这里的不等号由 Jensen 不等式得到.

由式 (9.22) 和式 (9.23) 即知式 (9.21) 右端是非负的. ■

定理 9.2 设 $L(\theta)=\log P(Y\mid\theta)$ 为观测数据的对数似然函数，$\theta^{(i)}\,(i=1,2,\cdots)$ 为 EM 算法得到的参数估计序列，$L(\theta^{(i)})\,(i=1,2,\cdots)$ 为对应的对数似然函数序列.

（1）如果 $P(Y\mid\theta)$ 有上界，则 $L(\theta^{(i)})=\log P(Y\mid\theta^{(i)})$ 收敛到某一值 L^*；

（2）在函数 $Q(\theta,\theta')$ 与 $L(\theta)$ 满足一定条件下，由 EM 算法得到的参数估计序列 $\theta^{(i)}$ 的收敛值 θ^* 是 $L(\theta)$ 的稳定点.

证明 （1）由 $L(\theta)=\log P(Y\mid\theta^{(i)})$ 的单调性及 $P(Y\mid\theta)$ 的有界性立即得到.

（2）证明从略，参阅文献［6］. ■

定理 9.2 关于函数 $Q(\theta,\theta')$ 与 $L(\theta)$ 的条件在大多数情况下都是满足的. EM 算法的收敛性包含关于对数似然函数序列 $L(\theta^{(i)})$ 的收敛性和关于参数估计序列 $\theta^{(i)}$ 的收敛性两层意思，前者并不蕴涵后者. 此外，定理只能保证参数估计序列收敛到对数似然函数序列的稳定点，不能保证收敛到极大值点. 所以在应用中，初值的选择变得非常重要，常用的办法是选取几个不同的初值进行迭代，然后对得到的各个估计值加以比较，从中选择最好的.

9.3 EM 算法在高斯混合模型学习中的应用

EM 算法的一个重要应用是高斯混合模型的参数估计. 高斯混合模型应用广泛，在许多情况下，EM 算法是学习高斯混合模型（Gaussian misture model）的有效方法.

9.3.1 高斯混合模型

定义 9.2（高斯混合模型） 高斯混合模型是指具有如下形式的概率分布模型：

$$P(y\mid\theta)=\sum_{k=1}^{K}\alpha_k\phi(y\mid\theta_k) \tag{9.24}$$

其中，α_k 是系数，$\alpha_k\geqslant 0$，$\sum_{k=1}^{K}\alpha_k=1$；$\phi(y\mid\theta_k)$ 是高斯分布密度，$\theta_k=(\mu_k,\sigma_k^2)$，

$$\phi(y\mid\theta_k)=\frac{1}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{(y-\mu_k)^2}{2\sigma_k^2}\right) \tag{9.25}$$

称为第 k 个分模型.

一般混合模型可以由任意概率分布密度代替式 (9.25) 中的高斯分布密度，我们只介绍最常用的高斯混合模型.

9.3.2 高斯混合模型参数估计的 EM 算法

假设观测数据 $y_1, y_2, \cdots, y_N$ 由高斯混合模型生成，

$$P(y \mid \theta) = \sum_{k=1}^{K} \alpha_k \phi(y \mid \theta_k) \tag{9.26}$$

其中，$\theta = (\alpha_1, \alpha_2, \cdots, \alpha_K; \theta_1, \theta_2, \cdots, \theta_K)$．我们用 EM 算法估计高斯混合模型的参数 θ．

1．明确隐变量，写出完全数据的对数似然函数

可以设想观测数据 y_j，$j = 1,2,\cdots,N$，是这样产生的：首先依概率 α_k 选择第 k 个高斯分布分模型 $\phi(y \mid \theta_k)$；然后依第 k 个分模型的概率分布 $\phi(y \mid \theta_k)$ 生成观测数据 y_j．这时观测数据 y_j，$j = 1,2,\cdots,N$，是已知的；反映观测数据 y_j 来自第 k 个分模型的数据是未知的，$k = 1,2,\cdots,K$，以隐变量 γ_{jk} 表示，其定义如下：

$$\gamma_{jk} = \begin{cases} 1, & \text{第 } j \text{ 个观测来自第 } k \text{ 个分模型} \\ 0, & \text{否则} \end{cases}$$

$$j = 1,2,\cdots,N; \quad k = 1,2,\cdots,K \tag{9.27}$$

γ_{jk} 是 0-1 随机变量．

有了观测数据 y_j 及未观测数据 γ_{jk}，那么完全数据是

$$(y_j, \gamma_{j1}, \gamma_{j2}, \cdots, \gamma_{jK}), \quad j = 1,2,\cdots,N$$

于是，可以写出完全数据的似然函数：

$$\begin{aligned} P(y, \gamma \mid \theta) &= \prod_{j=1}^{N} P(y_j, \gamma_{j1}, \gamma_{j2}, \cdots, \gamma_{jK} \mid \theta) \\ &= \prod_{k=1}^{K} \prod_{j=1}^{N} \left[\alpha_k \phi(y_j \mid \theta_k) \right]^{\gamma_{jk}} \\ &= \prod_{k=1}^{K} \alpha_k^{n_k} \prod_{j=1}^{N} \left[\phi(y_j \mid \theta_k) \right]^{\gamma_{jk}} \\ &= \prod_{k=1}^{K} \alpha_k^{n_k} \prod_{j=1}^{N} \left[\frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{(y_j - \mu_k)^2}{2\sigma_k^{\ 2}} \right) \right]^{\gamma_{jk}} \end{aligned}$$

式中，$n_k = \sum_{j=1}^{N} \gamma_{jk}$，$\sum_{k=1}^{K} n_k = N$．

那么，完全数据的对数似然函数为

$$\log P(y, \gamma \mid \theta) = \sum_{k=1}^{K} n_k \log \alpha_k + \sum_{j=1}^{N} \gamma_{jk} \left[\log\left(\frac{1}{\sqrt{2\pi}} \right) - \log \sigma_k - \frac{1}{2\sigma_k^2} (y_j - \mu_k)^2 \right] \tag{9.28}$$

2. EM 算法的 E 步：确定 Q 函数

$$
\begin{aligned}
Q(\theta,\theta^{(i)}) &= E[\log P(y,\gamma\mid\theta)\mid y,\theta^{(i)}] \\
&= E\left\{\sum_{k=1}^{K} n_k \log\alpha_k + \sum_{j=1}^{N}\gamma_{jk}\left[\log\left(\frac{1}{\sqrt{2\pi}}\right) - \log\sigma_k - \frac{1}{2\sigma_k^2}(y_j-\mu_k)^2\right]\right\} \\
&= \sum_{k=1}^{K}\left\{\sum_{j=1}^{N}(E\gamma_{jk})\log\alpha_k + \sum_{j=1}^{N}(E\gamma_{jk})\left[\log\left(\frac{1}{\sqrt{2\pi}}\right) - \log\sigma_k - \frac{1}{2\sigma_k^2}(y_j-\mu_k)^2\right]\right\}
\end{aligned}
$$

这里需要计算 $E(\gamma_{jk}\mid y,\theta)$，记为 $\hat{\gamma}_{jk}$.

$$
\begin{aligned}
\hat{\gamma}_{jk} &= E(\gamma_{jk}\mid y,\theta) = P(\gamma_{jk}=1\mid y,\theta) \\
&= \frac{P(\gamma_{jk}=1,y_j\mid\theta)}{\sum_{k=1}^{K}P(\gamma_{jk}=1,y_j\mid\theta)} \\
&= \frac{P(y_j\mid\gamma_{jk}=1,\theta)P(\gamma_{jk}=1\mid\theta)}{\sum_{k=1}^{K}P(y_j\mid\gamma_{jk}=1,\theta)P(\gamma_{jk}=1\mid\theta)} \\
&= \frac{\alpha_k\phi(y_j\mid\theta_k)}{\sum_{k=1}^{K}\alpha_k\phi(y_j\mid\theta_k)}, \quad j=1,2,\cdots,N;\ k=1,2,\cdots,K
\end{aligned}
$$

$\hat{\gamma}_{jk}$ 是在当前模型参数下第 j 个观测数据来自第 k 个分模型的概率，称为分模型 k 对观测数据 y_j 的响应度.

将 $\hat{\gamma}_{jk}=E\gamma_{jk}$ 及 $n_k=\sum_{j=1}^{N}E\gamma_{jk}$ 代入式 (9.28) 即得

$$
Q(\theta,\theta^{(i)}) = \sum_{k=1}^{K} n_k\log\alpha_k + \sum_{k=1}^{N}\hat{\gamma}_{jk}\left[\log\left(\frac{1}{\sqrt{2\pi}}\right) - \log\sigma_k - \frac{1}{2\sigma_k^2}(y_j-\mu_k)^2\right] \tag{9.29}
$$

3. 确定 EM 算法的 M 步

迭代的 M 步是求函数 $Q(\theta,\theta^{(i)})$ 对 θ 的极大值，即求新一轮迭代的模型参数：

$$
\theta^{(i+1)} = \arg\max_{\theta} Q(\theta,\theta^{(i)})
$$

用 $\hat{\mu}_k$，$\hat{\sigma}_k^2$ 及 $\hat{\alpha}_k$，$k=1,2,\cdots,K$，表示 $\theta^{(i+1)}$ 的各参数. 求 $\hat{\mu}_k$，$\hat{\sigma}_k^2$ 只需将式 (9.29) 分别对 μ_k，σ_k^2 求偏导数并令其为 0，即可得到；求 $\hat{\alpha}_k$ 是在 $\sum_{k=1}^{K}\alpha_k=1$ 条件下求偏导数并令其为 0 得到的. 结果如下：

$$\hat{\mu}_k = \frac{\sum_{j=1}^{N} \hat{\gamma}_{jk} y_j}{\sum_{j=1}^{N} \hat{\gamma}_{jk}}, \qquad k = 1,2,\cdots,K \tag{9.30}$$

$$\hat{\sigma}_k^2 = \frac{\sum_{j=1}^{N} \hat{\gamma}_{jk} (y_j - \mu_k)^2}{\sum_{j=1}^{N} \hat{\gamma}_{jk}}, \qquad k = 1,2,\cdots,K \tag{9.31}$$

$$\hat{\alpha}_k = \frac{n_k}{N} = \frac{\sum_{j=1}^{N} \hat{\gamma}_{jk}}{N}, \qquad k = 1,2,\cdots,K \tag{9.32}$$

重复以上计算，直到对数似然函数值不再有明显的变化为止.

现将估计高斯混合模型参数的 EM 算法总结如下：

算法 9.2（高斯混合模型参数估计的 EM 算法）

输入：观测数据 $y_1, y_2, \cdots, y_N$，高斯混合模型；

输出：高斯混合模型参数.

（1）取参数的初始值开始迭代

（2）E 步：依据当前模型参数，计算分模型 k 对观测数据 y_j 的响应度

$$\hat{\gamma}_{jk} = \frac{\alpha_k \phi(y_j \mid \theta_k)}{\sum_{k=1}^{K} \alpha_k \phi(y_j \mid \theta_k)}, \qquad j = 1,2,\cdots,N; \quad k = 1,2,\cdots,K$$

（3）M 步：计算新一轮迭代的模型参数

$$\hat{\mu}_k = \frac{\sum_{j=1}^{N} \hat{\gamma}_{jk} y_j}{\sum_{j=1}^{N} \hat{\gamma}_{jk}}, \qquad k = 1,2,\cdots,K$$

$$\hat{\sigma}_k^2 = \frac{\sum_{j=1}^{N} \hat{\gamma}_{jk} (y_j - \mu_k)^2}{\sum_{j=1}^{N} \hat{\gamma}_{jk}}, \qquad k = 1,2,\cdots,K$$

$$\hat{\alpha}_k = \frac{\sum_{j=1}^{N} \hat{\gamma}_{jk}}{N}, \qquad k = 1,2,\cdots,K$$

（4）重复第（2）步和第（3）步，直到收敛. ■