Preliminary

The probabilistic density of multi-dimensional Gaussian Distribution is:

$$N(x;\vec{u}_k,\Sigma_k)=\frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}}\cdot e^{\frac{1}{2}(x-\vec{u}_k)^T\Sigma^{-1}(x-\vec{u}_k)}$$

where $\vec{u}_k \in R^d$ is the mean vector and Σ_k is the covariance matrix for the kth Gaussian Distribution.

GMM

Hidden Variable Z

In order to clearly describe the data generating process of GMM, we introduce a random variable Z to auxiliary it, where Z belongs to a discrete distribution Q with values drawn from $k \in \{1,\dots,K\}$. However, the specific parameters of Q are inaccessible to us, thus the probability $\alpha_k = Q(Z=k)$ is **unknown**.

Two-step Sampling Method

With Z, the data sampling process can be divided into two parts. Firstly, we need to sample a specific $z_i = k. k \in \{1,\dots,K\}$ from $Z \sim Q$ to select the Gaussian Distribution N_k from which the data sample x_i will be generated. Secondly, we sample the kth $N(x;\vec{u}_k,\Sigma_k)$ to generate the data point.

Thus, two terms can be defined:
Complete Data: $\{(x_i,z_i)\}_{i=1}^n$
Incomplete Data: $\{x_i\}_{i=1}^n$

Modeling the distribution of incomplete data

The GMM takes the weighted-sum of a series of Gaussian Distribution to model its probabilistic density as follows

$$p(x;\Theta)=\sum_{k=1}^K \alpha_k \cdot \mathcal{N}(x;\vec{u}_k,\Sigma_k) \qquad (1)$$

In (1), the parameters are composed of $\Theta = \{\alpha_k,\vec{u}_k,\Sigma_k\}_{k=1}^K$. (1) need to satisfy the following properties:

$$\int_{-\infty}^{\infty} p(x;\Theta) = \int_{-\infty}^{\infty}\sum_{k=1}^K \alpha_k \cdot \mathcal{N}(x;\vec{u}_k,\Sigma_k) = \sum_{k=1}^K \alpha_k \int_{-\infty}^{\infty}\mathcal{N}(x;\vec{u}_k,\Sigma_k) = \sum_{k=1}^K \alpha_k = 1 \quad (2)$$

To (1), the parameter α_k needs to satisfy the condition that $\sum_{k=1}^K \alpha_k = 1, k \in \{1,\dots,K\}$, where α_k is the parameter of distribution Q, e.g. $\alpha_k = P(Z=k), k \in \{1,\dots,K\}$.
It turns out that **GMM models incomplete data distribution by calculating its edge distribution** where $p(x;\Theta) = \sum_z p(x,z;\Theta) = \sum_z p(z;\Theta)\cdot p(x|z;\Theta)$

In practice, the true value of the hidden variable z_i is inaccessible, therefore we can only model the incomplete data to solve the GMM problem.
Data: $\{x_i\}_{i=1}^n$

$$
\begin{aligned}
\mathcal{NLL} &= -log\prod_{i=1}^{n} p(x_i;\Theta) \qquad (3)\\
&= -\sum_{i=1}^{n} log\, p(x_i;\Theta)\\
&= -\sum_{i=1}^{n} log\, [\sum_{z_i} p(z_i;\Theta)\cdot p(x_i|z_i;\Theta)]\\
&= -\sum_{i=1}^{n} log\, [\sum_{k=1}^{K} \alpha_k \cdot p(x_i|z_i=k;\Theta)]
\end{aligned}
$$

The objective is

$$
\begin{aligned}
&min \;\; \mathcal{NLL}(\Theta) \qquad (4)\\
&s.t. \sum_{k=1}^{K}\alpha_k = 1
\end{aligned}
$$

It is difficult to optimize this NLL objective since that:
(1) In the NLL function, there are a series of add operations in the logarithmic function.
(2) Constraints exist.

To solve the problem, we need the **EM algorithm**.

EM Algorithm

The EM algorithm continues the idea of MLE, by continuously constructing the lower bound of the log-likelihood, and optimizing it to increase the lower bound so that after several iterations, the value of the log-likelihood function can approach the maximum value, thereby completing the parameter estimation task.

Preliminary for EM

Jensen's Inequality

In the context of probability theory, it is generally stated in the following form: if X is a random variable and ϕ is a convex function, then $\phi(E(X)) \leq E(\phi(X))$, notice that the equality holds if X is constant (degenerate random variable) or if ϕ is linear.

Derivation

According to that, **we could construct an Expectation for z_i in $\mathcal{NLL}$ to move the add operation out of the log operation** as

$$
\begin{aligned}
\mathcal{LL}(\Theta) &= \sum_{i=1}^{n} log[\sum_{z_i} p(x_i, z_i;\Theta)] \qquad (5)\\
&= \sum_{i=1}^{n} log[\sum_{z_i} p(z_i|x_i;\Theta_{t-1})\cdot \frac{p(x_i, z_i;\Theta)}{p(z_i|x_i;\Theta_{t-1})}]\\
&\geq \sum_{i=1}^{n}\sum_{z_i} p(z_i|x_i;\Theta_{t-1})\cdot log(\frac{p(x_i, z_i;\Theta)}{p(z_i|x_i;\Theta_{t-1})})
\end{aligned}
$$

The posterior can be easily calculated according to the Bayesian theory:

$$
p(z_i|x_i;\Theta_{t-1}) = \frac{p(x_i|z_i;\Theta_{t-1})\cdot p(z_i;\Theta_{t-1})}{\sum_{z_i} p(x_i|z_i;\Theta_{t-1})\cdot p(z_i;\Theta_{t-1})} \qquad (6)
$$

We might as well write the lower bound function in (5) as:

$$\mathcal{B}(\Theta,\Theta_{t-1}) = \sum_{i=1}^{n} E_{z_i|x_i}[log\ \frac{p(x_i,z_i;\Theta)}{p(z_i|x_i;\Theta_{t-1})}] \qquad (7)$$
$$= \sum_{i=1}^{n}\sum_{z_i} p(z_i|x_i;\Theta_{t-1})\cdot log[\frac{p(x_i,z_i;\Theta)}{p(z_i|x_i;\Theta_{t-1})}]$$

According to (5)

$$\mathcal{LL}(\Theta) \geq \mathcal{B}(\Theta,\Theta_{t-1}) \qquad (8)$$

Thus we could optimize the original objective by maximizing the $\mathcal{B}(\Theta,\Theta_{t-1})$.
Specifically, the EM algorithm can be divided into two steps.
Step 1(Expectation Step): Construct the lower bound function in (7)
Step 2(Maximum Step): $\Theta_t = \arg\max_\Theta \mathcal{B}(\Theta,\Theta_{t-1})$

##Convergence of EM algorithm
For (5), when $\Theta = \Theta_{t-1}$:

$$\frac{p(x_i,z_i;\Theta_{t-1})}{p(z_i|x_i;\Theta_{t-1})} = \frac{p(z_i|x_i;\Theta_{t-1})\cdot p(x_i;\Theta_{t-1})}{p(z_i|x_i;\Theta_{t-1})} = p(x_i;\Theta_{t-1})$$

where $p(x_i;\Theta_{t-1})$ is a constant for z_i.
Therefore

$$\mathcal{LL}(\Theta_{t-1}) = \mathcal{B}(\Theta_{t-1},\Theta_{t-1}) \leq \mathcal{B}(\Theta_t,\Theta_{t-1}) \leq \mathcal{LL}(\Theta_t), t \in \Big[1,\dots,\infty\Big]$$

Specific Process of EM used to solve GMM problem

Core: Construct the Expectation Model $\mathcal{B}(\Theta,\Theta_{t-1})$ as the lower bound and maximize it with constraints.

Given the estimation results of iteration $t-1$: $\Theta_{t-1} = \{\alpha_k^{t-1},\vec{u}_k^{t-1},\Sigma_k^{t-1}\}_{k=1}^K$, from (6) the posterior of z_i can be expressed as

$$p(z_i=k|x_i;\Theta_{t-1}) = \frac{p(x_i|z_i=k;\Theta_{t-1})\cdot p(z_i=k;\Theta_{t-1})}{\sum_{z_i}\Big[p(x_i|z_i=k;\Theta_{t-1})\cdot p(z_i=k;\Theta_{t-1})\Big]} \qquad (9)$$
$$= \frac{\alpha_k^{t-1}\cdot\mathcal{N}(x_i;\vec{u}_k^{t-1},\Sigma_k^{t-1})}{\sum_{k=1}^K \alpha_k^{t-1}\cdot\mathcal{N}(x_i;\vec{u}_k^{t-1},\Sigma_k^{t-1})}, k=1,\dots,K$$

which can be directly calculated and treated as a constant q_{ik}
The lower bound then can be expressed as

$$\mathcal{B}(\Theta,\Theta_{t-1}) = \sum_{i=1}^{n} E_{z_i|x_i;\Theta_{t-1}} \log(\frac{p(x_i,z_i;\Theta)}{p(z_i|x_i;\Theta_{t-1})}) \qquad (10)$$
$$= \sum_{i=1}^{n}\sum_{z_i} p(z_i|x_i;\Theta_{t-1})\cdot\log\Big[\frac{p(x_i,z_i;\Theta)}{p(z_i|x_i;\Theta_{t-1})}\Big]$$
$$= \sum_{i=1}^{n}\sum_{k=1}^{K} q_{ik}\cdot\log\Big[\frac{p(x_i|z_i=k;\Theta)\cdot p(z_i=k;\Theta)}{q_{ik}}\Big]$$
$$= \sum_{i=1}^{n}\sum_{k=1}^{K}\Big[q_{ik}\cdot\log p(x_i|z_i=k;\Theta) + q_{ik}\cdot\log\alpha_k - q_{ik}\cdot\log q_{ik}\Big]$$
$$= \sum_{i=1}^{n}\sum_{k=1}^{K}\Big[q_{ik}\cdot\log\Big[\frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_k|^{\frac{1}{2}}}\cdot e^{-\frac{1}{2}(x_i-\vec{u}_k)^T\Sigma_k^{-1}(x_i-\vec{u}_k)}\Big] + q_{ik}\cdot\log\alpha_k - q_{ik}\cdot\log q_{ik}\Big]$$

Since

$$\log\Big[\frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_k|^{\frac{1}{2}}}\cdot e^{-\frac{1}{2}(x_i-\vec{u}_k)^T\Sigma_k^{-1}(x_i-\vec{u}_k)}\Big]=-\frac{d}{2}\log 2\pi-\frac{1}{2}\log|\Sigma_k|-\frac{1}{2}(x_i-\vec{u}_k)^T\Sigma_k^{-1}(x_i-\vec{u}_k)$$

By removing the irrelevant items, the form of the lower bound objective can be expressed as

$$\mathcal{B}(\Theta,\Theta_{t-1})=\sum_{i=1}^{n}\sum_{k=1}^{K}q_{ik}\Big[-\frac{1}{2}\log\Big|\Sigma_k\Big|-\frac{1}{2}(x_i-\vec{u}_k)^T\Sigma_k^{-1}(x_i-\vec{u}_k)+\log\alpha_k\Big]$$

Given that

(1) If A is a square matrix of order n, x is an n-dimensional column vector, then

$$\frac{\partial(x^TAx)}{\partial x}=(A+A^T)x$$

Let the partial derivatives of $\mathcal{B}(\Theta,\Theta_{t-1})$ with respect to Θ be 0

$$\frac{\partial\mathcal{B}}{\partial\vec{u}_k}=-\sum_{i=1}^{n}q_{ik}\cdot\Sigma_k^{-1}(x_i-\vec{u}_k)=0$$
$$\Rightarrow\vec{u}_k=\frac{\sum_{i=1}^{n}q_{ik}\cdot x_i}{\sum_{i=1}^{n}q_{ik}},k=1,\ldots,K$$

As the same,

Given that

$$(1)\frac{\partial|A|}{\partial A}=|A|A^{-1}$$
$$(2)\frac{\partial tr\Big[f(A)\cdot B\Big]}{\partial A}=B^T\cdot\frac{\partial f(A)}{\partial A}$$

$(3)tr(AB)=tr(BA)$ holds for any matrices that meet the multiplication properties where $A\in R^{m\times n}$ and $B\in R^{n\times m}$

therefore

$$\begin{aligned}\frac{\partial(x_i-\vec{u}_k)^T\Sigma_k^{-1}(x_i-\vec{u}_k)}{\partial\Sigma_k}&=\frac{\partial tr\Big[(x_i-\vec{u}_k)^T\Sigma_k^{-1}(x_i-\vec{u}_k)\Big]}{\partial\Sigma_k}\\&=\frac{\partial tr\Big[\Sigma_k^{-1}(x_i-\vec{u}_k)(x_i-\vec{u}_k)^T\Big]}{\partial\Sigma_k}\\&=-(x_i-\vec{u}_k)(x_i-\vec{u}_k)^T\Sigma_k^{-2}\end{aligned}$$

$$\begin{aligned}\frac{\partial\mathcal{B}}{\partial\Sigma_k}&=\sum_{i=1}^{n}q_{ik}\cdot\Big[-\frac{1}{2}\Sigma_k^{-1}+\frac{1}{2}(x_i-\vec{u}_k)(x_i-\vec{u}_k)^T\Sigma_k^{-2}\Big]=0\\&\Rightarrow\sum_{i=1}^{n}q_{ik}\cdot\Big[\Sigma_k-(x_i-\vec{u}_k)(x_i-\vec{u}_k)^T\Big]=0\\\Sigma_k&=\frac{\sum_{i=1}^{n}q_{ik}\cdot(x_i-\vec{u}_k)(x_i-\vec{u}_K)^T}{\sum_{i=1}^{n}q_{ik}},k=1,\ldots,K\end{aligned}$$

For the parameters α_k of hidden variables Z of distribution Q, with constraints $\sum_{k=1}^{K}\alpha_k=1$

The Lagrange function can be written as

$$\mathcal{L}(\alpha_1,\ldots,\alpha_k,\lambda)=\sum_{i=1}^{n}\sum_{k=1}^{K}q_{ik}\log\alpha_k+\lambda(\sum_{k=1}^{K}\alpha_k-1)$$

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial \alpha} = \displaystyle\sum_{i=1}^{n} \frac{q_{ik}}{\alpha_k} + \lambda & = 0 \\ \dfrac{\partial \mathcal{L}}{\partial \lambda} = \displaystyle\sum_{k=1}^{K} \alpha_k - 1 & = 0 \end{cases}$$

$$\begin{aligned} &\sum_{k=1}^{K} \alpha_k = -\frac{\sum_{i=1}^{n} \sum_{k=1}^{K} q_{ik}}{\lambda} = 1 \\ &\Rightarrow \lambda = -\sum_{i=1}^{n} \sum_{k=1}^{K} q_{ik} \\ &\Rightarrow \alpha_k = \frac{\sum_{i=1}^{n} q_{ik}}{\sum_{i=1}^{n} \sum_{k=1}^{K} q_{ik}} = \frac{\sum_{i=1}^{n} q_{ik}}{n}, k = 1, \dots, Kv \end{aligned}$$

Thus

$$\arg\max_{\Theta} \mathcal{B}(\Theta, \Theta_{t-1}) = \begin{cases} q_{ik}^t = \dfrac{a_k^{t-1} \cdot \mathcal{N}(x_i; \vec{u}_k^{t-1}, \Sigma_k^{t-1})}{\sum_{k=1}^{K} a_k^{t-1} \cdot \mathcal{N}(x_i; \vec{u}_k^{t-1}, \Sigma_k^{t-1})}, & k = 1, \dots, K \\ \vec{u}_k^t = \dfrac{\sum_{i=1}^{n} q_{ik}^t \cdot x_i}{\sum_{i=1}^{n} q_{ik}^t}, & k = 1, \dots, K \\ \Sigma_k^t = \dfrac{\sum_{i=1}^{n} q_{ik}^t \cdot (x_i - \vec{u}_k^t)(x_i - \vec{u}_k^t)^T}{\sum_{i=1}^{n} q_{ik}^t}, & k = 1, \dots, K \\ \alpha_k^t = \dfrac{\sum_{i=1}^{n} q_{ik}^t}{n}, & k = 1, \dots, K \end{cases}$$

Summary

For the tth iteration

1. (**E Step**)With parameters estimated by the $t-1$th iteration: $\Theta_{t-1} = \{\alpha_k^{t-1}, \vec{u}_k^{t-1}, \Sigma_k^{t-1}\}_{k=1}^{K}$, constructing the lower bound funtion with the following form:

$$\mathcal{B}(\Theta, \Theta_{t-1}) = \sum_{i=1}^{n} \sum_{k=1}^{K} q_{ik} \left[-\frac{1}{2} \log \left| \Sigma_k \right| - \frac{1}{2} (x_i - \vec{u}_k)^T \Sigma_k^{-1} (x_i - \vec{u}_k) + \log \alpha_k \right]$$

2. (M Step)Maximize it to get Θ_t:

$$\arg\max_{\Theta} \mathcal{B}(\Theta, \Theta_{t-1}) = \begin{cases} q_{ik}^t = \dfrac{a_k^{t-1} \cdot \mathcal{N}(x_i; \vec{u}_k^{t-1}, \Sigma_k^{t-1})}{\sum_{k=1}^{K} a_k^{t-1} \cdot \mathcal{N}(x_i; \vec{u}_k^{t-1}, \Sigma_k^{t-1})}, & k = 1, \dots, K \\ \vec{u}_k^t = \dfrac{\sum_{i=1}^{n} q_{ik}^t \cdot x_i}{\sum_{i=1}^{n} q_{ik}^t}, & k = 1, \dots, K \\ \Sigma_k^t = \dfrac{\sum_{i=1}^{n} q_{ik}^t \cdot (x_i - \vec{u}_k^t)(x_i - \vec{u}_k^t)^T}{\sum_{i=1}^{n} q_{ik}^t}, & k = 1, \dots, K \\ \alpha_k^t = \dfrac{\sum_{i=1}^{n} q_{ik}^t}{n}, & k = 1, \dots, K \end{cases}$$

The relationship between K-Means and GMM

Constraints

1. The samples of incomplete data $\{x_i\}_{i=1}^{n}$ no longer belongs to one of the Gaussian Distribution according to the probability $Z \sim Q$ yet the posterior $q_{ik} = p(z_i = k | x_i; \Theta), k = 1, \dots, K$ has only one possible value of 1 with the other to be 0, which means in each iteration the sample will be assigned to one class with certainty 1 rather according to the posterior of $p(z_i | x_i, \Theta)$

2. The covariance matrix of each Gaussian Distribution is an identity matrix I.
3. The distribution of $Z \sim Q$ belongs to a uniform distribution, e.g. $\alpha_k = p(z_i = k; \Theta) = \frac{1}{K}, k = 1, \dots, K$

Thus the solution has the following form:

$$\arg\max_{\Theta} \mathcal{B}(\Theta, \Theta_{t-1}) = \begin{cases} q_{ik}^t &= \dfrac{\mathcal{N}(x_i; \vec{u}_k^{t-1}, \Sigma_k^{t-1})}{\sum_{k=1}^K \mathcal{N}(x_i; \vec{u}_k^{t-1}, \Sigma_k^{t-1})}, & k = 1, \dots, K \\ \vec{u}_k^t &= \dfrac{\sum_{i=1}^n q_{ik}^t \cdot x_i}{\sum_{i=1}^n q_{ik}^t}, & k = 1, \dots, K \\ \Sigma_k^t &= I, & k = 1, \dots, K \\ \alpha_k^t &= \dfrac{1}{K}, & k = 1, \dots, K \end{cases}$$

Since

$$\begin{aligned} \Sigma_k^t &= \frac{\sum_{i=1}^n q_{ik}^t \cdot (x_i - \vec{u}_k^t)(x_i - \vec{u}_k^t)^T}{\sum_{i=1}^n q_{ik}^t} \\ &= \frac{\sum_{i=1}^n \frac{\mathcal{N}(x_i; \vec{u}_k^{t-1}, \Sigma_k^{t-1})}{\sum_{k=1}^K \mathcal{N}(x_i; \vec{u}_k^{t-1}, \Sigma_k^{t-1})} \cdot (x_i - \vec{u}_k^t)(x_i - \vec{u}_k^t)^T}{\sum_{i=1}^n q_{ik}^t} \\ &\Rightarrow \mathcal{N}(x_i; \vec{u}_k^{t-1}, \Sigma_k^{t-1}) \propto \frac{1}{(x_i - \vec{u}_k^t)^T \cdot (x_i - \vec{u}_k^t)} \end{aligned}$$

Therefore the sample will be assigned to the nearest cluster with the distance metric of Euclidean distance.

The process of the K-Means algorithm

For the tth iteration

1. According to the Euclidean distance between $\{u_k\}_{k=1}^K$ and sample $\{x_i\}_{i=1}^n$ assigning each sample with a class label with the nearest principle.
2. Update the mean vector of each Gaussian Distribution $\{u_k\}_{k=1}^K$ with the mean values of samples belong to one cluster.