R Notebook

Code ▾

1.Read the EM Tutorial on Blackboard. It can be found under Outline/Clustering. 2.You do not have to include the following derivations in the submitted homework ???le. 3. Following the steps in the tutorial, write R codes for: a. The initialization step. Note that the parameters for each cluster, λ1 and λ2 have to be positive.

Hide

```
#initiation
em_init_B <- function(lambda1=-2,lambda2=3){
  #If the user specified standard deviation is negative, make them positive
  lambda1 <- abs(lambda1)
  lambda2 <- abs(lambda2)
  return(c(lambda1,lambda2))
  }
```

b. The E-step. If you know what you are doing, this step should be straightforward. You just need to use the dpois function.

Hide

```
# Poisson distribution with parameter λ which has to be positive
#dpois gives the (log) density
#ppois gives the (log) distribution function
#qpois gives the quantile function
#rpois generates random deviates.
#Invalid lambda will result in return value NaN, with a warning.
em_e <- function(x,lambda1,lambda2){
  p1 <- dpois(x, lambda1, log = FALSE)
  p2 <- dpois(x, lambda2, log = FALSE)
  return(p1/(p1+p2)) #probability of variable is in cluster1
  }
```

c. The M-step. Use your result from the question 2.

Hide

```
em_m <- function(x,z){
  lambda1 <- sum(z*x)/sum(z)
  lambda2 <- sum((1-z)*x)/sum(1-z)
  return(c(lambda1,lambda2))
  }
```

4. Combine all of the code in one big R code. Your code should take in a vector of Poisson observations and return the parameters of the model (λ1,λ2) and the expected cluster assignment for each sample (πi,k).

Hide

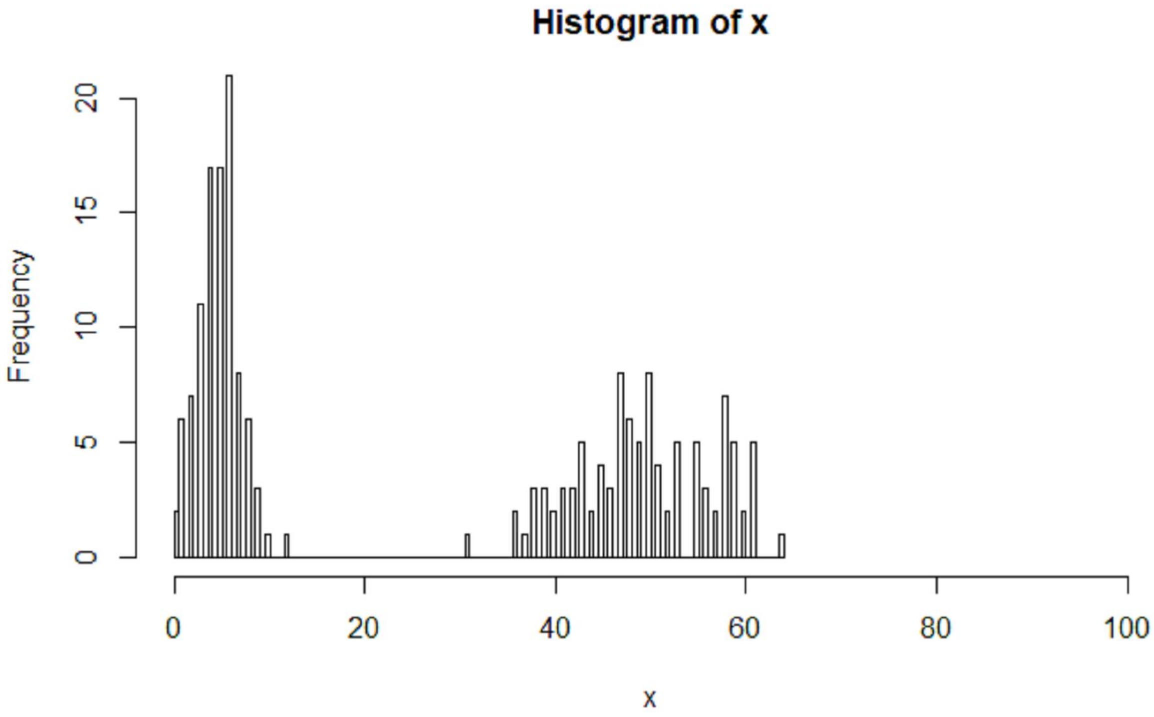
```
em_mix_gauss <- function(x,iter.max=100,conv.check=0){
init_param <- em_init_B() #initialize variables with the data driven method
lambda1 <- init_param[1]
lambda2 <- init_param[2]

previous_theta <- init_param
for(t in 1:iter.max){
e_result <- em_e(x,lambda1,lambda2) #E-step
m_result <- em_m(x,e_result) #M-step
lambda1 <- m_result[1]
lambda2 <- m_result[2]
#stop the algorithm if we achieved convergence
if(max(abs(m_result-previous_theta))<conv.check) break;
previous_theta <- m_result
}
return(list(z=e_result,theta=m_result))
}
```

5. Generate random Poisson samples with the following R code, and test how well your algorithm works in finding the clusters and the cluster parameters:

Hide

```
x1 <- rpois(100, 5)
x2 <- rpois(100, 50)
x <- c(x1,x2)
hist(x,xlim=c(0,100),breaks=200)
```



Hide

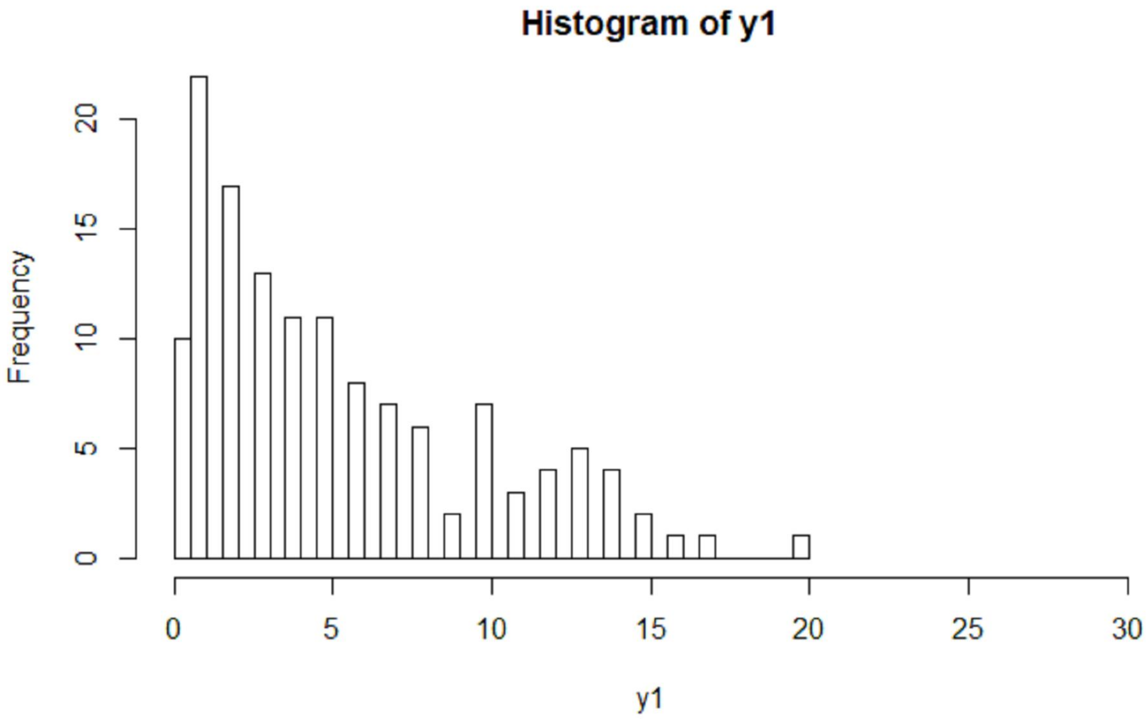
```
em_res <- em_mix_gauss(x)
em_res$theta
```

```
[1]  4.85 49.50
```

6. Download the DJI_vol.csv dataset from Blackboard. In this dataset, you will find the total number of volatile days that the Dow Jones Index had for each month. Here, the number of volatile days is defined as the number of days in which the absolute value of the daily return is higher than 1%.
7. Apply your EM clustering algorithm on the dataset. What are the cluster estimates? Finally, plot the cluster assignments in time. Are there any interesting trends? Summarize your findings in a couple of sentences.

Hide

```
data <- read.csv(file="C:/Users/zhang/Desktop/6240_r/DJI_vol.csv")
y <- data[2]
y1 <- as.numeric(unlist(y))
hist(y1,xlim=c(0,30),breaks=30)
```



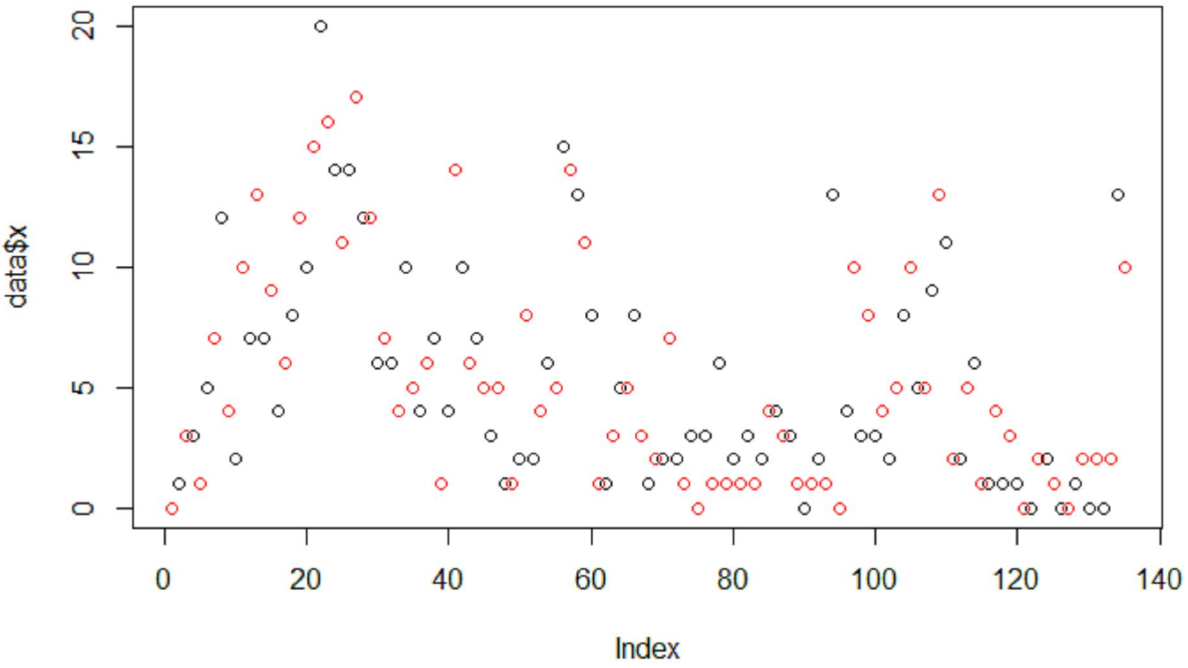
Hide

```
em_res <- em_mix_gauss(y1)
#cluster estimates
em_res$theta
```

```
[1] 2.126616 9.356977
```

Hide

```
plot(data$High_Vol_Days, data$x,col=em_res$theta)
```


data$x
Index