EM アルゴリズムについて

前川眞一
東京工業大学大学院　社会理工学研究科　人間行動システム専攻
090708,14

1　EM アルゴリズム

パラメタ $\boldsymbol{\xi}$ と $\boldsymbol{\theta}$ の同事後分布の密度関数 $h(\boldsymbol{\xi},\boldsymbol{\theta}|\boldsymbol{y})$ が与えられている場合に、そこから $\boldsymbol{\theta}$ を積分消去した $\boldsymbol{\xi}$ の周辺事後分布

$$g(\boldsymbol{\xi}|\boldsymbol{y}) = \int_{\boldsymbol{\Theta}} h(\boldsymbol{\xi},\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta} \tag{1}$$

の最頻値を求める問題を考える。

この問題は

$$\frac{\partial\ \log g(\boldsymbol{\xi}|\boldsymbol{y})}{\partial\,\boldsymbol{\xi}} = \frac{\partial\ \log \int_{\boldsymbol{\Theta}} h(\boldsymbol{\xi},\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}}{\partial\,\boldsymbol{\xi}} = 0 \tag{2}$$

の解を求めることと同値であるが、定積分が難しい場合には、その解を求めることは難しい。Dempster, et al(19xx) が提案した EM アルゴリズムは、上記の定積分をより簡単な形の定積分に置き換えることにより数値的に周辺最頻値を求める方法である。

いま、$k(\boldsymbol{\theta}|\boldsymbol{\xi},\boldsymbol{y})$ を $\boldsymbol{\xi}$ を与えたときの $\boldsymbol{\theta}$ の事後分布の密度関数とすれば、$\boldsymbol{\xi}$ の周辺事後分布の密度関数は、それぞれの密度関数の定義から

$$g(\boldsymbol{\xi}|\boldsymbol{y}) = \frac{h(\boldsymbol{\xi},\boldsymbol{\theta}|\boldsymbol{y})}{k(\boldsymbol{\theta}|\boldsymbol{\xi},\boldsymbol{y})} \tag{3}$$

と書くことが出来る。したがって、その対数は

$$\log g(\boldsymbol{\xi}|\boldsymbol{y}) = \log h(\boldsymbol{\xi},\boldsymbol{\theta}|\boldsymbol{y}) - \log k(\boldsymbol{\theta}|\boldsymbol{\xi},\boldsymbol{y}) \tag{4}$$

となる。

次に、パラメタ $\boldsymbol{\xi}$ の仮の値を $\boldsymbol{\xi}^{(\ell)}$ として、それが与えたれた時の $\boldsymbol{\theta}$ の事後分布の密度関数を $k(\boldsymbol{\theta}|\boldsymbol{\xi}^{(\ell)},\boldsymbol{y})$ とし、それに関して、(4) 式 の両辺の期待値をとったものを考える。しかし、実は、(4) 式 は $\boldsymbol{\theta}$ には依存しない量であるため、左辺はそのままであることに注意すると、

$$\log g(\boldsymbol{\xi}|\boldsymbol{y}) \quad = \quad Q(\boldsymbol{\xi}|\boldsymbol{\xi}^{(\ell)}) - H(\boldsymbol{\xi}|\boldsymbol{\xi}^{(\ell)}) \tag{5}$$

ただし、

$$Q(\boldsymbol{\xi}|\boldsymbol{\xi}^{(\ell)}) \quad = \quad \int_{\boldsymbol{\Theta}} \log h(\boldsymbol{\xi},\boldsymbol{\theta}|\boldsymbol{y}) \times k(\boldsymbol{\theta}|\boldsymbol{\xi}^{(\ell)},\boldsymbol{y})d\boldsymbol{\theta} \tag{6}$$

$$H(\boldsymbol{\xi}|\boldsymbol{\xi}^{(\ell)}) \quad = \quad \int_{\boldsymbol{\Theta}} \log k(\boldsymbol{\theta}|\boldsymbol{\xi},\boldsymbol{y}) \times k(\boldsymbol{\theta}|\boldsymbol{\xi}^{(\ell)},\boldsymbol{y})d\boldsymbol{\theta} \tag{7}$$

と書くことが出来る。

EM アルゴリズムとは、(1) 式 を直接に $\boldsymbol{\xi}$ に関して最大化するのではなく、$\boldsymbol{\xi}^{(\ell)}$ が与えられた場合に、

$$Q(\boldsymbol{\xi}^{(\ell+1)}|\boldsymbol{\xi}^{(\ell)}) > Q(\boldsymbol{\xi}^{(\ell)}|\boldsymbol{\xi}^{(\ell)}) \tag{8}$$

を満たすような $\boldsymbol{\xi}^{(\ell+1)}$ を逐次求めていきながら、最終的に (4) 式、したがって (1) 式 を最大にする $\boldsymbol{\xi}$ を求めていく方法である。$\boldsymbol{\xi}^{(\ell+1)}$ としては、例えば

$$\frac{\partial\, Q(\boldsymbol{\xi}|\boldsymbol{\xi}^{(\ell)})}{\partial\, \boldsymbol{\xi}} = 0 \tag{9}$$

の解を用いることが考えられる。ただし、$\boldsymbol{\xi}^{(\ell)}$ は定数として取り扱う。

なお、(8) 式 を満たすような $\boldsymbol{\xi}^{(\ell+1)}$ は (5) 式 をも増加させる、すなわち

$$d^{(\ell)} = \log g(\boldsymbol{\xi}^{(\ell+1)}|\boldsymbol{y}) - \log g(\boldsymbol{\xi}^{(\ell)}|\boldsymbol{y}) > 0 \tag{10}$$

であることは、以下のようにして示すことが出来る。上式は

$$\begin{aligned} d^{(\ell)} &= Q(\boldsymbol{\xi}^{(\ell+1)}|\boldsymbol{\xi}^{(\ell)}) - H(\boldsymbol{\xi}^{(\ell+1)}|\boldsymbol{\xi}^{(\ell)}) - \Big(Q(\boldsymbol{\xi}^{(\ell)}|\boldsymbol{\xi}^{(\ell)}) - H(\boldsymbol{\xi}^{(\ell)}|\boldsymbol{\xi}^{(\ell)})\Big) && (11)\\ &= Q(\boldsymbol{\xi}^{(\ell+1)}|\boldsymbol{\xi}^{(\ell)}) - Q(\boldsymbol{\xi}^{(\ell)}|\boldsymbol{\xi}^{(\ell)}) - \Big(H(\boldsymbol{\xi}^{(\ell+1)}|\boldsymbol{\xi}^{(\ell)}) - H(\boldsymbol{\xi}^{(\ell)}|\boldsymbol{\xi}^{(\ell)})\Big) && (12) \end{aligned}$$

とかけるが、その右辺の Q を含む部分は (8) 式 より正である。また、H を含む部分は、対数関数が凹関数であることを利用してジェンセンの不等式を用いれば、

$$\begin{aligned} H(\boldsymbol{\xi}^{(\ell+1)}|\boldsymbol{\xi}^{(\ell)}) - H(\boldsymbol{\xi}^{(\ell)}|\boldsymbol{\xi}^{(\ell)}) &= \int_{\boldsymbol{\Theta}} \log k(\boldsymbol{\theta}|\boldsymbol{\xi}^{(\ell+1)}, \boldsymbol{y}) \times k(\boldsymbol{\theta}|\boldsymbol{\xi}^{(\ell)}, \boldsymbol{y}) d\boldsymbol{\theta} \\ &\qquad - \int_{\boldsymbol{\Theta}} \log k(\boldsymbol{\theta}|\boldsymbol{\xi}^{(\ell)}, \boldsymbol{y}) \times k(\boldsymbol{\theta}|\boldsymbol{\xi}^{(\ell)}, \boldsymbol{y}) d\boldsymbol{\theta} && (13)\\ &= \int_{\boldsymbol{\Theta}} \log \frac{k(\boldsymbol{\theta}|\boldsymbol{\xi}^{(\ell+1)}, \boldsymbol{y})}{k(\boldsymbol{\theta}|\boldsymbol{\xi}^{(\ell)}, \boldsymbol{y})} \times k(\boldsymbol{\theta}|\boldsymbol{\xi}^{(\ell)}, \boldsymbol{y}) d\boldsymbol{\theta} && (14)\\ &\leq \log \int_{\boldsymbol{\Theta}} \frac{k(\boldsymbol{\theta}|\boldsymbol{\xi}^{(\ell+1)}, \boldsymbol{y})}{k(\boldsymbol{\theta}|\boldsymbol{\xi}^{(\ell)}, \boldsymbol{y})} \times k(\boldsymbol{\theta}|\boldsymbol{\xi}^{(\ell)}, \boldsymbol{y}) d\boldsymbol{\theta} && (15)\\ &= \log \int_{\boldsymbol{\Theta}} k(\boldsymbol{\theta}|\boldsymbol{\xi}^{(\ell+1)}, \boldsymbol{y}) d\boldsymbol{\theta} && (16)\\ &= \log(1) \;=\; 0 && (17) \end{aligned}$$

となり、負の値となる。したがって、(10) 式 が示せた。なお、上式に於いて、最終行は、k が密度関数であるため、その積分が 1 となることを用いた。

なお、上記の議論に於いて、同時事後分布を、パラメタが与えられたときのデータ $\boldsymbol{y}$ の同時分布の密度関数と $f(\boldsymbol{y}|\boldsymbol{\xi}, \boldsymbol{\theta})$ とパラメタの事前分布 $h(\boldsymbol{\theta}, \boldsymbol{\xi})$ とで表現すれば、

$$h(\boldsymbol{\xi}, \boldsymbol{\theta}|\boldsymbol{y}) \propto\; f(\boldsymbol{y}|\boldsymbol{\xi}, \boldsymbol{\theta}) h(\boldsymbol{\theta}, \boldsymbol{\xi}) = f(\boldsymbol{y}|\boldsymbol{\xi}, \boldsymbol{\theta}) h(\boldsymbol{\theta}|\boldsymbol{\xi}) h(\boldsymbol{\xi}) \tag{18}$$

となるが、これより

$$g(\boldsymbol{\xi}|\boldsymbol{y}) \propto \int_{\Theta} f(\boldsymbol{y}|\boldsymbol{\xi}, \boldsymbol{\theta}) h(\boldsymbol{\theta}|\boldsymbol{\xi}) d\boldsymbol{\theta}\; h(\boldsymbol{\xi}) \;= L(\boldsymbol{\xi})\; h(\boldsymbol{\xi}) \tag{19}$$

となる。したがって、パラメタ $\boldsymbol{\xi}$ に関する事前分布が一様分布の場合には、EM アルゴリズムは、所謂周辺尤度 $L(\boldsymbol{\xi})$ をを最大にするアルゴリズムと見なすことが出来る。実は、この形が本来の EM アルゴリズムであり、その意味で、$Q(\boldsymbol{\xi}|\boldsymbol{\xi}^{(\ell)})$ は期待対数完全データ尤度と呼ばれることがある。

EM アルゴリズムを用いて周辺事後分布の最頻値が簡単に求まるか否かは、関数 Q、すなわち、同時事後分布の密度関数の期待値が簡単に求められ、且つ、Q の値を増加させる、もしくは最大にす

る $\boldsymbol{\xi}^{(\ell+1)}$ が簡単に求められるか否かにかかっている。なお、この期待値を求める操作のことを E ステップと呼び、それを最大にする (もしくはその値を前回から増加させる) $\boldsymbol{\xi}^{(\ell+1)}$ の値を求める操作のことを M ステップと呼んでいる。

EM アルゴリズムを用いて得られるパラメタ $\boldsymbol{\xi}^{(\ell)}$ の系列は、以上に説明により、決して周辺事後分布の密度関数の値を減少させることがないため、最終的に得られる $\boldsymbol{\xi}$ の値は、周辺事後分布の最頻値であると考えられる。しかし、最頻値が複数存在する事後分布の場合、最も密度関数の値が大きい最頻値に到達するという保証はなく、最大値ではなく極大値に到達している可能性も否めない。また、EM アルゴリズムは、自動的に、周辺事後分布の最頻値付近における集中の度合い、例えば、事後標準偏差等を与えるものではない。しかし、周辺事後分布の要約として点推定値のみが必要な場合には、有用な方法である。

以下に、階層的線形モデルと潜在クラスモデルに関して、EM アルゴリズムを用いて周辺事後分布の最頻値を求める方法に関して記す。

註

ジェンセンの不等式 Jesen's inequality

$f(\boldsymbol{x})$ を $\boldsymbol{x}$ の密度関数、$r(\boldsymbol{x})$ を $\boldsymbol{x}$ の任意の実数値関数、$q(\boldsymbol{x})$ を $\boldsymbol{x}$ の任意の凹関数とするとき、

$$q\left(\int_X r(\boldsymbol{x})f(\boldsymbol{x})d\boldsymbol{x}\right) \geq \int_X q\left(r(\boldsymbol{x})\right) f(\boldsymbol{x})d\boldsymbol{x} \tag{20}$$

すなわち

$$q\left(\boldsymbol{\mathcal{E}}(r(\boldsymbol{x}))\right) \geq \boldsymbol{\mathcal{E}}(q\left(r(\boldsymbol{x})\right)) \tag{21}$$

である。

凹関数 concave function

任意の変数の値 $\boldsymbol{x}$ と $\boldsymbol{y}$ に於いて、以下の関係が成り立つ関数を凸関数 convex function と呼ぶ。ただし a は $0 < a < 1$ を満たす定数である。

$$q(a\boldsymbol{x} + (1-a)\boldsymbol{y}) \leq aq(\boldsymbol{x}) + (1-a)q(\boldsymbol{y}) \tag{22}$$

$-q(\boldsymbol{x})$ が凸関数の場合 q は凹関数と呼ばれる。

2 階層的線形モデル

Q=3 層の階層線形モデル

$$\begin{aligned}
\boldsymbol{y}|\boldsymbol{v}_1, \boldsymbol{C}_1 &\sim N(\boldsymbol{G}_0\boldsymbol{v}_1, \boldsymbol{C}_1) && (23)\\
\boldsymbol{v}_1|\boldsymbol{v}_2, \boldsymbol{C}_2 &\sim N(\boldsymbol{G}_1\boldsymbol{v}_2, \boldsymbol{C}_2) && (24)\\
\boldsymbol{v}_2 &\sim N(\boldsymbol{G}_2\boldsymbol{v}_3, \boldsymbol{C}_3) && (25)
\end{aligned}$$

を考え、パラメタを

$$\boldsymbol{\xi} = \{\boldsymbol{v}_2, \boldsymbol{C}_1, \boldsymbol{C}_2\}, \quad \boldsymbol{\theta} = \boldsymbol{v}_1 \tag{26}$$

と分割する。また、$(\boldsymbol{v}_3, \boldsymbol{C}_3)$ 並びに $\boldsymbol{C}_1, \boldsymbol{C}_2$ の事前分布のハイパーパラメタ $\boldsymbol{D}_1, \boldsymbol{D}_2$ は定数とする。

いま、このモデルの $\boldsymbol{\xi}$ の周辺事後分布から最頻値を求めたいものとする。$\boldsymbol{\xi}$ の周辺事後分布は、上記のモデルから $\boldsymbol{\theta}$ を積分消去した周辺尤度に $\boldsymbol{\xi}$ の事前分布の密度関数を掛け合わせることによって得られるが、周辺尤度が

$$y|\boldsymbol{v}_2, \boldsymbol{C}_1, \boldsymbol{C}_2 \quad \sim \quad N\left(\boldsymbol{G}_0\boldsymbol{G}_1\boldsymbol{v}_2,\ \boldsymbol{G}_0\boldsymbol{C}_2\boldsymbol{G}_0' + \boldsymbol{C}_1\right) \tag{27}$$

となり、$\left|\boldsymbol{G}_0\boldsymbol{C}_2\boldsymbol{G}_0' + \boldsymbol{C}_1\right|$ や $\left(\boldsymbol{G}_0\boldsymbol{C}_2\boldsymbol{G}_0' + \boldsymbol{C}_1\right)^{-1}$ 等が出現するため、簡単に十辺事後分布の最頻値を求めることができない。

そこで EM アルゴリズムを用いて、周辺事後分布の最頻値を求めることを考える。まず、$\boldsymbol{\xi}^{(\ell)}$ を与えたときの $\boldsymbol{\theta}$ の条件付き分布であるが、これは第 xx 章の (**??**) 式 より、

$$\boldsymbol{v}_1|\boldsymbol{y}, \boldsymbol{C}_1, \boldsymbol{C}_2, \boldsymbol{v}_2 \quad \sim \quad N(\boldsymbol{\theta}_1, \boldsymbol{\Theta}_1) \tag{28}$$

ただし、

$$\begin{aligned}
\boldsymbol{\theta}_1 &= \boldsymbol{\Theta}_1(\boldsymbol{C}_2{}^{-1}\boldsymbol{G}_1\boldsymbol{v}_2 + \boldsymbol{G}_0'\boldsymbol{C}_1{}^{-1}\boldsymbol{G}_0\widehat{\boldsymbol{v}_1}) && (29)\\
\boldsymbol{\Theta}_1 &= (\boldsymbol{G}_0'\boldsymbol{C}_1{}^{-1}\boldsymbol{G}_0 + \boldsymbol{C}_2{}^{-1})^{-1} && (30)\\
\widehat{\boldsymbol{v}_1} &= (\boldsymbol{G}_0'\boldsymbol{C}_1{}^{-1}\boldsymbol{G}_0)^{-1}\boldsymbol{G}_0'\boldsymbol{C}_1{}^{-1}\boldsymbol{y} && (31)
\end{aligned}$$

である。なお、上記の条件付き分布のパラメタ $\boldsymbol{\theta}_1, \boldsymbol{\Theta}_1, \widehat{\boldsymbol{v}_1}$ の中に現れる $\boldsymbol{C}_1, \boldsymbol{C}_2, \boldsymbol{v}_2$ は $\boldsymbol{\xi}^{(\ell)}$ の値である。次に、同時事後分布の密度関数の対数は、$h(\boldsymbol{\xi}) = h_v(\boldsymbol{v}_2)h_{C1}(\boldsymbol{C}_1)h_{C2}(\boldsymbol{C}_2)$ をパラメタの事前分布として

$$\begin{aligned}
\log h(\boldsymbol{\xi}, \boldsymbol{\theta}|\boldsymbol{y}) \quad = \quad & -\tfrac{1}{2}\Big[\log|\boldsymbol{C}_1| + (\boldsymbol{y} - \boldsymbol{G}_0\boldsymbol{v}_1)'\boldsymbol{C}_1{}^{-1}(\boldsymbol{y} - \boldsymbol{G}_0\boldsymbol{v}_1)\\
& + \log|\boldsymbol{C}_2| + (\boldsymbol{v}_1 - \boldsymbol{G}_1\boldsymbol{v}_2)'\boldsymbol{C}_2{}^{-1}(\boldsymbol{v}_1 - \boldsymbol{G}_1\boldsymbol{v}_1 2)\Big] + \log(h(\boldsymbol{\xi})) + const \qquad (32)
\end{aligned}$$

であるから、その $\boldsymbol{v}_1$ に関する期待値は、付録の公式から、

$$\begin{aligned}
Q(\boldsymbol{\xi}, \boldsymbol{\xi}^{(\ell)}) \quad = \quad & -\tfrac{1}{2}\Big[\log|\boldsymbol{C}_1| + (\boldsymbol{y} - \boldsymbol{G}_0\boldsymbol{\theta}_1)'\boldsymbol{C}_1{}^{-1}(\boldsymbol{y} - \boldsymbol{G}_0\boldsymbol{\theta}_1) + \mathrm{tr}\left(\boldsymbol{\Theta}_1\boldsymbol{G}_0'\boldsymbol{C}_1{}^{-1}\boldsymbol{G}_0\right)\\
& + \log|\boldsymbol{C}_2| + (\boldsymbol{\theta}_1 - \boldsymbol{G}_1\boldsymbol{v}_2)'\boldsymbol{C}_2{}^{-1}(\boldsymbol{\theta}_1 - \boldsymbol{G}_1\boldsymbol{v}_2)\Big] + \mathrm{tr}\left(\boldsymbol{\Theta}_1\boldsymbol{C}_2{}^{-1}\right) + \log(h(\boldsymbol{\xi})) + const \qquad (33)
\end{aligned}$$

となるが、これが E ステップに対応する。したがって、上式を $\boldsymbol{\xi}$ に関して最大化する手続きが M ステップに当たるが、たとえば $\boldsymbol{\xi}$ のうち $\boldsymbol{v}_2$ の事前分布が無情報分布である場合には、$\boldsymbol{v}_2$ に関しては、付録の残差の２乗和の分解公式から

$$\boldsymbol{v}_2 = \left(\boldsymbol{G}_1'\boldsymbol{C}_2{}^{-1}\boldsymbol{G}_1\right)^{-1}\boldsymbol{G}_1'\boldsymbol{C}_2{}^{-1}\boldsymbol{\theta}_1 \tag{34}$$

がその解となることが分かる。また、$\boldsymbol{v}_2$ の事前分布が正規分布である場合には、一般ガウス・マルコフモデルの場合と同様に平方を完成を行えば、その事後平均に相当するものが Q を最大とする解であることが分かる。

また、(33) 式 を以下のように変形すると

$$\begin{aligned}
Q(\boldsymbol{\xi}, \boldsymbol{\xi}^{(\ell)}) \quad = \quad & -\tfrac{1}{2}\left[\log|\boldsymbol{C}_1| + \mathrm{tr}\left(\boldsymbol{C}_1{}^{-1}\left((\boldsymbol{y} - \boldsymbol{G}_0\boldsymbol{\theta}_1)(\boldsymbol{y} - \boldsymbol{G}_0\boldsymbol{\theta}_1)' + \boldsymbol{G}_0\boldsymbol{\Theta}_1\boldsymbol{G}_0'\right)\right) + \log(h_{C1}(\boldsymbol{C}_1))\right]\\
& -\tfrac{1}{2}\left[\log|\boldsymbol{C}_2| + \mathrm{tr}\left(\boldsymbol{C}_2{}^{-1}\left((\boldsymbol{\theta}_1 - \boldsymbol{G}_1\boldsymbol{v}_2)(\boldsymbol{\theta}_1 - \boldsymbol{G}_1\boldsymbol{v}_2)' + \boldsymbol{\Theta}_1\right)\right) + \log(h_{C2}(\boldsymbol{C}_2))\right] \qquad (35)
\end{aligned}$$

となり、周辺尤度から導かれる周辺事後分布の密度関数よりも簡単な形をしていることが分かる。すなわち、$\boldsymbol{C}_1$ と $\boldsymbol{C}_2$ は独立に推定することが可能である。特に、 Q を最大とする $\boldsymbol{C}_1$ の構成要素である $\sigma_k,\ k = 1, 2, \ldots, m$ は、準自然共役事前分布である逆カイ分布を用いれば、それぞれ独立に解析的に求めることが出来る。また、 $\boldsymbol{C}_2$ の構成要素である $\boldsymbol{\Phi}$ も同様である。

3　潜在クラスモデル

潜在クラスモデルにおいては、パラメタ $\boldsymbol{\xi}$ と $\boldsymbol{\rho}$ を $\boldsymbol{\xi}$ とし、各固体のクラスへの所属を示す潜在変数 $\boldsymbol{d}_i, i = 1, 2, \ldots, N$ を $\boldsymbol{\theta}$ として取り扱う。したがって、前節で示した期待値をとるための積分を和で置き換える必要が生じる。

$\boldsymbol{\xi}^{(\ell)}$ を与えたときの $\boldsymbol{\theta}$ の事後分布は第 xx 章の (**??**) 式 で示した多項分布となる。

また、 (**??**) 式 に示した、事後分布の密度関数の対数は

$$\begin{aligned}
\log h(\boldsymbol{\xi}, \boldsymbol{\theta}|\boldsymbol{y}) &= \sum_{i=1}^{N}\sum_{q=1}^{Q} d_{iq} \log\left(f(y_i|\boldsymbol{\xi}_q)\ \rho_q\right) + \sum_{q=1}^{Q}(\alpha_q - 1)\log\rho_q\ + \log h(\boldsymbol{\xi}) && (36)\\
&= \sum_{q=1}^{Q}\sum_{i=1}^{N} d_{iq} \left(\log f(y_i|\boldsymbol{\xi}_q) + \log\rho_q\right) + \sum_{q=1}^{Q}(\alpha_q - 1)\log\rho_q\ + \log h(\boldsymbol{\xi}) && (37)\\
&= \sum_{q=1}^{Q}\sum_{i=1}^{N} d_{iq} \log f(y_i|\boldsymbol{\xi}_q) + \sum_{q=1}^{Q}\left(\sum_{i=1}^{N} d_{iq}\right)\log\rho_q + \sum_{q=1}^{Q}(\alpha_q - 1)\log\rho_q\ + \log h(\boldsymbol{\xi}) && (38)\\
&= \sum_{q=1}^{Q}\sum_{i=1}^{N} d_{iq} \log f(y_i|\boldsymbol{\xi}_q) + \sum_{q=1}^{Q}\left(\sum_{i=1}^{N} d_{iq} + \alpha_q - 1\right)\log\rho_q && (39)
\end{aligned}$$

となるが、$\boldsymbol{d}$ の事後分布より、

$$\boldsymbol{\mathcal{E}}(d_{iq}|y_i, \boldsymbol{\rho}, \boldsymbol{\xi}) = h_{iq} \tag{40}$$

であるから、Q は

$$Q(\boldsymbol{\xi}|\boldsymbol{\xi}^{(\ell)}) = \sum_{q=1}^{Q}\sum_{i=1}^{N} h_{iq} \log f(y_i f|\boldsymbol{\xi}_q) + \log h(\boldsymbol{\xi}) + \sum_{q=1}^{Q}\left(\sum_{i=1}^{N} h_{iq} + \alpha_q - 1\right)\log\rho_q \tag{41}$$

となる。

したがって、M step では、$\boldsymbol{\xi}^{(\ell)}$ の関数である $h_{iq}, i = 1, 2, \ldots, N, q = 1, 2, \ldots, Q$ を定数と見なして上式を最大化する事になるが、 (41) 式 の最初の項は、y_i という値の観測値が h_{iq} 個観測された場合の尤度関数

$$f_h(\boldsymbol{y}|\boldsymbol{\xi}_q) = \prod_{i=1}^{N} f(y_i|\boldsymbol{\xi}_q)^{h_{iq}} \tag{42}$$

の積の対数であると考えられる。また、事前分布として、各クラスのパラメタの事前分布は独立である

$$h(\boldsymbol{\xi}) = \prod_{q=1}^{Q} h(\boldsymbol{\xi}_q) \tag{43}$$

と仮定すると、結局、 (41) 式 の最初の 2 項は

$$h(\boldsymbol{\xi}_q|\boldsymbol{y},\boldsymbol{h}) \propto \prod_{i=1}^{N} f(y_i|\boldsymbol{\xi}_q)^{h_{iq}} \times h(\boldsymbol{\xi}_q) \tag{44}$$

、すなわち、y_i という値の観測値が h_{iq} 個観測された場合の事後分布の密度関数の対数となっていることが分かる。したがって、この事後分布の最頻値を与える $\boldsymbol{\xi}_q$ を求めれば、それが M step における Q を最大とするパラメタの値となる。

例えば、混合正規分布や switching regression の例では、$\boldsymbol{\xi}_q$ の事前分布として事前共役事前分布をもちれば、この Q を最大とする $\boldsymbol{\xi}$ の値は解析的に求められるため、M step における計算は簡単である。

また、生起確率のパラメタ $\boldsymbol{\rho}$ に関しては、その和が 1 であるという制約を考えれば、

$$\rho_q = \frac{\sum_{i=1}^{N} h_{iq} + \alpha_q - 1}{\sum_{q=1}^{Q} \left(\sum_{i=1}^{N} h_{iq} + \alpha_q - 1 \right)} \tag{45}$$

が Q を最大とするパラメタの値となる。