UNIVERSIDADE FEDERAL DO AMAZONAS
PRO REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
DEPARTAMENTO DE APOIO A PESQUISA
PROGRAMA INSTITUCIONAL DE INICIAÇÃO CIENTÍFICA

ALGORITMO E.M. EM ESTIMAÇÃO DE DENSIDADES.

Bolsista: Débora Pinto Pessoa, CNPq.

MANAUS
2009

UNIVERSIDADE FEDERAL DO AMAZONAS
PRO REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
DEPARTAMENTO DE APOIO A PESQUISA
PROGRAMA INSTITUCIONAL DE INICIAÇÃO CIENTÍFICA

RELATÓRIO FINAL
PIB – E – 0089/2008
ALGORITMO E.M. EM ESTIMAÇÃO DE DENSIDADES.

Bolsista: Débora Pinto Pessoa, CNPq.
Orientadora: PROFª. DRª. Amazoneida Sá Peixoto Pinheiro

MANAUS
2009

ALGORITMO E.M. EM ESTIMAÇÃO DE DENSIDADES

Débora Pinto Pessoa*
Amazoneida Sá Peixoto Pinheiro**

RESUMO

O estudo da distribuição normal é muito importante na estatística, pela sua característica de convergência das demais distribuições, muitas distribuições podem ser modeladas por misturas de normais.

O problema na estimação dos parâmetros no modelo de mistura de normais apresenta uma forma não convencional onde os métodos estatísticos convencionais não são suficientes, pois os cálculos se tornam muito complexos.

O algoritmo E.M. (*Expectation-Maximization*) é uma forma computacional que tem sido muito usada para solucionar numericamente o problema do cálculo de estimadores, pois ele maximiza a função de verossimilhança encontrando o Estimador de Máxima Verossimilhança (EMV). Dentro deste contexto este trabalho faz o estudo do algoritmo E.M. para mistura finita de normais, com o objetivo de verificar a qualidade das estimativas geradas pelo E.M., quando dois critérios são usados para inicialização do algoritmo: pelo método dos momentos e pelo método aleatório.

SUMÁRIO

1. INTRODUÇÃO

Em muitas atividades do cotidiano é necessário efetuar a classificação de objetos a classes previamente definidas, em que o objeto em questão é tudo aquilo que se deseja classificar, seja pessoa, imagens, plantas ou outro objeto que se pode classificar. A necessidade de classificação surge porque a população está misturada, os dados são distintos e, assim não se sabe como discriminá-las.

Exemplo: Em um diagnóstico médico classificar um paciente como portador de uma doença ou não; classificar uma imagem de raios-X como sendo uma classe de tumor maligno ou não.

A modelagem estatística desses problemas envolve muitas vezes misturas de densidades e estimar parâmetros de mistura de densidades é um procedimento complexo, uma vez que o objetivo é encontrar uma solução para o problema de complexidade das densidades misturadas e o Algoritmo E.M. é um método viável para esta finalidade.Este trabalho focará alguns modelos de misturas de normais com uma tentativa de explicar por simulação os exemplos postos acima.

Objetivo deste trabalho:

(i). Compreender o algoritmo E.M., suas potencialidades e restrições;

(ii). Estudar e implementar o algoritmo E.M. para mistura de normais;

(iii). Simular e analisar a qualidade das estimativas dentro da teoria estatística, diante de diferentes métodos de inicialização do E.M.

2. REVISÃO BIBLIOGRÁFICA

2.1. Misturas de densidades

Misturas são úteis para modelar dados heterogêneos, quando se sabe que as observações pertencem a um número finito de populações distintas, mas não sabemos como discriminá-las.

O estudo de dados através de misturas é mais expressivo, quando se encontra a função que expressa o verdadeiro comportamento dos dados as estimativas se tornam mais exatas, ou seja, quando aproximamos os dados para uma distribuição que não é a do seu modelo de origem, perdemos exatidão e o erro aleatório cresce e os estimadores não serão tão eficientes, pois estarão baseados em uma distribuição aproximada.

As misturas podem ser denotadas, segundo Mclachlan & Pell (2000), como uma soma ponderada de densidades.

$$\begin{aligned} f(x/\theta) &= \sum_{i=1}^{m} \pi_i f_i(x/\theta) \\ f(x) &= \pi_1 f_1(x) + ... + \pi_m f_m(x) \\ \sum_{i=1}^{m} \pi_i &= 1 \end{aligned} \qquad (1)$$

Onde,

m = número de componentes da mistura.

f_i = densidade da i-ésima população.

π_i = pesos ou proporção de cada densidade na mistura.

θ = vetor de parâmetros.

2.2 Misturas de Normais

A mistura de densidades de Normais, usando a expressão (1) pode ser escrita como:

$$f(x) = \pi_1 * (2\pi\sigma)^{-\frac{1}{2}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} + ... + \pi_m * (2\pi\sigma)^{-\frac{1}{2}} e^{-\frac{(x-\mu_m)^2}{2\sigma_m^2}} \qquad (2)$$

A figura 1, abaixo, apresentada simulações de misturas de 2 e 3 Normais. Nota-se que, nem sempre as misturas serão visíveis graficamente, os gráficos 1 e 2 são diferentes porém, são misturas de duas Normais, já o gráfico 3 é uma mistura de três Normais.

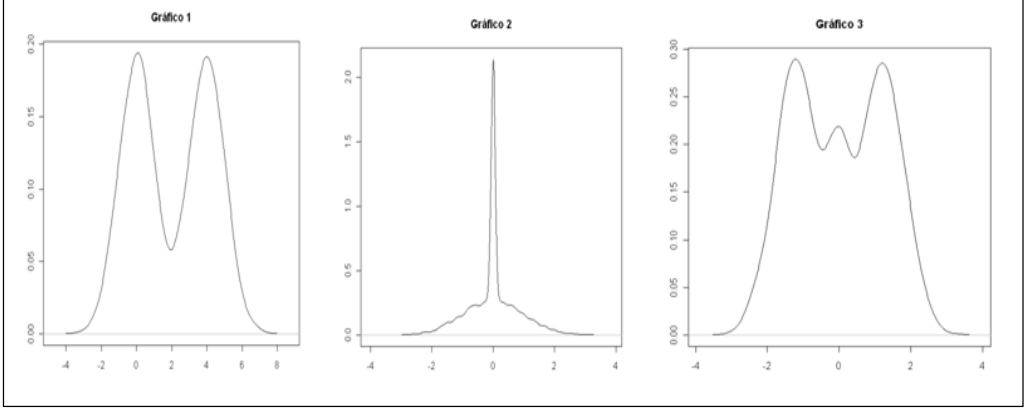

Figura 1: Mistura de duas e três Normais **Fonte: Simulação**

2.3 Estimadores de Máxima Verossimilhança

Parâmetros são medidas populacionais (Ex. Média, variância), e é papel da Inferência estudar medidas baseadas em uma amostra que expressem um mesmo comportamento dos parâmetros. A estimação de parâmetros é uma parte muito importante na teoria de inferência Estatística.

Definição: Seja θ uma medida da população, então qualquer estatística que assuma valores em Θ (conjunto em que θ toma seus valores) é um estimador para θ (Bolfarine, Heleno).

Para se achar o ponto $\hat{\theta}$ que maximiza verossimilhança é necessário derivar o logaritmo da mesma igualando-a a zero a raiz dessa equação será o estimador de máxima verossimilhança (EMV).

Esse método consiste em encontrar o ponto $\hat{\theta}$ que maximize a função de verossimilhança, derivando o logaritmo e igualando-a a zero encontrando a raiz da equação.

Baseando-se no estudo matemático de pontos críticos, máximos globais e locais encontraremos o máximo de $L(\theta; X)$.

$$\frac{\partial \log(\theta; X)}{\partial \theta} = 0 \tag{3}$$

Se a desigualdade abaixo for satisfeita então será o EMV (Estimador de Máxima Verossimilhança) de θ.

$$\frac{\partial^2 \log(\theta; X)}{\partial \theta^2} < 0 \quad \textit{quando} \;\; \theta = \hat{\theta} \tag{4}$$

3. MÉTODOS UTILIZADOS

3.1 Algoritmo E.M.

Estimar parâmetros de misturas de densidades é um procedimento complexo em que os métodos estatísticos tradicionais não são suficientes. Para tanto, um caminho estudado por Mclachlan (2000), é a estimação via algoritmo E.M.

O algoritmo E.M. é uma forma computacional que maximiza a função de verossimilhança, encontrando os Estimadores de Máxima Verossimilhança (EMV) em misturas de densidades.

O algoritmo E.M. é constituído de dois passos:

- Passo 1.

  E-Step

  Cálculo da Esperança do log da verossimilhança da densidade misturada.

  $$E[\sum_{i=1}^{n}\log(\sum_{i=1}^{m}\pi_i f_j(X_i/\theta_j))]$$

- Passo 2.

  M-Step

  Maximização da função.

  $$\arg\max E[\sum_{i=1}^{n}\log\pi_i f_j(x_i/\theta_j)]$$

No modelo de misturas de normais, quanto mais sub-populações pertencerem à mistura mais complexa será a verossimilhança e maior será o vetor de parâmetros a ser estimado.

No estudo de simulação, estudou-se 2 modelos com 2 e 3 componentes normais, sendo geradas 1000 amostras com 1000 iterações, com a condição de parada, isto é, o salto de uma iteração para outra é de 0.00001.

Os programas foram escritos na linguagem do software estatístico R, e estão em anexo neste trabalho.

4. Resultados e Discussões

As estimativas obtidas com o E.M. dependem muito dos valores inicias indicados no algoritmo. Nas simulações foram focados dois métodos para obtenção dos valores iniciais.

1. Valores iniciais Aleatórios

O método dos valores iniciais aleatórios consiste em da valores iniciais aleatoriamente da forma que acharmos melhor, sem se basear em nenhuma regra, esse método é rústico pois como os valores iniciais podem ser valores muito afastados do valor real a convergência do algoritmo fica prejudicada e as estimativas muito ruins.

2. Método dos momentos

O método dos momento é de cunho mais estatístico pois se baseia em medidas amostrais das populações misturadas, utilizando métodos de analise discriminante agrupando as observações e classificando-as com um certo erro em cada grupo (população), essa classificação é feita no software R através da função k-means.

O tamanho das populações foi extraído de partições de um conjunto de 1000 observações, os tamanhos das populações da mistura variam dependendo do peso de cada população.

Mistura de duas Normais - Medias afastadas							
		μ_0	μ_1	σ_0	σ_1	π_0	π_1
Valores iniciais aleatórios	Verdadeiro valor dos parâmetros	0	12	1	2	0,5	0,5
	valores iniciais	0,5	11,5	0,5	1,5	0,4	0,6
	estimativas geradas	-0,0659	11,83392	0,984061	1,89712	0,499856	0,500144
Método dos momentos	Verdadeiro valor dos parâmetros	0	12	1	2	0,5	0,5
	valores iniciais	-0,06484	11,83629	0,987704	2,15814	0,5	0,5
	estimativas geradas	-0,03932	12,02212	1,086164	2,071893	0,499999	0,500001

Tabela 1: Comparação de Estimativas para uma Normal misturada m=2 **Fonte: Simulação**

O uso de valores iniciais aleatórios é importante quando acompanhado de certa experiência quanto a população misturada, diante desse pensamento compara-se o método dos valores aleatórios como o método dos momentos. Nota-se pela tabela 1 que as estimativas geradas usando valores iniciais do

método dos momentos foram um pouco melhores que as do método dos valores inicias, porém diante da inconstância do método aleatório verifica-se que o método dos momentos é melhor.

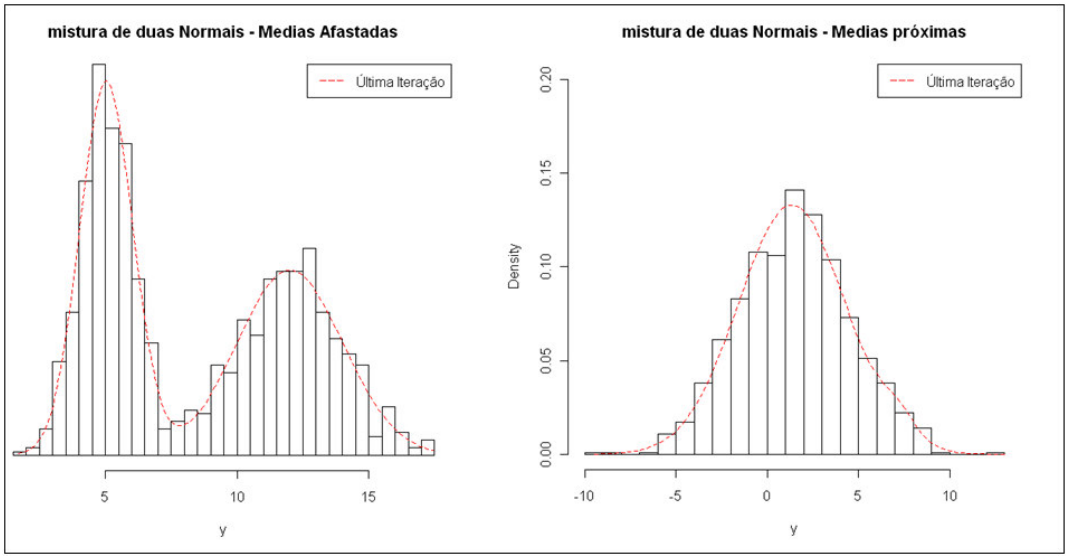

Figura 2: Mistura de duas Normais, medias afastadas e médias próximas **Fonte: Simulação**

Na figura 2 nota-se a falta de visualização da mistura quando as médias estão próximas e as dificuldades não ficam apenas na visualização como já foi dito, as estimativas também ficam menos eficientes, a linha vermelha diz respeito a estimação da densidade através dos parâmetros finais encontrados via algoritmo E.M.

Mistura de duas Normais - medias próximas							
		μ_0	μ_1	σ_0	σ_1	π_0	π_1
Valores iniciais aleatórios	Verdadeiro valor dos parâmetros	1	2	2	3	0,5	0,5
	Valores iniciais	0,7	1,8	2,5	3,2	0,4	0,6
	Estimativas geradas	1,20480	1,9876	3,8793	4,5303	0,5239	0,4760
Método dos momentos	Verdadeiro valor dos parâmetros	1	2	2	3	0,5	0,5
	Valores iniciais	-0,2287	3,8676	2,6737	2,4520	0,432	0,568
	Estimativas geradas	1,002695	2,09777	4,91303	5,07996	0,62713	0,37286

Tabela 2: Mistura de duas Normais - Medias próximas **Fonte: Simulação**

Quanto maior a proximidade das médias menos exatas serão as

estimativas geradas como verifica-se na tabela 2.

Mistura de três Normais - Medias distantes										
		μ_0	μ_1	μ_3	σ_0	σ_1	σ_2	π_0	π_1	π_2
Valores iniciais aleatórios	Verdadeiro valor dos parâmetros	-5	0	10	1	1	4	0,5	0,2	0,3
	Valores iniciais	4,5	0,8	10,6	2	0,7	3,3	0,6	0,1	0,3
	estimativas geradas	-5,018	-0,048	10,045	0,854	1,043	4,130	0,501	0,197	0,301
Método dos momentos	Verdadeiro valor dos parâmetros	-5	0	10	1	1	4	0,5	0,2	0,3
	valores iniciais	-4,941	1,001	11,204	0,999	2,170	3,203	0,505	0,251	0,244
	Estimativas geradas	-4,944	0,085	10,120	1,084	0,946	4,056	0,503	0,201	0,295

Verificou-se na tabela 1 que quando as medias das populações estão distantes a discriminação é mais fácil e as estimativas dos parâmetros mais exatas comparando um modelo de misturas onde as médias são próximas, tanto para um mistura com m=2 quanto com m=3, porém comparando a tabela3 com a tabela 1 nota-se que quanto maior a quantidade de componentes na mistura maior serão a quantidade de parâmetros e menos exatas as estimativas.

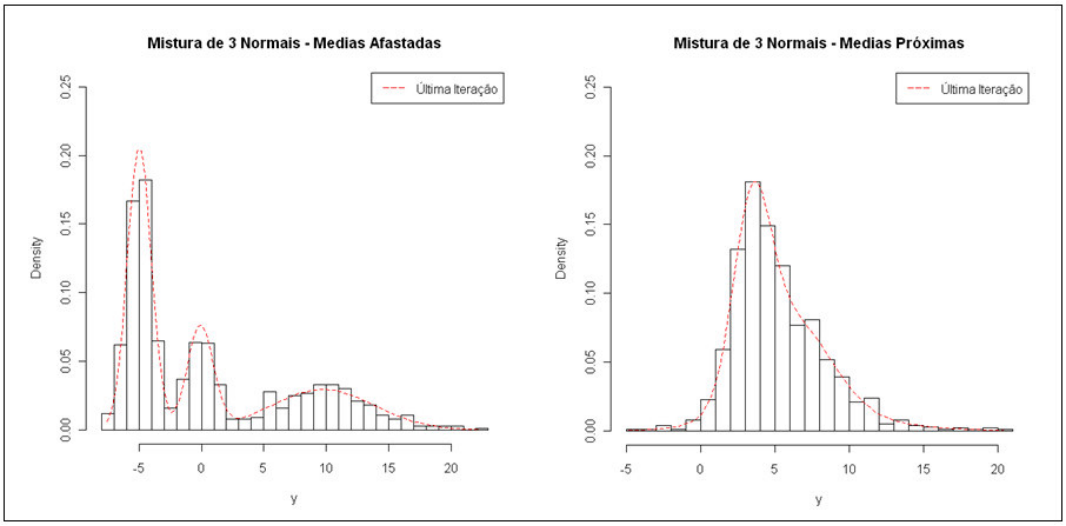

Figura 3: Mistura de três Normais **Fonte: Simulação**

A figura 3 mostra a densidade de uma mistura com três normais, verifica-se que a falta de visualização é um problema também para medias próximas com mistura de três normais.

Mistura de três Normais - Medias próximas										
		μ_0	μ_1	μ_3	σ_0	σ_1	σ_2	π_0	π_1	π_2
Valores iniciais aleatórios	Verdadeiro valor dos parâmetros	5	3	7	2	1	4	0,5	0,2	0,3
	Valores iniciais	4,5	3,8	7,6	2.7	0,7	3,3	0,3	0,3	0,4
	Estimativas geradas	5,908	3,117	6,639	1,867	1,079	4,062	0,391	0,308	0,300
Método dos momentos	Verdadeiro valor dos parâmetros	5	3	7	2	1	4	0,5	0,2	0,3
	Valores iniciais	6,370	2,879	10,742	1,072	1,343	2,016	0,398	0,475	0,127
	Estimativas geradas	5,952	3,452	7,811	2,721	1,23	5,19	0,499	0,388	0,112

Tabela 3: Mistura de três Normais - Medias próximas **Fonte: Simulação**

Como na mistura com duas populações, a mistura para três também tem o problema de medias próximas e isso existirá para m populações também, o acréscimo dessa classe de mistura é que as estimativas vão ficando cada vez mais imprecisas quando m cresce e tende para o infinito, porem para m=3 vê-se que os melhores valores iniciais são aqueles obtidos através do método dos momentos.

5. CONCLUSÃO

Conclui-se que na realização do presente estudo, verificou-se que o algoritmo EM é uma ótima ferramenta para estimar parâmetros de misturas de densidades normais. Diante das simulações, observou-se:

(i). As estimativas são boas, apesar de ficarem pouco afastadas do verdadeiro valor do parâmetro quando a quantidade de misturas no modelo cresce;

(ii). Os valores gerados para as estimativas se afastam do valor verdadeiro quando as medias das misturas são próximas e os desvios altos;

(iii). Na simulação de modelos misturados com distribuição normal, percebeu-se que o algoritmo convergiu rapidamente, apesar das observações acima;

(iv). A inicialização do EM com os momentos gerou estimativas melhores, como demonstrado por (MARQUES,L.A.V , 2009).

6. FONTES E REFERÊNCIAS BIBLIOGRÁFICAS

MCLACHLAN,G. & PELL, D –Finites mixtures models – Wiley & Sons, Nova York, 2000. 419p.

REDNER, A. R. & WALKAR, H. F.-Mixtures densities, maximum likelihood and the algorithm.Society for Industrial Aappplied Mathematics Review. (26), 195-329. 1984

PEREIRA, J. R. G. - Mistura finita de densidades com aplicações em reconhecimento de padrões. Tese de Doutorado, Unicamp, 2001.

BOLFARINE, H & SANDOVAL, M. C – Introdução à Inferência Estatística.

MARQUES,L.A.V – Um estudo sobre a inicialização do algoritmo E.M. aplicado em misturas finitas de normais assimétricas. Dissertação de Mestrado, Ufam, 2009

7. ANEXOS

Algoritmo 1 misturas de duas normais

```
# Agorito E.M.-------------------------------------------------------------------------------
## Mistura de duas normais-------------------------------------------------------------------
y <- c(rnorm(500,1,3),rnorm(500,2,3))
mu0 =1.002695
mu1 = 2.097774
var0 =4.913036
var1 =5.079962
prob =0.372861
sim = 1000
eps = 1/100000

# y assumido como uma mistura de 2 normais-----------------------------------------
mat.name = list(1:sim,c("mu0","mu1","var0","var1","prob"))
result.mat = matrix(,sim,5,dimnames=mat.name)
cat("Sim","\t","var0","\t","var1","\t","prob0","\t","erro","\n")
for(f in 1:sim){
        new.params <- c(mu0,mu1,var0,var1,prob)
        erro <- 1
        iter <-1
        maxiter <- 5
        hist (y,probability=T,nclass=30,ylim=c(0,0.20))
        xvals <- seq(from=min(y),to=max(y),length=100)      #gera 100 valores
        entre o min(y) e o max(y)
                while (erro > eps) {
                        if (iter <= maxiter) {
                                lines (xvals,prob*dnorm (xvals,mu0,sqrt(var0))+
                                (1-prob)*dnorm(xvals,mu1,sqrt(var1)),
                                lty=iter+1,col=iter+1)
                        }
```

```
 # Inicio do E-step--------------------------------------------------------------------------------
                    bayes <- (prob * dnorm(y, mu0, sqrt(var0))) / ((prob *
                    dnorm(y,mu0, sqrt(var0))) + ((1 - prob) * dnorm(y, mu1,
                    sqrt(var1))))
  # Início M-step
                    cat(f,"\t",round(var0,2),"\t",round(var1,2),"\t",round(erro,6),"\t",b
                    ayes[1],"\n")

                    mu0 <- sum(bayes * y)/ sum(bayes)
                    mu1 <- sum((1 - bayes) * y) / sum ((1 - bayes))
                    var0 <- sum(bayes * (y - mu0)^2)/ sum(bayes)
                    var1 <- sum((1 - bayes) * (y - mu1)^2) / sum((1 - bayes))
                    prob <- mean(bayes)
                    old.params <- new.params
                    new.params <- c(mu0, mu1,var0, var1,prob)
                    erro <- max(abs((old.params - new.params)/new.params))
                    iter <- iter + 1
                }
    legend ("topright",legend=c("Última Iteração"),lty=5,col=2)
    result.mat[f,] = new.params
}
print(result.mat)
mean_new.params<-c(mean(mu0),mean(mu1),mean(var0),mean(var1),
mean(prob))
print(mean_new.params)
cbind(mean_new.params)

##kmens-------------------------------------------------------------------------------------------
#colnames(y) <- c("x", "y")
(cl <- kmeans(y, 2))
#plot(y, col = cl$cluster)
#points(cl$centers, col = 1:2, pch = 8, cex=2)
```

Algoritmo 2 misturas de duas normais

```
#Agorito E.M.-------------------------------------------------------------------------------
## Mistura de três normais---------------------------------------------------------------
y <- c(rnorm(500,0,1),rnorm(200,5,3), rnorm(300,10,1))
mu0 = 0.03
mu1 = 4.81
mu2 = 9.9
var0 = 0.9
var1 = 4.0
var2 = 1.02
prob0 = 0.52
prob1 = 0.03
sim = 1000
eps = 1/100000

# y assumindo como uma mistura de 3 normais----------------------------------------
mat.name=list(1:sim,c("mu0","mu1","mu2","var0","var1","var2","prob0","prob1"))
result.mat = matrix(,sim,8,dimnames=mat.name)
cat("Sim","\t","var0","\t","var1","\t","var2","\t","prob0","\t","prob1","\t","erro","\n")
for(f in 1:sim){
    new.params <- c(mu0, mu1, mu2,var0, var1, var2, prob0, prob1)
    erro <- 1
    iter <-1
    maxiter <- 5
    hist (y,probability=T,nclass=30,ylim=c(0,0.25))
    xvals <- seq(from=min(y),to=max(y),length=100) #gera 100 valores entre o
    min(y) e o max(y)
    while (erro > eps) {
        if (iter <= maxiter) {
            lines (xvals,prob1*dnorm (xvals,mu1,sqrt(var1))+ prob0*dnorm
            (xvals,mu0,sqrt(var0))+(1-prob0-prob1)*dnorm(xvals,mu2,
            sqrt(var2)),lty=iter+1,col=iter+1)
        }
```

```
# Inicio do E-step----------------------------------------------------------------------------
        bayes1 <- (prob1 * dnorm(y,mu1,sqrt(var1))) / ((prob1 * dnorm(y, mu1,
        sqrt(var1)))+ (prob0*dnorm (y,mu0,sqrt(var0)))+((1-prob0-prob1)*
        dnorm(y,mu2,sqrt(var2))))
        bayes0 <- (prob0*dnorm(y,mu0,sqrt(var0)))/ ((prob1 *
        dnorm(y,mu1,sqrt(var1)))+ (prob0*dnorm (y,mu0,sqrt(var0)))+
        ((1-prob0-prob1)*dnorm(y,mu2,sqrt(var2))))
# Início M-step-------------------------------------------------------------------------------
        cat(f,"\t",round(var0,2),"\t",round(var1,2),"\t",round(var2,2),"\t",round(erro,
        6),"\t",bayes1[1],"\t",bayes0[1],"\n")
        mu0 <- sum(bayes0 * y) / sum(bayes0)
        mu1 <- sum(bayes1 * y)/ sum(bayes1)
        mu2 <- sum((1-bayes1-bayes0) *y)/ sum(1-bayes1-bayes0)
        var0 <- sum(bayes0 * (y - mu0)^2) / sum( bayes0)
        var1 <- sum(bayes1 * (y - mu1)^2)/ sum(bayes1)
        var2 <- sum((1-bayes0-bayes1)*(y-mu2)^2) / sum(1-bayes0-bayes1)
        prob0 <- mean(bayes0)
        prob1 <- mean(bayes1)
        old.params <- new.params
        new.params <- c(mu0, mu1,mu2, var0, var1,var2, prob0, prob1)
        erro <- max(abs((old.params - new.params)/new.params))
        iter <- iter + 1
    }
    legend ("topright",legend=c("Última Iteração"),lty=5,col=2)
    result.mat[f,] = new.params
}
print(result.mat)
parametros = c("mu0","mu1","mu2","var0","var1","var2","prob0","prob1")
means=c(mean(result.mat[,1]),mean(result.mat[,2]),mean(result.mat[,3]),mean(r
esult.mat[,4]),mean(result.mat[,5]),mean(result.mat[,6]),mean(result.mat[,7]))
cbind(parametros,means)
print(means)
```

8. CRONOGRAMA

Descrição	Ago	Set	Out	Nov	Dez	Jan	Fev	Mar	Abr	Mai	Jun	Jul
Revisão da Literatura	Sim	Sim	Sim	Sim	Sim	Sim	Sim	Sim	Sim	Sim	Sim	Não
Modelagem						Sim	Sim	Sim	Sim	Sim		
Estudos de simulação				Sim	Sim	Sim	Sim	Sim	Sim	Sim		
Elaboração do Resumo e Relatório Final											Sim	
Preparação da apresentação Final para o Congresso												Não