Expectation Maximization Algorithm

EM 算法是一种从不完全数据或者含有隐含变量（hidden variable）的数据集中求解概率模型参数的极大似然估计方法，采用迭代的方式，每次迭代分为两步：E 步：求期望（expectation）；M 步：求极大似然（maximization）。

1.从极大似然估计到 EM 算法

1.1 引出

在之前的学习过程中，我们知道在已知数据的分布而不知具体分布参数的时候，我们会使用极大似然估计来估计出该分布的参数 Θ，具体过程为：

1 写出似然函数 $L(\theta) = P(X|\theta), \theta \in \Theta$
2.对似然函数取对数，得到 log 形式$H(\theta) = logL(\theta) = log(P(X;\theta)), \theta \in \Theta$
3.对对数似然函数求导，令其为 0，得到似然方程
4.求解似然方程，得到所求参数

极大似然估计，只是一种概率论在统计学的应用，它是参数估计的方法之一。假设已知某个随机样本满足某种概率分布，但其中具体参数不清楚，参数估计就是通过若干次试验，观察其结果，利用结果推出参数的估计值。最大似然估计是建立在这样的思想上：已知某个参数能使这个样本出现的概率最大，我们当然不会再去选择其他小概率的样本，所以干脆就把这个参数作为估计的真实值。

与最大似然估计不同的是，EM 所处理的是不完备的数据，其中含有隐含变量，也就是说很难直接写出似然函数，我们需要通过隐含变量的介入，得到隐变量条件下的似然函数，再进一步进行求解。

形式化描述：假设我们有一个观测样本集$X =$ （$x_1, x_2, x_3, \dots, x_n$），这些样本属于不同的类别$Z = (z_1, z_2, \dots, z_m)$，即模型中的隐变量数据，联合分布$P(X, Z|\theta)$,条件分布$P(Z|X,\theta)$但任务是求模型$P(x,z)$的参数 Θ，此时因为隐变量Z的存在，使得观测样本不是完全数据，最大似然很难直接用于求解，自然地想法是如果我们知道隐变量Z，那么问题便会变得简单。此时问题变成

$$H(\theta) = \ln P(X|\theta)$$

$$H(\theta) = \ln \sum_Z P\,(X, Z|\theta)$$

对于（1）式，即为似然函数,我们的目标是去最大化（1）式，所以我们根据联合概率密度下求边缘概率密度的公式，于是我们得到了（2）式，显然去对一个和的 log 函数求导并不是一件容易的事情，于是我们引入隐含变量 Z 的分布$q(z)$，下面我们会对其进行具体的分析和推导.

1.2Jensen 不等式

对于一凸函数$f(x)$，我们有如下性质：

$$E[f(x)] \geq f(E[x])$$

通俗的讲就是对于一个凸函数，函数的期望大于等于期望的函数。

看下图：

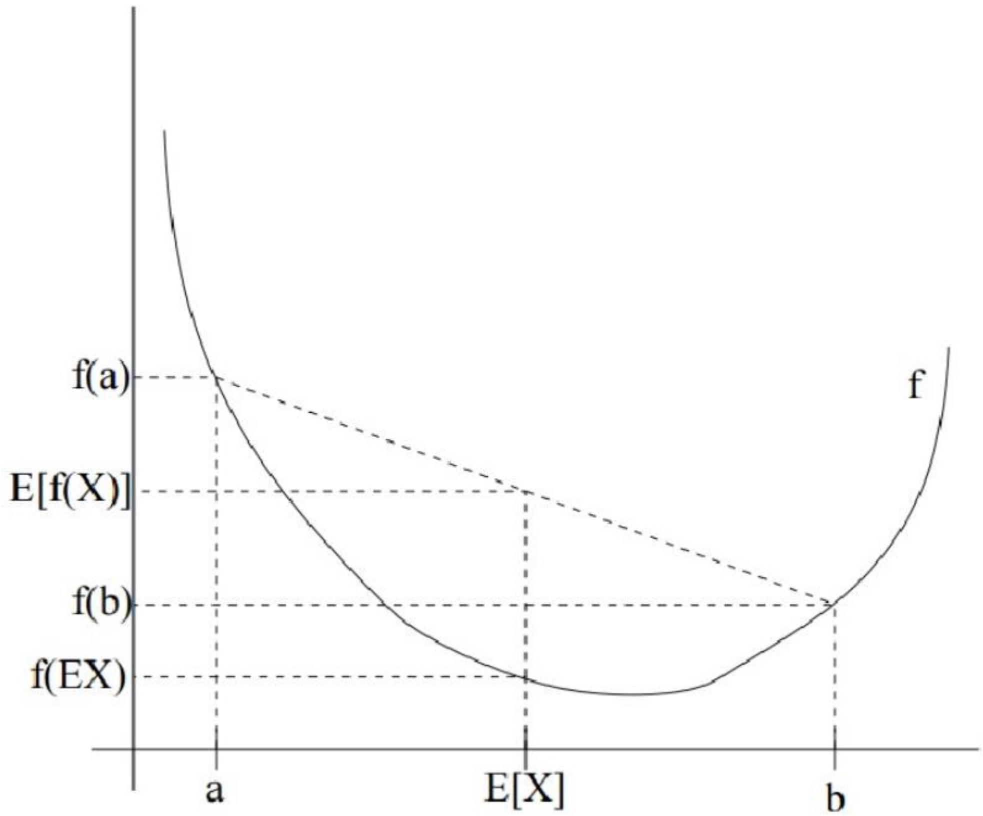

图 1.1

由上图可以简单证明： 挑选 a，b 中一点 c，可以将 c 表示成$c = ta + (1 - t)b$，那么可以得到$[tf(a) + (1 - t)f(b)] \geq f(c) = f(ta + (1 - t)b)$，遂简单得证，在此基础上也可以引申用数学归纳法进行完整证明。

2.EM 算法的数学推导

2.1 推导

通过引入隐含变量Z的分布函数$q(Z)$，我们将目标函数分解成下式：

$$H(\theta) = \ln \sum_{Z} P(X, Z|\theta)$$

$$H(\theta) = \mathcal{L}(q,\theta) + KL(q||p)$$

定义了$\mathcal{L}(q.\theta)$*和*$KL(q||p)$

其中$\mathcal{L}(q,\theta) = \sum_Z q\,(Z)\ln\frac{p(X,Z|\theta)}{q(Z)}$是概率分布 q(Z)的一个泛函，亦为参数$\theta$的一个函数；

而$KL(q||p) = -\sum_Z q\,(Z)\ln\frac{p(Z|X,\theta)}{q(Z)}$代表的是分布$q(Z)$和$p(Z|,X,\theta)$之间的 KL 散度。可以通过下图直观了解

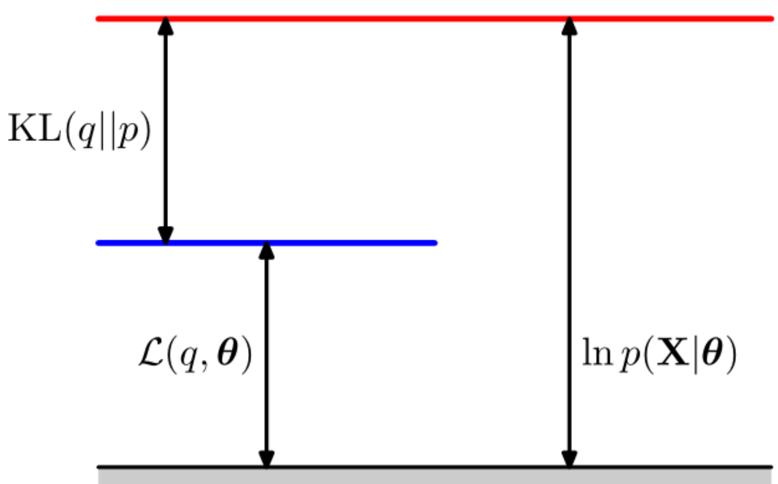

图 2.1

该图是对（2）式的分解说明，其对 Z 的任何分布都成立。

对（2）式予以证明：

$$\ln p(X,Z|\theta) = \ln p(Z|X,\theta) + \ln p(X|\theta)$$

将其代入$\mathcal{L}(q,\theta)$的表达式，可以将$KL(q||p)$消去，得到了目标$P(X|\theta)$，并且 Kull-Leibler 散度当且仅当$q(Z) = p(Z|X,\theta)$时大于等于 0,由此可得$\mathcal{L}(p,\theta)$是$\ln p(X|\theta)$的下界。

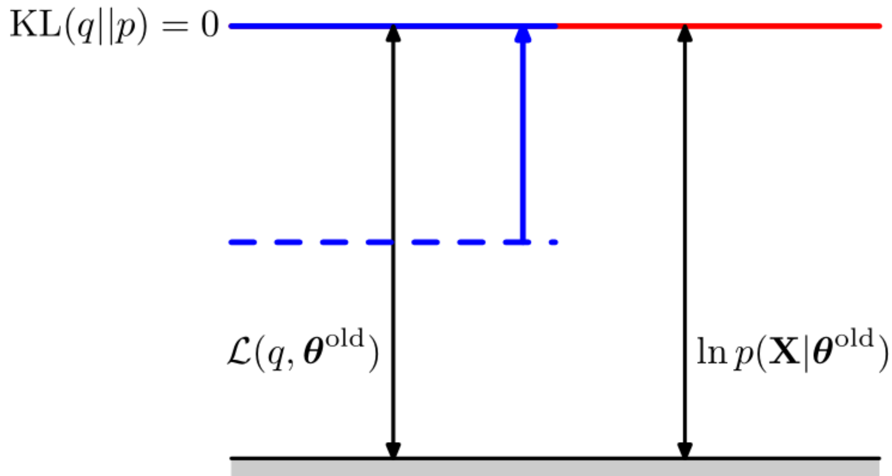

图 2.2

假设参数向量当前值为θ'，为上一轮的旧值，
在 E 步骤中，固定参数θ'让下界$\mathcal{L}(q,\theta)$关于$q(Z)$最大化，实际上，当$\mathcal{L}(q,\theta)$达到$\ln p(X|\theta')$时，即$KL(q||p)=0$，即$q(Z)$等价于$p(Z|x,\theta')$，此时对于目标（2）式可以转化为：

$$H(\theta)=\mathcal{L}(q,\theta)=\sum_{Z} q\,(Z)\ln\frac{p(X,Z|\theta)}{q(Z)}$$

$$=\sum_{Z} p\,(Z|X,\theta')\ln p(X,Z|\theta)-\sum_{Z} p\,(Z|X,\theta')\ln P(Z|x,\theta')=Q(\theta,\theta')+H(q(Z))$$

上式中可以看出$H(q(Z))$为$q(Z)$的信息熵为一个常数，所以通过 E 步我们得到了 Q 函数和$q(z)$的信息熵，所以接下来会在 M 步中将 Q 函数作为最大化目标。 >Q 函数：

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

即为完全数据的对数似然函数$\ln p(X,Z|\theta)$在给定条件X,θ'下关于条件分布$p(Z|X,\theta')$的期望

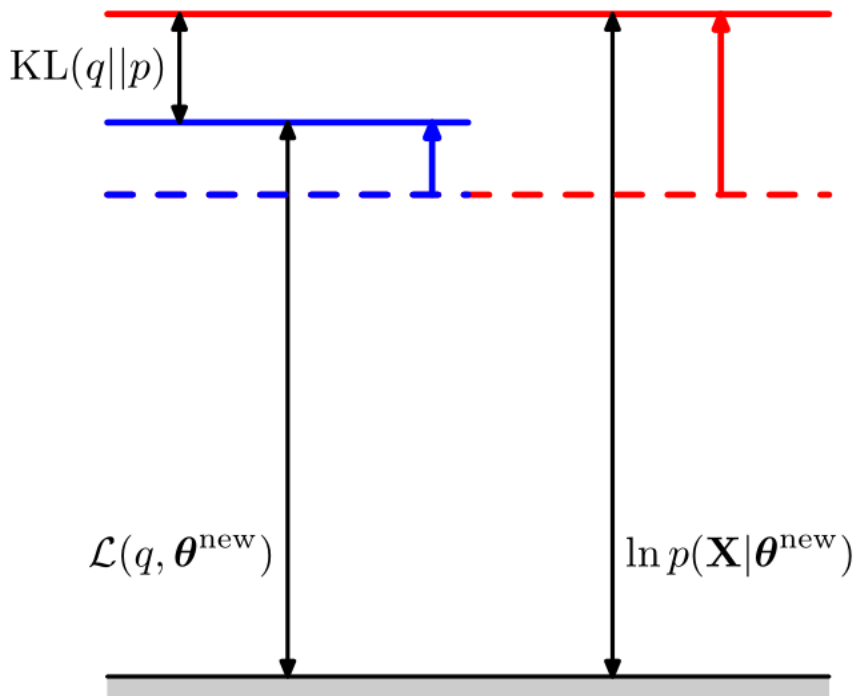

图 2.3

在 M 步骤中，固定$q'(Z)$，让下界$\mathcal{L}(q',\theta)$关于θ最大化，并得到更新后的参数值

$\theta = \underset{\theta}{\operatorname{argmax}} Q(\theta,\theta')$

由下图可知，$\mathcal{L}$会随着θ的更新而变大，但在这过程中$q(Z)$是保持不变的，所以其必不等价于$p(Z|X,\theta)$，所以其 KL 散度必大于 0，意味着似然函数的增量大于$\mathcal{L}$，直至其达到极大值。

稍微总结一下，EM 算法是通过迭代逐步近似来极大化$L(\theta)$的。可以通过下图来加深理解：

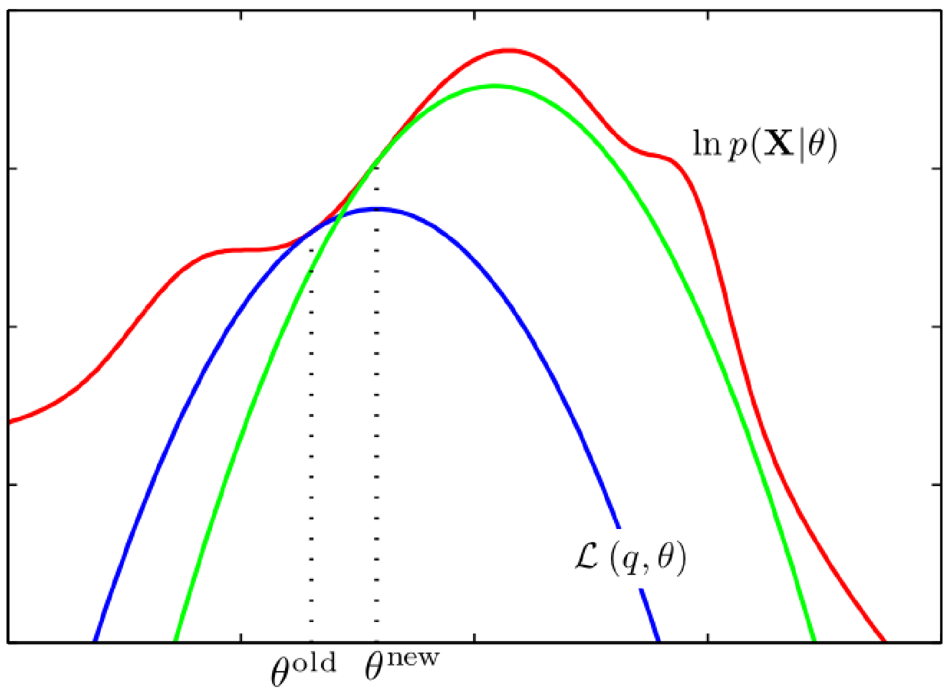

图 2.4

图中红色曲线为我们的目标曲线$\ln p(X|\theta)$，目标是求求其最大值，我们先给定一个参数初始值θ'，在 E 步骤中计算潜在变量上的后验概率分布$p(Z|X,\theta')$得到$\mathcal{L}(q,\theta)$更大的一个下界，与不完全数据的对数似然函数在θ'处相切，在图中显示为蓝色，然后在 M 步中最大化$\mathcal{L}(q,\theta)$更新θ值，通过下一轮的 E 步去得到新的下界，图中为绿色曲线。可以看到，EM 算法中的 E 步和 M 步都增大了对数似然函数的一个良好定义的下界的值，完整的 EM 循环会使得模型的参数向着使对数似然函数增大的方向去改变，因为目标对数似然函数是具有唯一的极大值，所以一直迭代下去是会找到不完整数据对数似然函数的极大值。

3.EM 算法的实际应用

3.1 EM 算法在 pLSA 中的应用

3.1.1 pLSA 的引出

pLSA 的概率图模型如下图：

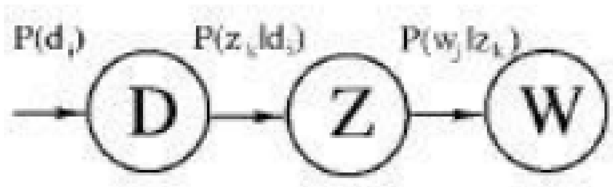

图 3.1

其中 D 代表文档，Z 代表隐含主题，W 代表观察到的单词，$p(d_i)$代表单词出现在文档d_i中的概率，$p(z_k|d_i)$代表文档d_i中出现在主题z_k下的单词的概率，$p(w_j|z_k)$代表给定主题z_k下出现单词w_j的概率。这其中主题z_k信息为隐藏变量。并且假设前提是每个主题在所有词项上的分布服从 Multinomial 分布，每篇文档在所有主题上的分布服从 Multinomial 分布。由此可得文档的生成过程：>文档生成过程：

（1）以$p(d_i)$的概率选中文档d_i

（2）以$p(z_k|d_i)$的概率选中文档中的主题z_k

（3）以$p(w_j|z_k)$的概率产生一个单词。

其中可观测的数据位单词文档对(d_i, w_j)，而主题信息z_k为隐含变量。于是观测数据的联合分布如下：

$$p(d_i, w_j) = p(d_i) = p(d_i)p(w_j|d_i)$$

$$p(w_j|d_i) = \sum_Z p\,(w_j|z_k, d_i) = \sum_Z p\,(w_j|z_k)p(z_k|d_i)$$

其中$p(z_k|d_i)$和$p(w_j|z_k)$分别对应两组 Multinomial 分布，下面我们用 EM 算法来对 pLSA 参数的详细推导。

3.1.2 pLAS 参数的 EM 算法推导

倘若我们试图用 MLE 方法来估计参数的话，我们会得到下面的式子：

$$L(\theta) = \sum_N \sum_M n\,(d_i, w_j)\log p(d_i, w_j) = \sum_N \sum_M n\,(d_i, w_j) \sum_Z \log p(d_i)p(w_j|d_i)$$

可以观察到上式中一共有$N * K + M * K$各自变量，倘若对这些自变量分别进行求偏导，然后再令其为零，求解难度是非常大的。，所以我们使用 EM 算法进行求解。

EM 算法步骤为：

（1）E 步：首先计算出 Q 函数，即为完全数据在条件分布下的期望

（2）M 步：计算出 Q 的极大值的参数θ

有：

$$p(z_k|d_i,w_j)=\frac{p(z_k|d_i)p(w_j|z_k)}{\sum_{l=1}^{K}p\,(z_l|d_i)p(w_j|z_l)}$$

其中$p(z_k|d_i)p(w_j|z_k)$是由上一轮的 M 步中固定住的，已知。
那么 Q 函数为：

$$Q=\sum_{N}\sum_{M}n\,(d_i,w_j)\sum_{Z}p\,(z_k|d_i,w_j)\log p(z_k|d_i)p(w_j|z_k)$$

对其进行最大化求参数操作，这是一个多元函数求极值的问题，可以使用拉格朗日乘子法，将条件机制问题转化成无条件极值问题，其约束条件为$\sum_{l=1}^{K}p\,(z_l|d_i)=1,\sum_{j=1}^{M}p\,(w_j|z_k)=1$

由此可以写出拉格朗日方程：

$$\mathcal{H}=Q+\sum_{k=1}^{K}\tau_k\,(1-\sum_{l=1}^{K}p\,(z_l|d_i))+\sum_{j=1}^{M}\rho_j\,(1-\sum_{j=1}^{M}p\,(w_j|z_k))$$

让$\mathcal{H}$分别对自变量$p(z_l|d_i),p(w_j|z_k)$求偏导可得：

$$\sum_{i=1}^{N}n\,(d_i,w_j)p(z_k|d_i,w_j)-\tau_k p(w_j|z_k)=0,1\le i\le M,1\le k\le K$$

$$\sum_{j=1}^{M}n\,(d_i,w_j)p(z_k|d_i,w_j)-\rho_i p(z_k|d_i)=0,1\le j\le N,1\le k\le K$$

解得：

$$\tau_i=-n(d_i),\rho_i=-\sum_{i=1}^{N}\sum_{j=1}^{M}n\,(d_i,w_j)p(z_k|d_i,w_j)$$

代入上式，可解得：

$$p(w_j|z_k)=\frac{\sum_{i=1}^{N}n\,(d_i,w_j)p(z_k|d_i,w_j)}{\sum_{i=1}^{N}\sum_{j=1}^{M}n\,(d_i,w_j)p(z_k|d_i,w_j)}$$

$$p(z_k|d_i)=\frac{\sum_{j=1}^{M}n\,(d_i,w_j)p(z_k|d_i.w_j)}{n(d_i)}$$

使用更新后的参数值，又进入 E 步骤，计算隐含变量z_k在当前估计的参数情况下的后验概率，如此这般，不断迭代，直至满足终止阈值条件。

3.2EM 算法在 GMM 中的应用

3.2.1GMM 模型简介

混合高斯模型的模型分布，认为随机变量$\mathcal{X}$服从一个多峰的高斯分布，其由多个高斯分布组合而成，概率图如下

Figure 9.9 This shows the same graph as in Figure 9.6 except that we now suppose that the discrete variables $\mathbf{z}_n$ are observed, as well as the data variables $\mathbf{x}_n$.

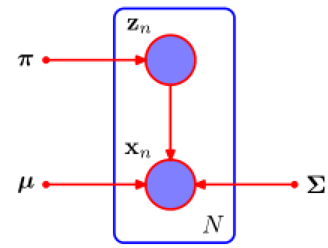

图 3.2.1

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

其中

$$\mathcal{N}(x|\mu_k, \Sigma_k) = \frac{1}{(2\pi)^{D/2}} \frac{1}{(|\Sigma_k|)^{1/2}} \exp\{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)\}$$

根据上面给出的概率密度函数，如果我们要从 GMM 的分布中随机取样，实际上可以分成两步：首先随即在 K 个 Component 的子高斯分布中选一个，每个 Coponent 被选中的概率就是π_k，再单独考虑从这个 Component 中随机选取样本点，在 PRML 中，引入了一个 K 维的二值随机变量 z，其中只有 1 维为 1，其余为 0，非零的维对应的就是 GMM 参数样本时被选中的 Component，其概率为π_k，即

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

3.2.2GMM 参数的 EM 推导

针对随机变量服从的混合高斯分布模型，有观察变量的对数似然函数，形式如下：

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

亦写出 Q 函数，形式为：

$$Q = E[\log P(X, \gamma|\theta)] = \sum_{i=1}^{N} \gamma_{jk} \{\ln \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)\}$$

其中γ_{jk}代表第 j 个观测样本来自第 k 个 Component 的后验概率，记为：

$$\gamma_{jk} = P(z_k = 1|X) = \frac{p(\gamma_{jk} = 1, x_j|\theta)}{\sum_{k=1}^{K} p\,(\gamma_{jk} = 1, x_j|\theta)} = \frac{p(x_j|\gamma_{jk} = 1, \theta)p(\gamma_{jk} = 1|\theta)}{\sum_{k=1}^{K} p\,(x_j|\gamma_{jk} = 1, \theta)p(\gamma_{jk} = 1|theta)}$$
$$= \frac{\pi_k \mathcal{N}(x|\mu_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k\, \mathcal{N}(x|\mu_k, \Sigma_k)}$$

令 Q 极大，分别对μ_k和Σ_k求偏导，可以解出：

$$\mu_k = \frac{\sum_{j=1}^{N} \gamma_{jk}\, x_j}{\sum_{j=1}^{N} \gamma_{jk}}$$

$$\Sigma_k = \frac{\sum_{j=1}^{N} \gamma_{jk}\, (x_j - \mu_k)(x_j - \mu_k)^T}{\sum_{j=1}^{N} \gamma_{jk}}$$

加上约束$\sum \pi_k = 1$，应用拉格朗日算子法，可以求得

$$\pi_k = \frac{N_k}{N} = \frac{\sum_{j=1}^{N_k} \gamma_{jk}}{N}$$

重复上述步骤，直至似然函数收敛。