# Inferência em modelos de mistura via algoritmo EM estocástico modificado

***Raul Caram de Assis***

UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA

PROGRAMA INTERINSTITUCIONAL DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

RAUL CARAM DE ASSIS

# INFERÊNCIA EM MODELOS DE MISTURA VIA ALGORITMO EM ESTOCÁSTICO MODIFICADO

Dissertação apresentada ao Departamento de Estatística – DEs-UFSCar e ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística.

Orientador: Prof. Dr. Luis Aparecido Milan

**UFSCar - São Carlos**

**Junho de 2017**

UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA

PROGRAMA INTERINSTITUCIONAL DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

RAUL CARAM DE ASSIS

# INFERENCE ON MIXTURE MODELS VIA MODIFIED STOCHASTIC EM ALGORITHM

Master dissertation submitted to the Departamento de Estatística – DEs-UFSCar and to the Instituto de Ciências Matemáticas e de Computação – ICMCUSP, in partial fulfilment of the requirements for the degree of the Master joint Graduate Program in Statistics.
Advisor: Prof. Dr. Luis Aparecido Milan

**UFSCar - São Carlos**

**June 2017**

**UNIVERSIDADE FEDERAL DE SÃO CARLOS**

Centro de Ciências Exatas e de Tecnologia
Programa Interinstitucional de Pós-Graduação em Estatística

## Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Dissertação de Mestrado do candidato Raul Caram de Assis, realizada em 02/06/2017:

Prof. Dr. Luis Aparecido Milan
UFSCar

Prof. Dr. Luiz Koodi Hotta
UNICAMP

Profa. Dra. Miriam Harumi Tsunemi
UNESP

*Dedico este trabalho aos meus pais que me incentivaram fortemente,*
*à Caroline que esteve ao meu lado*
*e a todos os meus colegas de turma que fizeram parte deste período especial.*

# AGRADECIMENTOS

Agradeço à minha família e à Caroline pelo suporte prestado durante o mestrado, ao Prof. Dr. Luis Aparecido Milan que me orientou durante a vigência de meu mestrado e aos demais membros das bancas de Qualificação e Defesa pelas sugestões: Prof. Dr. José Galvão Leite, Prof. Dr. Danilo Lourenço Lopes, Prof. Dr. Luiz Koodi Hotta e Profa. Dra. Miriam Harumi Tsunemi.

*“De tudo ficaram três coisas...*
*A certeza de que estamos começando...*
*A certeza de que é preciso continuar...*
*A certeza de que podemos ser interrompidos antes de terminar...*
*Façamos da interrupção um caminho novo...*
*Da queda, um passo de dança...*
*Do medo, uma escada...*
*Do sonho, uma ponte...*
*Da procura, um encontro!”*
*(Fernando Sabino)*

# RESUMO

RAUL, C. A. **Inferência em modelos de mistura via algoritmo EM estocástico modificado**. 2017. 83 p. Dissertação (Mestrado em Estatística – Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2017.

Apresentamos o tópico e a teoria de Modelos de Mistura de Distribuições, revendo aspectos teóricos e interpretações de tais misturas. Desenvolvemos a teoria dos modelos nos contextos de máxima verossimilhança e de inferência bayesiana. Abordamos métodos de agrupamento já existentes em ambos os contextos, com ênfase em dois métodos, o algoritmo EM estocástico no contexto de máxima verossimilhança e o Modelo de Mistura com Processos de Dirichlet no contexto bayesiano. Propomos um novo método, uma modificação do algoritmo EM Estocástico, que pode ser utilizado para estimar os parâmetros de uma mistura de componentes enquanto permite soluções com número distinto de grupos.

**Palavras-chave:** Modelos de mistura, Mistura de distribuições, Algoritmo EM, Cadeia de Markov, *Gibbs sampling*, Segmentação de imagens.

# ABSTRACT

RAUL, C. A. **Inference on mixture models via modified stochastic EM algorithm**. 2017. 83 p. Dissertação (Mestrado em Estatística – Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2017.

We present the topics and theory of Mixture Models in a context of maximum likelihood and Bayesian inferece. We approach clustering methods in both contexts, with emphasis on the stochastic EM algorithm and the Dirichlet Process Mixture Model. We propose a new method, a modified stochastic EM algorithm, which can be used to estimate the parameters of a mixture model and the number of components.

**Keywords:** Mixture models, Mixture of distributions, EM algorithm, Markov chain, Gibbs Sampling, Image segmentation.

# LISTA DE ILUSTRAÇÕES

# SUMÁRIO

CAPÍTULO

# 1

# INTRODUÇÃO

A importância dos modelos de mistura se deve a sua capacidade de representar subpopulações com características específicas dentro de uma população maior através de um modelo probabilístico.

Uma possível aplicação de modelos de mistura é na análise de agrupamentos, sendo utilizados em diversas áreas do conhecimento humano, como no agrupamento de indivíduos de uma mesma espécie em subgrupos baseados nas suas diversidades genéticas, o agrupamento de fenômenos astronômicos, o reconhecimento de objetos em imagens, o reconhecimento de padrões no movimento humano, a análise de grupos de eventos econômicos, o reconhecimento por máquinas de texto escrito a mão, entre muitos outros.

Segundo MacDonald (2017), em uma publicação de 1894, Karl Pearson analisou a amostra da razão entre a largura da cabeça e o comprimento do corpo de 1000 caranguejos obtida em Nápoles, no sul da Itália, pelo Professor W.F.R. Weldon, que exibia uma distribuição não-normal e assimétrica. A assimetria do histograma foi interpretado por K. Pearson como evidência de que a amostra continha duas espécies diferentes de caranguejos.

O modelo de mistura mais utilizado é o de mistura de normais, mas grande parte da abordagem se aplica a modelos de mistura de quaisquer distribuições.

Referências sobre o assunto podem ser encontradas em McLachlan e Krishnan (1996) e Titterington, Smith e Makov (1985).

## 1.1 Revisão Bibliográfica

Há diversos métodos que podem ser utilizados para estimar os parâmetros via verossimilhança. Um dos mais conhecidos deles é o algoritmo EM, publicado por Dempster, Laird e Rubin (1977). Em cada iteração do algoritmo EM, existem dois passos, sendo que o primeiro corresponde a calcular a esperança da log-verossimilhança como função dos valores não observados

condicionada nas observações sob a estimativa atual do parâmetro e o segundo corresponde a calcular a estimativa do parâmetro que maximiza a esperança encontrada no passo anterior. O EM se aplica a situações em que existem variáveis não observáveis mas que são parte fundamental do problema. Picard (2007) faz uma introdução à teoria do EM e à sua aplicação na estimação de parâmetros em modelos de mistura, assim como uma curta revisão de critérios de seleção de modelos ao se ajustar modelos com número diferente de componentes.

Celeux e Diebolt (1985) consideraram uma versão modificada do algoritmo EM, que cria uma Cadeia de Markov em que os estados desta cadeia são possíveis valores das variáveis não observáveis. A sequência de estimativas é uma Cadeia de Markov ergódica e converge em distribuição para a distribuição estacionária da cadeia. Ferguson (1973) introduziu o Processo de Dirichlet, um processo estocástico que define distribuições de Dirichlet a partir de partições finitas de um espaço. O processo de Dirichlet também pode ser pensado como uma generalização da distribuição de Dirichlet para uma dimensão infinita. Da mesma maneira que a distribuição de Dirichlet é conjugada *a priori* da multinomial, o Processo de Dirichlet é conjugado *a priori* de uma multinomial com infinitas categorias.

## 1.2 Objetivos

Nós propomos uma nova ferramenta para tratar de modelos de mistura que demonstra similaridades do Modelo de Mistura com o Processo de Dirichlet, por ser um algoritmo com capacidade de ajustar modelos com número diverso de grupos, e com o algoritmo EM estocástico por se situar em um ambiente de verossimilhança.

Por fim, fazemos uma análise do desempenho do algoritmo usando dados simulados e dados reais, e demonstramos o uso do algoritmo para agrupamento de observações e segmentação de imagens.

## 1.3 Estrutura do Trabalho

No capítulo 2 definimos o que é um modelo de mistura e qual a sua interpretação. No capítulo 3 discorremos sobre métodos de estimação dos parâmetros de modelos de mistura, no capítulo 4 introduzimos uma versão modificada do algoritmo EM estocástico que permite seleção de modelos com número diferente de componentes e no capítulo 5 aplicamos o algoritmo proposto em diferentes conjuntos de dados com uma discussão dos resultados obtidos. O capítulo 6 contém considerações finais e as conclusões sobre o trabalho.

CAPÍTULO

# 2

# MODELOS DE MISTURA

Um modelo de mistura é um modelo probabilístico em que a distribuição de probabilidade de interesse é uma distribuição obtida pela mistura de outras distribuições. O modo como tal mistura é feita é por uma média ponderada das distribuições que compõe o modelo.

Considere, por exemplo, o caso de uma variável aleatória $X$ cuja realização $x$ representa o tempo em minutos que um automóvel leva para se deslocar do ponto $A$ para o ponto $B$, para o qual existem dois trajetos possíveis $C$ e $C'$, e que o motorista escolhe o caminho $C$ com probabilidade $p$ e o caminho $C'$ com probabilidade $1-p$. Caso o motorista percorra o caminho $C$ a duração da viagem segue uma distribuição exponencial com $\lambda = 0,4$, portanto com média $1/0,4 = 2,5$ minutos (notação $X \sim Exp(0,4)$), enquanto que pelo caminho $C'$ esta duração segue uma exponencial com $\lambda' = 0,2$, portanto com média $1/0,2 = 5$ minutos (notação $X \sim Exp(0,2)$).

Temos pois, que a distribuição de $X$ é função das distribuições condicionais $X|C$ e $X|C'$, além da probabilidade de $C$. De fato, podemos escrever para um intervalo $(a,b)$, com $0 < a < b$, usando a lei da probabilidade total

$$P(a < X < b) = P(a < X < b|C)P(C) + P(a < X < b|C')P(C').$$

Como $C$ e $C'$ formam uma partição do espaço dos eventos, temos que $P(C) + P(C') = 1$, e a distribuição de $X$ se escreve como uma média ponderada das distribuições condicionais de $X|C$ e $X|C'$.

Se quisermos então saber qual a probabilidade de que um automóvel leve até 8 minutos para percorrer o caminho de A até B, devemos então calcular

$$\begin{aligned} P(0 < X < 8) &= P(a < X < b|C)p + P(a < X < b|C')(1-p) \\ &= (1 - e^{(-0,4\cdot 8)})p + (1 - e^{(-0,2\cdot 8)})(1-p). \end{aligned}$$

Neste exemplo, a variável dicotômica $S$ tal que $S = 1$ se $C$ ocorre e $S = 2$ se $C^c = C'$ ocorre, opera como um seletor de distribuições. E poderíamos substituir $P(a < X < b|C)$ por $P(a < X < b|S = 1)$ e $P(C)$ por $P(S = 1)$, assim como $P(a < X < b|C')$ por $P(a < X < b|S = 2)$ e $P(C')$ por $P(S = 2)$. Esta variável é responsável por enumerar os eventos possíveis $S = 1, 2$, dos quais a distribuição condicional de $X$ a cada um destes casos se define de modo específico, em nosso caso estes eventos são a ocorrência da escolha do caminho $C$ ou do $C'$.

A distribuição de $S$ caracteriza as probabilidades com a qual selecionamos uma distribuição entre várias pelas quais $X$ se realizará. No exemplo do automóvel, $S = 1$ significa que selecionamos a distribuição indexada pelo número 1 que corresponde à uma $Exp(0,4)$, e $S = 2$ que selecionamos a distribuição indexada pelo número 2 que corresponde à uma $Exp(0,2)$.

Caso houvesse mais de 2 trajetos, teríamos $S = 1, 2, \ldots, K$, cada elemento correspondendo à seleção de uma distribuição diferente.

Ao conjunto destas distribuições que são indexadas por números $1, 2, \ldots$ chamamos de distribuições componentes do modelo de mistura, ou simplesmente, componentes da mistura. Em um modelo de mistura dito finito existem $K < \infty$ distribuições componentes, o caso $K = 1$ se reduz a um único componente e se torna uma combinação de uma única distribuição, pode ser interpretado como se a única distribuição componente fosse selecionada com probabilidade 1.

Como cada evento $\{S = k\}$ possível de ocorrência representa a seleção de uma das distribuições componentes da mistura, se houver $K$ destas, então $S$ pode assumir exatamente $K$ valores. Por conveniência definiremos $S$ como uma variável aleatória discreta que toma valores no conjunto $\{1, 2, \ldots, K\}$, sendo que cada elemento deste conjunto representa um dos componentes da mistura, ou seja, $S$ é uma variável categórica. E o evento $\{S = k\}$ indica a seleção do componente $k$, e portanto da distribuição da variável aleatória $X|S = k$.

É importante notar que uma mistura de distribuições é também uma distribuição e que dentro de um modelo especificado para esta mistura é possível fazer inferência a partir de uma amostra para a mistura como um todo.

## 2.1 Formalização

Considere $K$ funções de distribuição $F_1(\cdot), \ldots, F_K(\cdot)$ e $S$ uma variável aleatória com distribuição Multinomial$(1; p_1, p_2, \ldots, p_K)$ para algum $K$ inteiro positivo. Deste modo, $P(S = k) = p_k$ para $1 \leq k \leq K$ e $\sum_{k=1}^{K} p_k = 1$.

Ao definirmos uma nova função $F(x)$ como uma média ponderada das distribuições $F_1(\cdot),\ldots,F_K(\cdot)$, isto é,

$$F(x)=\sum_{k=1}^{K} p_k F_k(x),\ \text{onde}\ p_k\geq 0\ \forall k \quad \text{e} \quad \sum_{k=1}^{K} p_k = 1, \tag{2.1}$$

é fácil ver que $F$ também satisfaz as propriedades que caracterizam uma função de distribuição, já que

1. $\lim_{x\to\infty} F(x) = \lim_{x\to\infty} \sum_{k=1}^{K} p_k F_k(x) = \sum_{k=1}^{K} p_k \lim_{x\to\infty} F_k(x) = \sum_{k=1}^{K} p_k = 1$;
2. $\lim_{x\to-\infty} F(x) = \lim_{x\to-\infty} \sum_{k=1}^{K} p_k F_k(x) = \sum_{k=1}^{K} p_k \lim_{x\to-\infty} F_k(x) = \sum_{k=1}^{K} p_k\, 0 = 0$;
3. $F$ é uma função não-decrescente. Se $x\leq y$ então $p_k F_k(x)\leq p_k F_k(y)$ para todo $k$. Então, $F(x)=\sum_{k=1}^{K} p_k F_k(x)\leq \sum_{k=1}^{K} p_k F_k(y) = F(y)$;
4. $F$ é contínua à direita, já que é combinação linear de funções contínuas à direita. Seja $(x_n)_{n\geq 1}$ uma sequência de pontos tal que $x_n\downarrow x\in\mathbb{R}$, então

$$\lim_{x_n\downarrow x} F(x_n) = \lim_{x_n\downarrow x} \sum_{k=1}^{K} p_k F_k(x_n) = \sum_{k=1}^{K} p_k \lim_{x_n\downarrow x} F_k(x_n) = \sum_{k=1}^{K} p_k F_k(x) = F(x).$$

Usando este fato, a equação (2.1) nos diz que uma média ponderada de funções distribuições também é uma função distribuição.

Temos também que (2.1) representa a função distribuição da variável aleatória $X$ em que

i) Consideramos a variável aleatória $S$, cuja realização é um índice $k$ em $\{1,2,\ldots,K\}$, segundo a distribuição definida por $P(S=k)=p_k$;

ii) Obtemos um valor proveniente da distribuição de índice $S=k$ selecionada em (i);

Considerando $(X_k)=(X|S=k)$, temos que para um boreliano $B$ qualquer

$$\begin{aligned} P(X\in B) &= \sum_{k=1}^{K} P(X\in B|S=k)P(S=k) = \sum_{k=1}^{K} P(X_k\in B)P(S=k) \\ &= \sum_{k=1}^{K} p_k P(X_k\in B), \end{aligned}$$

impondo $B=(-\infty,x]$ então $F_X(x)=\sum_{k=1}^{K} p_k F_{X_k}(x)$.

O parâmetro $\boldsymbol{p}=(p_1,\ldots,p_K)$ da distribuição multinomial define uma distribuição sobre os índices $\{1,\ldots,K\}$, em que cada um pode ser associado a uma das $K$ distribuições determinadas por $F_1,F_2,\ldots,F_K$, que chamamos de distribuições componentes da mistura.

Deste modo, uma interpretação natural de um modelo de mistura com $K$ componentes é como um processo de duas etapas, em que selecionamos uma distribuição componente a partir de uma distribuição multinomial $\boldsymbol{p}$, e em seguida obtemos uma realização da distribuição selecionada.

Impondo $p_k > 0$ para todo $k$, todas as funções de distribuição $F_k(\cdot)$ em (2.1) contribuem para a forma de $F(\cdot)$. Quanto maior o valor de $p_k$, mais a função de distribuição $F_k$ contribui para a formação de $F$, o parâmetro $p_k$ é chamado de "probabilidade da componente $k$", "proporção na mistura do componente $k$"ou "peso do componente $k$".

Se cada $X_k$ é uma variável aleatória absolutamente contínua, cuja distribuição é parametrizada por $\theta_k$, derivando (2.1) em relação à $x$ obtemos a densidade da mistura em função das densidades e pesos dos componentes

$$f(x|\boldsymbol{\theta},\boldsymbol{p}) = \sum_{k=1}^{K} p_k f_k(x|\theta_k), \tag{2.2}$$

em que consideramos $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_K)$ o vetor que contém todos os parâmetros das distribuições componente da mistura.

Se as densidades $f_k(\cdot)$ pertencem a mesma classe de famílias paramétricas, cada $f_k(x|\theta_k)$ se define pelo seu parâmetro $\theta_k$, então podemos escrever mais sucintamente

$$f(x|\boldsymbol{\theta},\boldsymbol{p}) = \sum_{k=1}^{K} p_k f(x|\theta_k). \tag{2.3}$$

Como exemplo mostramos nos gráficos da Figura 1 a densidade de uma normal padrão e a densidade de uma mistura de uma Normal$(0,1)$ com uma Normal$(5,1)$, com $p_1 = p_2 = 1/2$.

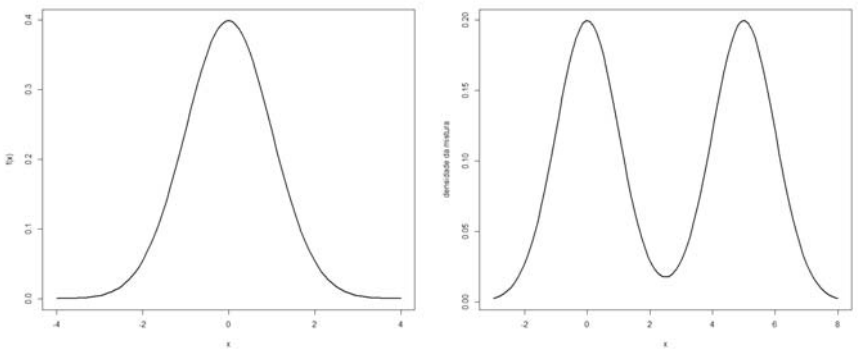

Figura 1 – Função densidade de uma Normal(0,1) e função da densidade da mistura com $X_1 \sim N(0,1)$, $X_2 \sim N(5,1)$, $p_1 = 1/2$ e $p_2 = 1/2$

A densidade da mistura no segundo gráfico da Figura 1 é

$$f(x|\boldsymbol{\mu},\boldsymbol{\sigma}^2,\boldsymbol{p}) = \frac{1}{2} f(x|\mu_1 = 0, \sigma_1^2 = 1) + \frac{1}{2} f(x|\mu_1 = 5, \sigma_1^2 = 1), \tag{2.4}$$

para todo $x \in \mathbb{R}$, em que $f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$, com $\mu \in \mathbb{R}$ e $\sigma > 0$

Podemos dizer que a função de distribuição da mistura é uma média ponderada das funções de distribuição das $K$ distribuições componentes, e, portanto, sempre se situa entre os extremos dos componentes. Note que

$$\begin{aligned}\max_{l;1\leq l\leq K} F_l(x) &= \max_{l;1\leq l\leq K} F_l(x)\sum_{k=1}^{K} p_k = \sum_{k=1}^{K} p_k \max_{l;1\leq l\leq K} F_l(x) \geq \sum_{k=1}^{K} p_k F_k(x) = F(x)\\ &\geq \sum_{k=1}^{K} p_k \min_{l;1\leq l\leq K} F_l(x) = \min_{l;1\leq l\leq K} F_l(x)\sum_{k=1}^{K} p_k = \min_k F_k(x) \quad \forall x \in \mathbb{R},\end{aligned}$$

e se $f_k(x)$ são as funções densidade de probabilidade ou massa de probabilidade da mistura, trocando $F_k$ por $f_k$ e $F$ por $f$, se obtém o resultado análogo,

$$\max_{l;1\leq l\leq K} f_l(x) \geq f(x) \geq \min_{l;1\leq l\leq K} f_l(x) \quad \forall x \in \mathbb{R}.$$

Usando a mistura definida em (2.4) a função de distribuição da mistura se encontra no ponto médio entre as funções distribuições dos 2 componentes para todo $x \in \mathbb{R}$.

A bimodalidade da densidade da mistura pode ser notada facilmente pelo gráfico da densidade, na Figura 1, mas também pode ser notada pelo gráfico da distribuição acumulada, na Figura 2. As retas que tangenciam o gráfico da distribuição acumulada da mistura possuem inclinação maior (contando em sentido anti-horário a partir do eixo $x$) por volta de $x = 0$ e $x = 5$.

Para que a distribuição da mistura esteja bem definida é necessário definir quais são as distribuições componentes e quais são as proporções da mistura de cada componente. Assim, no caso de uma mistura de distribuições da mesma família paramétrica, se há $K$ componentes devemos conhecer $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_K)$, em que $\theta_k$ é o parâmetro do componente $k$ e também $\mathbf{p} = (p_1,\ldots,p_K)$ em que $p_k$ indica a proporção da mistura do componente $k$. Doravante, definimos $\boldsymbol{\Psi} = (\boldsymbol{\theta},\mathbf{p})$ o vetor de parâmetros da mistura.

O espaço paramétrico de $\mathbf{p} = (p_1,\ldots,p_K)$ é o subconjunto de $\mathbb{R}^K$ definido pelas restrições

$$\begin{cases} p_k & \geq 0 \quad \forall\, 1 \leq k \leq K \\ \sum\limits_{i=1}^{K} p_i & = 1. \end{cases}$$

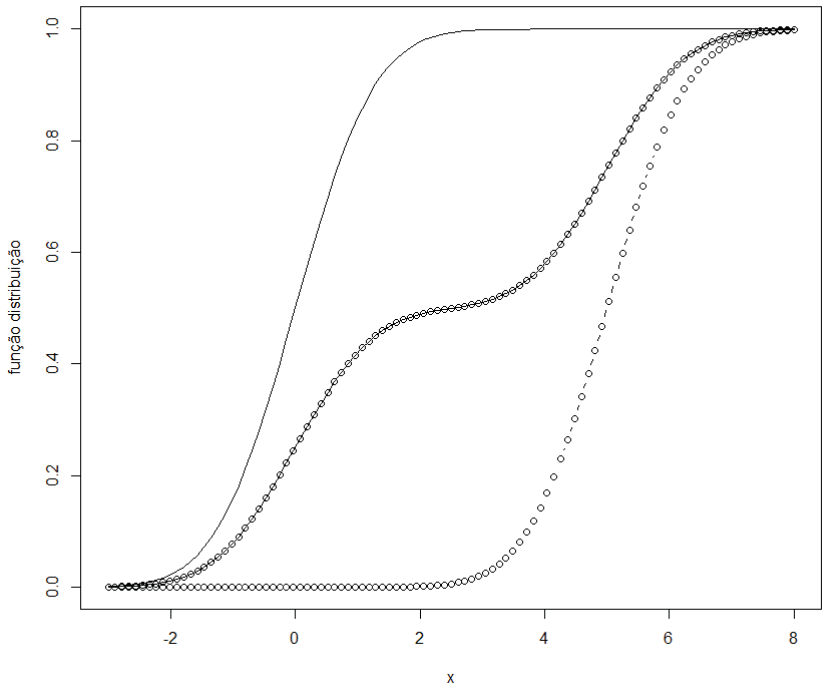


Figura 2 – Funções de distribuição dos componentes e da mistura.

Considerando $\mathbf{e}_i = (e_{i1}, \ldots, e_{iK})$, com $e_{ik} = 1$ se $i = k$ e $e_{ik} = 0$ se $i \neq k$, este espaço paramétrico é o subconjunto de $\mathbb{R}^K$, dos elementos $a_1\mathbf{e}_1 + a_2\mathbf{e}_2 + \ldots + a_K\mathbf{e}_K$ tais que $a_1, \ldots, a_K \geq 0$ e $a_1 + \ldots + a_k = 1$, também chamado de $(K-1)$-simplex. Uma vez que, apesar de ser expressado como uma combinação linear dos $K$ vetores $\mathbf{e}_k$ que formam uma base canônica para $\mathbf{R}^k$, se fixarmos quaisquer $K-1$ valores $a_{n_1}, \ldots, a_{n_{K-1}}$, então $a_{n_K} = 1 - (a_{n_1} + \ldots + a_{n_{K-1}})$ é determinado.

### 2.1.1 Sobre a Identificabilidade de uma Mistura

Uma mistura com $K > 1$ componentes, de maneira geral não é identificável. Para uma permutação $\eta = (\eta_1, \eta_2, \ldots, \eta_K)$ qualquer de $(1, 2, \ldots, K)$, seja $\boldsymbol{\theta}_\eta = (\theta_{\eta_1}, \theta_{\eta_2}, \ldots, \theta_{\eta_K})$ e $\mathbf{p}_\eta = (p_{\eta_1}, p_{\eta_2}, \ldots, p_{\eta_K})$. Então $F(\boldsymbol{\theta}, \mathbf{p}) = F(\boldsymbol{\theta}_\eta, \mathbf{p}_\eta)$ uma vez que a permutação $\eta$ implica somente numa mudança da ordem da soma de

$$F(x|\boldsymbol{\theta}, \mathbf{p}) = \sum_{k=1}^{K} p_k F(\theta_k).$$

Assim, qualquer uma das $K!$ permutações dos índices de 1 a $K$ deixa inalterado a distribuição da mistura.

No entanto se definirmos que deve haver uma ordenação qualquer sobre as estimativas do parâmetros, por exemplo de modo a satisfazer $\hat{\theta}_1 < \hat{\theta}_2 < \ldots < \hat{\theta}_K$, o modelo de mistura ajustado

se torna identificável, no caso de uma mistura de distribuições uniparamétricas e univariáveis da mesma família paramétrica, como uma mistura de Poissons ou de normais univariadas, pode-se rotular os componentes de acordo com as estimativas de seus parâmetros de modo que o componente 1 tenha a menor média estimada, o componente 2 a segunda menor média estimada e assim por diante.

Apesar disto, é conveniente que durante as iterações dos algoritmos não se imponham restrições às estimativas, e somente depois se necessário ordená-las segundo algum critério. Por exemplo, se tivermos estimativas, $\hat{\theta}_1$ e $\hat{\theta}_2$, dos parâmetros das distribuições componentes, $\theta_1$ e $\theta_2$, e impormos $\hat{\theta}_1 < \hat{\theta}_2$, é possível que as primeiras estimativas de $\theta_1$ sejam maiores que as de $\theta_2$ mas conforme as estimativas se aprimorem obtemos estimativas de $\theta_1$ menores que as de $\theta_2$. Se nos preocuparmos somente com o índice dos grupos após a finalização do processo de estimação não será necessário renomear os grupos durante o processo, mas apenas no fim como uma questão de encaixar as estimativas obtidas em um modelo teórico.

## 2.2 Abordagem Bayesiana

Em um contexto bayesiano é necessário definir uma distribuição *a priori* sobre o parâmetro $\boldsymbol{p}$, para isto é possível usar a distribuição de Dirichlet de ordem $K$, que tem densidade

$$f(p_1,\ldots,p_K|\alpha_1,\ldots,\alpha_K) = \frac{1}{B(\boldsymbol{\alpha})}\prod_{k=1}^{K} p_i^{\alpha_k-1}, \tag{2.5}$$

em que $\alpha_1,\ldots,\alpha_K > 0$ e a constante normalizadora é o inverso da função beta aplicada ao vetor $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_K)$, definida por

$$B(\boldsymbol{\alpha}) = \frac{\prod_{k=1}^{K}\Gamma(\alpha_k)}{\sum_{k=1}^{K}\Gamma(\alpha_k)}. \tag{2.6}$$

Tal distribuição tem como suporte o $(K-1)$-simplex. Para definir uma distribuição uniforme no suporte da Dirichlet é necessário usar o caso particular da distribuição Dirichlet simétrica com $\alpha_1 = \ldots = \alpha_K$ =1.

Note que se $\boldsymbol{p} = (p_1,\ldots,p_K)$ é a realização de uma variável aleatória com distribuição de Dirichlet, então $p_k > 0$ para todo $k$ e $p_1 + p_2 + \ldots + p_K = 1$. Assim interpretamos as realizações de uma distribuição de Dirichlet como um vetor de probabilidades.

Se as densidades $f_k(\cdot|\theta_k)$ são da mesma classe paramétrica (por exemplo, todas fdp's de variáveis com distribuição normal ou de Poisson), então toda a diferença entre as distribuições dos componentes reside na diferença entre seus parâmetros.

Neste caso, os parâmetros $\boldsymbol{p} = (p_1, \ldots, p_K)$ induzem uma distribuição discreta sobre $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$, uma vez que uma distribuição sobre a escolha dos componentes

$$P(S_i = k) = p_k$$

é equivalente neste caso a uma distribuição sobre a escolha dos parâmetros, pois escolher o componente $k$ através de uma variável categórica equivale a escolher o parâmetro $\theta_k$ do componente $k$ entre os possíveis valores de $\theta$. Definimos assim, uma distribuição discreta $G$, sobre o espaço paramétrico $\Theta$ de $\theta$, tal que

$$G(\theta_k) = P(\theta = \theta_k) = p_k \quad \text{para todo } k.$$

Deste modo, a distribuição da mistura com $K$ componentes que é dada por

$$f(\cdot|\boldsymbol{\Psi}) = \sum_{k=1}^{K} p_k f(\cdot|\theta_k)$$

pode ser escrita de modo equivalente levando em conta a distribuição do parâmetro $\theta$ no lugar da distribuição de uma variável categórica $S$, da seguinte forma,

$$f(\cdot|\boldsymbol{\Psi}) = \sum_{k=1}^{K} f(\cdot|\theta_k) G(\theta_k),$$

ou ainda,

$$f(\cdot|\boldsymbol{\Psi}) = \int_{\boldsymbol{\Theta}} f(\cdot|\theta) dG(\theta). \tag{2.7}$$

CAPÍTULO

3

# INFERÊNCIA EM MODELOS DE MISTURA

Até o capítulo anterior definimos o que é um modelo de mistura de distribuições, demos uma interpretação do modelo de mistura como um processo em que existem $K$ distribuições e uma variável categórica sobre tais distribuições e uma segunda interpretação algébrica em que o modelo de mistura tem como função de distribuição uma média ponderada das funções de distribuição dos $K$ componentes. Mostramos alguns exemplos de tais misturas de distribuições e que, quando temos uma mistura com distribuições da mesma família paramétrica, a distribuição da variável categórica $S$ sobre os componentes induz uma distribuição no espaço paramétrico $\Theta$ do parâmetro $\theta$ de cada componente.

A partir de uma amostra $\mathbf{x} = (x_1, \ldots, x_n)$ e sob a premissa de que tal amostra é proveniente de uma família de distribuições de probabilidade $\mathscr{P} = \{P_\psi | \psi \in \Psi\}$, em que cada $P_\psi$ é uma distribuição de um modelo de mistura com $K$ componentes como função do parâmetro $\psi$, o próximo passo é procurar a distribuição $P_\psi$ que melhor se ajusta aos dados segundo algum critério, uma vez que o modelo estatístico para mistura de distribuições está bem definido.

Por vezes, abordaremos o problema por outro ângulo. Se considerarmos $\mathbf{S} = (S_1, \ldots, S_n)$ como as variáveis não-observáveis, em que $S_i$ indica a proveniência de $x_i$ a um dos $K$ componentes, podemos trabalhar com uma distribuição de probabilidade sobre o conjunto $\mathscr{S}$, que definimos como o espaço de todos os possíveis valores de $S$, buscando o $S$ que melhor se ajusta às observações $\mathbf{x}$ segundo algum critério.

## 3.1 Algoritmo EM

O algoritmo EM foi proposto em Dempster, Laird e Rubin (1977) como um método iterativo para obter o máximo de uma função de verossimilhança. Muito do que tratamos aqui sobre o algoritmo EM é baseado no livro McLachlan e Peel (2000).

### 3.1.1 Introdução ao Algoritmo EM

Seja $\mathbf{X} = (X_1, \ldots, X_n)$ o vetor aleatório que corresponde às observações $\mathbf{x}$ com função de densidade $f(\mathbf{x}; \boldsymbol{\Psi})$, e $\boldsymbol{\Psi} = (\Psi_1, \ldots, \Psi_K)$ é vetor com os parâmetros $\Psi_k$ de cada componente. Por exemplo, se os componentes são normais univariadas então $\Psi_k = (p_k, \mu_k, \sigma_k^2)$ contém a proporção da mistura, a média e variância do componente $k$ respectivamente.

Além de $\mathbf{X}$ existe um vetor de variáveis aleatórias não observáveis $\mathbf{S} = (S_1, \ldots, S_n)$ em que a variável $S_i$ assume um entre $K$ valores, que indica a proveniência da observação $x_i$ a um dos componentes.

Fazendo uso deste contexto, definimos $\mathbf{C} = (\mathbf{X}, \mathbf{S})$ o vetor aleatório de dados completos, que contém tanto as variáveis observáveis quanto as não-observáveis, e $\mathbf{c}$ a sua realização e também $f_{\mathbf{C}}(\mathbf{c}; \boldsymbol{\Psi})$ como a densidade de $\mathbf{C}$ no ponto $\mathbf{c}$.

Definimos $\mathscr{X}$ e $\mathscr{S}$ como o espaço amostral das variáveis $\mathbf{X}$ e $\mathbf{S}$ respectivamente, portanto $\mathscr{C} = \mathscr{X} \times \mathscr{S}$ é o espaço amostral de $\mathbf{C}$. Considerando a projeção dos dados completos sobre os dados observáveis, $\Pi(\mathbf{x}, \mathbf{s}) = \mathbf{x}$ para todo $(\mathbf{x}, \mathbf{s}) \in \mathscr{C}$, como $f_{\mathbf{C}}(\mathbf{C}; \boldsymbol{\Psi})$ é a densidade conjunta de $\mathbf{X}$ e $\mathbf{S}$, temos

$$f(\mathbf{x}; \boldsymbol{\Psi}) = \int_{\Pi^{-1}(\mathbf{x})} f_{\mathbf{C}}(\mathbf{c}; \boldsymbol{\Psi})\, d\mathbf{c}$$

como uma fórmula para expressar $f(\mathbf{x}|\boldsymbol{\Psi})$ em função de $f_{\mathbf{C}}(\mathbf{c}|\boldsymbol{\Psi})$, em que estamos integrando sobre o conjunto $\{(\mathbf{x}, \mathbf{s}); \mathbf{s} \in \mathscr{S}\}$ ($\mathbf{x}$ está fixo). Como em nosso caso $\mathbf{S}$ é uma variável discreta a integral acima pode ser escrita como uma soma

$$f(\mathbf{x}; \boldsymbol{\Psi}) = \sum_{\mathbf{s} \in \mathscr{S}} f_{\mathbf{C}}(\mathbf{c}; \boldsymbol{\Psi}) = \sum_{\mathbf{s} \in \mathscr{S}} f_{\mathbf{C}}(\mathbf{x}; \mathbf{s}, \boldsymbol{\Psi}) P(\mathbf{S} = \mathbf{s}; \boldsymbol{\Psi}).$$

Como os dados completos dependem de variáveis não-observáveis $\mathbf{s}$ o valor de $\mathbf{c}$ é desconhecido, e portanto não podemos maximizar a função $\log f_{\mathbf{C}}(\mathbf{c}; \boldsymbol{\Psi})$ diretamente e obter a estimativa $\hat{\boldsymbol{\Psi}} = \operatorname{argmax} f_{\mathbf{C}}(\mathbf{c}; \boldsymbol{\Psi})$.

A abordagem do EM é trabalhar com o valor esperado de $\log f_{\mathbf{C}}(\mathbf{C}; \boldsymbol{\Psi})$ considerando o conhecimento dos valores observáveis $\mathbf{x}$ em relação à variável $\mathbf{C}|\mathbf{x}, \boldsymbol{\Psi}^t$, onde $\boldsymbol{\Psi}^t$ é a estimativa de $\boldsymbol{\Psi}$ na iteração $t$, e iterativamente aprimorar as estimativas dos parâmetros.

1. Defina $t = 0$ e escolha uma estimativa inicial $\boldsymbol{\Psi}^0$ de $\boldsymbol{\Psi}$;
2. (Passo E) Encontre $Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^t) = \mathbb{E}_{\boldsymbol{\Psi}^t}[\log f_{\mathbf{C}}(\mathbf{c}; \boldsymbol{\Psi})|\mathbf{x}]$;
3. (Passo M) Escolha $\boldsymbol{\Psi}^{t+1} = \operatorname{argmax} Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^t)$;
4. Incremente em uma unidade o valor de $t$ e retorne ao passo 2.

A esperança no passo E pode ser escrita como

$$\mathbb{E}_{\boldsymbol{\Psi}^t}[\log f_{\mathbf{C}}(\mathbf{c};\boldsymbol{\Psi})|\mathbf{x}] = \sum_s (\log f_{\mathbf{C}}(\mathbf{c};\boldsymbol{\Psi})|\mathbf{x})\, f_{\mathbf{C}}(\mathbf{c}|\mathbf{x},\boldsymbol{\Psi}^t).$$

Assim, fixado $\boldsymbol{\Psi}^t$, $Q$ é uma função de $\boldsymbol{\Psi}$ que maximizamos no passo M. O procedimento descrito produz uma sequência de estimativas $\boldsymbol{\Psi}^0, \boldsymbol{\Psi}^1, \boldsymbol{\Psi}^2, \ldots$, esta por sua vez induz outra sequência, a das log-verossimilhanças segundo cada estimativa,

$$\log L(\boldsymbol{\Psi}^0|\mathbf{x}), \log L(\boldsymbol{\Psi}^1|\mathbf{x}), L(\boldsymbol{\Psi}^2|\mathbf{x}), \ldots$$

que é de interesse para a próxima seção.

### 3.1.2 Monotonicidade do Algoritmo EM

Para estabelecer a convergência da sequência de estimativas geradas pelo EM vamos fazer uso de um resultado básico, retirado de Lima (2011), sobre convergência de sequências monótonas e limitadas.

**Teorema 1.** (LIMA, 2011, p. 121) Se uma sequência de valores reais $a_1, a_2, \ldots$ é tal que $a_{n+1} \geq a_n$ para todo $n$ e $a = \sup a_n < \infty$ então a sequência $(a_n)_{n\geq 1}$ converge para $a$.

*Demonstração.* Seja $\varepsilon > 0$ qualquer, então $a - \varepsilon < a$ não é limitante superior da sequência $a_n$. Logo $\exists n_0$ natural tal que $a - \varepsilon < a_{n_0} \leq a \Rightarrow a - \varepsilon < a_{n_0} < a + \varepsilon$. Como a sequência $a_n$ é não-decrescente e $a$ é supremo desta sequência, então para todo $n > n_0$ temos $a - \varepsilon < a_n < a + \varepsilon$. □

Com o propósito de aplicar este resultado no contexto do algoritmo EM, enunciamos o seguinte

**Teorema 2.** (MCLACHLAN; PEEL, 2000, p. 78–79) A sequência de estimativas $L(\boldsymbol{\Psi}^t|\mathbf{x}) = f(\mathbf{x};\boldsymbol{\Psi}^t)$ é uma função não-decrescente de $t$, isto é

$$L(\boldsymbol{\Psi}^{t+1}|\mathbf{x}) \geq L(\boldsymbol{\Psi}^t|\mathbf{x}) \quad \forall t \geq 0$$

*Demonstração.* Considere, usando a regra de Bayes, que

$$k(\mathbf{c}|\mathbf{x};\boldsymbol{\Psi}) = \frac{f_{\mathbf{C}}(\mathbf{c};\boldsymbol{\Psi})}{f(\mathbf{x};\boldsymbol{\Psi})}$$

é a densidade condicional de $\mathbf{C}$ dado que $\mathbf{X}$ assume o valor $\mathbf{x}$.

Definindo $\log L_{\mathbf{C}}(\boldsymbol{\Psi}|\mathbf{c}) = \log f_{\mathbf{C}}(\mathbf{c};\boldsymbol{\Psi})$ e aplicando a função log nos dois lados da equação obtemos

$$\log k(\mathbf{c}|\mathbf{x};\boldsymbol{\Psi}) = \log L_{\mathbf{C}}(\boldsymbol{\Psi}|\mathbf{c}) - \log L(\boldsymbol{\Psi}|\mathbf{x}).$$

Rearranjando os termos acima obtemos

$$\log L(\boldsymbol{\Psi}|\mathbf{x}) = \log L_{\mathbf{C}}(\boldsymbol{\Psi}|\mathbf{c}) - \log k(\mathbf{c}|\mathbf{x},\boldsymbol{\Psi}).$$

Tomando a esperança com relação à variável $\mathbf{C}|\mathbf{x} = (\mathbf{x},\mathbf{S})$ nos dois lados da equação (o termo no lado esquerdo é constante em função de $\mathbf{C}|\mathbf{x}$) e usando $\boldsymbol{\Psi}^t$ como estimativa para $\boldsymbol{\Psi}$, temos que

$$\begin{aligned} \log L(\boldsymbol{\Psi}|\mathbf{x}) &= \mathbb{E}_{\boldsymbol{\Psi}^t}[\log L_{\mathbf{C}}(\boldsymbol{\Psi}|\mathbf{c}) \mid \mathbf{x}] - \mathbb{E}_{\boldsymbol{\Psi}^t}[\log k(\mathbf{C}|\mathbf{x},\boldsymbol{\Psi}) \mid \mathbf{x}] \\ &= Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^t) - H(\boldsymbol{\Psi};\boldsymbol{\Psi}^t), \end{aligned} \tag{3.1}$$

em que $H(\boldsymbol{\Psi};\boldsymbol{\Psi}^t) = \mathbb{E}_{\boldsymbol{\Psi}^t}[\log k(\mathbf{C}|\mathbf{x},\boldsymbol{\Psi}) \mid \mathbf{x}]$. Usando (3.1) obtemos uma fórmula para a diferença de verossimilhança entre as iterações,

$$\begin{aligned} \log L(\boldsymbol{\Psi}^{t+1}|\mathbf{x}) - \log L(\boldsymbol{\Psi}^t|\mathbf{x}) =& [Q(\boldsymbol{\Psi}^{t+1};\boldsymbol{\Psi}^t) - Q(\boldsymbol{\Psi}^t;\boldsymbol{\Psi}^t)] \\ &- [H(\boldsymbol{\Psi}^{t+1};\boldsymbol{\Psi}^t) - H(\boldsymbol{\Psi}^t;\boldsymbol{\Psi}^t)]. \end{aligned} \tag{3.2}$$

Queremos demonstrar que $\log L(\boldsymbol{\Psi}^{t+1}|\mathbf{x}) - \log L(\boldsymbol{\Psi}^t|\mathbf{x}) \geq 0$. Por definição $\boldsymbol{\Psi}^{t+1} = \operatorname{argmax} Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^t)$, logo $Q(\boldsymbol{\Psi}^{t+1};\boldsymbol{\Psi}^t) \geq Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^t)$ para todo $\boldsymbol{\Psi}$, em particular $Q(\boldsymbol{\Psi}^{t+1};\boldsymbol{\Psi}^t) \geq Q(\boldsymbol{\Psi}^t;\boldsymbol{\Psi}^t)$, logo a primeira diferença em (3.2) é maior ou igual a zero.

Agora, é suficiente mostrar que a segunda diferença, que é subtraída, é menor ou igual a zero. Usando a definição de $H(\boldsymbol{\Psi};\boldsymbol{\Psi}^t)$, as propriedades da função log e a desigualdade de Jensen, temos

$$
\begin{aligned}
H(\boldsymbol{\Psi}^{t+1};\boldsymbol{\Psi}^{t}) - H(\boldsymbol{\Psi}^{t};\boldsymbol{\Psi}^{t}) &= \mathbb{E}_{\boldsymbol{\Psi}^{t}}[\log(k(\mathbf{C}|x;\boldsymbol{\Psi}^{t+1})/k(\mathbf{C}|x;\boldsymbol{\Psi}^{t})) \mid \mathbf{x}] \\
&\leq \log(\mathbb{E}_{\boldsymbol{\Psi}^{t}}[(k(\mathbf{c}|\mathbf{x};\boldsymbol{\Psi}^{t+1})/k(\mathbf{c}|\mathbf{x};\boldsymbol{\Psi}^{t})] \mid \mathbf{x}) \\
&= \log \sum_{\mathbf{s}\in\mathscr{S}} \frac{k(\mathbf{c}|\mathbf{x},\boldsymbol{\Psi}^{t+1})d\mathbf{x}}{k(\mathbf{c}|\mathbf{x};\boldsymbol{\Psi}^{t})} k(\mathbf{c}|\mathbf{x};\boldsymbol{\Psi}^{t}) \\
&= \log \sum_{\mathbf{s}\in\mathscr{S}} k(\mathbf{c}|\mathbf{x},\boldsymbol{\Psi}^{t+1}) \\
&= \log 1 \qquad (3.3) \\
&= 0. \qquad (3.4)
\end{aligned}
$$

Demonstrado isto, temos por efeito a partir de (3.2) que

$$\log L(\boldsymbol{\Psi}^{t+1}|\mathbf{x}) \geq \log L(\boldsymbol{\Psi}^{t}|\mathbf{x}) \text{ para todo } t \geq 0.$$

□

Assim, está provado que o algoritmo EM produz uma sequência de estimativas $(\boldsymbol{\Psi}^{t})_{t\geq 1}$, de modo que a sequências das log-verossimilhanças $(\log(L|\boldsymbol{\Psi}^{t}))_{t\geq 1}$ é não-decrescente, e portanto para qualquer problema com verossimilhança limitada esta última sequência converge.

Esta propriedade do algoritmo EM por si não significa que o algoritmo sempre convirja para um máximo global da função de verossimilhança, de fato, Wu (1983) demonstra que o algoritmo pode convergir para pontos estacionários, como máximos locais e pontos de sela dependendo da forma da função de verossimilhança e da estimativa inicial dos parâmetros.

### 3.1.3 Estimando o Máximo a posteriori com o Algoritmo EM

Uma das maneiras de se obter uma estimativa pontual do parâmetro $\boldsymbol{\Psi}$ é através do máximo *a posteriori*; definimos $\hat{\boldsymbol{\Psi}}_{\text{MAP}} = \text{argmax} \log p(\boldsymbol{\Psi}|\mathbf{x})$ que é um valor do parâmetro que maximiza a densidade *a posteriori*.

Seja $p(\boldsymbol{\Psi})$ a densidade *a priori* do parâmetro $\boldsymbol{\Psi}$, e $p(\boldsymbol{\Psi}|\mathbf{x})$ e $p(\boldsymbol{\Psi}|\mathbf{c})$ as densidades *a posteriori* de $\boldsymbol{\Psi}$ dados somente os valores das variáveis observáveis e os valores das variáveis observáveis e não-observáveis respectivamente.

Pelo teorema de Bayes, sabemos que

$$p(\boldsymbol{\Psi}|\mathbf{x}) \propto L(\boldsymbol{\Psi};\mathbf{x})\ p(\boldsymbol{\Psi}).$$

Deste modo, aplicando a logaritmo dos dois lados e ignorando um termo que não depende de $\boldsymbol{\Psi}$, temos

$$\log p(\boldsymbol{\Psi}|\mathbf{x}) = \log L(\boldsymbol{\Psi};\mathbf{x}) + \log p(\boldsymbol{\Psi}). \tag{3.5}$$

Podemos fazer uso do EM para estimar $\hat{\boldsymbol{\Psi}}_{\text{MAP}}$ com pequenas modificações no algoritmo EM descrito na seção 3.1.1,

1. Defina $t = 0$ e escolha uma estimativa inicial $\boldsymbol{\Psi}^0$ de $\boldsymbol{\Psi}$;
2. Encontre $\mathbb{E}_{\boldsymbol{\Psi}^t}[\log p(\boldsymbol{\Psi}|\mathbf{c})|\mathbf{x}] = Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^t) + \log p(\boldsymbol{\Psi})$;
3. Escolha $\boldsymbol{\Psi}^{t+1} = \text{argmax}\, \mathbb{E}_{\boldsymbol{\Psi}^t}[\log p(\boldsymbol{\Psi}|\mathbf{c})|\mathbf{x}]$;
4. Incremente em uma unidade o valor de $t$ e retorne ao passo 2.

### 3.1.4 Algoritmo EM Aplicado aos Modelos de Mistura

Se a densidade da mistura é

$$f(x|\boldsymbol{\Psi}) = \sum_{k=1}^{K} p_k f(x|\theta_k),$$

então, para uma amostra independente $\mathbf{x}$ a log-verossimilhança pode ser escrita como

$$\log L(\boldsymbol{\Psi}|\mathbf{x}) = \sum_{i=1}^{n} \log f(x_i|\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} p_k f(x_i|\theta_k).$$

que normalmente não tem solução analítica para encontrar o seu máximo. Quando consideramos a verossimilhança condicionada aos dados completos, uma vez que $\mathbb{I}(s_i = k) = \begin{cases} 0, \text{se } s_i \neq k \\ 1, \text{se } s_i = k \end{cases}$ e que $s_i = k$ para um único $k$, podemos expressar a verossimilhança de $(x_i, s_i)$ como

$$f(x_i, s_i|\boldsymbol{\Psi}) = f(x_i|s_i, \boldsymbol{\Psi}) P(S = s_i) = f(x_i|\theta_{s_i})\ p_{s_i} = \sum_{k=1}^{K} \mathbb{I}(s_i = k) f(x_i|\theta_k) p_k,$$

e de $\mathbf{c} = (\mathbf{x}, \mathbf{s})$ como

$$L_{\mathbf{C}}(\boldsymbol{\Psi}|\mathbf{c}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \mathbb{I}(s_i = k)\ f(x_i|\theta_k)\ p_k.$$

Como $\log\left(\sum_{k=1}^{K} \mathbb{I}(s_i = k)\ f(x_i|\theta_k)\ p_k\right) = \sum_{k=1}^{K} \mathbb{I}(s_i = k) \log\left(f(x_i|\theta_k)\ p_k\right)$ podemos escrever

$$\log L_{\mathbf{C}}(\boldsymbol{\Psi}|\mathbf{c}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \mathbb{I}(s_i = k)\ f(x_i|\theta_k)\ p_k = \sum_{i=1}^{n}\sum_{k=1}^{K} \mathbb{I}(s_i = k) \log(f(x_i|\theta_k)\ p_k)$$
$$= \sum_{i=1}^{n}\sum_{k=1}^{K} \mathbb{I}(s_i = k)[\log(p_k) + \log(f(x_i|\theta_k))]. \tag{3.6}$$

Assim pela definição no passo E, usando $\boldsymbol{\Psi}^t = (\mathbf{p}^t, \boldsymbol{\theta}^t)$ como a estimativa de $\boldsymbol{\Psi}$ na iteração $t$, temos

$$Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^t) = \sum_{i=1}^{n}\sum_{k=1}^{K} \mathbb{E}_{\boldsymbol{\Psi}^t}[\mathbb{I}(s_i = k)|\mathbf{x}][\log(p_k) + \log(f(x_i|\theta_k))]$$
$$= \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}^t[\log(p_k) + \log(f(x_i|\theta_k))]$$
$$= \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}^t \log(p_k) + \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}^t \log(f(x_i|\theta_k)), \tag{3.7}$$

em que

$$\tau_{ik}^t = P(s_i = k|x_i, \boldsymbol{\Psi}^t) = \frac{p_k^t f(x_i|\theta_k^t)}{\sum_{l=1}^{K} p_l^t f(x_i|\theta_l^t)} \tag{3.8}$$

é a estimativa, calculada na iteração $t$, da probabilidade de $x_i$ ser proveniente do componente de índice $k$ dado o seu valor e $\boldsymbol{\Psi}^t$, sendo que o sobrescrito $t$ indica a estimativa do parâmetro na iteração $t$ também para as estimativas das proporções dos componentes da mistura.

No passo M queremos maximizar $Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^t)$ com respeito a $\boldsymbol{\Psi} = (\boldsymbol{p}, \boldsymbol{\theta})$. Como as variáveis $p_k$ e $\theta_k$ aparecem em termos separados em (3.7), podemos calcular as estimativas de máxima verossimilhança separadamente,

$$\boldsymbol{p}^{t+1} = \underset{\boldsymbol{p}}{\operatorname{argmax}} \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}^t \log(p_k),$$

como $\sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}^t = n$ o termo a ser maximizado tem a forma da log-verossimilhança de uma distribuição multinomial, a menos de uma constante que podemos ignorar. Logo

$$p_k^{t+1} = \frac{\sum_{i=1}^{n} \tau_{ik}^t}{n}, \text{ para todo } k. \tag{3.9}$$

Como os parâmetros $\theta_k$'s não se restringem, como no caso dos pesos $p_k$'s que devem somar 1, a atualização de $\boldsymbol{\theta}$ pode ser feita para cada $\theta_k$ individualmente, tomando-se $k \in \{1, \ldots, K\}$ qualquer, temos

$$\theta_k^{t+1} = \underset{\theta_k \in \Theta}{\operatorname{argmax}} \sum_{i=1}^{n} \tau_{ik}^t \log(f(x_i|\theta_k)). \tag{3.10}$$

A fórmula acima é similar à do EMV para uma amostra aleatória simples de tamanho $n$ proveniente de uma distribuição $f(x|\theta_k)$, mas cada ponto amostral $x_i$ contribui com um peso $\tau_{ik}$.

Por exemplo, se aplicarmos o algoritmo EM à uma mistura de normais e $\theta_k = (\mu_k, \sigma_k^2)$, com $\mu_k$ sendo a média da distribuição componente $k$ e $\sigma_k^2$ sua variância, então

$$\mu_k^{t+1} = \frac{\sum_{i=1}^n \tau_{ik} x_i}{\sum_{i=1}^n \tau_{ik}}$$
$$[\sigma_k^2]^{t+1} = \frac{\sum_{i=1}^n \tau_{ik}(x_i - \mu_k^{t+1})^2}{\sum_{i=1}^n \tau_{ik}}.$$

Portanto, quando o algoritmo EM é aplicado a um modelo de mistura, as estimativas podem ser atualizadas, segundo os passos:

1. Faça $t \leftarrow 0$ e escolha $\mathbf{\Psi}^0$ uma estimativa inicial dos parâmetros da mistura;

2. Usando a estimativa atual $\mathbf{\Psi}^t = (\mathbf{p}^t, \boldsymbol{\theta}^t)$ calcule $\boldsymbol{\tau}_i^t = (\tau_1^t, \ldots, \tau_k^t)$ para $1 \leq i \leq n$, como descrito em (3.8);

3. Calcule $\boldsymbol{p}^{t+1}$ e $\boldsymbol{\theta}^{t+1}$ usando $(\boldsymbol{\tau}_i^t)_{1 \leq i \leq n}$ como em (3.9) e em (3.10);

4. Faça $t \leftarrow t+1$ e retorne ao passo 2.

Deste modo, o algoritmo EM aplicado a modelos de mistura consiste numa alternância entre utilizar as estimativas do parâmetro $\mathbf{\Psi}$ para atualizar as estimativas de probabilidade $\tau_{ik}$, e usar estas para atualizar as estimativas dos parâmetros.

### 3.1.5 Aplicação numa Mistura de duas Poissons

Simulamos uma amostra de tamanho $n = 50$ de duas Poissons com médias $\lambda_1 = 5$, $\lambda_2 = 15$ e pesos $p_1 = p_2 = 0,5$, e através do algoritmo EM obtivemos estimativas destes parâmetros, rodando 299 iterações a partir de estimativas iniciais dos parâmetros.

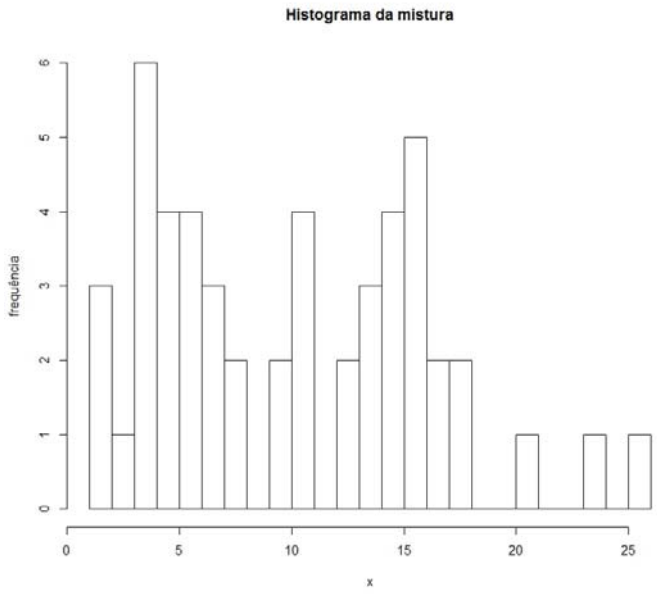


Figura 3 – Frequência dos dados

Começando com as estimativas $\lambda_1^1 = 11$, $\lambda_1^2 = 20$, $p_1^1 = 0,3$ e $p_2^1 = 0,7$, iteramos alternativamente, através das estimativas dos parâmetros da mistura, $(\lambda_1^t, \lambda_2^t, p_1^t, p_2^t)$, calculamos as estimativas das probabilidades $\tau_{ik}^t$, e com estas últimas calculamos novas estimativas $(\lambda_1^{t+1}, \lambda_2^{t+1}, p_1^{t+1}, p_2^{t+1})$.

O tamanho da subamostra que contém os pontos provenientes do componente 1, com $p_1 = 0,5$ é uma variável aleatória com distribuição Binomial$(50; 0,5)$, que tem média 25, nesta simulação 23 pontos amostrais foram gerados do primeiro componente e 27 do segundo.

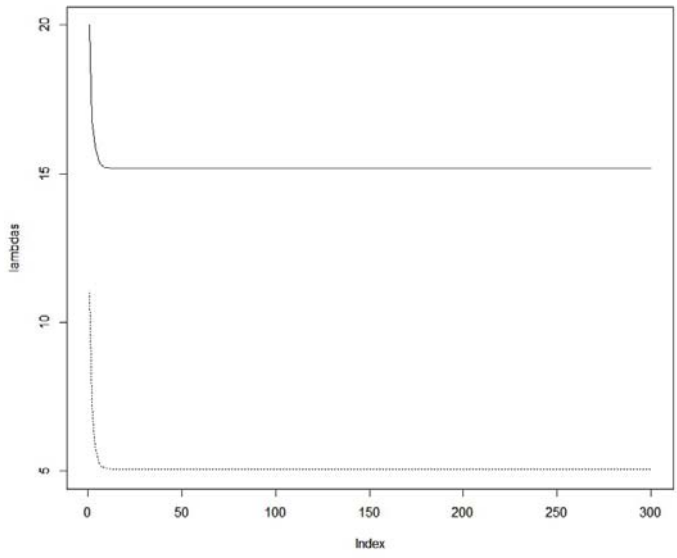


Figura 4 – Sequências de médias estimadas

A média dos pontos gerados pelo primeiro componente é $5,14$, a linha pontilhada que representa as estimativas de $\lambda_1$ se fixou em $5,07$, já a média dos pontos gerados pelo segundo componente é $15,04$, que é representado pela linha contínua, que se estabeleceu em $15,19$.

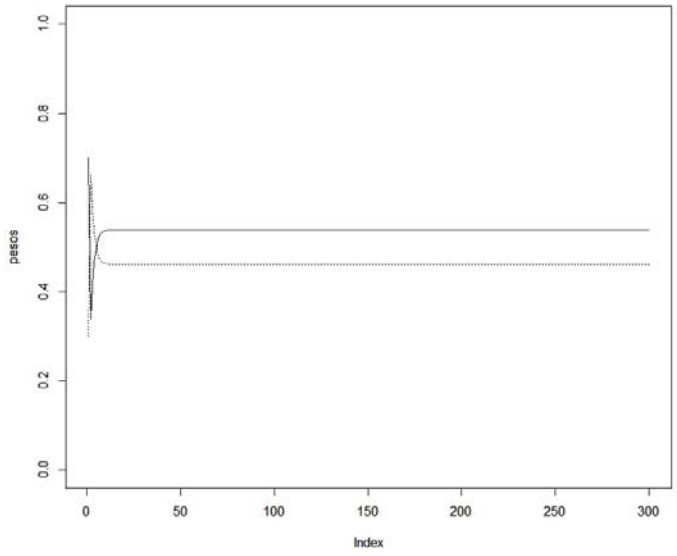


Figura 5 – Sequências de pesos estimados

Como o primeiro componente gerou menos pontos amostrais que o segundo, por uma questão da natureza aleatória da seleção de componentes dentro de uma mistura de distribuições, a estimativa de $p_2$ ficou um pouco acima da estimativa de $p_1$. O valor real de $p_1$ é 0,5, mas de fato na amostra 23/50 = 46% dos pontos são provenientes do componente 1, a estimativa de $p_1$ ficou em $0,46$, já a estimativa de $p_2$, que complementa a de $p_1$, ficou em $0,54$.

Para cada estimativa $\boldsymbol{\Psi}^t = (\boldsymbol{p}^t, \boldsymbol{\theta}^t)$ calculamos $\log L(\Psi^t|\mathbf{x})$, como as estimativas dos parâmetros se estabilizaram rapidamente e a estimativa da log-verossimilhança é função contínua das estimativas esta também se estabilizou rapidamente.

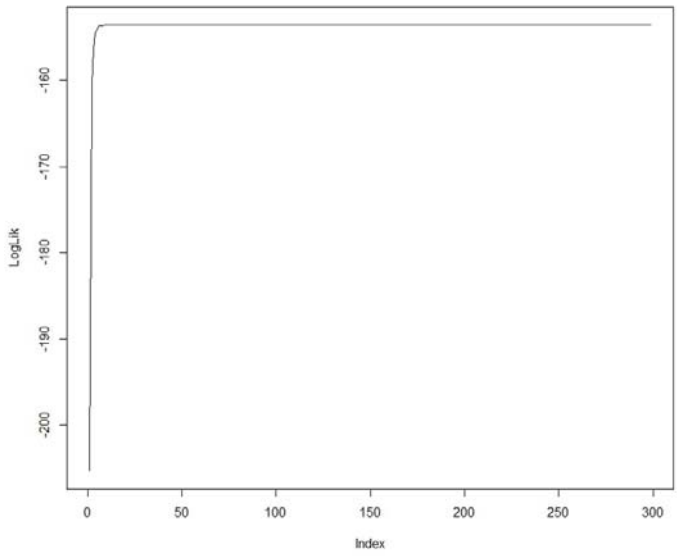


Figura 6 – Sequências de log-verossimilhanças obtidas

Note que, em consonância com o Teorema 2 da seção 3.1.2 que garante a convergência do EM, verifica-se pela Figura 6 que a sequência $\log L(\mathbf{x}|\mathbf{\Psi}^t)$ gerada é monótona crescente.

### 3.1.6 Algoritmo GEM

O algoritmo EM generalizado, também chamado de algoritmo GEM é uma variação do algoritmo EM que exige uma condição mais fraca quando se define a próxima estimativa $\mathbf{\Psi}^{t+1}$ no passo 3 do algoritmo.

Não exigimos uma escolha de $\mathbf{\Psi}^{t+1}$ que satisfaça $Q(\mathbf{\Psi}^{t+1};\mathbf{\Psi}^t) \geq Q(\mathbf{\Psi};\mathbf{\Psi}^t)$ para todo $\mathbf{\Psi}$, mas apenas que satisfaça $Q(\mathbf{\Psi}^{t+1};\mathbf{\Psi}^t) \geq Q(\mathbf{\Psi}^t;\mathbf{\Psi}^t)$.

Usando as fórmulas (3.2) e (3.3), e notando que a condição exigida pelo GEM é equivalente à $Q(\mathbf{\Psi}^{t+1},\mathbf{\Psi}^t) - Q(\mathbf{\Psi}^t,\mathbf{\Psi}^t) \geq 0$ concluímos que o GEM também produz uma sequência de estimativas $(\mathbf{\Psi}^t)_{t\geq 0}$ de modo que se tenha $\log L(\mathbf{\Psi}^{t+1}|\mathbf{y}) \geq \log L(\mathbf{\Psi}^t|\mathbf{y})$ para todo $t \geq 0$. Portanto, sob a mesma condição de que a função de verossimilhança seja limitada superiormente, existe um valor $L^*$ tal que $\lim_{t\to\infty} \log L(\mathbf{\Psi}^t) = \log L^*$, para qualquer escolha incial de $\mathbf{\Psi}^0$.

### 3.1.7 EM Estocástico

O algoritmo EM estocástico, criado por Celeux e Diebolt (1985) é uma versão modificada do algoritmo EM em que se gera uma cadeia de Markov no espaço de estados $\mathscr{S}$, das variáveis não-observáveis $\mathbf{S}$ que indicam a proveniência de cada ponto amostral à um dos componentes da mistura, com um número de grupos fixado $K$.

Iniciamos com uma escolha arbitrária de valores para as variáveis não-observáveis, a qual chamamos de $\mathbf{s}^0 = (s_1^0, \ldots, s_n^0)$, com $s_i \in \{1, 2, \ldots, K\}$ para todo $i$ e de modo que para todo $1 \leq k \leq K$ exista algum $x_i$ tal que $s_i = k$. Esta configuração $\mathbf{s}^0$ define uma partição dos dados $\mathbf{x}$ em $K$ grupos. Seja $A_k^t = \{x_i; x_i \in \mathbf{x} \text{ e } s_i^t = k\}$, assim $A_1^t, \ldots, A_K^t$ é a partição de $\mathbf{x}$ induzida por $\mathbf{s}^t$.

Se o nosso modelo de mistura tem como vetor de probabilidades (pesos) dos componentes $\mathbf{p} = (p_1, \ldots, p_K)$ e a amostra $\mathbf{x}$ é aleatória simples de tamanho $n$, então o vetor dos números de elementos provenientes de cada componente $k$ é o vetor aleatório

$$(n_1, \ldots, n_K) \sim \text{Multinomial}(n; p_1, \ldots, p_K),$$

e a estimativa de máxima verossimilhança para $p_k$ é $\frac{n_k}{n}$, $k = 1, \ldots, K$.

Além disso, os elementos $x_i$ em $\mathbf{x}$ tais que $s_i = k$ formam uma subamostra, que contém todas e somente as observações que provém do componente $k$. Usando cada subamostra que contém as observações que se originaram em cada componente, podemos estimar os parâmetros $\theta_k$ para cada $1 \leq k \leq K$.

A cada iteração substituímos $\mathbf{s}^t$ por um novo vetor $\mathbf{s}^{t+1}$. Ao par $(\mathbf{x}, \mathbf{s}^t)$ chamamos de amostra pseudo-completa ou dados ampliados na iteração $t$. Mantendo sempre $\mathbf{x}$ fixado selecionamos novos vetores $\mathbf{s}^t$ de acordo com as estimativas atuais.

O algoritmo consiste na alternância entre um passo M determinístico e um passo S estocástico.

1. Defina $t = 0$ e escolha $\mathbf{s}^0$ arbitrário com $K$ elementos únicos;

2. (Passo M) Usando a partição $A_1^t, A_2^t, \ldots, A_K^t$ induzida por $\mathbf{s}^t$, calcule as estimativas de máxima verossimilhança de $\boldsymbol{p}^t$ e $\boldsymbol{\theta}^t$ por

$$\begin{aligned} p_k^t &= \frac{\sum_{i=1}^n \mathbb{I}(s_i^t = k)}{n} = \frac{\#A_k^t}{n}, \\ \theta_k^t &= \underset{\theta_k \in \Theta_k}{\text{argmax}} \log f(A_k^t | \theta_k) = \underset{\theta_k \in \Theta_k}{\text{argmax}} \sum_{x_i \in A_k^t} \log f(x_i | \theta_k), \end{aligned}$$

   em que $\Theta_k$ é o espaço paramétrico de $\theta_k$;

3. Calcule as probabilidades de cada $x_i$ pertencer a cada componente segundo as estimativas calculadas no passo anterior,

$$\tau_{ik}^t = P(s_i = k | x_i, \boldsymbol{\Psi}^t) = \frac{p_k^t f(x_i | \theta_k^t)}{\sum_{l=1}^K p_l^t f(x_i | \theta_l^t)},$$

   para todo $k$;

4. (Passo S) Gere um novo vetor $\mathbf{s}^{t+1}$ segundo a distribuição estimada $\boldsymbol{\tau}^t = (\tau_{ik}^t)_{(i,k)\in\{1,\ldots,n\}\times\{1,\ldots,K\}}$, *i.e.*, cada $s_i^{t+1}$ é gerado segundo a distribuição Multinomial$(1;\tau_i^t)$ = Multinomial$(1;\tau_{i1}^t,\ldots,\tau_{iK}^t)$;

5. Faça $t \leftarrow t+1$ e retorne ao passo 2.

Note que na primeira iteração uma alternativa é, em vez de partimos de uma configuração $\mathbf{s}^0$ poderíamos começar diretamente com uma estimativa inicial $\boldsymbol{\Psi}^0 = (\boldsymbol{\theta}^0, \boldsymbol{p}^0)$ sem atribuir ainda valores às variáveis não-observáveis, e prosseguir daí para o passo 3 e em diante.

Para um $K$ fixado, representando o número de grupos, através da sequência de atualização das estimativas

$$\mathbf{s}^t \to \boldsymbol{\Psi}^t \to (\tau_{ik}^t) \to \mathbf{s}^{t+1},$$

como somente o passo S é estocástico, conhecendo $\mathbf{x}$, a distribuição sobre $\mathbf{s}^{t+1}$ dado $\mathbf{s}^t$ independe do conjunto dos $\mathbf{s}^l$'s com $l < t$. Então geramos uma cadeia de Markov $\mathbf{S}^0 = \mathbf{s}^0, \mathbf{S}^1 = \mathbf{s}^1, \mathbf{S}^2 = \mathbf{s}^2, \ldots$, no espaço de estados $\mathscr{S}_K = \{\mathbf{s}; \mathbf{s} \in \mathscr{S}$ e $\mathbf{s}$ tem $K$ elementos únicos$\}$ se não permitirmos que um elemento da partição se esvazie.

Esta cadeia possui um número finito de estados possíveis, admitindo que o tamanho da amostra é finito, assim pela teoria de cadeias de Markov quando a cadeia é irredutível e aperiódica ela possui uma distribuição estacionária $\boldsymbol{\pi}$ para o qual a distribuição da cadeia converge, independentemente do estado $\mathbf{s}^0$ inicial.

A cadeia ser irredutível significa que dados quaisquer dois estados $\mathbf{s}_1, \mathbf{s}_2 \in \mathscr{S}_K$, existe probabilidade positiva de transição de $\mathbf{s}_1$ para $\mathbf{s}_2$ em algum número $n_0 > 0$ de passos. Um estado da cadeia $\mathbf{s}_1 \in \mathscr{S}_K$ é aperiódico se não existe nenhum número maior que 1 que divida todos os elementos do conjunto $R$,

$$R = \{r;\ \text{há probabilidade positiva de que partindo de } \mathbf{s}_1 \text{ se retorne para } \mathbf{s}_1 \text{ em } r \text{ passos}\},$$

e dizemos que a cadeia é aperiódica se todos os seus estados o forem.

Mostraremos que no caso de uma mistura de distribuições de uma mesma família paramétrica que se situa dentro da família exponencial (ex.: mistura de normais, Poissons, gamas), ocorre a convergência para uma distribuição estacionária $\boldsymbol{\pi}$.

Considere um conjunto de dados $\mathbf{x}$ proveniente de uma mistura com função de densidade de probabilidade, ou função massa de probabilidade,

$$f = \sum_{k=1}^{K} p_k f(x|\theta_k)$$

em que o conjunto suporte, $\text{supp}(\theta) = \{x; f(x|\theta) > 0\}$, é constante em relação a $\theta$.

Para cada ponto $x_i$,

$$\begin{aligned} f(s_i = k|x_i, \boldsymbol{\Psi}) &\propto f(x_i|S_i = k, \boldsymbol{\Psi}) P(S_i = k|\boldsymbol{\Psi}) \\ &\propto f(x_i|\theta_k) p_k, \end{aligned}$$

como $x_i$ provém de uma distribuição $f(\cdot|\theta_q)$, para algum $q$, e $\text{supp}(\theta)$ não depende de $\theta$, então $\forall\theta\ f(x_i|\theta) > 0$.

Logo, se substituirmos $\theta_k$ e $p_k$ por estimativas $\theta_k^t$ e $p_k^t > 0$, temos $f(s_i = k|x_i, \theta_k^t, p_k^t) > 0$ para cada $k$.

Quando sorteamos um novo valor $s_i^{t+1}$, temos pelo passo (3) que

$$P(s_i^{t+1} = k|x_i, \boldsymbol{\Psi}^t) \propto f(x_i|\theta_k^t) p_k^t > 0, \text{ para todo } k.$$

Logo, dadas duas configurações $\mathbf{s}^t, \mathbf{u} \in \mathscr{S}_K$, temos que $P(s_i^{t+1} = u_i|x_i, \boldsymbol{\Psi}^t) > 0$ para todo $u_i \in \{1, \ldots, K\}$ e $\boldsymbol{\Psi}^t$, assim

$$P(\boldsymbol{S}^{t+1} = \mathbf{u}|\mathbf{x}, \boldsymbol{\Psi}^t) = \prod_{i=1}^{n} P(s_i^{t+1} = u_i|x_i, \boldsymbol{\Psi}^t) > 0,$$

o que implica que a probabilidade de transição entre dois estados quaisquer, mesmo em um único passo, é positiva.

Como consequência a cadeia é irredutível e aperiódica, o caso de uma mistura de distribuições da família exponencial segue como um caso particular.

Para um ponto $x \in \mathbb{R}^n$, uma distribuição da família exponencial têm densidade ou função massa de probabilidade da forma

$$f(x|\theta) = a(x) \exp(< \alpha(\theta), \beta(x) > - b(\theta)), \tag{3.11}$$

com $\alpha(\theta), \beta(x) \in \mathbb{R}^n$ e $<,>$ sendo o produto escalar usual. Como $\exp(y) > 0$ para todo $y \in \mathbb{R}$, é o termo $a(x)$ que define se $f(x|\theta)$ é positivo ou zero, assim se $f(x|\theta) > 0$ então para qualquer $\theta^*$, $f(x|\theta^*) > 0$.

Logo, para qualquer mistura em que as distribuições componentes se expressam como em (3.11), $\text{supp}(\theta)$ é função constante de $\theta$ e o resultado se aplica.

### 3.1.8 EM Estocástico Aplicado a Mistura de Poissons

Como um exemplo, geramos uma amostra de tamanho $n = 50$ de uma mistura de duas Poisson's com $\lambda_1 = 5$ e $\lambda_2 = 15$, $p_1 = p_2 = 0.5$.

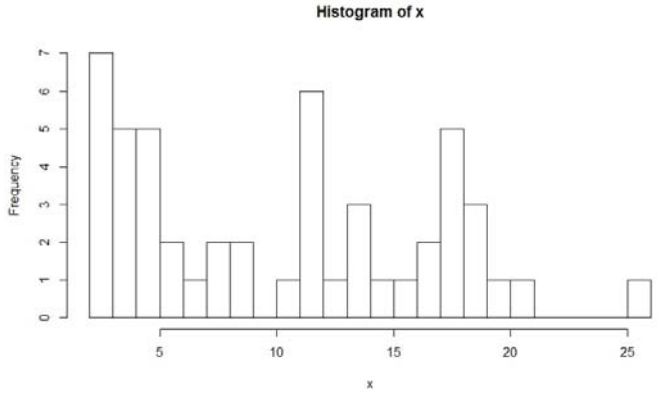


Figura 7 – Frequência dos dados da mistura de Poisson

Sorteamos então aleatoriamente as variáveis não-observáveis $\mathbf{s}^0$, com cada $s_i^0$ tomando valores em $\{1,2\}$, e a partir das estimativas dos parâmetros induzidas por cada configuração $\mathbf{s}^t$, re-sorteamos novas variáveis $\mathbf{s}^{t+1}$ e assim alternadamente. Gerando uma sequência de estimativas para $\lambda_1$, $\lambda_2$, $p_1$ e $p_2$.

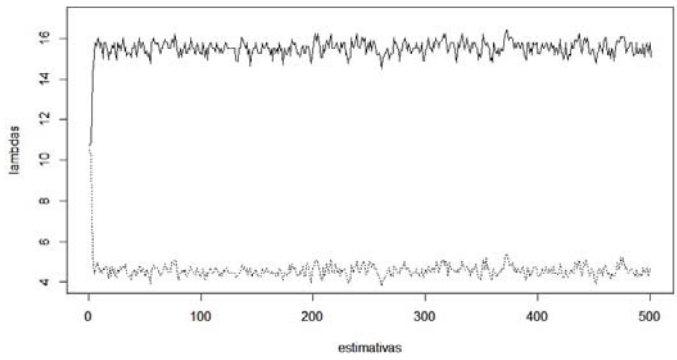


Figura 8 – Sequência de estimativas dos parâmetros $\lambda_1$ (linha pontilhada) e $\lambda_2$ (linha contínua)

Era esperado que nas primeiras iterações tivéssemos $\lambda_1^t \approx \lambda_2^t$, pois iniciamos o algoritmo atribuindo a cada ponto $x_i$ um rótulo $s_i$, tomando com probabilidades iguais $s_i = 1$ e $s_i = 2$.

Observe, na Figura 8 que a sequência de estimativas atingiu uma região de estabilidade com poucas iterações. A escolha de 500 iterações foi feita para que pudéssemos observar o comportamento da sequência nas primeiras iterações e ao longo do processo.

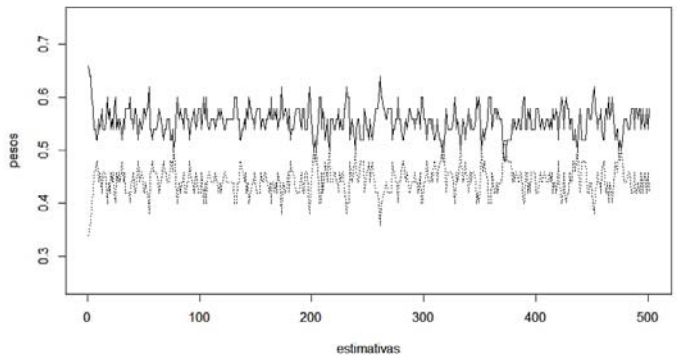


Figura 9 – Sequência de estimativas dos pesos $p_1$ (linha pontilhada) e $p_2$ (linha contínua)

As estimativas das probabilidades $p_1$ e $p_2$ se mantiveram próximas de $0,5$ durante todo o procedimento, note que como as estimativas $p_i^t$ são calculadas pela frequência relativa de pontos amostrais que estão atribuídos ao grupo $i$ na iteração $t$, devemos ter $p_1^t + p_2^t = 1$ para toda iteração $t$, o que explica a simetria dos gráficos em torno do eixo horizontal $y = 0,5$.

Cada configuração $(\mathbf{x}, \mathbf{s}^t)$ induz uma estimativa $\boldsymbol{\Psi}^t = (\boldsymbol{\lambda}^t, \mathbf{p}^t)$ que por sua vez define uma log-verossimilhança $\log L(\boldsymbol{\Psi}^t|\mathbf{x}) = \log f(\mathbf{x}|\boldsymbol{\Psi}^t)$.

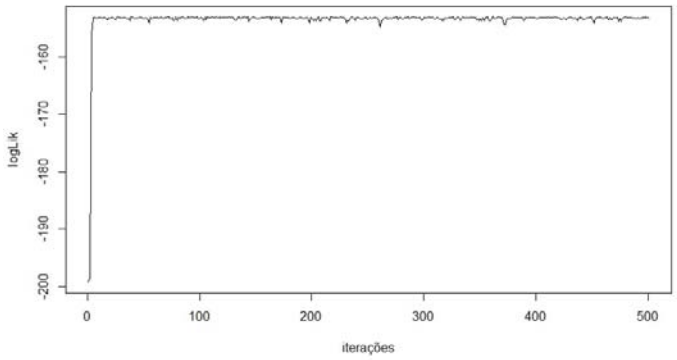


Figura 10 – Sequência de log-verossimilhanças obtidas

Como a log-verossimilhança depende dos parâmetros, assim que as estimativas destes se mantém estáveis em uma região, também se estabiliza a sequência de log-verossimilhanças.

Note que as iterações atingiram rapidamente uma região de estabilidade mas o EM Estocástico, ao contrário do EM tradicional, não garante a monotonicidade da sequência de log-verossimilhanças, e por vezes $\log L(\boldsymbol{\Psi}^{t+1}|\mathbf{x}) < \log L(\boldsymbol{\Psi}^t|\mathbf{x})$.

Podemos também olhar para a quantidade de pontos que foram classificados corretamente ao longo das iterações, com o devido cuidado de considerar somente a partição induzida por cada $\mathbf{s}^t$ e não os rótulos em si. Por exemplo, para uma amostra de 5 dados, a partição $\mathbf{s} = (1,1,2,2,2)$

é igual a de $\mathbf{u} = (2,2,1,1,1)$, então devemos contar os acertos um-a-um comparando os dois vetores diretamente e no caso de permutarmos os rótulos 2 e 1 em $\mathbf{u}$, uma permutação de 2-1 é dada pela função $f(s) = -s+3$ para $s = 1$ ou 2. Assim, sendo $\mathbf{s}$ o vetor com o verdadeiro valor das variáveis não-observáveis podemos contar o número de acertos em um vetor $\mathbf{s}^t$ por

$$m = \max\left\{\sum_{i=1}^{n}\mathbb{I}(s_i = s_i^t), \sum_{i=1}^{n}\mathbb{I}(s_i = -s_i^t + 3)\right\}.$$

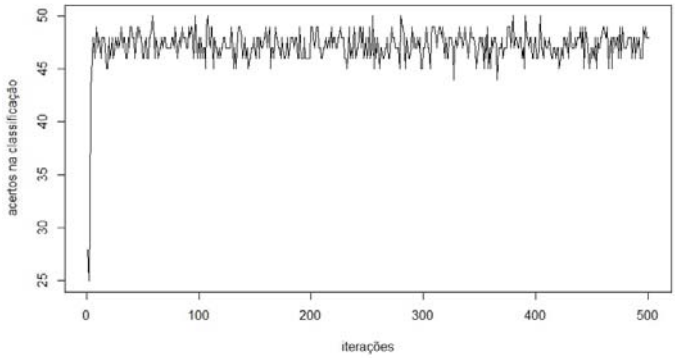


Figura 11 – Sequência de $m$ acertos obtidas nas classificações dos dados

A tendência do comportamento do EM estocástico ao começar com uma configuração arbitrária é buscar configurações que induzam estimativas dos parâmetros $\hat{\boldsymbol{\Psi}} = (\hat{\mathbf{p}}, \hat{\boldsymbol{\theta}})$ sob a qual a função de verossimilhança dos dados observados é maior.

Esta versão do EM tem a vantagem sobre o EM tradicional de ser menos sensível à estimativa inicial do parâmetros, oferecendo a possibilidade de se escapar de uma região no espaço paramétrico onde esteja um máximo local da função de log-verossimilhança. No entanto, o EM estocástico não possui a qualidade de ser numericamente estável, de modo que a cada nova iteração a estimativa dos parâmetros seja melhor, ou pelo menos tão boa quanto, no sentido de aumentar a log-verossimilhança da amostra condicionada às estimativas $\boldsymbol{\Psi}^t$. Por isso, ocasionalmente pode ser conveniente começar o algoritmo com o EM estocástico até a obtenção de uma boa estimativa e deste ponto passar a usar o EM tradicional fazendo um algoritmo híbrido EM - EM Estocástico.

Os autores Celeux e Diebolt (1985) propuseram um algoritmo, chamado de SAEM, em que a cada iteração, usando a estimativa atual $\boldsymbol{\Psi}^t$ dos parâmetros, calculamos tanto a próxima estimativa segundo o EM, $\boldsymbol{\Psi}_{EM}^{t+1}$, quanto segundo o EM Estocástico, $\boldsymbol{\Psi}_{SEM}^{t+1}$, e então se define a próxima estimativa como

$$\boldsymbol{\Psi}^{t+1} = (1-\gamma_{t+1})\boldsymbol{\Psi}_{EM}^{t+1} + \gamma_{t+1}\boldsymbol{\Psi}_{SEM}^{t+1},$$

em que $\gamma_0 = 1$, e $\lim_{t\to\infty} \gamma_t = 0$. Deste modo o SAEM começa com o EM Estocástico puro, e aos poucos vai dando mais peso às estimativas do EM, até que o impacto das estimativas do EM Estocástico seja negligenciável.

Sob as condições adicionais de $\lim_{t\to\infty} \gamma_{t+1}/\gamma_t = 1$ e $\sum_{t=1}^{\infty} \gamma_t = \infty$, a sequência $\boldsymbol{\Psi}^t$ converge para um máximo local da função de log-verossimilhança com probabilidade 1, evitando assim a possibilidade de convergência para pontos de sela.

### 3.1.9 EM Estocástico aplicado a Mistura de Normais

Simulamos uma amostra de uma mistura de duas normais bivariadas com médias $\mu_1 = (0,0)$ e $\mu_2 = (5,5)$ e matrizes de variância $\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ e $\Sigma_2 = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$ respectivamente, com 100 observações provenientes de cada normal, e atribuímos a cada ponto aleatoriamente a pertinência a um de dois grupos.

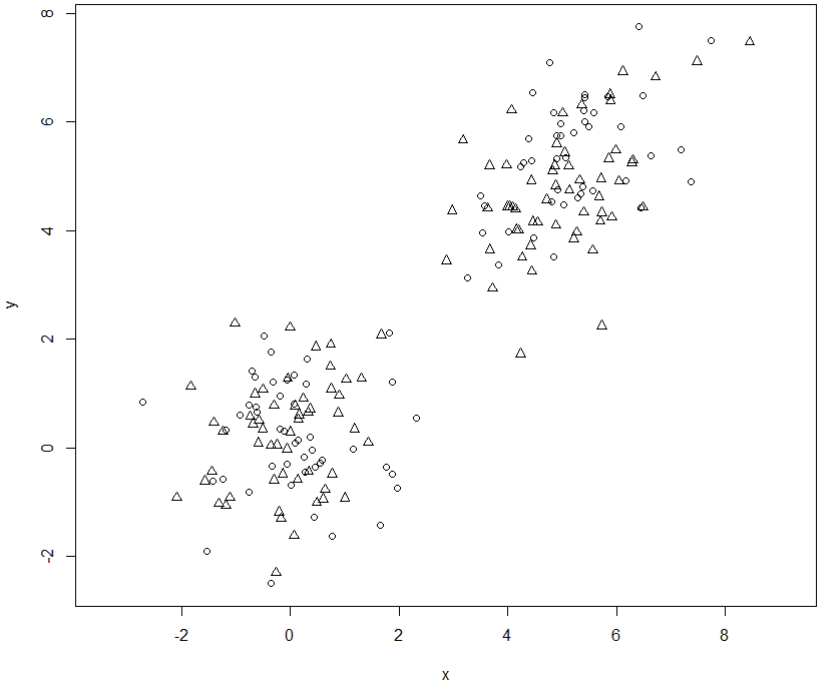


Figura 12 – Gráfico de dispersão dos pontos simulados. Pontos representados com a mesma figura foram atribuídos ao mesmo grupo.

A cada passo usando a amostra ampliada $(\mathbf{x}, \mathbf{s}^t)$ da iteração atual estimamos o parâmetro $\boldsymbol{\Psi}$ da seguinte maneira.

1. Definimos $t = 0$ e geramos arbitrariamente um vetor $\mathbf{s}^0 \in \{1, 2, \ldots, K\}^n$ com $K$ elementos únicos;

2. Calculamos as estimativa $\boldsymbol{p}^t = (p_1^t, \ldots, p_K^t)$ das proporções da mistura de cada grupo, que são obtidas utilizando os estimadores de máxima verossimilhança para as probabilidades de uma distribuição multinomial, que correspondem às frequências empíricas relativas das ocorrências de cada categoria $k$, que neste contexto se contam por $\#\{s_i; s_i = k, 1 \leq i \leq n\}$ . Isto é, para cada grupo $k$ fazemos $p_k^t = n_k/n = \#A_k^t$;

3. Calculamos $\boldsymbol{\theta}^t = (\theta_1^t, \ldots, \theta_K^t)$, a estimativa dos parâmetros dos componentes $k$, estimando cada $\theta_k$ considerando somente os pontos amostrais que foram atribuídos ao grupo $k$, *i.e.* aqueles com $s_i = k$, como uma subamostra de $\mathbf{x}$ que representa a amostra do componente $k$.

   No caso de uma mistura de normais temos,

$$\boldsymbol{\mu}_k^t = \frac{\sum_{i=1}^n \mathbb{I}(s_i = k) x_i}{\sum_{i=1}^n \mathbb{I}(s_i = k)} = \frac{\sum_{x_i; s_i = k} x_i}{\#A_k^t}$$
$$\boldsymbol{\Sigma}_k^t = \frac{\sum_{i=1}^n (x_i - \bar{x})^T (x_i - \bar{x}) \mathbb{I}(s_i = k)}{\sum_{i=1}^n \mathbb{I}(s_i = k)} = \frac{\sum_{x_i; s_i = k} (x_i - \bar{x})^T (x_i - \bar{x})}{\#A_k^t}$$

   para todo $k$. Lembrando que como a distribuição é bivariada, cada $x_i$ é um vetor $(1 \times 2)$;

4. Considerando $x_i$ e as estimativas dos parâmetros obtidas, $\boldsymbol{\Psi}^t = (\boldsymbol{\theta}^t, \boldsymbol{p}^t)$, a estimativa de probabilidade de cada $s_i$ é

$$\tau_{ik}^t = \tau_{ik}^t(\boldsymbol{\Psi}^t, x_i) = \frac{p_k^t \, f(x_i | \theta_k^t)}{\sum_{l=1}^K p_l^t \, f(x_i | \theta_l^t)}, \; \forall \; 1 \leq k \leq K.$$

   Assim, para um $i$ fixo, $\tau_{ik}^t$ define uma distribuição de probabilidade sobre $\{1, \ldots, K\}$. E sorteamos novos valores segundo a distribuição $P(s_i^{t+1} = k | x_i, \boldsymbol{\Psi}^t) = \tau_{ik}$ para todo $k$, obtendo assim $\mathbf{s}^{t+1}$;

5. Fazemos $t \leftarrow t + 1$ e retornamos ao passo 2.

Partindo desta amostra pseudo-completa inicial, as iterações se seguiram de tal modo.

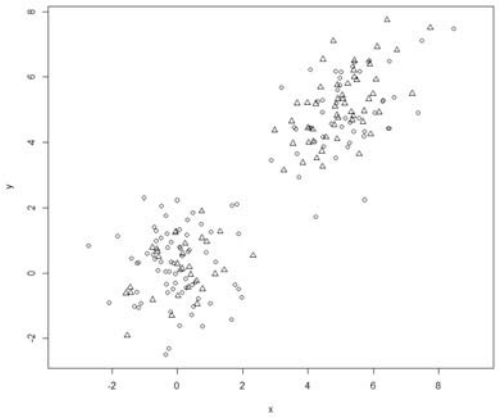


Figura 13 – iteração 15

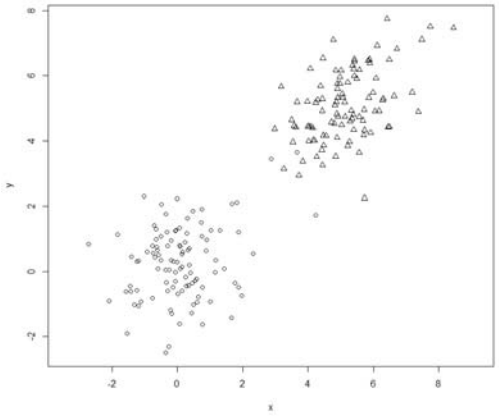


Figura 14 – iteração 30

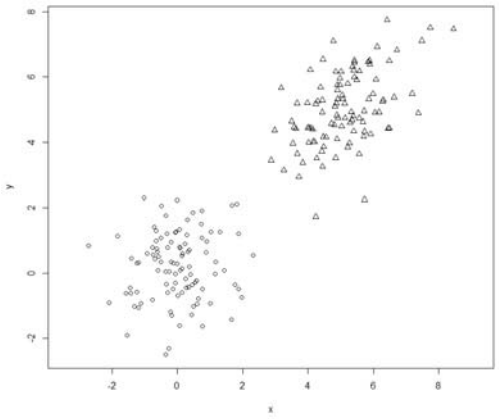


Figura 15 – iteração 35

Da teoria de cadeias de Markov, se a cadeia tem um número finito de estados, é aperiódica (tem probabilidade positiva de retornar ao mesmo estado sem que se exija que todos os tempos

de retorno sejam múltiplos de um mesmo número) e transitiva (dados quaisquer dois estados a probabilidade de partindo de um deles chegar-se no outro é positiva), então a distribuição da cadeia converge para a chamada distribuição estacionária da cadeia.

Nos nossos exemplos, cada ponto tem sempre uma probabilidade positiva de pertencer à cada um dos componentes, *i.e.* cada $\tau_{ik}^{t} > 0$ para qualquer $i$,$k$ e $t$, assim em qualquer ponto qualquer mudança de estados é possível, segue que a cadeia é transitiva e aperiódica, estão a distribuição sobre os $\mathbf{s}$'s converge para a distribuição estacionária.

## 3.2 Processo de Dirichlet

Um processo de Dirichlet (PD) é um processo estocástico, introduzido por Ferguson (1973) cuja realização pode ser interpretada como uma distribuição discreta com massa de probabilidade em infinitos pontos. Como vimos, a distribuição de Dirichlet pode ser interpretada como uma distribuição sobre distribuições Multinomiais, o Processo de Dirichlet é o análogo para uma dimensão infinita.

Em um contexto bayesiano não paramétrico podemos usar o PD como uma ferramenta para agrupar dados sem assumir um número fixo de componentes e permitir que novos grupos sejam criados ao introduzirmos novas observações.

Para desenvolvermos a teoria do Processo de Dirichlet faremos uso de alguns conceitos e definições:

**Definição 1.** Dado um conjunto $\Theta$ dizemos que $\tilde{T} = \{T_1, \ldots, T_m\}$ é uma partição finita de $\Theta$ se

- $T_i \cap T_j = \varnothing$ para todo $i$ e $j$
- $\bigcup_{i=1}^{m} T_i = \Theta$

**Definição 2.** Dizemos que $\mathscr{B}$ é uma $\sigma$-álgebra sobre $\Theta$ se

- $\Theta \in \mathscr{B}$
- se $A \in \mathscr{B}$ então $A^C \in \mathscr{B}$
- se $A_1, A_2, \ldots \in \mathscr{B}$ então $\bigcup_{i=1}^{\infty} A_i \in \mathscr{B}$

**Definição 3.** Dizemos que $(\Theta, \mathscr{B})$ é um espaço mensurável se $\mathscr{B}$ é uma $\sigma$-álgebra em $\Theta$ e os elementos de $\mathscr{B}$ são chamados de conjuntos mensuráveis.

**Definição 4.** Uma função de conjuntos $\mu : \mathscr{B} \to [0, \infty]$ é uma medida em $(\Theta, \mathscr{B})$ quando

- $\mu(\varnothing) = 0$

- Para uma sequência de conjuntos disjuntos $A_1, A_2, \ldots \in \mathscr{B}$ temos
$\mu(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$

Se ainda, $\mu(\Theta) = 1$ então $\mu$ é dita ser uma medida (ou distribuição) de probabilidade.

### 3.2.1 Medidas de Probabilidade Aleatórias

Se uma medida de probabilidade $G$ sobre um espaço $(\Theta, \mathscr{B})$ é escolhida aleatoriamente, então para todo $B \in \mathscr{B}$ temos que $G(B)$ é uma variável aleatória.

Por exemplo, considere $\Theta = \mathbb{R}$, $\mathscr{B}$ como os borelianos em $\mathbb{R}$ e uma variável aleatória $a$ em que $P(a = m) = 1/3$ para $m = 1, 2, 3$. Se $G \sim \mathrm{U}(0, a)$ então é um exemplo de medida de probabilidade aleatória, para $B = (1/2, 1)$ temos que

$$G(B) = \begin{cases} 1/2, \text{ com probabilidade 1/3} \\ 1/4, \text{ com probabilidade 1/3} \\ 1/6, \text{ com probabilidade 1/3.} \end{cases}$$

Escolhas diferentes de $G$ determinam probabilidades diferentes sobre o mesmo conjunto mensurável $B$. Neste exemplo, a distribuição de $G$ é uniforme sobre $\{\mathrm{U}(0,1), \mathrm{U}(0,2), \mathrm{U}(0,3)\}$.

### 3.2.2 Formalização do PD

Dizemos que $G$ está distribuído segundo um PD com parâmetro de concentração $\alpha$ e medida base $H$, *i.e.* $G \sim PD(\alpha, H)$, se $G$ é uma medida de probabilidade sobre o espaço mensurável $(\Theta, \mathscr{B})$ que satisfaz

$$(G(T_1), \ldots, G(T_k)) \sim \text{Dirichlet}(\alpha H(T_1), \ldots, \alpha H(T_k)) \tag{3.12}$$

para qualquer partição finita $\tilde{T} = \{T_1, \ldots, T_k\}$ de $\Theta$.

Note que tomando qualquer partição finita $\tilde{T}$ de $\Theta$, e qualquer medida de probabilidade $G$ sobre $\Theta$, e uma variável aleatória $\theta \sim G$, temos que

$$G(T_i) = P(\theta \in T_i) \geq 0 \text{ para todo } i$$
$$G(\Theta) = P(\theta \in \Theta) = \sum_{i=1}^{k} G(T_i) = 1.$$

Assim, $G$ induz uma distribuição sobre todas tais partições. Como $G$ é uma medida de probabilidade aleatória ($G$ é a realização de uma variável aleatória) as probabilidades $G(T_i)$ são variáveis

aleatórias. Em particular se $G$ está distribuído de modo a satisfazer (3.12) para algum $\alpha > 0$ e alguma distribuição $H$ sobre $\Theta$ então $G$ está distribuído segundo um PD.

Podemos definir, assim como faz Ferguson (1973), o PD em função de um único parâmetro $\tilde{\alpha} = \alpha H$ que não é uma medida de probabilidade exceto quando $\alpha = 1$. Neste caso podemos determinar $(\alpha, H)$ pelas fórmulas $\alpha = \tilde{\alpha}(\Theta)$ e $H = \tilde{\alpha}/\alpha$.

Por um resultado de Ferguson (1973), sabemos que se $G$ é realização de um Processo de Dirichlet então é uma distribuição discreta com massa de probabilidade em infinitos pontos com probabilidade 1. Isso nos diz que para alguma distribuição de probabilidade $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots)$ e alguma sequência $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots)$ que contém os pontos de $\Theta$ nos quais $G(\theta_i) = \pi_i$, então $G$ é da forma

$$G(\theta) = \sum_{i=1}^{\infty} \pi_i \delta(\theta = \theta_i), \text{ para todo } \theta \in \Theta$$

em que $\delta(\theta = \theta_i) = \mathbb{I}(\theta = \theta_i)$, e para qualquer $B \in \mathscr{B}$ temos

$$G(B) = \sum_{\substack{i=1,2,\ldots \\ \theta_i \in B}} \pi_i.$$

Aqui usamos a notação '$\theta|G \sim G$' para significar que dada a realização $G$ do PD, e portanto os parâmetros desta distribuição, a variável $\theta$ está distribuída conforme a distribuição $G$, isto é, $P(\theta = \theta_i|G) = \pi_i$ para todo $i$.

Se $V = (V_1, \ldots, V_k) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_k)$ [1] então $\mathbb{E}(V_i) = \dfrac{\alpha_i}{\sum_{l=1}^{k} \alpha_l}$ para todo $i$, assim para todo $B \in \mathscr{B}$ se considerarmos a partição $\{B, B^C\}$ temos

$$(G(B), G(B^C)) \sim \text{Dirichlet}(\alpha H(B), \alpha H(B^C))$$
$$\mathbb{E}(G(B)) = \frac{\alpha H(B)}{\alpha H(B) + \alpha H(B^C)} = H(B),$$

o que nos diz que em média uma realização do processo de Dirichlet tem a distribuição $H$. Também pelo fato de que $\text{Var}(V_i) = \dfrac{\alpha_i(\sum_{l=1}^{k} \alpha_l - \alpha_i)}{(\sum_{l=1}^{k} \alpha_l)^2(\sum_{l=1}^{k} \alpha_l - \alpha_i)}$ temos que

$$\text{Var}(G(B)) = \frac{\alpha H(B)(\alpha - \alpha H(B))}{\alpha^2(\alpha + 1)} = \frac{H(B)(1 - H(B))}{\alpha + 1}.$$

Isto nos mostra que a esperança de $G$ é função constante de $\alpha$ sendo determinada apenas por $H$, no entanto a variância depende de ambos os parâmetros e para um $\alpha$ fixado é mínima quando

[1] Com isso queremos na verdade dizer que $(V_1, \ldots, V_{k-1})$ segue a distribuição especificada em (2.5) com parâmetro $(\alpha_1, \ldots, \alpha_k)$

$H(B) = 1/2$. O fato de que $Var(G(B)) \propto 1/(\alpha+1)$ justifica o nome de $\alpha$ de parâmetro de concentração.

Além disso, as marginais de $(G(B), G(B^C))$ seguem uma distribuição Beta satisfazendo

$$G(B) \sim \text{Beta}(\alpha H(B), \alpha - \alpha H(B^C)).$$

### 3.2.3 Distribuição a posteriori

Se $G \sim PD(\alpha, H)$ e $\theta_1|G, \ldots, \theta_{t-1}|G \sim G$ então a distribuição de $G$ condicionada aos valores assumidos pelos $\theta$'s também é um PD com os seguintes parâmetros

$$G|\theta_1, \ldots, \theta_{t-1} \sim PD\left(\alpha + t - 1, \frac{\alpha H + \sum_{i=1}^{t-1} \delta(\theta = \theta_i)}{\alpha + t - 1}\right).$$

### 3.2.4 Preditiva a posteriori do PD

Blackwell e MacQueen (1973) criaram uma definição equivalente do PD como a distribuição limite de uma urna de Polya modificada. Quando em uma urna em um dado instante existem diferentes tipos de bolas, a probabilidade de se retirar aleatoriamente uma bola de um tipo específico corresponde à proporção de bolas daquele tipo na urna naquele instante.

Esta analogia é conhecida como a urna de Blackwell-McQueen, em que o espaço $\Theta$ é tratado como um espaço de cores para as bolas da urna e então a distribuição base do PD $H$ se torna uma distribuição sobre tais cores.

Suponha que começamos com uma urna sem bolas, no primeiro passo $t = 1$, com probabilidade 1, selecionamos uma cor $\theta_1$ de acordo com a distribuição $H$ e então inserimos uma bola na urna com a cor selecionada, a partir deste ponto a cada iteração $t \geq 2$ temos que $t-1$ é o número de bolas na urna e com probabilidade $\alpha/(\alpha+t-1)$ geramos uma nova cor $\theta_t$ segundo a distribuição $H$ e inserimos uma bola desta cor, e com probabilidade $1-\alpha/(\alpha+t-1)$ selecionamos aleatoriamente uma bola da urna, retornamos esta bola e inserimos uma nova da mesma cor.

Assim, para todo $t$, a distribuição de $\theta_t$ é uma mistura de $H$ com a distribuição da urna definida pelas $(t-1)$ bolas e

$$P(\theta_t \in B) = \frac{\alpha}{\alpha+t-1} H(B) + \frac{t-1}{\alpha+t-1} \sum_{i=1}^{t-1} \frac{\mathbb{I}(\theta_i \in B)}{(t-1)}$$

para todo $B \in \mathscr{B}$. Esta fórmula é também válida para $t = 1$, uma vez que a distribuição de $\theta_1$ se reduz à $H$.

Esta urna pode ser resumida no seguinte procedimento

1. Começe com $t = 1$ e gere $\theta_1 \sim H$;

2. Para $t > 1$, gere $\theta_t|\theta_1,\ldots,\theta_{t-1} \sim G_t(\theta_1,\ldots,\theta_{t-1})$;

em que $G_t(\theta_1,\ldots,\theta_{t-1}) = \dfrac{\alpha H + \sum_{i=1}^{t-1}\delta(\theta_t = \theta_i)}{\alpha + t - 1}$ e $\delta(\theta_t = \theta_i) = \mathbb{I}(\theta_t = \theta_i)$.

Note que $\theta_1$ é sempre gerado de $H$, enquanto que para $t > 1$, $\theta_t$ pode coincidir com um valor anterior ou assumir um novo valor gerado de $H$.

Blackwell e MacQueen (1973) demonstraram que $\lim_{t\to\infty} G_{t+1}(\theta_1,\ldots,\theta_t) = G \sim PD(\alpha,H)$ com probabilidade 1. Desta maneira, o procedimento descrito oferece uma maneira de se construir uma medida $G$ sobre $\Theta$ que é uma realização de um PD$(\alpha,H)$.

Note que, dados $\alpha$, $H$ e a configuração das urna no tempo $t$ ($G_t$), então a distribuição de $G_{t+1}$ independe do conjunto dos $G_u$'s com $u < t$ (não importa como que a urna no tempo atual tomou sua forma). Assim, condicionado em $\alpha$ e $H$, a sequência $G_1, G_2, \ldots$ é uma cadeia de Markov cujos estados são medidas de probabilidade aleatórias. Podemos portanto ver uma realização de um PD como o limite quase certo de uma sequência de distribuições em uma cadeia de Markov definida pela urna de Blackwell-MacQueen.

Observe que, se a distribuição base $H$ é contínua, a distribuição de $\theta_1$ é contínua, para $t > 1$ a distribuição condicional de $\theta_t$ é mista tendo um componente contínuo e outro discreto, e independente da forma de $H$ como seu peso tende à zero temos que $G$ é uma distribuição discreta, ou mais precisamente, como $G$ é aleatório dizemos que $G$ é uma distribuição discreta com probabilidade 1.

Como a proporção $\alpha/(\alpha+t-1)$ do componente com distribuição $H$ tende à zero quando $t$ tende à infinito, o limite de $G_t$ é o mesmo que o limite da distribuição empírica. Isto atribui uma característica de consistência ao PD. Uma vez que o limite da distribuição empírica é a distribuição teórica da qual provém os dados (Teorema de Glivenko-Cantelli).

### 3.2.5 Processo do Restaurante Chinês

Uma formulação equivalente da urna de Blackwell-MacQueen pode ser dada por una analogia chamada de Processo do Restaurante Chinês com parâmetro de concentração $\alpha > 0$ (PRC($\alpha$)).

Considere uma sequência infinita enumerável de mesas $M_1, M_2, M_3, \ldots$ que se encontram dentro de um restaurante chinês, sendo que cada mesa pode comportar um número infinito de indivíduos, e uma outra sequência da mesma natureza de clientes $C_1, C_2, \ldots$ que entram no restaurante, na ordem listada, e ocupam um lugar em uma das mesas um por um a cada intervalo unitário de tempo $\Delta t = 1$. Denotamos por $\#M_{i,t}$ o número de clientes sentados na mesa de índice $i$ no tempo $t$.

Suponha que a relação entre os clientes e as mesas obedece a tais regras

1. Para $t = 1$ o cliente $C_1$ se senta na mesa $M_1$ com probabilidade 1;

2. Para $t \geq 2$, sendo $\text{MAX} = \text{MAX}_{t-1} = \max\{M_i; \#M_{i,t-1} > 0\}$, com probabilidade $\alpha/(\alpha + t - 1)$ o cliente $C_t$ opta por sentar na mesa vazia $M_{\text{MAX}+1}$ e com probabilidade $\#M_{i,t-1}/(\alpha + t - 1)$, para $i \leq \text{MAX}_{t-1}$, opta por sentar na mesa ocupada $M_i$.

Note que dado que o cliente optou sentar-se numa mesa não-vazia, a probabilidade dele selecionar uma mesa $M_i$ é proporcional ao número de clientes anteriores que estão sentados naquela mesa. Assim,

$$P(C_t \text{ sentar na mesa } M_i | i \leq \text{MAX}_{t-1}) = \frac{\#M_{i,t-1}}{t-1}$$
$$P(C_t \text{ sentar em } M_{\text{MAX}_{t-1}+1}) = \frac{\alpha}{\alpha + t - 1}$$

que é a mesma distribuição que existe sobre as cores das bolas na urna de Blackwell-MacQueen se considerarmos que um cliente sentar-se em uma mesa ocupada é análogo a selecionarmos uma bola da urna e então a repormos e adicionarmos uma de mesma cor, e um cliente sentar-se numa mesa nova é análogo a selecionarmos uma nova cor de $H$ e inserirmos uma bola com esta cor.

Assim, o processo do Restaurante Chinês particiona os clientes de maneira equivalente ao modo em que a urna de Blackwell-MacQueen particiona as bolas. A diferença é que no processo da urna ainda atribuímos valores $\theta_i$ (cores) para cada grupo de bolas, enquanto o PRC só nos fornece a partição.

Seja $\gamma_i(t-1) = \dfrac{\#M_{i,t-1}}{t-1}$ para todo $i$ e $\Gamma = (\gamma_1, \gamma_2, \ldots)$, em que $\gamma_i = \lim_{t\to\infty} \gamma_i(t-1)$ representa a proporção aleatória de clientes na mesa $M_i$ quando o número de clientes tende a infinito. Como os clientes se sentam aleatoriamente nas mesas, $\Gamma$ é uma medida de probabilidade aleatória sobre as infinitas mesas. Construindo $\Gamma$ desta forma dizemos que $\Gamma \sim \text{PRC}(\alpha)$.

Para completarmos então o processo, podemos atribuir a cada mesa um valor $\theta_i \sim H$ independentemente. E então obtemos uma equivalência entre os dois processos.

### 3.2.6 Mistura de Distribuições com PD

Suponha que temos uma amostra $\mathbf{x} = (x_1, \ldots, x_n)$ e assumimos que nossa amostra é proveniente de um elemento de uma família de distribuições de probabilidade $(P_{\boldsymbol{p},\boldsymbol{\theta}})$, em que cada $P_{\boldsymbol{p},\boldsymbol{\theta}}$ é uma distribuição de um modelo de mistura com infinitos componentes como função dos parâmetro $\boldsymbol{p} = (p_1, p_2, \ldots)$ que indica os pesos dos componentes e $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots)$ que indica os parâmetros dos componentes. E ainda que cada componente de índice $k$ da mistura possui distribuição $F(\theta_k)$.

Assuma adicionalmente que a distribuição *a priori* sobre $\boldsymbol{p}$ é a distribuição PRC($\alpha$) e a distribuição *a priori* sobre cada $\theta_k$ é $H$.

Assim, de acordo com a distribuição de um PD$(\alpha, H)$ obtemos uma distribuição $G$ sobre um espaço mensurável $(\Theta, \mathscr{B})$ que faz o papel da distribuição de mistura como em (2.7). A partir de $G$ obtemos uma amostra i.i.d. $\mathbf{x} = (x_1, \ldots, x_n)$.

O modelo de mistura com PD une a teoria sobre o Processo de Dirichlet com a estrutura de um modelo de mistura sendo descrito pelas relações

$$\begin{aligned} G &\sim PD(\alpha, H) \\ \theta_i | G &\sim G, \qquad \text{para } 1 \leq i \leq n \\ x_i | \theta_i &\sim F(\theta_i), \text{ para } 1 \leq i \leq n. \end{aligned}$$

Uma das maneiras de fazer inferência usando um modelo de mistura com PD é usando um *Gibbs Sampling* colapsado. Este algoritmo é usado por Wood e Black (2008) para análise de agrupamentos com dados de disparos neuronais obtidos em registros eletrofisiológicos.

O procedimento possui uma certa semelhança com o EM estocástico começando com uma escolha inicial arbitrária de $\mathbf{s}_0$ para as variáveis não observáveis $\mathbf{S}$ e gerando uma cadeia de Markov em que o espaço de estados são as possíveis configurações das variáveis não observáveis.

Consideramos a distribuição *a priori* de escolha $\boldsymbol{\theta} \sim H$ e sequencialmente, de $i = 1$ até $n$, simulamos valores segundo a distribuição definida por

$$P(s_i = k | \mathbf{s}_{-i}, \mathbf{x}, \alpha, H) \propto p(x_i | \mathbf{x}_{-i}, H) P(s_i = k | \mathbf{s}_{-i}, \alpha). \tag{3.13}$$

Em que $\mathbf{x}_{-i}$ são os dados amostrais sem a observação $x_i$ e $\mathbf{s}_{-i}$ são os dados simulados para $\mathbf{s}$ sem o valor $s_i$. A distribuição de $s_i$ é discreta e finita, pois a cada passo ou alocamos $x_i$ em um grupo já existente $s_i \in \{1, \ldots, K\}$ ou alocamos $x_i$ em um novo grupo $s_i = K + 1$, com $K$ variável durante o processo pois se atualiza com a criação ou extinção de grupos.

Pelo PRC temos que

$$P(s_i = k | \mathbf{s}_{-i}, \alpha) = \begin{cases} \frac{n_k}{n-1+\alpha}, \text{ se o cluster } k \text{ não está vazio} \\ \frac{\alpha}{n-1+\alpha} \text{ para um cluster vazio} \end{cases}$$

e $p(x_i | \mathbf{x}_{-i}, H) = \int_\theta p(x_i | \theta) p(\theta | H, \mathbf{x}_{-i}) d\theta$ é a preditiva *a posteriori* de $x_i$.

Por exemplo, se estamos ajustando um modelo de mistura de normais com distribuição *a priori* conjugada Normal-Inv. Wishart, $H$, sobre $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, a distribuição de $x_i | \mathbf{x}_{-i}, H$ é uma t de Student e a distribuição de $s_i = k | \mathbf{s}_{-i}, \alpha$ é dado pelo Processo de Restaurante Chinês. Adaptando (3.13) para um procedimento MCMC fazemos

1. Faça $t = 0$ e comece com uma configuração $\mathbf{s}^0$ arbitrária e parâmetros $\alpha > 0$ e $H$ distribuição de probabilidade sobre $\Theta$;

2. Calcule $P(s_i^{t+1} = k|\mathbf{s}_{-i}^t, \alpha)$ e $p(x_i|\mathbf{x}_{-i}, H)$;

3. Para $i$ de 1 até $n$, simule novos valores $s_i^{t+1}$ segundo

$$P(s_i^{t+1} = k|\mathbf{s}_{-i}, \mathbf{x}, \alpha, H) \propto p(x_i|\mathbf{x}_{-i}, H)P(s_i^{t+1} = k|\mathbf{s}_{-i}, \alpha);$$

4. Faça $t \leftarrow t+1$ e retorne ao passo 2;

O procedimento descrito gera uma cadeia de Markov homogênea sobre as possíveis configurações $\mathbf{s}$ e pode ser usado para agrupamento dos dados sem imposição de um número específico de grupos.

CAPÍTULO

# 4

# ALGORITMO EM ESTOCÁSTICO COM PERTURBAÇÕES ALEATÓRIAS

Fazemos uma proposta por uma modificação do algoritmo EM estocástico que amplia o espaço de soluções possíveis para além do espaço de soluções com um número fixo de grupos.

Seja $\mathbf{x} = (x_1, \ldots, x_n)$ a amostra e $\mathbf{s} = (s_1, \ldots, s_n)$ o vetor não-observável, de modo que $s_i = k$ indica que a origem da observação $x_i$ é o grupo $k$, como no caso do algoritmo EM estocástico, cada configuração $\mathbf{s}^t$ define uma partição dos dados em tantos grupos quanto há elementos únicos na configuração.

Se o nosso modelo de mistura tem como vetor de probabilidades (pesos) dos componentes $\boldsymbol{p} = (p_1, \ldots, p_K)$ e a amostra $\mathbf{x}$ é i.i.d., então o vetor dos números de elementos provenientes de cada componente $k$ é o vetor aleatório

$$(n_1, \ldots, n_K) \sim \text{Multinomial}(n; p_1, \ldots, p_K),$$

e a estimativa de máxima verossimilhança para $p_k$ é $\frac{n_k}{n}$, $k = 1, \ldots, K$.

Além disso, os elementos $x_i$ em $\mathbf{x}$ tais que $s_i = k$ formam uma subamostra, que contém todas e somente as observações que provém do componente $k$. Usando cada subamostra que contém as observações que se originaram em cada componente, podemos estimar os parâmetros $\theta_k$ para cada $1 \leq k \leq K$.

O objetivo é recuperar (estimar) as informações não observadas a partir dos dados observáveis. Devemos responder às perguntas:

1. Quais pontos amostrais pertencem ao mesmo grupo?
2. Quais são os parâmetros de cada componente da mistura?
3. Quantos grupos existem?

Considere $\mathscr{S}$ o conjunto de todos os possíveis $\mathbf{s}$. No algoritmo EM estocástico fixamos o número de grupos desejados em um valor $K$ e buscamos uma solução ótima, ou próxima da ótima, dentro do subconjunto de $\mathscr{S}$ que satisfaz a condição de que existem exatamente $K$ valores únicos em $\mathbf{s}$. Isto é, a pergunta 3 deve ser respondida antes da aplicação do algoritmo.

O método proposto permite que sejam encontradas soluções diversas daquelas que o algoritmo EM estocástico encontraria partindo-se do mesmo ponto inicial ao buscar soluções com número de grupos diferentes do ponto inicial.

Acrescentamos ao algoritmo EM tradicional a possibilidade de um ponto amostral $x_i$ ser alocado em um novo grupo ao longo das iterações, assim como a possibilidade de um grupo deixar de existir após todos os seus pontos serem alocados em outros grupos. Inserimos uma variável $\xi$ que age como um fator de perturbação ao permitir este comportamento.

## 4.1 Descrição do Algoritmo EM Estocástico com PA

1. Inicie com $t = 0$, o valor inicial do fator de perturbação $\xi = \xi^0 \in (0,1)$, uma estimativa inicial do número de grupos $K^0 = K^* \geq 1$ e um vetor $\mathbf{s}_0$ em $\mathscr{S}_{K^*} = \{\mathbf{s} \in \mathscr{S}; s_i \in \{1,\ldots,K^*\}$ e para todo $k \in \{1,\ldots,K^*\}$ existe $s_i = k\}$;

2. Calcule as estimativas de máxima verossimilhança, $\boldsymbol{\Psi}^t = (\boldsymbol{\theta}^t, \boldsymbol{p}^t)$, dos parâmetros e probabilidades dos componentes usando os valores de $\mathbf{s}_t$ (de modo idêntico como é feito no algoritmo EM estocástico)

$$p_k^t = \frac{\sum_{i=1}^n \mathbb{I}(s_i^t = k)}{n}$$
$$\theta_k^t = \underset{\theta_k \in \Theta}{\operatorname{argmax}} \sum_{i=1}^n \mathbb{I}(s_i^t = k) \log f(x_i|\theta_k),$$

   para todo $k \in \{1,\ldots,K^*\}$;

3. Usando $\boldsymbol{\Psi}^t$, calcule as probabilidades

$$\tau_{ik}^t = P(s_i = k|x_i, \boldsymbol{\Psi}^t) = \frac{p_k^t f(x_i|\theta_k^t)}{\sum_{l=1}^K p_l^t f(x_i|\theta_l^t)},$$

   para todo $i \in \{1,\ldots,n\}$ e $k \in \{1,\ldots,K^*\}$;

4. Para cada ponto $x_i$ em $\mathbf{x}$ gere $u \sim \text{Unif}(0,1)$ de forma independente:

   a) caso $u < \xi^t$, isto é, com probabilidade $\xi^t$, aloque-o em um novo grupo fazendo $s_i^{t+1} = K^* + 1$ e atualize o número de grupos fazendo $K^* \leftarrow K^* + 1$;

   b) caso $u \geq \xi^t$ aloque-o em um grupo já existente, sorteando $s_i^t \sim \text{Multinomial}(1; \tau_{i1}^t, \ldots, \tau_{iK^*}^t)$,

deste modo criamos um novo vetor $\mathbf{s}_{t+1}$;

5. Como é possível que algum grupo tenha se esvaziado no passo anterior, fazemos $K^* \leftarrow$ # elementos únicos em $\mathbf{s}^{t+1}$;

6. Também é possível que um grupo de rótulo $k < K^*$ tenha se esvaziado, por exemplo se haviam 5 grupos e o grupo 2 se esvaziou, ficamos com os grupos 1,3,4 e 5, mas queremos renomeá-los para que não existam descontinuidades e se tornem os grupos 1,2,3 e 4 respectivamente. Assim, repetimos uma subrotina computacional que faz isso;

7. Uma mudança no tamanho de $K^*$ implica numa mudança na dimensão do espaço paramétrico de $\mathbf{\Psi}$, então de acordo com a mudança feita devemos mudar as dimensões de $\boldsymbol{\theta}^{t+1}$, $\boldsymbol{p}^{t+1}$ para que se adequem ao novo número de componentes;

8. Faça $\xi \leftarrow \xi_{t+1}$ e $t \leftarrow t+1$ e retorne ao passo 2.

Ao longo das iterações fazemos $\xi_t$ decrescer de modo que $\lim_{t\to\infty} \xi_t = 0$, assim o comportamento do algoritmo assintoticamente é igual ao do algoritmo EM estocástico mas ao longo do algoritmo pode haver mudanças no número de grupos existentes que são conduzidas pela estrutura dos dados.

Sabemos que, partindo de uma configuração de qualidade baixa, a tendência do algoritmo EM estocástico é encontrar configurações melhores, que fazem mais sentido com os dados no sentido que a nova configuração induza estimativas $\mathbf{\Psi}^t$ sob as quais a log-verossimilhança $\log L(\mathbf{x}|\mathbf{\Psi}^t)$ é maior do que era inicialmente. Mas apesar desta qualidade o algoritmo EM estocástico não alcança soluções com um número maior de grupos do que existiam em $\mathbf{s}^0$, ao permitirmos que um ponto amostral $x_i$ seja alocado em um novo grupo, se estivermos piorando a configuração pela natureza do algoritmo EM estocástico a tendência deste novo grupo é sumir, no entanto se estivermos melhorando a tendência deste novo grupo é permanecer, mas não poderíamos ter explorado este espaço sem uma perturbação aleatória.

Neste sentido, o método que aqui propomos funciona de maneira análoga a mutações aleatórias em genes de indivíduos, se estas introduzirem uma nova configuração benéfica em algum sentido ao genoma do indivíduo elas tendem a permanecer devido a um mecanismo estocástico que favorece a perpetuação destas, caso contrário tendem a desaparecer. Como essas mutações são aleatórias, esperamos que muitas sejam deletérias, e ao mesmo tempo em que estas mutações são responsáveis pela aparecimento de novas possibilidades, se a taxa de ocorrência de mutações for muito alta e constante, os indivíduos não conseguem se adaptar em tempo a elas, assim impomos que $\xi^t$, que em nossa analogia seria a taxa da ocorrência de mutações, decresça com as iterações até ser negligenciável.

A esta sequência $\xi^0, \xi^1, \xi^2, \ldots \to 0$ chamaremos de rotina de decaimento da variável $\xi$, recomendamos que tal rotina seja escolhida de modo que $\xi$ comece com um valor inicial

significante e convirja para 0 de modo lento, e que sejam realizadas um número suficientemente grande de iterações de modo que na última iteração do algoritmo, $t_f$, seja satisfeito $\xi^{t_f} \approx 0$.

Isto porque, se $\xi^t \approx 0$ para toda iteração $t$ a capacidade de criar novos grupos do algoritmo fica debilitada e o algoritmo se comporta já desde o princípio de modo próximo ao algoritmo EM estocástico. Da mesma maneira, se $\xi^t \leftarrow 0$ muito rapidamente existirão poucas chances do algoritmo encontrar uma partição com novos grupos onde ele possa se estabilizar.

Podemos escrever $\xi^t = \xi^0 g(t)$ para definir a rotina de decaimento de $\xi$ e chamar $f(t)$ de função de decaimento.

Recomendamos usar decaimentos exponenciais, que têm a forma $\xi^t = \xi^0 \exp(\log(k)t) = \xi^0 \exp(\log(k))^t = \xi^0 k^t$, com $0 < k < 1$, ou equivalentemente, $\log(k) < 0$. Por exemplo $\xi^t = \xi^0 \cdot 0,98^t$ ou $\xi^t = \xi^0 \exp(-0,1t)$.

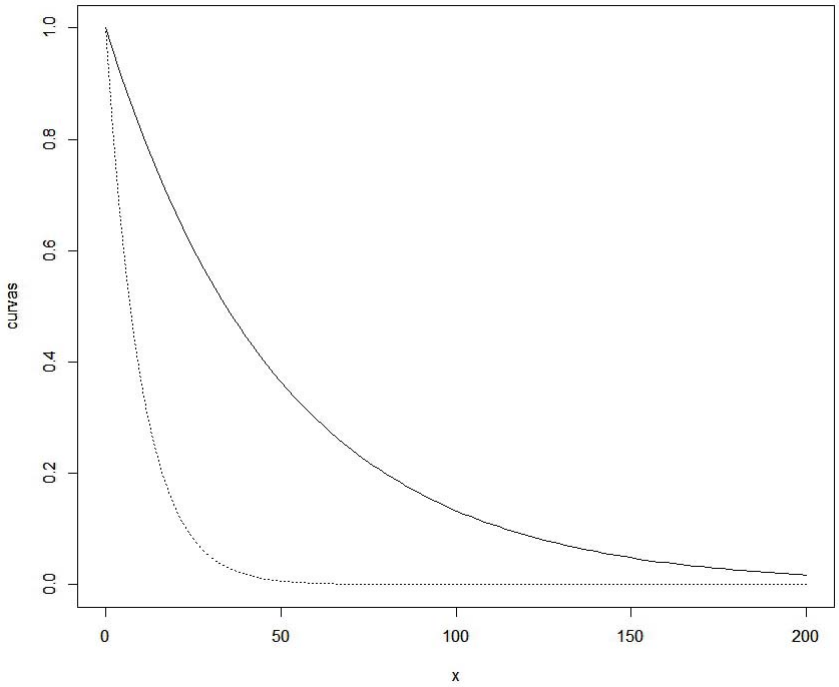


Figura 16 – Na linha contínua: $g(x) = 0,98^x$, na linha pontilhada $g(x) = \exp(-0,1\,x)$

Entre estas duas funções, por exemplo, a escolha que permite maior variabilidade no ajuste do modelo, partindo do mesmo $\mathbf{s}^0$, é $g(t) = 0,98^t$, que converge mais lentamente para 0. Se quiséssemos então permitir esta variabilidade maior e garantir que na última iteração temos $\xi^t < \xi^0\ 10^{-2}$ deveríamos então fazer pelo menos 228 iterações, caso jugássemos que a função de decaimento $g(t) = \exp(-0,1t)$ nos é suficiente para o nosso problema, então sob os mesmos critérios é suficiente realizar pelo menos 47 iterações.

Quanto mais rápida é a convergência, mais rápido o algoritmo se degenera e passa a

se comportar como o algoritmo EM estocástico sem as perturbações aleatórias e também mais rápido ele atinge uma estabilidade nas configurações $\mathbf{s}^t$, de modo que convém que este dois fatores sejam pesados ao se escolher uma rotina de decaimento.

Quando alocamos um ponto $x_i$ à um novo grupo, este grupo tenderá a permanecer caso existam pontos muito mais próximos de $x_i$ do que da média dos outros grupos, e tenderá a desaparecer caso esteja próximo da média de outro grupo. Podemos parar as iterações quando o algoritmo obtiver um comportamento estável, com pequenas diferenças nos resultados obtidos entre uma iteração e a seguinte, ou após um grande número de iterações.

A alternativa ao método proposto seria ajustar diferentes estimativas de densidade, cada um com um número $K$ de grupos diferentes e depois selecionar o melhor modelo segundo um método de critério de informação. O método proposto se torna útil em muitos problemas pois

1. Nem sempre é evidente a partir de uma análise exploratória dos dados quais valores possíveis de $K$ são razoáveis. Isto é verdade principalmente se os dados têm dimensão alta e/ou existem muitos grupos;

2. O algoritmo proposto tem potencial de buscar soluções com número de grupos diferentes em uma única execução.

**Exemplo 1:** Simulamos amostras de uma mistura de duas normais bivariadas com médias $\mu_1 = (0,0)$ e $\mu_2 = (5,5)$ e matrizes de variância $\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ e $\Sigma_2 = \begin{pmatrix} 1 & 0,5 \\ 0,5 & 1 \end{pmatrix}$ respectivamente, com 100 observações provenientes de cada normal, e atribuímos inicialmente todos os pontos à um mesmo grupo. Usamos $\xi_0 = 1/n = 0,01$ e $\xi_{t+1} = \xi_t \cdot 0,9$.

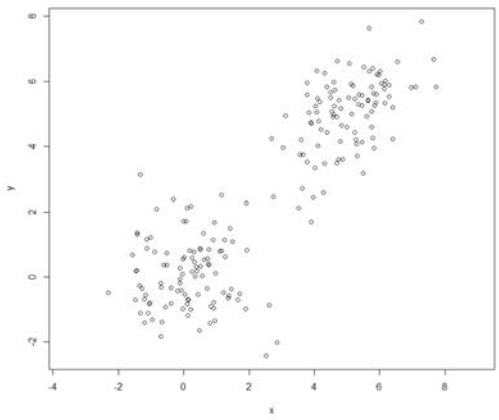


Figura 17 – Gráfico de dispersão dos pontos simulados. A amostra pseudo-completa inicial possui apenas um grupo.

Neste exemplo, o procedimento se torna:

1. Definimos $t = 0$, escolhemos $\mathbf{s}^0$ e $K^*$;

2. Calculamos as estimativas de máxima verossimilhança $\boldsymbol{\Psi}^t = (\mathbf{p}^t, \boldsymbol{\theta}^t)$ como descrito no passo 2 do algoritmo;

3. Usando $\boldsymbol{\Psi}^t$ calculamos as probabilidades $\tau_{ik}^t$ como no passo 3;

4. Para cada ponto $x_i$, gere $u \sim \text{unif}(0,1)$ independentemente, se $u < \xi_t$ alocamos $x_i$ em um novo grupo $K^* + 1$. Caso contrário simulamos um novo valor entre 1 e $K^*$ para $s_i^{t+1}$ segundo as estimativas $\tau_{ik}^t$;

5. Atualizamos o valor de $K^*$;

6. Fazemos $\xi_{t+1} \leftarrow 0.9 \cdot \xi_t$;

7. Fazemos $t \leftarrow t + 1$ e retornamos ao passo 2.

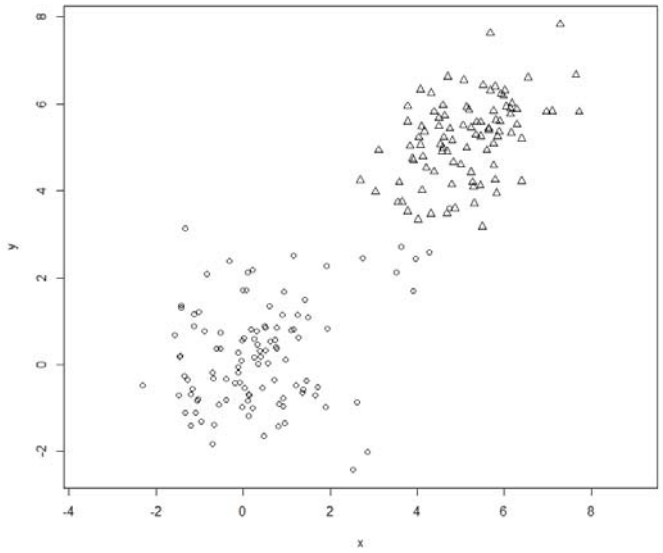


Figura 18 – Iteração 20.

Após algumas iterações já existe o número real de dois grupos distintos e a separação destes de modo aproximado corresponde aos grupos reais, sendo que após 30 iterações o ajuste é quase perfeito.

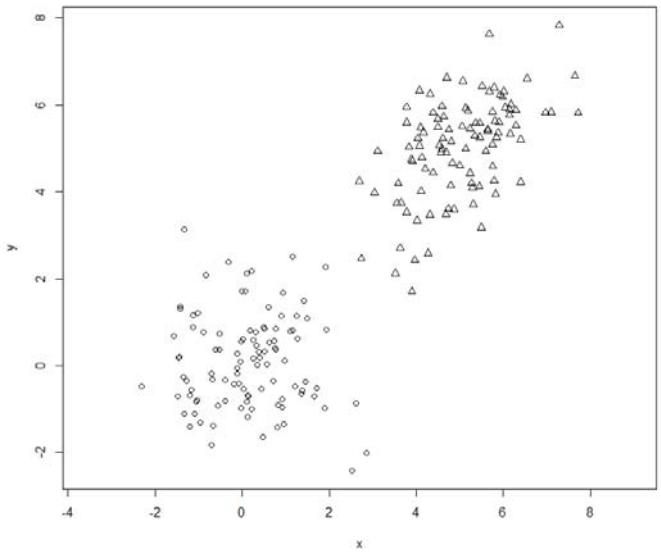


Figura 19 – Iteração 30.

**Exemplo 2:** Gerando uma nova amostra do mesmo modo como foi feito no Exemplo 1 e utilizando a mesma rotina para os valores de $\xi_t$, começamos agora com uma amostra pseudo-completa inicial de 4 grupos.

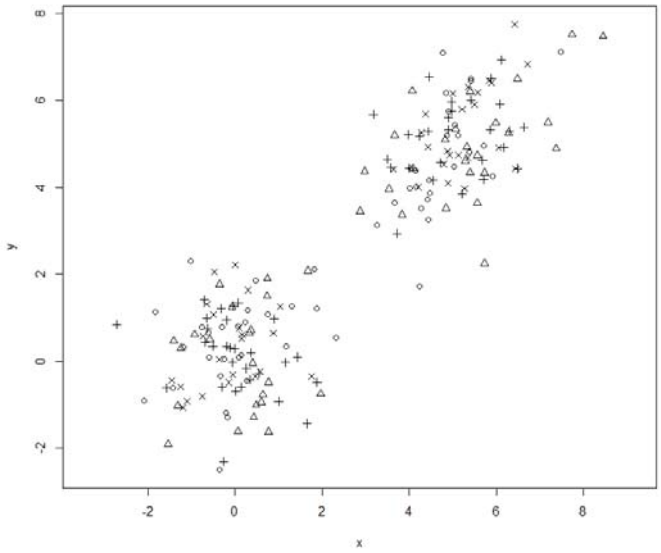


Figura 20 – Gráfico de dispersão dos pontos com pseudo-amostra inicial com 4 grupos.

Neste exemplo, visualmente é fácil distinguir os dois agrupamentos e separar quais pontos da amostra devem estar no mesmo grupo. Aqui partimos de uma amostra pseudo-completa arbitrária com 4 grupos, identificados pelos símbolos de ‘círculo’, ‘triângulo’, ‘cruz’ e ‘xis’,

esperamos que o algoritmo seja capaz de após algumas iterações reduzir o número de grupos e separar quase todos os pontos corretamente.

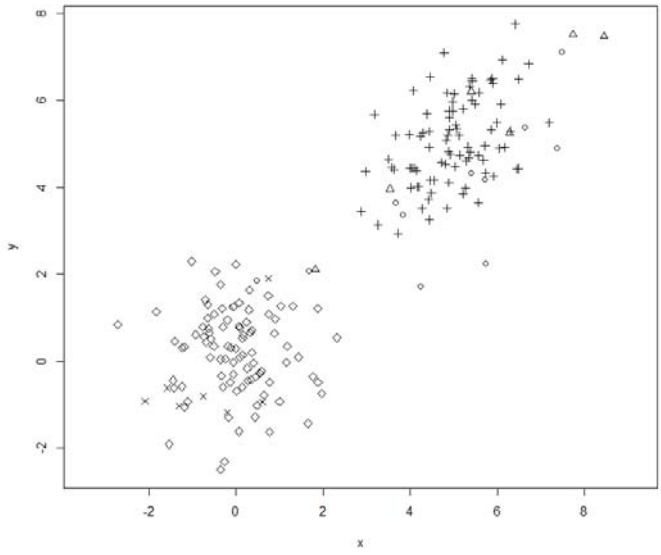


Figura 21 – Iteração 20

Após 20 iterações já existe uma prevalência forte de dois grupos, sendo que cada um destes já engloba a maior parte dos grupos verdadeiros. Com 40 iterações a separação obtida foi perfeita.

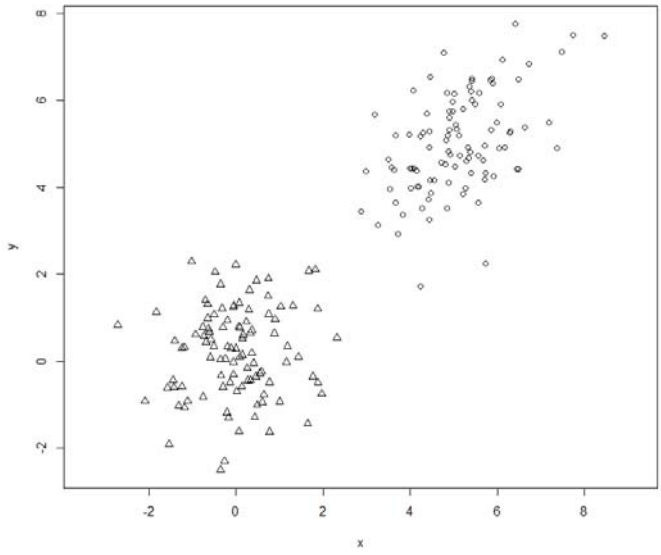


Figura 22 – Iteração 40

Observe que o algoritmo proposto possui a capacidade de aumentar ou diminuir o número de grupos durante a sua execução, permitindo a vantagem de se exigir menos do analista ao

não se comprometer com um número pré-fixado de grupos para o ajuste do modelo, fornecendo assim um meio de modelagem mais flexível para análise de agrupamentos.

Para a execução deste algoritmo, para a análise de modelos de mistura de qualquer distribuição paramétrica, somente é necessário que seja possível calcular as estimativas de máxima verossimilhança dos parâmetros. Não é necessário a especificação de uma distribuição *a priori* dos parâmetros $\mathbf{\Psi}$ e nem o conhecimento da distribuição *a posteriori* e da *a posteriori* preditiva.

## 4.2 Algoritmo EM Estocástico com Perturbações Aleatórias - Versão com *Gibbs Sampling*

Também é possível implementar o algoritmo usando as probabilidades condicionais, modificando cada $x_i$ de grupo e então atualizando as estimativas de $\mathbf{\Psi}$ segundo cada uma destas mudanças. Pelo teorema de Bayes

$$P(S_i = k|x_i, \mathbf{\Psi}) = P_{\mathbf{\Psi}}(S_i = k|x_i) \propto P_{\mathbf{\Psi}}(S_i = k) f_{\mathbf{\Psi}}(x_i|S_i = k) = P(S_i = k|\mathbf{\Psi})\ f(x_i|S_i = k, \mathbf{\Psi})$$
$$= p_k\ f(x_i|\theta_k), \forall k,$$

assim se pudermos estimar $p_k$ e $f(x_i|\theta_k)$ podemos estimar a distribuição condicional da esquerda, as estimativas das proporções da mistura podem ser obtidas da maneira usual e o segundo termo pode ser estimado inserindo a estimativa $\theta_k^t$ no lugar de $\theta_k$. Nesta versão modificamos cada $s_i$ e em seguida atualizamos as estimativas de $\mathbf{\Psi} = (\boldsymbol{\theta}, \boldsymbol{p})$, isto torna as iterações computacionalmente mais devagar, no entanto como as estimativas são atualizadas com mais frequência dentro da mesma iteração a convergência costuma ocorrer com menos iterações.

Esta versão que utiliza as probabilidades condicionais tem forma similar a do algoritmo de modelos de mistura com o Processo de Dirichlet descrita na seção (3.2.6), e procede do seguinte modo

1. Fazemos $t \leftarrow 0$, selecionamos uma amostra pseudo-completa $(\mathbf{x}, \mathbf{s}_0)$ inicial arbitrária e um número de grupos inciais $K^*$;
2. Calculamos $\mathbf{\Psi}^t$ usando $(\mathbf{x}, \mathbf{s}^t)$;
3. Calculamos $\tau_{ik}^t$, para $i$ de 1 até $n$ e $k$ de 1 até $K^*$, usando $\mathbf{\Psi}^t$;
4. A cada vez que mudamos um único ponto $x_i$ de grupo, vamos da estimativa $\mathbf{\Psi}^t(i)$ anterior a esta mudança para a nova estimativa $\mathbf{\Psi}^{t+1}(i)$ dentro da mesma iteração. Assim para $i$ de 1 até $n$ faça:

a) Simulamos $u \sim \text{uniforme}(0,1)$,

   i. se $u < \xi_t$ alocamos $x_i$ em um novo grupo $K^* + 1$ e faça $K^* \leftarrow K^* + 1$;

   ii. caso contrário calculamos a estimativa de máxima verossimilhança $\boldsymbol{\Psi}^t(i)$ e as estimativas de probabilidade $\boldsymbol{\tau}_i^t$ usando a amostra sem o ponto $x_i$ e alocamos $x_i$ em um dos $K^*$ grupos já existentes segundo uma multinomial$(1; \boldsymbol{\tau}_i^t)$;

b) Calculamos novas estimativas $\boldsymbol{\Psi}^{t+1}(i)$ usando o valor atual de $s_i^{t+1}$ no lugar de $s_i^t$;

c) Fazemos $K^* \leftarrow$ # elementos únicos em $\mathbf{s}^{t+1}$ e ajustamos os rótulos dos grupos para evitar descontinuidades como é feito no algoritmo EM estocástico com perturbações aleatórias;

5. Faça $t \leftarrow t+1$, $\xi^t \leftarrow \xi^{t+1}$ e retornamos ao passo 2.

CAPÍTULO

# 5

# SIMULAÇÃO E APLICAÇÕES

Para demonstrar o uso dos algoritmos na aplicação de modelos de mistura, fizemos um estudo de simulação com dados simulados de uma mistura de normais univariadas e Poissons.

Por meio dos dados e partindo de uma atribuição aleatória dos pontos amostrais à grupos obtivemos estimativas dos parâmetros de cada distribuição. Como um primeiro exemplo demonstrativo, aplicamos o algoritmo com um número fixo de grupos, que se reduz ao algoritmo EM Estocástico.

## 5.1 Simulação de mistura de duas distribuições com K=2 fixo

Fizemos simulações de uma mistura de duas Poissons com $p_1 = p_2 = 0,5$, $\lambda_1 = 5$ e $\lambda_2 = 10$ e com $n = 200$. Geramos 1000 amostras desta mistura e utilizamos para cada amostra o algoritmo EM estocástico com 300 iterações.

Para evitar problemas com a identificabilidade das distribuições componentes da mistura, atribuímos a estimativa de menor valor de $\lambda$ à distribuição de parâmetro $\lambda_1$.

Com $n = 200$, temos que os histogramas para $\lambda_1$ e $\lambda_2$ são dados nas Figuras 23 e 24.

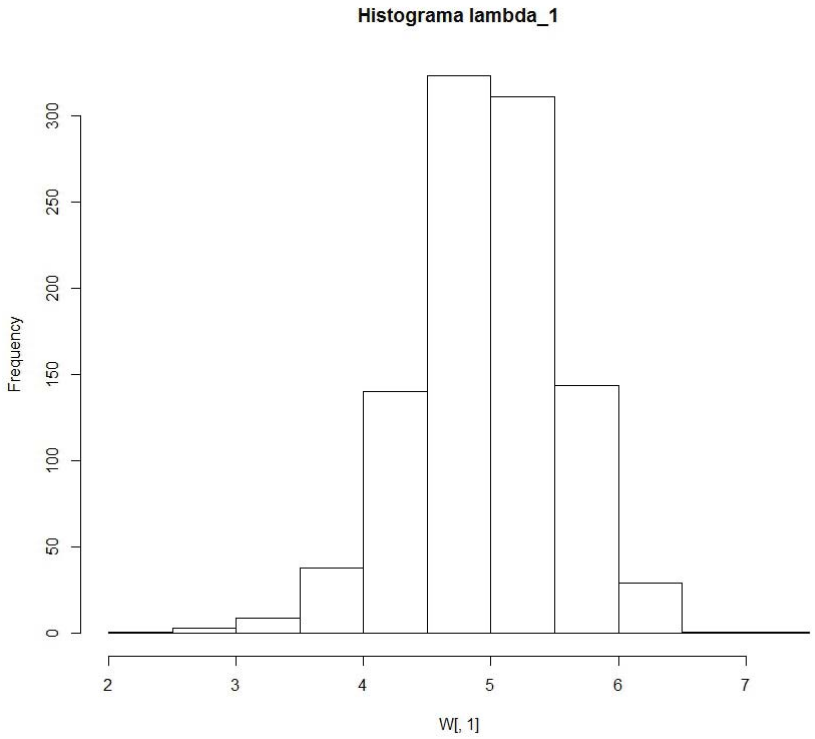


Figura 23 – Histograma dos valores de $\lambda_1$ com tamanho da amostra $n = 200$.

Os quantis referentes a $\lambda_1$ são dados por 0% = 2,44; 25% = 4,61; 50% = 4,99; 75% = 5,36 e 100% = 7,13. E para $\lambda_2$ temos

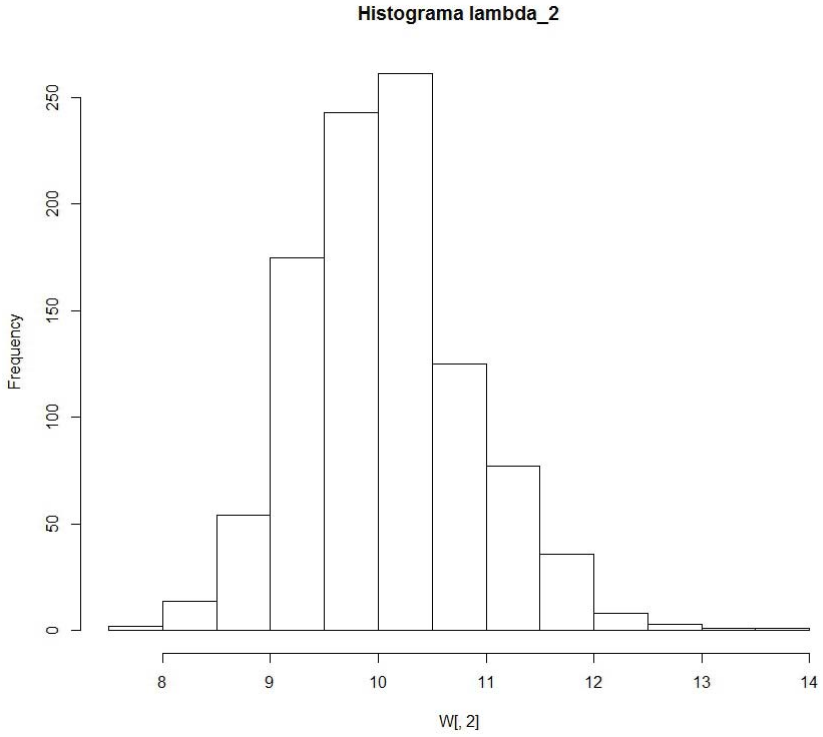


Figura 24 – Histograma dos valores de $\lambda_2$ com tamanho da amostra $n = 200$.

Os quantis referentes a $\lambda_2$ são dados por 0% = 7,99; 25% = 9,53; 50% = 10,02; 75% = 10,51 e 100% = 13,86.

A média das estimativas de $p_1$ foi 0,45 e de $p_2$ foi 0,55 e a média dos pontos classificados corretamente foi 74,03%.

A mediana das estimativas dos parâmetros ofereceu portanto uma estimativa muito próxima do valor real, a média das estimativas foi adequada e a classificação dos pontos foi boa mas de qualidade inferior.

## 5.2 Simulação de mistura de duas distribuições com K variável

Agora, fazemos um estudo de simulação em que o número de grupos $K$ não seja uma variável fixa.

Fizemos 500 simulações do seguinte modo: Geramos 200 valores (a cada simulação) de uma mistura de três variáveis aleatórias de Poisson com $p_1 = p_2 = p_3 = 1/3$, $\lambda_1 = 5$, $\lambda_2 = 15$, $\lambda_3 = 25$, e aplicamos o algoritmo EM estocástico com perturbações aleatórias, iniciando com $\xi^0 = 0,5$ e atribuindo aleatoriamente cada ponto a um de dez grupos. A rotina de decaimento de $\xi$ segue $\xi^t = \xi^0 \exp(-0,1 \cdot t)$, ou equivalentemente, $\xi^t = \xi^{t-1} \exp(-0,1)$ para todo $t \geq 1$.

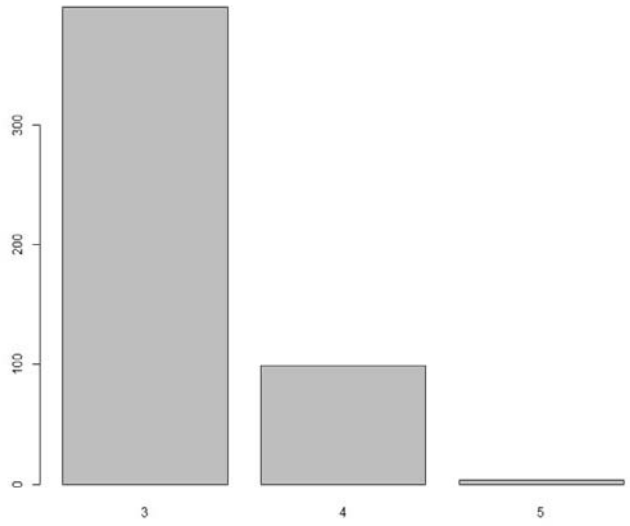


Figura 25 – Gráfico de barras do número de grupos estimados.

O algoritmo acertou o número de três componentes 398 de 500 vezes (79,6 %).

Em mais um exemplo de aplicação do algoritmo EM estocástico com perturbações aleatórias, considere uma amostra, proveniente de uma mistura de duas normais bivariadas,

de tamanho $n = 100$, com $\mu_1 = (3/2, 3/2)$, $\mu_2 = (3,4)$, $\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ e $\Sigma_2 = \begin{pmatrix} 0,4 & 0,3 \\ 0,3 & 0,3 \end{pmatrix}$, $p_1 = 0,34$ e $p_2 = 0,66$, com $\xi^0 = 4/100$ (em média criamos quatro novos grupos na primeira iteração) e $\xi^{t+1} = 0,7 \cdot \xi^t$ e executamos 400 iterações.

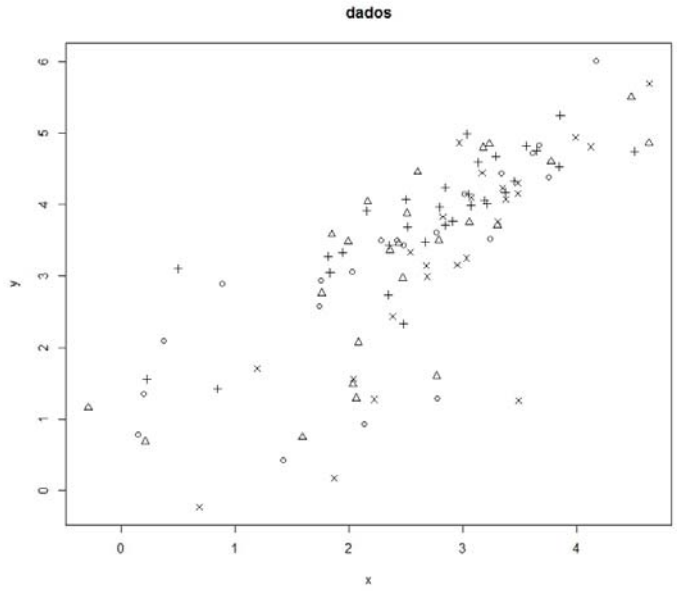


Figura 26 – Agrupamento inicial com quatro grupos.

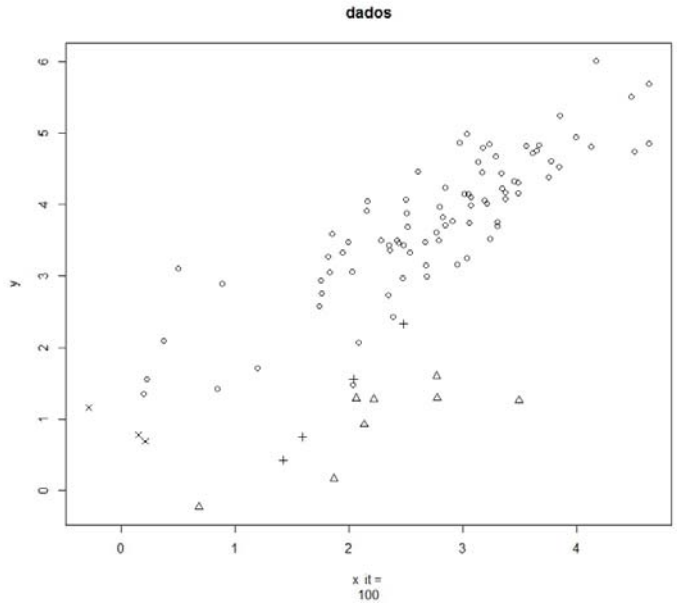


Figura 27 – Iteração #100.

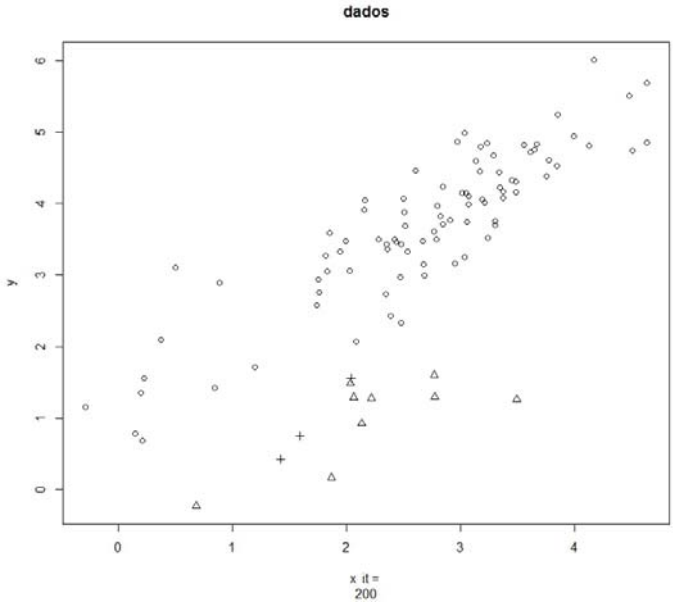


Figura 28 – Iteração #200.

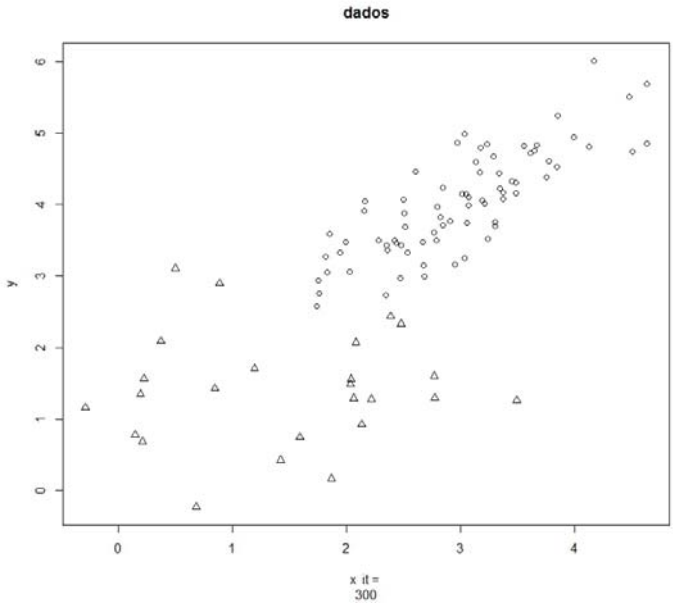


Figura 29 – Iteração #300.

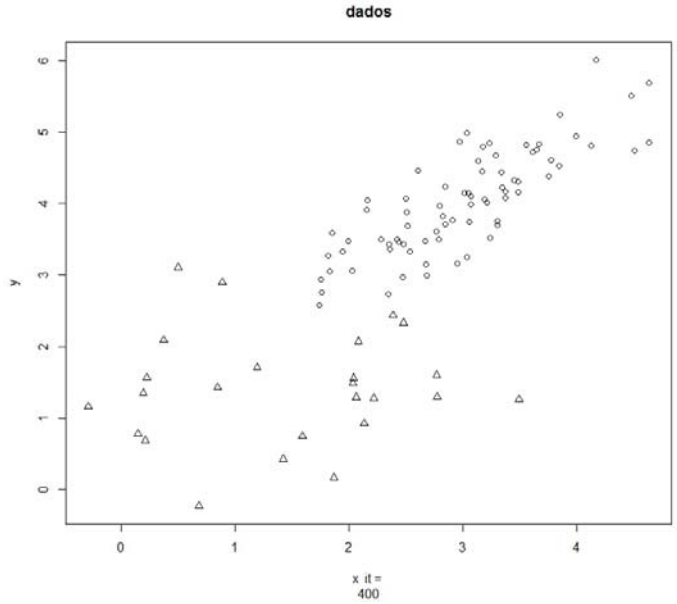


Figura 30 – Iteração #400.

As estimativas foram $\hat{\mu}_1 = (1,51; 1,56)$, $\hat{\mu}_2 = (3,04; 4,07)$, $\hat{\Sigma}_1 = \begin{bmatrix} 0,93 & 0,11 \\ 0,11 & 0,71 \end{bmatrix}$, $\hat{\Sigma}_2 = \begin{bmatrix} 0,47 & 0,37 \\ 0,37 & 0,47 \end{bmatrix}$, $\hat{p}_1 = 0,28$ e $\hat{p}_2 = 0,72$.

Usando a mesma amostra e a mesma rotina de decaimento para $\xi$, executamos o algoritmo partindo de um único grupo. Como em nossos experimentos observamos que o algoritmo em média leva mais iterações para ajustar o modelo quando começamos com uma estimativa inicial do número de grupos inferior ao real, rodamos desta vez 800 iterações.

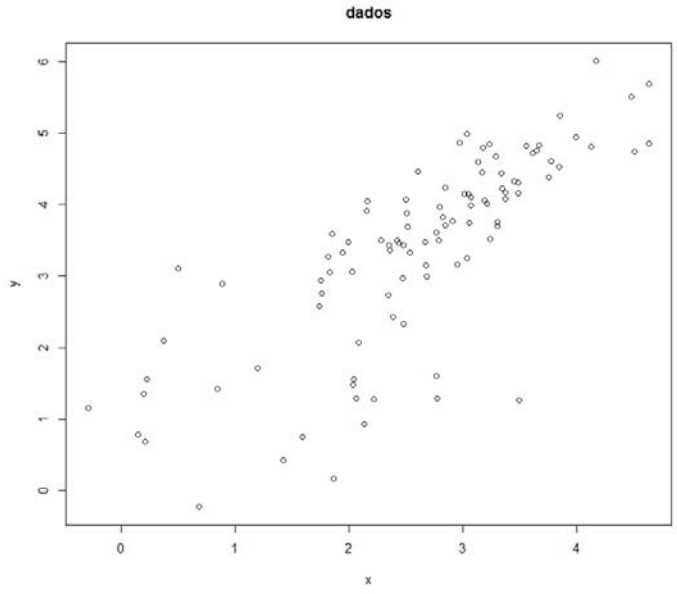


Figura 31 – Agrupamento inicial com um único grupo.

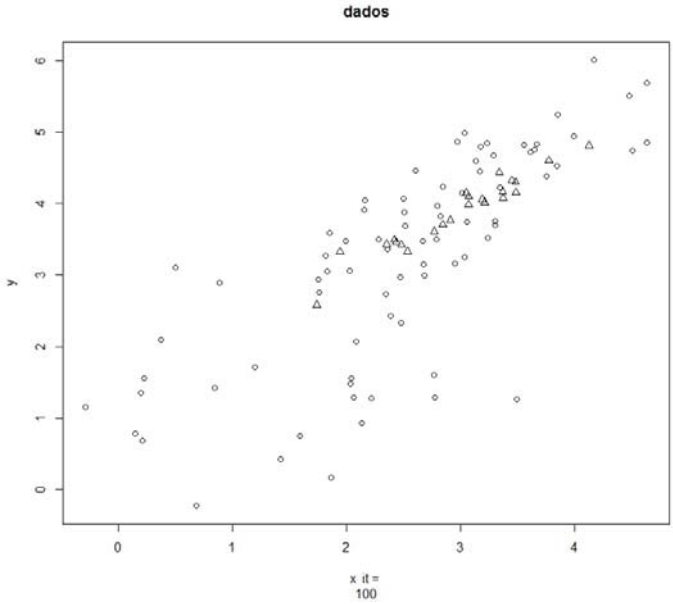


Figura 32 – Iteração #100.

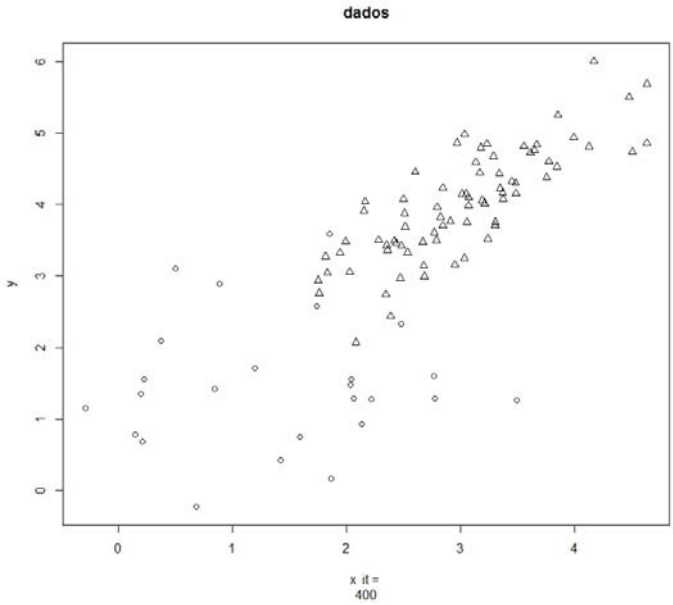


Figura 33 – Iteração #400.

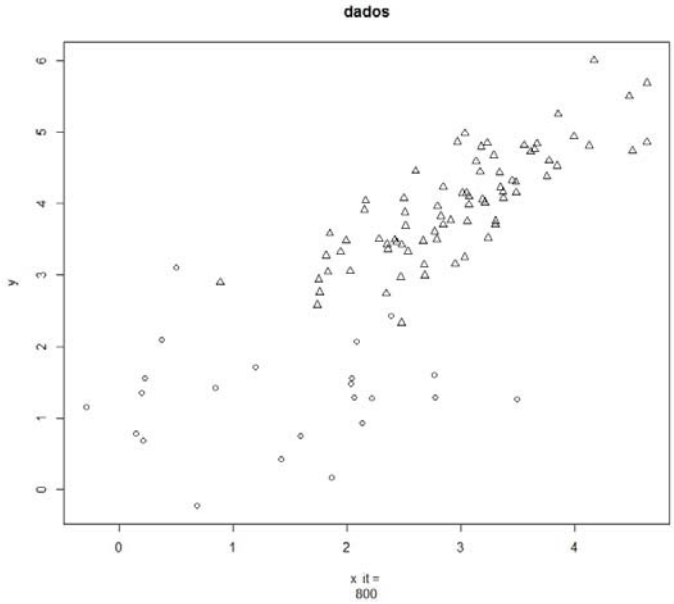


Figura 34 – Iteração #800.

As estimativas obtidas foram $\hat{\mu}_1 = (1,47; 1,40)$, $\hat{\mu}_2 = (2,97; 3,98)$, $\hat{\Sigma}_1 = \begin{bmatrix} 0,96 & 0,06 \\ 0,06 & 0,63 \end{bmatrix}$, $\hat{\Sigma}_2 = \begin{bmatrix} 0,56 & 0,46 \\ 0,46 & 0,57 \end{bmatrix}$, $\hat{p}_1 = 0,24$ e $\hat{p}_2 = 0,76$.

### 5.2.1 Gêiser Old Faithful

Aplicamos agora o algoritmo a um conjunto de dados de erupções do gêiser *Old Faithful* localizado no Parque Yellowstone em Wyoming, EUA. Na Figura 35, há o histograma da variável tempo entre erupções, em minutos.

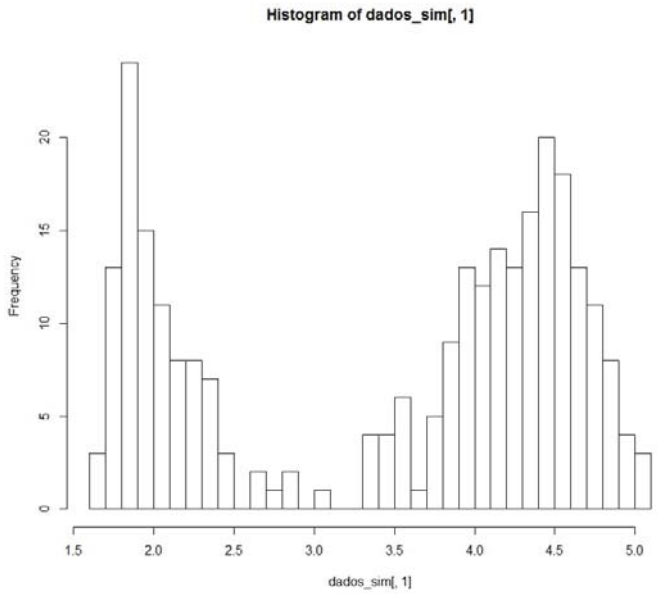


Figura 35 – Histograma do tempo entre erupções (em minutos).

No eixo x da Figura 36, temos o tempo de duração das erupções em minutos e no eixo *y* temos o tempo até a próxima erupção. Começamos o algoritmo com quatro grupos.

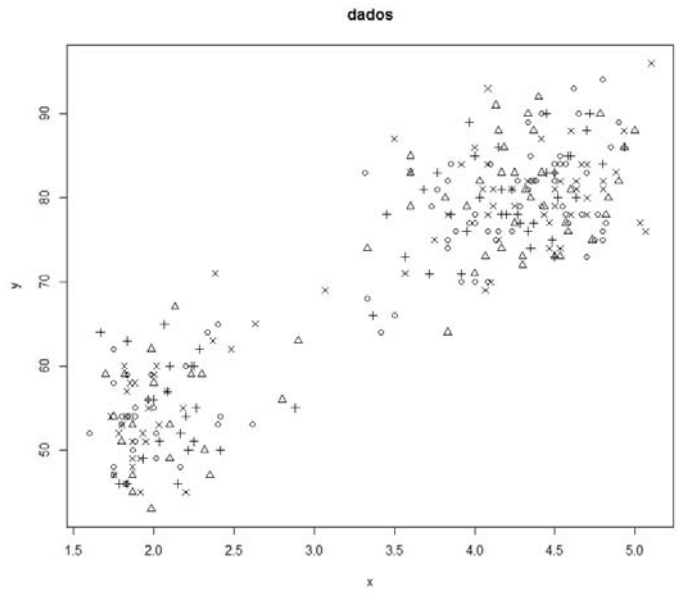


Figura 36 – Agrupamento inicial.

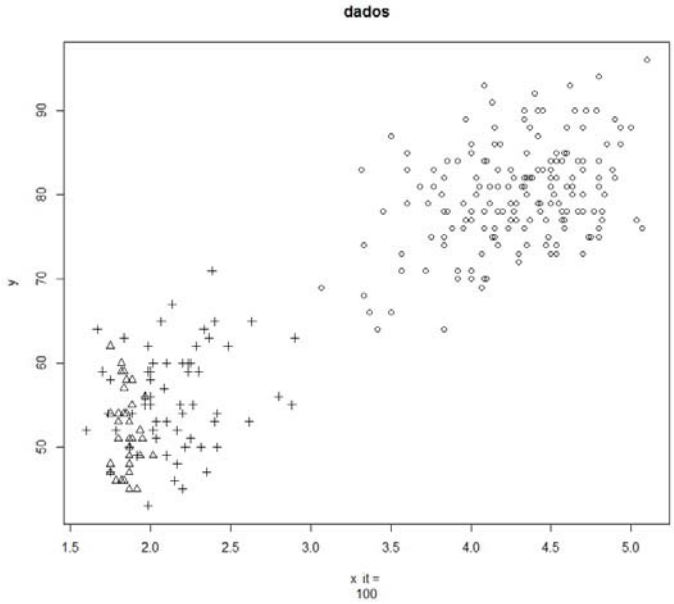


Figura 37 – Iteração #100.

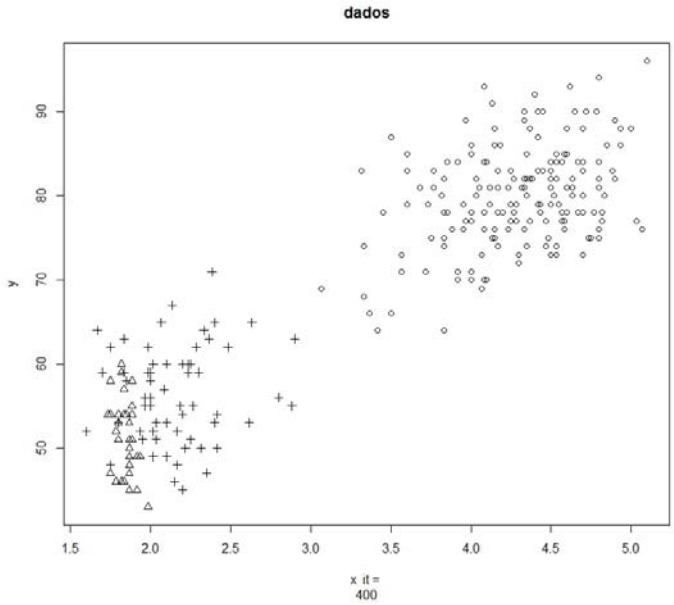


Figura 38 – Iteração #400.

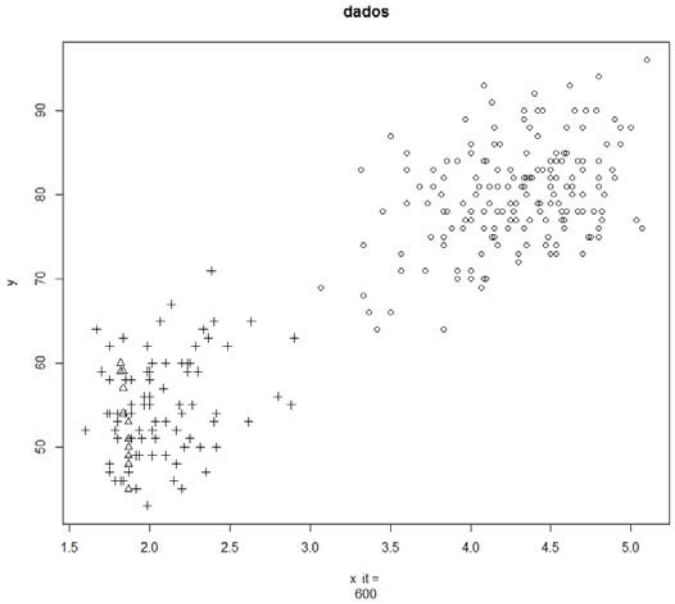


Figura 39 – Iteração #600.

Observamos que o algoritmo permanece com ajustes de três componentes. Isso se deve ao fato de que, apesar de visualmente identificarmos somente dois grupos, existe uma concentração anormal de erupções cujas durações são próximas de 1,9 minutos, o que influencia no ajuste do modelo. Este fato faz com que o algoritmo ajuste uma distribuição separada para estes pontos alta densidade nessa região de concentração, mas é difícil visualizar isto pelo diagrama de dispersão dos pontos porque essa região de concentração se encontra no meio de outra região onde a densidade dos pontos é menor.

## 5.3 Aplicação em Imagens

Uma imagem digital é constituída de um conjunto finito de pixels, cada um com sua localização no plano.

Se a imagem tem $n = m_1 \times m_2$ pixels podemos considerar uma grade $\mathscr{O} := \{(a,b) \mid 1 \leq a \leq m_1, 1 \leq b \leq m_2\} \subset \mathbb{Z}^2$ que contém os sítios dos pixels e uma função $f : \mathscr{O} \to I$ que mapeia cada pixel para a sua intensidade.

Existem três tipos comuns de imagens

1. Imagem binária, possui somente pixels em preto e em branco, neste caso $I = \{0,1\}$;
2. Imagem em tons de cinza, as intensidades variam gradualmente entre o preto e branco ao longo da escala de cinza. Neste caso podemos tomar $I = [0,1]$.
3. Imagem colorida, a intensidade de cada pixel pode ser visto como uma mistura de tons de vermelho, verde e azul. Neste caso usamos $I = [0,1]^3$.

A idéia é que, usando as intensidades dos pixels, nós agrupemos pixels de intensidade semelhantes segundo um modelo de mistura de distribuições normais.

Adicionalmente a levar em conta a intensidade dos pixels, seria preferível que caso um pixel se encaixe igualmente bem em dois grupos haja um critério que tenda a alocar pixels que estão espacialmente próximos no mesmo grupo.

Silva (2007) utilizou um modelo de mistura de normais com PD para segmentar imagens de ressonância magnética de modo que se distingam três componentes, correspondendo à massa branca, massa cinzenta e fluido cerebrospinal; como este tipo de imagem geralmente é exibida em tons de cinza foi utilizado uma mistura de normais univariadas. Aqui, como um exemplo, segmentamos uma imagem digital colorida e por isto usamos uma mistura de normais trivariadas.

### 5.3.1 Redução de Amostra e Aplicação

Em uma imagem digital de digamos $1000 \times 1000$ pixels, teremos $10^6$ pixels no total, se esta imagem for colorida então para cada pixel teremos 3 valores totalizando $3 \times 10^6$ valores. Muitas vezes, o tamanho do conjunto de dados pode tornar a execução do algoritmo no computador lenta, isso pode ser contornado se trabalharmos com um subconjunto da amostra desde que este subconjunto ainda contenha uma quantidade de informação bem significativa da população para propósitos de inferência.

Suponha que a nossa amostra é como descrita acima, se selecionarmos uma amostra de tamanho 1000 escolhendo dados amostrais aleatoriamente e sem reposição ainda teremos informação suficiente para obter estimativas dos parâmetros próximas aos seus valores reais e isto tornaria a execução do algoritmo muito mais rápida.

Outra possibilidade seria buscar uma subamostra da imagem com pixels espalhados pela imagem de maneira a representar todas as regiões da imagem com ênfase parecida. Isto pode ser obtido considerando o conjunto $\mathbf{x}$ dos pixels ordenado segundo suas posições, concatenando as linhas ou colunas de pixels segundo a sua ordem em um único vetor e escolhendo 1 a cada $M$ pixels.

Por exemplo, eis uma imagem digital colorida.

Figura 40 – Imagem colorida original.

Após executarmos o algoritmo 'EM Estocástico com Perturbações Aleatórias - Versão com *Gibbs Sampling*', que desenvolvemos para agrupamento dos pixels, formamos segmentos onde os pixels agrupados são similares entre si. Para propósito de distinção visual, cada segmento foi colorido inteiramente com a cor média de seus pixels constituintes.

Figura 41 – Imagem digital segmentada.

Podemos notar que as fronteiras que foram definidas pelos segmentos correspondem de modo aproximado às fronteiras reais entre objetos distintos da imagem ou entre partes com colorações diferentes do mesmo objeto. A segmentação também separou os objetos principais, o cachorro, a areia da praia e o mar.

CAPÍTULO

6

# CONSIDERAÇÕES FINAIS E CONCLUSÃO

Nesta monografia apresentamos a revisão bibliográfica relativa ao tema modelos de mistura assim como o obtenção do estimador de máxima verossimilhança para os parâmetros. Descrevemos os estimadores baseados nos algoritmos EM, tanto na versão básica quanto na sua versão estocástica.

Apresentamos também o estimador bayesiano baseado no Processo de Dirichlet implementado através do procedimento denominado “Processo do Restaurante Chinês".

No capítulo 5 desenvolvemos um algoritmo próprio e aplicamos a teoria em uma situação prática que é a segmentação de imagens digitais.

Implementamos uma versão modificada do algoritmo EM estocástico. Nessa versão modificada pretendemos que o mesmo incorpore qualidades do PD que permite a variação do número de grupos no procedimento de ajuste de modelo.

Realizamos um trabalho de simulação mais extensivo que testará a performance do algoritmo proposto quanto à precisão do estimador e sua qualidade na seleção de modelos.

# REFERÊNCIAS

BLACKWELL, D.; MACQUEEN, J. Ferguson Distributions Via Polya Urn Schemes. **The Annals of Statistics**, v.1, n. 2, p. 353–355, 1973. Citado nas páginas 52 e 53.

CELEUX, G.; DIEBOLT, J. The SEM Algorithm: A Probabilistic Teacher Algorithm derived from the EM Algorithm for the Mixture Problem. **Computational Statistics Quaterly**, v. 2, p. 73–82, 1985. Citado nas páginas 20, 39 e 45.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum Likelihood from Incomplete Data via the EM Algorithm. **Medical Image Analysis**, v. 39, n. 1, p. 1–38, 1977. Citado nas páginas 19 e 29.

FERGUSON, T. A Bayesian Analysis of some Non-Parametric Problems. **The Annals of Statistics**, v. 1, n. 2, p. 209–230, 1973. Citado nas páginas 20, 49 e 51.

LIMA, E. **Espaços Métricos**. [S.l.]: IMPA, 2011. Citado na página 31.

MACDONALD, P. Karl Pearson's Crab Data. Notas de Aula. 2017. Disponível em: <http://ms.mcmaster.ca/peter/mix/demex/excrabs.html>. Citado na página 19.

MCLACHLAN, G.; KRISHNAN, T. **The EM Algorithm and Extensions**. [S.l.]: Wiley Series in Probability and Statistics, 1996. Citado na página 19.

MCLACHLAN, G.; PEEL, D. **Finite Mixture Models**. [S.l.]: Wiley Series in Probability and Statistics, 2000. Citado nas páginas 29 e 31.

PICARD, F. An Introduction to Mixture Models. Relatório Técnico. 2007. Citado na página 20.

SILVA, A. F. D. A Dirichlet Process Mixture Model for Brain MRI Tissue Classification. **Medical Image Analysis**, v. 11, p. 169–182, 2007. Citado na página 78.

TITTERINGTON, D. M.; SMITH, A. F. M.; MAKOV, U. E. **Statistical Analysis of Finite Mixture Distributions**. [S.l.]: Wiley, 1985. Citado na página 19.

WOOD, F.; BLACK, M. A Nonparametric Bayesian Alternative to Spike Sorting. **Journal of Neuroscience Methods**, v. 173, n. 1, p. 1–12, 2008. Citado na página 55.