西南大學

本科生课程论文

论文题目 ： 基于 EM 算法的高斯混合模型聚类及应用

课程名称 ： 统计计算

任课教师 ： 刘传递

专　　业 ： 统计学

班　　级 ： 2020 级统计学

学　　号 ： 222020314011081

姓　　名 ： 余乐扬

2023 年 01 月 20 日

目录

基于 EM 算法的高斯混合模型聚类及应用

余乐扬

2020 级统计学

学号：222020314011081

摘　要：EM 聚类作为一种强大且可靠的聚类算法如今已被广泛运用在各个场合，例如在模式识别、NLP、图像分割与情感分类等诸多领域中都有大量使用，而 EM 聚类的基础 EM 算法本身便是无监督学习中的经典方法。所有 EM 聚类算法中应用最为广泛的是基于高斯混合模型的 EM 聚类算法，得益于正态分布优秀的理论性质与 EM 算法的泛用性，EM 聚类通常能在有限时间内完成，且效果较佳。EM 算法与 EM 聚类有较强的统计学理论支持，本文通过系统解析 EM 算法，进而导出将 EM 算法应用于聚类的理论推理。

关键词：**聚类分析；无监督学习；EM 算法；高斯混合模型**

1. 引言

EM 聚类算法基于 EM 算法（Expectation-Maximization Algorithm，期望最大化算法）实现，是一种强力而行之有效的聚类算法，常配合高斯混合模型进行聚类。由于正态分布及高斯混合分布的优良性质，一般能取得较为理想的聚类效果。

EM 算法长期在“机器学习十大算法”评比中榜上有名，是极其受欢迎的一种经典算法，而 EM 聚类亦是被广泛运用在各个领域，因此研究 EM 算法与基于高斯混合模型的 EM 聚类算法是有必要的。

记p元正态分布 PDF 为$\mathcal{N}$，GMM（高斯混合模型）：

$$\begin{cases} p(\boldsymbol{\theta}) = \sum_{i=1}^{N} \omega_i \, \mathcal{N}_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \\ \sum_{i=1}^{N} \omega_i = 1 \end{cases} \tag{1}$$

2. EM 算法

EM 聚类基于 EM 算法实现，研究 EM 聚类便不得不研究 EM 算法。

2.1. EM 算法概述及思想

EM 算法是一种用以寻求参数$\boldsymbol{\theta}$的极大似然估计或最大后验估计的优化迭代策略。通过考虑缺失数据，避开直接处理似然函数$L(\boldsymbol{\theta}|\boldsymbol{x})$迭代求解，转而利用较容易得到的密度函数$f_{(\boldsymbol{X},\boldsymbol{Z})|\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{z}|\boldsymbol{\theta})$与$f_{\boldsymbol{Z}|(\boldsymbol{X},\boldsymbol{\theta})}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\theta})$间接地迭代求解。

通过这种方式，EM 算法能简单地被执行并在保证每一步均单调上升的情况下，可靠地优化到局部最优解。

EM 算法是一种非梯度算法、典型的单调算法[1]，但这不能使 EM 算法保证收敛到全局最优解，譬如对于多峰的分布很可能会收敛到局部最优解，这是 EM 算法显著的不足。不过，针对 GMM 这一劣势不存在。

本文用$\boldsymbol{\theta}$表示待估参数，假设完全数据由随机向量$\boldsymbol{Y}=(\boldsymbol{X},\boldsymbol{Z})$产生，其中观测数据样本$\boldsymbol{x}$由随机变（向）量$\boldsymbol{X}$产生，缺失数据$\boldsymbol{z}$由随机变（向）量$\boldsymbol{Z}$产生，并称$\boldsymbol{Z}$为隐变量（Latent Variable），容易导出：

$$f_{\boldsymbol{Z}|(\boldsymbol{X},\boldsymbol{\theta})}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\theta}) = \frac{f_{\boldsymbol{Y}|\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{\theta})}{f_{\boldsymbol{X}|\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{\theta})} \tag{2}$$

用$Q(\boldsymbol{\theta}|\boldsymbol{\theta}_i)$表示在$\boldsymbol{X}=\boldsymbol{x}$与$\boldsymbol{\theta}_i$条件下，完全数据$\boldsymbol{Z}$的参数$\boldsymbol{\theta}$的联合对数似然函数的条件期望：

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}_i) &= \mathbb{E}_{\boldsymbol{Z}}(\log L(\boldsymbol{\theta}|\boldsymbol{Y}) \mid \boldsymbol{x},\boldsymbol{\theta}_i) \\ &= \mathbb{E}_{\boldsymbol{Z}}\big(\log f_{\boldsymbol{Y}|\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{\theta}) \,\big|\, \boldsymbol{x},\boldsymbol{\theta}_i\big) \end{aligned} \tag{3}$$

由条件期望公式，有：

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}_i) &= \mathbb{E}_{\boldsymbol{Z}}(l(\boldsymbol{\theta}|\boldsymbol{Y})| \,\boldsymbol{x},\boldsymbol{\theta}_i) \\ &= \int_{Z} \left[\log f_{\boldsymbol{Y}|\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{\theta})\right]\left[f_{\boldsymbol{Z}|(\boldsymbol{X},\boldsymbol{\theta})}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\theta}_i)\right] \mathrm{d}z \end{aligned} \tag{4}$$

上式强调了当给定$\boldsymbol{X}=\boldsymbol{x}$，$\boldsymbol{Z}$便是$\boldsymbol{Y}$中唯一随机的部分（因为“未知”而导致的随机）。

缺失数据可能并不是真的缺失了，他可能只是为了简化问题而引进的；通过隐变量$\boldsymbol{Z}$，EM 算法将收敛到关于$\boldsymbol{X}$的参数 $\boldsymbol{\theta}$ MLE 某个局部最优解。

2.1. EM 算法的实现与单调性

EM 算法的根本目的是寻找一个使得似然函数$L(\boldsymbol{\theta}|\boldsymbol{x})$最大化的参数$\boldsymbol{\theta}$，即 MLE。由于通过传统统计分析方法直

接求解 MLE 解析解在很多情况下并不方便，考虑通过迭代算法解出可接受误差范围内的数值解。利用相对容易得到的$f_{(\boldsymbol{X},\boldsymbol{Z})|\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{z}|\boldsymbol{\theta})$和$f_{\boldsymbol{Z}|(\boldsymbol{X},\boldsymbol{\theta})}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\theta})$，可以间接地求出使得$L(\boldsymbol{\theta}|\boldsymbol{x})$最大化的$\boldsymbol{\theta}$。

EM 算法的一般步骤分为 E 步与 M 步，通常的改进都是基于这两个步骤的[5][6]：以$\boldsymbol{\theta}_0$为初始值，令$t=0$,

- E 步：基于$\boldsymbol{\theta}_t$计算$Q(\boldsymbol{\theta}|\boldsymbol{\theta}_t)$；
- M 步：$\boldsymbol{\theta}_{t+1}=\arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}_t)$，转步 E 步并令$t\leftarrow t+1$直至收敛.

EM 算法可以被解释为一种类坐标下降法[1]，因为每次更新一些参数时，另一些参数是固定的：更新部分参数时，利用未更新参数更新。

EM 算法单调性：由于

$$\begin{aligned}l(\boldsymbol{\theta}|\boldsymbol{x})&=\mathbb{E}_{\boldsymbol{Z}}(l(\boldsymbol{\theta}|\boldsymbol{x})|\,\boldsymbol{x},\boldsymbol{\theta}_i)\\&=\mathbb{E}_{\boldsymbol{Z}}\left(\log\frac{f_{\boldsymbol{Y}|\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{\theta})}{f_{\boldsymbol{Z}|(\boldsymbol{X},\boldsymbol{\theta})}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\theta})}\middle|\,\boldsymbol{x},\boldsymbol{\theta}_i\right)\\&=Q(\boldsymbol{\theta}|\boldsymbol{\theta}_i)-\mathbb{E}_{\boldsymbol{Z}}\left(\log f_{\boldsymbol{Z}|(\boldsymbol{X},\boldsymbol{\theta})}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\theta})\,\middle|\,\boldsymbol{x},\boldsymbol{\theta}_i\right)\end{aligned}\tag{5}$$

因此定义$H(\boldsymbol{\theta}|\boldsymbol{\theta}_i)=\mathbb{E}_{\boldsymbol{Z}}\left(\log f_{\boldsymbol{Z}|(\boldsymbol{X},\boldsymbol{\theta})}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\theta})\,\middle|\,\boldsymbol{x},\boldsymbol{\theta}_i\right)$，

$$\begin{aligned}&l(\boldsymbol{\theta}_{i+1}|\boldsymbol{x})-l(\boldsymbol{\theta}_i|\boldsymbol{x})\\&=\mathbb{E}_{\boldsymbol{Z}}(l(\boldsymbol{\theta}_{i+1}|\boldsymbol{x})|\,\boldsymbol{x},\boldsymbol{\theta}_i)-\mathbb{E}_{\boldsymbol{Z}}(l(\boldsymbol{\theta}_i|\boldsymbol{x})|\,\boldsymbol{x},\boldsymbol{\theta}_i)\\&=[Q(\boldsymbol{\theta}_{i+1}|\boldsymbol{\theta}_i)-Q(\boldsymbol{\theta}_i|\boldsymbol{\theta}_i)]-[H(\boldsymbol{\theta}_{i+1}|\boldsymbol{\theta}_i)-H(\boldsymbol{\theta}_i|\boldsymbol{\theta}_i)]\end{aligned}\tag{6}$$

其中$Q(\boldsymbol{\theta}_{i+1}|\boldsymbol{\theta}_i)-Q(\boldsymbol{\theta}_i|\boldsymbol{\theta}_i)\geqslant 0$，这是 EM 算法的 M 步保证的；故只需要说明$\forall\widetilde{\boldsymbol{\theta}}\in\Theta,-\left[H(\widetilde{\boldsymbol{\theta}}|\boldsymbol{\theta}_i)-H(\boldsymbol{\theta}_i|\boldsymbol{\theta}_i)\right]\geqslant 0$则可证明 EM 的单调性：

$$\begin{aligned}&-\left[H(\widetilde{\boldsymbol{\theta}}|\boldsymbol{\theta}_i)-H(\boldsymbol{\theta}_i|\boldsymbol{\theta}_i)\right]\\&=\mathbb{E}_{\boldsymbol{Z}}\left(\log\frac{f_{\boldsymbol{Z}|(\boldsymbol{X},\boldsymbol{\theta})}(\boldsymbol{z}|\boldsymbol{x},\widetilde{\boldsymbol{\theta}})}{f_{\boldsymbol{Z}|(\boldsymbol{X},\boldsymbol{\theta})}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\theta}_i)}\middle|\,\boldsymbol{x},\boldsymbol{\theta}_i\right)\\&=\int_{Z}-\log\frac{f_{\boldsymbol{Z}|(\boldsymbol{X},\boldsymbol{\theta})}(\boldsymbol{z}|\boldsymbol{x},\widetilde{\boldsymbol{\theta}})}{f_{\boldsymbol{Z}|(\boldsymbol{X},\boldsymbol{\theta})}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\theta}_i)}f_{\boldsymbol{Z}|(\boldsymbol{X},\boldsymbol{\theta})}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\theta}_i)\,\mathrm{d}z\\&\geq-\log\int_{Z}f_{\boldsymbol{Z}|(\boldsymbol{X},\boldsymbol{\theta})}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\theta}_i)\,\mathrm{d}z\quad(Jensen\text{不等式})\\&=-\log 1=0\end{aligned}\tag{7}$$

从信息论角度观察，(7)实质是$Gibbs$不等式的运用。

针对 EM 算法的不足，国内外基于传统 EM 算法已提出 PX-EM 算法、ECME 算法、MCEM 算法等诸多 EM 算法的改良版本。

记$\boldsymbol{\theta}^*$为 EM 算法的不动点、收敛的局部最优解，EM 算法的收敛阶定义为：

$$\rho=\lim_{i\to\infty}\frac{\|\boldsymbol{\theta}_{i+1}-\boldsymbol{\theta}^*\|}{\|\boldsymbol{\theta}_i-\boldsymbol{\theta}^*\|}\tag{8}$$

除了应用于聚类分析，在 NLP、图像分割、机器学习（如 HMM、LDA 变分推断）与计算机图形学等诸多领域中 EM 算法都表现出了强大的效力[5]。关于应用场景将在后文提到。

3. EM 聚类理论

EM 算法作为一个“框架”可以被应用在多个场合，EM 聚类则是 EM 算法的一个典型应用。EM 聚类通常考虑某类模型作为条件或假设，如参数未知的 GMM（高斯混合模型）、隐马尔可夫过程，因此这是一种参数化的方法。通过 EM 算法解出这些参数（的 MLE），故这种聚类算法被称为 EM 聚类算法。

3.1. 基于 GMM 的 EM 聚类思想与框架

需要解释的是为什么考虑 EM 算法：试图直接对$L(\boldsymbol{\theta}|\boldsymbol{X})$偏导以得到参数解析解是极其麻烦的，即使模型已知（以(1)为定义的 GMM），记第j个正态分布均值向量为$\boldsymbol{\mu}_j$、协方差阵为$\boldsymbol{\Sigma}_j$，用n表示样本量：

$$\begin{aligned}&l(\boldsymbol{\theta}|\boldsymbol{X})\\&=\log\prod_{i=1}^{n}P_{GMM}(\boldsymbol{x}_i|\boldsymbol{\theta})\\&=\sum_{i=1}^{n}\log\sum_{j=1}^{N}\omega_jP(\boldsymbol{x}_i|\boldsymbol{\theta})\\&=\sum_{i=1}^{n}\log\sum_{j=1}^{N}\omega_j\frac{1}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Sigma}_j|}\exp\left(-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_j)\right)\end{aligned}\tag{9}$$

(9)式难以直接优化求解，因此考虑在 GMM 聚类中应用 EM 算法的想法是自然的[1]。

此外，值得说明的是 EM 算法的运用：如果某样本最可能是由第z个总体产生的，则他应该被归为第z个集群，这等价于通过最大后验概率确定样本所属的聚类。引入隐变量$\boldsymbol{Z}=(z_1,z_2,\cdots,z_n)$代表所属类别，令第$i$个样本点具有隐变量$z_i=j$，即第$i$个样本点属于第$j$类，由此类比(9)可以写出$l(\boldsymbol{\theta}|\boldsymbol{Y})$具体形式：

$$\begin{aligned}l(\boldsymbol{\theta}|\boldsymbol{Y})&=\log\prod_{i=1}^{n}P_{GMM}(z_i,\boldsymbol{x}_i|\boldsymbol{\theta})\\&=\sum_{i=1}^{n}\log\left(\omega_{z_i}P(z_i,\boldsymbol{x}_i|\boldsymbol{\theta})\right)\end{aligned}\tag{10}$$

其中$P_{GMM}(z_i,\boldsymbol{x}_i|\boldsymbol{\theta})$代表从属 GMM 的随机变量$\boldsymbol{X}_i=\boldsymbol{x}_i$且属第$z_i$集群的概率，$P(z_i,\boldsymbol{x}_i|\boldsymbol{\theta})$代表从属第$i$个集群的随机变量$\boldsymbol{X}_i=\boldsymbol{x}_i$的概率。

由(4)可知，在$\boldsymbol{X}$确定时$\boldsymbol{Z}$是唯一的变量，故(10)式对$\boldsymbol{Z}$取条件期望（实质是对$\boldsymbol{Z}$积分），得到$Q(\boldsymbol{\theta}|\boldsymbol{\theta}_t)$：

$$\begin{aligned}Q(\boldsymbol{\theta}|\boldsymbol{\theta}_t)&=\mathbb{E}_{\boldsymbol{Z}}(l(\boldsymbol{\theta}|\boldsymbol{Y})|\,\boldsymbol{x},\boldsymbol{\theta}_t)\\&=\sum_{\boldsymbol{Z}}l(\boldsymbol{\theta}|\boldsymbol{Y})P(\boldsymbol{Z}|\boldsymbol{x},\boldsymbol{\theta}_t)\\&=\sum_{\boldsymbol{Z}}\sum_{i=1}^{n}\log\left(\omega_{z_i}P(z_i,\boldsymbol{x}_i|\boldsymbol{\theta})\right)P(\boldsymbol{Z}|\boldsymbol{x}_i,\boldsymbol{\theta}_t)\\&=\sum_{j=1}^{N}\sum_{i=1}^{n}\log\left(\omega_jP(j,\boldsymbol{x}_i|\boldsymbol{\theta})\right)P(j|\boldsymbol{x}_i,\boldsymbol{\theta}_t)\end{aligned}\tag{11}$$

为计算$P(j|\boldsymbol{x}_i,\boldsymbol{\theta}_t)$，应用$Bayes$公式：

$$P(j|\boldsymbol{x}_i,\boldsymbol{\theta}_t)=\frac{P(j,\boldsymbol{x}_i|\boldsymbol{\theta}_t)}{\sum_{k=1}^{N}P(k,\boldsymbol{x}_i|\boldsymbol{\theta}_t)} \tag{12}$$

由(11)与(12)，至此得到了可运用 EM 算法的 GMM 下 EM 聚类的理论更新算法：

首先设置初始簇数N（可以参考 K-Means 的结果），并分别用随机的正态分布参数$\boldsymbol{\mu}_j^{(0)}$、$\boldsymbol{\Sigma}_j^{(0)}$与$\omega_j$初始化第$i$个簇（可以通过样本数据为这些正态分布提供一个良好的假设参数，但这不是必要的，因为 EM 聚类算法可以在较少的步骤内迅速优化这些参数），令$t=0$、所有的参数集结为$\boldsymbol{\theta}_0$，记：

$$Q_{GMM}(\boldsymbol{\theta}|\boldsymbol{\theta}_t)=\sum_{j=1}^{N}\sum_{i=1}^{n}\log\left(\omega_jP(j,\boldsymbol{x}_i|\boldsymbol{\theta})\right)\frac{P(j,\boldsymbol{x}_i|\boldsymbol{\theta}_t)}{\sum_{k=1}^{N}P(k,\boldsymbol{x}_i|\boldsymbol{\theta}_t)} \tag{13}$$

- E 步：基于$\boldsymbol{\theta}_t$计算$Q_{GMM}(\boldsymbol{\theta}|\boldsymbol{\theta}_t)$；
- M 步：$\boldsymbol{\theta}_{t+1}=\arg\max_{\boldsymbol{\theta}} Q_{GMM}(\boldsymbol{\theta}|\boldsymbol{\theta}_t)$，转步 E 步并令$t\leftarrow t+1$直至收敛.

3.2. 基于 GMM 的 EM 聚类算法的实现

鉴于本文的研究对象是针对 GMM 的 EM 聚类，基于 GMM 可以进一步具体地导出具体的更新算法，这主要体现在 M 步上。

首先对$Q(\boldsymbol{\theta}|\boldsymbol{\theta}_t)$分解[7]：

$$q_1(\boldsymbol{\theta}|\boldsymbol{\theta}_t)=\sum_{j=1}^{N}\sum_{i=1}^{n}\log(\omega_j)\,P(j|\boldsymbol{x}_i,\boldsymbol{\theta}_t) \tag{14}$$

$$q_2(\boldsymbol{\theta}|\boldsymbol{\theta}_t)=\sum_{j=1}^{N}\sum_{i=1}^{n}\log\big(P(j,\boldsymbol{x}_i|\boldsymbol{\theta})\big)\,P(j|\boldsymbol{x}_i,\boldsymbol{\theta}_t) \tag{15}$$

易见有$Q(\boldsymbol{\theta}|\boldsymbol{\theta}_t)=(15)+(16)$；

3.2.1. 最大化$\boldsymbol{\mu}_k$

$$\begin{aligned}\frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}_t)}{\partial\boldsymbol{\mu}_k}&=\frac{\partial q_2(\boldsymbol{\theta}|\boldsymbol{\theta}_t)}{\partial\boldsymbol{\mu}_k}\\&=\frac{\partial}{\partial\boldsymbol{\mu}_k}\sum_{i=1}^{n}\log\big(P(k,\boldsymbol{x}_i|\boldsymbol{\theta})\big)\,P(k|\boldsymbol{x}_i,\boldsymbol{\theta}_t)\\&=-\sum_{i=1}^{n}\boldsymbol{\Sigma}_k^{-1}\big(\boldsymbol{x}_i-\boldsymbol{\mu}_k^{(t)}\big)\,P(k|\boldsymbol{x}_i,\boldsymbol{\theta}_t)\\&\overset{let}{=}0\end{aligned} \tag{16}$$

$$\Rightarrow\boldsymbol{\mu}_k^{(t)}=\frac{\sum_{i=1}^{n}\boldsymbol{x}_iP(k|\boldsymbol{x}_i,\boldsymbol{\theta}_t)}{\sum_{i=1}^{n}P(k|\boldsymbol{x}_i,\boldsymbol{\theta}_t)} \tag{17}$$

多个一元正态分布组合而成的 GMM 情况：

$$\mu_k^{(t)}=\frac{\sum_{i=1}^{n}x_iP(k|x_i,\boldsymbol{\theta}_t)}{\sum_{i=1}^{n}P(k|x_i,\boldsymbol{\theta}_t)} \tag{18}$$

其中：

$$P(k,\boldsymbol{x}_i|\boldsymbol{\theta})=\omega_k\frac{1}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Sigma}_k|}\exp\left(-\frac{1}{2}\left(\boldsymbol{x}_i-\boldsymbol{\mu}_k\right)^T\boldsymbol{\Sigma}_k^{-1}\left(\boldsymbol{x}_i-\boldsymbol{\mu}_k\right)\right) \tag{19}$$

3.2.2. 最大化$\boldsymbol{\Sigma}_k$

同理于 3.2.1.：

$$\left\{\begin{aligned}&\frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}_t)}{\partial\boldsymbol{\Sigma}_k}=\frac{\partial q_2(\boldsymbol{\theta}|\boldsymbol{\theta}_t)}{\partial\boldsymbol{\Sigma}_k}\overset{let}{=}0\\&\Leftrightarrow\\&\sum_{i=1}^{n}\Big((\boldsymbol{x}_i-\boldsymbol{\mu}_k)(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T-\boldsymbol{\Sigma}_k^{(t)}\Big)\,P(k|\boldsymbol{x}_i,\boldsymbol{\theta}_t)\overset{let}{=}0\end{aligned}\right. \tag{20}$$

$$\Rightarrow\boldsymbol{\Sigma}_k^{(t)}=\frac{\sum_{i=1}^{n}P(k|\boldsymbol{x}_i,\boldsymbol{\theta}_t)(\boldsymbol{x}_i-\boldsymbol{\mu}_k)(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T}{\sum_{i=1}^{n}P(k|\boldsymbol{x}_i,\boldsymbol{\theta}_t)} \tag{21}$$

多个一元正态分布组合而成的 GMM 情况：

$$\sigma_k^{2(t)}=\frac{\sum_{i=1}^{n}(x_i-\mu_k)^2P(k|x_i,\boldsymbol{\theta}_t)}{\sum_{i=1}^{n}P(k|x_i,\boldsymbol{\theta}_t)} \tag{22}$$

3.2.3. 最大化ω_k

$$\begin{aligned}\frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}_t)}{\partial\omega_k}&=\frac{\partial q_1(\boldsymbol{\theta}|\boldsymbol{\theta}_t)}{\partial\omega_k}\\&=\frac{\partial}{\partial\omega_k}\sum_{i=1}^{n}\log(\omega_j)\,P(j|\boldsymbol{x}_i,\boldsymbol{\theta}_t)\\&=\sum_{i=1}^{n}\frac{1}{\omega_k}P(k|\boldsymbol{x}_i,\boldsymbol{\theta}_t)\\&\overset{let}{=}0\end{aligned} \tag{23}$$

由于存在$\sum_{i=1}^{n}\omega_i=1$的限制，因此与(24)联立解得：

$$\Rightarrow\omega_k^{(t)}=\frac{\sum_{i=1}^{n}P(k|\boldsymbol{x}_i,\boldsymbol{\theta}_t)}{n} \tag{24}$$

多个一元正态分布组合而成的 GMM 情况：

$$\omega_k^{(t)}=\frac{\sum_{i=1}^{n}P(k|x_i,\boldsymbol{\theta}_t)}{n} \tag{25}$$

3.2.4. 基于 GMM 的 EM 聚类算法的实现

综上所述，由(17)、(21)和(24)可以直接计算出 M 步更新参数的大小，注意到式中唯一非源自原始样本数据的统计量$P(k|\boldsymbol{x}_i,\boldsymbol{\theta}_t)$根据$Bayes$公式也是可计算的[7]，即(12)式。

$$P(k|\boldsymbol{x}_i,\boldsymbol{\theta}_t)=\frac{P(k,\boldsymbol{x}_i|\boldsymbol{\theta}_t)}{\sum_{j=1}^{N}P(j,\boldsymbol{x}_i|\boldsymbol{\theta}_t)}$$

综上所述，首先通过 3.1.中的方法初始化参数，接下来进行迭代：

- E 步：$\forall k\leqslant n$，计算$P(k|\boldsymbol{x}_i,\boldsymbol{\theta}_t)=\frac{P(k,\boldsymbol{x}_i|\boldsymbol{\theta}_t)}{\sum_{j=1}^{N}P(j,\boldsymbol{x}_i|\boldsymbol{\theta}_t)}$；
- M 步：$\forall k\leqslant n$，计算$\boldsymbol{\mu}_k^{(t)}$、$\boldsymbol{\Sigma}_k^{(t)}$与$\omega_k^{(t)}$，转步 E 步并令$t\leftarrow t+1$直至收敛，其中：

$$\left\{\begin{aligned}\boldsymbol{\mu}_k^{(t)}&=\frac{\sum_{i=1}^{n}\boldsymbol{x}_iP(k|\boldsymbol{x}_i,\boldsymbol{\theta}_t)}{\sum_{i=1}^{n}P(k|\boldsymbol{x}_i,\boldsymbol{\theta}_t)}\\\boldsymbol{\Sigma}_k^{(t)}&=\frac{\sum_{i=1}^{n}P(k|\boldsymbol{x}_i,\boldsymbol{\theta}_t)(\boldsymbol{x}_i-\boldsymbol{\mu}_k)(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T}{\sum_{i=1}^{n}P(k|\boldsymbol{x}_i,\boldsymbol{\theta}_t)}\\\omega_k^{(t)}&=\frac{\sum_{i=1}^{n}P(k|x_i,\boldsymbol{\theta}_t)}{n}\end{aligned}\right. \tag{26}$$

可以看出，EM 聚类算法实质是在每个总体都服从一个正态分布的假定下，通过 EM 算法解出这些正态总体的均值向量$\{\boldsymbol{\mu}_i\}_{i=1}^N$与协方差阵$\{\boldsymbol{\Sigma}_i\}_{i=1}^N$。

值得一提的是，从 EM 聚类的实现思路上可以看出 EM 聚类算法是一种软聚类算法，这意味着 EM 算法是将数据以一定的概率分配到各集群中的，一个样本点可能以不同的概率（程度）从属于不同的集群：EM 聚类算法只讨论某个样本点属于任意一个集群的概率，而不限制一个样本点只能属于一个簇。因此。EM 聚类算法是一种基于统计模型的、基于概率的聚类算法[4]，这与诸多硬聚类算法有本质不同。

4. EM 聚类的应用

理论上，足够复杂的 GMM 可以在任意精度内拟合任何分布[2]，因此 GMM 是一种非常强大的模型。基于 GMM 的 EM 聚类算法在今天得到广泛关注，与 GMM 和 EM 算法本身的性质密不可分。

相比其他聚类算法 EM 聚类有许多优势[3]，下主要将 EM 聚类与 K-Means 算法对比以说明 EM 聚类的长处。

1) 较之 K-Means 算法，EM 聚类算法不仅考虑“距离”（通过集群正态分布的均值），还考虑了分布的形状（通过集群正态分布的协方差阵控制），这使得在面临一些并非呈圆或球体分布的样本时 EM 聚类表现更好，应对复杂分布的适应性更强[3]，而传统 K-Means 的效果较差（这时除了 EM 聚类，还可以考虑基于密度的聚类算法）。
2) 由于 EM 聚类属软聚类算法，在多个集群距离较近时仍能有不错的聚类效果[6]，而 K-Means 在质心较近时则表现糟糕。
3) 算法步骤并不复杂，EM 算法的单调性使得每一次的迭代值都稳定上升，最终收敛到最优解，尽管这个最优解可能是局部的。
4) 相对而言对初始值不太敏感（GMM），人为设置参数的影响小，EM 算法的更新参数会在若干步骤内很快收敛到可接受误差范围内，相对而言 K-Means 算法对初始值依赖极其严重。
5) 算法运行中占用存储空间小。

同时，EM 聚类也有一些不足之处，其中部分问题可以通过改进 EM 算法解决，部分问题至今尚未有泛用的应对办法。

1) 噪声、离群点对结果可能有明显影响，这与 K-Means 算法的劣势一致，一些基于密度的聚类算法能解决这个问题[5]。
2) 计算量相对较大，导致不能很好处理大样本与高维数据，这时可以考虑一些 EM 算法的改进算法[5][6]。

参考文献

[1] **Jeff A. Bilmes**, "A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models"[J], *International Computer Science Institute*, 1998
[2] **Miin-Shen Yang** , **Chien-Yo Lai** and **Chih-Ying Lin**, "A robust EM clustering algorithm for Gaussian mixture models"[J], *Pattern Recognition*, vol , no, 2012, pp 3950–3961
[3] **Osama Abu Abbas**, "Comparisons Between Dara Clustering Algorithms"[J], *The International Arab Journal of Information Technology*, vol 5, no. 3, pp 320-325, July, 2008
[4] **李世锋**, **文志强**, **吴岳忠**, 基于改进的 GMM 参数估计的目标检测方法[J], *重庆工商大学学报(自然科学版)*, 2013, 30(05): 30-36.
[5] **史鹏飞**, 基于改进 EM 算法的混合模型参数估计及聚类分析[D], 西北大学, 2009
[6] **王凯南**, **金立左**, 基于高斯混合模型的 EM 算法改进与优化[J], *工业控制计算机*, 2017, 30(05): 115-116+118
[7] **张宏东**, EM 算法及其应用[D], 山东大学, 2014