HW3 第三次作业解答

更新：2025 年 11 月 18 日

Exercise 1

流行病学家对研究有 HIV 感染风险的个体性行为感兴趣。假设调查了 1500 名男同性恋者并询问在过去的 30 天里每人有多少次危险性行为。令 n_i 表示回答有 i 次危险性行为的人数，这里 $i = 1, \cdots, 16$。表 1 列出了他们的回答。

表 1: 回答有相应次数危险性行为的人数

性行为数 i	0	1	2	3	4	5	6	7	8
人数 n_i	379	299	222	145	109	95	73	59	45
性行为数 i	9	10	11	12	13	14	15	16	
人数 n_i	30	24	12	4	2	0	1	1	

Poisson 模型拟合这些数据的效果很差。假设这些人可以分为三组更为实际。首先，有一组人，无论出于什么原因，回答了有 0 次危险行为，即使是不真实。假定个体属于这一组的概率为 α。

个体属于第二组的概率为 β，他们声称有典型的行为。这些人的回答是真实的，且假定他们进行危险行为的次数服从参数为 μ 的 Poisson 分布。

最后，个体属于高危险组的概率为 $1 - \alpha - \beta$。这些人回答是真实的，且他们进行危险行为的次数服从参数为 λ 的 Poisson 分布。

模型的参数为 α, β, μ 和 λ。在 EM 的第 t 次迭代中，我们用 $\boldsymbol{\theta}^{(t)} = (\alpha^{(t)}, \beta^{(t)}, \mu^{(t)}, \lambda^{(t)})$ 表示当前参数值。观测数据的似然为

$$L(\boldsymbol{\theta}|n_0, \cdots, n_{16}) \propto \prod_{i=0}^{16} \left[\frac{\pi_i(\boldsymbol{\theta})}{i!}\right]^{n_i}$$

其中对于 $i = 1, \cdots, 16$，定义

$$\pi_i(\boldsymbol{\theta}) = \alpha \cdot \mathbb{I}(i = 0) + \beta \cdot \mu^i e^{-\mu} + (1 - \alpha - \beta) \cdot \lambda^i e^{-\lambda}$$

观测到的数据为 $n_0,\cdots,n_{16}$。完全数据为 $n_{z,0},n_{t,0},\cdots,n_{t,16}$ 和 $n_{p,0},\cdots,n_{p,16}$，其中 $n_{k,i}$ 表示在第 k 组中回答有 i 次危险性行为的人数且 $k=z,t,p$ 分别表示 0 组、典型组和性乱交组。因而 $n_0=n_{z,0}+n_{t,0}+n_{p,0}$，且对于 $i=1,\cdots,16$，有 $n_i=n_{t,i}+n_{p,i}$。令 $N=\sum_{i=0}^{16}n_i=1500$。

对 $i=1,\cdots,16$，定义

$$
\begin{aligned}
z_0(\boldsymbol{\theta}) &= \frac{\alpha}{\pi_0(\boldsymbol{\theta})}\\
t_i(\boldsymbol{\theta}) &= \frac{\beta\mu^i e^{-\mu}}{\pi_i(\boldsymbol{\theta})}\\
p_i(\boldsymbol{\theta}) &= \frac{(1-\alpha-\beta)\lambda^i e^{-\lambda}}{\pi_i(\boldsymbol{\theta})}
\end{aligned}
$$

分别对应于有 i 次危险行为的人属于各组的概率。

a 说明 EM 算法可给出如下更新：

$$
\begin{aligned}
\alpha^{(t+1)} &= \frac{n_0 z_0(\boldsymbol{\theta}^{(t)})}{N}\\
\beta^{(t+1)} &= \sum_{i=0}^{16}\frac{n_i t_i(\boldsymbol{\theta}^{(t)})}{N}\\
\mu^{(t+1)} &= \frac{\sum_{i=0}^{16} i n_i t_i(\boldsymbol{\theta}^{(t)})}{\sum_{i=0}^{16} n_i t_i(\boldsymbol{\theta}^{(t)})}\\
\lambda^{(t+1)} &= \frac{\sum_{i=0}^{16} i n_i p_i(\boldsymbol{\theta}^{(t)})}{\sum_{i=0}^{16} n_i p_i(\boldsymbol{\theta}^{(t)})}
\end{aligned}
$$

b 由观测数据估计模型的参数。

c 用任一可行的方法估计所估参数的标准误和它们两两之间的相关系数。

解答 (a) 按 z,t,p 分三组，得到的完整数据似然函数为

$$
L(\boldsymbol{\theta}|n_0,\cdots,n_{16}) \propto \prod_{i=0}^{16} n_i \log\left[\frac{\pi_i(\boldsymbol{\theta})}{i!}\right] = \alpha^{n_{z,0}}\cdot\prod_{i=0}^{16}\left[\frac{\beta\mu^i e^{-\mu}}{i!}\right]^{n_{t,i}}\cdot\prod_{i=0}^{16}\left[\frac{(1-\alpha-\beta)\lambda^i e^{-\lambda}}{i!}\right]^{n_{p,i}}
$$

由于 $i!$ 与参数 $\boldsymbol{\theta}$ 无关，故省略。然后，取对数得到对数似然函数为

$$
\begin{aligned}
l(\boldsymbol{\theta}) &\propto n_{z,0}\log\alpha + \sum_{i=0}^{16} n_{t,i}\left(\log\beta + i\log\mu - \mu\right) + \sum_{i=0}^{16} n_{p,i}\left(\log(1-\alpha-\beta) + i\log\lambda - \lambda\right)\\
&= n_{z,0}\log\alpha + \left(\sum_{i=0}^{16} n_{t,i}\right)\log\beta + \left(\sum_{i=0}^{16} i n_{t,i}\right)\log\mu - \left(\sum_{i=0}^{16} n_{t,i}\right)\mu\\
&\quad + \left(\sum_{i=0}^{16} n_{p,i}\right)\log(1-\alpha-\beta) + \left(\sum_{i=0}^{16} i n_{p,i}\right)\log\lambda - \left(\sum_{i=0}^{16} n_{p,i}\right)\lambda
\end{aligned}
$$

对于 EM 算法的 E 步，计算 $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \mathbb{E}[l(\boldsymbol{\theta})|n_i,\boldsymbol{\theta}^{(t)}]$，其中 $i=0,\cdots,16$。注意到随机变量只有 $n_{z,i},n_{t,i},n_{p,i}$，且它们的条件期望分别为

$$
\begin{cases}
\mathbb{E}[n_{z,0}|n_1,\cdots,n_{16},\boldsymbol{\theta}^{(t)}] = n_0 z_0(\boldsymbol{\theta}^{(t)})\\
\mathbb{E}[n_{t,i}|n_1,\cdots,n_{16},\boldsymbol{\theta}^{(t)}] = n_i t_i(\boldsymbol{\theta}^{(t)})\\
\mathbb{E}[n_{p,i}|n_1,\cdots,n_{16},\boldsymbol{\theta}^{(t)}] = n_i p_i(\boldsymbol{\theta}^{(t)})
\end{cases}
$$

这是因为 $n_{z,0}$ 服从二项分布 $B(n_0, z_0(\boldsymbol{\theta}^{(t)}))$，而对于 $i = 1, \cdots, 16$，$n_{t,i}$ 和 $n_{p,i}$ 分别服从二项分布 $B(n_i, t_i(\boldsymbol{\theta}^{(t)}))$ 和 $B(n_i, p_i(\boldsymbol{\theta}^{(t)}))$。

于是，**E** 步得到 $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ 为

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = {} & n_0 z_0(\boldsymbol{\theta}^{(t)}) \log \alpha + \left(\sum_{i=0}^{16} n_i t_i(\boldsymbol{\theta}^{(t)}) \right) \log \beta + \left(\sum_{i=0}^{16} n_i p_i(\boldsymbol{\theta}^{(t)}) \right) \log(1 - \alpha - \beta) \\ & + \left(\sum_{i=0}^{16} i n_i t_i(\boldsymbol{\theta}^{(t)}) \right) \log \mu - \left(\sum_{i=0}^{16} n_i t_i(\boldsymbol{\theta}^{(t)}) \right) \mu \\ & + \left(\sum_{i=0}^{16} i n_i p_i(\boldsymbol{\theta}^{(t)}) \right) \log \lambda - \left(\sum_{i=0}^{16} n_i p_i(\boldsymbol{\theta}^{(t)}) \right) \lambda \end{aligned}$$

下面进入 **M** 步，最大化 $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$。分别对 $\alpha, \beta, \mu, \lambda$ 求偏导并令其为零，得到

$$\begin{cases} \dfrac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial \alpha} = \dfrac{n_0 z_0(\boldsymbol{\theta}^{(t)})}{\alpha} - \dfrac{\sum_{i=0}^{16} n_i p_i(\boldsymbol{\theta}^{(t)})}{1 - \alpha - \beta} = 0 \\ \dfrac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial \beta} = \dfrac{\sum_{i=0}^{16} n_i t_i(\boldsymbol{\theta}^{(t)})}{\beta} - \dfrac{\sum_{i=0}^{16} n_i p_i(\boldsymbol{\theta}^{(t)})}{1 - \alpha - \beta} = 0 \\ \dfrac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial \mu} = \dfrac{\sum_{i=0}^{16} i n_i t_i(\boldsymbol{\theta}^{(t)})}{\mu} - \sum_{i=0}^{16} n_i t_i(\boldsymbol{\theta}^{(t)}) = 0 \\ \dfrac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial \lambda} = \dfrac{\sum_{i=0}^{16} i n_i p_i(\boldsymbol{\theta}^{(t)})}{\lambda} - \sum_{i=0}^{16} n_i p_i(\boldsymbol{\theta}^{(t)}) = 0 \end{cases}$$

容易解的

$$\mu^{(t+1)} = \frac{\sum_{i=0}^{16} i n_i t_i(\boldsymbol{\theta}^{(t)})}{\sum_{i=0}^{16} n_i t_i(\boldsymbol{\theta}^{(t)})} \qquad \lambda^{(t+1)} = \frac{\sum_{i=0}^{16} i n_i p_i(\boldsymbol{\theta}^{(t)})}{\sum_{i=0}^{16} n_i p_i(\boldsymbol{\theta}^{(t)})}$$

而对于 α 和 β，联立方程组可解得

$$\begin{aligned} \alpha^{(t+1)} &= \frac{n_0 z_0(\boldsymbol{\theta}^{(t)})}{n_0 z_0(\boldsymbol{\theta}^{(t)}) + \sum_{i=0}^{16} n_i t_i(\boldsymbol{\theta}^{(t)}) + \sum_{i=0}^{16} n_i p_i(\boldsymbol{\theta}^{(t)})} \\ \beta^{(t+1)} &= \frac{\sum_{i=0}^{16} n_i t_i(\boldsymbol{\theta}^{(t)})}{n_0 z_0(\boldsymbol{\theta}^{(t)}) + \sum_{i=0}^{16} n_i t_i(\boldsymbol{\theta}^{(t)}) + \sum_{i=0}^{16} n_i p_i(\boldsymbol{\theta}^{(t)})} \end{aligned}$$

又注意到 $\pi_i(\boldsymbol{\theta}) = \alpha \cdot \mathbb{I}(i = 0) + \beta \cdot \mu^i e^{-\mu} + (1 - \alpha - \beta) \cdot \lambda^i e^{-\lambda}$，所以有

$$\begin{aligned} z_0(\boldsymbol{\theta}) + t_0(\boldsymbol{\theta}) + p_0(\boldsymbol{\theta}) &= \frac{\alpha + \beta e^{-\mu} + (1 - \alpha - \beta) e^{-\lambda}}{\pi_0(\boldsymbol{\theta})} = 1 \\ t_i(\boldsymbol{\theta}) + p_i(\boldsymbol{\theta}) &= \frac{\beta \mu^i e^{-\mu} + (1 - \alpha - \beta) \lambda^i e^{-\lambda}}{\pi_i(\boldsymbol{\theta})} = 1, \quad i = 1, \cdots, 16 \end{aligned}$$

所以有

$$\begin{aligned} & n_0 z_0(\boldsymbol{\theta}^{(t)}) + \sum_{i=0}^{16} n_i t_i(\boldsymbol{\theta}^{(t)}) + \sum_{i=0}^{16} n_i p_i(\boldsymbol{\theta}^{(t)}) \\ & = n_0 [z_0(\boldsymbol{\theta}^{(t)}) + t_0(\boldsymbol{\theta}^{(t)}) + p_0(\boldsymbol{\theta}^{(t)})] + \sum_{i=1}^{16} n_i [t_i(\boldsymbol{\theta}^{(t)}) + p_i(\boldsymbol{\theta}^{(t)})] \\ & = n_0 + \sum_{i=1}^{16} n_i = N \end{aligned}$$

综上所述，得到更新公式为

$$
\begin{aligned}
\alpha^{(t+1)} &= \frac{n_0 z_0(\boldsymbol{\theta}^{(t)})}{N} \\
\beta^{(t+1)} &= \sum_{i=0}^{16} \frac{n_i t_i(\boldsymbol{\theta}^{(t)})}{N} \\
\mu^{(t+1)} &= \frac{\sum_{i=0}^{16} i n_i t_i(\boldsymbol{\theta}^{(t)})}{\sum_{i=0}^{16} n_i t_i(\boldsymbol{\theta}^{(t)})} \\
\lambda^{(t+1)} &= \frac{\sum_{i=0}^{16} i n_i p_i(\boldsymbol{\theta}^{(t)})}{\sum_{i=0}^{16} n_i p_i(\boldsymbol{\theta}^{(t)})}
\end{aligned}
$$

解答 (b) 多起点迭代，得到的最优估计为 $\boldsymbol{\theta}^* = (0.2294, 0.7706, 3.5013, 0.8959)$，最优的 $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)$ 函数值为 216.5306。

理论上而言 μ 和 λ 是不可识别的，不过从估计结果分析：典型行为的占比较大，且典型行为代表的平均次数大约为 $3-4$ 次。同时存在许多人群回答 0 次，而危险行为的占比几乎没有。这个结果比较符合直觉。

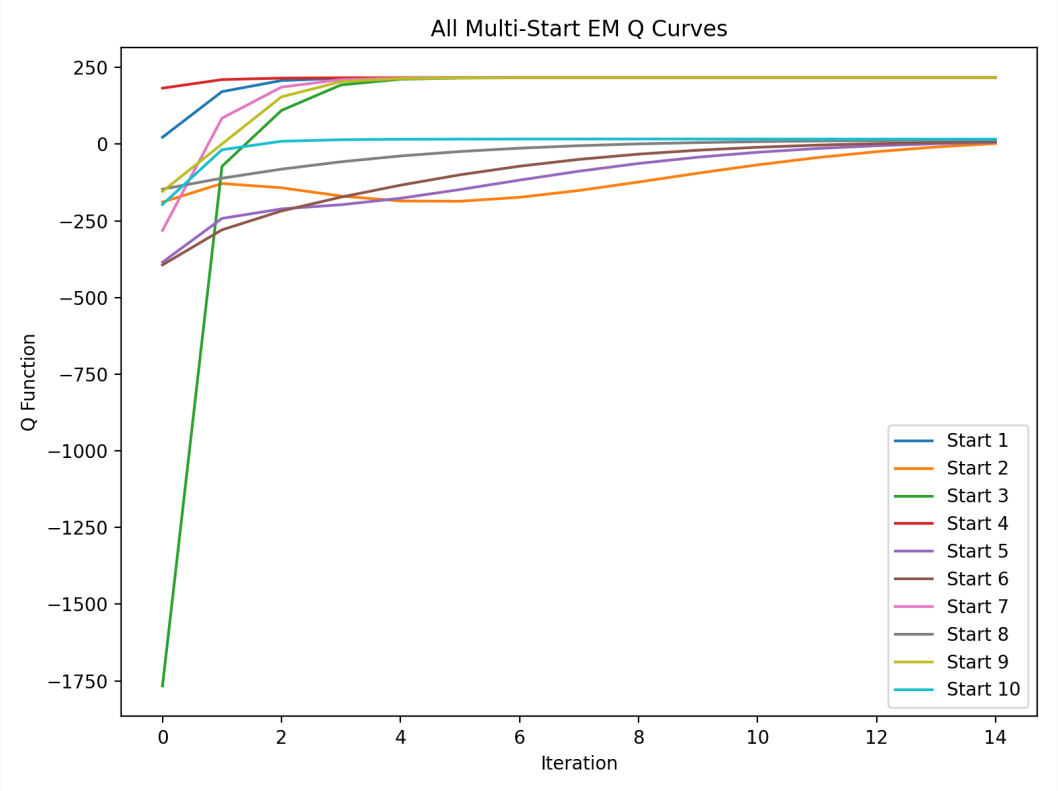

图 1: Multiple Initial Points for Params Estimated

程序运行日志见 `log/q1.log`。

上面结果的初始点均为从 $U(0,1)$ 随机抽取。若从 $\boldsymbol{\theta}_0 = (0.2, 0.6, 2, 3)$ 开设迭代（更符合直觉），最终结果为 $\boldsymbol{\theta}^* = (0.1353, 0.5645, 1.5575, 6.0575)$，而 $Q = 19.1054$。

解答 (c) 由问题 a 和 b 已经得到参数 $\alpha, \beta, \mu, \lambda$ 的估计 $\boldsymbol{\theta}^*$，下面使用 Louis 方法，求

$$\hat{\boldsymbol{i}}_{\boldsymbol{X}} = \hat{\boldsymbol{i}}_{\boldsymbol{Y}} - \hat{\boldsymbol{i}}_{\boldsymbol{Z}|\boldsymbol{X}}$$

由问题 a 已经求得 $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)$，故二阶导即为 $\hat{\boldsymbol{i}}_{\boldsymbol{Y}}$，先求主对角线元素

$$\begin{cases} -\dfrac{\partial^2 Q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{\partial \alpha^2} = \dfrac{n_0 z_0(\boldsymbol{\theta}^*)}{\alpha^2} + \dfrac{\sum_{i=0}^{16} n_i p_i(\boldsymbol{\theta}^*)}{(1-\alpha-\beta)^2} \\ -\dfrac{\partial^2 Q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{\partial \beta^2} = \dfrac{\sum_{i=0}^{16} n_i t_i(\boldsymbol{\theta}^*)}{\beta^2} + \dfrac{\sum_{i=0}^{16} n_i p_i(\boldsymbol{\theta}^*)}{(1-\alpha-\beta)^2} \\ -\dfrac{\partial^2 Q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{\partial \mu^2} = \dfrac{\sum_{i=0}^{16} i n_i t_i(\boldsymbol{\theta}^*)}{\mu^2} \\ -\dfrac{\partial^2 Q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{\partial \lambda^2} = \dfrac{\sum_{i=0}^{16} i n_i p_i(\boldsymbol{\theta}^*)}{\lambda^2} \end{cases}$$

而其他交叉项只有 α, β 之间非零

$$-\frac{\partial^2 Q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{\partial \alpha \partial \beta} = \frac{\sum_{i=0}^{16} n_i p_i(\boldsymbol{\theta}^*)}{(1-\alpha-\beta)^2}$$

如此我们得到矩阵 $\hat{\boldsymbol{i}}_{\boldsymbol{Y}} \in \mathbb{R}^{4\times 4}$ 的结果。

下面求 $\hat{\boldsymbol{i}}_{\boldsymbol{Z}|\boldsymbol{X}}$，由条件分布

$$n_{z,0}|n_0 \sim B(n_0, z_0(\boldsymbol{\theta})) \quad n_{t,i}|n_i \sim B(n_i, t_i(\boldsymbol{\theta})) \quad n_{p,i}|n_i \sim B(n_i, p_i(\boldsymbol{\theta}))$$

其中 $i = 0, \cdots, 16$。于是条件联合概率密度为

$$f_{Z|X}(z|x) = \binom{n_0}{n_{z,0}} z_0^{n_{z,0}} (1-z_0)^{n_0 - n_{z,0}} \cdot \prod_{i=0}^{16} \binom{n_i}{n_{t,i}} t_i^{n_{t,i}} (1-t_i)^{n_i - n_{t,i}} \cdot \prod_{i=0}^{16} \binom{n_i}{n_{p,i}} p_i^{n_{p,i}} (1-p_i)^{n_i - n_{p,i}}$$

于是

$$\begin{aligned} \log f_{Z|X}(z|x) = &\log \binom{n_0}{n_{z,0}} + n_{z,0} \log z_0 + (n_0 - n_{z,0}) \log(1-z_0) \\ &+ \sum_{i=0}^{16} \left[\log \binom{n_i}{n_{t,i}} + n_{t,i} \log t_i + (n_i - n_{t,i}) \log(1-t_i) \right] \\ &+ \sum_{i=0}^{16} \left[\log \binom{n_i}{n_{p,i}} + n_{p,i} \log p_i + (n_i - n_{p,i}) \log(1-p_i) \right] \end{aligned}$$

那么就有对 $\theta \in \{\alpha, \beta, \mu, \lambda\}$ 求导

$$\frac{\partial \log f_{Z|X}(z|x)}{\partial \theta} = \left[\frac{n_{z,0}}{z_0} - \frac{n_0 - n_{z,0}}{1-z_0} \right] \frac{\partial z_0}{\partial \theta} + \sum_{i=0}^{16} \left[\frac{n_{t,i}}{t_i} - \frac{n_i - n_{t,i}}{1-t_i} \right] \frac{\partial t_i}{\partial \theta} + \sum_{i=0}^{16} \left[\frac{n_{p,i}}{p_i} - \frac{n_i - n_{p,i}}{1-p_i} \right] \frac{\partial p_i}{\partial \theta}$$

由 $\mathrm{Var}(n_{z,0}|n_0) = n_0 z_0 (1-z_0)$, $\mathrm{Var}(n_{t,i}|n_i) = n_i t_i (1-t_i)$, $\mathrm{Var}(n_{p,i}|n_i) = n_i p_i (1-p_i)$ 可求

$$\hat{\boldsymbol{i}}_{\boldsymbol{Z}|\boldsymbol{X}} = \mathrm{Cov}\left(\frac{\partial \log f_{Z|X}(z|x)}{\partial \theta} \middle| X = (n_0, \cdots, n_{16}) \right) \in \mathbb{R}^{4\times 4}$$

如此可得参数估计的方差和协方差为 $[\hat{\boldsymbol{i}}_{\boldsymbol{X}}]^{-1} = [\hat{\boldsymbol{i}}_{\boldsymbol{Y}} - \hat{\boldsymbol{i}}_{\boldsymbol{Z}|\boldsymbol{X}}]^{-1}$。其中 $\hat{\boldsymbol{i}}_{\boldsymbol{Z}|\boldsymbol{X}}$ 的解析计算过于复杂，这里采用从 $f_{Z|X}$ 中抽样，计算样本协方差矩阵代替（求导代码由 AI 生成）。得到估计如下

$$\operatorname{Corr}(\hat{\alpha},\hat{\beta},\hat{\mu},\hat{\lambda}) = \begin{pmatrix} 1 & -0.979 & 0.180 & 9.80\times10^{4} \\ - & 1 & -0.162 & -8.79\times10^{4} \\ - & - & 1 & 6.85\times10^{4} \\ - & - & - & 1 \end{pmatrix}$$

为对称阵，故只展示上三角。标准误为 $[0.0133, 0.0131, 0.0589, 24.7]$。Louis 方法受限于 $\hat{\boldsymbol{i}}_{\boldsymbol{Z}|\boldsymbol{X}}$ 的估计，下面采用 Bootstrap 方法，对同一起点 $\boldsymbol{\theta}_0 = (0.2, 0.6, 2, 3)$ 开设迭代，数据重采样，得到相关系数的估计为

$$\operatorname{Corr}(\hat{\alpha},\hat{\beta},\hat{\mu},\hat{\lambda}) = \begin{pmatrix} 1 & -0.479 & 0.527 & 0.333 \\ - & 1 & -0.136 & -0.136 \\ - & - & 1 & 0.417 \\ - & - & - & 1 \end{pmatrix}$$

为对称阵，故只展示上三角。标准误为 $[0.01, 0.02, 0.09, 0.18]$。

Exercise 2

本书的网站里有从 $N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 分布抽取的 50 个三维数据点。某些数据点在一个或多个分量上有缺失值。50 个观测值里只有 27 个是完整的。

a 导出 $\boldsymbol{\mu}$ 和 $\boldsymbol{\Sigma}$ 联合极大似然估计的 EM 算法。最容易记起的是多元正态密度属于指数族。
b 由合适的初始点确定它们的极大似然估计。考查这个算法的表现，并评价所得的结果。
c 使用均值填充缺失值后直接求解 MLE，并与 b 问的结果进行比较。

解答 (a) 记第 i 个数据点为 $U_i \in \mathbb{R}^d$，其中 $d = 3$，那么有

$$U_i \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad i = 1, \cdots, n$$

其中 $n = 50, \boldsymbol{\mu} \in \mathbb{R}^d, \boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$。

E-Step

于是完整数据的对数似然函数的核函数为

$$l(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{n}(U_i - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(U_i - \boldsymbol{\mu})$$

于是在观测数据 $\boldsymbol{O}$ 的条件下，l 的期望为

$$\begin{aligned} Q(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}) &= \mathbb{E}[l(\boldsymbol{\mu}, \boldsymbol{\Sigma})|\boldsymbol{O}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}] \\ &= -\frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}[(U_i - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(U_i - \boldsymbol{\mu})|\boldsymbol{O}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}] \end{aligned}$$

针对某个 i，记 $U_i = (U_{i,o}^T, U_{i,m}^T)^T \in \mathbb{R}^d$，其中 $U_{i,o} \in \mathbb{R}^{d_o^i}$ 为观测到的分量，$U_{i,m} \in \mathbb{R}^{d_m^i}$ 为缺失的分量，且 $d_o^i + d_m^i = d$。注意到 $(U_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (U_i - \boldsymbol{\mu}) \in \mathbb{R}$ 为标量，则

$$(U_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (U_i - \boldsymbol{\mu}) = \operatorname{tr}\left\{(U_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (U_i - \boldsymbol{\mu})\right\} = \operatorname{tr}\left\{\boldsymbol{\Sigma}^{-1} (U_i - \boldsymbol{\mu})(U_i - \boldsymbol{\mu})^T\right\}$$

而

$$(U_i - \boldsymbol{\mu})(U_i - \boldsymbol{\mu})^T = U_i U_i^T - U_i \boldsymbol{\mu}^T - \boldsymbol{\mu} U_i^T + \boldsymbol{\mu}\boldsymbol{\mu}^T \tag{1}$$

于是我们只需计算 $\mathbb{E}[U_i | U_{i,o}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}]$ 和 $\mathbb{E}[U_i U_i^T | U_{i,o}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}]$ 即可。

注意到，此时 $U_i \sim N_d(\boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)})$，根据 U_i 的拆分，同样对 $\boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}$ 进行拆分

$$\boldsymbol{\mu}^{(t)} = \begin{pmatrix} \boldsymbol{\mu}_o^{(t)} \\ \boldsymbol{\mu}_m^{(t)} \end{pmatrix}, \quad \boldsymbol{\Sigma}^{(t)} = \begin{pmatrix} \boldsymbol{\Sigma}_{oo}^{(t)} & \boldsymbol{\Sigma}_{om}^{(t)} \\ \boldsymbol{\Sigma}_{mo}^{(t)} & \boldsymbol{\Sigma}_{mm}^{(t)} \end{pmatrix}$$

其中 $\boldsymbol{\mu}_o^{(t)} \in \mathbb{R}^{d_o^i}, \boldsymbol{\mu}_m^{(t)} \in \mathbb{R}^{d_m^i}$，$\boldsymbol{\Sigma}_{oo}^{(t)} \in \mathbb{R}^{d_o^i \times d_o^i}, \boldsymbol{\Sigma}_{om}^{(t)} = [\boldsymbol{\Sigma}_{mo}^{(t)}]^T \in \mathbb{R}^{d_o^i \times d_m^i}, \boldsymbol{\Sigma}_{mm}^{(t)} \in \mathbb{R}^{d_m^i \times d_m^i}$（对于不同的 i，因为 U_i 缺失方式不同，故对 $\boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}$ 的拆分方式也不同）。根据多元正态分布的条件分布，有

$$U_{i,m} | U_{i,o}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)} \sim N_{d_m^i}\left(\boldsymbol{\mu}_m^{(t)} + \boldsymbol{\Sigma}_{mo}^{(t)} (\boldsymbol{\Sigma}_{oo}^{(t)})^{-1} (U_{i,o} - \boldsymbol{\mu}_o^{(t)}), \boldsymbol{\Sigma}_{mm}^{(t)} - \boldsymbol{\Sigma}_{mo}^{(t)} (\boldsymbol{\Sigma}_{oo}^{(t)})^{-1} \boldsymbol{\Sigma}_{om}^{(t)}\right)$$

方便起见，简记为 $U_{i,m} | U_{i,o}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)} \sim N(\boldsymbol{\mu}_{m|o}^{(t)}, \boldsymbol{\Sigma}_{m|o}^{(t)})$。于是

$$\mathbb{E}[U_i | U_{i,o}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}] = \begin{pmatrix} U_{i,o} \\ \boldsymbol{\mu}_{m|o}^{(t)} \end{pmatrix} \tag{2}$$

下面求 $\mathbb{E}[U_i U_i^T | U_{i,o}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}]$，注意到

$$U_i U_i^T = \begin{pmatrix} U_{i,o} U_{i,o}^T & U_{i,o} U_{i,m}^T \\ U_{i,m} U_{i,o}^T & U_{i,m} U_{i,m}^T \end{pmatrix}$$

只需要分别求条件期望即可。容易得到

$$\begin{aligned}
&\mathbb{E}[U_{i,o} U_{i,o}^T | U_{i,o}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}] = U_{i,o} U_{i,o}^T \\
&\mathbb{E}[U_{i,o} U_{i,m}^T | U_{i,o}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}] = U_{i,o} \cdot \mathbb{E}[U_{i,m}^T | U_{i,o}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}] = U_{i,o} [\boldsymbol{\mu}_{m|o}^{(t)}]^T \\
&\mathbb{E}[U_{i,m} U_{i,o}^T | U_{i,o}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}] = \{\mathbb{E}[U_{i,o} U_{i,m}^T | U_{i,o}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}]\}^T = \boldsymbol{\mu}_{m|o}^{(t)} U_{i,o}^T
\end{aligned}$$

又注意到 $\operatorname{Cov}[U_{i,m} | U_{i,o}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}] = \boldsymbol{\Sigma}_{m|o}^{(t)}$，故有

$$\begin{aligned}
\mathbb{E}[U_{i,m} U_{i,m}^T | U_{i,o}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}] &= \operatorname{Cov}[U_{i,m} | U_{i,o}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}] \\
&\quad + (\mathbb{E}[U_{i,m} | U_{i,o}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}])(\mathbb{E}[U_{i,m} | U_{i,o}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}])^T \\
&= \boldsymbol{\Sigma}_{m|o}^{(t)} + \boldsymbol{\mu}_{m|o}^{(t)} [\boldsymbol{\mu}_{m|o}^{(t)}]^T
\end{aligned}$$

综合可得

$$\mathbb{E}[U_i U_i^T | U_{i,o}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}] = \begin{pmatrix} U_{i,o} U_{i,o}^T & U_{i,o} [\boldsymbol{\mu}_{m|o}^{(t)}]^T \\ \boldsymbol{\mu}_{m|o}^{(t)} U_{i,o}^T & \boldsymbol{\Sigma}_{m|o}^{(t)} + \boldsymbol{\mu}_{m|o}^{(t)} [\boldsymbol{\mu}_{m|o}^{(t)}]^T \end{pmatrix} \tag{3}$$

结合式 1 将式 2 和 3 代入，即可求得 $Q(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)})$，即 E 步完成。

M-Step (CM-Step)

注意，上面式 2 和 3 是针对固定 i 的结果，现记

$$\hat{U}_i = \mathbb{E}[U_i|U_{i,o}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}] \in \mathbb{R}^d \qquad \hat{R}_i = \mathbb{E}[U_i U_i^T|U_{i,o}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}] \in \mathbb{R}^{d\times d}$$

那么

$$\begin{aligned}\mathbb{E}[\mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}(U_i-\boldsymbol{\mu})(U_i-\boldsymbol{\mu})^T\right\}|U_{i,o}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}] &= \mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}[\hat{R}_i - \hat{U}_i\boldsymbol{\mu}^T - \boldsymbol{\mu}\hat{U}_i^T + \boldsymbol{\mu}\boldsymbol{\mu}^T]\right\} \\ &= \mathrm{tr}(\boldsymbol{\Sigma}^{-1}\hat{R}_i) - \mathrm{tr}(\boldsymbol{\Sigma}^{-1}\hat{U}_i\boldsymbol{\mu}^T) - \mathrm{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\hat{U}_i^T) + \mathrm{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\boldsymbol{\mu}^T)\end{aligned}$$

代入 Q 函数可得

$$\begin{aligned}Q(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}) &= -\frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^n \mathbb{E}[(U_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(U_i-\boldsymbol{\mu})|\boldsymbol{O}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}] \\ &= -\frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^n \left(\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\hat{R}_i) - \mathrm{tr}(\boldsymbol{\Sigma}^{-1}\hat{U}_i\boldsymbol{\mu}^T) - \mathrm{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\hat{U}_i^T) + \mathrm{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\boldsymbol{\mu}^T)\right) \\ &= -\frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\left(\mathrm{tr}(\boldsymbol{\Sigma}^{-1}S_2^{(t)}) - \mathrm{tr}(\boldsymbol{\Sigma}^{-1}S_1^{(t)}\boldsymbol{\mu}^T) - \mathrm{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}[S_1^{(t)}]^T) + n\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\boldsymbol{\mu}^T)\right)\end{aligned}$$

其中

$$S_1^{(t)} = \sum_{i=1}^n \hat{U}_i \in \mathbb{R}^d \qquad S_2^{(t)} = \sum_{i=1}^n \hat{R}_i \in \mathbb{R}^{d\times d}$$

上标 (t) 是因为 $\hat{U}_i, \hat{R}_i$ 均是由 $\boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}$ 和观测数据推导而来。

我们的目标是

$$\max_{\boldsymbol{\mu},\boldsymbol{\Sigma}} Q(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)})$$

同时优化 $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ 过于复杂，可以采用 $s = 2$ 的 ECM 算法，即 $(\boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}) \to (\boldsymbol{\mu}^{(t+1/2)}, \boldsymbol{\Sigma}^{(t)}) \to (\boldsymbol{\mu}^{(t+1/2)}, \boldsymbol{\Sigma}^{(t+2/2)}) = (\boldsymbol{\mu}^{(t+1)}, \boldsymbol{\Sigma}^{(t+1)})$。所以，下面我们固定 $\boldsymbol{\Sigma}$，考虑优化 $\boldsymbol{\mu}$。由 tr 的求导性质，可知

$$\frac{\partial \mathrm{tr}(\boldsymbol{A}\boldsymbol{X}^T)}{\partial \boldsymbol{X}} = \boldsymbol{A} \quad \frac{\partial \mathrm{tr}(\boldsymbol{A}\boldsymbol{X}\boldsymbol{B})}{\partial \boldsymbol{X}} = \boldsymbol{A}^T\boldsymbol{B}^T \quad \frac{\partial \mathrm{tr}(\boldsymbol{A}\boldsymbol{X}\boldsymbol{X}^T)}{\partial \boldsymbol{X}} = (\boldsymbol{A}+\boldsymbol{A}^T)\boldsymbol{X}$$

于是根据

$$\begin{aligned}&\frac{\partial Q(\boldsymbol{\mu}|\boldsymbol{\Sigma}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)})}{\partial \boldsymbol{\mu}} = -\frac{1}{2}\left(2n\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - 2\boldsymbol{\Sigma}^{-1}S_1^{(t)}\right) = 0 \\ &\Rightarrow \quad \boldsymbol{\mu}^{(t+1/2)} = \frac{1}{n}S_1^{(t)}\end{aligned}$$

又因为

$$\frac{\partial \log|\boldsymbol{X}|}{\partial \boldsymbol{X}} = [\boldsymbol{X}^{-1}]^T \qquad \frac{\partial \mathrm{tr}(\boldsymbol{A}\boldsymbol{X}^{-1}\boldsymbol{B})}{\partial \boldsymbol{X}} = -(\boldsymbol{X}^{-1}\boldsymbol{B}\boldsymbol{A}\boldsymbol{X}^{-1})^T$$

于是有

$$\begin{aligned}&\frac{\partial Q(\boldsymbol{\Sigma}|\boldsymbol{\mu}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)})}{\partial \boldsymbol{\Sigma}} = -\frac{n}{2}\boldsymbol{\Sigma}^{-1} + \frac{1}{2}\left(\boldsymbol{\Sigma}^{-1}S_2^{(t)}\boldsymbol{\Sigma}^{-1} - 2\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}[S_1^{(t)}]^T\boldsymbol{\Sigma}^{-1} + n\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\right) = 0 \\ &\Rightarrow \quad \boldsymbol{\Sigma}^{(t+1)} = \frac{1}{n}\left(S_2^{(t)} - 2\boldsymbol{\mu}^{(t+1/2)}[S_1^{(t)}]^T + n\boldsymbol{\mu}^{(t+1/2)}[\boldsymbol{\mu}^{(t+1/2)}]^T\right)\end{aligned}$$

综上所述，我们得到 CM-Step 的更新式为

$$\begin{cases} \boldsymbol{\mu}^{(t+1)} = \dfrac{1}{n} S_1^{(t)} \\ \boldsymbol{\Sigma}^{(t+1)} = \dfrac{1}{n}\left(S_2^{(t)} - \dfrac{1}{n} S_1^{(t)} [S_1^{(t)}]^T \right) \end{cases}$$

解答 (b,c) 采用样本均值填充缺失值，得到完整数据 $\tilde{U}_i,\ i = 1, \cdots, n$，于是对参数的 MLE 为

$$\hat{\boldsymbol{\mu}}_0 = \frac{1}{n}\sum_{i=1}^{n} \tilde{U}_i \in \mathbb{R}^d \qquad \hat{\boldsymbol{\Sigma}}_0 = \frac{1}{n-1}\sum_{i=1}^{n} (\tilde{U}_i - \hat{\mu}_0)(\tilde{U}_i - \hat{\mu}_0)^T \in \mathbb{R}^{d\times d}$$

计算其 Q 函数值为 $Q_0 = -29.7058$，下面我们以 $(\hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\Sigma}}_0)$ 为初始值进行优化。得到最终结果为

$$\hat{\boldsymbol{\mu}}^* = \begin{pmatrix} 0.8786 \\ 2.8502 \\ 9.0257 \end{pmatrix} \quad \hat{\boldsymbol{\Sigma}}^* = \begin{pmatrix} 1.4131 & 1.0016 & 1.3185 \\ 1.0016 & 0.7779 & 0.7043 \\ 1.3185 & 0.7043 & 2.5224 \end{pmatrix}$$

计算其 Q 函数值为 $Q^* = 0.1911$ 优于 Q_0。收敛过程见程序日志 `log/q2.log`。