・・・・・2.5・・・・・

最尤法とその計算アルゴリズム

ここでは，最尤推定値を数値計算で求める方法を 3 つ紹介する．

2.5.1 ニュートン・ラプソン法

いま，$g : \mathbb{R} \to \mathbb{R}$ を 2 階微分可能な関数とし，方程式 $g(x) = 0$ をみたす解 $x = c$ をみつけたい．そのために，c に近い x に対して，テーラー展開をする：

$$0 = g(c) \approx g(x) + \dot{g}(x)(x - c)$$

ただし，$\dot{g}(x) = dg/dx$ である．$\dot{g}(x) \neq 0$ のとき，これを c について解けば

$$c \approx x - \frac{g(x)}{\dot{g}(x)}$$

を得る．初期値 x_0 を取り，点列 $\{x_n\}_{n=1}^{\infty}$ を逐次的に

$$x_{n+1} = x_n - \frac{g(x_n)}{\dot{g}(x_n)}, \qquad n = 0, 1, \ldots \tag{2.7}$$

で定義する．そして，$|g(x_n)/\dot{g}(x_n)|$ が十分小さくなるまで操作を繰り返すとする．区間 $I = [a, b]$ 上で $\ddot{g}(x) > 0$ で，$g(a)g(b) < 0$ のとき，$g(x_0) > 0$ となる $x_0 \in (a, b)$ をひとつ見つければ，(2.7) で得られる点列 $\{x_n\}_{n=1}^{\infty}$ は I 上における零点 c に収束することが知られている[1)]．ニュートン・ラプソン法を用いて，尤度方程式の解として定義される最尤推定値を求めることができる．

いま，$\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ は同時確率密度関数または確率関数 $p(\boldsymbol{x}|\,\theta)$ を持つとする．ただし，$\theta \in \Theta \subset \mathbb{R}$ とする．$\boldsymbol{X} = \boldsymbol{x}$ が与えられたときの尤度関数は $L_n(\theta|\,\boldsymbol{x}) = p(\boldsymbol{x}|\,\theta)$ である．最尤推定値 $\widehat{\theta}$ は尤度方程式

$$S(\theta) = \frac{d}{d\theta} \log L_n(\theta|\,\boldsymbol{x}) = 0 \tag{2.8}$$

の解とする．さらに，k 回操作を行ったのちの θ の推定値を $\widehat{\theta}^{(k)}$ とする：

$$\widehat{\theta}^{(k+1)} = \widehat{\theta}^{(k)} + \frac{S(\widehat{\theta}^{(k)})}{H(\widehat{\theta}^{(k)})}$$

ただし，

$$H(\theta) = -\frac{d^2}{d\theta^2} \log L_n(\theta)$$

1) 杉浦「解析入門 I (東京大学出版会)」p.105 を参照.

である．この操作を $\widehat{\theta}^{(k+1)}$ と $\widehat{\theta}^{(k)}$ との差が十分小さくなるまで繰り返す.

ニュートン・ラプソン法を用いるために初期推定値 $\widehat{\theta}^{(0)}$ が必要である．初期推定値に何を用いるかによって，アルゴリズムは収束したりしなかったりする．また，尤度方程式 $S(\theta)=0$ が複数の解を持つ場合には，尤度方程式 (2.8) の解は尤度関数の極小点，極大点，鞍馬点 (saddle-point) に対応するので，$\widehat{\theta}^{(k)}$ は最尤推定値とは異なる点に収束する可能性がある．収束先が最尤推定値と異なるかどうかが不明な場合には，複数の初期値で試すとよい．また，初期値として，別の推定値 (別の推定量の実現値) を用いることもできる．たとえば，$\widehat{\theta}_n^{(0)}(\boldsymbol{X})$ が θ の十分よい推定量ならば，一段階推定量

$$\widehat{\theta}_n^{(1)} = \widehat{\theta}_n^{(0)} + \frac{S(\widehat{\theta}_n^{(0)})}{H(\widehat{\theta}_n^{(0)})}$$

は，最尤推定量と同じ性質を漸近的には同じ性質をもつことが知られている．正確に言えば，$\sqrt{n}(\widehat{\theta}_n^{(0)}-\theta)$ は正規分布に分布収束するならば，$\sqrt{n}(\widehat{\theta}_n^{(1)}-\widehat{\theta}_n) \xrightarrow{P} 0$ となる．ただし，$\widehat{\theta}_n$ は θ の最尤推定量である．

例 2.17　　確率密度関数

$$f_X(x|\,\theta) = \frac{1}{\pi\{1+(x-\theta)^2\}} I_{(-\infty,\infty)}(x), \qquad (-\infty<\theta<\infty)$$

を持つコーシー分布からの大きさ n のランダム標本を $X_1, X_2, \ldots, X_n$ とする．このとき，実現値 $x_1, x_2, \ldots, x_n$ に対する対数尤度関数は

$$\log L_n(\theta) = -\sum_{i=1}^n \log\{1+(x_i-\theta)^2\} - n\log\pi$$

となる．最尤推定値 $\widehat{\theta}_n$ は尤度方程式

$$S(\widehat{\theta}_n) = \sum_{i=1}^n \frac{2(x_i-\widehat{\theta}_n)}{1+(x_i-\widehat{\theta}_n)^2} = 0$$

の解である．$S(\theta)$ は θ の単調関数でないので，与えられた $(x_1, x_2, \ldots, x_n)$ に対して，尤度方程式は複数の解を持つ可能性がある．したがって，適切な初期値 $\widehat{\theta}^{(0)}$ を選ぶことが重要である．コーシー分布は $\mathbb{E}[X_1]$ が定義されないので，初期値として，標本平均を用いるのは適当ではない．X_1 の分布は θ に関して対称であることに注目して，標本中央値を初期値 $\widehat{\theta}^{(0)}$ として用いる．これを用いて，逐次的に $\widehat{\theta}^{(k)}, k=1, 2, \ldots$ を

$$\widehat{\theta}^{(k)} = \widehat{\theta}^{(k-1)} + \frac{S(\widehat{\theta}^{(k-1)})}{H(\widehat{\theta}^{(k-1)})}$$

で定める．ただし，

$$H(\theta) = 2\sum_{i=1}^n \frac{1-(x_i-\theta)^2}{\{1+(x_i-\theta)^2\}^2}$$

である．$\theta=10$ のコーシー分布から標本の大きさ $n=100$ のランダム標本に基づいて最尤推定値を求めた例が次である．

k	$\widehat{\theta}^{(k)}$	$logL_n(\widehat{\theta}^{(k)}) + 100\log(\pi)$
0	9.932387	11.95144
1	9.98055	11.9517
2	9.980323	11.9517
3	9.980323	11.9517

つぎに，母数の次元が p の場合を考える．母数 $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_p)'$ が p - 次元のとき，$\boldsymbol{\theta}$ の最尤推定値 $\widehat{\boldsymbol{\theta}}$ は尤度方程式 $\boldsymbol{S}(\boldsymbol{\theta}) = \boldsymbol{0}$ の解として定義される．ただし，

$$\boldsymbol{S}(\boldsymbol{\theta}) = \left(\frac{\partial}{\partial\theta_1}\log L_n(\boldsymbol{\theta}),\ \frac{\partial}{\partial\theta_2}\log L_n(\boldsymbol{\theta}),\ \dots,\ \frac{\partial}{\partial\theta_p}\log L_n(\boldsymbol{\theta}) \right)'$$

である．このとき，k 回目の逐次回 $\widehat{\boldsymbol{\theta}}^{(k)}$ は

$$\widehat{\boldsymbol{\theta}}^{(k)} = \widehat{\boldsymbol{\theta}}^{(k-1)} + [H(\widehat{\boldsymbol{\theta}}^{(k-1)})]^{-1} S(\widehat{\boldsymbol{\theta}}^{(k-1)}) \tag{2.9}$$

で定義される．ただし，$H(\boldsymbol{\theta})$ の (i, j) - 成分は

$$H_{ij}(\boldsymbol{\theta}) = -\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log L_n(\boldsymbol{\theta})$$

で定義される．

2.5.2 Fisher のスコア法

Newton-Raphson アルゴリズムの簡単な修正として，Fisher のスコアアルゴリズムがある．Fisher のスコアアルゴリズムは (2.9) の中の H の代わりに

$$H^*(\boldsymbol{\theta}) = \mathbb{E}_\theta[H(\boldsymbol{\theta})] = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\log L_n(\boldsymbol{\theta})\right]$$

である．ただし，

$$\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\log L_n(\boldsymbol{\theta})$$

の (i, j) - 成分は

$$\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log L_n(\boldsymbol{\theta})$$

である．したがって，k 回目の逐次解 $\widehat{\boldsymbol{\theta}}^{(k-1)}$ は

$$\widehat{\boldsymbol{\theta}}^{(k)} = \widehat{\boldsymbol{\theta}}^{(k-1)} + \left[H^*(\widehat{\boldsymbol{\theta}}^{(k-1)})\right]^{-1}\boldsymbol{S}(\widehat{\boldsymbol{\theta}}^{(k-1)})$$

で定義される．

例 2.18 (例 2.17 の続き)

$$H(\theta) = 2\sum_{i=1}^{n}\frac{1-(x_i-\theta)^2}{\{1+(x_i-\theta)^2\}^2}$$

から

$$H^*(\theta) = \frac{2n}{\pi} \int_{-\infty}^{\infty} \frac{1-(x-\theta)^2}{\{1+(x-\theta)^2\}^3}\, dx = \frac{n}{2} \tag{2.10}$$

から Fisher のスコアアルゴリズムは

$$\widehat{\theta}^{(k)} = \widehat{\theta}^{(k-1)} + \frac{4}{n}\sum_{i=1}^{n} \frac{x_i - \widehat{\theta}^{(k-1)}}{1+(x_i-\widehat{\theta}^{(k-1)})^2}$$

となる．最後に，(2.10) の計算をする．$z = x - \theta$ とおき，さらに $y = \tan\gamma$ とおけば，

$$\begin{aligned}
\int_{-\infty}^{\infty} \frac{1-y^2}{(1+y^2)^3}\, dy &= \int_{-\infty}^{\infty} \frac{2}{(1+y^2)^3}\, dy - \int_{-\infty}^{\infty} \frac{1}{(1+y^2)^2}\, dy \\
&= 2\int_{-\pi/2}^{\pi/2} \frac{1}{(1+\tan\gamma^2)^3}\frac{1}{\cos\gamma^2}\, d\gamma - \int_{-\pi/2}^{\pi/2} \frac{1}{(1+\tan\gamma^2)^2}\frac{1}{\cos\gamma^2}\, d\gamma \\
&= 2\int_{-\pi/2}^{\pi/2} \cos^4\gamma\, d\gamma - \int_{-\pi/2}^{\pi/2} \cos^2\gamma\, d\gamma, \\
&= 2\int_{-\pi/2}^{\pi/2} \left(\frac{\cos(4\gamma)+1}{8} + \cos(2\gamma) + \frac{1}{4}\right) d\gamma - \int_{-\pi/2}^{\pi/2} \left(\frac{\cos(2\gamma)+1}{2}\right) d\gamma \\
&= 2\left[\frac{\sin(4\gamma)}{32} + \frac{\gamma}{8} + \frac{\sin(2\gamma)}{2} + \frac{\gamma}{4}\right]_{-\pi/2}^{\pi/2} - \left[\frac{\sin(2\gamma)}{4} + \frac{\gamma}{4}\right]_{-\pi/2}^{\pi/2} \\
&= \frac{\pi}{4}
\end{aligned}$$

となることがわかる．

2.5.3 EM アルゴリズム

$(\mathcal{X}, \mathcal{A}, \mu)$ を測度空間とする．ここで，μ は σ - 有限な測度[2)]とする．$\mathcal{X}$ - 値確率変数 $\boldsymbol{X}$ は母数 $\theta\,(\theta \in \Theta)$ の確率測度 P_θ を持つ：

$$\mathbb{P}(\{\boldsymbol{X} \in A\}) = P_\theta(A), \qquad A \in \mathcal{A}$$

さらに，P_θ は μ に関する確率密度関数 $p(\cdot\,|\,\theta)$ をもつとする：

$$P_\theta(A) = \int_A p(\boldsymbol{x}|\,\theta)\,\mu(d\boldsymbol{x}), \qquad A \in \mathcal{A}$$

いま，$(\mathcal{X}, \mathcal{A})$ は隠れた空間で $\boldsymbol{X}$ のすべてを観測できないとする．実際には，ある可測空間 $(\mathcal{Y}, \mathcal{B})$ と可測関数 $T : \mathcal{X} \mapsto \mathcal{Y}$ が存在して，$\boldsymbol{Y} = T(\boldsymbol{X})$ のみが観測できるとする．ν を $(\mathcal{Y}, \mathcal{B})$ 上の σ - 有限な測度とする．すると T は $(\mathcal{Y}, \mathcal{B})$ 上の測度を誘導する：

$$Q_\theta(B) = P_\theta T^{-1}(B) = P_\theta(T^{-1}(B)), \qquad B \in \mathcal{B}$$

[2)] ある部分集合の列 $\{G_n\}_{n=1}^{\infty}$, $G_n \in \mathcal{A}$ で $\cup_{n=1}^{\infty} G_n = \mathcal{X}$ かつ各 n に対して $\mu(G_n) < \infty$ なるものが存在することである．

さらに，確率測度 Q_θ は測度 ν に関して確率密度関数 $q(\cdot\,|\,\theta)$ を持つとする：

$$Q_\theta(B) = \int_B q(\boldsymbol{y}|\,\theta)\,\nu(d\boldsymbol{y}), \qquad B \in \mathcal{B}$$

EM アルゴリズムは，観測 $\boldsymbol{Y} = \boldsymbol{y}$ が与えられたとき，θ の関数として $q(\boldsymbol{y}|\,\theta)$ を最大化することで θ の最尤推定値を求める方法である．

EM アルゴリズムは以下のように行う：$\boldsymbol{y}$ が与えられたとき，θ の初期推定値の $\widehat{\theta}^{(0)}$ から始める．$\widehat{\theta}^{(0)}$ より $P_{\widehat{\theta}^{(0)}}$ と $Q_{\widehat{\theta}^{(0)}} = P_{\widehat{\theta}^{(0)}}T^{-1}$ が初期の推定された確率測度となる．

E - 段階 (E - Step)：各 $\theta \in \Theta$ に対し，条件付き期待値

$$\phi_1(\theta) = \mathbb{E}_{P_{\widehat{\theta}^{(0)}}}\{\log p(\boldsymbol{X}|\,\theta)|\,T(\boldsymbol{X}) = \boldsymbol{y}\} \tag{2.11}$$

を求める．$P_{\widehat{\theta}^{(0)}}\{T(\boldsymbol{X}) = \boldsymbol{y}\} > 0$ の場合には，$T(\boldsymbol{X}) = \boldsymbol{y}$ が与えられたときの $\boldsymbol{X}$ の条件付分布を求め，それに関して関数 $\boldsymbol{x} \mapsto \log p(\boldsymbol{x}|\,\theta)$ の期待値を求めればよい．$P_{\widehat{\theta}^{(0)}}\{T(\boldsymbol{X}) = \boldsymbol{y}\} = 0$ のときは注意が必要であるが，(2.11) の正当化は可能である．以後は簡単のために，$P_{\widehat{\theta}^{(0)}}\{T(\boldsymbol{X}) = \boldsymbol{y}\} > 0$ の場合を考える．

M - 段階 (M - Step)：θ に関して

$$\phi_1(\theta)$$

を最大化する．最大を与える点 (存在すれば) を $\widehat{\theta}^{(1)}$ とおく．つぎに，$P_{\widehat{\theta}^{(0)}}$ の代わりに $P_{\widehat{\theta}^{(1)}}$ を用いる．

E - 段階 (E - Step)：各 $\theta \in \Theta$ に対して

$$\phi_2(\theta) = \mathbb{E}_{P_{\widehat{\theta}^{(1)}}}\{\log p(\boldsymbol{X}|\,\theta)|\,T(\boldsymbol{X}) = \boldsymbol{y}\}$$

を求める．

M - 段階 (M - Step)：θ に関して

$$\phi_2(\theta)$$

を最大化する．最大を与える点 (存在すれば) を $\widehat{\theta}^{(2)}$ とおく．

一般には，m - 段階 (m $= 1, 2, \ \ldots$) で

E - 段階 (E - Step)：各 $\theta \in \Theta$ に対し，条件付き期待値

$$\phi_m(\theta) = \mathbb{E}_{P_{\widehat{\theta}^{(m-1)}}}\{\log p(\boldsymbol{X}|\,\theta)|\,T(\boldsymbol{X}) = \boldsymbol{y}\} \tag{2.12}$$

を求める．

M - 段階 (M - Step)：θ に関して

$$\phi_m(\theta)$$

を最大化する．最大を与える点 (存在すれば) を $\widehat{\theta}^{(m)}$ とおく．

この操作を $\widehat{\theta}^{(m)}$ が $\widehat{\theta}^{(m-1)}$ とほとんど変化がなくなるまで繰り返す．

例 2.19 $X_1, X_2, \ldots, X_n$ は独立同一に指数分布

$$f(x|\,\theta) = \frac{1}{\theta}e^{-x/\theta}I_{(0,\,\infty)}(x)$$

に従うとする．ただし，$\theta > 0$ である．しかし，各 X_i, $i = 1, 2, \ldots, n$, は直接観測されず，各 X_i の整数部分のみが観測されるとする．すなわち，$Y_i = \lfloor X_i \rfloor$ である．$Y_1, Y_2, \ldots, Y_n$ の観測に基づいて θ の最尤推定値を求めよう．

この場合，$\mathcal{X} = (\mathbb{R}^+)^n$，$\mathcal{A}$ は $(\mathbb{R}^+)^n$ 上のボレル可測集合体であり，P_θ はルベーグ測度に関する確率密度関数

$$p(\boldsymbol{x}|\,\theta) = \Pi_{i=1}^n f(x_i|\,\theta) = \theta^{-n}\exp\bigl(-\sum_{i=1}^n x_i/\theta\bigr), \qquad \boldsymbol{x} = (x_1, x_2, \ldots, x_n)$$

を持つ．また，$\mathcal{Y} = \{0, 1, \ldots\}^n$ で $\mathcal{B}$ は $\mathcal{Y}$ のすべての部分集合の集まりからなる σ - 集合体である．さらに，関数 $T : \mathcal{X} \mapsto \mathcal{Y}$ は

$$T(\boldsymbol{x}) = (\lfloor x_1 \rfloor, \lfloor x_2 \rfloor, \ldots, \lfloor x_n \rfloor)$$

で定義される．

いま，

$$P(Y_i = y) = \int_y^{y+1} f(x|\,\theta)\,dx = e^{-y/\theta}(1 - e^{-1/\theta})$$

となるので

$$q(\boldsymbol{y}|\,\theta) = \Pi_{i=1}^n e^{-y_i/\theta}(1 - e^{-1/\theta}), \qquad \boldsymbol{y} = (y_1, y_2, \ldots, y_n)$$

となる．直接 $q(\boldsymbol{y}|\,\theta)$ を θ に関して最大化して，θ の最尤推定値を求めることはできるが，EM アルゴリズムを用いるとどうなるかを観てみよう．

まず，$\lfloor X_i \rfloor = y$ が与えられたとき，θ のもとでの X_i の条件付確率密度関数を求める：

$$k_\theta(x|\,y) = \frac{P(X_i = x,\, \lfloor X_i \rfloor = y)}{P(\lfloor X_i \rfloor = y)} = \frac{\theta^{-1}e^{-x/\theta}I_{[y,\,y+1)}(x)}{e^{-y/\theta}(1 - e^{-1/\theta})}$$

これより

$$\phi_1(\theta) = \mathbb{E}_{P_{\widehat{\theta}(0)}}[\log p(\boldsymbol{X}|\,\theta)|T(\boldsymbol{X}) = \boldsymbol{y}] = -\frac{1}{\theta}\mathbb{E}_{P_{\widehat{\theta}(0)}}\bigl[\sum_{i=1}^n X_i|T(\boldsymbol{X}) = \boldsymbol{y}\bigr] - n\log\theta$$

と

$$\begin{aligned}
\mathbb{E}_{P_{\widehat{\theta}^{(0)}}}[X_i|T(\boldsymbol{X})=\boldsymbol{y}] &= \frac{1}{\widehat{\theta}^{(0)}e^{-y/\widehat{\theta}^{(0)}}(1-e^{-1/\widehat{\theta}^{(0)}})}\int_y^{y+1} xe^{-x/\widehat{\theta}^{(0)}}\,dx \\
&= \frac{1}{e^{-y/\widehat{\theta}^{(0)}}(1-e^{-1/\widehat{\theta}^{(0)}})}\left\{\left[-xe^{-x/\widehat{\theta}^{(0)}}\right]_y^{y+1}+\int_y^{y+1} e^{-x/\widehat{\theta}^{(0)}}\,dx\right\} \\
&= \frac{1}{e^{-y/\widehat{\theta}^{(0)}}(1-e^{-1/\widehat{\theta}^{(0)}})}\Big\{-(y+1)e^{-(y+1)/\widehat{\theta}^{(0)}}+ye^{-y/\widehat{\theta}^{(0)}} \\
&\qquad -\widehat{\theta}^{(0)}e^{-(y+1)/\widehat{\theta}^{(0)}}+\widehat{\theta}^{(0)}e^{-y/\widehat{\theta}^{(0)}}\Big\} \\
&= \frac{1}{e^{-y/\widehat{\theta}^{(0)}}(1-e^{-1/\widehat{\theta}^{(0)}})}\left\{(y+\widehat{\theta}^{(0)})e^{-y/\widehat{\theta}^{(0)}}(1-e^{-1/\widehat{\theta}^{(0)}})-e^{-(y+1)/\widehat{\theta}^{(0)}}\right\} \\
&= y+\widehat{\theta}^{(0)}-\frac{e^{-1/\widehat{\theta}^{(0)}}}{1-e^{-1/\widehat{\theta}^{(0)}}} \\
&= y+\widehat{\theta}^{(0)}-\frac{1}{e^{1/\widehat{\theta}^{(0)}}-1}
\end{aligned}$$

となる．よって，E 段階は

$$\phi_1(\theta)=n\left(-\log\theta+\frac{1}{\theta(e^{1/\widehat{\theta}^{(0)}}-1)}-\frac{\bar{y}_n+\widehat{\theta}^{(0)}}{\theta}\right)$$

となる．ただし，$\bar{y}_n=(1/n)\sum_{i=1}^n y_i$ である．次に，M 段階は上の式を θ に関して最大化する：

$$\widehat{\theta}^{(1)}=\mathrm{argmax}_\theta\phi_1(\theta)=\widehat{\theta}^{(0)}+\bar{y}_n-\frac{1}{e^{1/\widehat{\theta}^{(0)}}-1}$$

となる．したがって，EM アルゴリズムは

$$\widehat{\theta}^{(m)}=\mathrm{argmax}_\theta\phi_m(\theta)=\widehat{\theta}^{(m-1)}+\bar{y}_n-\frac{1}{e^{1/\widehat{\theta}^{(m-1)}}-1}$$

で与えられる．

つぎに，EM アルゴリズムがどうしてうまく働くかを観る．$\boldsymbol{Y}=\boldsymbol{y}$ が与えられたとき，$\widehat{\theta}$ を θ の最尤推定値とし，Θ の内部上で関数

$$\theta\mapsto q(\boldsymbol{y}|\,\theta),\qquad \theta\in\Theta$$

は微分可能とする．このとき，

$$\left.\frac{\partial}{\partial\theta}q(\boldsymbol{y}|\,\theta)\right|_{\theta=\widehat{\theta}}=0$$

である．

いま，$\widehat{\theta}^{(\infty)}$ を Θ の内点とし，EM アルゴリズムの収束先とする．このとき，$\widehat{\theta}^{(\infty)}$ は関数

$$\theta\mapsto\mathbb{E}_{P_{\widehat{\theta}^{(\infty)}}}\{\log p(\boldsymbol{X}|\,\theta)|T(\boldsymbol{X})=\boldsymbol{y}\} \tag{2.13}$$

を最大化する．(2.13) は Θ の内部で微分可能とし，期待値と微分記号の交換が可能とすれば，$\theta = \widehat{\theta}^{(\infty)}$ において

$$\frac{\partial}{\partial\theta}\mathbb{E}_{P_{\widehat{\theta}^{(\infty)}}}\{\log p(\boldsymbol{X}|\,\theta)|T(\boldsymbol{X})=\boldsymbol{y}\}\bigg|_{\theta=\widehat{\theta}^{(\infty)}} = \mathbb{E}_{P_{\widehat{\theta}^{(\infty)}}}\left\{\frac{\partial}{\partial\theta}\log p(\boldsymbol{X}|\,\theta)|\,T(\boldsymbol{X})=\boldsymbol{y}\right\}\bigg|_{\theta=\widehat{\theta}^{(\infty)}} = 0 \tag{2.14}$$

となる．ここで

$$\dot{l}_n(\theta|\,\boldsymbol{x}) = \frac{\partial}{\partial\theta}\log p(\boldsymbol{x}|\,\theta)$$

とおけば，適当な仮定のもとで

$$\begin{aligned}
\mathbb{E}_{P_\theta}\{\dot{l}_n(\theta|\,\boldsymbol{X})|T(\boldsymbol{X})=\boldsymbol{y}\} &= \left\{\int_{T(\boldsymbol{x})=\boldsymbol{y}} \dot{l}_n(\theta|\,\boldsymbol{x})p(\boldsymbol{x}|\,\theta)\mu(d\boldsymbol{x})\right\} \\
&= \left\{\int_{T(\boldsymbol{x})=\boldsymbol{y}} \left(\frac{\partial}{\partial\theta}p(\boldsymbol{x}|\,\theta)\right)\mu(d\boldsymbol{x})\right\} \\
&= \left\{\frac{\partial}{\partial\theta}\int_{T(\boldsymbol{x})=\boldsymbol{y}} p(\boldsymbol{x}|\,\theta)\mu(d\boldsymbol{x})\right\} \\
&= \left(\frac{\partial}{\partial\theta}q(\boldsymbol{y}|\,\theta)\right) \\
&= \frac{\partial}{\partial\theta}q(\boldsymbol{y}|\,\theta)
\end{aligned}$$

となる．よって，$q(\boldsymbol{y}|\,\theta) > 0$ となる $\boldsymbol{y}$ に対して，(2.14) は

$$\frac{\partial}{\partial\theta}\log q(\boldsymbol{y}|\,\theta)\bigg|_{\theta=\widehat{\theta}^{(\infty)}} = 0 \tag{2.15}$$

を意味する．したがって，最尤推定値が一意に存在するならば，$\widehat{\theta} = \widehat{\theta}^{(\infty)}$ となる．

$\theta = \widehat{\theta}^{(\infty)}$ なる解をもつ方程式

$$\mathbb{E}_{P_\theta}\{\dot{l}_n(\theta|\,\boldsymbol{X})|T(\boldsymbol{X})=\boldsymbol{y}\} = 0$$

を自己一致方程式 (self - consistency equation) という．

以上の議論からつぎの場合には，EM アルゴリズムはうまく機能されるかは保障されていない．

1. 最大値を与える点が Θ の内点に含まれない．
2. 尤度関数がその最大を取る点で微分可能ではない．
3. スコア方程式 (2.15) が複数の解を持ち，そのいくつかは尤度関数を最大にしない．

最後に，EM アルゴリズムの各段階で，対数尤度関数 $\log q(\boldsymbol{y}|\,\theta)$ は非減少であることを示す：

$$\log q(\boldsymbol{y}|\,\widehat{\theta}^{(m)}) > \log q(\boldsymbol{y}|\,\widehat{\theta}^{(m-1)}), \qquad m = 0,\, 1,\, \ldots \tag{2.16}$$

$T(\boldsymbol{X})=\boldsymbol{y}$ を与えたとき，$\boldsymbol{X}$ の条件付確率密度関数 $k_\theta(\boldsymbol{x}|\,\boldsymbol{y})$ は次で与えられる：

$$k_\theta(\boldsymbol{x}|\,\boldsymbol{y})=\frac{p(\boldsymbol{x}|\,\theta)}{q(\boldsymbol{y}|\,\theta)}I_{T^{-1}(\boldsymbol{y})}(\boldsymbol{x})$$

$T(\boldsymbol{x})=\boldsymbol{y}$, $p(\boldsymbol{x}|\,\theta)>0$, $q(\boldsymbol{y}|\,\theta)>0$ の場合，

$$\log q(\boldsymbol{y}|\,\theta)=\log p(\boldsymbol{x}|\,\theta)-\log k_\theta(\boldsymbol{x}|\,\boldsymbol{y})$$

となる．

以下では，最尤推定値の候補は $q(\boldsymbol{y}|\,\theta)>0$ をみたしていなければいけないので，$q(\boldsymbol{y}|\,\theta)>0$ を仮定して議論を進める．これは，そうでなければ，$\log q(\boldsymbol{y}|\,\theta)>0$ を最大にしないことからわかる．

各 m に対し

$$\begin{aligned}\log q(\boldsymbol{y}|\,\theta)&=\mathbb{E}_{P_{\widehat{\theta}^{(m-1)}}}\{\log q(\boldsymbol{y}|\,\theta)|T(\boldsymbol{X})=\boldsymbol{y}\}\\&=\mathbb{E}_{P_{\widehat{\theta}^{(m-1)}}}\{\log p(\boldsymbol{X}|\,\theta)-\log k_\theta(\boldsymbol{X}|\,\boldsymbol{y})|T(\boldsymbol{X})=\boldsymbol{y}\}\\&=\mathbb{E}_{P_{\widehat{\theta}^{(m-1)}}}\{\log p(\boldsymbol{X}|\,\theta)|T(\boldsymbol{X})=\boldsymbol{y}\}-\mathbb{E}_{P_{\widehat{\theta}^{(m-1)}}}\{\log k_\theta(\boldsymbol{X}|\,\boldsymbol{y})|T(\boldsymbol{X})=\boldsymbol{y}\}\end{aligned}\tag{2.17}$$

となる．上式の最右辺の各項を別々に評価していく．

まず，$\widehat{\theta}^{(m)}$ の定義から

$$\mathbb{E}_{P_{\widehat{\theta}^{(m-1)}}}\{\log p(\boldsymbol{X}|\,\widehat{\theta}^{(m)})|T(\boldsymbol{X})=\boldsymbol{y}\}-\mathbb{E}_{P_{\widehat{\theta}^{(m-1)}}}\{\log p(\boldsymbol{X}|\,\widehat{\theta}^{(m-1)})|T(\boldsymbol{X})=\boldsymbol{y}\}\geq 0\tag{2.18}$$

がわかる．

次に，(2.17) の最右辺の 2 項目を評価する：

$$\begin{aligned}&\mathbb{E}_{P_{\widehat{\theta}^{(m-1)}}}\{\log k_{\widehat{\theta}^{(m)}}(\boldsymbol{X}|\,\boldsymbol{y})|T(\boldsymbol{X})=\boldsymbol{y}\}-\mathbb{E}_{P_{\widehat{\theta}^{(m-1)}}}\{\log k_{\widehat{\theta}^{(m-1)}}(\boldsymbol{X}|\,\boldsymbol{y})|T(\boldsymbol{X})=\boldsymbol{y}\}\\&\qquad=\mathbb{E}_{P_{\widehat{\theta}^{(m-1)}}}\left\{\log\frac{k_{\widehat{\theta}^{(m)}}(\boldsymbol{X}|\,\boldsymbol{y})}{k_{\widehat{\theta}^{(m-1)}}(\boldsymbol{X}|\,\boldsymbol{y})}\middle|T(\boldsymbol{X})=\boldsymbol{y}\right\}\\&\qquad\leq\log\mathbb{E}_{P_{\widehat{\theta}^{(m-1)}}}\left\{\frac{k_{\widehat{\theta}^{(m)}}(\boldsymbol{X}|\,\boldsymbol{y})}{k_{\widehat{\theta}^{(m-1)}}(\boldsymbol{X}|\,\boldsymbol{y})}\middle|T(\boldsymbol{X})=\boldsymbol{y}\right\}\end{aligned}\tag{2.19}$$

となる．最後の不等号は Jensen の不等式よりわかる．いま，

$$g(\boldsymbol{y})=\mathbb{E}_{P_{\widehat{\theta}^{(m-1)}}}\left\{\frac{k_{\widehat{\theta}^{(m)}}(\boldsymbol{X}|\,\boldsymbol{y})}{k_{\widehat{\theta}^{(m-1)}}(\boldsymbol{X}|\,\boldsymbol{y})}\middle|T(\boldsymbol{X})=\boldsymbol{y}\right\}$$

とおく．条件付き期待値の定義から任意の $B\in\mathcal{B}$ に対し

$$\int_B g(\boldsymbol{y})\,dQ_{\widehat{\theta}^{(m-1)}}(\boldsymbol{y})=\int_{T^{-1}(B)}\frac{k_{\widehat{\theta}^{(m)}}(\boldsymbol{X}|\,\boldsymbol{y})}{k_{\widehat{\theta}^{(m-1)}}(\boldsymbol{X}|\,\boldsymbol{y})}\,dP_{\widehat{\theta}^{(m-1)}}(\boldsymbol{x})\tag{2.20}$$

となる．しかし

$$\frac{k_{\widehat{\theta}^{(m)}}(\boldsymbol{x}|\,\boldsymbol{y})}{k_{\widehat{\theta}^{(m-1)}}(\boldsymbol{x}|\,\boldsymbol{y})}=\frac{p(\boldsymbol{x}|\,\widehat{\theta}^{(m)})}{q(T(\boldsymbol{x})|\,\widehat{\theta}^{(m)})}\cdot\frac{q(T(\boldsymbol{x})|\,\widehat{\theta}^{(m-1)})}{p(\boldsymbol{x}|\,\widehat{\theta}^{(m-1)})}$$

に注意すれば，(2.20) の右辺は

$$
\begin{aligned}
\int_{T^{-1}(B)} \frac{k_{\widehat{\theta}^{(m)}}(\boldsymbol{X}|\,\boldsymbol{y})}{k_{\widehat{\theta}^{(m-1)}}(\boldsymbol{X}|\,\boldsymbol{y})}\, dP_{\widehat{\theta}^{(m-1)}} &= \int_{T^{-1}(B)} \frac{p(\boldsymbol{x}|\,\widehat{\theta}^{(m)})}{q(T(\boldsymbol{x})|\,\widehat{\theta}^{(m)})} q(T(\boldsymbol{x})|\,\widehat{\theta}^{(m-1)})\, \mu(d\boldsymbol{x}) \\
&= \int_{T^{-1}(B)} \frac{q(T(\boldsymbol{x})|\,\widehat{\theta}^{(m-1)})}{q(T(\boldsymbol{x})|\,\widehat{\theta}^{(m)})}\, dP_{\widehat{\theta}^{(m)}}(\boldsymbol{x}) \\
&= \int_{B} \frac{q(\boldsymbol{y}|\,\widehat{\theta}^{(m-1)})}{q(\boldsymbol{y}|\,\widehat{\theta}^{(m)})}\, dQ_{\widehat{\theta}^{(m)}}(\boldsymbol{y}) \\
&= \int_{B} 1 \cdot\, dQ_{\widehat{\theta}^{(m-1)}}(\boldsymbol{y})
\end{aligned}
$$

となる．これより

$$
g(\boldsymbol{y}) = \mathbb{E}_{P_{\widehat{\theta}^{(m-1)}}} \left\{ \frac{k_{\widehat{\theta}^{(m)}}(\boldsymbol{X}|\,\boldsymbol{y})}{k_{\widehat{\theta}^{(m-1)}}(\boldsymbol{X}|\,\boldsymbol{y})} | T(\boldsymbol{X}) = \boldsymbol{y} \right\} = 1, \qquad a.e. \quad Q_{\widehat{\theta}^{(m-1)}}
$$

となる．この式と (2.19) から

$$
\mathbb{E}_{P_{\widehat{\theta}^{(m-1)}}}\{\log k_{\widehat{\theta}^{(m)}}(\boldsymbol{X}|\,\boldsymbol{y})|T(\boldsymbol{X}) = \boldsymbol{y}\} - \mathbb{E}_{P_{\widehat{\theta}^{(m-1)}}}\{\log k_{\widehat{\theta}^{(m-1)}}(\boldsymbol{X}|\,\boldsymbol{y})|T(\boldsymbol{X}) = \boldsymbol{y}\} \leq 0 \tag{2.21}
$$

がわかる．よって，(2.18) と (2.21) から (2.17) は示せた．