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

ESTIMAÇÃO DE MODELOS DE MISTURA DE EXPONENCIAIS: UM ESTUDO DE SIMULAÇÃO

PROGRAMA DE INICIAÇÃO CIENTÍFICA NA UFAM

Bolsista: Renata Evangelista Monteiro, CNPq

MANAUS

2009

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

RELATÓRIO FINAL

PIB – E – 0076/2008

ESTIMAÇÃO DE MODELOS DE MISTURA DE EXPONENCIAIS: UM ESTUDO DE SIMULAÇÃO

PROGRAMA DE INICIAÇÃO CIENTÍFICA NA UFAM

Bolsista: Renata Evangelista Monteiro
Orientadora: Profa. Dra. Amazoneida Sá Peixoto Pinheiro

MANAUS

2009

Estimação de Modelos de Misturas de Exponenciais: Um estudo de Simulação

Renata Evangelista Monteiro*

Resumo

Uma grande variedade de fenômenos aleatórios pode ser modelada por mistura de distribuições exponenciais. Especificamente, em estudos de confiabilidade estes modelos tem recebido grande atenção nos últimos anos devido sua grande flexibilidade na estimação no tempo de vida de produtos.

As funções de verossimilhança encontradas para as misturas são de grande complexidade, tornando impossível a determinação analítica destes estimadores. Dentro deste contexto, este trabalho tem por objetivo estudar a estimação dos modelos de mistura de exponenciais pelo algoritmo EM (Expectation - Maximization).

Foram estudados três modelos de misturas com 2, 3 e 10 componentes densidades, com o objetivo de verificar a convergência dos valores, isto é, a qualidade das estimativas obtidas e a relação com o número de componentes densidades no modelo.

As simulações foram realizadas no software estatístico R e apresentaram boas estimativas para modelos mais simples, ou seja, modelos que contem menor número de densidades.

---

* e-mail: evangelista_renata@hotmail.com. Este trabalho foi financiado pelo CNPq.

Sumário

1. Introdução 5

2. Revisão Bibliográfica..... .....6 2.1 Mistura de densidades.....6 2.2 Mistura Finita de Exponenciais.....6

3. Metodologia.....8

3.1 Métodos de estimação.....8

3.2 Algoritmo EM.....8

4. Resultados e Discussões.....10

5. Conclusões.....14

6. Referências Bibliográficas.....15

7. Anexos.....16

8. Cronograma de Atividades.....19

1. Introdução

Misturas finitas de distribuições são ferramentas importantes para modelagem de populações heterogêneas com aplicações em diversas áreas do conhecimento, como biologia, genética, medicina, psiquiatria, economia, engenharia e marketing. Devido sua grande flexibilidade, os modelos de mistura tem recebido grande atenção nos últimos anos, tanto do ponto de vista prático como teórico.

Em confiabilidade, produtos produzidos sob as mesmas condições técnicas terão tempos de vidas estudados segundo uma distribuição exponencial. Entretanto, os principais componentes envolvidos na fabricação de eletro – eletrônicos são adquiridos muitas vezes por fornecedores diferentes. Essa característica da cadeia produtiva pode gerar equipamentos com tempos de vidas diferentes, mesmo sendo manufaturados na mesma indústria e sob as mesmas condições técnicas.

No presente estudo empregamos as funções de verossimilhança, através das quais são encontrados os estimadores de máxima verossimilhança para o tempo de vida de tais produtos. Neste sentido, não utilizamos dados reais mas, dados simulados de modelos misturas com 2, 3 e 10 componentes densidades de exponenciais. Os principais objetivos desta pesquisa são:

(i). Estudar o algoritmo EM para implementação do programa de misturas;

(ii). Verificar a qualidade das estimativas obtidas;

(iii). Relação da qualidade das estimativas com o número de densidades componentes na mistura.

2. Revisão Bibliográfica

2.1 Mistura de Densidades

Segundo McLachlan & Peel (2000) define-se uma mistura finita de densidades:

$$f_i(.\mid \theta_i)\,, \qquad \text{i = 1,..., n}$$

indexadas pelos vetores de parâmetros desconhecidos θ_i, i = 1,..., n como sendo a densidade:

$$f(.\mid \theta_1,...,\theta_n) = \sum_{i=1}^{n} p_i\, f_i(.\mid \theta_i) \tag{2.1}$$

onde os pesos $0 < p_i < 1$, i = 1,..., n, são parâmetros desconhecidos, com $\sum_{i=1}^{n} p_i = 1$.

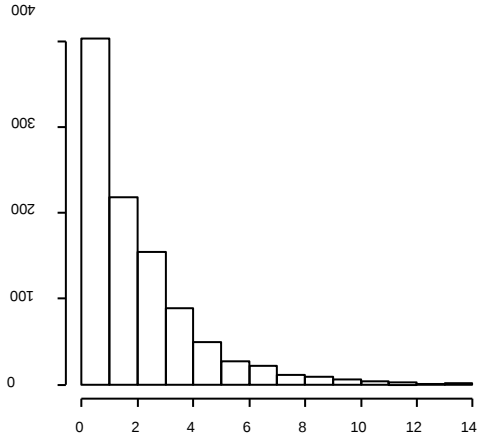
2.2 Mistura Finita de Exponenciais

Abordaremos a situação onde cada densidade na mistura é uma exponencial conforme a definição de Dantas (2004), e é dada por:

$$f(x\mid \lambda) = \lambda e^{-\lambda x}, \qquad \lambda > 0 \;\; e \;\; x > 0 \tag{2.2}$$

O parâmetro λ indica a taxa de ocorrência por unidade de medida, que neste caso pode ser tempo. Se X tem densidade (2.2) utilizamos a notação X~Exp(λ).

Figura 1: Gráfico da distribuição exponencial λ=1/2



Segundo Redner (1984), mistura finita de exponenciais é uma soma ponderada de funções densidades de exponenciais, onde cada parcela da soma é atribuído um peso. Podemos escrever a densidade (2.1) como:

$$f(x \mid \theta) = p_1\lambda_1 e^{-x\lambda_1} + p_2\lambda_2 e^{-x\lambda_2} \tag{2.3}$$

$\theta = (\lambda_1, \lambda_2, p_1, p_2)$ vetor de parâmetros a ser estimado, onde os pesos p_i também são parâmetros desconhecidos, com $\sum_{i=1}^{n} p_i = 1$.

m = nº. de componentes densidades.

p = proporção das densidades na mistura.

Representação Gráfica de uma mistura com duas densidades exponenciais com diferentes parâmetros com n = 1000

Figura 2:

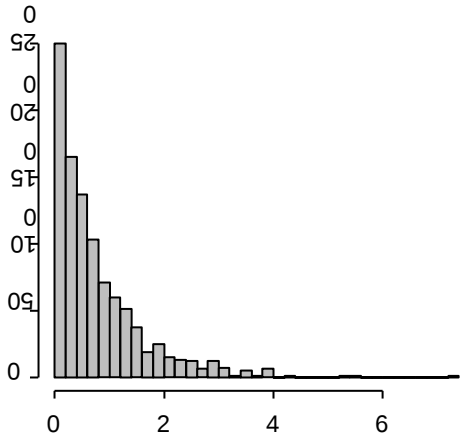
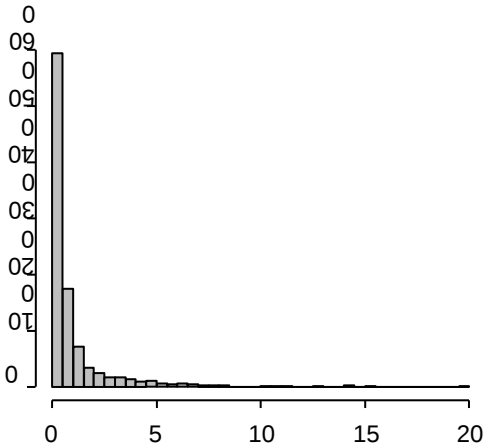


Figura 3:



n = nº. de observações da amostra aleatória gerada no R

Na figura 2 $f(x_1,..., x_n \mid \theta) = 0{,}6\, f_1 + 0{,}4\, f_2$, onde $f_1 \sim Exp(\lambda = 1)$ e $f_2 \sim Exp(\lambda = 2)$

.

Na figura 3 $f(x_1,..., x_n \mid \theta) = 0{,}7\, f_1 + 0{,}3\, f$, onde $f_1 \sim Exp(\lambda = 3)$ $f_2 \sim Exp(\lambda = 2/5)$.

Para estimação dessa mistura utilizaremos o método de máxima verossimilhança (MV) e o algoritmo EM. O EM é empregado quando não há solução analítica para encontrar o vetor de parâmetros. O algoritmo EM é uma forma computacional que maximiza a função de verossimilhança encontrando os estimadores de máxima verossimilhança em misturas de densidades (EMV).

3. Metodologia

3.1 Método de Máxima Verossimilhança

Sejam $X_1,..., X_n$ observações de uma amostra aleatória de tamanho n da variável aleatória X com função de densidade $f(x \mid \theta)$, com $\theta \in \Theta$, onde Θ é o espaço paramétrico. A função de verossimilhança de θ correspondente a amostra aleatória é dada por:

$L(. \mid \theta_i) = f_1(. \mid \theta_1) \times f_2(. \mid \theta_2) \times ... \times f_i(. \mid \theta_i), \qquad i = 1,..., n$

$$L(. \mid \theta_i) = \prod_{i=1}^{n} f_i(. \mid \theta_i) \tag{3.1}$$

Para uma mistura com duas densidades exponenciais a seguinte função de verossimilhança:

$$L(x_1, ..., x_n \mid \theta) = \prod_{i=1}^{n} \left\{ p\, \lambda_1 e^{-\lambda_1 x_i} + (1-p)\, \lambda_2 e^{-\lambda_2 x_i} \right\} \tag{3.2}$$

Aplica-se o log em (3.2), temos:

$$l(x_1, ..., x_n \mid \theta) = \prod_{i=1}^{n} \left\{ p\, \lambda_1 e^{-\lambda_1 x_i} + (1-p)\, \lambda_2 e^{-\lambda_2 x_i} \right\} \tag{3.3}$$

Para encontrar os valores dos parâmetros que maximizam a função calcula-se as derivadas parciais da função (3.3) em relação aos parâmetros $\lambda_1, \lambda_2 \; e \; p_i$ e resolver seus respectivos sistemas igualando a zero.

$$\frac{\partial \log L(x_1,...,x_n \mid \theta)}{\partial \lambda_1} = \sum_{i=1}^{n} \frac{pe^{-\lambda_1 x_i} - p\lambda_1 x_i e^{-\lambda_1 x_i}}{p\lambda_1 e^{-\lambda_1 x_i} + (1-p)\lambda_2 e^{-\lambda_2 x_i}} \tag{3.4}$$

$$\frac{\partial \log L(x_1,...,x_n \mid \theta)}{\partial \lambda_2} = \sum_{i=1}^{n} \frac{(1-p)e^{-\lambda_2 x_i} - (1-p)\lambda_2 x_i e^{-\lambda_2 x_i}}{p\lambda_1 e^{-\lambda_1 x_i} + (1-p)\lambda_2 e^{-\lambda_2 x_i}} \tag{3.5}$$

$$\frac{\partial \log L(x_1,...,x_n \mid \theta)}{\partial p} = \sum_{i=1}^{n} \frac{\lambda_1 e^{-\lambda_1 x_i} + \lambda_2 e^{-\lambda_2 x_i}}{p\lambda_1 e^{-\lambda_1 x_i} + (1-p)\lambda_2 e^{-\lambda_2 x_i}} \tag{3.6}$$

Podemos observar que há um vetor com quatro parâmetros a ser estimado, à medida que for aumentando o número de componentes à quantidade de parâmetros a ser estimado aumentará. A partir daí, vai surgindo dificuldades para estimação do vetor de parâmetros, assim tornando impossível a obtenção da solução analítica.

3.2 Algoritmo EM

O algoritmo EM é usado para determinação de parâmetros a partir de dados incompletos, $x = (x_1,..., x_n)$, ou seja, observações não conhecidas.

Introduzimos a variável $z = (z_1,..., z_n)$ para tornar os dados completos de modo que Θ possa ser estimado utilizando as técnicas de estimação convencional de máxima verossimilhança.

$$\log L(\theta) = \log f(X, Z, \mid \theta) = \sum_{ij} \log f(x_i, z_j \mid \theta) \tag{3.7}$$

O algoritmo EM é constituído de dois passos:

1. Cálculo da Esperança do log da verossimilhança da mistura de exponenciais:

$$E\left[\sum_{i=i}^{n} \log\left(\sum_{i=1}^{m} p_i f_j(x_i / \theta_j)\right)\right] \tag{3.8}$$

2. Maximização da função:

$$\operatorname{argmax} E\left[\sum_{i=1}^{n} \log p_i \quad f\left(x_i / \theta_j\right)\right] \tag{3.9}$$

O processo de estimação pelos métodos convencionais torna-se complexo a partir de misturas com mais de duas exponenciais, ou seja, a função de verossimilhança torna-se mais difícil de ser encontrada. Para a verificação deste fato, estudamos modelos com 2,3 e 10 componentes.

Neste trabalho simulamos 1000 amostras de tamanho n= 50, 200, 500 e 1000. Os valores iniciais foram os mesmos dos parâmetros, uma vez que estamos interessados na qualidade das estimativas geradas pelo EM e sua relação com o número de componentes na mistura.

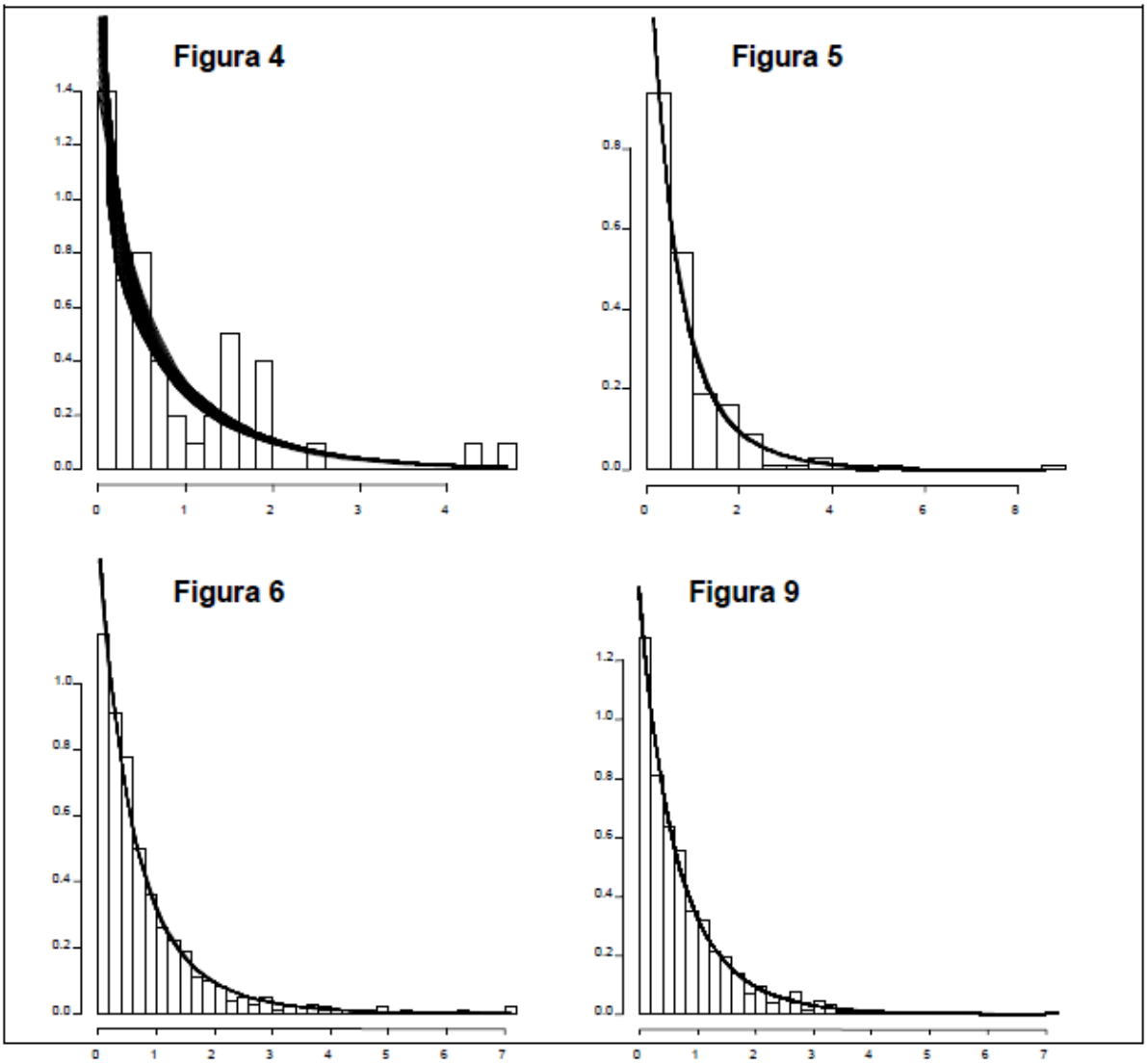
4. Resultados e Discussões

Observamos que as estimativas sao boas para uma mistura com duas densidades, a medida que n aumenta a tendencia dessas estimativas e ficar mais proxima dos verdadeiros valores (parametros), como mostra a tabela 1.

Tabela 1: Mistura finita com duas densidades exponenciais

n	$\lambda_1=1$	$\lambda_2=2$	$P_1=0,6$	$P_2=0,4$	lk
50	1,034274	3,737436	0,575035	0,424965	-9,555104
200	0,988536	3,066328	0,566773	0,433227	-1,020301
500	0,981397	2,667549	0,575724	0,424276	-2,988870
1000	0,996650	2,257631	0,593727	0,406274	-6,550547

Representação Gráfica da mistura com duas densidades

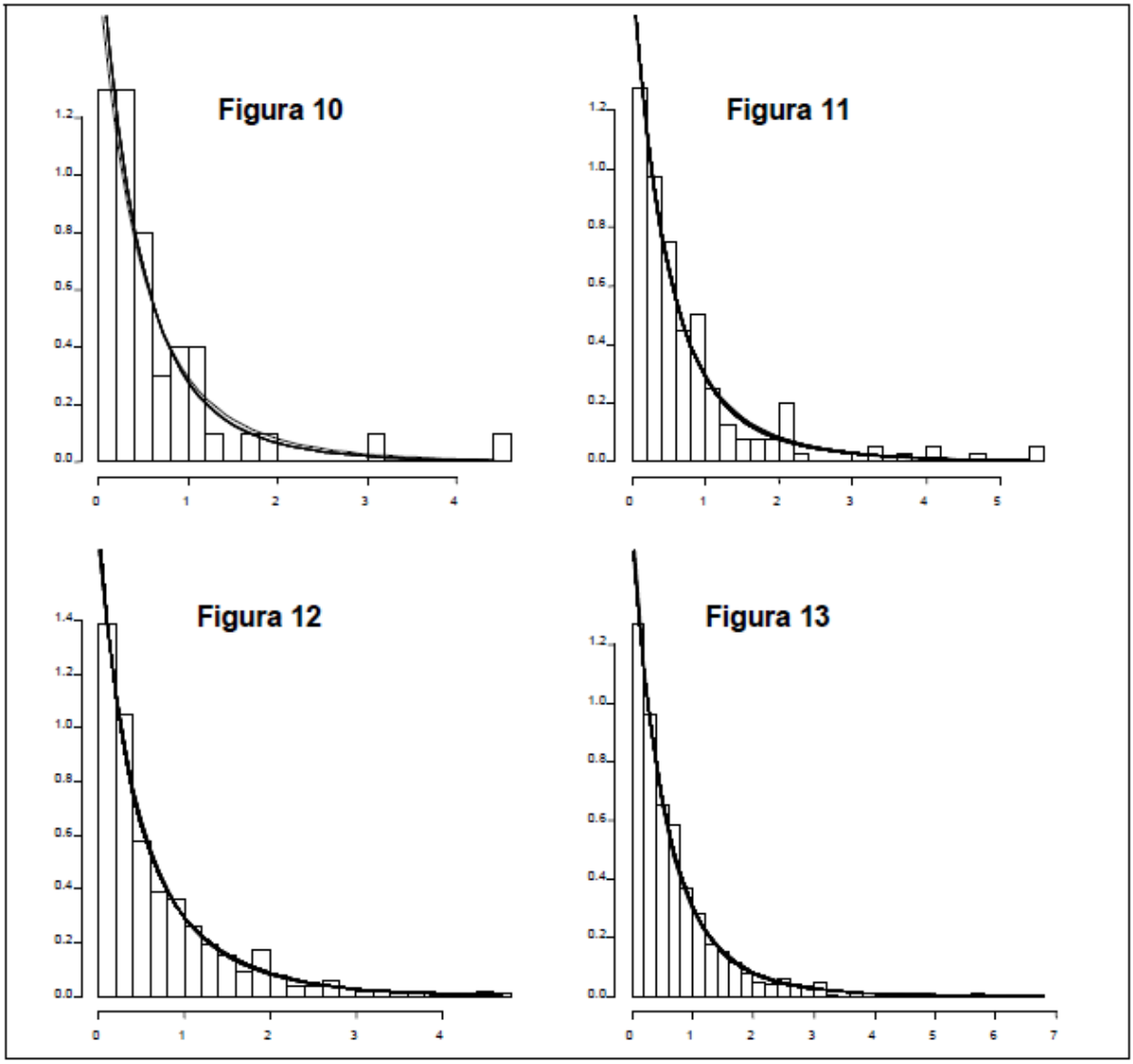


Para uma mistura com tres densidades podemos perceber que para λ_3 as estimativas nao sao boas.

Tabela 2: Mistura finita com três exponenciais

n	λ_1=1	λ_2=2	λ_3=3	P_1=0,5	P_2=0,3	P_3=0,2	lk
50	1,064187	1,941702	10,444067	0,443465	0,350127	0,206409	-1,606873
200	0,970954	1,926400	6,712473	0,432069	0,363104	0,204826	-74,968420
500	0,980091	1,899547	4,341376	0,454194	0,347591	0,198215	-255,631200
1000	0,987728	1,935867	3,523894	0,469370	0,328041	0,202589	-540,553200

Representação Gráfica da mistura com três exponenciais



Para uma mistura com dez densidades as estimativas vao se distanciando dos verdadeiros valores, e no caso de λ_{10} as estimativas nao sao boas.

Tabela 3: Mistura finita de exponenciais com dez densidades exponenciais (Parâmetros dos lambdas)

n	λ_1=1	λ_2=2	λ_3=3	λ_4=4	λ_5=5	lk
50	1,394680	2,578987	3,687570	4,664708	5,544820	44,551000
200	1,061663	2,362674	3,691623	4,753037	5,515397	133,098100
500	0,998266	2,197106	3,512133	4,517316	5,454249	230,805500
1000	0,977205	2,095127	3,364649	4,423071	5,365900	443,055500
n	λ_6=6	λ_7=7	λ_8=8	λ_9=9	λ_{10}=10	
50	6,399906	7,254544	8,323006	10,032768	88,903557	
200	6,136134	6,711868	7,435945	8,628251	26,200855	
500	6,198064	6,836652	7,457935	8,403719	40,135635	
1000	6,213740	6,810708	7,489169	8,379354	17,121233	

Tabela 4: Mistura finita de exponenciais com dez densidades exponenciais (Parâmetros das proporções)

n	P_1=0,15	P_2=0,10	P_3=0,05	P_4=0,10	P_5=0,10
50	0,1502062	0,1078753	0,0526697	0,1015134	0,0997764
200	0,1371880	0,1087358	0,0530044	0,1033074	0,1032626
500	0,1343577	0,1075811	0,0528464	0,1033151	0,1024385
1000	0,1331060	0,1081881	0,0532271	0,1022607	0,1019486
n	P_6=0,10	P_7=0,10	P_8=0,10	P_9=0,10	P_{10}=0,10
50	0,0985156	0,0973413	0,0964048	0,0960840	0,0996127
200	0,1021883	0,1011613	0,0995337	0,0980328	0,0935853
500	0,1024918	0,1023877	0,1014882	0,1002260	0.0928671
1000	0,1024955	0,1022616	0,1013592	0,1001949	0,0949577

5. Conclusões

Os modelos de mistura de exponenciais estudados no presente trabalho apresentaram:

- A estimação dos parâmetros é menos eficiente à medida que o número de densidades componentes aumenta na mistura;
- As estimativas dos pesos se aproximam dos verdadeiros valores nas três misturas;
- O algoritmo E.M é uma alternativa para modelos de mistura de exponenciais, havendo necessidade de aumentar o número de observações para modelos mais complexos.

6. Referências Bibliográficas

MCLACHLAN, G. & PEEL, D. – Finite mixture models. Wiley & Sons, Nova York, 2000. 419 p.

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

BOLFARINE, Heleno & SANDOVAL, Mônica Carneiro – Introdução à Inferência Estatística. SBM Editora, 2000.

DANTAS, C. A. Probabilidade: Um Curso Introdutório. Editora da Universidade de São Paulo, 2004.

TORGO, Luiz. Introdução a Programação R. Faculdade de economia. Universidade do Porto, 2006.

7. Anexos

Programas implementados no software R

```
rexpmix <- function(n,lambda,peso){
# Função para gerar amostras de uma densidade mistura de exponenciais
# n: tamanho da mostra
# lambda: vetor com os parâmetros das densidades exponenciais
# peso: vetor de pesos na mistura
# y: amostra
# z: rótulos
    y <- vector()
    k <- length(peso)
    z<-sample(k,n,prob=peso,replace=TRUE)
    for (i in 1:n){
        y[i] <- rexp(1,lambda[z[i]])
    }
        return(list(y=y,z=z))
}
```

```
mixexp <-function (y, lambdaini, pesoini,maxiter,eps){
# Programa para estimação por MV em um modelo de misturas de exponenciais
#
#-------------------------------------------------------
# Input
#-------------------------------------------------------
#  y: vetor n-dimensional com a amostra
#  lambdaini: vetor k-dimensional com os valores iniciais para os parâmetros das
exponenciais
#  probini: vetor k-dimensional com os valores iniciais para os pesos na mistura
#-------------------------------------------------------
     xvals<-seq(from=min(y), to=max(y), by=0.10)
     m <- length(xvals)
     n <- length(y)


   k <- length(pesoini)
   lk <- 0
   lambda.novo <- c(lambdaini)
   peso.novo <- c(pesoini)
   err <- 1
   iter<- 0
   hist(y, probability=T, nclass=30)
   bayes <- matrix(0,n,k)
   while( err>eps) {
  lines (xvals, dexpmix(xvals,lambda.novo,peso.novo),lty=1)
  for (j in 1:k){
  bayes[,j] <- peso.novo[j]*dexp(y,lambda.novo[j])
            }
           bayes <- bayes/rowSums(bayes)
lambda.velho <- lambda.novo
peso.velho <- peso.novo
 lambda.novo  <- colSums(bayes)/(y%*%bayes)
         peso.novo <- colSums(bayes)/n
```

```
lk.velho <- sum(log(dexpmix(y,lambda.velho,peso.velho)))
lk.novo <- sum(log(dexpmix(y,lambda.novo,peso.novo)))
#print('log-likelihood')
#print(lk.novo)
          err <- abs((lk.velho-lk.novo)/lk.novo)
#         err<- max(abs((velho.param-novo.param)/novo.param))
          iter<-iter+1
#         print('iteração')
#         print(iter)
#         print('Erro')
#         print(err)
    }
    pesomed <-  rowMeans(p)
    return(list(lk=lk,shape=shape,p=p,shapemed=shapemed,pesomed=pesomed))
   }
```

```
#     legend(4,.25,legend=c("iter 1", "iter 2", "iter 3", "iter 4"), lty=1:4)
    return(list(lambda=lambda.novo, peso=peso.novo, loglik=lk.novo))
  # f(x|lambda): densidade exponencial de parâmetro lambda
  # y: vetor com os pontos onde a densidade mistura de exponenciais será avaliada
  # lambda: vetor com os parâmetros das densidades exponenciais
  n <- length(y)
  k <- length(peso)
  peso <- matrix(peso,k,1)
  f <- matrix(0,n,k)
  for (j in 1:k){
      f[,j] <- dexp(y,lambda[j])
  }
  g <- f%*%peso
  return(g)
}
experimento <- function(n,nsimul,lambda,peso,maxiter,eps){
```

```
    lk <-0
    k <- length(peso)
    shape <- matrix(0,k,nsimul)
    p  <- matrix(0,k,nsimul)
        for (i in 1:nsimul){
            print(i)
             y <- rexpmix(n,lambda,peso)$y
             aaa <- mixexp(y, lambdaini=lambda, pesoini=peso,maxiter, eps)
             lk[i] <- aaa$loglik
             shape[,i] <- aaa$lambda
             p[,i] <- aaa$peso
          }
    }
    g <- f%*%peso
    return(g)
}

shapemed <-  rowMeans(shape)
pesomed <-  rowMeans(p)
   return(list(lk=lk,shape=shape,p=p,shapemed=shapemed,pesomed=pesomed))
  }
```

8. Cronograma

Descrição	Ago	Set	Out	Nov	Dez	Jan	Fev	Mar	Abr	Mai	Jun	Jul
Revisão da Literatura	X	X	X	X	X	X	X	X				
Modelagem						X	X	X	X	X		
Estudos de simulações				X	X	X	X	X	X	X		
Elaboração do resumo e relatório final											X	
Preparação da apresentação final para o Congresso												X