10-1

1 对有限正态混合总体，以$k=3$为例，即

$$X \stackrel{d}{=} p_1N(\mu_1,\sigma^2)+p_2N(\mu_2,\sigma^2)+p_3N(\mu_3,\sigma^2)$$

其中$p_1<p_2<p_3$. 生成X的一组样本,

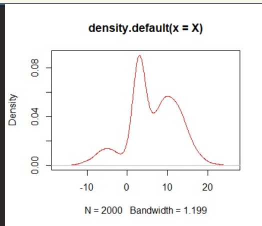
(1) 使用EM算法估计参数$(p_1,p_2,\mu_1,\mu_2,\mu_3,\sigma^2)$.

(2) 试使用方差估计的某个方法，估计(1)中各估计的标准差的估计.

分析过程：这是一个正态混合参数估计问题，$\theta=(p,u,\sigma^2)$为待估参数，其中 $p=(p_1,p_2,p_3)$ $u=(u_1,u_2,u_3)$。假设 $p_1=0.1$，$p_2=0.3$，$p_3=0.6$，$u_1=-5$，$\sigma_1^2=3$，$u_2=3$，$\sigma_2^2=1$，$u_3=10$，$\sigma_3^2=4$. 生成X的一组样本 $X=(x_1,\cdots,x_N)$，$N=2000$，初始化参数，利用EM算法迭代求解

```r
set.seed(101)
N<-1000              ###生成X的一组样本
n1<-rbinom(1,size=N,prob=0.1)
n2<-rbinom(1,size=N,prob=0.3)
n3<-N-n1-n2
X1<-rnorm(n1,-5,3)
X2<-rnorm(n2,3,1)
X3<-rnorm(n3,10,4)
X0<-c(X1,X2,X3)
X0<-X[sample(1:N)]
plot(density(X),col="red")

EM<-function(X){
##initial value规定初始值###
p10<-0.1
p20<-0.3
p30<-0.6
mu10<--5.2
sigma10<-3.1
mu20<-3.3
sigma20<-1.2
mu30<-9.3
sigma30<-4.4
para<-c(p10,p20,p30,mu10,sigma10,mu20,sigma20,mu30,sigma30)
tol<-1e-8
para.old<-para+1
###
for(j in 1:1000){
  #count<-count+1
  vp10<-(para[1]*dnorm(X,para[4],para[5]))/(para[1]*dnorm(X,para[4],para[5])
                                           +para[2]*dnorm(X,para[6],para[7])
                                           +para[3]*dnorm(X,para[8],para[9]))
  vp20<-(para[2]*dnorm(X,para[6],para[7]))/(para[1]*dnorm(X,para[4],para[5])
                                           +para[2]*dnorm(X,para[6],para[7])
                                           +para[3]*dnorm(X,para[8],para[9]))
  vp30<-(para[3]*dnorm(X,para[8],para[9]))/(para[1]*dnorm(X,para[4],para[5])
```




```r
                                           +para[2]*dnorm(X,para[6],para[7])
                                           +para[3]*dnorm(X,para[8],para[9]))

  phi1<-sum(vp10)
  phi2<-sum(vp20)
  phi3<-sum(vp30)
  p11<-phi1/N
  p21<-phi2/N
  p31<-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)
  sigma11<-sqrt(phixmu1/phi1)
  phixmu2<-sum((X-mu21)^2*vp20)
  sigma21<-sqrt(phixmu2/phi2)
  phixmu3<-sum((X-mu31)^2*vp30)
  sigma31<-sqrt(phixmu3/phi3)

  para<-c(p11,p21,p31,mu11,sigma11,mu21,sigma21,mu31,sigma31)

  if (sqrt(sum((para-para.old)^2))/sqrt(sum(para.old^2))<tol)
  {break}
  para.old<-para
}
  return(list(estimate=para,iter=j,tol=tol))
}
#print(list(estimate=para,iter=j,tol=tol))

B <- 50
n <- N
theta.b <- matrix(0,nrow=B,ncol=9)
theta.b[1,]<-EM(X0)$estimate
```

```r
B <- 50
n <- N
theta.b <- matrix(0,nrow=B,ncol=9)
theta.b[1,]<-EM(X0)$estimate
for (b in 2:B) {
  i <- sample(1:n, size = n, replace = TRUE)
  Xnew <- X0[i]
  theta.b[b,] <- EM(Xnew)$estimate
}
print(theta.b)
sqrt(diag(cov(theta.b)))        #计算样本协方差矩阵
```

(1).

```
> print(list(estimate=para,iter=j,tol=tol))
$estimate
[1]  0.1086194  0.2966098  0.5947708
[4] -4.9767161  3.0172355  2.9069963
[7]  0.9630215 10.2146407  3.9936090

$iter
[1] 92

$tol
[1] 1e-08
```

结论：图中estimate数据即为各参数的估计

(2).

```
> sqrt(diag(cov(theta.b)))
[1] 0.00977055 0.02499198 0.02629029 0.39873862
[5] 0.33572892 0.08494352 0.07279765 0.24781455
[9] 0.15521946
```

10.2

2.(1)假设一种灯泡的寿命服从某个参数为θ的指数分布. 为了估计θ，我们测试了 n 个灯泡直至他们失效，记失效的时间是 $y_1,\dots,y_n$; 另外还有一组独立的实验测试了m个灯泡, 但是单个灯泡的失效时间没有被记录, 只记录下来在时刻 t时共有 r个灯泡失效. 试(给出)使用EM算法求θ的极大似然估计的迭代公式.

提示：可观测数据：$y_1, \cdots, y_n, r$

完全数据：$y_1, \cdots, y_n, \underbrace{z_1, \cdots, z_m}$，其中$z_1, \cdots, z_m$为另一组实验中$m$个灯泡的失效时间.

r个在t时刻失效

$m-r$个在t时刻未失效.

(2) 自己设定参数(如上)生成数据，并给出用EM算法求得的$\hat\theta_{MLE}$.

分析：(1). $E(X)=\frac{1}{\theta}$，$f(x)=\theta e^{-\theta x} \quad x>0.$，$F(x)=\begin{cases}1-e^{-\theta x}, & x\geq 0\\ 0, & x<0\end{cases} \longrightarrow 1-e^{-\theta x} = 1-\frac{1}{e^{\theta x}} = \frac{e^{\theta x}-1}{e^{\theta x}}$

$$L(\theta|X) = \theta^{n+m} e^{-\theta\left(\sum_{i=1}^{n} y_i + \sum_{i=1}^{n} z_i\right)}.$$

其中，$X=(y,z)$.

$$\Rightarrow \log L(\theta|X) = (n+m)\log\theta - \theta\left(n\bar{y} + \sum_{i=1}^{m} z_i\right)$$

① E步：$Q(\theta|\theta^{(t)}) = E_X\{\log L(\theta|X) \mid y, \theta_t\}$

$$= (m+n)\log\theta - \theta\Big[n\bar{y} + \underbrace{(m-r)}_{t\text{时刻未失效}}\underbrace{\left(t+\tfrac{1}{\theta^{(t)}}\right)}_{\text{估计失效时间}} + \overset{\text{失效}}{r}\left(\tfrac{1}{\theta^{(t)}} - t h_t\right)\Big] \qquad (*)$$

② M步：由 $\frac{\partial Q}{\partial\theta}=0 \Rightarrow \frac{m+n}{\theta} - n\bar{y} + (m-r)\left(t+\frac{1}{\theta^{(t)}}\right) + r\left(\frac{1}{\theta^{(t)}} - t h_t\right) = 0$，$h_t = \frac{e^{-t\theta_t}}{1-e^{-t\theta_t}}$

$$\Rightarrow \theta^{(t+1)} = (m+n)\cdot\left[n\bar{y} + (m-r)\left(t+\frac{1}{\theta^{(t)}}\right) + r\left(\frac{1}{\theta^{(t)}} - t h_t\right)\right]$$

得到EM算法迭代公式

(*)式：
$$E\left(\sum_{i=1}^{m} z_i \mid r, \theta^{(t)}, x\right) = \sum_{i=1}^{m} E(z_i \mid r, \theta^{(t)}, x)$$

$$= \sum_{i=1}^{r} E(z_i \mid z_i<t, x, \theta^{(t)}) + \sum_{i=r+1}^{m} E(z_i \mid z_i \geq t, x, \theta^{(t)})$$

$$= r\,\underset{Ⓐ}{\frac{\int_0^t z f(z,\theta^{(t)})\,dz}{\int_0^t f(z,\theta^{(t)})\,dz}} + (m-r)\,\underset{Ⓑ}{\frac{\int_t^{+\infty} z f(z,\theta^{(t)})\,dz}{\int_t^{+\infty} f(z,\theta^{(t)})\,dz}}$$

$$A = \frac{\int_0^t z\,\cancel{\theta^{(t)}} e^{-\theta^{(t)}z}\,dz}{\int_0^t \cancel{\theta^{(t)}} e^{-\theta^{(t)}z}\,dz} = \frac{\cancel{-\tfrac{1}{\theta^{(t)}}}\int_0^t z\,d e^{-\theta^{(t)}z}}{\int_0^t d e^{-\theta^{(t)}z}\cdot\left(\cancel{-\tfrac{1}{\theta^{(t)}}}\right)} = \frac{t e^{-\theta^{(t)}t} - \int_0^t e^{-\theta^{(t)}z}\,dz}{e^{-\theta^{(t)}t}-1}$$

$$\left(\int_0^t e^{-\theta^{(t)}z}\,dz = -\frac{1}{\theta^{(t)}}\left(e^{-t\theta^{(t)}}-1\right)\right) \qquad = \frac{t e^{-\theta^{(t)}t}}{e^{-\theta^{(t)}t}-1} + \frac{1}{\theta^{(t)}}$$

$$B = \frac{\int_t^{+\infty} z\,\cancel{\theta^{(t)}} e^{-\theta^{(t)}z}\,dz}{\int_t^{+\infty} \cancel{\theta^{(t)}} e^{-\theta^{(t)}z}\,dz} = \frac{\int_t^{+\infty} z\,d e^{-\theta^{(t)}z}}{\int_t^{+\infty} d e^{-\theta^{(t)}z}} = \frac{z e^{-\theta^{(t)}z}\Big|_t^{+\infty} - \int_t^{+\infty} e^{-\theta^{(t)}z}\,dz}{e^{-\theta^{(t)}z}\Big|_t^{+\infty}}$$

$$= \frac{0 - t e^{-\theta^{(t)}t} + \frac{1}{\theta^{(t)}}\left(0 - e^{-t\theta^{(t)}}\right)}{0 - e^{-\theta^{(t)}t}} = t + \frac{1}{\theta^{(t)}}$$

2). R code :

```r
##2
set.seed(101)
N<-1000
n<-600
m<-400
t<-5
r<-160
```

生成样本

```r
X1<-rexp(n,1/10)    #设定样本服从theta=10的指数分布
X2<-rexp(m,1/10)
X0<-c(X1,X2)
X0<-X0[sample(1:N)]
plot(density(X0))

theta0<-9
para<-theta0
tol<-1e-8
para.old<-para+1
###
```

进行迭代

```r
for(j in 1:1000){
  ht<-exp(-t*para)/(1-exp(-t*para))
  theta1<-(m+n)/(sum(X1)+(m-r)*(t+1/para)+r*(1/para-t*ht))
  para<-theta1

  if (sqrt(sum((para-para.old)^2))/sqrt(sum(para.old^2))<tol)
  {break}
  para.old<-para
}
list(estimate=para,iter=j,tol=tol)
```

```
> list(estimate=para,iter=j,tol=tol)
$estimate
[1] 0.1034652

$iter
[1] 14

$tol
[1] 1e-08
```

结论：estimate 即为得到的样本θ估计值.