X: observed data → $X=(x_1, x_2, \cdots x_N)$

(X, Z): complete data

θ: parameter → $\theta=\{p_1, p_2, \cdots p_k, \mu_1, \mu_2 \cdots \mu_k, \Sigma_1, \Sigma_2 \cdots \Sigma_k\}$

$$\hat{\theta}_{MLE}=\arg\max_\theta \log P(X)=\arg\max_\theta \log \prod_{i=1}^N P(x_i)=$$

$$=\arg\max_\theta \sum_{i=1}^N \log P(x_i)$$

$$=\arg\max_\theta \sum_{i=1}^N \log \underbrace{\sum_{k=1}^K p_k N(x_i|\mu_k, \Sigma_k)}_{\text{混合高斯分布}}$$

无法直接得到解析解，所以用EM算法求近似解.

使用EM算法.

$$\theta^{t+1}=\arg\max_\theta \underbrace{E_{Z|X,\theta^t}[\log P(X, Z|\theta)]}_{Q(\theta, \theta^t)}$$

$$Q(\theta, \theta^t)=\int_Z \log P(X, Z|\theta) P(Z|X, \theta^t) dZ$$

$$=\sum_Z \log \prod_{i=1}^N P(x_i, z_i|\theta) \prod_{i=1}^N P(z_i|x_i, \theta^t)$$

$$=\sum_{z_1, z_2 \cdots z_N} \sum_{i=1}^N \log P(x_i, z_i|\theta) \cdot \prod_{i=1}^N P(z_i|x_i, \theta^t)$$

z_1：第一个样本属于各个分布的概率

$$=\sum_{z_1, z_2 \cdots z_N} [\log P(x_1, z_1|\theta)+\log P(x_2, z_2|\theta)+\cdots+\log P(x_N, z_N|\theta)] \prod_{i=1}^N P(z_i|x_i, \theta^t)$$

$$\sum_{z_1 \cdots z_N} \log P(x_1, z_1|\theta) \prod_{i=1}^N P(z_i|x_i, \theta^t)$$

$$=\sum_{z_1 \cdots z_N} \log P(x_1, z_1|\theta) P(z_1|x_1, \theta^t) \prod_{i=2}^N P(z_i|x_i, \theta^t)$$

$$=\sum_{z_1} \log P(x_1, z_1|\theta) P(z_1|x_1, \theta^t) \underbrace{\sum_{z_2 \cdots z_N} \prod_{i=2}^N P(z_i|x_i, \theta^t)}_{=1}$$

$$=\sum_{z_1} \log P(x_1, z_1|\theta) P(z_1|x_1, \theta^t)$$

$$=\sum_{z_1} \log P(x_1, z_1|\theta) \cdot P(z_1|x_1, \theta^t)+\cdots+$$

$$\sum_{z_N} \log P(x_N, z_N|\theta) \cdot P(z_N|x_N, \theta^t)$$

$$=\sum_{i=1}^N \sum_{z_i} \log P(x_i, z_i|\theta) \cdot P(z_i|x_i, \theta^t)$$