No.

Date EM

与Kmeans不同，高斯混合聚类采用概率模型表达聚类。先介绍EM：

考虑 $p(X,Z|\theta)=\prod_{n=1}^{N} p(x_n,z_n)$，其中：

observed：$X=\{x_n\}_{n=1}^{N}$ ；latent：$Z=\{z_n\}_{n=1}^{N}$

Goal: Estimate θ via MLE (or MAP).

$$\hat{\theta}=\arg\max_{\theta}\log p(X|\theta)=\arg\max_{\theta}\log\sum_{Z}p(X,Z|\theta) \quad (\text{discrete})$$

$$=\arg\max_{\theta}\log\int_{Z}p(X,Z|\theta)dZ. \quad (\text{continuous})$$

Z:

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

$$\geq\sum_{Z}q(Z)\log\frac{p(X,Z|\theta)}{q(Z)}$$ (since $\log(\cdot)$ is concave; Jensen)

$$=\sum_{Z}q(Z)\log p(X,Z|\theta)\underbrace{-\sum_{Z}q(Z)\log q(Z)}$$ constant! since irrelevant with θ

(if $q(Z)=p(Z|X,\theta)$，则为等号)

故令 $q(Z)=p(Z|X,\theta)$ (独立)：$\log p(X|\theta)=[\mathbb{E}[\log p(X,Z|\theta)]]^{*}+\text{const}$

* lower-bound factor! EM就是最大化它！

EM流程：E：以当前θ^t推后验 $P(Z|X,\theta^t)$，计算：

$$Q(\theta|\theta^t)=\mathbb{E}_{Z|X,\theta^t}[\log p(Z,X|\theta)]=\sum_{Z}\underbrace{p(Z|X,\theta^t)}_{\text{后验}}\log\underbrace{p(X,Z|\theta)}_{\text{先验}}$$

M: $\theta^{t+1}=\arg\max_{\theta}\{Q(\theta|\theta^t)\underbrace{+[\log P(\theta)]}\}$ 若加：MAP；不加：MLE

latent variable；cluster中，$|Z|=k$ k类

$\sum_{k=1}^{K}\pi_k=1.$

GMM：假设生成N个点：$x_n, n=1,2,\dots,N$

首先从K个高斯中选一个。此选择基于混合成分先验概率π的：

$z_n\sim \text{Multinomial}(z_n|\pi)^{\triangle}$ (多项式分布)

然后生成点服从 $\mathcal{N}(x_n|\mu_k,\Sigma_k)$，其中$\mu_k$是第k个高斯均值，$\Sigma_k$是第k个高斯的协方差 (suppose $z_n=k$)

△：说白了就是在 $\{\pi_i\}_{i=1}^{K}$ 的概率下，抽一个 z_i

No.

Date

上述描述了在有π_i, μ_i, Σ_i下，如何抽n个点出来。

那么学GMM呢？考虑：

observed：$X=\{x_1,\cdots,x_N\}$ latent：$Z=\{z_1,\cdots,z_N\}$

$z_i \in \{1,\cdots,K\}$ $\Rightarrow$ N点分K类

则 complete-data log-likelihood：

$$\log p(X,Z|\theta)=\sum_{i=1}^{N}\log \pi_{z_i}+\log N(x_i|\mu_{z_i},\Sigma_{z_i})$$

则 E-step：posterior：$\gamma_{ik}=p(z_i=k|x_i,\theta^{old})$

$$\text{则}\ \gamma_{ik}=\frac{p(x_i|z_i=k,\theta^{old})\,p(z_i=k|\theta^{old})}{p(x_i|\theta^{old})} \quad \text{Bayes}$$

$$=\frac{\pi_k^{old}N(x_i|\mu_k^{old},\Sigma_k^{old})}{\sum_{j=1}^{K}\pi_j^{old}N(x_i|\mu_j^{old},\Sigma_j^{old})} \quad \text{Bring in Gaussian}$$

M-step：$Q(\theta,\theta^{old})=\mathbb{E}_{Z|X,\theta^{old}}[\log p(X,Z|\theta)]$

$$=\sum_{i=1}^{N}\sum_{k=1}^{K}\gamma_{ik}[\underbrace{\log\pi_k}_{\text{先验}}+\underbrace{\log N(x_i|\mu_k,\Sigma_k)}_{\text{后验}}]$$, constraint：$\sum_k\pi_k=1$

$$\mathcal{L}=Q(\theta,\theta^{old})+\lambda(1-\sum_{k=1}^{K}\pi_k)$$

$$\frac{\partial\mathcal{L}}{\partial\pi_k}=\sum_{i=1}^{N}\frac{\gamma_{ik}}{\pi_k}-\lambda=0 \Rightarrow \pi_k\propto\sum_{i=1}^{N}\gamma_{ik}$$

又$\sum_k\pi_k=1$，故$\sum_{k=1}^{K}\sum_{i=1}^{N}\gamma_{ik}=1$

推出：

$$\begin{cases}\pi_k^{new}=\frac{1}{N}\sum_{i=1}^{N}\gamma_{ik}\\ \mu_k^{new}=\frac{\sum_{i=1}^{N}\gamma_{ik}x_i}{\sum_{i=1}^{N}\gamma_{ik}}\\ \Sigma_k^{new}=\frac{\sum_{i=1}^{N}\gamma_{ik}(x_i-\mu_k^{new})(x_i-\mu_k^{new})^T}{\sum_{i=1}^{N}\gamma_{ik}}\end{cases}$$

；同理：$\frac{\partial Q}{\partial\mu_k}=\sum_{i=1}^{N}\gamma_{ik}\Sigma_k^{-1}(x_i-\mu_k)=0$

$$\frac{\partial Q}{\partial\Sigma_k^{-1}}=\frac{1}{2}\sum_{i=1}^{N}\gamma_{ik}[\Sigma_k-(x_i-\mu_k)(x_i-\mu_k)^T]=0$$

KOKUYO

No.

Date

附：PCA：主成分分析

Input：$D=\{x_1, x_2, \cdots x_m\}$ $\in R^d$，低维空间维数 d'

Algorithm：中心化：$x_i \leftarrow x_i - \frac{1}{m}\sum_{i=1}^{m} x_i$

计算协方差矩阵：XX^T

特征值分解

取最大 d' 个特征值对应的 eigenvector $w_1, w_2, \cdots, w_{d'}$

$\Rightarrow$ Output $w^* = (w_1, w_2, \cdots, w_{d'}) \in R^{d' \times d}$.

$x_i \cdot w^{*T}$ 后便进入低维空间

PCA是最常用的降维方法!