PRML の 9 章の数式の補足

サイボウズ・ラボ 光成滋生

2011 年 8 月 26 日

1 概要

この文章は『パターン認識と機械学習』（以下 PRML）の 9 章の式変形を一部埋めたものです. 間違い, 質問などございましたら herumi@nifty.com または twitterID:herumi までご連絡ください. 面倒なので特に紛らわしいと思わない限り $\boldsymbol{x}$ を x と書いたりします. また対数尤度関数を F と書くことが多いです.

2 復習

よく使ういくつかの式を書いておく. どれも今までに既に示したものである.
https://github.com/herumi/prml/raw/master/prml2.pdf,
https://github.com/herumi/prml/raw/master/prml3.pdf を参照.

2.1 行列の公式

$$
\begin{aligned}
&x^{\boldsymbol{T}}Ax = \mathrm{tr}(Axx^{\boldsymbol{T}}),\\
&\frac{\partial}{\partial A}\log|A| = (A^{-1})^{\boldsymbol{T}},\\
&\frac{\partial}{\partial x}\log|A| = \mathrm{tr}(A^{-1}\frac{\partial}{\partial x}A),\\
&\frac{\partial}{\partial A}\mathrm{tr}(A^{-1}B) = -(A^{-1}BA^{-1})^{\boldsymbol{T}}.
\end{aligned}
$$

2.2 微分

関数 f に対して対数関数の微分は

$$(\log f)' = \frac{f'}{f}.$$

よって逆に

$$f' = f\cdot(\log f)'.$$

ガウス分布など対数の微分が分かりやすいときによく使う.

2.3 ガウス分布

$$\mathcal{N} = \mathcal{N}(x|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2}}|\Sigma|^{-1/2}\exp(-\frac{1}{2}(x-\mu)^{\boldsymbol{T}}\Sigma^{-1}(x-\mu)).$$

期待値と分散について

$$\begin{aligned}
&E[x] = \mu,\\
&\mathrm{cov}[x] = \Sigma,\\
&E[xx^{\boldsymbol{T}}] = \mu\mu^{\boldsymbol{T}} + \Sigma,\\
&E[x^{\boldsymbol{T}}x] = \mu^{\boldsymbol{T}}\mu + \mathrm{tr}(\Sigma).
\end{aligned}$$

最後の式は 3 番目から出る.

$$E[x_i^2] = (\mu\mu^{\boldsymbol{T}})_{ii} + \Sigma_{ii} = \mu_i^2 + \Sigma_{ii}.$$

よって

$$E[x^{\boldsymbol{T}}x] = \sum_i E[x_i^2] = \mu^{\boldsymbol{T}}\mu + \mathrm{tr}(\Sigma).$$

3 混合ガウス分布

離散的な潜在変数を用いた混合ガウス分布の定式化. K 次元 2 値確率変数 z を考える（どれか一つの成分のみが 1 であとは 0）. つまり

$$\sum_k z_k = 1.$$

z の種類は K 個である. $0 \leq \pi_k \leq 1$ という係数を用いて

$$p(z_k = 1) = \pi_k$$

という確率分布を与える.

$$p(z) = \prod_k \pi_k^{z_k}.$$

$$p(x|z_k = 1) = \mathcal{N}(x|\mu_k, \Sigma_k)$$

なので

$$P(x|z) = \prod_k \mathcal{N}(x|\mu_k, \Sigma_k)^{z_k}.$$

これらを合わせて

$$\begin{aligned}
p(x) &= \sum_z p(z)p(x|z)\\
&= \sum_z \prod_k (\pi_k \mathcal{N}(x|\mu_k, \Sigma_k))^{z_k}\\
&\quad z_k \text{ はどれか一つのみが } 1 \text{（そのとき } \pi_k\text{）であとは } 0 \text{ なので}\\
&= \sum_k \pi_k \mathcal{N}(x|\mu_k, \Sigma_k).
\end{aligned}$$

x が与えられたときの z の条件付き確率 $p(z_k = 1|x)$ を $\gamma(z_k)$ とする.

$$\gamma(z_k) = \frac{p(z_k = 1)p(x|z_k = 1)}{\sum_j p(z_j = 1)p(x|z_j = 1)} = \frac{\pi_k \mathcal{N}(x|\mu_k, \Sigma_k)}{\sum_j \pi_j \mathcal{N}(x|\mu_j, \Sigma_j)}.$$

これを混合要素 k が観測値 x に対する負担率という.

4 混合ガウス分布の EM アルゴリズム

混合ガウス分布において観測したデータ集合を $X^{\boldsymbol{T}} = \{x_1, \ldots, x_N\}$, 対応する潜在変数を $Z^{\boldsymbol{T}} = \{z_1, \ldots, z_N\}$ とする. X は $N \times D$ 行列で Z は $N \times K$ 行列.

対数尤度関数の最大点の条件をもとめる.

$$F = \log p(X|\boldsymbol{\pi}, \mu, \Sigma) = \sum_{n=1}^{N} \log(\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n|\mu_j, \Sigma_j))$$

とする.

$$\frac{\partial}{\partial \mu} \log \mathcal{N}(x|\mu, \Sigma) = \frac{\partial}{\partial \mu}(-\frac{1}{2}(x-\mu)^{\boldsymbol{T}} \Sigma^{-1}(x-\mu)) = \Sigma^{-1}(x-\mu)$$

より

$$\frac{\partial}{\partial \mu}\mathcal{N} = \mathcal{N} \cdot (\frac{\partial}{\partial \mu} \log \mathcal{N}) = \mathcal{N} \cdot \Sigma^{-1}(x-\mu).$$

もちろんガウス分布の微分は普通にそのまましてもよい. だが今回は対数をとってから微分をとった方が, 微分してでてくる $\mathcal{N}$ が $\gamma(z_{nk})$ の一部となることを見通しやすいのでそうしてみた. さて $\mathcal{N}_{nk} = \mathcal{N}(x_n|\mu_k, \Sigma_k)$ とおいて

$$\begin{aligned}
\frac{\partial}{\partial \mu_k} F &= \sum_n \frac{\pi_k \frac{\partial}{\partial \mu_k}\mathcal{N}_{nk}}{\sum_j \pi_j \mathcal{N}_{nj}} \\
&= \sum_n (\frac{\pi_k \mathcal{N}_{nk}}{\sum_j \pi_j \mathcal{N}_{nj}}) \frac{\partial}{\partial \mu_k} \log \mathcal{N}_{nk} \\
&= \sum_n \gamma(z_{nk}) \frac{\partial}{\partial \mu_k} \log \mathcal{N}_{nk} \\
&= \Sigma_k^{-1}(\sum_n \gamma(z_{nk})(x_n - \mu_k)) = 0.
\end{aligned}$$

よって

$$\sum_n \gamma(z_{nk}) x_n - (\sum_n \gamma(z_{nk}))\mu_k = 0.$$

$$N_k = \sum_n \gamma(z_{nk})$$

とおくと

$$\mu_k = \frac{1}{N_k} \sum_n \gamma(z_{nk}) x_n.$$

これは μ_k が X の重みつき平均であることを示している.

次に Σ_k に関する微分を考える.

$$\mathcal{N} = \mathcal{N}(x|\mu, \Sigma)$$

のとき

$$\log \mathcal{N} = -\frac{D}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}\mathrm{tr}(\Sigma^{-1}(x-\mu)(x-\mu)^{\boldsymbol{T}})$$

なので $\Sigma^{\boldsymbol{T}} = \Sigma$ だから

$$\frac{\partial}{\partial\Sigma}(\log\mathcal{N}) = -\frac{1}{2}(\Sigma^{-1}) + \frac{1}{2}(\Sigma^{-1}(x-\mu)(x-\mu)^{\boldsymbol{T}}\Sigma^{-1}).$$

よって μ_k の微分と同様にして

$$\begin{aligned}\frac{\partial}{\partial\Sigma_k}F &= \sum_n \gamma(z_{nk})\frac{\partial}{\partial\Sigma_k}\log\mathcal{N}_{nk} \\ &= \sum_n \gamma(z_{nk})(-\frac{1}{2}(\Sigma_k^{-1}) + \frac{1}{2}(\Sigma_k^{-1}(x_n-\mu_k)(x_n-\mu_k)^{\boldsymbol{T}}\Sigma_k^{-1})) = 0.\end{aligned}$$

よって

$$\sum_n \gamma(z_{nk})(I - (x_n-\mu_k)(x_n-\mu_k)^{\boldsymbol{T}}\Sigma_k^{-1}) = 0.$$

$$\Sigma_k = \frac{1}{N_k}\sum_n \gamma(z_{nk})(x_n-\mu_k)(x_n-\mu_k)^{\boldsymbol{T}}.$$

最後に π_k に関する微分を考える．$\sum_k \pi_k = 1$ の制約を入れる．

$$G = F + \lambda(\sum_k \pi_k - 1)$$

とすると

$$\frac{\partial}{\partial\pi_k}G = \sum_n \frac{\mathcal{N}_{nk}}{\sum_j \pi_j\mathcal{N}_{nj}} + \lambda = \sum_n \gamma(z_{nk})/\pi_k + \lambda = N_k/\pi_k + \lambda = 0.$$

つまり

$$N_k = -\lambda\pi_k.$$

よって

$$N = \sum_k N_k = \sum_k(-\lambda\pi_k) = -\lambda.$$

よって

$$\pi_k = \frac{N_k}{-\lambda} = \frac{N_k}{N}.$$

5 混合ガウス分布再訪

$$p(z) = \prod_k \pi_k^{z_k},$$

$$p(x|z) = \prod_k \mathcal{N}(x|\mu_k,\Sigma_k)^{z_k}$$

より

$$F = \log p(X, Z|\mu, \Sigma, \boldsymbol{\pi}) = \log(\prod_{n,k} \pi_k^{z_{nk}} \mathcal{N}(x_n|\mu_k, \Sigma_k)^{z_{nk}})$$
$$= \sum_{n,k} z_{nk}(\log \pi_k + \log \mathcal{N}_{nk})).$$

z_n は $(0, 0, \ldots, 1, 0, \ldots, 0)$ の形で $\sum_k \pi_k = 1$ の制約条件を入れると上式の微分を考えると

$$G = F + \lambda(\sum_k \pi_k - 1)$$

として

$$\frac{\partial}{\partial \pi_k} G = \sum_n z_{nk} \frac{1}{\pi_k} + \lambda = (\sum_n z_{nk})/\pi_k + \lambda = 0.$$

よって

$$\pi_k = -\frac{1}{\lambda} \sum_n z_{nk}.$$

$$\sum_k \pi_k = -\frac{1}{\lambda} \sum_{n,k} z_{nk} = -\frac{N}{\lambda} = 1.$$

よって $\lambda = -N$. つまり

$$\pi_k = \frac{1}{N} \sum_n z_{nk}.$$

完全データ集合についての対数尤度関数の最大化は解けるが, 潜在変数が分からない場合の不完全データに関する対数尤度関数の最大化は困難. この場合は潜在変数の事後分布に関する完全データ尤度関数の期待値を考える.

$$p(Z|X, \mu, \Sigma, \boldsymbol{\pi}) = \frac{p(X, Z|\mu, \Sigma, \pi)}{p(X|\mu, \Sigma, \boldsymbol{\pi})} \propto \prod_{n,k} (\pi_k \mathcal{N}_{nk})^{z_{nk}}.$$

$$E[z_{nk}] = \frac{\sum_{z_n} z_{nk} \prod_j (\pi_j \mathcal{N}_{nj})^{z_{nj}}}{\sum_{z_n} \prod_j (\pi_j \mathcal{N}_{nj})^{z_{nj}}} = \frac{\pi_k \mathcal{N}_{nk}}{\sum_j \pi_j \mathcal{N}_{nj}} = \gamma(z_{nk}).$$

よって

$$F = E_Z[\log p(X, Z|\mu, \Sigma, \boldsymbol{\pi})] = \sum_{n,k} \gamma(z_{nk})(\log \pi_k + \log \mathcal{N}_{nk}).$$

まずパラメータ μ, Σ, $\boldsymbol{\pi}$ を適当に決めて負担率 $\gamma(z_{nk})$ を求め, それを fix して μ_k, Σ_k, π_k について F を最大化. 今までと同様にできる. $F' = F + \lambda(\sum_k \pi_k - 1)$ として

$$\frac{\partial}{\partial \pi_k} F' = \sum_n \gamma(z_{nk})(1/\pi_k) + \lambda = 0$$

より

$$\sum_n \gamma(z_{nk}) = \lambda \pi_k.$$

$$\sum_{n,k} \gamma(z_{nk}) = -\lambda(\sum_k \pi_k) = -\lambda = N$$

より

$$\pi_k = \frac{1}{N}\sum_n \gamma(z_{nk}) = \frac{N_k}{N}.$$

$$\frac{\partial}{\partial \mu_k}F = \sum_n \gamma(z_{nk})(-\Sigma_k^{-1}(x_n - \mu_k)) = \Sigma_k^{-1}(\sum_n \gamma(z_{nk})x_n - (\sum_n \gamma(z_{nk}))\mu_k) = 0.$$

よって

$$\mu_k = \frac{1}{N_k}\sum_n \gamma(z_{nk})x_n.$$

$$\frac{\partial}{\partial \Sigma_k}F = \sum_n \gamma(z_{nk})\frac{\partial}{\partial \Sigma_k}\log \mathcal{N}_{nk} = 0$$

として同様（流石に略）.

6 K-means との関連

式 (9.43) は不正確. E ではなく ϵE を考えないと (9.43) の右辺にはならない. 式 (9.40) を E とおく.

$$E = \sum_{n,k}\gamma(z_{nk})(\log \pi_k + \log \mathcal{N}(x_n|\mu_k, \Sigma_k)).$$

ϵE に

$$\mathcal{N}(x|\mu_k, \Sigma_k) = \frac{1}{(2\pi\epsilon)^{D/2}}\exp(-\frac{1}{2\epsilon}||x - \mu_k||^2)$$

を代入する.

$$\epsilon E = \sum_{n,k}\gamma(z_{nk})(\epsilon \log \pi_k - \frac{D}{2}\epsilon \log(2\pi\epsilon) - \frac{1}{2}||x_n - \mu_k||^2).$$

$\epsilon \to 0$ で

$$\gamma(z_{nk}) \to r_{nk},$$

$$\epsilon \log \pi_k \to 0,$$

$$\epsilon \log(2\pi\epsilon) \to 0$$

より

$$\epsilon E \to -\frac{1}{2}\sum_{n,k} r_{nk}||x_n - \mu_k||^2 = -J.$$

よって期待完全データ対数尤度の最大化は J の最小化と同等.

7 混合ベルヌーイ分布

$x = (x_1, \ldots, x_D)^{\boldsymbol{T}}$, $\mu = (\mu_1, \ldots, \mu_D)^{\boldsymbol{T}}$ とする.

$$p(x|\mu) = \prod_{i=1}^{D}\mu_i^{x_i}(1 - \mu_i)^{(1-x_i)}.$$

$E[x] = \mu$ は容易に分かる.

$$E[x_i x_j] = \begin{cases} \mu_i \mu_j & (i \neq j) \\ \mu_i & (i = j). \end{cases}$$

よって

$$\mathrm{cov}[x]_{ij} = E[(x-\mu)(x-\mu)^{\boldsymbol{T}}]_{ij} = E[x_i x_j] - (\mu\mu^{\boldsymbol{T}})_{ij} = (\mu_i - \mu_i^2)\delta_{ij}$$

より

$$\mathrm{cov}[x] = \mathrm{diag}(\mu_i(1-\mu_i)).$$

$\mu = \{\mu_1, \ldots, \mu_K\}$, $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_K\}$ として次の混合分布を考えよう.

$$p(x|\mu_k) = \prod_i \mu_{ki}^{x_i}(1-\mu_{ki})^{(1-x_i)}.$$

$$E[x] = \int x p(x|\mu)\, dx = \sum_k \pi_k \int x p(x|\mu_k)\, dx = \sum_k \pi_k E_k[x] = \sum_k \pi_k \mu_k.$$

$$E_k[xx^{\boldsymbol{T}}] = \mathrm{cov}_k[x] + \mu_k \mu_k{}^{\boldsymbol{T}} = \Sigma_k + \mu_k \mu_k{}^{\boldsymbol{T}}$$

より

$$\mathrm{cov}[x] = E[(x-E[x])(x-E[x])^{\boldsymbol{T}}] = E[xx^{\boldsymbol{T}}] - E[x]E[x]^{\boldsymbol{T}} = \sum_k \pi_k(\Sigma_k + \mu_k \mu_k{}^{\boldsymbol{T}}) - E[x]E[x]^{\boldsymbol{T}}.$$

データ集合 $X = \{x_1, \ldots, x_N\}$ が与えられたとき, 対数尤度関数は

$$\log p(X|\mu, \boldsymbol{\pi}) = \sum_n \log(\sum_k \pi_k p(x_n|\mu_k)).$$

対数の中に和があるので解析的に最尤解をもとめられない. EM アルゴリズムを使う. x に対応する潜在変数を $z = (z_1, \ldots, z_K)^{\boldsymbol{T}}$ を導入する. どれか一つのみ 1 でその他は 0 のベクトルである. z の事前分布を

$$p(z|\pi) = \prod_k \pi_k^{z_k}$$

とする. z が与えられたときの条件付き確率は

$$p(x|z, \mu) = \prod_k p(x|\mu_k)^{z_k}.$$

$$p(x, z|\mu, \boldsymbol{\pi}) = p(x|z, \mu)p(z|\boldsymbol{\pi}) = \prod_k (\pi_k p(x|\mu_k))^{z_k}.$$

よって

$$p(x|\mu, \boldsymbol{\pi}) = \sum_z p(x, z|\mu, \boldsymbol{\pi}) = \sum_k \pi_k p(x|\mu_k).$$

完全データ対数尤度関数は $X = \{x_n\}$, $Z = \{z_n\}$ として

$$\begin{aligned} \log p(X, Z|\mu, \boldsymbol{\pi}) &= \sum_{n,k} z_{nk} \underbrace{(\log \pi_k + \sum_i x_{ni} \log \mu_{ki} + (1-x_{ni}) \log(1-\mu_{ki}))}_{=: A_{nk}} \\ &= \sum_{n,k} z_{nk} A_{nk}. \end{aligned}$$

$$
\begin{aligned}
E[z_{nk}] &= \frac{\sum_{z_n} z_{nk} \prod_j (\pi_j p(x_n|\mu_j))^{z_{nj}}}{\sum_{z_n} \prod_j (\pi_j p(x_n|\mu_j)^{z_{nj}}} \\
&\quad (z_{nk} = 1 \text{ となるものだけとるので}) \\
&= \frac{\pi_k p(x_n|\mu_k)}{\sum_j \pi_j p(x_n|\mu_j)}
\end{aligned}
$$

を $\gamma(z_{nk})$ とおく. すると

$$
E_Z[\log p(X, Z)|\mu, \boldsymbol{\pi})] = \sum_{n,k} \gamma(z_{nk}) A_{nk}.
$$

$$
N_k = \sum_n \gamma(z_{nk}), \quad \bar{x}_k = \frac{1}{N_k} \sum_n \gamma(z_{nk}) x_n
$$

とおく.

$$
\begin{aligned}
F &= E_Z[\log p(X, Z|\mu, \boldsymbol{\pi})] \\
&= \sum_k (\log \pi_k)(\sum_n \gamma(z_{nk})) + \sum_{k,i} \log \mu_{ki} (\sum_n \gamma(z_{nk}) x_{ni}) + \sum_{k,i} \log(1 - \mu_{ki})(\sum_n \gamma(z_{nk})(1 - x_{ni})) \\
&= \sum_k N_k \log \pi_k + \sum_{k,i} N_k \bar{x}_{ki} \log \mu_{ki} + \sum_{k,i} \log(1 - \mu_{ki}) N_k (1 - \bar{x}_{ki}).
\end{aligned}
$$

μ_{ki} に関する最大化.

$$
\begin{aligned}
\frac{\partial}{\partial \mu_{ki}} F &= N_k \bar{x}_{ki} \frac{1}{\mu_{ki}} + \frac{-1}{1 - \mu_{ki}} N_k (1 - \bar{x}_{ki}) \\
&= \frac{N_k}{\mu_{ki}(1 - \mu_{ki})} (\bar{x}_{ki}(1 - \mu_{ki}) - (1 - \bar{x}_{ki})\mu_{ki}) = 0.
\end{aligned}
$$

よって

$$
\bar{x}_{ki} - \bar{x}_{ki}\mu_{ki} - \mu_{ki} + \bar{x}_{ki}\mu_{ki} = \bar{x}_{ki} - \mu_{ki} = 0.
$$

よって

$$
\mu_k = \bar{x}_k.
$$

π_k に関する最適化. $G = F + \lambda(\sum_k \pi_k - 1)$ とすると

$$
\frac{\partial}{\partial \pi_k} G = \frac{N_k}{\pi_k} + \lambda = 0.
$$

よって

$$
N_k = -\lambda \pi_k, \quad N = \sum_k N_k = -\lambda \sum_k \pi_k = -\lambda.
$$

つまり $\lambda = -N$ となり

$$
\pi_k = \frac{N_k}{N}.
$$

$0 \leq p(x_n|\mu_k) \leq 1$ より

$$
\log p(X|\mu, \boldsymbol{\pi}) = \sum_n \log(\sum_k \pi_k p(x_n|\mu_k)) \leq \sum \log(\sum_k \pi_k) = 0.
$$

よって尤度関数が発散することはない.

8 ベイズ線形回帰に関する EM アルゴリズム

EM アルゴリズムに基づいてベイズ線形回帰を考えてみる. w を潜在関数と見なしてそれを最大化する方針を採る.

$$p(w|t) = \mathcal{N}(w|m_N, S_N)$$

で w の事後分布が求まっているとする.

$$p(t|w,\beta) = \prod_n \mathcal{N}(t_n|w^{\boldsymbol{T}}\phi(x_n), \beta^{-1}),$$

$$p(w|\alpha) = \mathcal{N}(w|0, \alpha^{-1}I)$$

であった. このとき完全データ対数尤度関数は

$$\log p(t, w|\alpha, \beta) = \log p(t|w,\beta) + \log p(w|\alpha).$$

なので

$$\begin{aligned} F &= E[\log p(t, w|\alpha, \beta)] \\ &= E[\sum_n (\frac{1}{2}\log(\frac{\beta}{2\pi}) - \frac{\beta}{2}(t_n - w^{\boldsymbol{T}}\phi_n)^2) + \frac{M}{2}\log(\frac{\alpha}{2\pi}) - \frac{\alpha}{2}w^{\boldsymbol{T}}w] \\ &= \frac{M}{2}\log(\frac{\alpha}{2\pi}) - \frac{\alpha}{2}E[w^{\boldsymbol{T}}w] + \frac{N}{2}\log(\frac{\beta}{2\pi}) - \frac{\beta}{2}\sum_n E[(t_n - w^{\boldsymbol{T}}\phi_n)^2]. \end{aligned}$$

α に関する最大化

$$\frac{\partial}{\partial\alpha}F = \frac{M}{2}\frac{1}{\alpha} - \frac{1}{2}E[w^{\boldsymbol{T}}w] = 0.$$

よって

$$\alpha = \frac{M}{E[w^{\boldsymbol{T}}w]} = \frac{M}{m_N{}^{\boldsymbol{T}}m_N + \mathrm{tr}(S_N)}.$$

β に関する最大化

$$\frac{\partial}{\partial\beta}F = \frac{N}{2}\frac{1}{\beta} - \frac{1}{2}\sum_n E[(t_n - w^{\boldsymbol{T}}\phi_n)^2] = 0.$$

よって

$$\frac{1}{\beta} = \frac{1}{N}\sum_n E[(t_n - w^{\boldsymbol{T}}\phi_n)^2].$$

9 一般の EM アルゴリズム

潜在変数をもつ確率モデルの最尤解を求めるための一般的手法. X を確率変数, Z を潜在変数, θ をパラメータとする. 完全データ対数尤度関数 $\log p(X, Z|\theta)$ の最適化は容易であるという仮定の元で目的は $p(X|\theta) = \sum_Z p(X, Z|\theta)$ の最大化.

Z に対する分布を $q(Z)$ とする

$$p(X, Z|\theta) = p(Z|X, \theta)p(X|\theta).$$

$$\mathcal{L}(q,\theta) = \sum_Z q(Z)\log\frac{p(X,Z|\theta)}{q(Z)},$$

$$\mathrm{KL}(q||p) = -\sum_Z q(Z)\log\frac{p(Z|X,\theta)}{q(Z)}$$

とおく. $\mathrm{KL}(q||p)$ は $q(Z)$ と事後分布 $p(Z|X,\theta)$ との距離なので常に 0 以上
(cf. `https://github.com/herumi/prml/raw/master/prml3.pdf`).

$$\begin{aligned}\mathcal{L}(q,\theta) + \mathrm{KL}(q||p) &= \sum_Z q(Z)\log\frac{p(X,Z|\theta)}{p(Z|X,\theta)}\\ &= \sum_Z q(Z)\log p(X|\theta)\\ &= \log p(X|\theta).\end{aligned}$$

よって

$$\begin{aligned}\log p(X|\theta) &= \mathcal{L}(q,\theta) + \mathrm{KL}(q||p)\\ &\geq \mathcal{L}(q,\theta).\end{aligned}$$

したがって $\mathcal{L}(q,\theta)$ は $\log p(X|\theta)$ の下界.

パラメータの現在の値が θ^o だったときに

E ステップでは θ^o を固定して $\mathcal{L}(q,\theta)$ を $q(Z)$ について最大化する. $\log p(X|\theta)$ は q によらないのでそれは $\mathrm{KL}=0$ のとき, つまり

$$q(Z) = p(Z|X,\theta^o)$$

のときである.

M ステップでは $q(Z)$ を固定して $\mathcal{L}(q,\theta)$ を θ について最大化する. その θ を θ^n とする. 最大値になっていなければ, 必ず $\mathcal{L}$ が増加し, $\log p(X|\theta)$ も増える. このときの $\mathrm{KL}(q||p)$ は θ^o を使って計算されていた(そして値は 0)ので新しい θ^n を使って計算し直すと通常正となる.

$q(Z) = p(Z|X,\theta^o)$ より

$$\begin{aligned}q(Z) &= \sum_Z q(Z)\log\frac{p(X,Z|\theta)}{q(Z)}\\ &= \sum_Z p(Z|X,\theta^o)\log p(X,Z|\theta) - \sum_Z p(Z|X,\theta^o)\log p(Z|X,\theta^o)\\ &\quad \mathcal{Q}(\theta,\theta^o) = \sum_Z p(Z|X,\theta^o)\log p(X,Z|\theta) \text{ とおいて}\\ &= \mathcal{Q}(\theta,\theta^o) + \theta\text{に非依存}.\end{aligned}$$

つまり $\mathcal{L}(q,\theta)$ の最大化は $\mathcal{Q}(\theta,\theta^o)$ の最大化に等しい.

10 混合ガウス分布のオンライン版 EM アルゴリズム

各 EM のステップで一つのデータ点のみの更新を行うことを考える. これは m 番目のデータ以外を潜在変数とする EM アルゴリズムとみなすことができる.

E ステップでは分布 $p(Z|X,\theta^o)$ を求める必要があるが, M ステップで必要な μ_k,Σ_k, π_k の更新式の右辺を見ると必要なデータは $\gamma(z_{nk})$ のみであることが分かる. つまりそれらの差分さえ分かればアルゴリズムを書き下すことができる.

$$N_k = \sum_n \gamma(z_{nk})$$

を m 番目の値だけ更新する. 新しい値を N'_k とすると

$$N'_k = \sum_{n\neq m} \gamma(z_{nk}) + \gamma'(z_{mk}) = N_k + \gamma'(z_{mk}) - \gamma(z_{mk}).$$

$d := \gamma'(z_{mk}) - \gamma(z_{mk})$ とおくと $N'_k = N_k + d$. $\pi_k = N_k/N$ なので

$$\pi'_k = \frac{N'_k}{N} = \frac{N_k + d}{N_k} = \pi_k + \frac{d}{N}.$$

$\mu_k = (1/N_k)\sum_n \gamma(z_{nk})x_n$ より

$$\mu'_k = \frac{1}{N'_k}(\sum_{n\neq m} \gamma(z_{nk})x_n + \gamma'(z_{mk})x_m).$$

よって

$$N'_k\mu'_k = N_k\mu_k - \gamma(z_{mk})x_m + \gamma'(z_{mk})x_m = (N'_k - d)\mu_k + dx_m = N'_k\mu_k + d(x_m - \mu_k).$$

よって

$$\mu'_k = \mu_k + \frac{d}{N'_k}(x_m - \mu_k).$$

$S := \Sigma_k = (1/N_k)\sum_n \gamma(z_{nk})(x_n - \mu_k)(x_n - \mu_k)^{\boldsymbol{T}}$ より $S' := \Sigma'_k$ とすると

$$N'_kS' = \sum_{n\neq m} \gamma(z_{nk})(x_n - \mu'_k)(x_n - \mu'_k)^{\boldsymbol{T}} + \gamma'(z_{mk})(x_m - \mu'_k)(x_m - \mu'_k)^{\boldsymbol{T}}.$$

以下式変形をひたすら行う.

$$x_m - \mu'_k = x_m - \mu_k - \frac{d}{N'_k}(x_m - \mu_k) = (1 - \frac{d}{N'_k})(x_m - \mu_k) = \frac{N_k}{N'_k}(x_m - \mu_k).$$

$$x_n - \mu'_k = (x_n - \mu_k) - \frac{d}{N'_k}(x_m - \mu_k).$$

$A := (x_m - \mu_k)(x_m - \mu_k)^{\boldsymbol{T}}$ とおくと

$$(x_n - \mu'_k)(x_n - \mu'_k)^{\boldsymbol{T}} = (x_n - \mu_k)(x_n - \mu_k)^{\boldsymbol{T}} - 2\frac{d}{N'_k}(x_n - \mu_k)(x_m - \mu_k)^{\boldsymbol{T}} + \frac{d^2}{N'^2_k}A.$$

$$\sum_{n\neq m} \gamma(z_{nk})(x_n - \mu_k)(x_n - \mu_k)^{\boldsymbol{T}} = N_kS - \gamma(z_{mk})(x_m - \mu_k)(x_m - \mu_k)^{\boldsymbol{T}} = N_kS - \gamma(z_{mk})A.$$

$$\begin{aligned}\sum_{n\neq m} \gamma(z_{nk})(x_n - \mu_k) &= \sum_{n\neq m} \gamma(z_{nk})x_n - \sum_{n\neq m} \gamma(z_{nk})\mu_k = N_k\mu_k - \gamma(z_{mk})x_m - (N_k - \gamma(z_{mk}))\mu_k \\ &= -\gamma(z_{mk})(x_m - \mu_k).\end{aligned}$$

よって $\gamma := \gamma(z_{mk})$ とおくと

$$\begin{aligned}
N'_k S' &= N_k S - \gamma A + 2\frac{d}{N'_k}\gamma A + (N_k - \gamma)\frac{d^2}{N'^2_k}A + (\gamma + d)A\frac{N_k^2}{N'^2_k} \\
&= N_k S + \frac{A}{N'^2_k}(-\gamma(N_k + d)^2 + 2d\gamma(N_k + d) + (N_k - \gamma)d^2 + (\gamma + d)N_k^2) \\
&= N_k S + \frac{A}{N'^2_k}(-\gamma N_k^2 - 2d\gamma N_k - \gamma d^2 + 2d\gamma N_k + 2d^2\gamma + N_k d^2 - \gamma d^2 + \gamma N_k^2 + dN_k^2) \\
&= N_k S + \frac{A}{N'^2_k}N_k d(N_k + d) = N_k S + \frac{AN_k d}{N'_k}.
\end{aligned}$$

よって

$$S' = \frac{N_k}{N'_k}(S + \frac{d}{N'_k}(x_m - \mu_k)(x_m - \mu_k)^{\boldsymbol{T}}).$$