EM algorithm for mixture models

Department of Statistics

November 15, 2023

University of Seoul

Mixture distribution

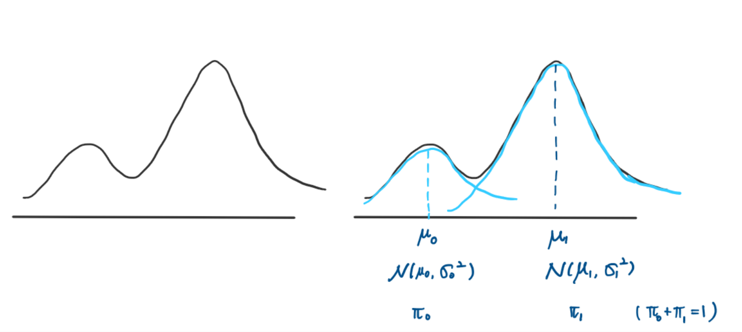

Figure 1: (Left) density of data distributions; (Right) Modeling of a two-component mixture distribution

Two component mixture

- $X|Z = 1 \sim N(\mu_1, \sigma_1^2)$
- $X|Z = 0 \sim N(\mu_0, \sigma_0^2)$
- $Z \sim \text{Bin}(1, \pi)$, $\pi \in (0, 1)$

Denote the pdf of normal distribution with mean μ and variance σ^2 by $\phi(\cdot; \mu, \sigma)$. Then, the pdf of (X, Z) is given by

$$\begin{aligned} f_{X,Z}(x, z; \theta) &= f_{X|Z}(x|z) \times f_Z(z) \\ &= \phi(x; \mu_1, \sigma_1^2)^z \phi(x; \mu_0, \sigma_0^2)^{1-z} \pi^z (1 - \pi)^{1-z}, \end{aligned} \tag{1}$$

where $\theta = (\mu_1, \sigma_1^2, \mu_0, \sigma_0^2, \pi)$.

(Density)

$$\Pr(X \le x, Z = z) \quad = \quad \underbrace{\int_{-\infty}^{x} \phi(x; \mu_z, \sigma_z) dx}_{\Pr(X \le x | Z = z)} \Pr(Z = z),$$

$$\begin{aligned}
\Pr(X \le x) &= \Pr(X \le x, Z = 0) + \Pr(X \le x, Z = 1) \\
&= (1 - \pi) \int_{-\infty}^{x} \phi(x; \mu_0, \sigma_0) dx + \pi \int_{-\infty}^{x} \phi(x; \mu_1, \sigma_1) dx \\
&= \int_{-\infty}^{x} \underbrace{[(1 - \pi)\phi(x; \mu_0, \sigma_0) + \pi\phi(x; \mu_1, \sigma_1)]}_{\text{pdf}} dx
\end{aligned}$$

The marginal pdf of X is given by

$$
\begin{aligned}
f_X(x;\theta) &= \int f_{X,Z}(x,z;\theta)du(z) \\
&= \pi\phi(x;\mu_1,\sigma_1^2) + (1-\pi)\phi(x;\mu_0,\sigma_0^2),
\end{aligned}
$$

where $u(z)$ is a counting measure on $\{0,1\}$.

Likelihood and MLE

Let $\{(x_i, z_i)\}_{i=1}^n$ be IID random sample of (X, Z). (1) defines the (complete) loglikelihood as

$$\begin{aligned} l^c(\theta) &= \sum_{i=1}^n \Big(z_i \log \phi(x_i; \mu_1, \sigma_1^2) + (1 - z_i) \log \phi(x_i; \mu_0, \sigma_0^2) \\ & z_i \log(\pi) + (1 - z_i) \log(1 - \pi) \Big). \end{aligned}$$

Assume that z is observed. Let $B_1 = \{i : z_i = 1\}$, $n_1 = |B_1|$, $B_0 = \{i : z_i = 0\}$ and $n_0 = |B_0|$ then

$$\begin{aligned} l^c(\theta) &= \sum_{i \in B_1} \log \phi(x_i; \mu_1, \sigma_1^2) + \sum_{i \in B_0} \log \phi(z_i; \mu_0, \sigma_0^2) \\ &\quad + n_1 \log(\pi) + n_0 \log(1 - \pi), \end{aligned}$$

where $n_2 = n - n_1$. Thus, the loglikelihood function is separable, and the MLE is given by

- $\hat{\mu}_1 = \bar{x}_1 = \sum_{i \in B_1} x_i / n_1$ and $\hat{\sigma}_1^2 = \sum_{i \in B_1} (x_i - \bar{x}_1)^2 / n_1$
- $\hat{\mu}_0 = \bar{x}_0 = \sum_{i \in B_0} x_i / n_0$ and $\hat{\sigma}_0^2 = \sum_{i \in B_0} (x_i - \bar{x}_0)^2 / n_0$
- $\hat{\pi} = n_1 / (n_1 + n_0)$

Likelihood and MLE

Suppose that $\{z_i\}$ is missing. What is the MLE of the considered model? Note that the likelihood should be defined by observations. The (observed) likelihood is given by

$$\begin{aligned} L^o(\theta) &= \prod_{i=1}^{n} f(x_i;\theta) \\ &= \prod_{i=1}^{n} \left(\pi\phi(x_i;\mu_1,\sigma_1^2) + (1-\pi)\phi(x_i;\mu_0,\sigma_0^2)\right). \end{aligned}$$

The maximizer of $L^o(\theta)$ is the MLE of the model. Hereafter, denote $\log L^o(\theta)$ by $l^o(\theta)$.

Difficulty of obtaining the MLE

Likelihood function of the normal mixture model has numerous poles, points not defined with an infinite limit.

- Choose an arbitrary $j \in \{1, \cdots, n\}$ and let $\mu_1 = x_j$. In addition fix μ_0 and $\sigma_0^2 > 0$ arbitrary.
- As $\sigma_1^2 \to 0$ we know that

$$\phi(x_i; \mu_1, \sigma_1^2) \to \begin{cases} \infty & \text{for } i = j \\ 0 & \text{for } i \neq j, \end{cases}$$

  which implies

$$\prod_{i=1}^{n} \left(\pi\phi(x_i; \mu_1, \sigma_1^2) + (1-\pi)\phi(x_i; \mu_0, \sigma_0^2)\right) \to \infty$$

That is, there exist at least n more poles in the likelihood function.

Difficulty in obtaining the MLE

The objective function is nonconvex with numerous saddle points. This problem is closely related to the identifiability problem on indexing the groups (or clusters).

- Suppose that we have estimate $\hat{\pi} = 0.3$, $\hat{\mu}_1 = 1$, $\hat{\mu}_0 = 0$, $\sigma_1 = 1$ and $\sigma_0 = 1.5$. But even if we let $\hat{\pi} = 0.7$, $\hat{\mu}_1 = 0$, $\hat{\mu}_0 = 1$, $\sigma_1 = 1.5$ and $\sigma_0 = 1$, the likelihood does not change. This means that the model is not identifiable.
- There are $M!$ combinations of parameter pairs in the M component mixture problem.

EM algorithm for Two-component mixture

In the complete loglikelihood, we treat z_i as a random variable. For convenience, consider a one-sample complete loglikelihood.

$$
\begin{aligned}
l^c(\theta) &= z_1 \log \phi(x_1; \mu_1, \sigma_1^2) + (1 - z_i) \log \phi(x_1; \mu_0, \sigma_0^2) \\
&\quad + z_1 \log(\pi) + (1 - z_1) \log(1 - \pi)
\end{aligned}
$$

Since x_1 is observed, $\phi(x_1; \mu_1, \sigma_1^2)$ and $\phi(x_1; \mu_0, \sigma_0^2)$ are functions of $(\mu_1, \sigma_1^2, \mu_0, \sigma_0^2)$. If we set a distribution of z_1 (Bernoulli dist.), we can compute

$$
\begin{aligned}
\mathrm{E}l^c(\theta) &= \mathrm{E}(z_1) \log \phi(x_1; \mu_1, \sigma_1^2) + \mathrm{E}(1 - z_i) \log \phi(x_1; \mu_0, \sigma_0^2) \\
&\quad + \mathrm{E}(z_1) \log(\pi) + \mathrm{E}(1 - z_1) \log(1 - \pi)
\end{aligned}
$$

We can maximize $\mathrm{E}l^c(\theta)$ with respect to θ.

- Set an initial estimate $\theta^{(0)} = (\mu_1, \sigma_1^2, \mu_0, \sigma_0^2, \pi)$ and $t = 0$.
- Expectation step: compute $\mathrm{E}_{Z|X,\theta^{(t)},}[l^c(\theta)]$

$$
\begin{aligned}
E_{Z|X,\theta^{(t)}}[l^c(\theta)] \quad = \quad & \sum_{i=1}^{n} \Big[(E_{Z|X,\theta^{(t)}} z_i) \log \phi(x_i; \mu_1, \sigma_1^2) \\
& + (E_{Z|X,\theta^{(t)}} (1 - z_i)) \log \phi(x_i; \mu_0, \sigma_0^2) \\
& + (E_{Z|X,\theta^{(t)}} z_i) \log(\pi) + (1 - (E_{Z|X,\theta^{(t)}} z_i)) \log(1 - \pi) \Big]
\end{aligned}
$$

Moreover,

$$
E_{Z|X,\theta^{(t)}}[z_i] = \Pr(z_i = 1|x_i) = \frac{\pi\phi(x_i; \mu_1, \sigma_1^2)}{\pi\phi(x_i; \mu_1, \sigma_1^2) + (1 - \pi)\phi(x_i; \mu_0, \sigma_0^2)}.
$$

EM algorithm for two-component mixture

- Maximization step: maximize $\mathrm{E}_{Z|X,\theta^{(t)},}[l^c(\theta)]$ with respect to θ. Denote $\mathrm{E}_{Z|X,\theta^{(t)}}[z_i]$ by $\hat{z}_i$ simply. Then,

$$\begin{aligned}\mathrm{E}_{Z|X,\theta^{(t)},}[l^c(\theta)] &= \sum_{i=1}^{n}\Big[\hat{z}_i \log\phi(x_i;\mu_1,\sigma_1^2) \\ &\quad +(1-\hat{z}_i)\log\phi(x_i;\mu_0,\sigma_0^2) \\ &\quad +\hat{z}_i\log(\pi)+(1-\hat{z}_i)\log(1-\pi)\Big].\end{aligned}$$

The maximizer is given by $\hat{\mu}_1=\sum_{i=1}^n w_i x_i$, $\hat{\sigma}_1^2=\sum_{i=1}^n w_i(x_i-\hat{\mu}_1)^2$, and $\hat{\pi}=\sum_{i=1}\hat{z}_i/\sum_{i=1}^n\hat{z}_i$ where $w_i=\hat{z}_i/\sum_{i=1}^n\hat{z}_i$.

EM algorithm for two-component mixture

- Maximization step: Obtain

$$\hat{\theta} = \underset{\theta}{\text{argmax}}\ \mathrm{E}_{Z|X,\theta^{(t)}}[l^c(\theta)]$$

  and update $\hat{\theta} \to \theta^{(t+1)}$ and $t \to t+1$

- Repeat E-step and M-step until the solution converges.

EM algorithm for two-component mixture

- For each step, the solution achieves higher observed likelihood $L^o(\theta)$.
- The solution converges a local maximum of the observed likelihood function.
- Varying initial values, we can try to investigate many local maxima.

Note that the EM algorithm is a special case of the MM algorithm since $\mathrm{E}_{Z|X}[l^c(\theta)]$ is a majorized function of the observed likelihood function at the current solution.

Notation

- x: observed variable
- z: missing (latent) variable
- (x, z): complete variable
- θ: parameter of the density function of (x, z).

Let

$$f_{Z|X}(z|x;\theta) = \frac{f(x, z; \theta)}{f_X(x;\theta)}.$$

Here, we omit the subscript X, Z in $f(x, z; \theta)$.

Since only x is observed, the (observed) likelihood is defined by

$$L^o(\theta) = \prod_{i=1}^{n} f_X(x_i),$$

where x_is are random samples following f_X. The MLE is obtained by maximizing $\log L^o(\theta)$. However, maximization is frequently difficult due to the form of the loglikelihood function. The typical case is the normal mixture model.

EM algorithm seeks to find the maximizer of $L^o(\theta)$. The starting point is the expectation for the missing variables

First, let $Q(\theta|\theta^{(t)})$ be a conditional expectation of the complete loglikelihood for missing values.

$$
\begin{aligned}
Q(\theta|\theta^{(t)}) &= \mathrm{E}_{Z|X,\theta^{(t)}}\left[\log L^c(\theta)\right] \\
&= \mathrm{E}_{Z|X,\theta^{(t)}}\left[\sum_{i=1}^{n} \log f(x_i, z_i; \theta)\right] \\
&= \sum_{i=1}^{n} \mathrm{E}_{Z_i|X_i,\theta^{(t)}}\left[\log f(x_i, z_i; \theta)\right] \\
&= \sum_{i=1}^{n} \int (\log f(x_i, z; \theta)) f_{Z|X}(z|x_i; \theta^{(t)}) dz.
\end{aligned}
$$

Let $Z \sim$ Logis$(\mu, 1)$ and $X = I(z \geq 0)$. Let $Y = (X, Z)$

- PDF of Y: simply denote pdf of z by $f(z)$.

$$f(x, z) = f(z)I(z < x, x = 0) + f(z)I(z \geq x, x = 1)$$

- Suppose that x is observed. $Q(\theta|\theta^{(t)})$:

$$\begin{aligned} f_{Z|X}(z|x) &= \frac{f(x, z)}{f(x)} \\ &= \frac{f(z)I(z < x, x = 0) + f(z)I(z \geq x, x = 1)}{\Pr(X = x)} \end{aligned}$$

(When $X = 0$, $Z|X$ follows the truncated logistic distribution.)

EM algorithm

- Set an intial $\theta^{(t)}$ for $t = 0$.
- E step: Compute $Q(\theta|\theta^{(t)})$.
- M step: Maximize $Q(\theta|\theta^{(t)})$ w.r.t. θ, and update $\theta^{(t+1)}$ as the maximizer and let $t \to t+1$.
- Repeat the E step and M step until the solutions converge.

Example 1 (Peppered moths)

The peppered moth's coloring is believed to be determined by a single gene with three possible alleles: C, I, and T. C is dominant to I and I is dominant to T. Thus,

(Phenotype)

- C: CC, CI, CT,
- I: II, IT
- T: TT.

We can only observe n_C, n_I, n_T among $n = n_C + n_I + n_T$. Our goal is to estimate the proportion of moths with each genotype.

Figure 2: Pappered moths: Carbonia(left), insularia(middle), typical(right)

Let $z = (n_{CC}, n_{CI}, n_{CT}, n_{II}, n_{IT}, n_{TT})$ and $x = (n_C, n_I, n_T)$. Let the allele frequencies in the population be p_C, p_I and p_T and assume that the probabilities of genotype CC, CI, CT, II, IT, TT are given bt p_C^2, $2p_Cp_I$, $2p_Cp_T$, p_I^2, $2p_Ip_T$ and p_T^2. Note that z is not observed. The complete likelihood (x, z) is given by

$$
\begin{aligned}
\Pr(X = x, Z = z) &= \binom{n}{n_{CC}\ n_{CI}\ n_{CT}\ n_{II}\ n_{IT}\ n_{TT}} \\
&\times (p_C^2)^{n_{CC}} (2p_Cp_I)^{n_{CI}} (2p_Cp_T)^{n_{CT}} (p_I^2)^{n_{II}} (2p_Ip_T)^{n_{IT}} (p_T^2)^{n_{TT}} \\
&\times I(n_{CC} + n_{CI} + n_{CT} = n_C, n_{II} + n_{IT} = n_I, n_{TT} = n_T)
\end{aligned}
$$

$$
\begin{aligned}
l^c(\theta) &= \log Pr(X = x, Z = z) \\
&= 2n_{CC}\log(p_C) + n_{CI}\log(2p_Cp_I) + n_{CT}\log(2p_Cp_T) \\
&\quad +2n_{II}\log(p_I) + n_{IT}\log(2p_Ip_T) + 2n_{TT}\log(p_T) + \text{const.}
\end{aligned}
$$

(E-step)

For given $\hat{p}_C$, $\hat{p}_I$, $\hat{p}_T$

$$
\begin{aligned}
\mathrm{E}(N_{CC}|n_C, n_I, n_T) &= \frac{n_C \times \hat{p}_C^2}{\hat{p}_C^2 + 2\hat{p}_C\hat{p}_I + 2\hat{p}_C\hat{p}_T} \\
\mathrm{E}(N_{CI}|n_C, n_I, n_T) &= \frac{n_C \times 2\hat{p}_C\hat{p}_I}{\hat{p}_C^2 + 2\hat{p}_C\hat{p}_I + 2\hat{p}_C\hat{p}_T} \\
\mathrm{E}(N_{CT}|n_C, n_I, n_T) &= \frac{n_C \times 2\hat{p}_C\hat{p}_T}{\hat{p}_C^2 + 2\hat{p}_C\hat{p}_I + 2\hat{p}_C\hat{p}_T}
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{E}(N_{II}|n_C, n_I, n_T) &= \frac{n_I \times \hat{p}_I^2}{\hat{p}_I^2 + 2\hat{p}_I\hat{p}_T} \\
\mathrm{E}(N_{IT}|n_C, n_I, n_T) &= \frac{n_I \times 2\hat{p}_I\hat{p}_T}{\hat{p}_I^2 + 2\hat{p}_I\hat{p}_T} \\
\mathrm{E}(N_{TT}|n_C, n_I, n_T) &= n_T\hat{p}_T^2
\end{aligned}
$$

Denote the conditional expectations by c_j for $j = 1, \cdots, 6$ in turn. Thus,

$$
\begin{aligned}
Q(\theta|\hat{\theta}) = \mathrm{E}_{Z|X} l^c(\theta) &= 2c_1 \log(p_C) + c_2 \log(2p_C p_I) + c_3 \log(2p_C p_T) \\
&\quad + 2c_4 \log(p_I) + c_5 \log(2p_I p_T) + 2c_6 \log(p_T) + \text{const.} \\
&= (2c_1 + c_2 + c_3) \log p_C + (c_2 + 2c_4 + c_5) \log p_I \\
&\quad + (c_3 + c_5 + 2c_6) \log(p_T) + \text{ const'}
\end{aligned}
$$

(M-step)

$$\begin{aligned} \max \quad & (2c_1 + c_2 + c_3)\log p_C + (c_2 + 2c_4 + c_5)\log p_I + (c_3 + c_5 + 2c_6)\log(p_T) \\ \text{subject to} \quad & p_C + p_I + p_T = 1 \\ & p_C, p_I, p_T \geq 0 \end{aligned}$$

$$\begin{aligned} \hat{p}_C &= \frac{2c_1 + c_2 + c_3}{2\sum_{j=1}^{6} c_j} \\ \hat{p}_I &= \frac{c_2 + 2c_4 + c_5}{2\sum_{j=1}^{6} c_j} \\ \hat{p}_T &= 1 - \hat{p}_C - \hat{p}_I \end{aligned}$$

Example 2 (Risk for HIV infection)

Suppose 1500 gay men were surveyed and each was asked how many risky sexual encounters in the previous 30 days.

Encounters, k	0	1	2	3	4	5	6	7	8
Frequency, n_k	379	299	222	145	109	95	73	59	45
Encounters, k	9	10	11	12	13	14	15	16	
Frequency, n_k	30	24	12	4	2	0	1	1	-

Table 1: Frequency table

Because a single Poisson distribution does not fit the data well, we consider a Poisson mixture model consisting of three populations: $c = 1$ denotes the population 1 following poisson (μ_1); $c = 2$ denotes the population 2 (more risky group) following poisson (μ_2); $c = 3$ denotes zero-response group to a sensitive question.

Model

- $y|c = 1 \sim \text{Poisson}(\mu_1)$
- $y|c = 2 \sim \text{Poisson}(\mu_2)$
- $\Pr(y|c = 3) = I(y = 0)$ (Dirac measure)
- $\Pr(c = j) = \pi_j$ for $j = 1, 2, 3$.

Denote the conditional distribution of $y|c = j$ by $f_j(y)$

Likelihood

$$
\begin{aligned}
\Pr(y=0) &= \sum_{j=1}^{3} \Pr(y=0|c=j)\Pr(c=j) \\
&= \pi_1 \exp(-\mu_1) + \pi_2 \exp(-\mu_2) + \pi_3
\end{aligned}
$$

For $k \geq 1$,

$$
\begin{aligned}
\Pr(y=k) &= \sum_{j=1}^{3} \Pr(y=k|c=j)\Pr(c=j) \\
&= \pi_1 \frac{\mu_1^k \exp(-\mu_1)}{k!} + \pi_2 \frac{\mu_2^k \exp(-\mu_2)}{k!}
\end{aligned}
$$

Likelihood

Let $B_k = \{i : y_i = k\}$ then

$$
\begin{aligned}
l(\mu_1, \mu_2, \pi_1, \pi_2, \pi_3) &= \sum_{i=1}^{n} \log \Pr(y = y_i) \\
&= \sum_{i \in B_0} \log \Pr(y = y_i) + \sum_{i \in B_1} \log \Pr(y = y_i) + \cdots \\
&= \sum_{i \in B_0} \log \Pr(y = 0) + \sum_{i \in B_1} \log \Pr(y = 1) + \cdots \\
&= \sum_{i \in B_0} \log \Pr(y = 0) + \sum_{k=1}^{\infty} \sum_{i \in B_k} \log \Pr(y = k)
\end{aligned}
$$

Loglikelihood

Let $n_k = |B_k|$ then the loglikelihood is given by

$$
\begin{aligned}
l(\mu_1, \mu_2, \pi_1, \pi_2, \pi_3) &= n_0 \log\left(\pi_1 \exp(-\mu_1) + \pi_2 \exp(-\mu_2) + \pi_3\right) \\
&\quad + \sum_{k=1}^{\infty} n_k \log\left(\pi_1 \frac{\mu_1^k \exp(-\mu_1)}{k!} + \pi_2 \frac{\mu_2^k \exp(-\mu_2)}{k!}\right),
\end{aligned}
$$

where $\mu_1, \mu_2, \pi_1, \pi_2, \pi_3 > 0$ and $\pi_1 + \pi_2 + \pi_3 = 1$.

n_k for $k \geq 0$ are given in Table 1.

Maximum Likelihood Estimator

$$
\begin{aligned}
(\hat{\mu}_1, \hat{\mu}_2, \hat{\pi}_1, \hat{\pi}_2, \hat{\pi}_3) \quad = \quad & \text{argmax} \quad l(\mu_1, \mu_2, \pi_1, \pi_2, \pi_3) \\
& \text{subject to } \mu_1, \mu_2, \pi_1, \pi_2, \pi_3 > 0 \\
& \qquad\qquad\qquad \pi_1 + \pi_2 + \pi_3 = 1.
\end{aligned}
$$

Complete loglikelihood

Let a complete observation be (y_i, c_i) for $i = 1, \cdots, n$. Then

$$\begin{aligned}\Pr(y = y_i, c = c_i) &= \Pr(y = y_i | c = c_i) \Pr(c = c_i) \\ &= (\pi_1 f_1(y_i))^{I(c_i=1)} (\pi_2 f_2(y_i))^{I(c_i=2)} (\pi_3 f_3(y_i))^{I(c_i=3)}.\end{aligned}$$

The complete loglikelihood is given by

$$l^c(\mu_1, \mu_2, \pi_1, \pi_2) = \sum_{i=1}^{n} \sum_{j=1}^{3} I(c_i = j)(\log \pi_j + \log f_j(y_i))$$

Let $B_k = \{i : y_i = k\}$ then

$$\begin{aligned} l^c(\mu_1, \mu_2, \pi_1, \pi_2) &= \sum_{k=0}^{\infty} \sum_{i \in B_k} \sum_{j=1}^{3} I(c_i = j)(\log \pi_j + \log f_j(y_i)) \\ &= \sum_{k=0}^{\infty} \sum_{i \in B_k} \sum_{j=1}^{3} I(c_i = j)(\log \pi_j + \log f_j(k)) \\ &= \sum_{k=0}^{\infty} \sum_{i \in B_k} \Big(I(c_i = 1)(\log \pi_1 + k \log \mu_1 - \mu_1 - \log k!) \\ &\qquad + I(c_i = 2)(\log \pi_2 + k \log \mu_2 - \mu_2 - \log k!) \\ &\qquad + I(c_i = 3, k = 0) \log(\pi_3) \Big), \end{aligned}$$

where $\pi_3 = 1 - \pi_1 - \pi_2$.

(conditional prob.)

$$
\begin{aligned}
\Pr(c_i = 1 | y_i = 0) &= \frac{\Pr(y_i = 0 | c_i = 1) \Pr(c_i = 1)}{\sum_{j=1}^{3} \Pr(y_i = 0 | c_i = j) \Pr(c_i = j)} \\
&= \frac{\pi_1 e^{-\mu_1}}{\pi_1 e^{-\mu_1} + \pi_2 e^{-\mu_2} + \pi_3} \\
\Pr(c_i = 2 | y_i = 0) &= \frac{\pi_2 e^{-\mu_2}}{\pi_1 e^{-\mu_1} + \pi_2 e^{-\mu_2} + \pi_3} \\
\Pr(c_i = 3 | y_i = 0) &= \frac{\pi_3}{\pi_1 e^{-\mu_1} + \pi_2 e^{-\mu_2} + \pi_3}
\end{aligned}
$$

(conditional prob.)

For $k \geq 1$

$$
\begin{aligned}
\Pr(c_i = 1 | y_i = k) &= \frac{\Pr(y_i = k | c_i = 1) \Pr(c_i = 1)}{\sum_{j=1}^{3} \Pr(y_i = k | c_i = j) \Pr(c_i = j)} \\
&= \frac{\pi_1 \mu_1^k e^{-\mu_1}}{\pi_1 \mu_1^k e^{-\mu_1} + \pi_2 \mu_2^k e^{-\mu_2}} \\
\Pr(c_i = 2 | y_i = k) &= 1 - \Pr(c_i = 1 | y_i = k) \\
\Pr(c_i = 3 | y_i = k) &= 0
\end{aligned}
$$

(E-step)

$$
\begin{aligned}
& \mathrm{E}_{c|y}(l^c(\mu_1, \mu_2, \pi_1, \pi_2, \pi_3)) \\
= \; & \sum_{k=0}^{\infty} \sum_{i \in B_k} \sum_{j=1}^{3} \mathrm{E}(I(c_i = j)|y = k)(\log \pi_j + \log f_j(k)) \\
= \; & n_0 \Big(\Pr(c_1 = 1|y_1 = 0)(\log \pi_1 - \mu_1) + \Pr(c_1 = 2|y_1 = 0)(\log \pi_2 - \mu_1) \\
& \quad + \Pr(c_1 = 3|y_1 = 0) \log \pi_3 \Big) \\
& + \sum_{k=1}^{\infty} n_k \Big(\Pr(c_1 = 1|y_1 = k)(\log \pi_1 + k \log \mu_1 - \mu_1) \\
& \quad + \Pr(c_1 = 2|y_1 = k)(\log \pi_2 + k \log \mu_2 - \mu_2) \Big) \\
& + \text{const.}
\end{aligned}
$$

(E-step)

For a given $\mu_1^{(t)}, \mu_2^{(t)}, \pi_1^{(t)}, \pi_2^{(t)}, \pi_3^{(t)}$, denote $\hat{c}_{jk} = \Pr(c_1 = j|y_1 = k)$, which is a real number computed by the conditional prob.

$$
\begin{aligned}
& \mathrm{E}_{c|y}(l^c(\mu_1, \mu_2, \pi_1, \pi_2, \pi_3)) \\
= \quad & n_0\hat{c}_{10}(\log \pi_1 - \mu_1) + n_0\hat{c}_{20}(\log \pi_2 - \mu_2) + n_0 c_{30} \log \pi_3 \\
& + \sum_{k=1}^{\infty} n_k c_{1k}(\log \pi_1 + k \log \mu_1 - \mu_1) + \sum_{k=1}^{\infty} n_k c_{2k}(\log \pi_2 + k \log \mu_2 - \mu_2) \\
= \quad & \left(\sum_{k=0}^{\infty} n_k \hat{c}_{1k}\right) \log \pi_1 + \left(\sum_{k=0}^{\infty} n_k \hat{c}_{2k}\right) \log \pi_2 + n_0 \hat{c}_{30} \log \pi_3 \\
& \left(\sum_{k=0}^{\infty} k n_k \hat{c}_{1k}\right) \log \mu_1 - \left(\sum_{k=1}^{\infty} n_k \hat{c}_{1k}\right) \mu_1 + \left(\sum_{k=0}^{\infty} k n_k \hat{c}_{2k}\right) \log \mu_2 - \left(\sum_{k=1}^{\infty} n_k \hat{c}_{2k}\right) \mu_2
\end{aligned}
$$

(E-step) In summary,

$$\begin{aligned} & \mathrm{E}_{c|y}(l^c(\mu_1, \mu_2, \pi_1, \pi_2, \pi_3)) \\ = \quad & s_1 \log \pi_1 + s_2 \log \pi_2 + s_3 \log \pi_3 + s_4 \log \mu_1 - s_5 \mu_1 + s_6 \log \mu_2 - s_7 \mu_2, \end{aligned}$$

where $s_1 = \sum_{k=0}^{\infty} n_k \hat{c}_{1k}$, $s_2 = \sum_{k=0}^{\infty} n_k \hat{c}_{2k}$ and $s_3 = n_0 \hat{c}_{30}$, $s_4 = \sum_{k=0}^{\infty} k n_k \hat{c}_{1k}$, $s_5 = \sum_{k=1}^{\infty} n_k \hat{c}_{1k}$, $s_6 = \sum_{k=0}^{\infty} k n_k \hat{c}_{2k}$, and $s_7 = \sum_{k=1}^{\infty} n_k \hat{c}_{2k}$.

(M-step)

Since $\mathrm{E}_{c|y}(l^c(\mu_1, \mu_2, \pi_3, \pi_1, \pi_2))$ is separable, (π_1, π_2, π_3), μ_1 and μ_2 are independently obtained.

$$\begin{aligned}
\pi_1^{(t+1)} &= s_1/(s_1+s_2+s_3), \quad \pi_2^{(t+1)} = s_2/(s_1+s_2+s_3), \quad \pi_3^{(t+1)} = s_3/(s_1+s_2+s_3) \\
\mu_1^{(t+1)} &= s_4/s_5, \quad \mu_2^{(t+1)} = s_6/s_7.
\end{aligned}$$

Convergence

Note that

$$\log f_X(x;\theta) = \log f(x,z;\theta) - \log f_{Z|X}(z|x;\theta)$$

Therefore,

$$\begin{aligned} & \mathrm{E}_{Z|X,\theta^{(t)}}\left[\log f_X(x;\theta)\right] \\ = \quad & \mathrm{E}_{Z|X,\theta^{(t)}}\left[\log f(x,z;\theta)\right] - \mathrm{E}_{Z|X,\theta^{(t)}}\left[\log f_{Z|X}(z|x;\theta)\right]. \end{aligned}$$

Note that $\log f_X(x|\theta) = \mathrm{E}_{Z|X,\theta^{(t)}}\left[\log f_X(x;\theta)\right]$.

Convergence

So, we can write that

$$\log f_X(x;\theta) \quad = \quad Q(\theta|\theta^{(t)}) - H(\theta|\theta^{(t)}),$$

where $H(\theta|\theta^{(t)}) = \mathrm{E}_{Z|X,\theta^{(t)}}\left[\log f_{Z|X}(z|x;\theta)\right]$.

Actually,

$$KL(f_{Z|X}(z|x,\theta^{(t)})||f_{Z|X}(z|x;\theta)) = \mathrm{E}_{Z|(X,\theta^{(t)})}\log\frac{f(z|x;\theta^{(t)})}{f(z|x;\theta)}$$
$$= \quad \mathrm{E}_{Z|X,\theta^{(t)}}\log f(z|x,\theta^{(t)}) - H(\theta|\theta^{(t)}) \geq 0$$

(KL divergence: the equality holds when $\theta = \theta^{(t)}$)

So that

$$
\begin{aligned}
& \log f_X(x;\theta) + \mathrm{E}_{Z|X,\theta^{(t)}} \log f(z|x,\theta^{(t)}) \\
= \;& Q(\theta|\theta^{(t)}) + \mathrm{E}_{Z|X,\theta^{(t)}} \log f(z|x,\theta^{(t)}) - H(\theta|\theta^{(t)}) \\
\geq \;& Q(\theta|\theta^{(t)})
\end{aligned}
$$

That is, $Q(\theta|\theta^{(t)}) - \mathrm{E}_{Z|X,\theta^{(t)}} \log f(z|x,\theta^{(t)})$ is the minorized function of $\log f_X(x;\theta)$. Because $\mathrm{E}_{Z|X,\theta^{(t)}} \log f(z|x,\theta^{(t)})$ is constant, the maximization of $Q(\theta|\theta^{(t)})$ increases $\log f_X(x,\theta)$.

(Similar results:) We will investigate $H(\theta^{(t)}|\theta^{(t)}) - H(\theta|\theta^{(t)}) \geq 0$ for all θ.

$$
\begin{aligned}
& H(\theta^{(t)}|\theta^{(t)}) - H(\theta|\theta^{(t)}) \\
= \;& \mathrm{E}_{Z|X,\theta^{(t)}}\left[\log f_{Z|X}(z|x,\theta^{(t)}) - \log f_{Z|X}(z|x,\theta)\right] \\
= \;& \int -\log\left[\frac{f_{Z|X}(z|x,\theta)}{f_{Z|X}(z|x;\theta^{(t)})}\right] f_{Z|X}(z|x;\theta^{(t)})d\mathbf{z} \\
\geq \;& -\log \int f_{Z|X}(z|x,\theta)d\mathbf{z} = 0.
\end{aligned}
$$

(The last inequality holds from Jensen's inequality. Explain the inequality through the maximum likelihood method.)

Therefore, we know that $-H(\theta|\theta^{(t)}) \geq -H(\theta^{(t)}|\theta^{(t)})$ for an arbitrary θ

Consider an arbitrary $\theta^{(t+1)}$ satisfying

$$Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta^{(t)}|\theta^{(t)}),$$

then

$$Q(\theta^{(t+1)}|\theta^{(t)}) - H(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta^{(t)}|\theta^{(t)}) - H(\theta^{(t)}|\theta^{(t)}),$$

which is rewritten by

$$\log f_X(x|\theta^{(t+1)}) \geq \log f_X(x|\theta^{(t)}).$$

We can find a sequence of $\theta^{(t)}$ where observed (log)likelihood is increasing.

Since $-H(\theta|\theta^{(t)}) \geq 0$ and $-H(\theta^{(t)}|\theta^{(t)}) = 0$, $Q(\theta|\theta^{(t)})$ is minorized function of $\log f(x|\theta)$ at $\theta^{(t)}$. See the below Figure that illustrates the EM algorithm.

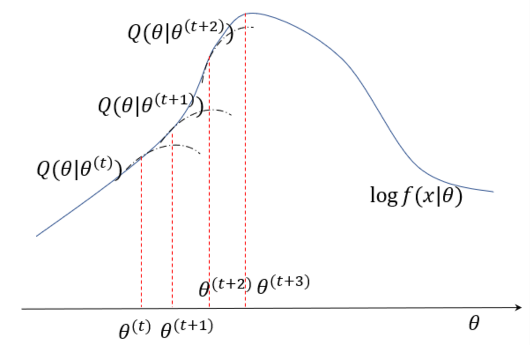

Figure 3: Solutions of EM algorithm

HW

- Derive the MLE of μ and Σ of a multivariate normal distribution. (hint: use the matrix derivatives)
- Write EM algorithm for a Gaussian Mixture Model (GMM).
- What is the selection method for the optimal number of a Gaussian Mixture Model?
- Discuss the usefulness of model-based clustering compared to distance-based models. (ex: when the categorical variables are included in the data, how to apply the K means clustering in the case? In addition, refer to navie bayes method.)
- Submit Python code for the normal mixture model.
- For a convergence criterion, investigate the sensitivity of results according to the selection of initial values.