

Statistical Computing

Chap. 2.1: Cases for EM

LIU, Ran

March 13, 2025

LIU, Ran - Department of Statistics @BNU

目录

EM 简述

两枚硬币正面概率估算

多项分布参数的 EM 算法

多项式分布的特点

正态分布参数 EM 估计

二项泊松混合模型的 EM 估计

实际例子



Li Q. Han Department of Statistics @BNU

EM 简述

我们考虑这样一个问题：完整数据是 Y ，但是 Y 只能被部分观测。我们假设 $Y = (X, Z)$ ，其中 X 是观测数据， Z 是缺失数据。

我们写出完全数据的似然函数

$$p(Y | \theta) = p(X, Z | \theta) = p(X | \theta) \cdot p(Z | X, \theta)$$

拆分出观测数据的似然函数：

$$\log p(X | \theta) = \log p(Y | \theta) - \log p(Z | X, \theta)$$

要关于 θ 最大化左边的部分，等价于最大化右边的部分。但是 Z 未知，无法优化，于是我们对两边分别求 $\mathbb{E}_Z [\cdots | X, \theta^{(t)}]$ ：

$$\begin{aligned} & \mathbb{E}_Z \left(\log p(X | \theta) | X, \theta^{(t)} \right) \\ &= \mathbb{E}_Z \left(\log p(Y | \theta) | X, \theta^{(t)} \right) - \mathbb{E}_Z \left(\log p(Z | X, \theta) | X, \theta^{(t)} \right) \end{aligned}$$

$$\log p(X | \theta) = Q(\theta | \theta^{(t)}) - H(\theta | \theta^{(t)}) \quad (\text{此时式子中不含 } Z)$$

观察 $H(\theta | \theta^{(t)})$, 我们有

$$\begin{aligned} H(\theta^{(t)} | \theta^{(t)}) - H(\theta | \theta^{(t)}) &= \mathbb{E}\{\log p(z|x, \theta^{(t)}) - \log p(z|x, \theta) | x, \theta^{(t)}\} \\ &= \int -\log \left[\frac{p(z|x, \theta)}{p(z|x, \theta^{(t)})} \right] p(z|x, \theta^{(t)}) dz \\ &\geq -\log \int p(z|x, \theta) dz \\ &= 0. \end{aligned}$$

于是我们只需要每次迭代保证, $Q(\theta^{(t+1)} | \theta^{(t)}) \geq Q(\theta^{(t)} | \theta^{(t)})$, 就可以实现:

$$\ell(\theta | X) = \log p(X | \theta) = Q(\theta | \theta^{(t)}) \uparrow - H(\theta | \theta^{(t)}) \downarrow$$

其中 $Q(\theta | \theta^{(t)}) = \mathbb{E}_Z (\log p(Y | \theta) | X, \theta^{(t)})$.

Liu, Ran - Department of Statistics @BNU

目录

EM 简述

两枚硬币正面概率估算

多项分布参数的 EM 算法

多项式分布的特点

正态分布参数 EM 估计

二项泊松混合模型的 EM 估计

实际例子



Department of Statistics @BNU

两枚硬币出现正面概率的 EM 算法

假设有 A, B 两枚硬币，其中正面朝上的概率分别为 p_A, p_B ，这两个参数是感兴趣的待估参数。

设计 6 组试验，每次实验投掷 5 次硬币，第 i 组试验结果为 $X_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}) (i = 1, \dots, 6)$ ， $x_{ij} = 1$ 表示硬币出现正面， $x_{ij} = 0$ 表示硬币出现反面。

令 Z_i 表示第 i 次投的硬币。 $Z_i = 1$ 表示是 A 硬币投的结果， $Z_i = 0$ 表示是 B 硬币投的结果。

两枚硬币出现正面概率的 EM 算法

如果知道每一组实验结果 X_i 是 A 硬币投的结果还是 B 硬币投的结果, 也就是观测到 Z_i 。

基于完整数据 $Y_i = (X_i, Z_i), i = 1, 2, \dots, 6$ 的对数似然函数为

$$\begin{aligned} \ell(p_A, p_B | Y) = & \sum_{k=1}^6 \{ Z_k (n_k \log p_A + (5 - n_k) \log(1 - p_A)) \\ & + (1 - Z_k) (n_k \log p_B + (5 - n_k) \log(1 - p_B)) \}. \end{aligned}$$

这里 $n_k = \sum_{j=1}^5 x_{kj}$. (黑板推导)

如果 $Z_i (i = 1, 2, \dots, 6)$ 可观测, 不难得到

$$\hat{p}_A = \frac{\sum_{k=1}^6 n_k Z_k}{5 \sum_{k=1}^6 Z_k}, \quad \hat{p}_B = \frac{\sum_{k=1}^6 n_k (1 - Z_k)}{5 \sum_{k=1}^6 (1 - Z_k)}.$$

也就是参数 p_A, p_B 的估计, 只需分别统计 A, B 硬币投的结果出现正面的次数, 然后除以分别投的总次数。

LIU, Ran - Department of Statistics @BNU

由于 $Z = (Z_1, \dots, Z_6)$ 没有观测到, 基于观测数据 $X = (X_1, \dots, X_6)$ 和 $p_A^{(i-1)}, p_B^{(i-1)}$, 条件期望为

$$\begin{aligned} Q(p_A, p_B; p_A^{(i-1)}, p_B^{(i-1)}) &= E(\ell(p_A, p_B | Y) | X; p_A^{(i-1)}, p_B^{(i-1)}) \\ &= \sum_{k=1}^6 \{E(Z_k | X; p_A^{(i-1)}, p_B^{(i-1)})(n_k \log p_A + (5 - n_k) \log(1 - p_A)) \\ &\quad + (1 - E(Z_k | X; p_A^{(i-1)}, p_B^{(i-1)}))(n_k \log p_B + (5 - n_k) \log(1 - p_B))\}. \end{aligned}$$

LIU, Ran - Department of Statistics @BNU

关于 p_A, p_B , 极大化 $Q(p_A, p_B; p_A^{(i-1)}, p_B^{(i-1)})$, 可得

$$\begin{aligned} p_A^{(i)} &= \frac{\sum_{k=1}^6 n_k E(Z_k | X; p_A^{(i-1)}, p_B^{(i-1)})}{5 \sum_{k=1}^5 E(Z_k | X; p_A^{(i-1)}, p_B^{(i-1)})} \\ p_B^{(i)} &= \frac{\sum_{k=1}^6 n_k (1 - E(Z_k | X; p_A^{(i-1)}, p_B^{(i-1)}))}{5 \sum_{k=1}^5 (1 - E(Z_k | X; p_A^{(i-1)}, p_B^{(i-1)}))}. \end{aligned}$$

LIU, Ran - Department of Statistics @BNU

对 $E(Z_k|X; p_A^{(i-1)}, p_B^{(i-1)})$, 不难推导可得

$$E(Z_k|X; p_A^{(i-1)}, p_B^{(i-1)}) = \frac{(p_A^{(i-1)})^{n_k} (1 - p_A^{(i-1)})^{5-n_k}}{(p_A^{(i-1)})^{n_k} (1 - p_A^{(i-1)})^{5-n_k} + (p_B^{(i-1)})^{n_k} (1 - p_B^{(i-1)})^{5-n_k}}$$

若是将每组硬币正面出现的次数当作 $X = \{n_1, \dots, n_6\}$:

$$\begin{aligned} E(Z_k|X; p_A^{(i-1)}, p_B^{(i-1)}) &= \left\{ \frac{C_5^{n_k}}{n_k!(5-n_k)!} (p_A^{(i-1)})^{n_k} (1 - p_A^{(i-1)})^{5-n_k} \right\} \\ &/ \left\{ \frac{C_5^{n_k}}{n_k!(5-n_k)!} (p_A^{(i-1)})^{n_k} (1 - p_A^{(i-1)})^{5-n_k} \right. \\ &\left. + \frac{C_5^{n_k}}{n_k!(5-n_k)!} (p_B^{(i-1)})^{n_k} (1 - p_B^{(i-1)})^{5-n_k} \right\}. \end{aligned}$$

估算是一样的。

LIU, Ran - Department of Statistics @BNU

上述计算过程有问题吗？

其实假定了 $p(z = 1) = p(z = 0) = \frac{1}{2}$.

LIU, Ran - Department of Statistics @BNU

假设观测到的数据为 $X_1 = (1, 1, 1, 0, 1)$, $X_2 = (0, 1, 0, 1, 1)$, $X_3 = (1, 0, 1, 1, 0)$, $X_4 = (0, 0, 1, 1, 1)$, $X_5 = (1, 0, 1, 0, 1)$ and $X_6 = (1, 1, 0, 0, 0)$, 用 EM 算法估计 p_A 和 p_B .

```
E <- function(nk, a, b){  
  A <- (choose(5, nk)/(factorial(nk)*factorial(5 - nk)))  
    * a^nk * (1 - a)^(5 - nk)  
  B <- (choose(5, nk)/(factorial(nk)*factorial(5 - nk)))  
    * b^nk * (1 - b)^(5 - nk)  
  E <- A/(A + B)  
  return(E)  
}
```

LIU, Ran - Department of Statistics @BNU

```
n <- c(4,3,3,3,3,2)
max.iter <- 100    #初值与最大循环数
pa <- c();pa[1] <- 0.6
pb <- c();pb[1] <- 0.4
for (i in 1:max.iter){#迭代
  EZ <- E(n, pa[i], pb[i])
  pa[i+1] <- sum(n*EZ)/(5*sum(EZ))
  pb[i+1] <- sum(n*(1-EZ))/(5*sum(1-EZ))
  if(abs(pa[i+1] - pa[i]) < 1e-8 &
     abs(pb[i+1] - pb[i]) < 1e-8) break
}
```

LIU, Ran - Department of Statistics @BNU

```
1 import math
2 def E(nk, a, b):
3     A = (a ** nk) * ((1 - a) ** (5 - nk))
4     B = (b ** nk) * ((1 - b) ** (5 - nk))
5     E = A / (A + B)
6     return E
7
8 n = [4, 3, 3, 3, 3, 2]
9 max_iter = 100
10 pa, pb = [0.6], [0.4]
11 for i in range(max_iter):
12     EZ = E(n, pa[i], pb[i])
13     pa_next = sum([nk * EZ_i for nk, EZ_i in zip(n, EZ)]) / (5
14 * sum(EZ))
15     pb_next = sum([(nk * (1 - EZ_i)) for nk, EZ_i in zip(n, EZ)
16 ]) / (5 * sum(1 - EZ))
17
18     if abs(pa_next - pa[i]) < 1e-8 and abs(pb_next - pb[i]) < 1
19 e-8:
20         break
21     pa.append(pa_next)
22     pb.append(pb_next)
```

目录

EM 简述

两枚硬币正面概率估算

多项分布参数的 EM 算法

多项式分布的特点

正态分布参数 EM 估计

二项泊松混合模型的 EM 估计

实际例子



Department of Statistics @BNU

多项分布参数的 EM 算法

假设 $x = (x_1, \dots, x_m)$ 服从多项分布，也就是

$$p(x|p_1, \dots, p_m) = \frac{n!}{x_1! \dots x_m!} p_1^{x_1} \dots p_m^{x_m}.$$

令 $m = 4$, $(p_1, \dots, p_4) = (0.5 - \theta/2, \theta/4, \theta/4, 0.5)$. 观测数据的数据为 $x = (x_1, x_2, x_3, x_4)^T$, 这里 $x_{34} = x_3 + x_4$, 感兴趣参数 θ 的估计。

潜变量 x_3, x_4 没有观测到，基于完整数据 $y = (x, z)$, 即 $(x_1, x_2, x_3, x_4)^T$, 对数似然函数为

$$\begin{aligned} \ell(\theta | y) &= \log L(\theta | y) \\ &= x_1 \log\left(\frac{1}{2} - \frac{\theta}{2}\right) + x_2 \log\left(\frac{\theta}{4}\right) + x_3 \log\left(\frac{\theta}{4}\right) + \text{const.} \end{aligned}$$

(黑板推导)

基于观测到的数据和上次迭代估计 $\theta^{(i-1)}$, 计算对数似然的条件期望,

$$\begin{aligned} E(\ell(\theta | y) | x; \theta^{(i-1)}) &= E(\log L(\theta | y) | x, \theta^{(i-1)}) \\ &= x_1 \log\left(\frac{1}{2} - \frac{\theta}{2}\right) + x_2 \log\left(\frac{\theta}{4}\right) + E(x_3 | x, \theta^{(i-1)}) \log\left(\frac{\theta}{4}\right) + \text{const.} \end{aligned}$$

下面计算 $E(x_3 | x, \theta^{(i-1)})$, 由于

$$E(x_3 | x; \theta^{(i-1)}) = \sum_{z=0}^{x_{34}} z p(z | x; \theta^{(i-1)}) = \sum_{z=0}^{x_{34}} z \frac{p(x, z; \theta^{(i-1)})}{p(x; \theta^{(i-1)})},$$

对于分子, 我们有

$$\begin{aligned} p(x, z; \theta^{(i-1)}) &= p(x_1, x_2, z, x_{34} - z; \theta^{(i-1)}) \\ &= \frac{n!}{x_1! x_2! z! (x_{34} - z)!} \left(\frac{1}{2} - \frac{\theta^{(i-1)}}{2}\right)^{x_1} \left(\frac{\theta^{(i-1)}}{4}\right)^{x_2} \left(\frac{\theta^{(i-1)}}{4}\right)^z \left(\frac{1}{2}\right)^{x_{34} - z}; \end{aligned}$$

对于分母，我们利用了多项式分布中，其中两项合并也是多项式分布的性质，则有

$$\begin{aligned} p(x; \theta^{(i-1)}) &= p(x_1, x_2, x_{34}; \theta^{(i-1)}) \\ &= \frac{n!}{x_1! x_2! (x_{34})!} \left(\frac{1}{2} - \frac{\theta^{(i-1)}}{2}\right)^{x_1} \left(\frac{\theta^{(i-1)}}{4}\right)^{x_2} \left(\frac{\theta^{(i-1)}}{4} + \frac{1}{2}\right)^{x_{34}}. \end{aligned}$$

基于 $p(x, z; \theta^{(i-1)})$ 和 $p(x; \theta^{(i-1)})$ 的表达式，可得

$$\frac{p(x, z; \theta^{(i-1)})}{p(x; \theta^{(i-1)})} = \frac{(x_{34})!}{z! (x_{34} - z)!} \left(\frac{\theta^{(i-1)}}{4}\right)^z \left(\frac{1}{2}\right)^{x_{34}-z} / \left(\frac{\theta^{(i-1)}}{4} + \frac{1}{2}\right)^{x_{34}},$$

上式说明 $z|x, \theta^{(i-1)}$ 服从二项分布，比对可知参数。

Liu, Ran - Department of Statistics @BNU

对一个二项分布，我们有：

$$P(X = k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

把 k 的合在一起再跟原式对齐

$$\frac{n!}{k!(n-k)!} \left(\frac{p}{1-p}\right)^k (1-p)^n = \frac{(x_{34})!}{z!(x_{34}-z)!} \left(\frac{\theta^{(i-1)}}{2}\right)^z \left(\frac{1}{2}\right)^{x_{34}} / \left(\frac{\theta^{(i-1)}}{4} + \frac{1}{2}\right)^{x_{34}}$$

只关注与 z 相关的部分，看作是一个以 z 为参数， $n = x_{34}$ 的二项式分布，我们令参数对应相等：

$$\frac{\theta^{(i-1)}}{2} = \frac{p}{1-p},$$

Liu, Ran - Department of Statistics @BNU

则我们有

$$p = \frac{\theta^{(i-1)}}{2 + \theta^{(i-1)}}, \quad 1 - p = \frac{2}{2 + \theta^{(i-1)}}.$$

所以我们有

$$E(x_3|x; \theta^{(i-1)}) = np = x_3 \frac{\theta^{(i-1)}}{2 + \theta^{(i-1)}}$$

值得一提的是，除了上面一步一步推导证明 $z|x, \theta^{(i-1)}$ 服从二项分布以外，也可以直接这样看待： x_3 在 x 已知的条件下，肯定为二项式分布，此时属于第三类别和第四类别的概率成比例于联合概率中二者的概率，也就是 $\theta/4$ 和 0.5 ，这个比例除以他们之和，即为条件概率。

Liu, Ran - Department of Statistics @BNU

关于对数似然的条件期望进一步写为

$$\begin{aligned} Q(\theta \mid \theta^{(i-1)}) &= E(\ell(\theta \mid y) \mid x; \theta^{(i-1)}) \\ &= x_1 \log\left(\frac{1}{2} - \frac{\theta}{2}\right) + x_2 \log\left(\frac{\theta}{4}\right) + (x_3 + x_4) \frac{\theta^{(i-1)}}{\theta^{(i-1)} + 2} \log\left(\frac{\theta}{4}\right) + \text{const.} \end{aligned}$$

LIU, Ran - Department of Statistics @BNU

关于 θ 极大化 $Q(\theta \mid \theta^{(i-1)})$ 得到下次迭代 $\theta^{(i)}$, 对 $Q(\theta \mid \theta^{(i-1)})$ 关于 θ 求导可得

$$\begin{aligned} Q'(\theta \mid \theta^{(i-1)}) &= -\frac{x_1}{1-\theta} + \frac{x_2}{\theta} + \frac{E(x_3 \mid x; \theta^{(i-1)})}{\theta} \\ &= -\frac{x_1}{1-\theta} + \frac{x_2}{\theta} + \frac{x_3}{\theta} \times \frac{\theta^{(i-1)}}{\theta^{(i-1)} + 2} = 0. \end{aligned}$$

因此

$$\theta^{(i)} = \frac{x_2 + E(x_3 \mid x; \theta^{(i-1)})}{x_1 + x_2 + E(x_3 \mid x; \theta^{(i-1)})}.$$

通过不断迭代, 最终得到 θ 的估计.

Liu, Ran - Department of Statistics @BNU

如果数据完全观测到，也就是 x_3 被观测，则参数 θ 的极大似然估计为

$$\hat{\theta} = \frac{x_2 + x_3}{x_1 + x_2 + x_3}.$$

对比完全数据的极大似然估计 $\hat{\theta}$ 和 EM 算法第 i 次迭代估计 $\theta^{(i)}$, EM 算法本质上相当于把基于完整数据得到的极大似然估计量中没有观测到的数据，采用观测到数据和上次迭代估计 $\theta^{(i-1)}$ 预测。

LIU, Ran - Department of Statistics @BNU

如果观测到的数据为 $x_{obs} = (x_1, x_2, x_{34})^T = (38, 34, 125)^T$ ，求参数的 EM 估计。

#观测数据

```
x1 <- 38
```

```
x2 <- 34
```

```
x34 <- 125
```

```
#初值与最大循环数
```

```
max.iter <- 100
```

```
htheta <- rep(0, max.iter); htheta[1] <- 0
```

```
#迭代
```

```
for (i in 1:max.iter){
```

```
hx3 <- 0.25*x34*htheta[i]/(0.25*htheta[i] + 0.5)
```

```
htheta[i+1] <- (x2 + hx3)/(x1 + x2 + hx3)
```

```
if (abs(htheta[i+1] - htheta[i]) < 1e-8)
```

```
  break
```

```
}
```

Liu, Ran - Department of Statistics @BNU

```
1 # 观测数据
  x1 = 38
3 x2 = 34
  x34 = 125
5
  # 初值与最大循环数
7 max_iter = 100
  htheta = [0] * max_iter
9 htheta[0] = 0
11
  # 迭代
  for i in range(max_iter):
13     hx3 = 0.25 * x34 * htheta[i] / (0.25 * htheta[i] + 0.5)
     htheta[i+1] = (x2 + hx3) / (x1 + x2 + hx3)
15
     if abs(htheta[i+1] - htheta[i]) < 1e-8:
17         break
```

Liu, Ran - Department of Statistics @BNU

目录

EM 简述

两枚硬币正面概率估算

多项分布参数的 EM 算法

多项式分布的特点

正态分布参数 EM 估计

二项泊松混合模型的 EM 估计

实际例子



Department of Statistics @BNU

多项式分布的特点

- 多项式分布其中两个类别合在一起，新分布也是多项式分布。
- 多项式分布的其中两个类别，他们加起来等于一个常数的话，那么它在这个他们两个在这个常数下的条件概率分布成比例于他们在原有的多项式分布中各自的概率，然后他们在这个二项式分布真实的概率为它们在联合概率中的这个概率除以它们的求和。

Liu, Ran - Department of Statistics @BNU

多项式分布其中两个类别合在一起，新分布也是多项式分布。

假设 $x = (x_1, \dots, x_m)$ 服从多项分布，也就是

$$p(x|p_1, \dots, p_m) = \frac{n!}{x_1! \dots x_m!} p_1^{x_1} \dots p_m^{x_m}.$$

那么 $x^* = (x_1 + x_2, x_3, \dots, x_m)$ 也服从多项式分布：

$$p(x^*|p_1, \dots, p_m) = \frac{n!}{(x_1 + x_2)! \dots x_m!} (p_1 + p_2)^{x_1 + x_2} \dots p_m^{x_m}.$$

LIU, Ran - Department of Statistics @BNU

证明:

$$\begin{aligned}
 p(x_1 + x_2 = y, x_3, \dots, x_m | \sim) &= \sum_{x_2=0}^y p(x_1 = y - x_2, x_2 = x_2, \dots | \sim) \\
 &= \sum_{x_2=0}^y \frac{n!}{(y - x_2)! x_2! \dots x_m!} p_1^{y-x_2} p_2^{x_2} \dots \\
 &= \frac{n!}{y! x_3! \dots x_m!} (p_1 + p_2)^y p_3^{x_3} \dots
 \end{aligned}$$

其中我们利用了二项式公式:

$$(p_1 + p_2)^y = \sum_{k=0}^y \binom{y}{k} p_1^{y-k} p_2^k.$$

LIU, Ran - Department of Statistics @BNU

多项式分布的其中两个类别，他们加起来等于一个常数的话，那么它在这个他们两个在这个常数下的条件概率分布成比例于他们在原有的多项式分布中各自的概率，然后他们在这个二项式分布真实的概率为它们在联合概率中的这个概率除以它们的求和。

假设 $x = (x_1, \dots, x_m)$ 服从多项分布，也就是

$$p(x|p_1, \dots, p_m) = \frac{n!}{x_1! \dots x_m!} p_1^{x_1} \dots p_m^{x_m}.$$

那么 $x_1, x_2 | x_1 + x_2 = y, x_3, \dots, x_m$ 也服从多项式分布：

$$p(x_1, x_2 | x_1 + x_2 = y, x_3, \dots, x_m, p) = C_y^{x_1} \left(\frac{p_1}{p_1 + p_2} \right)^{x_1} \left(\frac{p_2}{p_1 + p_2} \right)^{y-x_1},$$

$$x_1 = 0, 1, \dots, y.$$

Liu, Ran - Department of Statistics @BNU

证明：

$$\begin{aligned}
 p(x_1, x_2 | x_1 + x_2 = y, x_3, \dots, x_m, p) &= \frac{p(x_1 = z, x_2 = y - z, x_3, \dots, x_m)}{p(x_1 + x_2 = y, x_3, \dots, x_m)} \\
 &= \frac{\frac{n!}{z!(y-z)!x_3!\dots x_m!} p_1^{z_1} p_2^{y-z} p_3^{x_3} \dots}{\frac{n!}{y!x_3!\dots x_m!} (p_1 + p_2)^y p_3^{x_3} \dots} \\
 &= \frac{y!}{z!(y-z)!} \left(\frac{p_1}{p_1 + p_2}\right)^{z_1} \left(\frac{p_2}{p_1 + p_2}\right)^{z_2}
 \end{aligned}$$

由证明过程可知其实在未知 $x_3, \dots, x_m$ 的情况下, $x_1, x_2 | x_1 + x_2 = y$ 也服从多项式分布。

LIU, Ran - Department of Statistics @BNU

若没有 $x_1 + x_2 = y$, 那么 x_1, x_2 不服从多项式分布。简单来说, 他俩的和可以为 0 到 n 的所有值。他们的概率计算应为:

$$p(x_1, x_2) = \sum_{y=0}^n p(x_1, x_2 | x_1 + x_2 = y) p(x_1 + x_2 = y)$$

更多有趣的性质, 可查阅:

<https://online.stat.psu.edu/stat504/book/export/html/667>

LIU, Ran - Department of Statistics @BNU

目录

EM 简述

两枚硬币正面概率估算

多项分布参数的 EM 算法

多项式分布的特点

正态分布参数 EM 估计

二项泊松混合模型的 EM 估计

实际例子



Department of Statistics @BNU

正态分布参数 EM 估计

对来自正态总体 $N(\mu, \sigma^2)$ 的完整数据 $y = (x_1, x_2, \dots, x_n)^\tau$, 假设观测到数据 $x = (x_1, \dots, x_m)$, 隐含数据为 $z = (x_{m+1}, \dots, x_n)$. 则基于完整数据 y 的极大似然函数为

$$L(y; \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\sum_{j=1}^n \frac{(x_j - \mu)^2}{2\sigma^2}\right\}.$$

相应的对数极大似然函数为

$$\begin{aligned} l(y; \mu, \sigma^2) &= \log L(y; \mu, \sigma^2) \\ &= -\frac{n}{2} \log \sigma^2 - \frac{n}{2} \log 2\pi - \frac{1}{2\sigma^2} \sum_{j=1}^n x_j^2 + \frac{\mu}{\sigma^2} \sum_{j=1}^n x_j - \frac{n\mu^2}{2\sigma^2}. \end{aligned}$$

这个例子比较简单, 隐含数据和观测数据在参数已知下独立, 互不干扰。

基于观测数据的条件对数似然为

$$\begin{aligned} Q(\mu, \sigma^2; \mu^{(i-1)}, \sigma^{2(i-1)}) &= E(l(y; \mu, \sigma^2) | x; \mu^{(i-1)}, \sigma^{2(i-1)}) \\ &= -\frac{n}{2} \log \sigma^2 - \frac{n}{2} \log 2\pi - \frac{1}{2\sigma^2} \sum_{j=1}^m x_j^2 + \frac{\mu}{\sigma^2} \sum_{j=1}^m x_j - \frac{n\mu^2}{2\sigma^2} \\ &\quad - \frac{1}{2\sigma^2} E\left\{ \sum_{j=m+1}^n x_j^2 | x; \mu^{(i-1)}, \sigma^{2(i-1)} \right\} \\ &\quad + \frac{\mu}{\sigma^2} E\left\{ \sum_{j=m+1}^n x_j | x; \mu^{(i-1)}, \sigma^{2(i-1)} \right\}. \end{aligned}$$

LIU, Ran - Department of Statistics @BNU

对 $Q(\mu, \sigma^2; \mu^{(i-1)}, \sigma^{2(i-1)})$ 关于 μ, σ^2 求导可得

$$\begin{aligned}
 & \frac{\partial Q(\mu, \sigma^2; \mu^{(i-1)}, \sigma^{2(i-1)})}{\partial \mu} \\
 &= \frac{1}{\sigma^2} \sum_{j=1}^n x_j + \frac{1}{\sigma^2} E\left\{ \sum_{j=m+1}^n x_j | x; \mu^{(i-1)}, \sigma^{2(i-1)} \right\} - \frac{n\mu}{\sigma^2} = 0 \\
 & \frac{\partial Q(\mu, \sigma^2; \mu^{(i-1)}, \sigma^{2(i-1)})}{\partial \sigma^2} \\
 &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{j=1}^n x_j^2 - \frac{\mu}{\sigma^4} \sum_{j=1}^n x_j + \frac{n\mu^2}{2\sigma^4} \\
 & \quad + \frac{1}{2\sigma^4} E\left\{ \sum_{j=m+1}^n x_j^2 | x; \mu^{(i-1)}, \sigma^{2(i-1)} \right\} - \frac{\mu}{\sigma^4} E\left\{ \sum_{j=m+1}^n x_j | x; \mu^{(i-1)}, \sigma^{2(i-1)} \right\} \\
 &= 0.
 \end{aligned}$$

LIU, Ran - Department of Statistics @BNU

所以第 i 次迭代估计 $\theta^{(i)}$ 为

$$\begin{aligned}\mu^{(i)} &= \frac{1}{n} \left\{ \sum_{j=1}^m x_j + E \left(\sum_{j=m+1}^n x_j | x; \mu^{(i-1)}, \sigma^{2(i-1)} \right) \right\} \\ \sigma^{2(i)} &= \frac{1}{n} \left\{ \sum_{j=1}^m x_j^2 + E \left(\sum_{j=m+1}^n x_j^2 | x; \mu^{(i-1)}, \sigma^{2(i-1)} \right) \right\} - (\mu^{(i)})^2.\end{aligned}$$

给定第 $i-1$ 步的 $(\mu^{(i-1)}, \sigma^{2(i-1)})$, 且未观测到的隐含数据和观测到的数据独立, 则

$$\begin{aligned}E \left(\sum_{j=m+1}^n x_j | x; \mu^{(i-1)}, \sigma^{2(i-1)} \right) &= (n-m) \hat{\mu}^{(i-1)}, \\ E \left(\sum_{j=m+1}^n x_j^2 | x; \mu^{(i-1)}, \sigma^{2(i-1)} \right) &= (n-m) (\hat{\mu}^{(i-1)2} + \sigma^{2(i-1)}).\end{aligned}$$

若未观测的隐含数据和观测数据相关, 则根据变量相关性计算 $E(\sum_{j=m+1}^n x_j | x; \mu^{(i-1)}, \sigma^{2(i-1)})$ 和 $E(\sum_{j=m+1}^n x_j^2 | x; \mu^{(i-1)}, \sigma^{2(i-1)})$.

设 $y = (x_1, x_2, \dots, x_{1000})^\tau$ 来自于正态总体 $N(2, 4)$, 其中观测到的数据为 $x = (x_1, \dots, x_{600})$, 未观测到的数据为 $z = (x_{601}, \dots, x_{1000})$, 通过 EM 算法给出参数 μ, σ^2 的估计。

LIU, Ran - Department of Statistics @BNU

```
set.seed(1) #生成数据
n <- 1000;m <- 600
x <- rnorm(n, mean = 2, sd = 2)
sx <- sum(x[1:m]); sx2 <- sum(x[1:m]^2)
max.iter <- 100 #最大循环数
hmu <- rep(0, max.iter)
hsigma2 <- rep(0, max.iter)
hmu[1] <- 0;hsigma2[1] <- 1 #初值
for (i in 1:max.iter){
  s1 <- sx + (n - m)*hmu[i]
  s2 <- sx2 + (n - m)*(hmu[i]^2 + hsigma2[i])
  hmu[i+1] <- s1/n
  hsigma2[i+1] <- s2/n - hmu[i+1]^2
  if(abs(hmu[i+1] - hmu[i]) < 1e-8
    & abs(hsigma2[i+1] - hsigma2[i]) < 1e-8) break
}
```

Liu, Ran - Department of Statistics @BNU

```
1 import numpy as np
  np.random.seed(1)  # 生成数据
3 n = 1000
  m = 600
5 x = np.random.normal(loc=2, scale=2, size=n)
  sx = np.sum(x[:m])
7 sx2 = np.sum(x[:m]**2)
  max_iter = 100  # 最大循环数
9 hmu = np.zeros(max_iter)
  hsigma2 = np.zeros(max_iter)
11 hmu[0] = 0
  hsigma2[0] = 1  # 初值
13
  for i in range(max_iter):
15     s1 = sx + (n - m) * hmu[i]
     s2 = sx2 + (n - m) * (hmu[i]**2 + hsigma2[i])
17     hmu[i+1] = s1 / n
     hsigma2[i+1] = s2 / n - hmu[i+1]**2
19     if abs(hmu[i+1] - hmu[i]) < 1e-8 and abs(hsigma2[i+1] -
        hsigma2[i]) < 1e-8:
        break
```

目录

EM 简述

两枚硬币正面概率估算

多项分布参数的 EM 算法

多项式分布的特点

正态分布参数 EM 估计

二项泊松混合模型的 EM 估计

实际例子



Department of Statistics @BNU

二项泊松混合模型的 EM 估计

我们观测 n 个人的 B 站用户等级，一共 7 个等级，等级为 $i (i = 0, 1, \dots, 6)$ 的人数是 n_i ，则 $n = \sum_{i=0}^6 n_i$ 。观测数据如下表所示

用户等级	0	1	2	3	4	5	6
人数	n_0	n_1	n_2	n_3	n_4	n_5	n_6

假定已注册的人的 B 站等级服从参数为 λ 的泊松分布，感兴趣的参数是参数 λ 和没有注册的概率 ξ 。这里例子中，设定了刚注册的和没有注册的用户等级都为 0。

零膨胀模型 (Zero-Inflated Model) 常用于处理具有过度零值的数据，其中数据集中包含大量的零值观测。

令 n_A 表示没有注册的人数, 则 $n_B = n_0 - n_A$ 表示刚注册等级为 0 的用户, n_A 和 n_B 为未观测到的隐变量. 观测数据为 $x = (n_0, n_1, \dots, n_6)$.

以下根据 EM 算法估计 (ξ, λ) , 基于完全数据 $y = (x, z)$ 的似然函数为

$$\begin{aligned} L(y; \xi, \lambda) \\ = \frac{(n_A + n_B + n_1 + \dots + n_6)!}{n_A! n_B! n_1! \dots n_6!} \xi^{n_A} [e^{-\lambda} (1 - \xi)]^{n_B} \prod_{k=1}^6 \left[\frac{\lambda^k e^{-\lambda}}{k!} (1 - \xi) \right]^{n_k}. \end{aligned}$$

相当于先随机分组, 再乘以每个人出现在这个组别的概率 (多项式分布)。

这里其实有一个问题是, 看作 8 个类别的多项式分布时, 其中概率为泊松分布概率的话, 总体概率求和不等于 1。但我们可以假设是无穷类别的多项式分布, 与此同时有更多的观测变量, $n_7, \dots, n_\infty = 0$, 这样代入式子中没有任何影响。

对数似然函数为

$$l(y; \xi, \lambda) = n_A \log \xi - n_B \lambda + n_B \log(1 - \xi) + \sum_{k=1}^6 n_k [-\lambda + k \log \lambda + \log(1 - \xi)] + cost.$$

基于观测数据 x 和上次迭代估计 $\xi^{(i-1)}, \lambda^{(i-1)}$, 对数似然的条件期望为

$$\begin{aligned} Q(\xi, \lambda; \xi^{(i-1)}, \lambda^{(i-1)}) &= E(l(y; \xi, \lambda) | x; \xi^{(i-1)}, \lambda^{(i-1)}) \\ &= E(n_A | x; \xi^{(i-1)}, \lambda^{(i-1)}) \log \xi - E(n_B | x; \xi^{(i-1)}, \lambda^{(i-1)}) \lambda \\ &\quad + E(n_B | x; \xi^{(i-1)}, \lambda^{(i-1)}) \log(1 - \xi) \\ &\quad + \sum_{k=1}^6 n_k [-\lambda + k \log \lambda + \log(1 - \xi)] + cost. \end{aligned}$$

LIU, Ran - Department of Statistics @BNU

(黑板推导)

对 $Q(\xi, \lambda; \xi^{(i-1)}, \lambda^{(i-1)})$ 关于 ξ, λ 求导, 可得

$$\begin{aligned}
 & \frac{\partial Q(\xi, \lambda; \xi^{(i-1)}, \lambda^{(i-1)})}{\partial \xi} \\
 = & \frac{E(n_A|x; \xi^{(i-1)}, \lambda^{(i-1)})}{\xi} - \frac{E(n_B|x; \xi^{(i-1)}, \lambda^{(i-1)}) + n_1 + \cdots + n_6}{1 - \xi} = 0 \\
 & \frac{\partial Q(\xi, \lambda; \xi^{(i-1)}, \lambda^{(i-1)})}{\partial \lambda} \\
 = & -\{E(n_B|x; \xi^{(i-1)}, \lambda^{(i-1)}) + n_1 + \cdots + n_6\} + \frac{1}{\lambda} \sum_{k=1}^6 kn_k = 0.
 \end{aligned}$$

Liu, Ran - Department of Statistics @BNU

根据上式不难得出第 i 次迭代结果

$$\xi^{(i)} = \frac{E(n_A|x; \xi^{(i-1)}, \lambda^{(i-1)})}{n}, \quad \lambda^{(i)} = \frac{\sum_{k=1}^6 kn_k}{n - E(n_A|x; \xi^{(i-1)}, \lambda^{(i-1)})}.$$

对于上式 $E(n_A|x; \xi^{(i-1)}, \lambda^{(i-1)})$, 通过推导 (n_0 已知, 第一和第二类别的比例已知, 求和为 1), 不难进一步化简得

$$E(n_A|x; \xi^{(i-1)}, \lambda^{(i-1)}) = \frac{n_0 \xi^{(i-1)}}{\xi^{(i-1)} + (1 - \xi^{(i-1)})e^{-\lambda^{(i-1)}}}.$$

按照上述算法反复迭代, 得到参数 ξ, λ 的 EM 估计。

Liu, Ran - Department of Statistics @BNU

目录

EM 简述

两枚硬币正面概率估算

多项分布参数的 EM 算法

多项式分布的特点

正态分布参数 EM 估计

二项泊松混合模型的 EM 估计

实际例子



Li Q, Han Department of Statistics @BNU

实际例子

假设我们有一个大型的商圈，每天每家店各个时间段的营业额数据，如何刻画客户群体？

一天得到的数据是：

- ① 小明奶茶店：..., 11: 30 进账 500 元, 11: 47 进账 1000 元, 12: 20 进账 800 元...;
- ② 小芳电影院：..., 17: 30 进账 300 元, 19: 30 进账 700 元, 20: 45 进账 900 元...;
- ③ 小红超市：..., 12: 30 进账 1000 元, 13: 20 进账 1270 元, 14: 50 进账 1150 元...

首先要做的就是数据清洗，将他们合并成同一时间分割。

在任意时间点，进商场人群的数量（未知或部分未知）服从泊松分布：

学生 $X_{1k} \sim Poi(\lambda_{1k})$;

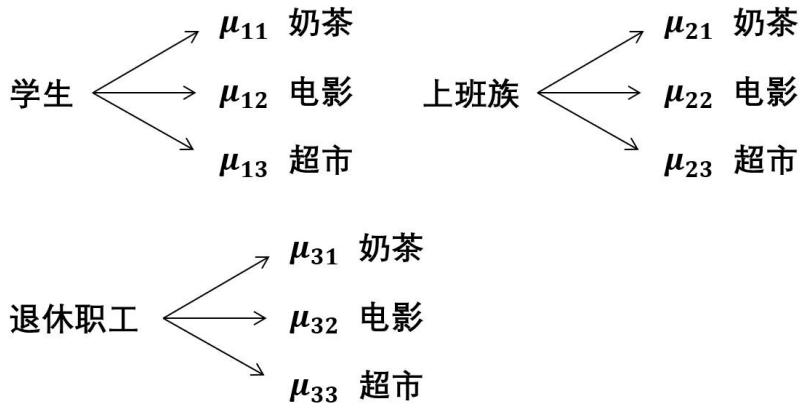
上班族 $X_{2k} \sim Poi(\lambda_{2k})$;

退休职工 $X_{3k} \sim Poi(\lambda_{3k})$;

其中 k 是时间，从早上 10 点到晚上 10 点，一共 12 小时，即 $k = 1, 2, \dots, 12$ 。

Liu, Ran - Department of Statistics @BNU

每种人群针对不同的店有不同的消费系数（人均费用）



Statistical Computing - Department of Statistics, HKUST

第 k 时间里, 奶茶店的销售额为 y_{1k} , 影城的销售额为 y_{2k} , 超市的销售额为 y_{3k} , 则我们有:

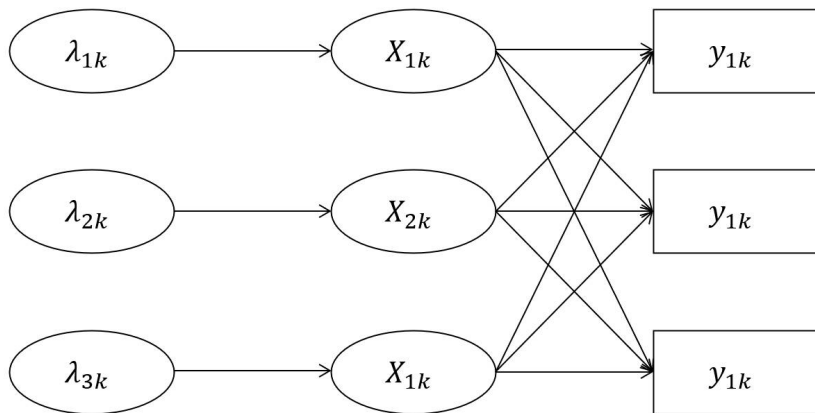
$$y_{1k} = X_{1k}\mu_{11} + X_{2k}\mu_{21} + X_{3k}\mu_{31} + \varepsilon_{1k}, \quad \varepsilon_{1k} \sim N(0, \sigma_1^2)$$

$$y_{2k} = X_{1k}\mu_{12} + X_{2k}\mu_{22} + X_{3k}\mu_{32} + \varepsilon_{2k}, \quad \varepsilon_{2k} \sim N(0, \sigma_2^2)$$

$$y_{3k} = X_{1k}\mu_{13} + X_{2k}\mu_{23} + X_{3k}\mu_{33} + \varepsilon_{3k}, \quad \varepsilon_{3k} \sim N(0, \sigma_3^2)$$

LIU, Ran - Department of Statistics @BNU

简易关系图：



Liu, Yan - Department of Statistics (2021)

我们先假设 Y 已知, X 已知。记 $\theta = (\sigma, \mu, \lambda)$, 我们有(黑板推导)

$$\begin{aligned}
 L(\theta | X, Y) &= \prod_{k=1}^{12} p(Y_k | X_k, \theta) \cdot p(X_k | \theta) \\
 &= \prod_{k=1}^{12} p(y_{1k} | X_k, \theta) \cdot p(y_{2k} | X_k, \theta) \cdot p(y_{3k} | X_k, \theta) \\
 &\quad \cdot p(X_{1k} | \theta) \cdot p(X_{2k} | \theta) \cdot p(X_{3k} | \theta) \\
 &= \prod_{k=1}^{12} \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{(y_{1k} - (X_{1k}\mu_{11} + X_{2k}\mu_{21} + X_{3k}\mu_{31}))^2}{2\sigma_1^2}\right) \\
 &\quad \cdot \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{(y_{2k} - (X_{1k}\mu_{12} + X_{2k}\mu_{22} + X_{3k}\mu_{32}))^2}{2\sigma_2^2}\right) \\
 &\quad \cdot \frac{1}{\sqrt{2\pi}\sigma_3} \exp\left(-\frac{(y_{3k} - (X_{1k}\mu_{13} + X_{2k}\mu_{23} + X_{3k}\mu_{33}))^2}{2\sigma_3^2}\right) \\
 &\quad \cdot \frac{e^{-\lambda_{1k}} \lambda_{1k}^{X_{1k}}}{X_{1k}!} \cdot \frac{e^{-\lambda_{2k}} \lambda_{2k}^{X_{2k}}}{X_{2k}!} \cdot \frac{e^{-\lambda_{3k}} \lambda_{3k}^{X_{3k}}}{X_{3k}!}
 \end{aligned}$$

取对数，可以进一步得到：

$$\begin{aligned}\ell(\theta|X, Y) \propto & \sum_{k=1}^{12} \left(-\ln \sigma_1 - \frac{(y_{1k} - (X_{1k}\mu_{11} + X_{2k}\mu_{21} + X_{3k}\mu_{31}))^2}{2\sigma_1^2} \right. \\ & - \ln \sigma_2 - \frac{(y_{2k} - (X_{1k}\mu_{12} + X_{2k}\mu_{22} + X_{3k}\mu_{32}))^2}{2\sigma_2^2} \\ & - \ln \sigma_3 - \frac{(y_{3k} - (X_{1k}\mu_{13} + X_{2k}\mu_{23} + X_{3k}\mu_{33}))^2}{2\sigma_3^2} \\ & - \lambda_{1k} + X_{1k} \ln \lambda_{1k} \\ & - \lambda_{2k} + X_{2k} \ln \lambda_{2k} \\ & \left. - \lambda_{3k} + X_{3k} \ln \lambda_{3k} \right)\end{aligned}$$

LIU, Ran - Department of Statistics @BNU

$$\frac{\partial \ell(\theta|X, Y)}{\partial \lambda_{1k}} = -1 + \frac{X_{1k}}{\lambda_{1k}} = 0 \quad \Rightarrow \quad \hat{\lambda}_{1k}^{(t)} = X_{1k}$$

$$\frac{\partial \ell(\theta|X, Y)}{\partial \lambda_{2k}} = -1 + \frac{X_{2k}}{\lambda_{2k}} = 0 \quad \Rightarrow \quad \hat{\lambda}_{2k}^{(t)} = X_{2k}$$

$$\frac{\partial \ell(\theta|X, Y)}{\partial \lambda_{3k}} = -1 + \frac{X_{3k}}{\lambda_{3k}} = 0 \quad \Rightarrow \quad \hat{\lambda}_{3k}^{(t)} = X_{3k}$$

$$\frac{\partial \ell(\theta|X, Y)}{\partial \mu_{11}} = \sum_{k=1}^{12} \frac{(y_{1k} - (X_{1k}\mu_{11} + X_{2k}\mu_{21} + X_{3k}\mu_{31})) X_{1k}}{\sigma_1^2} = 0$$

$$\Rightarrow 0 = \sum_{k=1}^{12} [X_{1k}(y_{1k} - X_{2k}\mu_{21} - X_{3k}\mu_{31}) - X_{1k}^2\mu_{11}]$$

$$\Rightarrow \hat{\mu}_{11}^{(t)} = \frac{\sum_{k=1}^{12} X_{1k} (y_{1k} - X_{2k}\mu_{21}^{(t-1)} - X_{3k}\mu_{31}^{(t-1)})}{\sum_{k=1}^{12} X_{1k}^2}$$

使用 ECM(coordinate ascent), 按照同样的推导可得:

$$\hat{\mu}_{21}^{(t)} = \frac{\sum_{k=1}^{12} X_{2k} \left(y_{1k} - X_{1k} \mu_{11}^{(t)} - X_{3k} \mu_{31}^{(t-1)} \right)}{\sum_{k=1}^{12} X_{2k}^2}$$

$$\hat{\mu}_{31}^{(t)} = \frac{\sum_{k=1}^{12} X_{3k} \left(y_{1k} - X_{1k} \mu_{11}^{(t)} - X_{2k} \mu_{21}^{(t)} \right)}{\sum_{k=1}^{12} X_{3k}^2}$$

那么 $\hat{\mu}_{22}^{(t)}$, $\hat{\mu}_{23}^{(t)}$, $\hat{\mu}_{32}^{(t)}$, $\hat{\mu}_{33}^{(t)}$ 同理可得。

LIU, Ran - Department of Statistics @BNU

若 $\{X_{ik}\}, i = 1, 2, 3; \quad k = 1, 2, \dots, 12$ 未知或者部分未知呢？

以 X_{1k} 相关的为例

$$\mathbb{E} \left(X_{1k} \mid Y, \theta^{(t)} \right) = \quad (a)$$

$$\mathbb{E} \left(X_{1k} X_{2k} \mid Y, \theta^{(t)} \right) = \quad (b)$$

$$\mathbb{E} \left(X_{1k}^2 \mid Y, \theta^{(t)} \right) = \quad (c)$$

LIU, Ran - Department of Statistics @BNU

我们来看看 (a)

$$\begin{aligned} p(X_{1k} | Y, \theta^{(t)}) &\propto P(Y, X_{1k} | \theta^{(t)}) \\ &\propto \sum_{X_{2k}, X_{3k}} p(Y_k, X_{1k}, X_{2k}, X_{3k} | \theta^{(t)}) \\ &= \sum_{X_{2k}, X_{3k}} p(Y_k | X_{1k}, X_{2k}, X_{3k}, \theta^{(t)}) p(X_{1k}, X_{2k}, X_{3k} | \theta^{(t)}) \\ &= \sum_{X_{2k}, X_{3k}} p(y_{1k} | X_{1k}, X_{2k}, X_{3k}, \theta^{(t)}) \\ &\quad p(y_{2k} | X_{1k}, X_{2k}, X_{3k}, \theta^{(t)}) \\ &\quad p(y_{3k} | X_{1k}, X_{2k}, X_{3k}, \theta^{(t)}) \\ &\quad p(X_{1k} | \theta^{(t)}) p(X_{2k} | \theta^{(t)}) p(X_{3k} | \theta^{(t)}) \end{aligned}$$

会发现根本没法算，所以这 E 步用 MC 来计算，即 MCEM。

但是此时隐变量的后验分布并不是简单就能采样的分布，所以我们会采用 MH 采样得到 $X_{1k}^{(m)}, X_{2k}^{(m)}, X_{3k}^{(m)}$ 。此时的目标分布为

$$p(X_{1k}, X_{2k}, X_{3k} \mid Y, \theta^{(t)}) \propto P(X_{1k}, X_{2k}, X_{3k}, Y \mid \theta^{(t)})$$

提案分布可为: $\text{Poi}(\lambda_{1k}^{(t)}), \text{Poi}(\lambda_{2k}^{(t)}), \text{Poi}(\lambda_{3k}^{(t)})$

LIU, Ran - Department of Statistics @BNU

设已采样到: $\{X_{1k}^{(m)}, X_{2k}^{(m)}, X_{3k}^{(m)}\}$, $m = 1, 2, \dots, M$

$$\mathbb{E} \left(X_{1k} \mid Y, \theta^{(t)} \right) = \frac{1}{M} \sum_{m=1}^M X_{1k}^{(m)}$$

$$\mathbb{E} \left(X_{1k} X_{2k} \mid Y, \theta^{(t)} \right) = \frac{1}{M} \sum_{m=1}^M X_{1k}^{(m)} X_{2k}^{(m)}$$

$$\mathbb{E} \left(X_{1k}^2 \mid Y, \theta^{(t)} \right) = \frac{1}{M} \sum_{m=1}^M \left(X_{1k}^{(m)} \right)^2$$

代入 $\hat{\mu}$ 的估计中即可得。(也可直接用 $\lambda_{1k}^{(t)}, \lambda_{2k}^{(t)}, \lambda_{3k}^{(t)}$, 不过不太准确)。

- 若改成神经网络，按样 M 步就不能用 ECM 了，而是梯度下降 ($\hat{\mu}$ 不能直接求得)。E 步还是一样用 MH 采样，只是联合后验分布里面变成 $y - f(x)$ 了。
- 值得注意的是，若 X 全未知，可能会有可识别性的问题。

LIU, Ran - Department of Statistics @BNU