度方程式を解くことで$\boldsymbol{u}$を求める.

$$\frac{\partial}{\partial \boldsymbol{u}} \{\bar{\boldsymbol{r}} \cdot \boldsymbol{\theta}(\boldsymbol{u}) - \psi(\boldsymbol{\theta}(\boldsymbol{u}))\} = \boldsymbol{0}$$

$$\boldsymbol{B}(\boldsymbol{u})\{\bar{\boldsymbol{r}} - \boldsymbol{\eta}(\boldsymbol{u})\} = \boldsymbol{0}$$

$$\boldsymbol{B}(\boldsymbol{u}) = \frac{\partial \boldsymbol{\theta}(\boldsymbol{u})}{\partial \boldsymbol{u}} = \left(\frac{\partial \boldsymbol{\theta}_i(\boldsymbol{u})}{\partial u_j}\right)$$

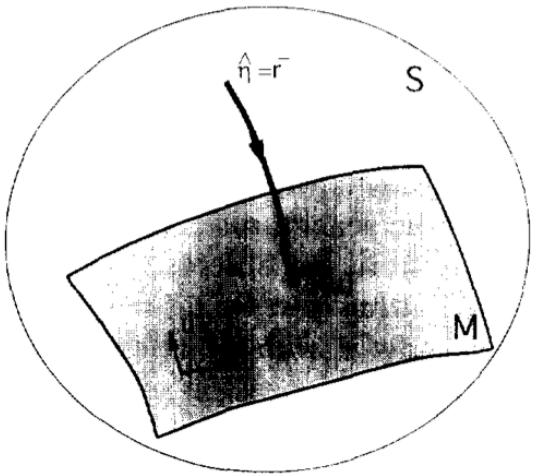


FIGURE 5. Maximum likelihood estimation.

曲線指数型分布族での推定

・ 直積空間S_T^*でKLダイバージェンスは

$$K\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{T}}^{*}||\boldsymbol{\theta}_{\boldsymbol{T}}^{*}(\boldsymbol{u})\right)=\sum_{t=1}^{T}K\left(\widehat{\boldsymbol{\theta}}_{t}||\boldsymbol{\theta}_{\boldsymbol{t}}(\boldsymbol{u})\right)=-\sum H\left(\widehat{\boldsymbol{\theta}}_{t}\right)-\sum\widehat{\boldsymbol{\eta}}_{\boldsymbol{t}}\cdot\boldsymbol{\theta}_{\boldsymbol{t}}(\boldsymbol{u})+\sum\psi\left(\boldsymbol{\theta}_{\boldsymbol{t}}(\boldsymbol{u})\right)$$

・最尤推定量$\widehat{\boldsymbol{u}}$は$\boldsymbol{M}^{*}$における$\boldsymbol{\theta}_{\boldsymbol{T}}^{*}(\boldsymbol{u})$のうち，観測点$\widehat{\boldsymbol{\eta}}_{\boldsymbol{T}}^{*}=\bar{\boldsymbol{r}}^{*}\in S_{\boldsymbol{T}}^{*}$に最も近い点.

・尤度式$\sum_{t=1}^{T}\boldsymbol{B}(\boldsymbol{x}_{\boldsymbol{t}},\boldsymbol{u})\{\check{\boldsymbol{r}}_{\boldsymbol{t}}-\boldsymbol{\eta}(\boldsymbol{x}_{\boldsymbol{t}},\boldsymbol{u})\}=\boldsymbol{0}$

ただし，$\boldsymbol{B}(\boldsymbol{x}_{\boldsymbol{t}},\boldsymbol{u})=\frac{\partial\boldsymbol{\theta}(\boldsymbol{x}_{\boldsymbol{t}},\boldsymbol{u})}{\partial\boldsymbol{u}}$

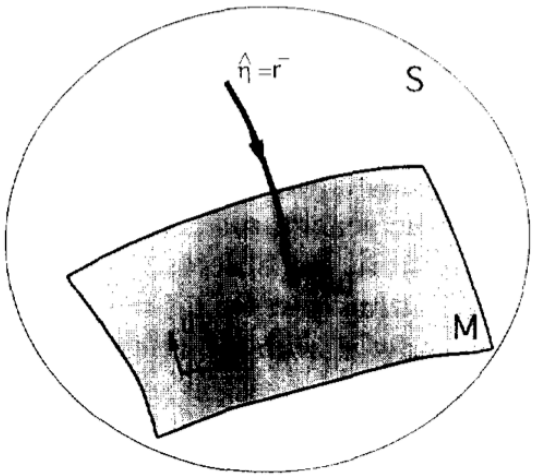


FIGURE 5. Maximum likelihood estimation.

部分観測

・確率変数$\boldsymbol{r}$の一部が観測できない場合を考える.
例)NNの隠れ層の出力は観測されない.
・この場合，観測されない変数も適切に決定する必要がある.

$\boldsymbol{r} = (\boldsymbol{s_v}, \boldsymbol{s_h})$とおく.
$\boldsymbol{s_v}$：観測可能な統計値
$\boldsymbol{s_h}$：観測不可な欠損値

・この時，観測点$\hat{\boldsymbol{\eta}} = \boldsymbol{r} = (\boldsymbol{s_v}, \boldsymbol{s_h})$を一意に特定することができず，任意の$\boldsymbol{s_h}$をとる.

・この候補点は観測データの部分多様体Dを形成する.

$$D = \{\hat{\boldsymbol{\eta}} | \hat{\boldsymbol{\eta}} = \boldsymbol{r}(\boldsymbol{s_v}, \boldsymbol{s_h}), \boldsymbol{s_v}\text{は観測値(固定)}, \ \boldsymbol{s_h}\text{は任意}\}$$

・多くの場合， $\boldsymbol{r}$は$\boldsymbol{s_h}$に対して線形
  ・ $\boldsymbol{r}(\boldsymbol{s_v}, \boldsymbol{s_{1h}}), \boldsymbol{r}(\boldsymbol{s_v}, \boldsymbol{s_{2h}}) \in D$， $0 < t < 1$に対し$t\boldsymbol{r}(\boldsymbol{s_v}, \boldsymbol{s_{1h}}) + (1-t)\boldsymbol{r}(\boldsymbol{s_v}, \boldsymbol{s_{2h}}) \in D$

・ hの次元の部分多様体を形成し，データ部分多様体，観測部分多様体と呼ばれる.

・Dが離散ベクトル変数の時，Dは多数の候補点からなるが，η座標の線形部分多様体に拡張する.

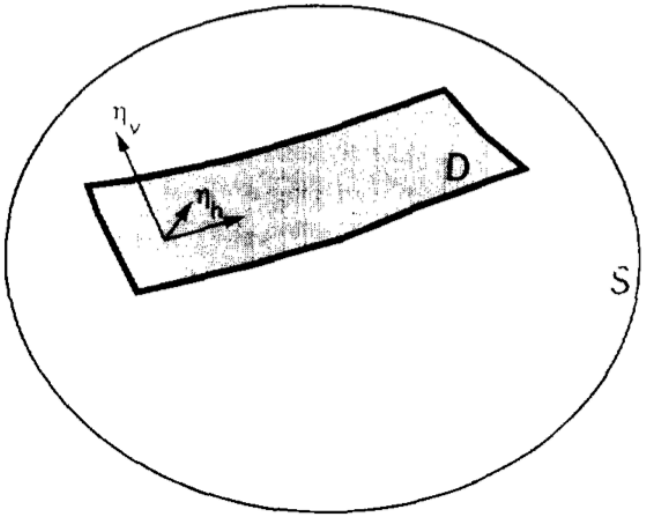


FIGURE 6. Data submanifold D : η =fixed.

部分観測

・η_hはDの内部座標系の役割を果たす.

・ Dがη座標で線形である場合，非特異行列Aを用いて$\tilde{\boldsymbol{\eta}} = A\boldsymbol{\eta}, \tilde{\boldsymbol{r}} = A\boldsymbol{r}$と線形変換し$\tilde{\boldsymbol{\eta}} = (\tilde{\boldsymbol{\eta}}_{\boldsymbol{v}}, \tilde{\boldsymbol{\eta}}_{\boldsymbol{h}})$に分割できる．またこの時$\tilde{\boldsymbol{\theta}} = A^{-1}\boldsymbol{\theta}$であり，確率分布は以下の通り

$$p\left(\tilde{\boldsymbol{r}}; \tilde{\boldsymbol{\theta}}\right) = \exp\left\{\tilde{\boldsymbol{\theta}} \cdot \tilde{\boldsymbol{r}} - \tilde{\psi}\left(\tilde{\boldsymbol{\theta}}\right)\right\}$$

・$\tilde{\boldsymbol{\theta}}$が新しい正規化パラメータとなり，$\tilde{\boldsymbol{\eta}} = \frac{\partial \tilde{\psi}}{\partial \boldsymbol{\theta}}$が期待値パラメータとなる.

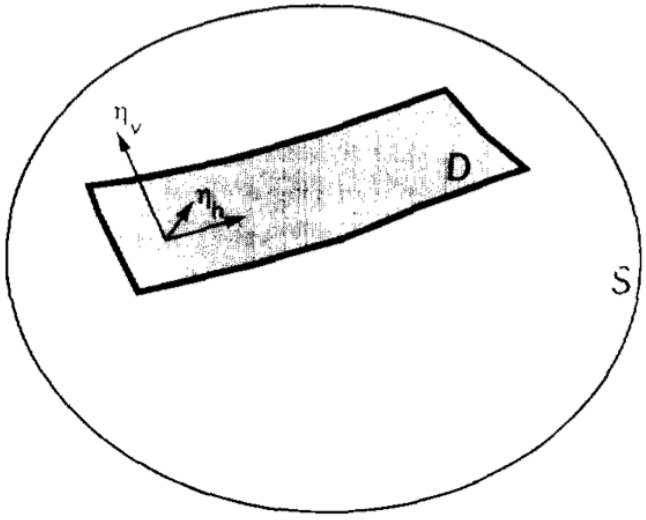


FIGURE 6. Data submanifold D : η =fixed.

例：正規分布の乗法モデルの部分観測

- $x_1, \ldots, x_T$： $N(u, u^2)$に従う確率変数.
- $r_v = \frac{1}{t}\sum x_t$ (観測可能)， $r_h = \frac{1}{T}\sum x_t^2$(観測不可)

- モデル多様体Mは曲線指数型分布族． η座標は$\eta_v = \mu = u, \eta_h = \mu^2 + \sigma^2 = 2u^2$
- データ多様体Dは$D = \{N(\mu, \sigma^2) | \mu = \tilde{r}, \sigma$は任意$\}$

- $\sum x_t$と$\sum x_t^2$が観測されるとき観測点は一意だが，この場合正確な観測点はD上にあることしかわからない.

- 最尤推定量は$\hat{\eta}$をm射影することで得られる,

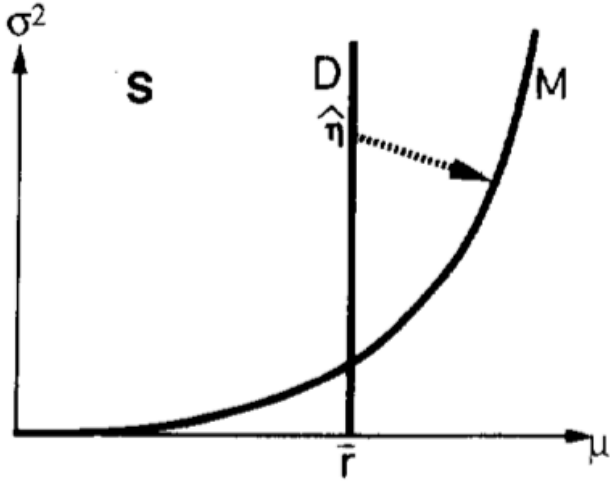


FIGURE 7. Example of data submanifold *D* and observed point $\hat{\eta}$.

例：確率的パーセプトロンのデータ部分多様体

・ 時刻tの確率変数$\boldsymbol{r}(t)=\left(\boldsymbol{r}_{1,\boldsymbol{k},t},\boldsymbol{r}_{2,\boldsymbol{k},t}\right)$において，観測点$\hat{\boldsymbol{\eta}}$が次のように表される．

$$\hat{\boldsymbol{\eta}}_{1,\boldsymbol{k}}=\boldsymbol{r}_{1,\boldsymbol{k},t}=y_t\delta_{\boldsymbol{k}}(\boldsymbol{z}_t),\quad \hat{\boldsymbol{\eta}}_{2,\boldsymbol{k}}=\boldsymbol{r}_{2,\boldsymbol{k},t}=\delta_{\boldsymbol{k}}(\boldsymbol{z}_t)$$

・y_tが観測可能，$\delta_{\boldsymbol{k}}(\boldsymbol{z}_t)$が観測不可能であり，$\boldsymbol{z}_t$により候補点が存在．

・η座標で線形結合をとり，$\delta_{\boldsymbol{k}}(\boldsymbol{z}_t)$につけた重みを$\alpha_{\boldsymbol{k}}$とおくと，観測多様体$D_t$は

$$D_t=\left\{\eta\middle|\eta_{1,\boldsymbol{k}}=y_t\alpha_{\boldsymbol{k}},\eta_{\boldsymbol{2},\boldsymbol{k}}=\alpha_{\boldsymbol{k}}\right\}$$

ただし，$\sum_{\boldsymbol{k}}\alpha_{\boldsymbol{k}}=1$

・$\alpha_{\boldsymbol{k}}$は$\boldsymbol{z_t}=\boldsymbol{k}$の確率と解釈しても良いが，観測不可能．

・データ部分多様体は

$$D_T^*=D_1\times\cdots\times D_T$$

EM(Expectation and Maximizing)アルゴリズム

・部分的に観測されたデータDまたはD_T^*から良い推定量$\hat{\boldsymbol{u}}$を得る手法.
・真のパラメータ$\boldsymbol{u}$と欠損データを同時に反復的に推定する.
・ $\boldsymbol{u}$に依存する分布の候補があれば$\hat{\boldsymbol{r}}$を推測することができ， $\hat{\boldsymbol{r}}$があれば$\boldsymbol{u}$が推測可能

方法

0. 準備．任意の初期推測値$\hat{\boldsymbol{u}}_{\boldsymbol{0}}$を選ぶ．初期推測分布$\hat{P}_0 \in M$は$\boldsymbol{\theta}(\hat{\boldsymbol{u}}_{\boldsymbol{0}})$で与えられる.

1. Eステップ．候補確率分布$\hat{P}_{\boldsymbol{i}} \in M$に基づき， $\boldsymbol{r}$の条件付き期待値を計算する.

$$\hat{\boldsymbol{\eta}}_{(\boldsymbol{i})} = E\left[\boldsymbol{r}\middle|\boldsymbol{s}_{\boldsymbol{v}}; \hat{\boldsymbol{P}}_{\boldsymbol{i}}\right]$$

これが座標$\hat{\boldsymbol{\eta}}_{(\boldsymbol{i})}$で表される観測点$\hat{Q}_i \in D$の$i$番目の候補となる.

2. Mステップ．推定された観測点$\hat{Q}_i$から最尤推定量$\hat{\boldsymbol{u}}_{\boldsymbol{i+1}}$を算出，パラメータを更新.
これはKLダイバージェンス$K\left(\hat{Q}_{\boldsymbol{i}}||P\right)$の最小化または対数尤度の最大化で求める.

パラメータが収束するまでEステップとMステップを繰り返す.

EM(Expectation and Maximizing)アルゴリズム

・部分観測データが$\bar{\boldsymbol{r}}$でまとめられない場合，積空間S_T^*にデータ多様体D_T^*が存在.
この時，Eステップは次のように推定する.

$$\hat{\boldsymbol{r}}_T^* = E\left[\boldsymbol{r}_T^* \middle| s_{v,1}, \ldots, s_{v,T}; \hat{P}_i\right]$$

ただし，$\boldsymbol{r_t}$は$s_{v,t}$のみと相関があるため，$\hat{\boldsymbol{\eta}}_t(i) = E\left[\boldsymbol{r}_t \middle| s_{v,t}; \hat{P}_i\right]$

・尤度は1回のEステップとMステップで増加する.
→尤度関数の局所最大に収束する(最尤推定量に収束するとは限らない).

・対数尤度を直接最大化する勾配法に比べ，大域的収束特性を持つ.

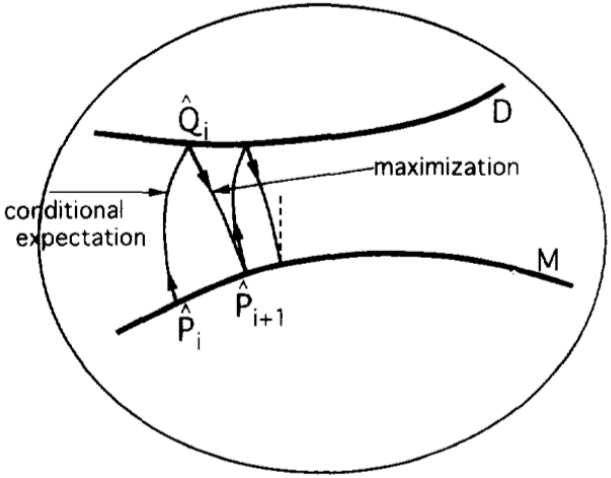


FIGURE 8. *EM* algorithm.

em(exponential mixture)アルゴリズム

・モデルは多様体Mに，観測値は多様体Dにあるため， DとMのダイバージェンスを最小にする$\hat{P} \in M, \hat{Q} \in D$の組を探索する． $\hat{Q}$は最尤推定量となる．

$$K(\hat{Q}||\hat{P}) = \min_{P \in M, Q \in D} K(Q||P)$$

・与えられたQに対し$K(Q||P)$を最小化する$\hat{P} \in M$はQのMへのm-射影で与えられる．
  ・m-射影は$\hat{P}$とQとを結ぶm-測地線により与えられる．
  ・指数型分布族では曲線がη座標に線形であるとき，m-測地線となる．
  ・m-測地線は$\hat{P}$でMと直交．直行性はフィッシャー・リーマン情報量で定義．

・与えられたPに対し$K(Q||P)$を最小化する$\hat{Q} \in D$はPのDへのe-射影で与えられる．
  ・m-射影はPと$\hat{Q}$とを結ぶe-測地線により与えられる． e-測地線は$\hat{Q}$でDと直交．
  ・指数型分布族では曲線がθ座標に線形であるとき，e-測地線となる．

・観測される部分多様体がD_T^*である場合，$K(Q^*||P^*) = \sum_{t=1}^{T} K(Q_t||P_t)$として各成分$P_t$を$D_t$にe-射影することにより与えられる．各 tに成分的に適用できる．

em(exponential mixture)アルゴリズム

方法

0. 準備．任意の初期推測値$\hat{\boldsymbol{u}}_{\boldsymbol{0}}$を選ぶ．初期推測分布$\hat{P}_0 \in M$は$\boldsymbol{\theta}(\hat{\boldsymbol{u}}_{\boldsymbol{0}})$で与えられる．

1. eステップ．$\hat{P}_{\boldsymbol{i}}$をDにe-射影して，$K(Q||\hat{P}_{\boldsymbol{i}})$，$Q \in D$を最小化する$\hat{Q}_i$を与える．

2. mステップ．$\hat{Q}_{\boldsymbol{i}}$をMにm-射影して，$K(\hat{Q}_{\boldsymbol{i}}||P)$，$P \in M$を最小化する$\hat{P}_{i+1}$を与える．

パラメータが収束するまでEステップとMステップを繰り返す．

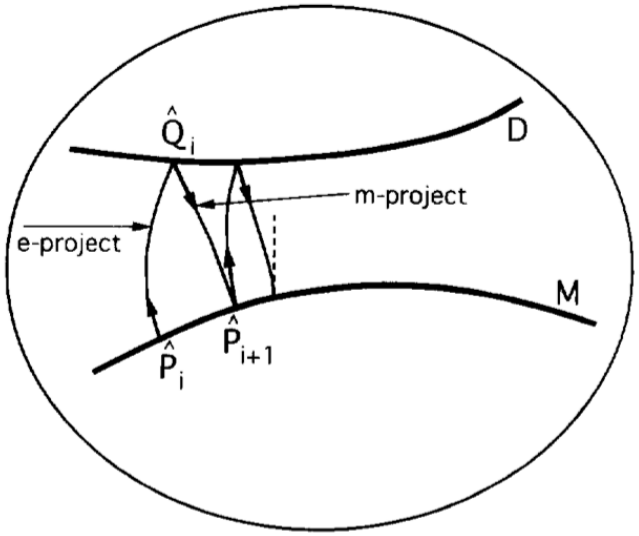


FIGURE 9. *em* algorithm.

e-射影とm-射影の特性

・定理1. $\hat{Q}$を観測点，MをS中のモデル多様体とすると，$\hat{Q}$のm-射影は尤度を最大化する最尤推定量$\theta(\hat{u}) \in M$を与える．このm-射影はMがe-平坦であるとき一意である．(Amari,1995)

・定理2. Dがm-平坦な部分多様体であるとき，PのDへのe-射影は一意である．またDをη座標で$D = \{\boldsymbol{\eta} | \boldsymbol{\eta} = (\boldsymbol{\eta_v}, \boldsymbol{\eta_h}), \boldsymbol{\eta_v} = \check{\boldsymbol{r}}_{\boldsymbol{v}}, \boldsymbol{\eta_h}$は任意$\}$に分離し，点$P$の$\theta$座標を$\boldsymbol{\theta^P} = (\boldsymbol{\theta_v^P}, \boldsymbol{\theta_h^P})$，$D$へのe-射影を$Q^*$の$\eta$座標，$\theta$座標をそれぞれ$(\boldsymbol{\eta_v^*}, \boldsymbol{\eta_h^*})$，$(\boldsymbol{\theta_v^*}, \boldsymbol{\theta_h^*})$とすると，以下が成り立つ．

1. Q^*の可視部分$\boldsymbol{\eta_v^*}$は$\boldsymbol{\eta_v^*} = \check{\boldsymbol{r}}_{\boldsymbol{v}}$で与えられる．
2. Q^*の$\boldsymbol{\theta_h^*}$はe-射影のもとで普遍に保たれる．$\boldsymbol{\theta_h^P} = \boldsymbol{\theta_h^*}$
3. Q^*における隠れ変数$\boldsymbol{r_h}$の条件付き確率はPにおける確率と等しい．

$$p(\boldsymbol{r_h} | \check{\boldsymbol{r}}_{\boldsymbol{v}}, P) = p(\boldsymbol{r_h} | \check{\boldsymbol{r}}_{\boldsymbol{v}}, Q^*)$$

4. Q^*における隠れ変数$\boldsymbol{r_h}$の条件付き期待値はPにおける期待値と等しい．

$$E_P[\boldsymbol{r_h} | \check{\boldsymbol{r}}_{\boldsymbol{v}}] = E_{Q^*}[\boldsymbol{r_h} | \check{\boldsymbol{r}}_{\boldsymbol{v}}]$$

e-射影とm-射影の特性

・証明
$Q^* \in D$なので1の$\boldsymbol{\eta}_v^* = \check{\boldsymbol{r}}_v$は成り立つ.

KLダイバージェンスは

$$\begin{aligned} K(Q||P) &= E_Q\left[\log\frac{p(\boldsymbol{r};\boldsymbol{\theta}_Q)}{p(\boldsymbol{r};\boldsymbol{\theta}_P)}\right] \\ &= E_Q\left[\left(\boldsymbol{\theta}_v^Q - \boldsymbol{\theta}_v^P\right)\cdot\boldsymbol{r}_v + \left(\boldsymbol{\theta}_h^Q - \boldsymbol{\theta}_h^P\right)\cdot\boldsymbol{r}_h\right] - \psi(\boldsymbol{\theta}_Q) + \psi(\boldsymbol{\theta}_P) \\ &= \left(\boldsymbol{\theta}_v^Q - \boldsymbol{\theta}_v^P\right)\cdot\check{\boldsymbol{r}}_v + \left(\boldsymbol{\theta}_h^Q - \boldsymbol{\theta}_h^P\right)\cdot\boldsymbol{\eta}_h^Q - \psi(\boldsymbol{\theta}_Q) + \psi(\boldsymbol{\theta}_P) \end{aligned}$$

部分多様体Dにおいて$K(Q||P)$を最小化するため，微分により

$$\frac{\partial}{\partial\boldsymbol{\eta}_h}K(Q||P) = \frac{\partial\boldsymbol{\theta}_v}{\partial\boldsymbol{\eta}_h}\cdot\check{\boldsymbol{r}}_v + \frac{\partial\boldsymbol{\theta}_h}{\partial\boldsymbol{\eta}_h}\cdot\boldsymbol{\eta}_h + \left(\boldsymbol{\theta}_h - \boldsymbol{\theta}_h^P\right) - \frac{\partial\psi(\boldsymbol{\theta}_Q)}{\partial\boldsymbol{\theta}_v}\frac{\partial\boldsymbol{\theta}_v}{\partial\boldsymbol{\eta}_h} - \frac{\partial\psi(\boldsymbol{\theta}_Q)}{\partial\boldsymbol{\theta}_h}\frac{\partial\boldsymbol{\theta}_h}{\partial\boldsymbol{\eta}_h} = 0$$

ここで，$\frac{\partial\psi(\boldsymbol{\theta}_Q)}{\partial\boldsymbol{\theta}_v} = \boldsymbol{\eta}_v = \check{\boldsymbol{r}}_v$，$\frac{\partial\psi(\boldsymbol{\theta}_Q)}{\partial\boldsymbol{\theta}_h} = \boldsymbol{\eta}_h$より，2の$\boldsymbol{\theta}_h^P = \boldsymbol{\theta}_h^*$が成立.

e-射影とm-射影の特性

・証明
Pにおける$\boldsymbol{r_h}$の条件付き確率は

$$
\begin{aligned}
p(\boldsymbol{r_h}|\check{\boldsymbol{r}}_v, \boldsymbol{\theta_P}) &= \frac{p(\boldsymbol{r_h}, \check{\boldsymbol{r}}_v; \boldsymbol{\theta_P})}{p(\check{\boldsymbol{r}}_v; \boldsymbol{\theta_P})} \\
&= \frac{\exp\{\boldsymbol{\theta_v^P} \cdot \check{\boldsymbol{r}}_v + \boldsymbol{\theta_h^P} \cdot \boldsymbol{r_h} - \psi\}}{\int \exp\{\boldsymbol{\theta_v^P} \cdot \check{\boldsymbol{r}}_v + \boldsymbol{\theta_h^P} \cdot \boldsymbol{r_h} - \psi\}\, d\boldsymbol{\mu}(\boldsymbol{r_h})} \\
&= \exp\{\boldsymbol{\theta_h^P} \cdot \boldsymbol{r_h} - \tilde{\psi}\}
\end{aligned}
$$

正規化係数$\tilde{\psi}$は$\boldsymbol{\theta_h^P}$と$\check{\boldsymbol{r}}_v$に依存し，$\boldsymbol{\theta_v^P}$に依存しない．また，PとQで$\boldsymbol{\theta_h}$は共通であるため，$p(\boldsymbol{r_h}|\check{\boldsymbol{r}}_v, P) = p(\boldsymbol{r_h}|\check{\boldsymbol{r}}_v, Q^*)$が示され，その結果$E_P[\boldsymbol{r_h}|\check{\boldsymbol{r}}_v] = E_{Q^*}[\boldsymbol{r_h}|\check{\boldsymbol{r}}_v]$となる．

EMアルゴリズムの幾何学的構造

・定理1，またはp.21により，Mステップとmステップが同じ手順で$\hat{\boldsymbol{u}}$を与えている．

・Eステップの幾何学的な解釈のため，$Q(\boldsymbol{\eta_v}, \boldsymbol{\eta_h})$に対し次の変換$\mathcal{F}: S \to S$を定義．

$$\mathcal{F}: (\boldsymbol{\eta_v}, \boldsymbol{\eta_h}) \to \left(\boldsymbol{\eta_v}, \boldsymbol{s_Q}(\boldsymbol{\eta_v})\right)$$

$$\boldsymbol{s_Q}(\boldsymbol{r_v}) = E_Q[\boldsymbol{r_h}|\boldsymbol{r_v}]$$

これは，$\boldsymbol{\eta_v}$は不変，$\boldsymbol{\eta_h}$は$\boldsymbol{r_v} = \boldsymbol{\eta_v}$を条件とした時の$\boldsymbol{r_h}$の期待値に置き換える変換．
$\boldsymbol{\eta_v} = \check{\boldsymbol{r}}_{\boldsymbol{v}}$を固定した時 $\mathcal{F}$はDを自身に写像する．

・EステップはPをDにe-射影し，得られたQ^*を$\mathcal{F}Q^*$に変換することを意味している．

よってEMアルゴリズムは次のようにかける．

定理3. EMアルゴリズムは次の2つのステップで表される．

1. Eステップ．$K\left(\mathcal{F}^{-1}(Q)||\hat{P}_i\right)$，$Q \in D$を最小化する$\hat{Q}_i \in D$を与える．
2. Mステップ．$K\left(\hat{Q}_i||P\right)$，$P \in M$を最小化する$\hat{P}_{i+1} \in M$を与える．

$\mathcal{F}$が恒等写像の時，EMとemのアルゴリズムが等しい．これは$Q \in D$での$\boldsymbol{r_h}$の条件付き期待値が$Q \in D$での $\boldsymbol{r_h}$の無条件期待値と等しい時である．

EMアルゴリズムの幾何学的構造

定理4. EMとemアルゴリズムは，条件付き期待値$\boldsymbol{s_Q(r_v)} = E_Q[\boldsymbol{r_h}|\boldsymbol{r_v}]$が任意の$Q \in D$においても$\boldsymbol{r_h}$に線形であるならば，等価である．

証明

$Q(\boldsymbol{\eta_v}, \boldsymbol{\eta_h}) \in D$の時，$\boldsymbol{\eta_v} = E_Q[\boldsymbol{r_v}]$，$\boldsymbol{\eta_h} = E_Q[\boldsymbol{r_h}] = E_Q E_Q[\boldsymbol{r_h}|\boldsymbol{r_v}] = E_Q\left[\boldsymbol{s_Q}(\boldsymbol{r_v})\right]$とおく．
$\boldsymbol{s_Q}(\boldsymbol{r_v})$が一次関数の時，$\boldsymbol{s_Q}(\boldsymbol{r_v}) = \boldsymbol{a} + \boldsymbol{B}\boldsymbol{r_v}$と定数$\boldsymbol{a}$，定数行列$\boldsymbol{B}$で記述される．
この時$\boldsymbol{\eta_h} = E_Q\left[\boldsymbol{s_Q}(\boldsymbol{r_v})\right] = \boldsymbol{s_Q}\left(E_Q(\boldsymbol{r_v})\right) = \boldsymbol{s_Q}(\boldsymbol{\eta_v})$より等価である．

逆に，$\mathcal{F}$が恒等写像であると仮定すると，$\boldsymbol{\eta_h} = E_Q\left[\boldsymbol{s_Q}(\boldsymbol{r_v})\right] = \boldsymbol{s_Q}\left(E_Q(\boldsymbol{r_v})\right)$が任意の$Q \in D$に対して成立．$Q = \{q(\boldsymbol{r_v}, \boldsymbol{r_h})\}$を$\eta_v = \eta_{v,0}$で定義される$D_0$に属する任意の分布，$P = \{p(\boldsymbol{r_v}, \boldsymbol{r_h})\}$を$D_0$へのe-射影が$Q$に等しくなるような任意の分布とする．$P$と$Q$を結ぶe-測地線上の分布の属$Q_t$，$0 \le t \le 1$を設定し，$Q_0 = Q$，$Q_1 = P$とする．$Q_t$は$S$における指数属で，$Q_t$のe-射影は$Q$であるため，すべての$Q_t$は同じ分布条件$q(\boldsymbol{r_v}, \boldsymbol{r_h})$を持つ．従って分布$Q_t = \{q_t(\boldsymbol{r_v}, \boldsymbol{r_h})\}$は$q_t(\boldsymbol{r_v}, \boldsymbol{r_h}) = p(\boldsymbol{r_v}; t)q(\boldsymbol{r_h}|\boldsymbol{r_v})$と表される．
$\boldsymbol{s_{Q_t}}(\boldsymbol{r_v})$は条件付き分布を通してのみ$Q$に依存するので，$\boldsymbol{s_{Q_t}}(\boldsymbol{r_v}) = \boldsymbol{s_Q}(\boldsymbol{r_v})$．
よって$\int s_Q(\boldsymbol{r_v})p(\boldsymbol{r_v}; t)d\boldsymbol{r_v} = s_Q\left[\int \boldsymbol{r_v}p(\boldsymbol{r_v}, t)d\boldsymbol{r_v}\right]$．$p$は条件$\int p(\boldsymbol{r_v}, t)d\boldsymbol{r_v} = 1$下で任意．
従って上の式は$\boldsymbol{s_Q}(\boldsymbol{r_v}) = \boldsymbol{a} + \boldsymbol{B}\boldsymbol{r_v}$と同様，$\boldsymbol{r_v}$の一次関数であることを示している．

EMとemアルゴリズムが異なる例：正規分布の乗法モデルの部分観測

・簡単のために$T=2$を考え，写像$\mathcal{F}$を探索する．

・ D上の点を$Q=N(\bar{r},\sigma^2)$とする．

・ $r_v=\frac{1}{2}(x_1+x_2)=\bar{r}$が与えられると条件付き確率分布$p_Q(x_1,x_2|\bar{r})$は正規化定数$c(\bar{r})$を用いて

$$p_Q(x_1,x_2|\bar{r})=c(\bar{r})\exp\left\{-\frac{(x_1-\bar{r})^2+(x_2-\bar{r})^2}{2\sigma^2}\right\}\times\delta\left(\frac{x_1+x_2}{2}-\bar{r}\right)$$

・これをx_2に関して積分し，$p_Q(x_1\,|\bar{r})=c\exp\left\{-\frac{(x_1-\bar{r})^2}{2\sigma^2}\right\}$．よって$\bar{r}$がわかると分散は$\frac{\sigma^2}{2}$に減少する．$x_2$も同様．

・よって$r_v=\bar{r}$を条件とするr_hの条件付き期待値は$s_Q(\bar{r})=E_Q\left[\frac{x_1^2+x_2^2}{2}\middle|\bar{r}\right]=\bar{r}^2+\frac{\sigma^2}{2}$
これは無条件の期待値$\eta_h=\bar{r}^2+\sigma^2$とは異なる．

・これより$\mathcal{F}$は恒等写像ではなく，$Q=N(\bar{r},\sigma^2)$を$\mathcal{F}Q=N\left(\bar{r},\frac{\sigma^2}{2}\right)$に写像する．

EMとemアルゴリズムが異なる例：正規分布の乗法モデルの部分観測

・ $Q = N(u, u^2) \in M$のDへのe-射影Q^*は， $\theta_v = \frac{2}{u}$， $\theta_h = \frac{1}{u^2}$より$T = 2$で$\eta_v^* = \bar{r}$，
$\theta_h^* = \theta_h^P = -u^2$が成り立つため， $Q^* = N(\bar{r}, u^2)$.

・一方，Eステップでは$E_p[r_h | r_v = \bar{r}] = \bar{r}^2 + \frac{u^2}{2}$より$\hat{Q} = N\left(\bar{r}, \frac{u^2}{2}\right)$.

これは，$\mathcal{F}Q^* = \hat{Q}$を表す.

・emアルゴリズムは$P^* = N(\bar{r}, \bar{r}^2)$に収束し，最適な解$\hat{u}^* = \bar{r}$を幾何学的に求めるが
EMアルゴリズムは最尤推定量$\hat{u}$に収束する． $\hat{u} = (\sqrt{3} - 1)\bar{r} \approx 0.73\bar{r}$
これはKLダイバージェンスの最小値ではない

・一般に， $N > 2$， $Q^* = N(\bar{r}, u^2)$に対して
$\mathcal{F}Q^* = N\left(\bar{r}, \frac{T-1}{T}u^2\right)$

・emの解$\hat{u}^* = \bar{r}$に対してEMの解は
$\hat{u} = \frac{1}{2}\left(\sqrt{T^2 + 4T} - T\right)\bar{r} \approx \left(1 - \frac{1}{T}\right)\bar{r}$

・よってTが大きいとき，漸近的に等価であり，現実的には差はない.

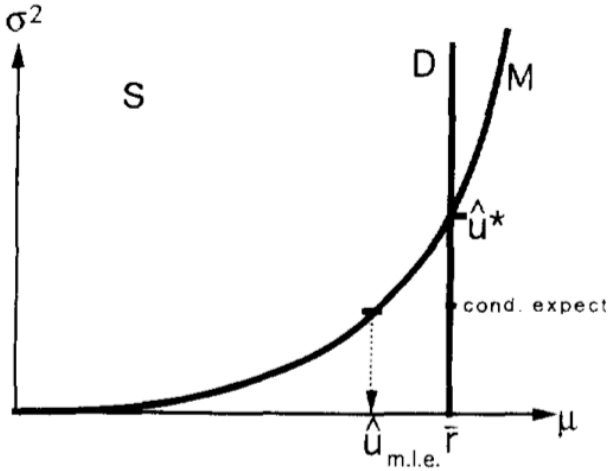


FIGURE 10. Example where *EM* and *em* algorithms are different.

EMとemのどちらが自然か

・ $\mathcal{F}$が恒等写像でない場合，異なる推定値を与える．
・EMは最尤推定量を与えるので，統計学的には自然．
・NNで教師データを近似する場合，統計的な推論ではない．その振る舞いはデータ多様体Dにまとめられるため，ダイバージェンスを最小化するemが自然．

・指数型分布族から関数空間へ枠組みを拡張することにより，二つのアルゴリズムが等価となる．
・また，指数型分布族でもTが大きい場合に漸近的に等価であることを証明する．
・またこの時，条件付き期待値を計算するEステップの方がe-射影より計算量が大きい可能性がある．

EMとemの等価性を保証する線形化トリック

・$\boldsymbol{s_Q(r_v)}$が線形である時，等価となる．
・拡張フレームワークで等価であることを示す．
・yを0か1の値をとる確率変数，$f(y)$をyに対して線形な関数とする

$$f(y) = \{f(1) - f(0)\}y + f(0)$$

・この方法は有限集合K上で値をとる変数であれば，どんな関数でも使える．
Kの要素をmとし，mを指標とする新しいベクトル変数$\boldsymbol{k(r)}$を導入する．

$$k(\boldsymbol{r}) = \{k_m(\boldsymbol{r}), m \in K, k_{\boldsymbol{m}}(\boldsymbol{r}) = \delta(\boldsymbol{r} - \boldsymbol{m})\}$$

・$f(\boldsymbol{r}) = \sum_m f(\boldsymbol{m}) k_{\boldsymbol{m}}(\boldsymbol{r})$ より，$f(\boldsymbol{r})$は確率変数$\boldsymbol{k(r)}$に対して線形

・よって可視確率変数が2値の場合，可視確率変数が拡張ベクトル$k(\boldsymbol{r}) = \{k_m(\boldsymbol{r})\}$の場合，条件付き期待値は$\boldsymbol{k(r)}$に対して線形であり，EMとemが等価．

→2値確率パーセプトロン、エキスパートネットの混合、ボルツマンマシン、正規混合モデルにおいて、EMアルゴリズムとemアルゴリズムが等価であることがわかる

EMとemの等価性を保証する線形化トリック

・連続変数の場合にも線形化のトリックを用いる.

$$f(\boldsymbol{r}) = \int f(\boldsymbol{m})\delta(\boldsymbol{r}-\boldsymbol{m})d\boldsymbol{m}$$

・この場合，確率変数$\boldsymbol{r}$は一般化された関数に拡張され，$\boldsymbol{r}$は一般化関数空間の要素.

・T個のサンプル$\boldsymbol{r_1},\ldots,\boldsymbol{r_T}$が観測された時，これらは経験分布に要約される.

$$\hat{p}_{emp}(\boldsymbol{r}) = \frac{1}{T}\sum_{t=1}^{T}\delta(\boldsymbol{r}-\boldsymbol{r_t})$$

・隠れ変数を関数空間で定式化し，EMとemが関数空間で等価であることを証明する
$S=\{p(\boldsymbol{r_v},\boldsymbol{r_h})\}$を確率変数$\boldsymbol{r}=(\boldsymbol{r_v},\boldsymbol{r_h})$の密度関数の関数空間とする.
$\boldsymbol{u}$で指定されるパラメトリックモデル$M=\{p(\boldsymbol{r_v},\boldsymbol{r_h};\boldsymbol{u})\}$を考え，$\boldsymbol{r}^*_{\boldsymbol{v,T}}=(\boldsymbol{r}_{\boldsymbol{v,1}},\ldots,\boldsymbol{r}_{\boldsymbol{v,T}})$が観測された時，関数空間にまとめると

$$\hat{p}_{emp}(\boldsymbol{r_v}) = \frac{1}{T}\sum_{t=1}^{T}\delta(\boldsymbol{r_v}-\boldsymbol{r}_{\boldsymbol{v,t}})$$

これは観測データに基づく経験分布.

7. 拡張フレームワークにおけるEMとemの等価性

EMとemの等価性を保証する線形化トリック

$p(\boldsymbol{r_v}, \boldsymbol{r_h}) = p(\boldsymbol{r_v})p(\boldsymbol{r_h}|\boldsymbol{r_v})$と分解．$p(\boldsymbol{r_v})$は経験分布で与えられるが$p(\boldsymbol{r_h}|\boldsymbol{r_v})$は自由．
従って観測データ多様体Dは関数空間において$D = \{\hat{p}_{emp}(\boldsymbol{r_v})p(\boldsymbol{r_h}|\boldsymbol{r_v})\}$
$p(\boldsymbol{r_v}, \boldsymbol{r_h}; \boldsymbol{u})$の$D$へのe-射影は$K[q(\boldsymbol{r_v}, \boldsymbol{r_h})||p(\boldsymbol{r_v}, \boldsymbol{r_h}; \boldsymbol{u})]$，$q(\boldsymbol{r_v}, \boldsymbol{r_h}) \in D$を最小化するものである．e-射影の$Q^*$が次式で表せることを示せば良い．

$$\begin{cases} T = 1\text{の時}：q^*(\boldsymbol{r_v}, \boldsymbol{r_h}) = \hat{p}_{emp}(\boldsymbol{r_v})p(\boldsymbol{r_h}|\boldsymbol{r_v}; \boldsymbol{u}) \\ \text{一般的な}T：\prod_{t=1}^{T} \hat{p}_{emp}(\boldsymbol{r_{v,t}})p(\boldsymbol{r_{h,t}}|\boldsymbol{r_v}; \boldsymbol{u}) \end{cases}$$

隠れた確率変数を関数$\delta(\boldsymbol{r_h} - \boldsymbol{m})$に拡張すると，その条件付き確率は

$$E[\delta(\boldsymbol{r_h} - \boldsymbol{m})|\boldsymbol{r_v}; \boldsymbol{u}] = \int \delta(\boldsymbol{r_h} - \boldsymbol{m})p(\boldsymbol{m}|\boldsymbol{r_v}; \boldsymbol{u})d\boldsymbol{m} = p(\boldsymbol{r_h}|\boldsymbol{r_v}; \boldsymbol{u})$$

よって二つのアルゴリズムは同じである．

EMとemの等価性を保証する線形化トリック

定理5．EMとemは関数空間において等価である．その手順は

E(e)ステップ．条件付き確率$p\left(\boldsymbol{r_h}\middle|\boldsymbol{r_v};\widehat{\boldsymbol{u}}_{(i)}\right)$を求める．

M(m)ステップ．$E_{Q_i}[\log p(\boldsymbol{r_h}|\boldsymbol{r_v};\boldsymbol{u})] = \sum_{t=1}^{T}\int p\left(\boldsymbol{r_{h,t}}\middle|\boldsymbol{r_{v,t}};\widehat{\boldsymbol{u}}_{(i)}\right)\log p\left(\boldsymbol{r_{h,t}}\middle|\boldsymbol{r_{v,t}};\boldsymbol{u}\right)d\boldsymbol{r_{h,t}}$を最大化する$\boldsymbol{u}$を探索し，$\widehat{\boldsymbol{u}}_{(i+1)}$とする．

ただしこの方法では，$\boldsymbol{r_{v,1}},\ldots,\boldsymbol{r_{v,T}}$を統計量にまとめることなく全て保持する必要がある．

例-非線形確率パーセプトロン

・すべてのNNが指数関数的な族だとは限らない.

・ fをシグモイド関数$f(u)=\frac{1}{1+\exp(-u)}$とする.

・ i番目の隠れユニット，最終出力はそれぞれ

$$z_i = f(\boldsymbol{w_i}\cdot\boldsymbol{x}) + n_i,\quad y = (\boldsymbol{v}\cdot\boldsymbol{z}) + n$$

$\boldsymbol{x}$：入力
$\boldsymbol{z}$：隠れユニットの出力
y：最終出力
n_i, n：それぞれ$N(0,\sigma^2)$に従う独立したランダムノイズ

・この時の確率は次のようになる.

$$p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{u}) = c\exp\left[-\frac{1}{2\rho^2}\sum\{z_i - f(\boldsymbol{w_i}\cdot\boldsymbol{x})\}^2\right]$$

$$p(y|\boldsymbol{z};\boldsymbol{u}) = c'\exp\left[-\frac{1}{2\rho^2}\{y - f(\boldsymbol{v}\cdot\boldsymbol{z})\}^2\right]$$

従って共同確率分布の対数は

$$\rho^2 l(y,\boldsymbol{z}|\boldsymbol{x};\boldsymbol{u}) = \sum z_i f(\boldsymbol{w_i}\cdot\boldsymbol{x}) + y f(\boldsymbol{v}\cdot\boldsymbol{z}) - \frac{1}{2}\{f(\boldsymbol{v}\cdot\boldsymbol{z})\} + k(y,\boldsymbol{z}) - \psi$$

これは$\theta(\boldsymbol{u},\boldsymbol{x})$と$r$の形でまとめることができず，$p(y,\boldsymbol{z}|\boldsymbol{x};\boldsymbol{u})$は指数型分布族に属さないが，関数の多様体を導入することで，情報幾何学とEMアルゴリズムが適用できる.

例-非線形確率パーセプトロン

・確率的出力z_i, yをそれぞれ期待値に置き換えることで通常のアナログパーセプトロンに還元される.
よって通常の多層パーセプトロンッモ確率的手法で学習できる.

$$\boldsymbol{f}_t = (f_{1,t}, \cdots, f_{k,t})$$
$$f_{j,t} = f(\hat{\boldsymbol{w}}_j \cdot \boldsymbol{x}_t)$$

・EMアルゴリズムは以下のとおり

Eステップ. 時刻tの観測値(y_t, x_t)に基づき, $\boldsymbol{z_t}$の分布を計算

$$p(\boldsymbol{z_t}|y_t, \boldsymbol{x_t}; \hat{\boldsymbol{u}}_i) = \frac{\exp\left\{-\frac{1}{2\sigma^2}[|\boldsymbol{z}_t - \boldsymbol{f}_t|^2 + \{y_t - f(\hat{\boldsymbol{v}}_i \cdot \boldsymbol{z}_t)\}^2]\right\}}{\int \exp\left\{-\frac{1}{2\sigma^2}[|\boldsymbol{z}_t - \boldsymbol{f}_t|^2 + \{y_t - f(\hat{\boldsymbol{v}}_i \cdot \boldsymbol{z}_t)\}^2]\right\} d\boldsymbol{z}_t}$$

Mステップ. 以下の式を最大化する$\boldsymbol{u}$を求め, $\hat{\boldsymbol{u}}_{i+1}$とする.

$$\sum_t \int p(\boldsymbol{z_t}|y_t, \boldsymbol{x_t}; \hat{\boldsymbol{u}}_i) \log p(\boldsymbol{z_t}|y_t, \boldsymbol{x_t}; \boldsymbol{u})\, d\boldsymbol{z_t}$$

これは, 最尤推定量に収束する.

8. 学習手順

・EMとemは時間tごとに部分観測$\boldsymbol{r_{v,t}}$または$\boldsymbol{s_{v,t}}$が利用できる場合，バッチアルゴリズムではデータ$\boldsymbol{r_{v,t}}$，$t = 1, \ldots, T$を使用するのに対し，オンライン学習形式で書き直すことができる．
・学習は一般に収束は遅くなるが，アルゴリズムは単純．
・環境構造の変化に対してロバスト．

・全データがSにおける観測点$\hat{\boldsymbol{\eta}}$に集約される場合の学習アルゴリズムを考える．

1.eステップ(Eステップ)．現在の$\hat{P}_t$のD_{t+1}へのe-射影($\hat{\boldsymbol{u}}_t$に基づく$\boldsymbol{s_{v,t+1}}$を条件とした$\hat{\boldsymbol{r}}_{t+1}$の条件付き期待値)を計算．$\hat{\boldsymbol{r}}_{t+1}$が得られ，これを次の式で修正．

$$\hat{\boldsymbol{\eta}}_{t+1} = (1 - \varepsilon_t)\hat{\boldsymbol{\eta}}_t + \varepsilon_t \hat{\boldsymbol{r}}_{t+1}$$

2.Mステップ．$\hat{\boldsymbol{\eta}}_{t+1}$から最尤推定量を計算し，$\hat{\boldsymbol{u}}_{t+1}$とすると，$\hat{P}_{t+1}$が求まる．

$\hat{\boldsymbol{u}}_t$：時刻t における推定量
$\hat{\boldsymbol{\eta}}_t$：時刻$t$ における観測点の推定
$\boldsymbol{s_{v,t+1}}$：時刻$t+1$における部分観測点
ε_t：定数または$\varepsilon_t \sim \frac{c}{t}$のような減少数列

8. 学習手順

・最尤推定の計算は通常簡単ではなく，勾配法を用いる.
・ $\hat{\boldsymbol{\eta}}_{t+1}$の対数尤度は$l = \boldsymbol{\theta}(\boldsymbol{u}) \cdot \hat{\boldsymbol{\eta}}_{t+1} - \psi(\boldsymbol{u})$であり，その勾配は

$$\frac{\partial l}{\partial \boldsymbol{u}} = \boldsymbol{B}\{\hat{\boldsymbol{\eta}}_{t+1} - \boldsymbol{\eta}(\boldsymbol{u})\}$$

ただし， $\boldsymbol{B}$は行列$\boldsymbol{B} = \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{u}}$.

・従って増分Mステップにおける勾配法は

$$\hat{\boldsymbol{u}}_{t+1} = \hat{\boldsymbol{u}}_t + \varepsilon_t \boldsymbol{B}\{\hat{\boldsymbol{\eta}}_{t+1} - \boldsymbol{\eta}(\hat{\boldsymbol{u}}_t)\}$$

よってどの加速度法でも適用可能.

・スコアリング法では$\hat{\boldsymbol{u}}_{t+1} = \hat{\boldsymbol{u}}_t + \varepsilon_t G_t^* \boldsymbol{B}\{\hat{\boldsymbol{\eta}}_{t+1} - \boldsymbol{\eta}(\hat{\boldsymbol{u}}_t)\}$
G_t^*は$\hat{\boldsymbol{u}}_t$におけるフィッシャー情報行列の逆行列

$\hat{\boldsymbol{u}}_t$：時刻tにおける推定量
$\hat{\boldsymbol{\eta}}_t$：時刻$t$における観測点の推定
$\boldsymbol{s}_{v,t+1}$：時刻$t+1$における部分観測点
ε_t：定数または$\varepsilon_t \sim \frac{c}{t}$のような減少数列

8. 学習手順

・観測データが一つの$\hat{\boldsymbol{\eta}}$にまとめられず，積空間S_T^*を考慮する必要がある場合

1.eステップ(Eステップ)．　θ座標を$(\boldsymbol{x}_{t+1}, \boldsymbol{u})$とするの$\hat{P}_t$の$D_{t+1}$へのe-射影($\hat{\boldsymbol{r}}_{\boldsymbol{t+1}}$の条件付き期待値)を計算．　$\hat{\boldsymbol{r}}_{\boldsymbol{t+1}}$が得られ，これを次の式で修正.

2.増分Mステップ．　$\hat{\boldsymbol{u}}_{t+1} = \hat{\boldsymbol{u}}_t + \varepsilon \boldsymbol{B}(\boldsymbol{x}_{t+1}, \hat{\boldsymbol{u}}_t)\{\hat{\boldsymbol{r}}_{t+1} - \boldsymbol{\eta}(\hat{\boldsymbol{u}}_t, \boldsymbol{x}_t)\}$

この場合，対数尤度は$l = \sum_t \{\boldsymbol{\theta}(\boldsymbol{x}_t, \boldsymbol{u}) \cdot \hat{\boldsymbol{r}}_t - \psi(\boldsymbol{x}_t, \boldsymbol{u})\}$

eステップでは観測変数$\boldsymbol{s}_{\boldsymbol{v},t+1}$と$\hat{\boldsymbol{u}}_{\boldsymbol{t}}$に基づく$\boldsymbol{x}_{t+1}$から$\hat{\boldsymbol{r}}_{\boldsymbol{t+1}}$を推定
増分Mステップでは古いデータは破棄し，最新データ$\hat{\boldsymbol{r}}_{\boldsymbol{t+1}}$のみから勾配を計算

$\hat{\boldsymbol{u}}_{\boldsymbol{t}}$：時刻$t$ における推定量
$\hat{\boldsymbol{\eta}}_{\boldsymbol{t}}$：時刻$t$ における観測点の推定
$\boldsymbol{s}_{\boldsymbol{v},t+1}$：時刻$t+1$における部分観測点
ε_t：定数または$\varepsilon_t \sim \frac{c}{t}$のような減少数列

例-確率的パーセプトロンの学習

・EMとemは同じ．E(e)ステップにおいて，観測されたy_{t+1}を条件とする期待値は

$$\hat{r}_{2,\mathbf{k}} = E[\delta_{\mathbf{k}}(\mathbf{z})] = \mathrm{Prob}\{z = \mathbf{k} \mid y_{t+1}, \mathbf{u}_t\}$$
$$= \frac{\varphi(y_{t+1}, \mathbf{k}\cdot\mathbf{v}_t)\Pi\,\varphi(k_i, \mathbf{x}\cdot\mathbf{w}_{i,t})}{\sum_{\mathbf{k}}\varphi(y_{t+1}, \mathbf{k}\cdot\mathbf{v}_t)\Pi\,\varphi(k_i, \mathbf{x}\cdot\mathbf{w}_{i,t})}$$

$$\hat{r}_{1,\mathbf{k}} = \begin{cases} 0, & y_{t+1} = 0 \\ \hat{r}_{2,\mathbf{k}}, & y_{t+1} = 1. \end{cases}$$

Mステップは最尤推定の計算である．増分アルゴリズムでは

$$l = \boldsymbol{\theta}(\boldsymbol{u})\cdot\hat{\boldsymbol{r}} - \psi\big(\boldsymbol{\theta}(\boldsymbol{u})\big)$$

$\frac{\partial l}{\partial \boldsymbol{u}}$を計算し，

$$\frac{\partial l}{\partial v} = \sum_{\mathbf{k}} \hat{r}_{2,\mathbf{k}}\{y - \varphi(1\cdot\mathbf{k}\cdot\mathbf{v})\}\mathbf{k},$$
$$\frac{\partial l}{\partial \mathbf{w}_i} = \sum_{\mathbf{k}} \delta(k_i)\hat{r}_{2,\mathbf{k}}\mathbf{x} - \frac{\mathbf{x}\exp(\mathbf{w}_i\cdot\mathbf{x})}{1+\exp(\mathbf{w}_i\cdot\mathbf{x})}.$$

$\hat{\boldsymbol{u}}_{t+1} = \hat{\boldsymbol{u}}_t + \varepsilon_t\frac{\partial l}{\partial \boldsymbol{u}}$で$\hat{\boldsymbol{u}}_{t+1}$が更新できる．

例-確率的パーセプトロンの学習

・注意点1.ノイズのない（決定論的）ネットワークで y = 0 または 1 で例を生成した場合、 $|w_i|$と|v|は学習によりすべて無限大に発散する。このような場合、正規化によってそれらの大きさを注意深く制御する必要がある。そこで、このような場合にも学習は有効である。

・注意点2. y=0,1 を出力とする確率的パーセプトロンを用いてアナログ入出力関係 $\bar{y} = f(\boldsymbol{x})$ を学習する場合、例 $(\bar{y}_1, \boldsymbol{x}_1), \ldots, (\bar{y}_T, \boldsymbol{x}_T)$の$\bar{y}_t$の値はバイナリではなく実数である。このような場合の$\bar{y}_t$は、 $\boldsymbol{x}_t$が入力されたとき、 $\bar{y}_t$がyt=1の確率であると解釈される。つまり、 $\boldsymbol{x}_t$が多数回入力されたとき、yt=1 が相対的な頻度$\bar{y}_t$で発生すると解釈する。条件付き期待値は次式で与えられる。

$$\hat{r}_{1,k} = \bar{y}_{t+1} E[\delta_k(\mathbf{z}) | y_{t+1} = 1, \mathbf{u}_t],$$

$$\hat{r}_{2,k} = \bar{y}_{t+1} E[\delta_k(\mathbf{z}) | y_{t+1} = 1, \mathbf{u}_t] + (1 - \bar{y}_{t+1}) E[\delta_k(\mathbf{z}) | y_{t+1} = 0, \mathbf{u}_t].$$

11. 結論

・EMアルゴリズムとemアルゴリズムの情報幾何学的知見が構築された.

・両者はTが大きくなると漸近的に等価になり，実用的なほとんどのケースで等価.

・拡張関数空間においても等価

・これによりEMの学習過程を定式化することが可能になった.