STAT 510 — Mathematical Statistics

Assignment: Problem Set 3 **Due Date:** September 30 2022, 11:59 PM

Note: The solution is provided by Mingxuan Cui, who has done the best job in HW3.

1. K-component Gaussian Mixture Model

$$f(x|\theta) = \sum_{k=1}^{K} \frac{\pi_k}{\sqrt{det(2\pi\Sigma_k)}} \exp[-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)],\, x \in \mathbb{R}^p$$

where $\theta = \{(\pi_k,\, \mu_k,\, \Sigma_k)_{k=1}^K\}$.

(a) E-Step.

Complete data log-likelihood is:

$$\begin{aligned} l(\theta|X,Z) &= \sum_{i=1}^{n} \log g(x_i, z_i|\theta) \\ &= \sum_{i=1}^{n} \log \left[\prod_{k=1}^{K} \left(\pi_k p_k(x_i|\mu_k,\Sigma_k)\right)^{z_{ik}} \right] \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\left(\log \pi_k + \log p_k(x_i|\mu_k,\Sigma_k)\right) \end{aligned}$$

where

$$p_k(x_i|\mu_k,\Sigma_k) = \frac{1}{\sqrt{det(2\pi\Sigma_k)}} \exp[-\frac{1}{2}(x_i-\mu_k)^T\Sigma_k^{-1}(x_i-\mu_k)],\, \forall 1 \le i \le n,\, 1 \le k \le K.$$

Since $Z = (z_{ik})_{1\le i\le n; 1\le k\le K}$ is latent variable and can't be observed, we will take expectation with respect to $Z \sim p(\cdot|X,\theta^{(t)})$ at iteration t. To be specific, we need compute the expected label $Z^{(t)} = (z_{ik}^{(t)})$ given X, $\theta^{(t)}$.

$$\begin{aligned} z_{ik}^{(t)} =& \mathbb{P}(z_{ik} = 1|X,\theta^{(t)}) \\ =& \frac{\mathbb{P}(x_i|z_{ik}=1,\theta^{(t)}) \cdot \mathbb{P}(z_{ik}=1|\theta^{(t)})}{\sum_{k=1}^K \mathbb{P}(x_i|z_{ik}=1,\theta^{(t)}) \cdot \mathbb{P}(z_{ik}=1|\theta^{(t)})} \\ =& \frac{p_k(x_i|\mu_k^{(t)},\Sigma_k^{(t)}) \cdot \pi_k^{(t)}}{\sum_{k=1}^K p_k(x_i|\mu_k^{(t)},\Sigma_k^{(t)}) \cdot \pi_k^{(t)}},\, \forall 1 \le i \le n, 1 \le k \le K. \end{aligned}$$

Then we have the expression of Q function:

$$\begin{aligned} Q(\theta|\theta^{(t)}) =& \mathbb{E}_Z[l(\theta|X,Z)|X,\theta^{(t)}] \\ =& \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}^{(t)} \log p_k(x_i|\mu_k,\Sigma_k) \\ =& \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}^{(t)} [\log \pi_k - \frac{1}{2}\log(2\pi) - \frac{1}{2}\log det(\Sigma_k) - \frac{1}{2}(x_i-\mu_k)^T\Sigma_k^{-1}(x_i-\mu_k)] \end{aligned}$$

(b) M-Step

Try to maximize Q over π_k, μ_k and Σ_k.

For π_k, because we have a constraint $\sum_{k=1}^{K} \pi_k = 1$, this is a constrained optimization problem and shoule be reformulated first. Replace π_K with $1 - \sum_{k=1}^{K-1} \pi_k$ and Q function becomes:

$$Q(\theta|\theta^{(t)}) = \sum_{i=1}^{n} \left(\sum_{k=1}^{K-1} z_{ik}^{(t)} \log \pi_k + z_{iK}^{(t)} \log \left(1 - \sum_{k=1}^{K-1} \pi_k \right) \right) + Q'(\mu_k, \Sigma_k|\theta^{(t)}),$$

where Q' is a function not containing θ. Then take derivative to all $\pi_k (1 \leq k \leq K-1)$:

$$\frac{\partial Q}{\partial \pi_k}(\theta|\theta^{(t)}) = \sum_{i=1}^{n} \left(z_{ik}^{(t)} \frac{1}{\pi_k} - z_{iK}^{(t)} \frac{1}{1 - \sum_{k=1}^{K-1} \pi_k} \right) = 0, \forall 1 \leq k \leq K-1.$$

This means $\frac{\sum_{i=1}^{n} z_{ik}^{(t)}}{\pi_k} = \lambda$ (λ is a constant) for each $1 \leq k \leq K-1$ and we have an equation

$$\lambda = \sum_{i=1}^{n} \frac{z_{iK}^{(t)}}{1 - \sum_{k=1}^{K-1} \pi_k} = \sum_{i=1}^{n} \frac{1 - \sum_{k=1}^{K-1} z_{ik}^{(t)}}{1 - \sum_{k=1}^{K-1} \pi_k} = \frac{n - \sum_{k=1}^{K-1} \lambda \pi_k}{1 - \sum_{k=1}^{K-1} \pi_k}$$

Thus $\lambda = n$ and

$$\pi_k^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} z_{ik}^{(t)}.$$

For μ_k and Σ_k, this is an unconstrained optimization problem and we can directly take derivative. For μ_k:

$$\begin{aligned} \frac{\partial Q}{\partial \mu_k}(\theta|\theta^{(t)}) &= \sum_{i=1}^{n} z_{ik}^{(t)} \cdot (-\frac{1}{2}) \cdot 2\Sigma_k^{-1}(x_i - \mu_k) \cdot (-1) \\ &= \sum_{i=1}^{n} z_{ik}^{(t)} \Sigma_k^{-1}(x_i - \mu_k) = 0 \end{aligned}$$

Then

$$\mu_k^{(t+1)} = \frac{\sum_{i=1}^{n} z_{ik}^{(t)} x_i}{\sum_{i=1}^{n} z_{ik}^{(t)}}$$

For Σ_k, the derivative is not easy to find. We also need some reformulation first.

$$\begin{aligned} Q(\theta|\theta^{(t)}) &= \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik}^{(t)} \left[\frac{1}{2} \log det(\Sigma_k^{-1}) - \frac{1}{2} Tr \left((x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k) \right) \right] + Q''(\pi_k, \mu_k|\theta^{(t)}) \\ &= \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik}^{(t)} \left[\frac{1}{2} \log det(\Sigma_k^{-1}) - \frac{1}{2} Tr \left((x_i - \mu_k)(x_i - \mu_k)^T \Sigma_k^{-1} \right) \right] + Q''(\pi_k, \mu_k|\theta^{(t)}) \\ &= F(\Sigma_k^{-1}|\theta^{(t)}) + Q''(\pi_k, \mu_k|\theta^{(t)}) \end{aligned}$$

where Q'' is a function not containing Σ_k. Because $\frac{\partial Q}{\partial \Sigma_k} = 0$ is equivalent to $\frac{\partial F}{\partial \Sigma_k^{-1}} = 0$, then we need to set:

$$\frac{\partial F}{\partial \Sigma_k^{-1}}(\theta|\theta^{(t)}) = \sum_{i=1}^{n} z_{ik}^{(t)} [\frac{1}{2}\Sigma_k - \frac{1}{2}(x_i - \mu_k)(x_i - \mu_k)^T] = 0$$

Then

$$\Sigma_k^{(t+1)} = \frac{\sum_{i=1}^{n} z_{ik}^{(t)} (x_i - \mu_k^{(t+1)})(x_i - \mu_k^{(t+1)})^T}{\sum_{i=1}^{n} z_{ik}^{(t)}}.$$

Now we have a summary of explicit expression of $\theta^{(t+1)}$:

$$\begin{cases} \pi_k^{(t+1)} &= \frac{1}{n}\sum_{i=1}^{n} z_{ik}^{(t)} \\ \mu_k^{(t+1)} &= \frac{\sum_{i=1}^{n} z_{ik}^{(t)} x_i}{\sum_{i=1}^{n} z_{ik}^{(t)}} \\ \Sigma_k^{(t+1)} &= \frac{\sum_{i=1}^{n} z_{ik}^{(t)} (x_i - \mu_k^{(t+1)})(x_i - \mu_k^{(t+1)})^T}{\sum_{i=1}^{n} z_{ik}^{(t)}} \end{cases}$$

(c) Implementation.

To save pages, I will move the code to the appendix.

(d) Data.

Implement the code to the given dataset, I find that this method is sensitive to the initial value of $Z = (Z_{ik})$. To show this, I will use two different initial values and observe the outcome.

```
data = read.table("iris_data.txt",header = T, sep = ",")
X = as.matrix(data)
K=3
n=150
Z1 = matrix(rep(0,n*K), nrow = n)
for(i in 1:(n/K)){Z1[(K*(i-1)+1):(K*i),] = diag(K)}
print(Z1[1:6,]) #First several rows of Z
```

```
##      [,1] [,2] [,3]
## [1,]    1    0    0
## [2,]    0    1    0
## [3,]    0    0    1
## [4,]    1    0    0
## [5,]    0    1    0
## [6,]    0    0    1
```

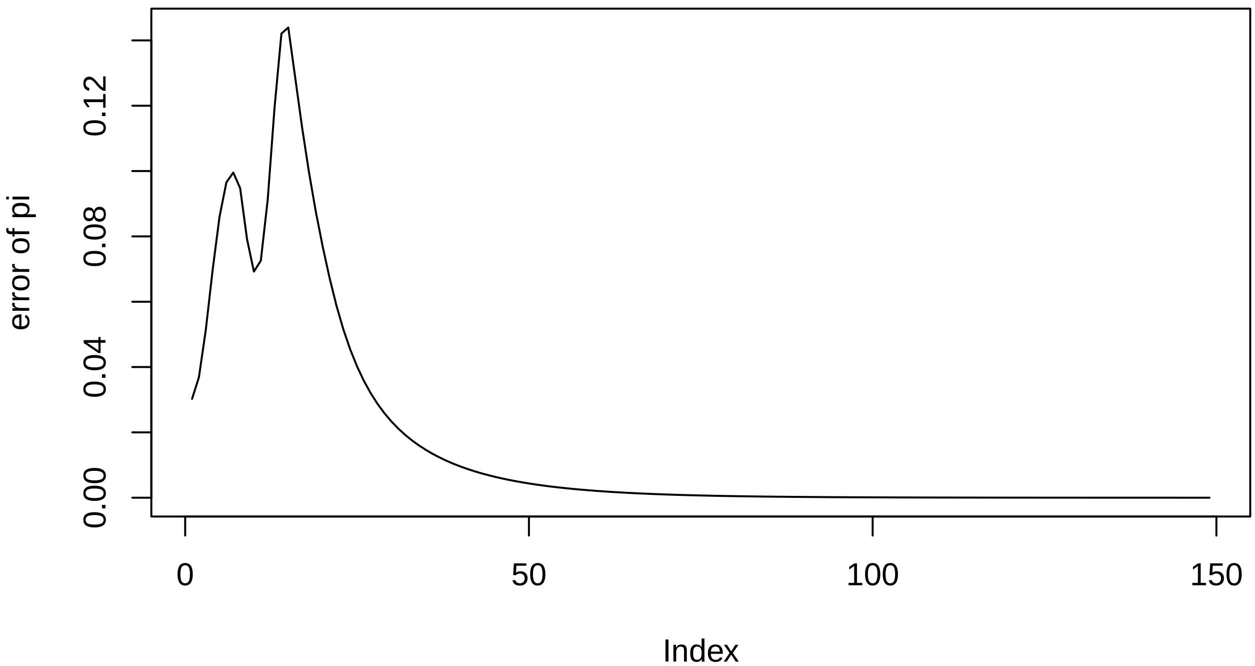
```
result1 = EM(X,Z1,3)
print(result1) #Number of iteration, pi, mu and sigma respectively
```

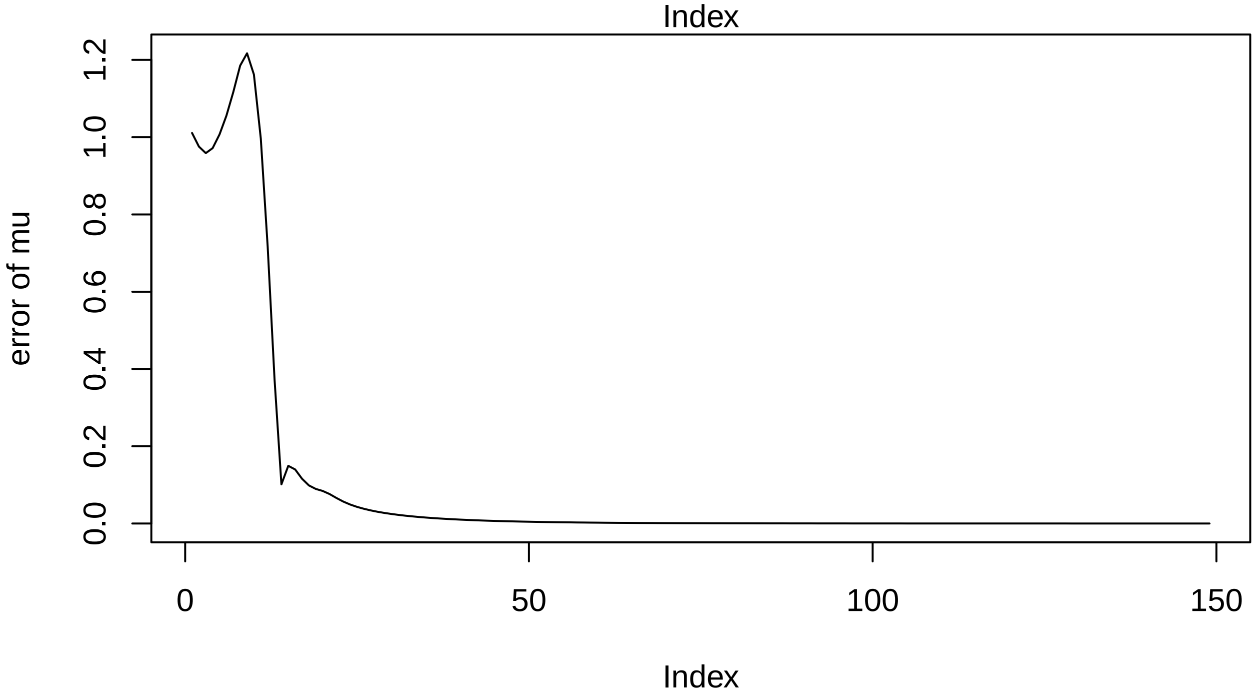
```
## [[1]]
## [1] 149
##
## [[2]]
## [1] 0.3331644 0.3544448 0.3123908
##
## [[3]]
## [[3]][[1]]
## sepal_length  sepal_width petal_length  petal_width
##    5.0062550    3.4185635    1.4640822    0.2439709
##
## [[3]][[2]]
## sepal_length  sepal_width petal_length  petal_width
##     6.232336     2.954583     5.103595     1.875528
##
## [[3]][[3]]
## sepal_length  sepal_width petal_length  petal_width
```

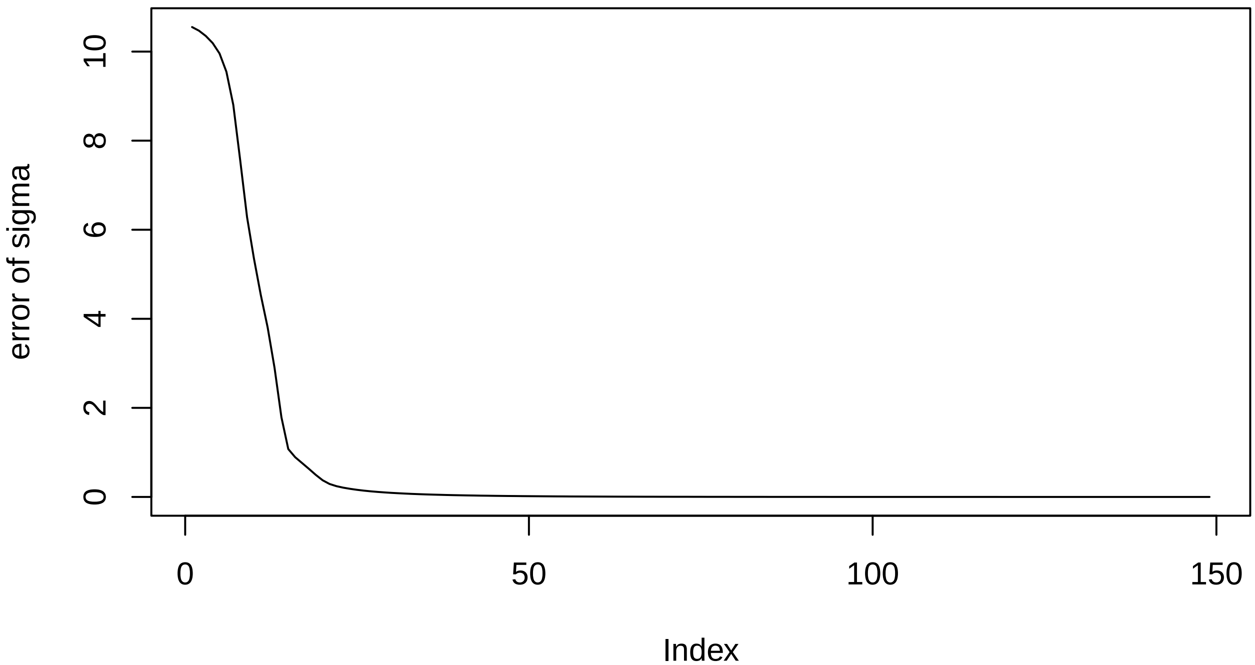
```
##     6.294707     2.777994     4.679856     1.448868
##
##
## [[4]]
## [[4]][[1]]
##      sepal_length sepal_width petal_length petal_width
## [1,]   0.12169658  0.09805657  0.015782142 0.010355502
## [2,]   0.09805657  0.14171785  0.011361356 0.011245226
## [3,]   0.01578214  0.01136136  0.029505301 0.005591309
## [4,]   0.01035550  0.01124523  0.005591309 0.011267830
##
## [[4]][[2]]
##      sepal_length sepal_width petal_length petal_width
## [1,]   0.27518836  0.06929879   0.22077599  0.12622201
## [2,]   0.06929879  0.06498699   0.06353632  0.05246808
## [3,]   0.22077599  0.06353632   0.30758577  0.17812752
## [4,]   0.12622201  0.05246808   0.17812752  0.15552869
##
## [[4]][[3]]
##      sepal_length sepal_width petal_length petal_width
## [1,]    0.6156021  0.18585293    0.7248133  0.22540274
## [2,]    0.1858529  0.14379533    0.1909303  0.06992094
## [3,]    0.7248133  0.19093026    1.0025906  0.31448807
## [4,]    0.2254027  0.06992094    0.3144881  0.10899371
```

The convergence of parameters can be shown in the following figures:

```
GMM_convergence(X,Z1,3,result1)
```







Now change the initial value

```
Z2 = matrix(rep(0,n*K), nrow = n)
for(i in 1:(n/K)){Z2[(K*(i-1)+1):(K*i),] = matrix(c(1/2,1/3,0,1/2,1/3,0,0,1/3,1), nrow = 3)}
print(Z2[1:6,]) #First several rows of Z
```

```
##           [,1]      [,2]      [,3]
## [1,] 0.5000000 0.5000000 0.0000000
## [2,] 0.3333333 0.3333333 0.3333333
## [3,] 0.0000000 0.0000000 1.0000000
## [4,] 0.5000000 0.5000000 0.0000000
## [5,] 0.3333333 0.3333333 0.3333333
## [6,] 0.0000000 0.0000000 1.0000000
```

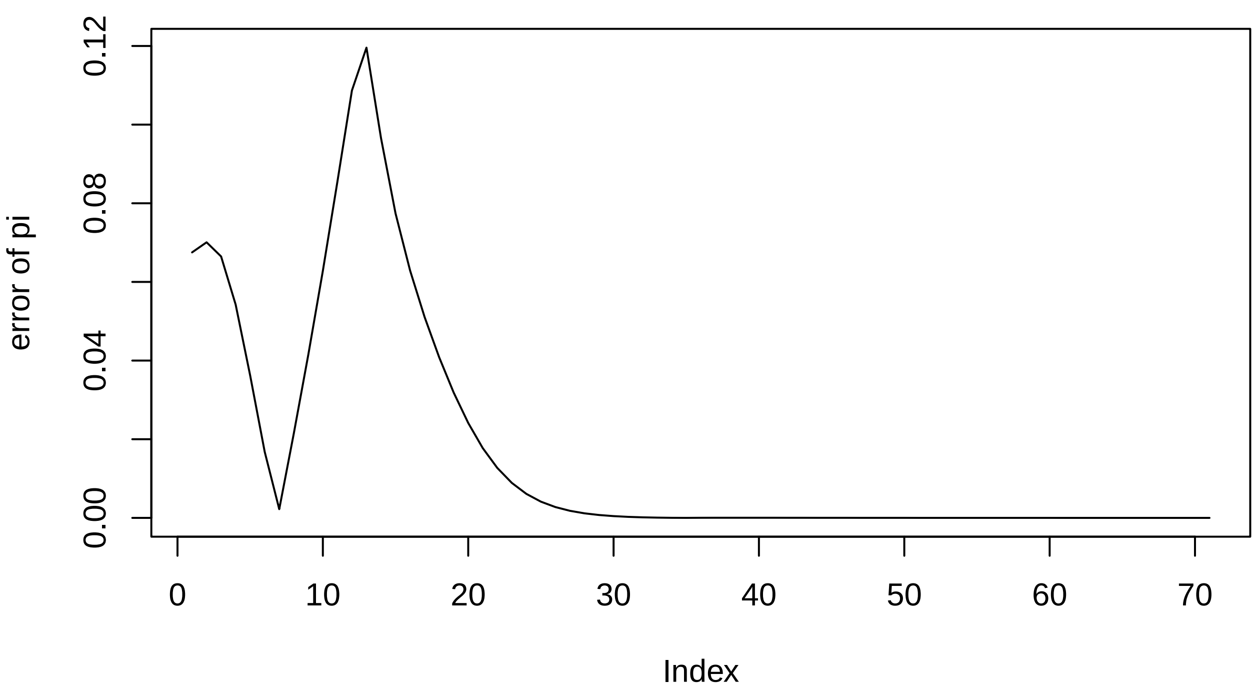
```
result2 = EM(X,Z2,3)
print(result2) #Number of iteration, $\pi_k$, $\mu_k$ and $\sigma_k$ respectively
```

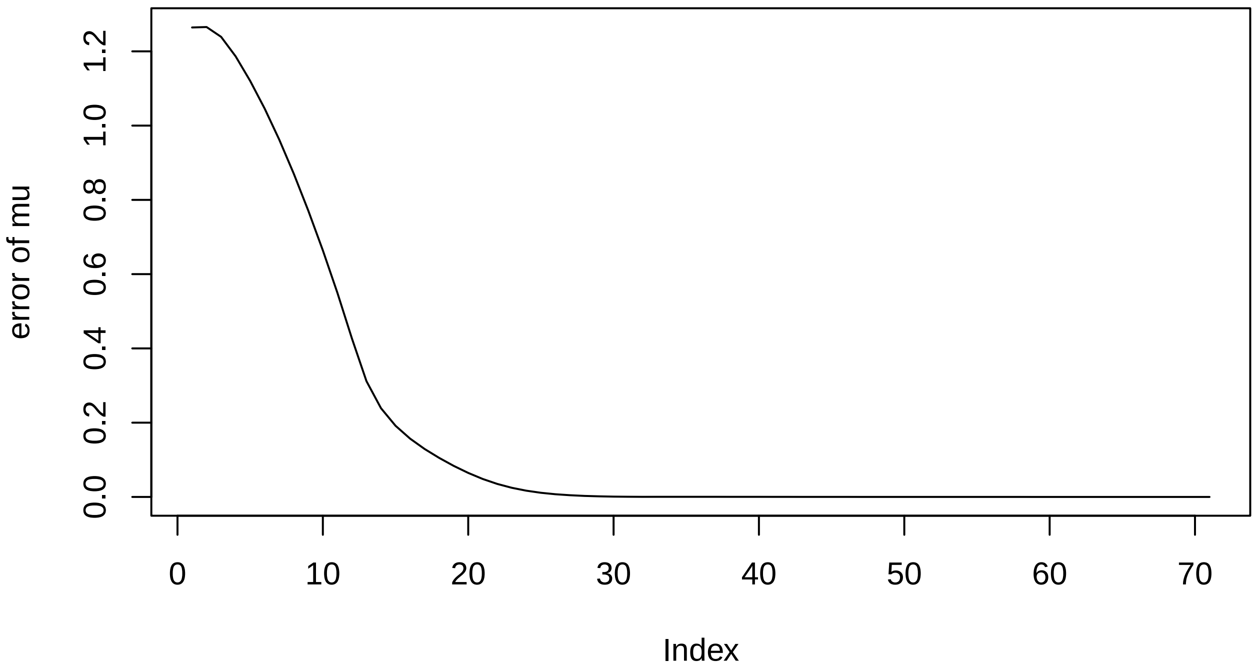
```
## [[1]]
## [1] 71
##
## [[2]]
## [1] 0.2509852 0.2509852 0.4980296
##
## [[3]]
## [[3]][[1]]
## sepal_length  sepal_width petal_length  petal_width
##     6.105250     2.868403     4.800285     1.707349
##
## [[3]][[2]]
## sepal_length  sepal_width petal_length  petal_width
##     6.105250     2.868403     4.800285     1.707349
##
## [[3]][[3]]
## sepal_length  sepal_width petal_length  petal_width
```

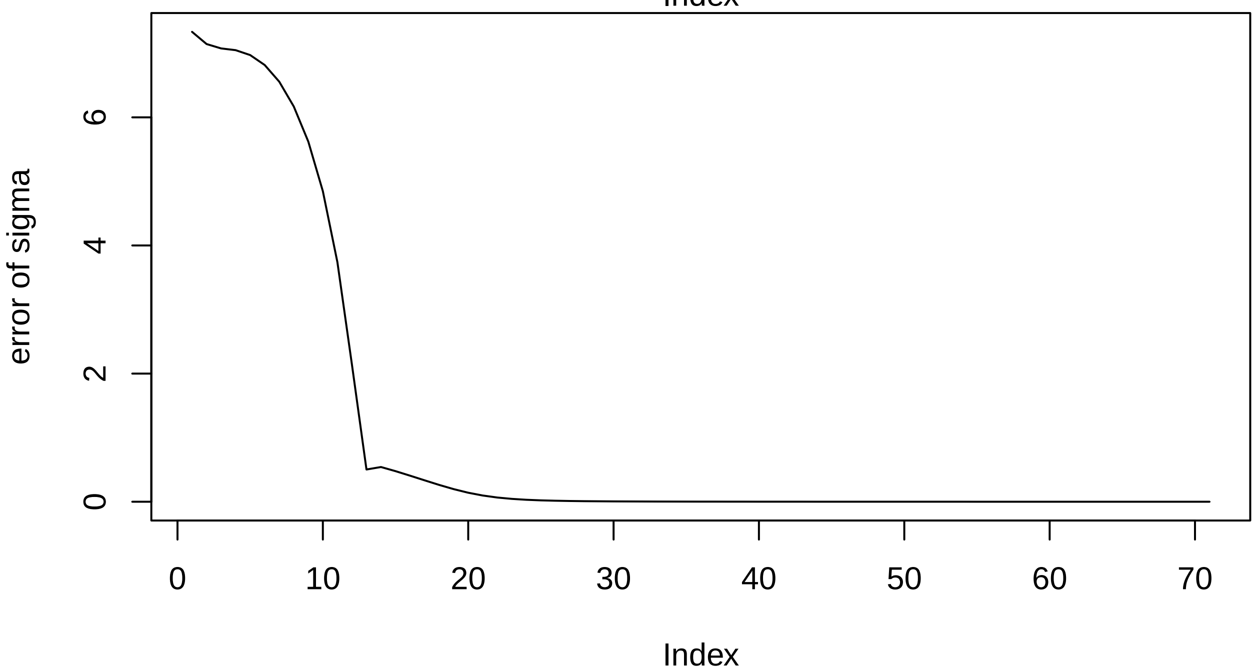
```
##    5.5793446    3.2410658    2.7088062    0.6859596
##
##
## [[4]]
## [[4]][[1]]
##      sepal_length sepal_width petal_length petal_width
## [1,]    0.3159792  0.10662102    0.3259702  0.16795076
## [2,]    0.1066210  0.09998979    0.1425304  0.08840541
## [3,]    0.3259702  0.14253036    0.5215954  0.28407311
## [4,]    0.1679508  0.08840541    0.2840731  0.19332872
##
## [[4]][[2]]
##      sepal_length sepal_width petal_length petal_width
## [1,]    0.3159792  0.10662102    0.3259702  0.16795076
## [2,]    0.1066210  0.09998979    0.1425304  0.08840541
## [3,]    0.3259702  0.14253036    0.5215954  0.28407311
## [4,]    0.1679508  0.08840541    0.2840731  0.19332872
##
## [[4]][[3]]
##      sepal_length sepal_width petal_length petal_width
## [1,]    0.9103215  -0.0874078    1.6597169   0.5920634
## [2,]   -0.0874078   0.2044855   -0.3940786  -0.1333546
## [3,]    1.6597169  -0.3940786    3.4878357   1.2270425
## [4,]    0.5920634  -0.1333546    1.2270425   0.4431090
```

The convergence plot is:

```
GMM_convergence(X,Z2,3,result2)
```




error of mu
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
10
20
30
40
50
60
70
Index


error of sigma
0
2
4
6
0
10
20
30
40
50
60
70
Index

2. Missing data problem

$X = (X_1, X_2, X_3)$ be a (three-category) multinomial random variable with probability mass function

$$f(x|\theta) = \frac{(x_1 + x_2 + x_3)!}{x_1!x_2!x_3!}(\frac{4+\theta}{6})^{x_1}(\frac{1-\theta}{3})^{x_2}(\frac{\theta}{6})^{x_3},\ x = (x_1, x_2, x_3),$$

where $\theta \in (0, 1)$ is the unknown parameter.

(a). EM algorithm

I'll try to further divide the first category into two categories X_{11} and X_{12}, with probabilities $\frac{2}{3}$ and $\frac{\theta}{6}$, so the model became

$$\text{Multi}(\frac{2}{3}, \frac{\theta}{6}, \frac{1-\theta}{3}, \frac{\theta}{6})$$

and the new probability mass function will be

$$g(x|\theta) = \frac{(x_{11} + x_{12} + x_2 + x_3)!}{x_{11}!x_{12}!x_2!x_3!}(\frac{2}{3})^{x_{11}}(\frac{\theta}{6})^{x_{12}}(\frac{1-\theta}{3})^{x_2}(\frac{\theta}{6})^{x_3},\ \ x = (x_{11}, x_{12}, x_2, x_3).$$

Then the complete log-likelihood function is

$$l(\theta|x) = (x_{12} + x_3)\log\theta + x_2\log(1-\theta).$$

expressions unrelated to θ can be treated as constant and has been removed from function l.

Our goal is compute the Q-function

$$Q(\theta|\theta^{(t)}) = \mathbb{E}_{\theta(t)} l(\theta|x).$$

and try to maximize it.

E-Step

Since $\theta^{(t)}$ only affect the value of x_{12}, we just need to compute the expected value of x_{12} given $\theta^{(t)}$. What's more, the distribution of x_{12} is binomial distribution $b(x_1, \frac{\theta}{4+\theta})$, so we can easily write down the expectation:

$$x_{12}^{(t)} = \mathbb{E}_{\theta^{(t)}}(x_{12}) = \frac{\theta x_1}{4+\theta}.$$

M-Step

Take the value $x_{12}^{(t)}$ into the expression of $l(\theta|x)$, we have

$$Q(\theta|\theta^{(t)}) = (x_{12}^{(t)} + x_3)\log\theta + x_2\log(1-\theta)$$

Take first derivative of Q with respect to θ and set it to 0:

$$\frac{\partial Q}{\partial\theta}(\theta|\theta(t)) = \frac{x_{12}^{(t)} + x_3}{\theta} - \frac{x_2}{1-\theta} = 0,$$

we have the maximal point as the new parameter:

$$\theta^{(t+1)} = \frac{x_{12}^{(t)} + x_3}{x_{12}^{(t)} + x_2 + x_3}$$

where $x_{12}^{(t)}$ is given before.

(b). Inplement

```r
E <- function(theta,x1){
  return(x1*theta/(4+theta))
}

M <- function(x12,x2,x3){
  return((x12+x3)/(x12+x2+x3))
}

EM <- function(x1,x2,x3){
  theta0 = 0
  theta1 = theta0
  ite = 0
  #Initialization
  x11 = 0
  x12 = x1 - x11

  while (TRUE) {
    x12 = E(theta0,x1)
    theta1 = M(x12,x2,x3)
    if(abs(theta1 - theta0)<1e-6) break
    theta0 = theta1
    ite = ite + 1
  }

  return(c(theta1, ite))
}
```

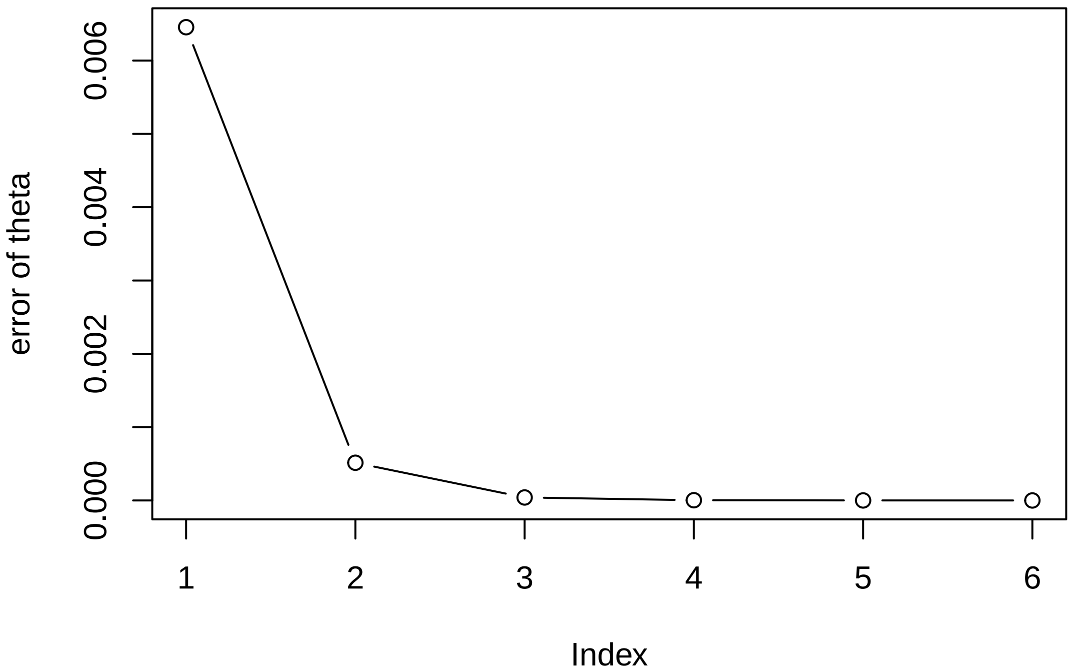
Implement the code to data:

```r
result =  EM(42,10,15)
theta = result[[1]]
ite = result[[2]]
print(theta)
```

```
## [1] 0.6783523
```

Now I want to rerun the function to show the convergence of $\{\theta^{(t)}\}_{t>0}$:

```r
EM_convergence(42,10,15,ite,theta)
```


error of theta
0.000
0.002
0.004
0.006
1
2
3
4
5
6
Index

Appendix

EM algorithm in problem 1.

```
normal <- function(x,mu,sigma){
  return (1/sqrt(det(2*pi*sigma))*exp(-1/2*t(x-mu) %*% solve(sigma) %*%(x-mu)))
}

EM <- function(X,Z0,K){
n = nrow(X) #Number of data points
p = ncol(X) #Number of features
Z = Z0 #Matrix to store $Z_{ik}$
#Parameters from last iteration
pi0 = rep(1/K, K)
mu0 = list()
for(k in 1:K) {mu0[[k]] = rep(0,p)}
sigma0 = list()
for(k in 1:K) {sigma0[[k]] = diag(p)}
#Parameters to be renewed
pi1 = pi0
mu1 = mu0
sigma1=sigma0

ite = 0

while(TRUE){
  for(k in 1:K) {pi1[k] = 1/n*sum(Z[,k])}
  for(k in 1:K) {
    sum = rep(0,p)
    for (i in 1:n) {
      sum = sum + Z[i,k] * X[i,]
    }
    if (sum(Z[,k]) == 0) {mu1[[k]]=rep(0,p)}
    else {mu1[[k]] = sum/sum(Z[,k])}
  }

  for(k in 1:K){
    sum = matrix(rep(0,p*p), nrow = p)
      for(i in 1:n){
        sum = sum + Z[i,k] * (X[i,]-mu1[[k]]) %*% t((X[i,]-mu1[[k]]))
      }
    if(sum(Z[,k]) == 0) {sigma1[[k]] = matrix(rep(0,p*p, nrow = p))}
    else {sigma1[[k]] = sum / sum(Z[,k])}
  }

  a = norm(pi1-pi0, type = "2")
  b = sum(norm(mu1[[k]]-mu0[[k]], type = "2"))
  c = 0
  for (k in 1:K) {
    c = c + norm(sigma1[[k]] - sigma0[[k]], type = "2")
  }

  if(max(a,b,c)<1e-6) break
```

```
  for(i in 1:n){
    sum = 0
    for(k in 1:K){
      sum = sum + normal(X[i,],mu1[[k]], sigma1[[k]]) * pi1[k]
    }
    for(k in 1:K) {Z[i,k] = normal(X[i,],mu1[[k]], sigma1[[k]]) * pi1[k] / sum}
  }

  pi0 = pi1
  mu0 = mu1
  sigma0 = sigma1
  ite = ite + 1

}
  result = list(ite, pi1, mu1, sigma1)
  return(result)
}
```