# 1　EM アルゴリズム

対数尤度 $\log q(\boldsymbol{X};\boldsymbol{\theta})$ の下界は以下のようになり，これを $f$ とおく．

$$f(\boldsymbol{X};\boldsymbol{\theta}) = \int q(\boldsymbol{Z}) \log p(\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\theta}) d\boldsymbol{Z} \tag{1}$$

この下界を最大化するパラメータ $\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}$ を求める．まず，

$$q(\boldsymbol{Z}) = \prod_{n=1}^{N}\prod_{k=1}^{K} \gamma_{nk}^{z_{nk}}$$

$$p(\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\theta}) = \prod_{n=1}^{N}\prod_{k=1}^{K} (\pi_k N(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}))^{z_{nk}}$$

であり，これを式 (1) に代入して整理すると，次のようになる．

$$\begin{aligned}
f(\boldsymbol{X};\boldsymbol{\theta}) &= \int \prod_{n=1}^{N}\prod_{k=1}^{K} \gamma_{nk}^{z_{nk}} \log \prod_{n=1}^{N}\prod_{k=1}^{K} (\pi_k N(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}))^{z_{nk}} d\boldsymbol{Z} \\
&= \int \prod_{n=1}^{N}\prod_{k=1}^{K} \gamma_{nk}^{z_{nk}} \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk} \log(\pi_k N(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1})) d\boldsymbol{Z} \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} \int \prod_{n'=1}^{N}\prod_{k'=1}^{K} \gamma_{n'k'}^{z_{n'k'}} z_{nk} \log(\pi_k N(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1})) d\boldsymbol{Z} \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma_{nk} \log(\pi_k N(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1})) \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma_{nk} \left\{ \log \pi_k + \frac{1}{2}\log|\boldsymbol{\Lambda}_k| - \frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Lambda}_k (\boldsymbol{x}_n - \boldsymbol{\mu}_k) \right\} + const
\end{aligned} \tag{2}$$

まず，最適な $\boldsymbol{\mu}, \boldsymbol{\Lambda}$ を計算するために，式 (2) をそれぞれのパラメータに関して偏微分する．

$$\begin{aligned}
\frac{\partial f}{\partial \boldsymbol{\mu}_k} &= -\frac{1}{2}\sum_{n=1}^{N} \gamma_{nk} \boldsymbol{\Lambda}_k (\boldsymbol{x}_n - \boldsymbol{\mu}_k) \\
\therefore\ \ \boldsymbol{\mu}_k^* &= \frac{\sum_{n=1}^{N} \gamma_{nk} \boldsymbol{x}_n}{\sum_{n=1}^{N} \gamma_{nk}} = \frac{S_k\left[\boldsymbol{x}\right]}{S_.\left[1\right]} \\
\frac{\partial f}{\partial \boldsymbol{\Lambda}_k} &= \frac{1}{2}\sum_{n=1}^{N} \gamma_{nk} \left\{ \boldsymbol{\Lambda}_k - (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T \right\} \\
\therefore\ \ \boldsymbol{\Lambda}_k^* &= \frac{\sum_{n=1}^{N} \gamma_{nk} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T}{\sum_{n=1}^{N} \gamma_{nk}} = \frac{S_k\left[\boldsymbol{x}\boldsymbol{x}^T\right]}{S_k[1]} - \boldsymbol{\mu}_k \boldsymbol{\mu}_k^T
\end{aligned}$$

$\boldsymbol{\pi}$ については，$\sum_{k=1}^{K} \pi_k = 1$ の条件からラグランジュの未定乗数法を用いる．すなわち，次の式について

各 $\pi_k$ で微分することを考える．

$$g(\boldsymbol{X},\lambda;\boldsymbol{\theta})= f(\boldsymbol{X};\boldsymbol{\theta})+\lambda(\sum_{k=1}^{K}\pi_k-1)$$
$$\therefore\quad \frac{\partial g}{\partial \pi_k}=\sum_{n=1}^{N}\frac{\gamma_{nk}}{\pi_k}+\lambda$$

これをすべての $\pi_k$ について足し合わせることで，$\lambda=-N$ を得る．よって，

$$\pi_k^*=\frac{-\sum_{n=1}^{N}\gamma_{nk}}{\lambda}=\frac{S_k[1]}{N}$$

## 2 変分ベイズ

因子 $q(\boldsymbol{\pi})$，$q(\boldsymbol{\mu},\boldsymbol{\Lambda})$ の最適解は次のように書けることを確認しておく．

$$\log q^*(\boldsymbol{\pi})=\log p(\boldsymbol{\pi})+\langle\log p(\boldsymbol{Z}\mid\boldsymbol{\pi})\rangle_{q(\boldsymbol{Z})}+const \tag{3}$$
$$\log q^*(\boldsymbol{\mu},\boldsymbol{\Lambda})=\log p(\boldsymbol{\mu},\boldsymbol{\Lambda})+\langle\log p(\boldsymbol{X}\mid\boldsymbol{Z},\boldsymbol{\mu},\boldsymbol{\Lambda})\rangle_{q(\boldsymbol{Z})}+const \tag{4}$$

まず $q^*(\boldsymbol{\pi})$ を求める．式 (3) 中に現れる確率密度関数を書き下すと，それぞれ次のようになる．

$$p(\boldsymbol{\pi})=Dir(\boldsymbol{\pi}\mid\boldsymbol{\alpha}_0)=\frac{\Gamma(\sum_{k=1}^{K}\alpha_{0k})}{\prod_{k=1}^{K}\Gamma(\alpha_{0k})}\prod_{k=1}^{K}\pi_k^{\alpha_{0k}-1}$$
$$p(\boldsymbol{Z}\mid\boldsymbol{\pi})=\prod_{n=1}^{N}\prod_{k=1}^{K}\pi_k^{z_{nk}}$$

また，$q^*(\boldsymbol{Z})=\prod_{n=1}^{N}\prod_{k=1}^{K}\gamma_{nk}^{z_{nk}}$ より，

$$\langle z_{nk}\rangle_{q(\boldsymbol{Z})}=\gamma_{nk} \tag{5}$$

これらを利用すると，$\log q^*(\boldsymbol{\pi})$ は次のようになる．

$$\begin{aligned}\log q^*(\boldsymbol{\pi})&=\sum_{k=1}^{K}(\alpha_{0k}-1)\log\pi_k+\sum_{k=1}^{K}\sum_{n=1}^{N}\gamma_{nk}\log\pi_k+const\\&=\sum_{k=1}^{K}(\alpha_{0k}+S_k[1]-1)\log\pi_k+const\\&=\sum_{k=1}^{K}(\alpha_k-1)\log\pi_k+const\end{aligned}$$

ただし，$\alpha_k=\alpha_{0k}+S_k[1]$ とする．最後に両辺の指数をとり，適当に正規化してやることによって，$q^*(\boldsymbol{\pi})$ はディリクレ分布になる．

$$q^*(\boldsymbol{\pi})=Dir(\boldsymbol{\pi}\mid\boldsymbol{\alpha})$$

次に $q^*(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ を求める．まず，式 (4) 中に現れる確率密度関数を書き下すと，それぞれ次のようになる．

$$
\begin{aligned}
p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) &= \prod_{k=1}^{K} N(\boldsymbol{\mu}_k \mid \boldsymbol{m}_0, (\beta_0 \boldsymbol{\Lambda}_k)^{-1}) W(\boldsymbol{\Lambda}_k \mid \boldsymbol{W}_0, \nu_0) \\
p(\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) &= \prod_{n=1}^{N} \prod_{k=1}^{K} N(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1})^{z_{nk}}
\end{aligned}
$$

これらの密度関数と式 (5) を用いると，式 (4) は次のようになる．

$$
\begin{aligned}
\log q^*(\boldsymbol{\mu}, \boldsymbol{\Lambda}) &= \sum_{k=1}^{K} \left\{ \log N(\boldsymbol{\mu}_k \mid \boldsymbol{m}_0, (\beta_0 \boldsymbol{\Lambda}_k)^{-1}) + \log W(\boldsymbol{\Lambda}_k \mid \boldsymbol{W}_0, \nu_0)) \right\} \\
&\quad + \sum_{k=1}^{K} \sum_{n=1}^{N} \gamma_{nk} \log N(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}) + const \\
&= \sum_{k=1}^{K} \left[ \frac{1}{2} \left\{ \log |\beta_0 \boldsymbol{\Lambda}_k| - (\boldsymbol{\mu}_k - \boldsymbol{m}_0)^T (\beta_0 \boldsymbol{\Lambda})_{\boldsymbol{k}} (\boldsymbol{\mu}_{\boldsymbol{k}} - \boldsymbol{m}_0) \right\} \right. \\
&\quad + \frac{1}{2} \left\{ \sum_{n=1}^{N} \gamma_{nk} \log |\boldsymbol{\Lambda}_k| - (\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Lambda} (\boldsymbol{x}_n - \boldsymbol{\mu}_k) \right\} \\
&\quad \left. + \left\{ \frac{\nu_0 - D - 1}{2} \log |\boldsymbol{\Lambda}_k| - \frac{1}{2} Tr(\boldsymbol{W}_0^{-1} \boldsymbol{\Lambda}_k) \right\} \right] + const \\
&= \sum_{k=1}^{K} \left[ \frac{1}{2} \underbrace{\left\{ -(\boldsymbol{\mu}_k - \boldsymbol{m}_0)^T (\beta_0 \boldsymbol{\Lambda})_{\boldsymbol{k}} (\boldsymbol{\mu}_{\boldsymbol{k}} - \boldsymbol{m}_0) - (\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Lambda} (\boldsymbol{x}_n - \boldsymbol{\mu}_k) \right\}}_{(a)} \right. \\
&\quad + \left\{ \underbrace{\frac{1}{2} \sum_{n=1}^{N} \gamma_{nk} \log |\boldsymbol{\Lambda}_k| + \frac{\nu_0 - D - 1}{2} \log |\boldsymbol{\Lambda}_k|}_{(b)} - \frac{1}{2} Tr(\boldsymbol{W}_0^{-1} \boldsymbol{\Lambda}_k) \right\} \\
&\quad \left. + \underbrace{\frac{1}{2} \log |\beta_0 \boldsymbol{\Lambda}_k|}_{(c)} \right] + const
\end{aligned} \tag{6}
$$

次に，式 (6) 中の (a), (b), (c) について変形していく．

$$
\begin{aligned}
(a) &= \boldsymbol{\mu}_k^T (\underbrace{\beta_0 + \sum_{n=1}^{N} \gamma_{nk}}_{S_k[1]}) \boldsymbol{\Lambda}_k \boldsymbol{\mu}_k - 2 \boldsymbol{\mu}_k^T \boldsymbol{\Lambda}_k (\beta_0 m_0 + \underbrace{\sum_{n=1}^{N} \gamma_{nk}}_{S_k[1]}) + \boldsymbol{m}_0^T \beta_0 \boldsymbol{\Lambda}_k \boldsymbol{m}_0 + \sum_{n=1}^{N} \gamma_{nk} \boldsymbol{x}_n^T \boldsymbol{\Lambda}_k \boldsymbol{x}_n \\
&= \boldsymbol{\mu}_k^T (\underbrace{\beta_0 + S_k[1]}_{\beta_k}) \boldsymbol{\Lambda}_k \boldsymbol{\mu}_k - 2 \boldsymbol{\mu}_k^T \boldsymbol{\Lambda}_k (\underbrace{\beta_0 + S_k[1]}_{\beta_k}) \underbrace{\frac{\beta_0 m_0 + S_k[\boldsymbol{x}]}{\beta_0 + S_k[1]}}_{\boldsymbol{m}_k} + \boldsymbol{m}_0^T \beta_0 \boldsymbol{\Lambda}_k \boldsymbol{m}_0 + \sum_{n=1}^{N} \gamma_{nk} \boldsymbol{x}_n^T \boldsymbol{\Lambda}_k \boldsymbol{x}_n \\
&= (\boldsymbol{\mu}_k - \boldsymbol{m}_k)^T \beta_k \boldsymbol{\Lambda}_k (\boldsymbol{\mu}_k - \boldsymbol{m}_k) - \boldsymbol{m}_k^T \beta_k \boldsymbol{\Lambda}_k \boldsymbol{m}_k + \boldsymbol{m}_0^T \beta_0 \boldsymbol{\Lambda}_k \boldsymbol{m}_0 + \sum_{n=1}^{N} \gamma_{nk} \boldsymbol{x}_n^T \boldsymbol{\Lambda}_k \boldsymbol{x}_n
\end{aligned} \tag{7}
$$

$$
\begin{aligned}
(b) &= \frac{1}{2}S_k[1]\log|\mathbf{\Lambda}_k| + \frac{\nu_0 - D - 1}{2}\log|\mathbf{\Lambda}_k| \\
&= \frac{\nu_0 + S_k[1] - D - 1}{2}\log|\mathbf{\Lambda}_k| \\
&= \frac{\nu_k - D - 1}{2}\log|\mathbf{\Lambda}_k| \qquad (8)
\end{aligned}
$$

$$
(c) = \frac{1}{2}\log|\mathbf{\Lambda}_k| + const \qquad (9)
$$

ここで，

$$
\beta_k = \beta_0 + S_k[1], \quad \boldsymbol{m}_k = \frac{\beta_0 m_0 + S_k[\boldsymbol{x}]}{\beta_0 + S_k[1]}, \quad \nu_k = \nu_0 + S_k[1]
$$

とおいた．式 (7),(8),(9) を式 (6) に戻して整理すると，

$$
\begin{aligned}
\log q^*(\boldsymbol{\mu}, \mathbf{\Lambda}) = \sum_{k=1}^{K} \Bigg[ & \left\{ \frac{1}{2}\log|\mathbf{\Lambda}_k| - \frac{1}{2}(\boldsymbol{\mu}_k - \boldsymbol{m}_k)^T \beta_k \mathbf{\Lambda}_k (\boldsymbol{\mu}_k - \boldsymbol{m}_k) \right\} \\
& - \frac{1}{2}\left\{ \underbrace{Tr(\boldsymbol{W}_0^{-1}\mathbf{\Lambda}_k) - \boldsymbol{m}_k^T \beta_k \mathbf{\Lambda}_k \boldsymbol{m}_k + \boldsymbol{m}_0^T \beta_0 \mathbf{\Lambda}_k \boldsymbol{m}_0 + \sum_{n=1}^{N} \gamma_{nk} \boldsymbol{x}_n^T \mathbf{\Lambda}_k \boldsymbol{x}_n}_{(d)} \right\} \\
& + \frac{\nu_k - D - 1}{2}\log|\mathbf{\Lambda}_k| \Bigg] + const \qquad (10)
\end{aligned}
$$

次に，式 (10) 中の (d) について変形する．

$$
\begin{aligned}
(d) &= Tr(\boldsymbol{W}_0^{-1}\mathbf{\Lambda}_k) - Tr(\boldsymbol{m}_k^T \beta_k \mathbf{\Lambda}_k \boldsymbol{m}_k) + Tr(\boldsymbol{m}_0^T \beta_0 \mathbf{\Lambda}_k \boldsymbol{m}_0) + Tr(\sum_{n=1}^{N} \gamma_{nk} \boldsymbol{x}_n^T \mathbf{\Lambda}_k \boldsymbol{x}_n) \\
&= Tr(\boldsymbol{W}_0^{-1}\mathbf{\Lambda}_k) - Tr(\boldsymbol{m}_k \boldsymbol{m}_k^T \beta_k \mathbf{\Lambda}_k) + Tr(\boldsymbol{m}_0 \boldsymbol{m}_0^T \beta_0 \mathbf{\Lambda}_k) + Tr(\underbrace{\sum_{n=1}^{N} \gamma_{nk} \boldsymbol{x}_n \boldsymbol{x}_n^T}_{S_k[\boldsymbol{x}\boldsymbol{x}^T]} \mathbf{\Lambda}_k) \\
&= Tr((\boldsymbol{W}_0^{-1} - \boldsymbol{m}_k \boldsymbol{m}_k^T \beta_k + \boldsymbol{m}_0 \boldsymbol{m}_0^T \beta_0 + S_k[\boldsymbol{x}\boldsymbol{x}^T])\mathbf{\Lambda}_k) \\
&= Tr(W_k^{-1}\mathbf{\Lambda}_k) \qquad (11)
\end{aligned}
$$

ここで，$W_k^{-1} = \boldsymbol{W}_0^{-1} - \boldsymbol{m}_k \boldsymbol{m}_k^T \beta_k + \boldsymbol{m}_0 \boldsymbol{m}_0^T \beta_0 + S_k[\boldsymbol{x}\boldsymbol{x}^T]$ とした．最後に，式 (11) を (10) に戻して，

$$
\begin{aligned}
\log q^*(\boldsymbol{\mu}, \mathbf{\Lambda}) = \sum_{k=1}^{K} \Bigg[ & \left\{ \frac{1}{2}\log|\mathbf{\Lambda}_k| - \frac{1}{2}(\boldsymbol{\mu}_k - \boldsymbol{m}_k)^T \beta_k \mathbf{\Lambda}_k (\boldsymbol{\mu}_k - \boldsymbol{m}_k) \right\} \\
& + \left\{ \frac{\nu_k - D - 1}{2}\log|\mathbf{\Lambda}_k| - \frac{1}{2}Tr(\boldsymbol{W}_k^{-1}\mathbf{\Lambda}_k) \right\} \Bigg] + const \\
\therefore q^*(\boldsymbol{\mu}, \mathbf{\Lambda}) = \prod_{k=1}^{K} & N(\boldsymbol{\mu}_k \mid \boldsymbol{m}_k, (\beta_0 \mathbf{\Lambda}^{-1})) W(\mathbf{\Lambda}_k \mid \boldsymbol{W}_k, \nu_k) \qquad (12)
\end{aligned}
$$

## 3 最適なクラスター数 K

クラスター数 $K$ を変化させて複数回実行することで，最適な $K$ を決定することを考える．モデルの質を判断するための指標として，まず尤度を考えることができるが，尤度はパラメータの数が多いほど大きくなる傾向があり，そのまま利用することは難しい．このような問題に対処するために，パラメータ数による問題を考慮するような判断指標を設定することが考えられるが，そのような指標として代表的なものが AIC である．AIC は，尤度に加えてモデルのパラメータ数をペナルティとして考慮しているので，単純な尤度におけるパラメータ数の問題を軽減することが期待できる．

表 3 は，$K$ を変化させながら学習を行い，尤度と AIC の計算を行った結果である．前述の通り，尤度は $K$ の増加に伴って増加しているが，AIC にはそのような傾向は見られず，AIC によると最適なクラスター数 $K$ は 4 である．このことは，図 1 に示す，$K = 4$ に設定して EM アルゴリズムを実行した結果からも見て取ることができる．

表 1　クラスター数の変化に対する尤度と AIC

| $K$ | 尤度 | AIC |
|---|---|---|
| 1 | -68789.6033 | 137597.2065 |
| 2 | -63032.5094 | 126103.0181 |
| 3 | -60010.1741 | 120078.2530 |
| 4 | -56631.7363 | 113341.4635 |
| 5 | -56625.4261 | 113348.6788 |
| 6 | -56620.5372 | 113358.8811 |
| 7 | -56614.7430 | 113367.3005 |
| 8 | -56619.1064 | 113396.0193 |

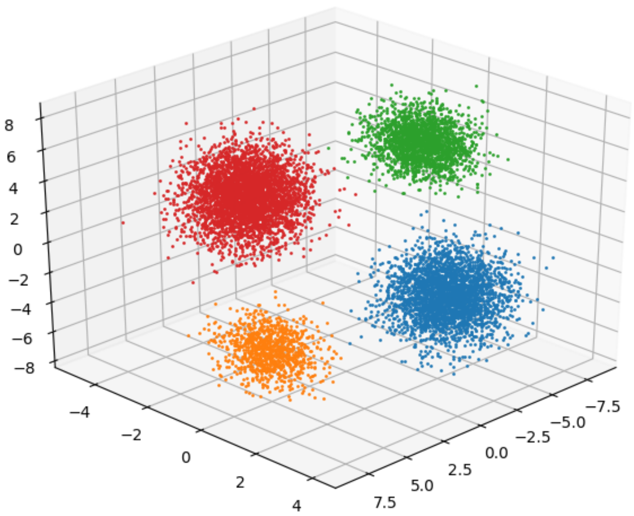


図 1　$K = 4$ の EM アルゴリズムによる分類結果