2019302030053-胡哲-第五次作业

EM算法习题

第一题

首先生成样本，可以利用均匀分布生成0.2,0.3,0.5的概率。

之后利用EM算法进行迭代

E Step

注意由Bayes定理得到

概率函数

$$\frac{p(x_j|z_{ij}=1,\theta^{(t)})\,p(z_{ij}=1|\theta^{(t)})}{\sum_{i=1}^{k} p(x_j|z_{ij}=1,\theta^{(t)})\,p(z_{ij}=1|\theta^{(t)})}$$

$$E[Z_{ij}|\underbrace{x}_{(x_1,\cdots,x_n)},\theta^{(t)}] = P(\underbrace{Z_{ij}=1}_{x_j\text{来源于}f_i}|x,\theta^{(t)}) = \frac{p_i^{(t)} f_i(x_j)}{\sum_{i=1}^k p_i^{(t)} f_i(x_j)} := z_{ij}^{(k)}$$

因此

$$Q(\theta|\theta^{(t)}) = \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n} z_{ij}^{(t)} log p_i}_{\mathrm{I}} - \underbrace{\frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{n} z_{ij}^{(t)}[log\sigma^2 + (x_j-\mu_i)^2/\sigma^2]}_{\mathrm{II}} + const.$$

M Step

最大化$Q(\theta|\theta^{(t)})$得到

$$p_i^{(t+1)} = \frac{1}{n}\sum_{j=1}^{n} z_{ij}^{(t)}, \qquad \mu_i^{(t+1)} = \frac{\sum_{j=1}^n z_{ij}^{(t)} x_j}{\sum_{j=1}^n z_{ij}^{(t)}}$$

$$(\sigma^2)^{(t+1)} = \sum_{i=1}^{k}\sum_{j=1}^{n} z_{ij}^{(t)}(x_j-\mu_i^{(t+1)})^2/n$$

迭代公式同上图。

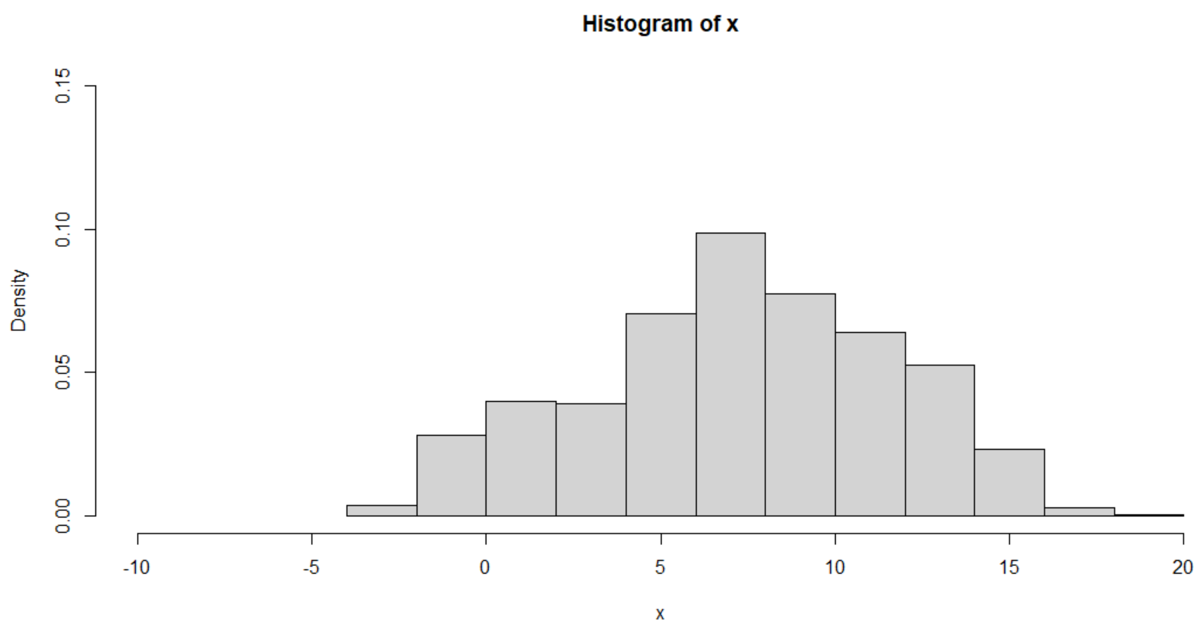
标准差估计利用bootstrap方法。

具体代码如下：

```
#### 题目一 ####
# 生成数据
n <- 1000
x1 <- rnorm(n, 1, 2)
x2 <- rnorm(n, 12, 2)
x3 <- rnorm(n, 7 , 2)
u <- runif(n)
k1 <- as.integer(u <0.2) #vector of 0’s and 1’s
k2 <- as.integer(u>=0.2 & u<=0.5)
x <- k1 * x1 + k2 * x2 + (1-k1-k2)*x3#the mixture
hist(x, prob=TRUE, xlim=c(-10,20), ylim=c(0,0.15))

em_mixnorm <- function(x,seed=520){
set.seed(seed)
# 真实概率(0.2,0.3,0.5)
# 初始化
alpha10 <- 0.16
alpha20 <- 0.36
alpha30 <- 0.48
mu10 <- 1.2
mu20 <- 11.6
mu30 <- 6.8
sigma0 <- 2.3
para <- c(alpha10,alpha20,alpha30,mu10,mu20,mu30,sigma0)
tol <- 1e-8
para.old <- para+1
for (j in 1:1000) {
  vp10<-(para[1]*dnorm(x,para[4],para[7]))/(para[1]*dnorm(x,para[4],para[7])
                                          +para[2]*dnorm(x,para[5],para[7])
                                          +para[3]*dnorm(x,para[6],para[7]))
  vp20<-(para[2]*dnorm(x,para[5],para[7]))/(para[1]*dnorm(x,para[4],para[7])
                                          +para[2]*dnorm(x,para[5],para[7])
                                          +para[3]*dnorm(x,para[6],para[7]))
  vp30<-(para[3]*dnorm(x,para[6],para[7]))/(para[1]*dnorm(x,para[4],para[7])
                                           +para[2]*dnorm(x,para[5],para[7])
                                           +para[3]*dnorm(x,para[6],para[7]))

  phi1<-sum(vp10)
  phi2<-sum(vp20)
  phi3 <- sum(vp30)
  alpha11<-phi1/n
  alpha21<-phi2/n
  alpha31 <- phi3/n
  phix1<-sum(x*vp10)
  mu11<-phix1/phi1
  phix2<-sum(x*vp20)
```

```
  mu21<-phix2/phi2
  phix3 <- sum(x*vp30)
  mu31<-phix3/phi3

  phixmu1<-sum((x-mu11)^2*vp10)
  phixmu2<-sum((x-mu21)^2*vp20)
  phixmu3 <- sum((x-mu31)^2*vp30)
  sigma1 <- sqrt(sum(phixmu1,phixmu2,phixmu3)/sum(phi1,phi2,phi3))

  para<-c(alpha11,alpha21,alpha31,mu11,mu21,mu31,sigma1)

  if (sqrt(sum((para-para.old)^2))/sqrt(sum(para.old^2))<tol) break
  para.old<-para

}

data.frame(estimate=para,iter=j,tol=tol)
}

# bootstrap 估计标准差
B <- 200 #number of replicates
N <- length(x) #sample size
R <- matrix(0,nrow = B,ncol = 7) #storage for replicates
#bootstrap estimate of standard error of R
for (b in 1:B) {
  #randomly select the indices
  i <- sample(1:N, size = N, replace = TRUE)

   x <- x[i]
  result <- em_mixnorm(x)
  for (j in 1:7)R[b,j] <- result[j,1] # 矩阵储存估计的七个参数
}
# 输出标准差的估计
print(em_mixnorm(x))
print(round(apply(R, 2, sd),3))
```

本题中，X的概率密度直方图如图：



输出结果如下:

参数	真实值	初始值	估计值	标准差
p_1	0.2	0.16	0.18	0.007
p_2	0.3	0.36	0.33	0.011
p_3	0.5	0.48	0.49	0.015
μ_1	1	1.2	1.1	0.008
μ_2	12	11.6	12.2	0.059
μ_3	7	6.5	7.3	0.035
σ^2	2	2.3	0.1	0.015

第二题

思路：

设缺失数据的失效时间为$z_1, z_2, \ldots z_m$，则可以得到联合似然函数：

$$\log L(\theta|x) = n(\log - \theta\overline{y}) + \sum_{i=m}^{m}(\log\theta - \theta z_i)$$

取条件期望可得：

$$E(\log L(\theta|x)|y,\theta_t) = (m+n)\theta - \theta[n\overline{y} + r(1/\theta_t) - tg_t) + (m-r)(t+1/\theta_t) \quad g_t = \frac{exp(-t\theta_t)}{1-exp(-t\theta_t)}$$

则θ_t迭代公式如下：

$$\theta_{t+1} = (m+n)[n\overline{y} + r(1/\theta_t - tg_t) + (m-r)(t+1/\theta_t)]^{-1}$$

具体代码如下：

```
# 生成数据
n <- 300
m <- 200
r <- 80
theta <- 1/3
t <- 1.5 # t的生成要根据r，m和theta决定的分位数函数
y <- rexp(n,theta)
theta0 <- 1
for (j in 1:3000) {
  ht <- exp(-t*theta0)/(1-exp(-t*theta0))
  theta1 <- (n+m)/(sum(y)+(m-r)*(t+1/theta0)+r*(1/theta0-t*ht)) #迭代公式
  theta0 <- theta1
}
print(theta0)
```

θ初值为1，真实值为1/3，估计值为0.3362416