第9章　聚类与EM算法

2025年11月

本讲介绍

聚类算法与主成分分析均为机器学习中的无监督算法

本章主要介绍K - means聚类、高斯混合模型及其估计方法、EM算法基本原理

重点内容：

1. 聚类算法的原理、优化步骤和收敛性分析
2. 高斯混合模型的结构、参数意义以及基于EM算法的极大似然估计步骤
3. 理解EM算法原理：通过构建下界的迭代优化来解决含隐变量的概率模型

本讲关键问题

（1）如何定义K - means聚类的目标函数，怎样实现数据点的聚类分配和聚类中心的优化？

（2）广义K - means和软分配的原理是什么？有何优势？

（3）高斯混合模型如何定义？其参数含义是什么？

（4）EM算法的一般形式和原理是什么？

（5）怎样利用EM算法对高斯混合模型进行极大似然估计？

（6）K - means算法与高斯混合的EM算法有何关系？

（7）如何选取聚类数量？

关键词： K - means算法、高斯混合模型、 EM算法

一、K - means聚类

给定数据集 $\{x_1, \ldots, x_N\}$，包含N个随机D维变量x的观测值，目标是将其划分为K个聚类（簇），K为给定值。

簇由一组数据点组成，我们认为同簇数据点之间的距离与簇外点之间的距离相比较小。

设第k个聚类中心为μ_k，$k = 1, \ldots, K$。我们的目标是找到数据点的聚类分配以及聚类中心，每个数据点到其最近的聚类中心的距离平方和最小。

一、K - means聚类

引入二值指示变量$r_{nk} \in \{0,1\}$（$k = 1,\dots,K$）描述数据点的聚类分配，若x_n被分配到聚类k，则$r_{nk} = 1$，$j \neq k$时$r_{nj} = 0$（“1 - of - K表示”）。

定义目标函数（失真度量）：

$$J = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk} \|x_n - \mu_k\|^2$$

该目标函数表示每个数据点到其分配向量μ_k的距离平方和。我们的目标是找到$\{r_{nk}\}$和$\{\mu_k\}$使J最小。

一、 K - means聚类 – 优化步骤

通过迭代程序最小化J，每次迭代分别优化$\{r_{nk}\}$和$\{\mu_k\}$。

- 为$\{\mu_k\}$选择初始值。
- 第一阶段（E步）：固定$\{\mu_k\}$，优化$\{r_{nk}\}$。

  J是$\{r_{nk}\}$的线性函数，有闭式解：

$$r_{nk} = \begin{cases} 1, & \text{if } k = \arg\min_j \|x_n - \mu_j\|^2 \\ 0, & \text{otherwise} \end{cases}$$

  相当于将数据点分配到最近的聚类中心，在几何上等价于基于聚类中心的垂直平分线划分空间。
- 第二阶段（M步）：固定$\{r_{nk}\}$，优化$\{\mu_k\}$。

  对μ_k求一阶条件：

$$\frac{\partial J}{\partial \mu_k} = 2\sum_{n=1}^{N} r_{nk}(x_n - \mu_k) = 0$$

  解得，$\mu_k = \frac{\sum_n r_{nk} x_n}{\sum_n r_{nk}}$，即$\mu_k$为分配给聚类$k$的所有数据点的均值。
- 两阶段优化交替进行，直至收敛。

一、 K - means聚类

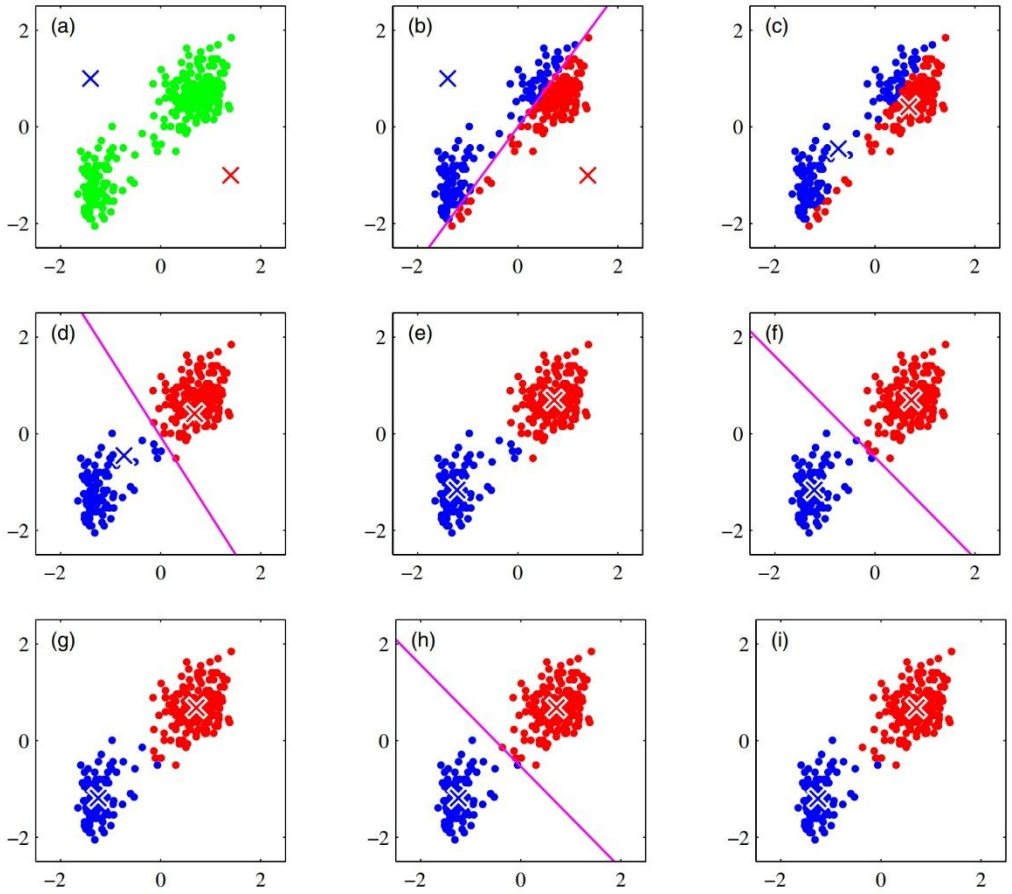

图10.1 K-means算法的示意图

(a) 初始的簇中心μ_1和μ_2分别由红色和蓝色的交叉标记表示。

(b) 第一步中，根据离哪个簇中心更近，每个数据点被分配到红色簇或蓝色簇。等价于根据两个簇中心的垂直平分线（粉线）将点分类。

(c) 第二步中，每个簇中心重新计算为分配到该簇的数据点的均值。

(b)–(i) 展示了从初始到最终收敛的过程。

一、 K - means聚类 – 收敛性

K - means算法保证目标函数J单调下降并最终收敛。

但它不是全局优化方法，可能陷入局部最小值，不同初始聚类中心会导致不同结果。

实际通常多次运行K - means算法，随机初始化聚类中心，选择使J最小的结果。

一、 K - means聚类 – 局限性

K - means算法直接实现可能较慢，因为E步需计算每个原型向量（聚类中心）与每个数据点之间的欧几里得距离。

欧几里得距离测度对非连续变量适用性有限，如变量为类别标签时，欧几里得距离不合适。

基于欧几里得距离也使得聚类中心计算对异常值敏感。

一、 K - means聚类 – 广义K-means

广义K - means引入一般化距离度量$V(x,y)$，目标函数变为

$$\tilde{J}=\sum_{n=1}^{N}\sum_{k=1}^{K}r_{nk}V(x_n,\mu_k)$$

对应K - medoids算法。$V(x,y)$需满足：

- 对称性：$V(x,y)=V(y,x)$。
- 非负性：$V(x,y)\geq 0$。
- 同一性：当且仅当$x=y$，$V(x,y)=0$。
- 三角不等式：$V(x,y)\leq V(x,z)+V(y,z)$。

一、 K - means聚类 – 广义K-means

K-medoids算法的E步与K-means算法相同，对于给定的簇中心μ_k，每个数据点被分配到最近的聚类中心，计算复杂度为$O(KN)$。

K - medoids算法M步更复杂，为适用于任何距离度量$V(\cdot,\cdot)$，常将簇中心限制为簇中某个数据点，每个簇需在N_k个点上进行离散搜索，计算$O(N_k^2)$次$V(\cdot,\cdot)$。

一、 K - means聚类 – 硬分配与软分配

K-means 强制性地将每个数据点分配到一个聚类，但对于某些点（如距离多个中心较近的点），这种“硬分配”可能不够合理。

通过采用概率视角可以实现“软分配”，即基于每个数据点属于不同聚类的概率来分配权重。

二、高斯混合模型 – 模型定义

定义高斯混合分布

$$p(x)=\sum_{k=1}^{K}\pi_k\mathcal{N}(x|\mu_k,\Sigma_k),\quad 0\le\pi_k\le 1,\sum_{k=1}^{K}\pi_k=1$$

其中π_k是混合系数，$p(x)$表示为正态分布的线性叠加。

引入K维二值随机变量$z=(z_1,\dots,z_K)'$，$z_k\in\{0,1\}$，$\sum_k z_k=1$。z的可能状态有K种。令

$$p(z_k=1)=\pi_k$$

$$p(x|z_k=1)=\mathcal{N}(x|\mu_k,\Sigma_k)$$

那么x的边际分布

$$p(x)=\sum_z p(x,z)=\sum_{k=1}^{K}\pi_k\mathcal{N}(x|\mu_k,\Sigma_k)$$

二、高斯混合模型 - 责任权重

若x服从高斯混合分布。我们可以使用如下步骤生成随机样本：

1. 从z的离散分布中采样以确定z的值
2. 根据采样得到的z_k，从条件分布$p(x|z_k = 1)$ 中生成 x

如果有观测x，根据贝叶斯定理，推断隐藏变量z的后验分布$p(z|x)$：

$$\gamma(z_k) \equiv p(z_k = 1|x) = \frac{p(z_k = 1)p(x|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(x|z_j = 1)}$$
$$= \frac{\pi_k \mathcal{N}(x|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x|\mu_j, \Sigma_j)}$$

π_k是$z_k = 1$的先验概率，$\gamma(z_k)$是在观察到x后$z_k = 1$的后验概率。 $\gamma(z_k)$被称为**责任权重**（responsibility），可以理解为第k个高斯成分对观测值 x 的解释程度。

二、高斯混合模型 – 责任权重

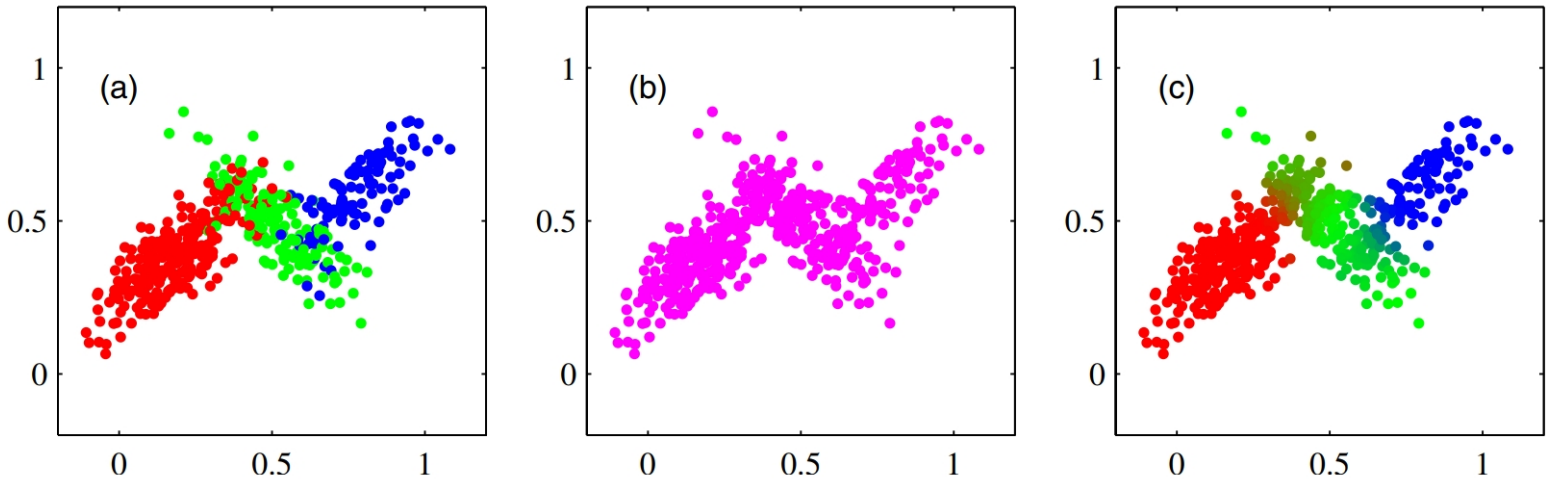

图10.2 从三个高斯混合模型中抽取500个样本

(a) 联合分布$p(z)p(x|z)$，其中z的三个状态分别用红色、绿色和蓝色表示

(b) 边际分布$p(x)$

(c) 相同的样本，颜色表示与数据点x_n相关的责任权重 $\gamma(z_{nk})$

二、高斯混合模型 – 极大似然估计

假设数据集$\{x_1, \ldots, x_N\}$，我们希望用高斯混合模型对其拟合。在独立同分布假设下，$\{x_1, \ldots, x_N\}$的对数似然函数为：

$$\ln p(X|\pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k) \right)$$

其中

$$\pi = (\pi_1, \ldots, \pi_K)$$

由于对数和的存在，直接最大化似然函数并无解析解。

二、高斯混合模型 – 极大似然估计

$$\frac{\partial \mathcal{N}(x|\mu,\Sigma)}{\partial \mu} = \mathcal{N}(x|\mu,\Sigma)\Sigma^{-1}(x-\mu)$$

将$\ln p(X|\pi,\mu,\Sigma)$ 对 μ_k求导得到一阶条件：

$$0 = \sum_{n=1}^{N} \underbrace{\frac{\pi_k \mathcal{N}(x_n|\mu_k,\Sigma_k)}{\sum_{j=1}^{K}\pi_j \mathcal{N}(x_n|\mu_j,\Sigma_j)}}_{\gamma(z_{nk})} \Sigma_k^{-1}(x_n-\mu_k)$$

$\gamma(z_{nk})$是数据点 x_n 属于第k个正态分布的后验概率。假设$\gamma(z_{nk})$ 给定，

$$\mu_k = \frac{\sum_{n=1}^{N}\gamma(z_{nk})x_n}{\sum_{n=1}^{N}\gamma(z_{nk})} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})x_n \qquad (1)$$

其中$N_k = \sum_{n=1}^{N}\gamma(z_{nk})$可以解释为分配给第$k$个簇的有效点数，$\mu_k$通过对所有数据进行加权平均得到，数据点$x_n$的权重是$\gamma(z_{nk})/\sum_{n=1}^{N}\gamma(z_{nk})$。

$$\frac{\partial \mathcal{N}(x|\mu,\Sigma)}{\partial \Sigma} = \frac{1}{2}\mathcal{N}(x|\mu,\Sigma)\left[\Sigma^{-1}(x-\mu)(x-\mu)^T\Sigma^{-1} - \Sigma^{-1}\right]$$

同理，将$\ln p(X|\pi,\mu,\Sigma)$对Σ_k求导的一阶条件得出：

$$0 = \sum_{n=1}^{N}\gamma(z_{nk})\left[\Sigma_k^{-1}(x_n-\mu_k)(x_n-\mu_k)^T\Sigma_k^{-1} - \Sigma_k^{-1}\right]$$

$$\Sigma_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(x_n-\mu_k)(x_n-\mu_k)^T \qquad (2)$$

二、高斯混合模型 - 极大似然估计

引入拉格朗日乘子来求解最优π_k：

$$\ln p(X|\pi,\mu,\Sigma)+\lambda\left(\sum_{k=1}^{K}\pi_k-1\right)$$

对π_k求导得到：

$$0=\sum_{n=1}^{N}\frac{\mathcal{N}(x_n|\mu_k,\Sigma_k)}{\sum_{j=1}^{K}\pi_j\mathcal{N}(x_n|\mu_j,\Sigma_j)}+\lambda=\sum_{n=1}^{N}\frac{\gamma(z_{nk})}{\pi_k}+\lambda$$

对两边同时乘以π_k并对k求和，得到$N+\lambda=0$，

$$\pi_k=\frac{N_k}{N} \tag{3}$$

第k个成分的混合系数估计（先验概率）由该成分的平均责任权重构成。

二、高斯混合模型 – EM算法

上述结果并不是μ_k 、Σ_k和π_k的闭式解，因为$\gamma(z_{nk})$仍依赖于这些参数。然而，可以通过逐步迭代得到最大似然解。

高斯混合模型的EM算法：

1. 设置μ_k 、Σ_k和π_k初值；
2. E 步：基于当前参数，计算每个数据点对各个高斯分量的责任权重：

$$\gamma(z_{nk}) = \frac{\pi_k N(x_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x_n|\mu_j, \Sigma_j)}$$

3. M 步：基于$\gamma(z_{nk})$，按照式(1)，(2)，(3)更新参数μ_k 、Σ_k和π_k 。
4. E步和M步交替进行，直到对数似然函数的值收敛（变化量小于某个阈值）。

可以证明，对数似然函数在迭代过程中单调递增，收敛到局部最优。

二、高斯混合模型 – EM算法

- 与K-means算法相比，EM算法通常需要迭代更多次才能收敛，每次迭代的计算量也大得多。
- 通常会先运行K-means算法，以便为高斯混合模型找到一个合适的初值，再通过EM算法进行优化。
- 协方差矩阵可以初始化为K-means算法所找到的簇内样本协方差，混合系数可以设置为分配给各簇的数据点的比例。
- 对数似然函数通常存在多个局部最大值，因此EM算法并不一定能找到全局最优解。

三、EM算法的一般形式 – EM算法

期望最大化算法（Expectation Maximization, EM）是一种用于寻找含有隐藏变量概率模型的最大似然解的通用技术（Dempster et al., 1977; McLachlan and Krishnan, 1997）。

本节介绍一般形式的EM算法，并证明其在高斯混合模型中确实能够最大化似然函数。

三、EM算法的一般形式 – EM算法

考虑概率模型$p(X,Z \mid \theta)$，观测变量$X=\{x_1,\cdots,x_N\}$，隐藏变量$Z=\{z_1,\cdots,z_N\}$，参数θ。完整数据集$\{X,Z\}$不可观测，仅能获得不完整数据X。

我们的目标是最大化对数似然函数求解最优θ：

$$\ln p(X|\theta)=\ln\left\{\sum_{z_1,\cdots,z_N} p(X,Z|\theta)\right\}$$

或

$$\ln p(X|\theta)=\ln\left\{\int p(X,Z|\theta)dz_1\cdots dz_N\right\}$$

由于多重求和/积分的存在，直接优化对数似然函数会很复杂。EM 算法引入辅助函数$q(Z)$来处理隐藏变量。

三、EM算法的一般形式 – EM算法

Jensen不等式：

$$\ln p(X|\theta) = \ln\left\{\sum_Z q(Z)\frac{p(X,Z|\theta)}{q(Z)}\right\} \geq \sum_Z q(Z)\ln\frac{p(X,Z|\theta)}{q(Z)}$$

$q(Z)$为任意概率分布，当$q(Z) = p(Z \mid X,\theta)$时等号成立。

定义

$$F(q,\theta) = \sum_Z q(Z)\ln\frac{p(X,Z|\theta)}{q(Z)}$$

Jensen不等式表明$\ln p(X|\theta)$的下界$F(q,\theta)$可以由任意概率分布$q(Z)$构造。

EM算法通过最大化下界$F(q,\theta)$实现$\ln p(X|\theta)$的最优化。

给定初值$q_{\text{old}},\theta_{\text{old}}$，通过如下步骤更新$q,\theta$

1. $q_{\text{new}} = \arg\max_q F[q,\theta_{\text{old}}]$
2. $\theta_{\text{new}} = \arg\max_\theta F[q_{\text{new}},\theta]$

三、EM算法的一般形式 – EM算法

- 第一步：注意到，对于任意θ，当$q(Z) = p(Z \mid X, \theta)$时，$F(q, \theta)$达到上界。因此，给定参数 θ_{old}

$$q_{\text{new}}(Z) = p(Z|X, \theta_{\text{old}})$$

即选择Z的后验分布作为$q(Z)$。

- 第二步：给定$q_{\text{new}}(Z)$，

$$F[q_{\text{new}}, \theta] = \sum_Z p(Z|X, \theta_{\text{old}}) \ln p(X, Z|\theta) - \sum_Z p(Z|X, \theta_{\text{old}}) \ln p(Z|X, \theta_{\text{old}})$$

最大化$F[q_{\text{new}}, \theta]$求解$\theta_{\text{new}}$等价于：

$$\theta_{\text{new}} = \arg\max_\theta Q(\theta|\theta_{\text{old}})$$

$$Q(\theta|\theta_{\text{old}}) = \sum_Z p(Z|X, \theta_{\text{old}}) \ln p(X, Z|\theta)$$

三、EM算法的一般形式 – EM算法

总结：

1. **E步（期望步骤）**：给定参数θ_{old}，计算后验分布$p(Z \mid X, \theta_{\text{old}})$和期望函数：

$$Q(\theta|\theta_{\text{old}}) = \sum_{Z} p(Z|X, \theta_{\text{old}}) \ln p(X, Z|\theta)$$

2. **M步（最大化步骤）**：更新参数θ_{new}

$$\theta_{\text{new}} = \arg\max_{\theta} Q(\theta|\theta_{\text{old}})$$

3. 令$\theta_{\text{old}} = \theta_{\text{new}}$，重复迭代步骤直至参数收敛。

$$\begin{aligned} \ln p(X|\theta_{\text{new}}) &\geq \sum_{Z} p(Z|X, \theta_{\text{old}}) \ln \frac{p(X, Z|\theta_{\text{new}})}{p(Z|X, \theta_{\text{old}})} \\ &\geq \sum_{Z} p(Z|X, \theta_{\text{old}}) \ln \frac{p(X, Z|\theta_{\text{old}})}{p(Z|X, \theta_{\text{old}})} = \ln p(X|\theta_{\text{old}}) \end{aligned}$$

除非参数已经收敛到局部最优，EM算法的每轮迭代都能保证对数似然函数单调增加，表明 EM 算法能够稳定收敛到局部最优解。

三、EM算法的一般形式 – 高斯混合模型

我们现在将 EM 算法应用于高斯混合模型。我们的目标是最大化观测数据X的对数似然函数求解$\theta = (\pi, \mu, \Sigma)$：

$$\ln p(X|\theta) = \sum_{n=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k) \right),$$

由于k的求和在对数内部，直接求解优化问题十分复杂。

三、EM算法的一般形式 – 高斯混合模型

考虑完整数据集 $\{X, Z\}$ 的似然函数：

$$p(X, Z|\theta) = \prod_{n=1}^{N} \prod_{k=1}^{K} [\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)]^{z_{nk}}$$

其中 z_{nk}表示z_n的第k个分量。

取对数后得到：

$$\ln p(X, Z|\theta) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \{\ln \pi_k + \ln \mathcal{N}(x_n|\mu_k, \Sigma_k)\}$$

该形式的似然函数属于指数族分布，显著简化了最优化步骤。

$$Q(\theta|\theta_{\text{old}}) = \sum_{n=1}^{N} \sum_{k=1}^{K} E(z_{nk}|X, \theta_{\text{old}}) \{\ln \pi_k + \ln \mathcal{N}(x_n|\mu_k, \Sigma_k)\}$$

三、EM算法的一般形式 – 高斯混合模型

后验概率：

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

$$= \prod_{n=1}^{N} \frac{p(x_n,z_n|\theta)}{p(x_n|\theta)} = \prod_{n=1}^{N} p(z_n|x_n,\theta)$$

意味着$\{z_1,\cdots,z_n\}$的后验分布是独立的。

在后验概率$p(Z \mid X,\theta_{\text{old}})$下，$z_{nk}$的期望为：

$$E[z_{nk}|X,\theta_{\text{old}}] = p(z_{nk}=1|x_n,\theta_{\text{old}}) = \gamma(z_{nk}|\theta_{\text{old}})$$

$$\mathcal{Q}(\theta|\theta_{\text{old}}) = \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(z_{nk}|\theta_{\text{old}}) \{\ln \pi_k + \ln \mathcal{N}(x_n|\mu_k,\Sigma_k)\}$$

三、EM算法的一般形式 – 高斯混合模型

固定$\gamma(z_{nk}|\theta_{\text{old}})$，选择$\mu_k$、$\Sigma_k$和$\pi_k$使得$\mathcal{Q}(\theta|\theta_{\text{old}})$最大化。

$$\frac{\partial \mathcal{Q}(\theta|\theta_{\text{old}})}{\partial \mu_k} = \sum_{n=1}^{N} \gamma(z_{nk}|\theta_{\text{old}}) \Sigma_k^{-1}(x_n - \mu_k) = 0$$

$$\frac{\partial \mathcal{Q}(\theta|\theta_{\text{old}})}{\partial \Sigma_k} = 0 \Rightarrow$$

$$\sum_{n=1}^{N} \gamma(z_{nk}|\theta_{\text{old}}) \left[\Sigma_k^{-1}(x_n - \mu_k)(x_n - \mu_k)^T \Sigma_k^{-1} - \Sigma_k^{-1}\right] = 0$$

在$\sum_k \pi_k = 1$约束下

$$\frac{\partial \mathcal{Q}(\theta|\theta_{\text{old}}) + \lambda\left(\sum_{k=1}^{K} \pi_k - 1\right)}{\partial \pi_k} = \sum_{n=1}^{N} \frac{\gamma(z_{nk}|\theta_{\text{old}})}{\pi_k} + \lambda = 0$$

μ_k、Σ_k和π_k的闭式解同式(1)、(2)和(3)。

三、EM算法的一般形式 – 高斯混合模型与K-means的关系

K-means算法将每个数据点唯一地分配到一个聚类中（硬分配），而高斯混合模型将每个数据点按照后验概率分配到所有聚类中（软分配）。

实际上，可以验证K-means算法是高斯混合的EM算法的一个特例。

三、EM算法的一般形式 – 高斯混合模型与K-means的关系

- 考虑K个高斯混合模型，其中每一组的协方差矩阵$\Sigma_k = \epsilon I$，I是单位阵。（将ϵ视为一个常数，而不是需要估计的参数。）

$$p(x|\mu_k, \Sigma_k) = \frac{1}{(2\pi\epsilon)^{1/2}} \exp\left\{-\|x-\mu_k\|^2 / 2\epsilon\right\}$$

对于观测x_n，z_{nk}的后验概率为

$$\gamma(z_{nk}) = \frac{\pi_k \exp\left\{-\|x_n-\mu_k\|^2 / 2\epsilon\right\}}{\sum_j \pi_j \exp\left\{-\|x_n-\mu_j\|^2 / 2\epsilon\right\}}$$

- 假设 $i = \underset{j}{\operatorname{argmin}} \|x_n - \mu_j\|^2$。如果$\epsilon \to 0$，那么$\gamma(z_{nk})$的分母收敛至0的速度将由第$i$项主导

$$\gamma(z_{nk}) \approx \frac{\pi_k \exp\left\{-\|x_n-\mu_k\|^2 / 2\epsilon\right\}}{\pi_i \exp\left\{-\|x_n-\mu_i\|^2 / 2\epsilon\right\}} \to \begin{cases} 1, & \text{if } k = i \\ 0, & \text{otherwise} \end{cases}$$

也即x_n被分配到与之距离最近的聚类中，$\gamma(z_{nk}) = r_{nk}$，与K-means算法一致。

三、EM算法的一般形式 – 高斯混合模型与K-means的关系

- μ_k的EM估计被简化为K-means的均值估计

$$\mu_k = \frac{\sum_n \gamma(z_{nk}) x_n}{\sum_n \gamma(z_{nk})} = \frac{\sum_n r_{nk} x_n}{\sum_n r_{nk}}$$

- π_k的EM估计被简化为分配给聚类k的数据点的比例

$$\pi_k = \frac{\sum_n \gamma(z_{nk})}{N} = \frac{\sum_n r_{nk}}{N}$$

- 最后，$\mathcal{Q}(\theta|\theta_{\text{old}})$变为

$$\mathcal{Q}(\theta|\theta_{\text{old}}) \to -\frac{1}{2\epsilon} \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|x_n - \mu_k\|^2 + \text{constent}$$

最大化$\mathcal{Q}(\theta|\theta_{\text{old}})$等价于最小化K-means算法的失真度量 J。

四、案例 – 中国股票市场数据聚类分析

考虑99只上交所股票（2016年1月1日——2024年9月30日）。用聚类方法将这99只股票按照收益率相关性特征进行分类。我们认为属于同一簇股票的收益率相关性高于不同簇的股票收益相关性。

将$1-\rho(x,y)$作为距离测度（但这不是一个合格的距离测度？），使用K-medoids 算法得到聚类。Matlab中的指令为

$$[ind, C, sumd, D, midx] = kmedoids(X, k, 'Distance', 'correlation')$$

X：$n \times p$变量

k：聚类数量

$'Distance', 'correlation'$：用相关系数作为距离

ind：数据点所属的聚类索引

C：每一行代表聚类中心（从X中选取）

$sumd$：每个聚类集合到其中心的距离之和

D：每个数据点到每个聚类中心的距离

$midx$：X中被选作聚类中心的索引

四、案例 – 中国股票市场数据聚类分析

表10.1 K-medoids算法，K=20

证券代码	证券简称	行业	证券代码	证券简称	行业	证券代码	证券简称	行业
600011	华能国际	公用事业	600006	东风股份	汽车	600084	中信尼雅	食品饮料
600021	上海电力	公用事业	600066	宇通客车	汽车	600004	白云机场	交通运输
600023	浙能电力	公用事业	600081	东风科技	汽车	600009	上海机场	交通运输
600027	华电国际	公用事业	600104	上汽集团	汽车	600054	黄山旅游	社会服务
600053	九鼎投资	非银金融	600010	包钢股份	钢铁	600055	万东医疗	医药生物
600076	康欣新材	轻工制造	600019	宝钢股份	钢铁	600079	人福医药	医药生物
600080	金花股份	医药生物	600022	山东钢铁	钢铁	600085	同仁堂	医药生物
600082	海泰发展	房地产	600058	五矿发展	商贸零售	600129	太极集团	医药生物
600088	中视传媒	传媒	600111	北方稀土	有色金属	600007	中国国贸	房地产
600099	林海股份	汽车	600117	西宁特钢	钢铁	600008	首创环保	公用事业
600107	美尔雅	纺织服装	600121	郑州煤电	煤炭	600012	皖通高速	交通运输
600119	长江投资	交通运输	600123	兰花科创	煤炭	600017	日照港	交通运输
600128	苏豪弘业	商贸零售	600126	杭钢股份	钢铁	600018	上港集团	交通运输
600130	波导股份	电子	600063	皖维高新	基础化工	600020	中原高速	交通运输
600029	南方航空	交通运输	600075	新疆天业	基础化工	600026	中远海能	交通运输
600115	中国东航	交通运输	600078	澄星股份	基础化工	600033	福建高速	交通运输
600038	中直股份	国防军工	600089	特变电工	综合	600035	楚天高速	交通运输
600072	中船科技	电力设备	600096	云天化	基础化工	600037	歌华有线	传媒
600100	同方股份	计算机	600105	永鼎股份	通信	600039	四川路桥	建筑装饰
600118	中国卫星	国防军工	600110	诺德股份	电力设备	600050	中国联通	通信
600073	光明肉业	食品饮料	600114	东睦股份	机械设备	600051	宁波联合	综合
600097	开创国际	农林牧渔	600116	三峡水利	公用事业	600052	东望时代	公用事业
600108	亚盛集团	农林牧渔	600056	中国医药	医药生物	600057	厦门象屿	交通运输
600127	金健米业	农林牧渔	600062	华润双鹤	医药生物	600064	南京高科	房地产
600030	中信证券	非银金融	600083	*ST博信	综合	600067	冠城新材	电力设备
600031	三一重工	机械设备	600000	浦发银行	银行	600094	大名城	房地产
600059	古越龙山	食品饮料	600015	华夏银行	银行	600098	广州发展	公用事业
600060	海信视像	家用电器	600016	民生银行	银行	600101	明星电力	公用事业
600061	国投资本	非银金融	600028	中国石化	石油石化	600103	青山纸业	轻工制造
600095	湘财股份	非银金融	600036	招商银行	银行	600106	重庆路桥	交通运输
600109	国金证券	非银金融	600048	保利发展	房地产	600113	浙江东日	商贸零售
600120	浙江东方	非银金融	600131	国网信通	计算机	600125	铁龙物流	交通运输
600070	*ST富润	综合	600071	凤凰光学	电子	600180	瑞茂通	交通运输

四、案例 – 中国股票市场数据聚类分析

第二种方法假设数据的相关系数矩阵是由多因子模型构造的。根据Oh and Patton, 2023，假设变量被分成K组，变量之间的相关系数由如下方式构造：

- 如果变量i,j属于同一组k：$\rho(i,j)=\left(\lambda_k^M\right)^2+\left(\lambda_k^C\right)^2$，
- 如果变量i属于组k，变量j属于组l：$\rho(i,j)=\lambda_k^M\,\lambda_l^M$。

该模型也可理解为，股票市场存在一个“市场因子”f^M和K个“行业因子”$f_1^C,\ldots,f_K^C$，这K+1个因子单位正交。组k的市场因子载荷是λ_k^M，行业因子载荷是λ_k^C。

假设股票收益率服从多维正态分布$N(\mu,\Gamma\,\Lambda\,\Gamma)$，其中Γ是对角元素为标准差的对角线矩阵，Λ是由上述多因子模型构造的相关系数矩阵。

μ和Γ可用样本均值和标准差估计。待估参数$\theta=(\lambda_1^M,\ldots,\lambda_K^M,\lambda_1^C,\ldots,\lambda_K^C)'$。

四、案例 – 中国股票市场数据聚类分析

Oh and Patton, 2023基于EM迭代算法给出了这类模型的估计方法：

1. 给定分组$\hat{R}^{(s)}$，基于似然函数最优化参数θ：

$$\hat{\theta}^{(s+1)} = \arg\max_{\theta} L(\theta, \hat{R}^{(s)}),$$

2. 给定参数参数$\hat{\theta}^{(s+1)}$，最优化似然函数选择最优分组：

$$\hat{R}^{(s+1)} = \arg\max_{R} L(\hat{\theta}^{(s+1)}, R).$$

3. 迭代更新参数$\hat{\theta}^{(s)}$和$\hat{R}^{(s)}$直至收敛。该算法可以保证收敛到局部最优解，可以使用 10 个随机的初始值来提高估计量的准确性。

Oh and Patton, 2023提出用BIC准则来寻找最优的聚类总数。

四、案例 – 中国股票市场数据聚类分析

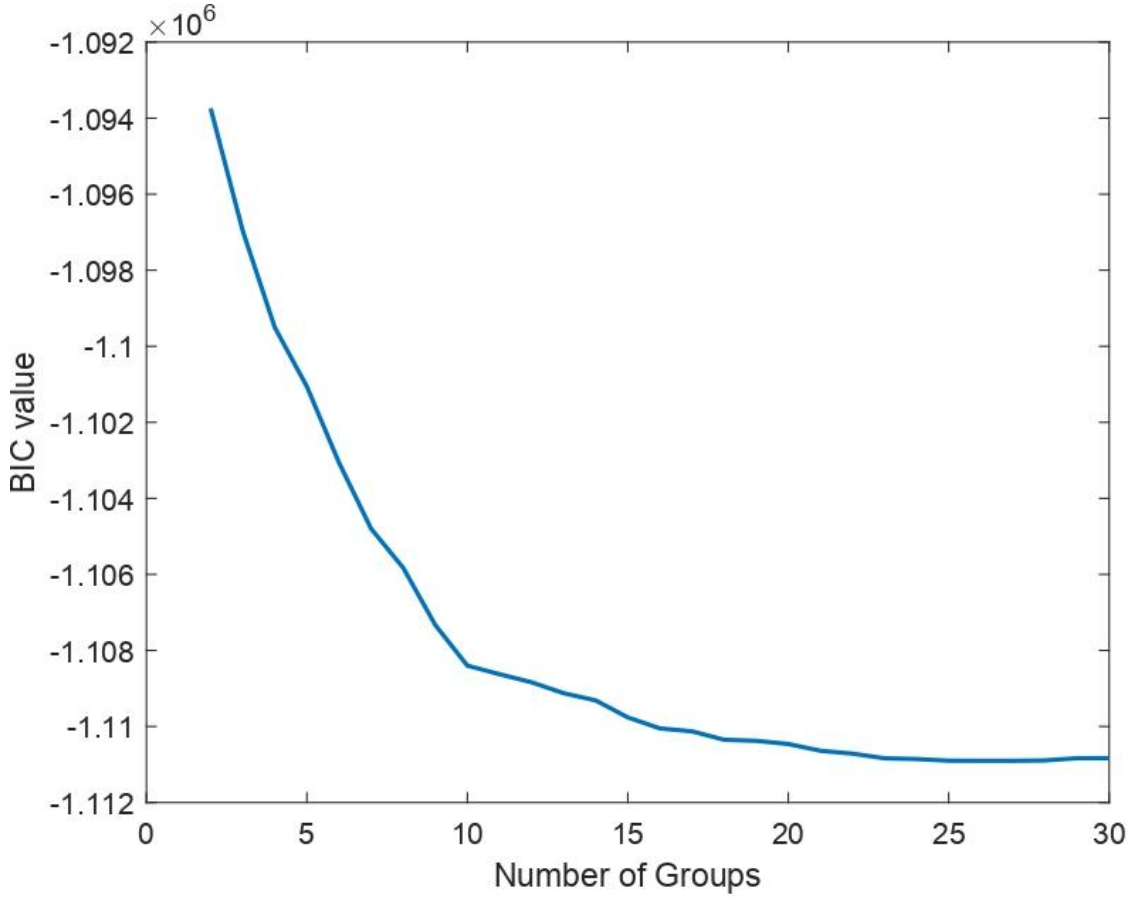

图10.3 BIC criterion across values of K

四、案例 – 中国股票市场数据聚类分析

表10.2 EM算法，K=27

证券代码	证券简称	行业	证券代码	证券简称	行业	证券代码	证券简称	行业
600000	浦发银行	银行	600011	华能国际	公用事业	600063	皖维高新	基础化工
600015	华夏银行	银行	600021	华电国际	公用事业	600075	新疆天业	基础化工
600016	民生银行	银行	600023	浙能电力	公用事业	600089	特变电工	综合
600036	招商银行	银行	600027	华电国际	公用事业	600096	云天化	基础化工
600004	白云机场	交通运输	600066	宇通客车	汽车	600029	南方航空	交通运输
600009	上海机场	交通运输	600104	上汽集团	汽车	600115	中国东航	交通运输
600007	中国国贸	房地产	600070	*ST 富润	综合	600073	光明肉业	食品饮料
600048	保利发展	房地产	600071	凤凰光学	电子	600108	亚盛集团	农林牧渔
600094	大名城	房地产	600083	*ST 博信	综合	600127	金健米业	农林牧渔
600030	中信证券	非银金融	600031	三一重工	机械设备	600105	永鼎股份	通信
600061	国投资本	非银金融	600059	古越龙山	食品饮料	600110	诺德股份	电力设备
600109	国金证券	非银金融	600060	海信视像	家用电器	600114	东睦股份	机械设备
600020	中原高速	交通运输	600038	中直股份	国防军工	600037	歌华有线	传媒
600033	福建高速	交通运输	600072	中船科技	电力设备	600088	中视传媒	传媒
600035	楚天高速	交通运输	600118	中国卫星	国防军工	600100	同方股份	计算机
600055	万东医疗	医药生物	600053	九鼎投资	非银金融	600051	宁波联合	综合
600056	中国医药	医药生物	600095	湘财股份	非银金融	600052	东望时代	公用事业
600062	华润双鹤	医药生物	600120	浙江东方	非银金融	600067	冠城新材	电力设备
600079	人福医药	医药生物	600064	南京高科	房地产	600076	康欣新材	轻工制造
600085	同仁堂	医药生物	600125	铁龙物流	交通运输	600080	金花股份	医药生物
600129	太极集团	医药生物	600180	瑞茂通	交通运输	600082	海泰发展	房地产
600010	包钢股份	钢铁	600012	皖通高速	交通运输	600084	中信尼雅	食品饮料
600019	宝钢股份	钢铁	600018	上港集团	交通运输	600097	开创国际	农林牧渔
600022	山东钢铁	钢铁	600026	中远海能	交通运输	600099	林海股份	汽车
600058	五矿发展	商贸零售	600028	中国石化	石油石化	600106	重庆路桥	交通运输
600111	北方稀土	有色金属	600050	中国联通	通信	600107	美尔雅	纺织服装
600117	西宁特钢	钢铁	600039	四川路桥	建筑装饰	600113	浙江东日	商贸零售
600121	郑州煤电	煤炭	600078	澄星股份	基础化工	600119	长江投资	交通运输
600123	郑州煤电	煤炭	600116	三峡水利	公用事业	600128	苏豪弘业	商贸零售
600126	杭钢股份	钢铁	600131	国网信通	计算机	600130	波导股份	电子
600006	东风股份	汽车	600098	广州发展	公用事业	600008	首创环保	公用事业
600081	东风科技	汽车	600101	明星电力	公用事业	600057	厦门象屿	交通运输
600054	黄山旅游	社会服务	600017	日照港	交通运输	600103	青山纸业	轻工制造

谢 谢!