统计中的计算方法 · 课后作业（2）

梁子龙（15300180026）

2018 年 4 月 4 日

作业 1 *已知数据* $(1,1)$，$(1,-1)$，$(-1,1)$，$(-1,-1)$，$(?,2)$，$(?,2)$，$(?,-2)$，$(?,-2)$，$(2,?)$，$(2,?)$，$(-2,?)$，$(-2,?)$ *服从二元正态分布，试用 EM 方法估计缺失数据，并估计正态分布的参数．取多个初始值，观察得到的结果．*

答. 利用 R 编写 EM 算法实验，得到结果．实验发现，不论取任何初始值，最终均值均收敛到

$$\tilde{\mu}_1 = \tilde{\mu}_2 = 0,$$

当协方差矩阵初始值取 $\Sigma = I_2$ 时，求得结果为

$$\Sigma = \begin{pmatrix} 2.500 & 0 \\ 0 & 2.500 \end{pmatrix}.$$

但当均值或者协方差矩阵的初值做一些微小扰动，协方差矩阵收敛到的值会发生变化．如取 $\mu_1^{(0)} = \mu_2^{(0)} = 0.01$ 时，求得结果为

$$\Sigma = \begin{pmatrix} 2.667 & 1.333 \\ 1.333 & 2.667 \end{pmatrix}.$$

取 Σ 元素皆取 1 时，求得结果为

$$\Sigma = \begin{pmatrix} 2.667 & -1.333 \\ -1.333 & 2.667 \end{pmatrix}.$$

这表明上述三种情况可能是 EM 算法中对数似然函数的鞍点，都是局部极大值点． □

作业 2 *假设数据来自两个 Poisson 分布的混合分布，给出参数估计的 EM 算法，并在给定的数据上实验．*

答. Poisson 分布的概率函数为

$$f(x,\lambda) = \mathrm{e}^{-\lambda}\frac{\lambda^x}{x!}, \quad x \in \mathbb{N}. \tag{1}$$

对于 k 个 Poisson 分布组成的混合分布，设 τ_j 为第 j 个分布的的生成概率，我们引入变量

$$z_{ij}=\begin{cases}1 & \text{若样本 } x_i \text{ 由参数为 } \lambda_j \text{ 的分布生成;}\\ 0 & \text{若样本 } x_i \text{ 不由参数为 } \lambda_j \text{ 的分布生成;}\end{cases}\qquad j=1,2,\cdots,k. \tag{2}$$

此时，对于一个样本 $\{x_i\}_{i=1}^n$ 来说，对数似然函数可以表示为

$$\begin{aligned}\log L(\theta;x,z)&=\log\left(\prod_{i=1}^n\prod_{j=1}^k\left(\tau_j f(x_i;\lambda_j)\right)^{z_{ij}}\right)\\&=\sum_{i=1}^n\sum_{j=1}^k z_{ij}\left(\log\tau_j+x_i\log\lambda_j-\lambda_j-\log x_i!\right).\end{aligned}\tag{3}$$

为使用 EM 算法，需对引入的变量 z_{ij} 求期望进行迭代．设 $T_{ij}^{(t)}=\mathrm{E}(z_{ij}|x_i,\theta^{(t)})$，那么依期望计算公式，有

$$\begin{aligned}T_{ij}^{(t)}=\mathrm{P}(z_{ij}=1|x_i,\theta^{(t)})&=\frac{\mathrm{P}(z_{ij}=1,x_i,\theta^{(t)})}{\mathrm{P}(x_i,\theta^{(t)})}\\&=\frac{\tau_j^{(t)}f(x_i;\lambda_j^{(t)})}{\sum_{l=1}^k\tau_l^{(t)}f(x_i;\lambda_l^{(t)})}.\end{aligned}\tag{4}$$

利用计算得到的 $T_{ij}^{(t)}$，代入对数似然函数 (3)，得到

$$\begin{aligned}Q(\theta|\theta^{(t)})&=\mathrm{E}(\log L(\theta;x,z))\\&=\sum_{i=1}^n\sum_{j=1}^k T_{ij}^{(t)}\left(\log\tau_j+x_i\log\lambda_j-\lambda_j-\log x_i!\right).\end{aligned}\tag{5}$$

下面考察参数的迭代更新．首先考察 τ_j．将已经得到的 $Q(\theta|\theta^{(t)})$ 对 τ_j 求极值，得到一个条件极值问题，并利用 Lagrange 乘数法，得到

$$\begin{cases}\dfrac{\partial Q}{\partial\tau_j}+\alpha=\dfrac{1}{\tau_j}\displaystyle\sum_{i=1}^n T_{ij}^{(t)}+\alpha=0,\quad j=1,2,\ldots,k;\\ \displaystyle\sum_{j=1}^k\tau_j=1.\end{cases}\tag{6}$$

求解该方程组，得到

$$\tau_j^{(t+1)}=\frac{\sum_{i=1}^n T_{ij}^{(t)}}{n}.\tag{7}$$

接下来考察 λ_j. 为对其求极值，令

$$\begin{aligned}\frac{\partial Q}{\partial \lambda_j} &= \frac{\partial}{\partial \lambda_j}\left(\sum_{i=1}^{n} T_{ij}^{(t)}\left(x_i \log \lambda_j - \lambda_j\right)\right) \\ &= \sum_{i=1}^{n} T_{ij}^{(t)} \frac{x_i}{\lambda_j} - \sum_{i=1}^{n} T_{ij}^{(t)} = 0,\end{aligned} \tag{8}$$

得到

$$\lambda_j^{(t+1)} = \frac{\sum_{i=1}^{n} T_{ij}^{(t)} x_i}{\sum_{i=1}^{n} T_{ij}^{(t)}}. \tag{9}$$

至此，所需估计参数的迭代过程已构建完毕．现将以上 EM 算法总结如下：

1. 适当选取参数的初始值；
2. E-Step：依 (4) 求得 $T_{ij}^{(t)}$；
3. M-Step：依 (7), (9) 求得 $\tau_j^{(t+1)}, \lambda_j^{(t+1)}$；
4. 重复以上两步，直至参数结果收敛．

表 1: Poisson 分布混合分布的估计

	$j=1$	$j=2$
τ_j	0.430	0.570
λ_j	1.696	5.876

根据以上 EM 算法对数据 `assignment02-data.csv` 进行迭代约 70 次后，估计得到的实验数据如表 1 所示，大致的概率函数如图 1 所示．

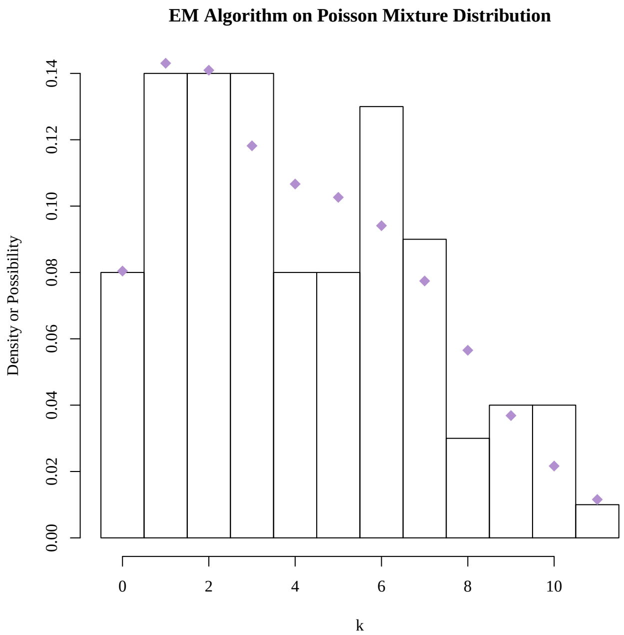

图 1: Poisson 分布混合分布的估计. 图中的直方图为原始样本的分布，紫色点图为混合分布的概率函数值.