Expectation-Maximization (EM) algorithm

1 Expectation-Maximization (EM) algorithm

The purpose of EM algorithm is to get the parameter estimation of mixed model contained hidden z variable: $\theta_{MLE} = \arg\max_{\theta} \log p(x|\theta)$. EM algorithm is to use iteration method to solve this problem:

$$\theta^{t+1} = \arg\max_{\theta} \int_z \log\left[p(x,z|\theta)\right] p(z|x,\theta^t)dz = \mathbb{E}_{z|x,\theta^t}\left[\log p(x,z|\theta)\right] \tag{1}$$

Therefore,

- E step: compute the expectation of $\log p(x,z|\theta)$ under the $p(x,z|\theta^t)$ distribution.
- M step: compute the the parameters to maximum the expectation in the E step.

1.1 EM Algorithm Proof

Lemma 1.1. *Jensen Inequality: when $f(x)$ is a convex function, the Jensen Inequality is*

$$\begin{aligned} &x_i \in \textit{discrete distribution}, \quad f\left(\sum_j \lambda_j x_j\right) \leq \sum_j \lambda_j f(x_j) \\ &x_i \in \textit{continuous distribution}, \quad f(\mathbb{E}(x)) \leq \mathbb{E}(f(x)) \Rightarrow f\left(\int xp(x)dx\right) \leq \int f(x)p(x)dx \end{aligned} \tag{2}$$

Lemma 1.2. *Given $\theta^{t+1} = \arg\max_{\theta} \int_z \log\left[p(x,z|\theta)\right] p(z|x,\theta^t)dz$, need to proof $\log p(x|\theta^t) \leq \log p(x|\theta^{t+1})$,* *Proof.*

$$
\begin{aligned}
&\because \quad p(x|\theta) = \frac{p(x,z|\theta)}{p(z|x,\theta)} \\
&\therefore \quad \log p(x|\theta) = \log p(x,z|\theta) - \log p(z|x,\theta) \\
&\text{left: } \int_z p(z|x,\theta^t) \log p(x|\theta) dz = \log p(x|\theta) \\
&\text{right: } \underbrace{\int_z p(z|x,\theta^t) \log p(x,z|\theta) dz}_{Q(\theta,\theta^t)} - \underbrace{\int_z p(z|x,\theta^t) \log p(z|x,\theta) dz}_{H(\theta,\theta^t)} \\
&\because \quad \theta^{t+1} = \arg\max_\theta \int_z \log \left[p(x,z|\theta)\right] p(z|x,\theta^t) dz \\
&\therefore \quad Q(\theta^{t+1},\theta^t) \geq Q(\theta^t,\theta^t) \\
&(1) \quad \text{Using KL divergence to proof} \\
&\therefore \quad H(\theta^{t+1},\theta^t) - H(\theta^t,\theta^t) = \int_z p(z|x,\theta^t) \log p(z|x,\theta^{t+1}) dz - \int_z p(z|x,\theta^t) \log p(z|x,\theta^t) dz \\
&= \int_z p(z|x,\theta^t) \log \left[\frac{p(z|x,\theta^{t+1})}{p(z|x,\theta^t)}\right] dz = -KL(p(z|x,\theta^{t+1})||p(z|x,\theta^t)) \leq 0 \\
&\because \quad Q(\theta^{t+1},\theta^t) \geq Q(\theta^t,\theta^t), H(\theta^{t+1},\theta^t) \leq H(\theta^t,\theta^t) \\
&\therefore \quad \log p(x|\theta^t) \leq \log p(x|\theta^{t+1}) \\
&(2) \;\; \text{Using Jensen Inequality to proof} \\
&\therefore \quad H(\theta^{t+1},\theta^t) - H(\theta^t,\theta^t) = \int_z p(z|x,\theta^t) \log p(z|x,\theta^{t+1}) dz - \int_z p(z|x,\theta^t) \log p(z|x,\theta^t) dz \\
&= \int_z p(z|x,\theta^t) \log \left(\frac{p(z|x,\theta^{t+1})}{p(z|x,\theta^t)}\right) dz \\
&\leq \log \left(\int_z p(z|x,\theta^t) \frac{p(z|x,\theta^{t+1})}{p(z|x,\theta^t)} dz\right) = \log \left(\int_z p(z|x,\theta^{t+1}) dz\right) = \log(1) = 0 \\
&\because \quad Q(\theta^{t+1},\theta^t) \geq Q(\theta^t,\theta^t), H(\theta^{t+1},\theta^t) \leq H(\theta^t,\theta^t) \\
&\therefore \quad \log p(x|\theta^t) \leq \log p(x|\theta^{t+1})
\end{aligned}
\tag{3}
$$

QED

1.2 EM Derivation

$$
\begin{aligned}
\because \quad p(x|\theta) &= \frac{p(x,z|\theta)}{p(z|x,\theta)} \\
\therefore \quad \log p(x|\theta) &= \log p(x,z|\theta) - \log p(z|x,\theta) \\
&= \log \frac{p(x,z|\theta)}{q(z)} - \log \frac{p(z|x,\theta)}{q(z)} \\
\therefore \quad \mathbb{E}_{q(z)}\left(\log p(x|\theta)\right) &= \mathbb{E}_{q(z)}\left(\log \frac{p(x,z|\theta)}{q(z)}\right) - \mathbb{E}_{q(z)}\left(\log \frac{p(z|x,\theta)}{q(z)}\right) \\
\int_z q(z) \log p(x|\theta) dz &= \int_z q(z) \log \frac{p(x,z|\theta)}{q(z)} dz - \int_z q(z) \log \frac{p(z|x,\theta)}{q(z)} dz \\
\log p(x|\theta) &= \underbrace{\int_z q(z) \log \frac{p(x,z|\theta)}{q(z)} dz}_{\text{ELBO}} + \text{KL}(p(z|x,\theta)||q(z))
\end{aligned}
\tag{4}
$$

ELBO (Evidence Lower Bound) is the Lower bound, so $\log p(x|\theta) \geq$ ELBO. When $q(z)$ has the same the distribution of

$p(z|x,\theta)$, both sides of inequality are equal. The purpose of EM algorithm is to maximum the ELBO.

$$
\begin{aligned}
&\hat{\theta} = \arg\max_\theta \text{ELBO} = \arg\max_\theta \int_z q(z) \log \frac{p(x,z|\theta)}{q(z)} dz \\
&\because \ \text{ELBO} \leq \log p(x|\theta) \\
&\therefore \ \text{when } q(z) \text{ has the same distribution of posterior } p(z|x,\theta^t) \text{ , ELBO obtains the maximal value.} \\
&\therefore \ \hat{\theta} = \arg\max_\theta \int_z p(z|x,\theta^t) \log \frac{p(x,z|\theta)}{p(z|x,\theta^t)} dz \\
&\qquad = \arg\max_\theta \int_z p(z|x,\theta^t) \log p(x,z|\theta) dz - \arg\max_\theta \int_z p(z|x,\theta^t) \log p(z|x,\theta^t) dz \\
&\because \ \arg\max_\theta \int_z p(z|x,\theta^t) \log p(z|x,\theta^t) dz = C \\
&\therefore \ \hat{\theta} = \arg\max_\theta \int_z p(z|x,\theta^t) \log p(x,z|\theta) dz
\end{aligned}
\tag{5}
$$

From Jensen Inequality,

$$
\begin{aligned}
\log p(x|\theta) &= \log \left[\int_z p(x,z|\theta) dz \right] \\
&= \log \left[\int_z \frac{p(x,z|\theta) q(z)}{q(z)} dz \right] \\
&= \log \mathbb{E}_{q(z)} \left[\frac{p(x,z|\theta))}{q(z)} \right] \\
&\leq \underbrace{\mathbb{E}_{q(z)} \left[\log \frac{p(x,z|\theta))}{q(z)} \right]}_{\text{ELBO}}
\end{aligned}
\tag{6}
$$

In the end, having the observed data Y, latent variable Z, complete data $X = (Y, Z)$.

- E step: Given y and pretending for the moment that θ^t is correct, formulate the distribution for the complete data x:

$$
f(x|y,\theta^t). \tag{7}
$$

  Then, calculate the Q-function:

$$
\begin{aligned}
Q(\theta,\theta^t) &= \mathbb{E}_{z|x,\theta^t} \left[\log p(x,z|\theta) \right] \\
&= \int_z p(z|x,\theta^t) \log p(x,z|\theta) dz
\end{aligned}
\tag{8}
$$

- M step: Maximum $Q(\theta,\theta^t)$ with regard θ^t:

$$
\theta^{t+1} = \arg\max_\theta Q(\theta,\theta^t) \tag{9}
$$

1.3 Generalized EM

- E step:

$$
q^{t+1}(z) = \arg\max_q \int_z q^t(z) \log \frac{p(x,z|\theta)}{q^t(z)} dz, \text{fixed } \theta \tag{10}
$$

- M step:

$$
\hat{\theta} = \arg\max_\theta \int_z q^{t+1}(z) \log \frac{p(x,z|\theta)}{q^{t+1}(z)} dz, \text{fixed } q \tag{11}
$$

1.4 Gaussian Mixture Model (GMM)

Firstly, identify the variables and parameters, having K Gaussian distributions.

$$
\begin{aligned}
&\text{observed data: } X = (x_1, x_2, \cdots, x_n) \\
&\text{latent variiable: } Z = (z_1, z_2, \cdots, z_n) \\
&\text{paramters: } \theta = p, \mu, \Sigma \\
\text{where: } &p = (p_1, p_2, \cdots, p_k) \\
&\mu = (\mu_1, \mu_2, \cdots, \mu_k) \\
&\sigma = (\Sigma_1, \Sigma_2, \cdots, \Sigma_k) \\
&p_k = \begin{cases} 1, \ \text{if } x_i \in \phi_k \\ 0, \ \text{otherwise} \end{cases},
\end{aligned}
\tag{12}
$$

Some probability formulas

$$\begin{aligned} p(x_i, Z|\theta) &= \sum_k p_k \phi(x_i|\theta_k) \\ p(X, Z|\theta) &= \prod_{i=1}^{N} \sum_{k=1}^{K} p_k \mathcal{N}(x_i|\mu_k, \Sigma_k) \\ p(z = C_k|\theta) &= p_k \\ p(x_i|z_j = C_k, \theta) &= \mathcal{N}(x_i|\mu_k, \Sigma_k) \end{aligned} \tag{13}$$

where $p_k \geq 0,\ \sum_k^K p_k = 1$, $\phi(x|\theta)$ is Gaussian distribution, $\theta_k = (\mu_k, \Sigma_k)$. Therefore,

$$\phi(x|\theta_k) = \frac{1}{\sqrt{(2\pi)^n |\Sigma_k|}} \exp\left(-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)\right) \tag{14}$$

Samples are $X = (x_1, x_2, \cdots, x_n)$, Z is the hidden variable,
the learning parameters are $\theta = \{p_1, p_2, \cdots, p_k, \mu_1, \mu_2, \cdots, \mu_k, \Sigma_1, \Sigma_2, \cdots, \Sigma_k\}$. Using MLE to get θ,

$$\begin{aligned} \theta_{\text{MLE}} &= \arg\max_\theta \log p(X) = \arg\max_\theta \sum_{i=1}^{N} \log p(x_i) \\ &= \arg\max_\theta \sum_{i=1}^{N} \log \sum_{k=1}^{K} p_k \mathcal{N}(x_i|\mu_k, \Sigma_k) \end{aligned} \tag{15}$$

1.4.1 E-Step

The Q-function is,

$$
\begin{aligned}
& Q(\theta,\theta^t) = \mathbb{E}_{Z|X,\theta^t}\left[\log P(X,Z|\theta)\right] = \sum_Z P(Z|X,\theta^t)\log P(X,Z|\theta) \\
\because\quad & \log P(X,Z|\theta) = \sum_{i=1}^{N}\log P(x_i,z_i|\theta), P(Z|X,\theta^t) = \prod_{i=1}^{N} P(z_i|x_i,\theta^t) \\
\therefore\quad & Q(\theta,\theta^t) = \sum_Z\left[\sum_{i=1}^{N}\log P(x_i,z_i|\theta)P(Z|X,\theta^t)\right] \\
& = \sum_Z\left[\log P(x_1,z_1|\theta)P(Z|X,\theta^t) + \cdots + \log P(x_n,z_n|\theta)P(Z|X,\theta^t)\right] \\
\because\quad & \sum_Z\left(\log P(x_1,z_1|\theta)P(Z,X|\theta^t)\right) = \sum_{z_1,z_2,\cdots,z_k}\left(\log P(x_1,z_1|\theta)P(Z,X|\theta^t)\right) \\
& = \sum_{z_1}\left(\log P(x_1,z_1|\theta)P(z_1,x_1|\theta^t)\right)\underbrace{\sum_{z_2,z_3,\cdots,z_k}\prod_{i=2}^{N}P(z_i|x_i,\theta^t)}_{\Delta} \\
\because\quad & \sum_{z_2,z_3,\cdots,z_k}\prod_{i=2}^{N}P(z_i|x_i,\theta^t) = \underbrace{\sum_{z_2}P(z_2|x_2,\theta^t)}_{=1}\cdots\underbrace{\sum_{z_n}P(z_n|x_n,\theta^t)}_{=1} \\
\therefore\quad & \sum_Z\left(\log P(x_1,z_1|\theta)P(Z,X|\theta^t)\right) = \sum_{z_1}\left(\log P(x_1,z_1|\theta)P(z_1,x_1|\theta^t)\right) \\
\therefore\quad & Q(\theta,\theta^t) = \sum_{i=1}^{N}\sum_{z_i}\left(\log P(x_i,z_i|\theta)P(z_i,x_i|\theta^t)\right) \\
& = \sum_{i=1}^{N}\sum_{k=1}^{K}\log P(x_i,z_i=C_k|\theta)P(z_i=C_j|x_i,\theta) \\
& = \sum_{i=1}^{N}\sum_{k=1}^{K}\log p_k\phi(x_i,\mu_k,\Sigma_k)P(z_i=C_j|x_i,\theta) \\
& = \sum_{i=1}^{N}\sum_{k=1}^{K}\left[\log p_k + \log\phi(x_i,\mu_k,\Sigma_k)\right]P(z_i=C_k|x_i,\theta) \\
\because\quad & P(z_i=C_j|x_i,\theta) = \frac{P(x_i,z_i=C_j|\theta)}{P(x_i|\theta)} = \frac{P(x_i,z_i=C_j|\theta)}{\sum_{k=1}^{K}P(x_i,z_i=C_k|\theta)} \\
& = \frac{P(x_i|z_i=C_j,\theta)P(z_i=C_j|\theta)}{\sum_{k=1}^{K}P(x_i|z_i=C_k,\theta)P(z_i=C_k|\theta)} \\
& = \frac{\phi(x_i|\mu_j,\Sigma_j)p_j}{\sum_{k=1}^{K}\phi(x_i|\mu_k,\Sigma_k)p_k} = \gamma_{ij} \\
\therefore\quad & Q(\theta,\theta^t) = \sum_{i=1}^{N}\sum_{k=1}^{K}\left[\log p_k + \log\phi(x_i,\mu_k,\Sigma_k)\right]\gamma_{ij}
\end{aligned} \tag{16}
$$

1.4.2 M-step

$$
\theta^{t+1} = \arg\max_\theta Q(\theta,\theta^t) = \arg\max_\theta\sum_{i=1}^{N}\sum_{k=1}^{K}\left[\log p_k + \log\phi(x_i,\mu_k,\Sigma_k)\right]\gamma_{ij}, \theta = (p_k,\mu_k,\Sigma_k) \tag{17}
$$

Therefore,

$$\begin{aligned}
p_k^{t+1} &= \arg\max_{p_k} \sum_{i=1}^{N}\sum_{k=1}^{K}\left[\log p_k + \log\phi(x_i,\mu_k,\Sigma_k)\right]\gamma_{ij}, \\
&\text{s.t.} \sum_{k=1}^{K} p_k = 1. \\
\therefore\ \ L(p_k,\lambda) &= \sum_{i=1}^{N}\sum_{k=1}^{K}\left[\log p_k + \log\phi(x_i,\mu_k,\Sigma_k)\right]\gamma_{ij} + \lambda(\sum_{k=1}^{K} p_k - 1) \\
\frac{\partial L(p_k,\lambda)}{\partial p_k} &= \sum_{i=1}^{N}\frac{1}{p_k}P(z_i = C_k|x_i,\theta) + \lambda = 0 \\
&\sum_{i=1}^{N} P(z_i = C_k|x_i,\theta) + p_k\lambda = 0 \\
&\sum_{k=1}^{K}\sum_{i=1}^{N} P(z_i = C_k|x_i,\theta) + \sum_{k=1}^{K} p_k\lambda = 0 \\
&\sum_{i=1}^{N}\underbrace{\sum_{k=1}^{K} P(z_i = C_k|x_i,\theta)}_{=1} + \underbrace{\sum_{k=1}^{K} p_k}_{=1}\lambda = 0 \\
&\lambda = -N \\
\therefore\ \ p_k^{t+1} &= \frac{\sum_{i=1}^{N} P(z_i = C_k|x_i,\theta)}{N}
\end{aligned} \tag{18}$$

Compute μ_k^{t+1}

$$\begin{aligned}
\mu_k^{t+1} &= \arg\max_{\mu_k} \sum_{i=1}^{N}\sum_{k=1}^{K}\left[\log p_k + \log\phi(x_i,\mu_k,\Sigma_k)\right]\gamma_{ij} \\
\therefore\ \ L(\mu_k) &= \sum_{i=1}^{N}\sum_{k=1}^{K}\left[\log p_k + \log\phi(x_i,\mu_k,\Sigma_k)\right]\gamma_{ij} \\
&= \arg\max_{\mu_k} \sum_{i=1}^{N}\sum_{k=1}^{K}\left[\log\left[\frac{1}{\sqrt{(2\pi)^n|\Sigma_k|}}\right] - \frac{1}{2}(x_i-\mu_k)^T\Sigma_k^{-1}(x_i-\mu_k)\right]\gamma_{ij} \\
\therefore\ \ \frac{\partial L}{\partial \mu_k} &= \sum_{i=1}^{N}\frac{\partial(-\frac{1}{2}(x_i-\mu_k)^T\Sigma_k^{-1}(x_i-\mu_k))}{\partial\mu_k}\gamma_{ij} = 0 \\
\because\ \ \frac{\partial u^T v}{\partial x} &= u^T\frac{\partial v}{\partial x} + v^T\frac{\partial u}{\partial x}, u = u(x), v = v(x) \\
\therefore\ \ \frac{\partial L}{\partial \mu_k} &= \sum_{i=1}^{N}(x_i - u_k)^T\Sigma_k^{-1}\gamma_{ij} = 0 \\
\therefore\ \ \sum_{i=1}^{N} x_i\Sigma_k^{-1}\gamma_{ij} &= \sum_{i=1}^{N}\mu_k\Sigma_k^{-1}\gamma_{ij} \\
\therefore\ \ \mu_k^{t+1} &= \frac{\sum_{i=1}^{N} x_i\gamma_{ij}}{\sum_{i=1}^{N}\gamma_{ij}}
\end{aligned} \tag{19}$$

Compute Σ_k^{t+1}

$$
\begin{aligned}
&\Sigma_k^{t+1} = \arg\max_{\Sigma_k} \sum_{i=1}^{N} \sum_{k=1}^{K} \left[\log p_k + \log \phi(x_i, \mu_k, \Sigma_k)\right] \gamma_{ij} \\
\therefore\quad & L(\Sigma_k) = \sum_{i=1}^{N} \sum_{k=1}^{K} \left[\log p_k + \log \phi(x_i, \mu_k, \Sigma_k)\right] \gamma_{ij} \\
&= \arg\max_{\Sigma_k} \sum_{i=1}^{N} \sum_{k=1}^{K} \left[\log\left[\frac{1}{\sqrt{(2\pi)^n |\Sigma_k|}}\right] - \frac{1}{2}(x_i-\mu_k)^T \Sigma_k^{-1}(x_i-\mu_k)\right] \gamma_{ij} \\
&= \arg\min_{\Sigma_k} \sum_{i=1}^{N} \left(\log|\Sigma_k| + (x_i-\mu_k)^T \Sigma_k^{-1}(x_i-\mu_k)\right) \gamma_{ij} \\
\because\quad & \frac{\partial |A|}{\partial A} = |A|A^{-1}, \frac{\partial \log|A|}{\partial A} = A^{-1} \\
\therefore\quad & \frac{\partial L}{\partial \Sigma_k} = \sum_{i=1}^{N} \frac{\partial}{\partial \Sigma_k} \left(\log|\Sigma_k| + (x_i-\mu_k)^T \Sigma_k^{-1}(x_i-\mu_k)\right) \gamma_{ij} = 0 \\
\because\quad & (x_i-\mu_k)^T \Sigma_k^{-1}(x_i-\mu_k) \text{ is a scalar} \\
\therefore\quad & (x_i-\mu_k)^T \Sigma_k^{-1}(x_i-\mu_k) \\
&= tr\left((x_i-\mu_k)^T \Sigma_k^{-1}(x_i-\mu_k)\right) \\
&= tr\left(\Sigma_k^{-1}(x_i-\mu_k)^T (x_i-\mu_k)\right) \\
\because\quad & \frac{\partial tr(AB)}{\partial A} = B^T \\
\therefore\quad & \frac{\partial tr\left(\Sigma_k^{-1}(x_i-\mu_k)^T(x_i-\mu_k)\right)}{\partial \Sigma_k} = (x_i-\mu_k)^T(x_i-\mu_k)\frac{\partial \Sigma_k^{-1}}{\partial \Sigma} = -(x_i-\mu_k)^T(x_i-\mu_k)\Sigma_k^{-2} \\
\therefore\quad & \frac{\partial L}{\partial \Sigma_k} = \sum_{i=1}^{N} \left(\Sigma_k^{-1} - (x_i-\mu_k)^T(x_i-\mu_k)\Sigma_k^{-2}\right) \gamma_{ij} = 0 \\
& \sum_{i=1}^{N} \Sigma_k^{-1}\gamma_{ij} = \sum_{i=1}^{N} (x_i-\mu_k)^T(x_i-\mu_k)\Sigma_k^{-2}\gamma_{ij} \\
& \sum_{i=1}^{N} \Sigma_k \gamma_{ij} = \sum_{i=1}^{N} (x_i-\mu_k)^T(x_i-\mu_k)\gamma_{ij} \\
\therefore\quad & \Sigma_k^{t+1} = \frac{\sum_{i=1}^{N}(x_i-\mu_k)^T(x_i-\mu_k)\gamma_{ij}}{\sum_{i=1}^{N}\gamma_{ij}}
\end{aligned} \tag{20}
$$

In the end,

$$
\begin{aligned}
\gamma_{ij} &= \frac{\phi(x_i|\mu_j, \Sigma_j)p_j}{\sum_{k=1}^{K}\phi(x_i|\mu_k, \Sigma_k)p_k} \\
p_k^{t+1} &= \frac{\sum_{i=1}^{N} P(z_i = C_k|x_i, \theta)}{N} \\
\mu_k^{t+1} &= \frac{\sum_{i=1}^{N} x_i\gamma_{ij}}{\sum_{i=1}^{N}\gamma_{ij}} \\
\Sigma_k^{t+1} &= \frac{\sum_{i=1}^{N}(x_i-\mu_k)^T(x_i-\mu_k)\gamma_{ij}}{\sum_{i=1}^{N}\gamma_{ij}}
\end{aligned} \tag{21}
$$