UM TUTORIAL SOBRE ESTIMAÇÃO EM MODELOS DE MISTURA

Laísa Ribeiro ZUCARELI[1]
Erlandson Ferreira SARAIVA[2]
Adriano Kamimura SUZUKI[3]

- RESUMO: Nos últimos anos houve um crescente interesse em modelagens utilizando modelos de mistura de distribuições devido às suas flexibilidades. Neste artigo, apresentamos um tutorial sobre modelos de mistura e o método de estimação baseado na função de verossimilhança. Como uma das principais questões no procedimento de estimação dos parâmetros de um modelo de mistura é a determinação do número de componentes k, apresentamos uma discussão sobre a determinação de k utilizando os critérios de seleção de modelos AIC, AICc e BIC. Para verificar a performance dos critérios de seleção de modelos na determinação de k, desenvolvemos um estudo de simulação Monte Carlo considerando os modelos de mistura de distribuições normal, Poisson e Gama. Os resultados obtidos mostram uma complementaridade entre os critérios e que estes devem ser utilizados com certa cautela, pois apresentam porcentagens de acerto inferiores a 70% na maioria dos casos estudados. Aplicamos o método de estimação e os critérios de seleção de modelos a três conjuntos de dados reais.

- PALAVRAS-CHAVE: Modelo de mistura; estimador de Máxima verossimilhança; algoritmo EM estocástico; seleção de modelos.

[1]Universidade Federal de Mato Grosso do Sul - UFMS, Faculdade de Engenharias, Arquitetura e Urbanismo e Geografia, CEP: 79.070-900, Campo Grande, MS, Brasil. E-mail: *laisazucareli@gmail.com*
[2]Universidade Federal de Mato Grosso do Sul - UFMS, Instituto de Matemática, CEP: 79.070-900, Campo Grande, MS, Brasil. E-mail: *erlandson.saraiva@ufms.br*
[3]Universidade de São Paulo - USP, Departamento de Matemática Aplicada e Estatística, CEP: 13.566-590, São Carlos, SP, Brasil. E-mail: *suzuki@icmc.usp.br*

1 Introdução

Modelos de mistura de distribuições são utilizados para modelar fenômenos cujas observações são provenientes de uma população composta por k subpopulações, onde k pode ser conhecido ou desconhecido. Cada subpopulação pode ser modelada por uma densidade pertencente a alguma família de distribuições paramétricas. A densidade de cada subpopulação é chamada de componente da mistura e é ponderada pela frequência relativa da subpopulação em relação a toda população. Além disso, modelos de mistura fornecem uma conveniente e flexível família de distribuições para ajustar dados que não são bem modelados por qualquer família padrão de distribuições paramétricas, e também podem ser utilizadas como uma alternativa paramétrica a métodos não paramétricos de estimação de densidades (STEPHENS, 2000a,b).

Recentemente, este tipo de modelagem tem sido utilizada em diversas aplicações, tais como, em análises genômicas (BROËT *et al.*, 2008; SARAIVA e MILAN, 2012), análises epidemiológicas (DIEBOLT e ROBERT, 1994; GREEN e RICHARDSON, 2002), análises econômicas (JEDIDI *et al*, 1997; ALLENBY *et al.*, 1998), análise de finanças (LAMOREUX e LATSRAPS, 1994; ROBERT *et al.*, 2000), segmentação de imagens (ZHOU e ZHU, 2018), entre outras.

Neste artigo, apresentamos um tutorial sobre modelo de mistura e o procedimento de estimação baseado na função de verossimilhança. Apresentamos a função de verossimilhança para um modelo de mistura e mostramos que a maximização desta função equivale a uma maximização ponderada, onde os pesos são dados pelas probabilidades *a posteriori* das variáveis latentes. Dessa forma, para obter as estimativas de máxima verossimilhança para os parâmetros de interesse, consideramos a função de verossimilhança obtida de forma condicionada nos dados observados e não-observados (latentes), denominada de função de verossimilhança aumentada. Obtemos as estimativas de forma iterativa utilizando o algoritmo SEM (*Stochastic Expectation Maximization*).

Como um dos principais pontos no procedimento de estimação dos parâmetros de um modelo de mistura é a determinação do número de componentes, apresentamos uma discussão sobre a determinação de k utilizando os critérios de seleção de modelos AIC, AICc e BIC. Nesta abordagem, os parâmetros do modelo são estimados para diferentes valores de k previamente fixados e então um valor de k é selecionado baseado no menor valor AIC, AICc e BIC dentre os valores de k considerados. Desenvolvemos um estudo de simulação Monte Carlo para verificar a performance dos critérios de seleção de modelos na determinação de k utilizando diferentes modelos de mistura de distribuições e diferentes tamanhos amostrais.

Os resultados obtidos mostram uma baixa porcentagem de acerto dos três critérios. Somente em casos com médias bem separadas, segundo o critério de Behboodian (1970), o critério BIC apresentou porcentagem de acerto superior a 50%. Porém, em todos os casos simulados a porcentagem de acerto ficou abaixo de 70%. Estes resultados mostram que os três critérios devem ser utilizados com certa cautela para determinação do número de componentes de um modelo de mistura.

Como ilustração, aplicamos o procedimento de estimação e os critérios de seleção de modelos a três conjuntos de dados reais. O primeiro conjunto de dados é referente a velocidade de Galáxias que já foi análisado por diversos autores, tais como, Roeder e Wasserman (1990), Richardson e Green (1997), Stephens (2000A), Saraiva *et al.*, (2014), entre outros. O segundo refere-se a dados de expressão gênica, analisados por Baldi e Long (2001) e Saraiva *et al.*, (2016). O Terceiro, é um conjunto de dados referente a atividade enzimática no sangue obtido no website https://people.maths.bris.ac.uk/∼mapjg/mixdata. Além disso, disponibilizamos no material suplementar (SM) o procedimento de estimação e o estudo de simulação para modelos de mistura de distribuições Poisson e Gama.

O restante do artigo esta organizado da seguinte maneira. Na seção 2, descrevemos o modelo de mistura e o procedimento de estimação via máxima verossimilhança. Um estudo de simulação considerando o modelo de mistura de distribuições normais e diferentes tamanhos amostrais é apresentado na Seção 3. Na Seção 4, aplicamos o procedimento de estimação e os critérios de seleção de modelos a três conjuntos de dados reais. Por fim, a Seção 5 apresenta algumas conclusões com base nos resultados obtidos.

2 Modelos de mistura

Considere uma população composta de k subpopulações. Sejam $w_1, \ldots, w_k$ as frequências relativas das k subpopulações em relação a toda população, com $0 \leq w_j \leq 1$ e $\sum_{j=1}^{k} w_j = 1$. Assuma que cada subpopulação j desta população é modelada por uma distribuição $F(\theta_j)$ indexada pelo parâmetro θ_j (escalar ou vetor), para $j = 1, \ldots, k$.

Sob estas condições, é natural supor que o processo para obtenção de uma amostra aleatória $Y_1, \ldots, Y_n$ de tamanho n desta população consiste em selecionar uma subpopulação j com probabilidade w_j e em seguida retirar desta subpopulação um valor y_i, para $i = 1, \ldots, n$ e $j = 1, \ldots, k$. Assim, podemos representar cada elemento da amostra pelo par (y_i, c_i), onde c_i assume um dos valores pertencente ao conjunto $\{1, 2, \ldots, k\}$ com probabilidade $\{w_1, w_2, \ldots, w_k\}$, respectivamente, e $c_i = j$ indica que y_i foi retirado da subpopulação j, para $i = 1, \ldots, n$ e $j = 1, \ldots, k$. Dessa forma, temos que

$$Y_i | c_i = j, \theta_j \sim F(\theta_j) \text{ e } P(C_i = j | \mathbf{w}) = w_j, \tag{1}$$

para $i = 1, \ldots, n$ e $j = 1, \ldots, k$. Uma amostra observada é representada pelo par $(\mathbf{y}, \mathbf{c})$, onde $\mathbf{y} = (y_1, \ldots, y_n)$ é o vetor de observações independentes e $\mathbf{c} = (c_1, \ldots, c_n)$ é o vetor de variáveis indicadoras.

A distribuição conjunta de $(\mathbf{Y}, \mathbf{C})$ é dada por

$$P(\mathbf{Y} = \mathbf{y}, \mathbf{C} = \mathbf{c} | \boldsymbol{\theta}, \mathbf{w}) = P(\mathbf{Y} = \mathbf{y} | \mathbf{c}, \boldsymbol{\theta}) P(\mathbf{C} = \mathbf{c} | \mathbf{w}).$$

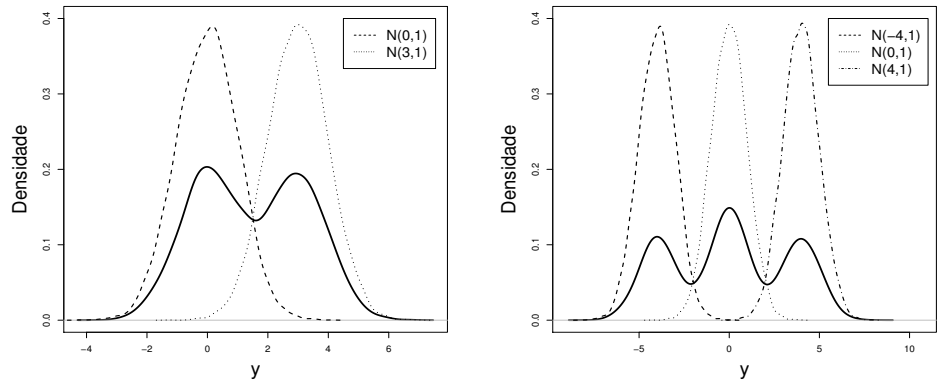
Porém, em muitos problemas práticos, as variáveis indicadoras $\mathbf{c}$ são variáveis

não-observáveis (latentes). Ou seja, somente os valores y_i's são observados. Dessa forma, cada observação y_i tem probabilidade w_j de ter sido retirada da subpopulação j, para $i = 1, \ldots, n$ e $j = 1, \ldots, k$. A função densidade de probabilidade marginal para $Y_i = y_i$ é dada por uma mistura de k componentes (subpopulação) da seguinte forma

$$f(y_i|\boldsymbol{\theta}, \mathbf{w}) = \sum_{j=1}^{k} w_j f(y_i|\theta_j), \tag{2}$$

onde $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ é o vetor de parâmetros e $f(\cdot|\theta_j)$ é a função densidade de probabilidade da distribuição $F(\cdot)$, para $i = 1, \ldots, n$. Cada função densidade de probabilidade $f(\cdot|\theta_j)$ é denominada de componente da mistura.

A Figura 1, mostra o gráfico da função densidade de probabilidade de dois modelos de misturas de distribuições normais. Nestes gráficos, a linha sólida representa a mistura e as linhas tracejadas representam as componentes separadas. Note que, no modelo de mistura de distribuições normais, temos k componentes e as observações pertencentes a uma determinada componente são normalmente distribuidas.



(a) $0,5\mathcal{N}(0,1) + 0,5\mathcal{N}(3,1)$. (b) $0,3\mathcal{N}(-4,1) + 0,4\mathcal{N}(0,1) + 0,3\mathcal{N}(4,1)$.

Figura 1 - Mistura de distribuições normais.

2.1 Probabilidade de alocação

Seja **y** uma amostra observada do modelo de mistura em (2). Assuma que o modelo de mistura em (2) seja completamente conhecido, *i.e.*, conhecemos os valores dos parâmetros $\boldsymbol{\theta}$, dos pesos **w** e de k.

Considere que o interesse seja em classificar cada observação y_i em relação as componentes. Ou seja, temos interesse em saber qual a probabilidade da observação y_i ter sido gerada pela componente j, para $i = 1, \ldots, n$ e $j = 1, \ldots, k$.

Para obtermos estas probabilidades, considere o modelo em (1), onde $\mathbf{c} = (c_1, \ldots, c_n)$ são as variáveis indicadoras. Dessa forma, nosso interesse é calcular a probabilidade de $C_i = j$ dado que observamos y_i, isto é, $P(C_i = j|y_i, \boldsymbol{\theta}, \mathbf{w})$. Utilizando o teorema de Bayes, as probabilidades de interesse são dadas por

$$\begin{aligned} P(C_i = j|y_i, \boldsymbol{\theta}, \mathbf{w}) &= \frac{P(Y_i = y_i, c_i = j|\boldsymbol{\theta}, \mathbf{w})}{P(Y_i = y|\boldsymbol{\theta}, \mathbf{w})} \\ &= \frac{P(Y_i = y_i|c_i = j, \boldsymbol{\theta})P(C_i = j|\mathbf{w})}{P(Y_i = y_i|\boldsymbol{\theta}, \mathbf{w})}, \end{aligned} \tag{3}$$

para $i = 1, \ldots, n$ e $j = 1 \ldots, k$.

Na expressão em (3), $P(C_i = j|\mathbf{w})$ é a probabilidade *a priori* de y_i ser amostrado da componente j. De (1), temos que esta probabilidade é igual ao tamanho da componente em relação a toda a população, *i.e.*, $P(C_i = j|\mathbf{w}) = w_j$. Além disso, se Y_i é uma variável aleatória discreta, então $P(Y_i = y_i|c_i = j, \boldsymbol{\theta}) = f(y_i|\theta_j)$, em que $f(\cdot)$ é a função de probabilidade. Dessa forma, $P(Y_i = y_i|\boldsymbol{\theta}, \mathbf{w})$ é dado pelo modelo de mistura em (2). Portanto, podemos escrever a probabilidade em (3) da seguinte maneira

$$P(C_i = j|y_i, \boldsymbol{\theta}, \mathbf{w}) = \frac{w_j f(y_i|\theta_j)}{\sum_{j=1}^{k} w_j f(y_i|\theta_j)}, \tag{4}$$

para $i = 1, \ldots, n$ e $j = 1, \ldots, k$. A expressão em (4) torna-se conveniente pois é válida se Y_i é uma variável aleatória contínua (ver, por exemplo, FRÜHWIRTH-SCHNATTER (2006), Capítulo 2, página 26).

2.2 Estimação dos parâmetros

Considere agora que $\mathbf{y}$ seja uma amostra observada do modelo de mistura em (2) e que os parâmetros $\boldsymbol{\theta}$ e os pesos $\mathbf{w}$ sejam quantidades desconhecidas e temos interesse em estimá-las. Assim, dado $\mathbf{y}$, a função de verossimilhança para $\boldsymbol{\theta}$ e $\mathbf{w}$ é dada por

$$L(\boldsymbol{\theta}, \mathbf{w}|\mathbf{y}) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta}) = \prod_{i=1}^{n} \left[\sum_{j=1}^{k} w_j f(y_i|\theta_j) \right].$$

A função log-verossimilhança é dada por

$$l(\boldsymbol{\theta}, \mathbf{w}|\mathbf{y}) = \log \left\{ \prod_{i=1}^{n} \left[\sum_{j=1}^{k} w_j f(y_i|\theta_j) \right] \right\} = \sum_{i=1}^{n} \log \left\{ \left[\sum_{j=1}^{k} w_j f(y_i|\theta_j) \right] \right\}. \tag{5}$$

O procedimento usual para obtenção dos estimadores de máxima verossimilhança consiste em obter as derivadas parciais de $l(\cdot)$ em relação a θ_j e igualar o resultado a zero, para $j = 1, \ldots, k$. Por exemplo, derivando (5) em relação a θ_j

obtemos

$$\begin{aligned}\frac{dl(\boldsymbol{\theta}, \mathbf{w}|\mathbf{y})}{d\theta_j} &= \sum_{i=1}^{n} \frac{1}{\sum_{j=1}^{k} w_j f(y_i|\theta_j)} w_j \frac{df(y_i|\theta_j)}{d\theta_j} \\ &= \sum_{i=1}^{n} \frac{w_j f(y_i|\theta_j)}{\sum_{j=1}^{k} w_j f(y_i|\theta_j)} \frac{1}{f(y_1|\theta_j)} \frac{df(y_i|\theta_j)}{d\theta_j} \\ &= \sum_{i=1}^{n} \frac{w_j f(y_i|\theta_j)}{\sum_{j=1}^{k} w_j f(y_i|\theta_j)} \frac{d\log\left[f(y_i|\theta_j)\right]}{d\theta_j}, \end{aligned} \tag{6}$$

para $j = 1, \ldots, k$.

O estimador de maxima verossimilhança $\hat{\theta}_j$ de θ_j é dado pela solução da equação

$$\frac{dl(\boldsymbol{\theta}, \mathbf{w}|\mathbf{y})}{d\theta_j} = 0,$$

para $j = 1, \ldots, k$.

Porém, note em (6), que o procedimento de maximização consiste de um processo de maximização ponderada da log-verossimilhança. Cada observação y_i tem um peso associado a componente j dado por

$$w_{ij}^{*} = \frac{w_j f(y_i|\theta_j)}{\sum_{j=1}^{k} w_j f(y_i|\theta_j)}. \tag{7}$$

No entanto, estes pesos dependem dos parâmetros que estamos querendo estimar. Dessa forma, não conseguiremos obter uma expressão matemática "fechada" que possibilite a maximização direta da log-verossimilhança.

Além disso, note que os pesos em (7) são as probabilidades de $C_i = j$ dados y_i, $\boldsymbol{\theta}$ e $\mathbf{w}$ apresentados na equação (4). Ou seja, embora os valores das variáveis indicadoras c_i's sejam não-observados em um modelo de mistura, a probabilidade de $C_i = j$ dado y_i, $\boldsymbol{\theta}$ e $\mathbf{w}$ está intrinsicamente presente no procedimento de estimação dado em (7). Isto permite-nos inserir as variáveis não-observáveis em modelo de mistura como sendo quantidades a serem estimadas.

2.2.1 Dados completos e função verossimilhança

Seja $(\mathbf{y}, \mathbf{c})$ os dados completos, sendo $\mathbf{y} = (y_1, \ldots, y_n)$ os dados observados e $\mathbf{c} = (c_1, \ldots, c_n)$ uma configuração conhecida. Assim, a função de verossimilhança para $\boldsymbol{\theta}$ e $\mathbf{w}$ dado $(\mathbf{y}, \mathbf{c})$ é dada por

$$
\begin{aligned}
L(\boldsymbol{\theta}, \mathbf{w}|\mathbf{y}, \mathbf{c}) &= P(\mathbf{Y} = \mathbf{y}|\mathbf{c}, \boldsymbol{\theta}) P(\mathbf{C} = \mathbf{c}|\mathbf{w}) \\
&= \prod_{i=1}^{n} f(y_i|c_i, \boldsymbol{\theta}) P(C_i = c_i|\mathbf{w}). \\
&= \prod_{i=1}^{n} \prod_{j=1}^{k} \left[w_j f(y_i|\theta_j)\right]^{z_{ij}} \\
&= \prod_{j=1}^{k} w_j^{n_j} L(\theta_j|\mathbf{y})
\end{aligned}
$$

em que, z_{ij} é uma variável indicadora, tal que, $z_{ij} = 1$ se $c_i = j$ e $z_{ij} = 0$ caso contrário, $n_j = \#\{y_i; c_i = j\}$ é quantidade de observações pertencentes a componente j e $L(\theta_j|\mathbf{y}) = \prod_{\{y_i; c_i = j\}} f(y_i|\theta_j)$ é a função de verosimilhança da componente j, para $i = 1, \ldots, n$ e $j = 1, \ldots, k$. Assim, temos que, condicional em $(\mathbf{y}, \mathbf{c})$, a função de verossimilhança é dada pelo produto de k funções de verossimilhanças locais.

A função log-verossimilhança é dada por

$$
l(\boldsymbol{\theta}, \mathbf{w}|\mathbf{y}, \mathbf{c}) = \log \left\{ \prod_{j=1}^{k} w_j^{n_j} L(\theta_j|\mathbf{y}) \right\} = \sum_{j=1}^{k} \left[n_j \log(w_j) + l(\theta_j|\mathbf{y}) \right],
$$

em que, $l(\theta_j|\mathbf{y}) = \log\left[L(\theta_j|\mathbf{y})\right]$.

Derivando $l(\boldsymbol{\theta}, \mathbf{w}|\mathbf{y}, \mathbf{c})$ em relação a θ_j, obtemos

$$
\frac{dl(\boldsymbol{\theta}, \mathbf{w}|\mathbf{y}, \mathbf{c})}{d\theta_j} = \frac{dl(\theta_j|\mathbf{y})}{d\theta_j},
$$

para $j = 1, \ldots, k$. O estimador de máxima verossimilhança $\hat{\theta}_j$ de θ_j é dado pela solução da equação

$$
\frac{dl(\theta_j|\mathbf{y})}{d\theta_j} = 0,
$$

para $j = 1, \ldots, k$. Ou seja, o procedimento de estimação dos k parâmetros se reduz a k independentes problemas de estimação. Neste ponto, é sempre necessário verificar que a segunda derivada é negativa para garantir que a solução é um ponto de máximo. Ou seja, para θ_j univariado, devemos ter

$$
\left.\frac{d^2 l(\theta_j|\mathbf{y})}{d\theta_j^2}\right|_{\theta_j = \hat{\theta}_j} < 0.
$$

Para θ_j multivariado, devemos ter a matriz Hessiana negativa definida.

Utilizando os multiplicadores de Lagrange, o estimador de máxima verossimilhança $\hat{w}_j$ de w_j é dado por

$$\hat{w}_j = \frac{n_j}{n},$$

para $j = 1, \ldots, k$.

2.3 Algoritmo SEM

Utilizando os resultados descritos nas Seções 2.1 e 2.2, obtemos um procedimento de estimação iterativo, em que, estimamos os parâmetros $(\boldsymbol{\theta}, \mathbf{w})$ condicional em uma configuração $\mathbf{c}$ e então atualizamos $\mathbf{c}$ condicional no valor estimado $(\hat{\boldsymbol{\theta}}, \hat{\mathbf{w}})$ de $(\boldsymbol{\theta}, \mathbf{w})$.

Assim, utilizando a equação em (7), atualizamos os valores das variáveis indicadoras c_i's de acordo com o algoritmo 1.

Algoritmo 1. Para $i = 1, \ldots, n$ e $j = 1, \ldots, k$:

(i) considere $\mathbf{z}_i = (z_{i1}, z_{i2}, \ldots, z_{ik})$ um vetor indicador, tal que, $z_{ij} = 0$ ou $z_{ij} = 1$;

(ii) gere $\mathbf{z}_i \sim Multinomial(1, 1, \mathbf{w}_i^*)$, onde $\mathbf{w}_i^* = (w_{i1}^*, \ldots, w_{ik}^*)$ e w_{ij}^* é dado em (7) para $\theta_j = \hat{\theta}_j$;

(iii) se $z_{ij} = 1$, então faça $c_i = j$;

Dessa forma, podemos definir o procedimento de estimação dos parâmetros de um modelo de mistura por três "passos":

(i) condicional nos valores das variáveis indicadoras $\mathbf{c}$, obtemos as estimativas $\hat{\boldsymbol{\theta}}$ e $\hat{\mathbf{w}}$ dos parâmetros $\boldsymbol{\theta}$ e $\mathbf{w}$;

(ii) condicional em $\mathbf{y}$, $\hat{\boldsymbol{\theta}}$ e $\hat{\mathbf{w}}$, atualizamos os valores dos pesos w_{ij}^* dado em (7).

(iii) atualizamos os valores das variáveis indicadoras $\mathbf{c}$ utilizando o algoritmo 1.

Estes três "passos" definem o algoritmo SEM (*Stochastic Expectation Maximization*). Uma descrição formal do algoritmo SEM no contexto de variáveis não-observáveis e maiores detalhes podem ser encontradas em Dempster *et al.*, (1977), Redner e Walker (1985), Meng (1997), MacLachlan e Pell (2000), Frühwirth-Schnatter (2006) e em suas referências.

Como o algoritmo SEM é baseado na repetição alternada dos passos (i), (ii) e (iii) descritos acima, então é necessário especificar um critério de parada para o algoritmo. Como a cada iteração do algoritmo temos um valor para log-verossimilhança observada $l(\hat{\boldsymbol{\theta}}, \hat{\mathbf{w}}|\mathbf{y})$, então um critério de parada é dado pela diferença absoluta

$$\left| l\left(\hat{\boldsymbol{\theta}}^{(s)}, \hat{\mathbf{w}}^{(s)}|\mathbf{y}\right) - l\left(\hat{\boldsymbol{\theta}}^{(s-1)}, \hat{\mathbf{w}}^{(s-1)}|\mathbf{y}\right)\right| \tag{8}$$

ou pela diferença relativa absoluta

$$\left| \frac{l\left(\hat{\boldsymbol{\theta}}^{(s)}, \hat{\mathbf{w}}^{(s)} | \mathbf{y}\right) - l\left(\hat{\boldsymbol{\theta}}^{(s-1)}, \hat{\mathbf{w}}^{(s-1)} | \mathbf{y}\right)}{l\left(\hat{\boldsymbol{\theta}}^{(s-1)}, \hat{\mathbf{w}}^{(s-1)} | \mathbf{y}\right)} \right|, \tag{9}$$

onde $\left(\hat{\boldsymbol{\theta}}^{(s)}, \hat{\mathbf{w}}^{(s)}\right)$ são as estimativas dos parâmetros na s-ésima iteração do algoritmo e $l(\cdot)$ é dada em (5). Dessa forma, se a diferença observada em (8) ou (9) for menor do que um valor de tolerância ϵ, $\epsilon > 0$, previamente especificado, então o algoritmo é parado e consideramos $\left(\hat{\boldsymbol{\theta}}, \hat{\mathbf{w}}\right) = \left(\hat{\boldsymbol{\theta}}^{(s)}, \hat{\mathbf{w}}^{(s)}\right)$ como sendo as estimativas dos parâmetros $(\boldsymbol{\theta}, \mathbf{w})$. Obviamente, os critérios em (8) e (9) não são os únicos critérios de parada. Por exemplo, também poderíamos utilizar como critério de parada uma mudança percentual nos valores dos parâmetros após cada iteração. Neste artigo, optamos por utilizar como critério de parada do algoritmo SEM a diferença relativa absoluta dada em (9) devido ao fato desta ser adimensional.

Desta forma o algoritmo SEM é dado pelos seguintes "passos".

Algoritmo SEM. Inicie o algoritmo com uma configuração $\mathbf{c}^{(0)} = \left(c_1^{(0)}, \ldots, c_n^{(0)}\right)$ para as variáveis não-observáveis. Para a s-ésima iteração do algoritmo, $s = 1, \ldots$:

(i) obtenha as estimativas de máxima verossimilhança $\hat{\boldsymbol{\theta}}^{(s)} = \left(\hat{\theta}_1^{(s)}, \ldots, \hat{\theta}_k^{(s)}\right)$ e $\hat{\mathbf{w}}^{(s)} = \left(\hat{w}_1^{(s)}, \ldots, \hat{w}_k^{(s)}\right)$ condicional em $\mathbf{c}^{(s-1)}$;

(ii) se $\left| \frac{l\left(\hat{\boldsymbol{\theta}}^{(s)}, \hat{\mathbf{w}}^{(s)} | \mathbf{y}\right) - l\left(\hat{\boldsymbol{\theta}}^{(s-1)}, \hat{\mathbf{w}}^{(s-1)} | \mathbf{y}\right)}{l\left(\hat{\boldsymbol{\theta}}^{(s-1)}, \hat{\mathbf{w}}^{(s-1)} | \mathbf{y}\right)} \right| < \epsilon$ pare o algoritmo. Caso contrário, vá para o item (iii);

(iii) condicionado em $\hat{\boldsymbol{\theta}}^{(s)}$ e $\hat{\mathbf{w}}^{(s)}$, atualize $\mathbf{c} = (c_1, \ldots, c_n)$ de acordo com o algoritmo 1 e obtenha $\mathbf{c}^{(s)}$. Faça $s = s + 1$ e retorne ao passo (i).

2.3.1 Exemplo: Mistura de distribuições Normais

Considere o modelo de mistura dado em (2) com $f(y_i|\theta_j)$ sendo a densidade da distribuição normal com média μ_j e variância σ_j^2, *i.e.*,

$$f(y_i|\theta_j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left\{ -\frac{1}{2\sigma_j^2}(y_i - \mu_j)^2 \right\},$$

onde $\theta_j = (\mu_j, \sigma_j^2)$, $i = 1, \ldots, n$ e $j = 1, \ldots, k$.

Condicional em uma configuração $\mathbf{c}$, a função log-verossimilhança para uma componente j é dada por

$$l(\theta_j|\mathbf{y},\mathbf{c}) \propto -\frac{n_j}{2}\log(\sigma_j^2) - \frac{1}{2\sigma_j^2}\sum_{\{y_i;c_i=j\}}(y_i-\mu_j)^2.$$

Derivando $l(\theta_j|\mathbf{y},\mathbf{c})$ com relação a μ_j e igualando o resultado a zero, obtemos como estimador da média

$$\hat{\mu}_j = \bar{y}_j = \frac{1}{n_j}\sum_{\{y_i;c_i=j\}} y_i,$$

i.e., a média das observações alocadas na componente j, para $j = 1,\ldots,k$.

De forma análoga, o estimador $\hat{\sigma}_j^2$ de σ_j^2 é dado pela variância das observações alocadas na componente j, *i.e.*,

$$\hat{\sigma}_j^2 = s_j^2 = \frac{1}{n_j-1}\sum_{\{y_i;c_i=j\}}(y_i-\hat{y}_j)^2,$$

para $j = 1,\ldots,k$.

Algoritmo SEM para mistura de distribuições normais. Inicie o algoritmo com uma configuração $\mathbf{c}^{(0)} = \left(c_1^{(0)},\ldots,c_n^{(0)}\right)$ para as variáveis não-observáveis. Para a s-ésima iteração do algoritmo, $s = 1,\ldots$:

(i) obtenha as estimativas de máxima verossimilhança $\hat{\mu}_j^{(s)}$, $\hat{\sigma}_j^{2(s)}$ e $\hat{w}_j^{(s)}$ condicional em $\mathbf{c}^{(s-1)}$;

(ii) se $\left|\frac{l\left(\hat{\boldsymbol{\theta}}^{(s)},\hat{\mathbf{w}}^{(s)}|\mathbf{y}\right)-l\left(\hat{\boldsymbol{\theta}}^{(s-1)},\hat{\mathbf{w}}^{(s-1)}|\mathbf{y}\right)}{l\left(\hat{\boldsymbol{\theta}}^{(s-1)},\mathbf{w}^{(s-1)}|\mathbf{y}\right)}\right| < \epsilon$ pare o algoritmo. Caso contrário, vá para o item (iii);

(iii) gere $\mathbf{z}_i \sim Multinomial(1,1,\mathbf{w}_i^*)$, onde $\mathbf{w}_i^* = (w_{i1}^*,\ldots,w_{ik}^*)$ e

$$w_{ij}^* \propto \hat{w}_j^{(s)}\exp\left\{-\frac{1}{2\hat{\sigma}_j^{2(s)}}\left(y_i-\hat{\mu}_j^{(s)}\right)^2\right\},$$

para $i = 1,\ldots,n$ e $j = 1,\ldots,k$. Se $z_{ij} = 1$, faça $c_i^{(s)} = j$, para $i = 1,\ldots,n$ e $j = 1,\ldots,k$. Faça $s = s+1$ e retorne ao passo (i).

2.4 Dificuldades práticas e não identificabilidade

Embora seja simples de se implementar um algoritmo SEM, este pode apresentar algumas dificuldades práticas para a obtenção dos estimadores de máxima verossimilhança. A primeira, é que pode ser difícil encontrar um estimador de máximo global da verossimilhaça numericamente. Em muitos casos, tal como em

um modelo de mistura de distribuições normais, um estimador de máximo global pode não existir; como discutido por MacLachlan e Peel (2000).

Além disso, dada uma amostra de tamanho n, existe uma probabilidade positiva, dada por $(1 - w_j)^n \neq 0$, de que a componente j não tenha observações associadas. Nesta caso, os estimadores de máxima verossimilhança não podem ser calculados. Isto pode ser facilmente notado, observando os estimadores $\hat{\mu}_j$ e $\hat{\sigma}_j^2$ de um modelo de mistura de distribuições normais descrito na Seção 2.3.1.

O algoritmo SEM também pode apresentar convergência lenta. Diversos artigos discutem a não convergência do algoritmo, principalmente quando o tamanho da amostra é pequeno ou as componentes não sejam bem separadas, ver por exemplo Finch *et al.* (1989). Alguns autores, tais como Karlis e Xekalaki (2003) e Biernacki (2003), discutem como escolher os valores iniciais do algoritmo para aumentar as chances de convergência do algoritmo. Neste texto, obtemos a configuração inicial $\mathbf{c}^{(0)}$ para as variáveis latentes utilizando o algoritmo *k-means* (MACQUEEN, 1967).

Além dos fatos citados acima, o modelo de mistura em (2) é invariante em relação a marcação j ($j = 1, \ldots, k$) das componentes. Isto implica que os parâmetros das componentes são não identificáveis marginalmente, isto é, não podemos distinguir, por exemplo, a componente 1 (θ_1) da componente 2 (θ_2) a partir da função de verossimilhança, devido ao fato delas serem permutáveis. Este problema é conhecido na literatura como *"label switching"*. Para maiores detalhes sobre *label switching* ver Stephens (2000b) e suas referências.

Dessa forma, para garantir a identificabilidade do modelo e ser capaz de desenvolver um procedimento de estimação para os parâmetros de cada componente, devemos considerar um único tipo de marcação. Por exemplo, podemos marcar as componentes de acordo com a ordem crescente das médias, $\mu_1 < \mu_2 < \ldots < \mu_k$, em que, a primeira componentes é a que tem média μ_1, a segunda é a que tem média μ_2 e sucessivamente até a última componente que tem média μ_k. Este procedimento também pode ser feito utilizando os pesos w_j ou as variâncias σ_j^2 ou qualquer outro parâmetro de referência, para $j = 1, \ldots, k$. Neste artigo, seguimos o padrão da literatura e consideramos a marcação das componentes segundo a ordem crescente das médias (ver por exemplo, RICHARDSON e GREEN, 1997, STEPHENS, 2000a e SARAIVA *et al.*, 2016).

2.5 Número de componentes

Em muitas situações práticas, a amostra aleatória $Y_1, \ldots, Y_n$ é proveniente de um modelo de mistura de distribuições, em que o número de componentes k é desconhecido. Nestes casos, é necessário um procedimento que possibilite estimar k e os parâmetros das componentes.

Na literatura existem duas abordagens distintas para estimação de k. Na primeira, os parâmetros do modelo são estimados para diferentes valores de k previamente fixados; e então um valor de k é selecionado segundo algum critério de seleção de modelos, tais como o AIC (*Akaike information criterion*) ou o BIC (*Bayesian information criterion*).

Na segunda abordagem, o valor de k é estimado de forma conjunta com os parâmetros do modelo. Neste caso, algoritmos mais sofisticados tais como o algoritmo *reversible-jump* (RICHARDSON e GREEN, 1997), o algoritmo *birth-death* (STEPHENS, 2000a) ou o algoritmo *split-merge* (SARAIVA*et al.*, 2014) são utilizados para a estimação.

Nesta seção, focamos na primeira forma de abordagem, isto é, nos procedimentos de estimação de k baseados em critérios de seleção de modelos. Como a função log-verossimilhança de um modelo de mistura "aumenta" quando adicionamos novas componentes, então os critérios de seleção de modelos a serem utilizados devem levar em consideração algum tipo de penalização da função log-verossimilhança de modo a evitar a escolha de modelos com grande número de parâmetros, ou seja, um grande número de componentes.

Dois critérios de seleção de modelos com essa característica são o AIC e o BIC, dados respectivamente, por

$$AIC_k = -2l(\hat{\boldsymbol{\theta}}, \hat{\mathbf{w}}|\mathbf{y}) + 2d_k \text{ e } BIC_k = -2l(\hat{\boldsymbol{\theta}}, \hat{\mathbf{w}}|\mathbf{y}) + d_k \log(n), \tag{10}$$

em que $l(\hat{\boldsymbol{\theta}}, \hat{\mathbf{w}}|\mathbf{y})$ é a função log-verossimilhança dada em (5) com $\left(\hat{\boldsymbol{\theta}}, \hat{\mathbf{w}}\right)$ sendo as estimativas de máxima verossilhança obtidas pelo algoritmo SEM e d_k é a quantidade de parâmetros "livres" do modelo com k componentes. Por exemplo, um modelo de mistura de distribuições normais com $k = 2$ componentes, tem $d_2 = 5$, correspondentes aos parâmetros μ_1, μ_2, σ_1, σ_2 e w_1. O modelo com menor valor AIC ou BIC é o modelo que melhor ajusta os dados observados.

Além dos critérios AIC e BIC também consideramos o AIC modificado, proposto por Bozdogan (1987), dado por

$$AICc = -2l(\hat{\boldsymbol{\theta}}, \hat{\mathbf{w}}|\mathbf{y}) + 2d_k + 2\frac{d_k(d_k+1)}{n-d_k-1}.$$

Analogamente, quanto menor o valor de AICc melhor o ajuste do modelo aos dados.

3 Estudo de simulação

Nesta seção, apresentamos os resultados de um estudo de simulação considerando o modelo de mistura de distribuições normais. Para desenvolvermos as simulações fixamos o tamanho da amostra $n = 100, 200, 500$ e 1.000.

Para geração dos dados simulados fixamos os parâmetros dos modelos conforme descritos na Tabela 1. Os valores foram escolhidos de forma a explorar diferentes tipos de modelos de mistura em relação ao número de componentes e a distância das médias das componentes. Com relação a distância entre as médias das componentes utilizamos o critério proposto por Behboodian (1970) para que duas componentes de uma mistura de distribuições normais sejam consideradas unimodal, que é dado por $\Delta = (\mu_j - \mu_{j'}) \leq 2min(\sigma_j, \sigma_{j'})$, em que j e j' representam duas componentes adjacentes.

Os conjuntos de dados A_1, A_2 e A_3 possuem duas componentes, sendo que no conjunto de dados A_1 temos $\Delta = 2$, isto é, as duas componentes são consideradas unimodal. Para os conjuntos de dados A_2 e A_3, temos $\Delta > 2$, isto é, as componentes definem um modelo bimodal. Os conjuntos de dados A_4 e A_5 apresentam três componentes, sendo que, A_4 têm as componentes adjacentes 1 e 2 com $\Delta > 2$ e componentes 2 e 3 com $\Delta = 2$; enquanto que A_5 têm as componentes adjacentes 1 e 2 e 2 e 3 com $\Delta > 2$. O conjunto de dados A_6 têm quatro componentes sendo $\Delta > 2$ para cada duas componentes adjacentes. Para os conjuntos de dados A_1 à A_6 consideramos $\sigma = 1$. No Apêndice A do SM apresentamos os resultados das simulações com $\sigma \neq 1$.

Tabela 1 - Número de componentes e valores para os parâmetros usados para geração dos dados simulados

Dados simulados	Número de componenets	Valores dos parâmetros
A_1	$k_{real} = 2$	$\mu_1 = 0, \quad \mu_2 = 2$ $\sigma_1 = 1, \quad \sigma_2 = 1$ $w_1 = 0,50, \quad w_2 = 0,50$
A_2	$k_{real} = 2$	$\mu_1 = 0, \quad \mu_2 = 3$ $\sigma_1 = 1, \quad \sigma_2 = 1$ $w_1 = 0,50, \quad w_2 = 0,50$
A_3	$k_{real} = 2$	$\mu_1 = 0, \quad \mu_2 = 4$ $\sigma_1 = 1, \quad \sigma_2 = 1$ $w_1 = 0,50, \quad w_2 = 0,50$
A_4	$k_{real} = 3$	$\mu_1 = -3, \quad \mu_2 = 0, \quad \mu_3 = 2$ $\sigma_1 = 1, \quad \sigma_2 = 1, \quad \sigma_3 = 1$ $w_1 = 0,30, \quad w_2 = 0,40, \quad w_3 = 0,30$
A_5	$k_{real} = 3$	$\mu_1 = -3, \quad \mu_2 = 0, \quad \mu_3 = 3$ $\sigma_1 = 1, \quad \sigma_2 = 1, \quad \sigma_3 = 1$ $w_1 = 0,30, \quad w_2 = 0,40, \quad w_3 = 0,30$
A_6	$k_{real} = 4$	$\mu_1 = -4, \quad \mu_2 = 0, \quad \mu_3 = 4, \quad \mu_4 = 8$ $\sigma_1 = 1, \quad \sigma_2 = 1, \quad \sigma_3 = 1, \quad \sigma_4 = 1$ $w_1 = 0,25, \quad w_2 = 0,25, \quad w_3 = 0,25, \quad w_4 = 0,25$

Os dados foram gerados da seguinte maneira. Para $i = 1, \ldots, n$,

(i) Gere U_i de uma distribuição Uniforme em $(0, 1)$, $U_i \sim \mathcal{U}(0, 1)$; se $\sum_{j'=1}^{j-1} w_j < u_i \leq \sum_{j'=1}^{j} w_j$, gere $Y_i \sim \mathcal{N}(\mu_j; \sigma_j^2)$, com valores para os parâmetros fixados de acordo com os especificados na Tabela 1, e $w_0 = 0$, para $j = 1, \ldots, k$.

(ii) Para "guardar" de qual componente cada observação foi gerada, definimos o vetor $G = (G_1, \ldots, G_n)$, tal que, $G_i = j$ se $Y_i \sim N(\mu_j, \sigma_j^2)$, para $j = 1, \ldots, k$.

Para verificar o desempenho dos métodos de seleção de modelos AIC, $AICc$ e BIC na escolha do número de componentes k de um modelo de mistura, geramos $M = 1.000$ conjuntos de dados diferentes para cada conjunto de dados A_m, $m = 1, \ldots, 6$, e calculamos para cada um dos M conjuntos de dados gerados o valor AIC, $AICc$ e BIC. Para os conjunto de dados A_1, A_2 e A_3 consideramos $k \in \{1, 2, 3, 4, 5\}$; e para os conjuntos de dados A_4, A_5 e A_6 consideramos $k \in \{1, \ldots, 2k_{real}\}$.

A Tabela 2 mostra a proporção de vezes que um modelo com k componentes foi selecionado como sendo o melhor modelo dentre os considerados, para A_1 a A_6.

Tabela 2 - Proporção de valores de k selecionados para A_1 a A_6

Dados	k_{real}	k	Tamanho amostral											
			100			200			500			1.000		
			AIC	*AICc*	*BIC*	*AIC*	*AICc*	*BIC*	*AIC*	*AICc*	*BIC*	*AIC*	*AICc*	*BIC*
A_1	2	1	0,327	0,362	0,638	0,281	0,292	0,595	0,220	0,223	0,481	0,151	0,152	0,360
		2	0,280	0,301	0,285	0,279	0,287	0,295	0,281	0,284	0,332	0,302	0,302	0,376
		3	0,177	0,180	0,060	0,191	0,187	0,084	0,203	0,206	0,135	0,188	0,188	0,158
		4	0,116	0,094	0,016	0,141	0,135	0,019	0,152	0,150	0,033	0,200	0,199	0,084
		5	0,100	0,063	0,001	0,108	0,099	0,007	0,144	0,137	0,019	0,159	0,159	0,022
A_2	2	1	0,090	0,111	0,364	0,042	0,044	0,214	0,004	0,005	0,038	0,000	0,000	0,001
		2	0,394	0,438	0,487	0,385	0,404	0,550	0,376	0,379	0,634	0,371	0,372	0,578
		3	0,243	0,247	0,119	0,276	0,282	0,183	0,267	0,267	0,221	0,241	0,241	0,238
		4	0,166	0,133	0,026	0,175	0,161	0,044	0,199	0,195	0,089	0,201	0,201	0,120
		5	0,107	0,071	0,004	0,122	0,109	0,009	0,154	0,154	0,018	0,187	0,186	0,063
A_3	2	1	0,000	0,000	0,019	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
		2	0,417	0,455	0,705	0,391	0,410	0,661	0,344	0,347	0,613	0,347	0,351	0,562
		3	0,263	0,283	0,218	0,296	0,295	0,241	0,249	0,250	0,233	0,267	0,268	0,261
		4	0,184	0,165	0,051	0,185	0,183	0,076	0,252	0,252	0,127	0,196	0,196	0,117
		5	0,136	0,097	0,007	0,128	0,112	0,022	0,155	0,151	0,027	0,190	0,185	0,060
A_4	3	1	0,106	0,140	0,431	0,062	0,069	0,291	0,004	0,004	0,060	0,000	0,000	0,007
		2	0,294	0,345	0,407	0,300	0,318	0,487	0,216	0,221	0,486	0,179	0,182	0,375
		3	0,241	0,249	0,126	0,255	0,264	0,173	0,285	0,288	0,308	0,300	0,300	0,397
		4	0,147	0,136	0,028	0,161	0,152	0,036	0,237	0,239	0,117	0,200	0,199	0,134
		5	0,125	0,090	0,007	0,125	0,116	0,009	0,152	0,148	0,025	0,183	0,181	0,065
		6	0,097	0,040	0,001	0,097	0,081	0,004	0,106	0,100	0,004	0,138	0,138	0,022
A_5	3	1	0,100	0,148	0,479	0,051	0,059	0,313	0,003	0,004	0,049	0,001	0,001	0,001
		2	0,196	0,236	0,308	0,113	0,132	0,294	0,025	0,026	0,156	0,003	0,003	0,045
		3	0,287	0,314	0,189	0,332	0,348	0,303	0,387	0,395	0,561	0,363	0,364	0,585
		4	0,182	0,162	0,021	0,220	0,216	0,073	0,250	0,252	0,185	0,264	0,266	0,245
		5	0,127	0,091	0,003	0,166	0,150	0,014	0,188	0,184	0,038	0,207	0,207	0,092
		6	0,108	0,049	0,000	0,118	0,095	0,003	0,147	0,139	0,011	0,162	0,159	0,032
A_6	4	1	0,000	0,000	0,046	0,000	0,000	0,002	0,000	0,000	0,000	0,000	0,000	0,000
		2	0,003	0,011	0,131	0,000	0,000	0,010	0,000	0,000	0,000	0,000	0,000	0,000
		3	0,088	0,152	0,317	0,042	0,054	0,206	0,002	0,002	0,030	0,000	0,000	0,002
		4	0,279	0,347	0,364	0,311	0,352	0,540	0,325	0,337	0,611	0,334	0,338	0,581
		5	0,241	0,244	0,112	0,275	0,285	0,197	0,265	0,269	0,260	0,233	0,235	0,262
		6	0,164	0,131	0,026	0,170	0,164	0,037	0,161	0,158	0,070	0,178	0,178	0,111
		7	0,129	0,077	0,004	0,118	0,093	0,006	0,155	0,147	0,024	0,145	0,143	0,036
		8	0,096	0,038	0,000	0,084	0,052	0,002	0,092	0,087	0,005	0,110	0,106	0,008

Para o conjunto de dados A_1, onde as componentes 1 e 2 são consideradas unimodal pelo critério de Behboodian (1970), $\Delta = 2$, os três critérios apresentam maior proporção no modelo com $k = 1$ para $n \in \{100, 200\}$; para $n = 500$, os critérios AIC e $AICc$ apresentam maior proporção no verdadeiro modelo ($k = 2$) e o BIC apresenta maior proporção em $k = 1$; para $n = 1.000$ os três critérios indicam o verdadeiro modelo, $k = 2$, com maior proporção. Aumentando o tamanho da amostra n, a proporção de vezes que o AIC e o BIC selecionam o verdadeiro modelo também aumenta. Porém, as proporções de vezes que os três critérios selecionam o verdadeiro modelo são todas inferiores a $0,38$ indicando uma baixa performance em selecionar o número de componentes k do verdadeiro modelo.

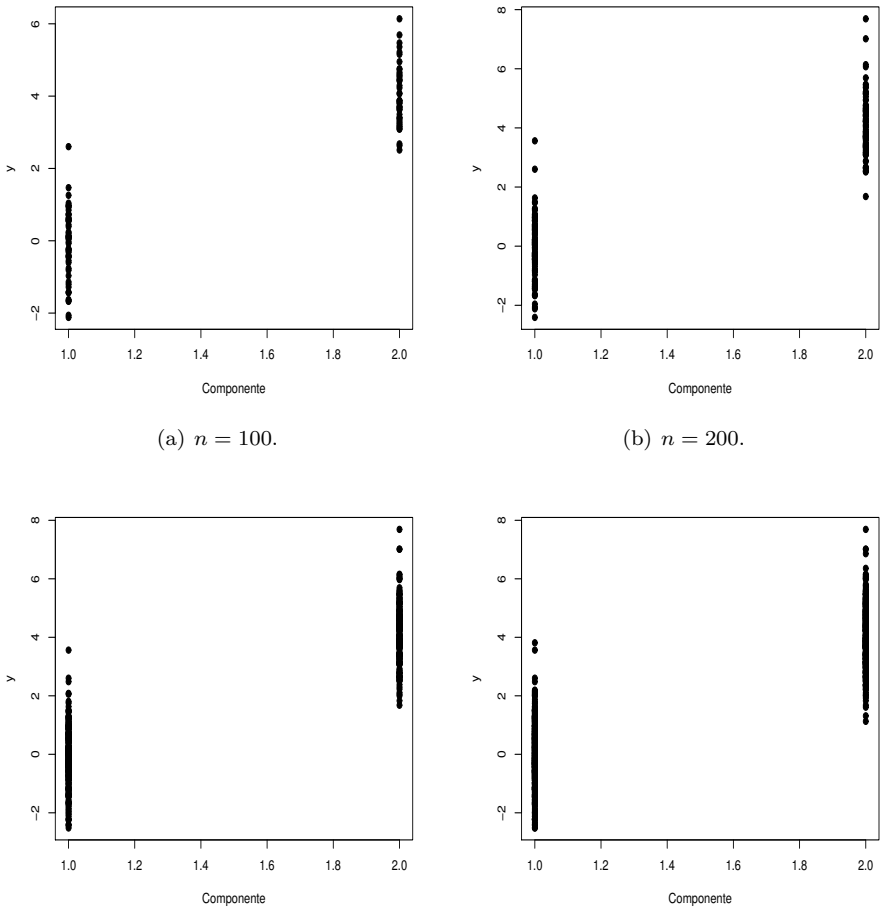
Para o conjunto de dados A_2, os três critérios apresentam maior proporção no valor de k referente ao verdadeiro modelo. Entretanto, note que, exceto para $n = 100$, somente o critério BIC apresenta proporção maior do que $0,50$. Neste caso, o BIC também apresenta um aumento na proporção de vezes que seleciona o verdadeiro modelo quando aumentamos o tamanho da amostra. Já os critérios AIC e $AICc$ apresentam uma diminuição na proporção.

Para A_3, onde as duas componentes apresentam médias bem separadas, novamente somente o BIC apresenta proporção em $k = 2$ (verdadeiro modelo) maior do que $0,50$. Além disso, note que para este caso, as proporções em $k = 2$ para os três critérios diminuem quando aumentamos o tamanho da amostra e os critérios tendem a favorecer modelos com mais componentes. Isto ocorre, pois ao aumentarmos o tamanho da amostra n temos mais observações na região onde as componentes se "misturam". Como ilustração, apresentamos na Figura 2 um conjunto de dados gerados sob as condições do conjunto de dados A_3. Note que, à medida que aumentamos o tamanho amostral n mais observações são geradas em regiões onde as componentes se "misturam". Como os critérios são calculados com base na função log-verossimilhança dada em (5), que é uma ponderação de cada observação em relação as componentes, então os métodos tendem a favorecer modelos com mais componentes.

Para A_4, com $k = 3$, sendo as componentes 1 e 2 com $\Delta = 2$ e componentes 2 e 3 com $\Delta > 2$, os três critérios apresentam maior proporção em modelos com $k = 2$ para $n \in \{100, 200\}$ e maiores proporções em $k = 3$ (verdadeiro modelo) para $n \in \{500, 1.000\}$. Para $n = \{100, 200\}$, AIC e $AICc$ apresentam maior proporção sob o modelo com $k = 3$ (verdadeiro modelo) do que o BIC. Porém, para $n = \{500, 1.000\}$ o BIC apresenta maior proporção.

Para o conjunto de dados A_5, as proporções sob o verdadeiro modelo pelos critérios AIC e $AICc$ para $n = \{100, 200\}$ é maior do que o BIC; porém para $n \in \{500, 1.000\}$ o BIC apresenta maior proporção. Além disso, note que, somente o critério BIC com $n = 500$ e 1.000 apresenta proporção sob o verdadeiro modelo maior do que $0,50$.

Para A_6, onde as médias das componentes são bem separadas, os três critérios apresentam maior proporção no valor de k referente ao verdadeiro modelo. Entretanto, exceto para $n = 100$, somente o critério BIC apresenta proporção maior do que $0,50$.



(a) $n = 100$. (b) $n = 200$.

(c) $n = 500$. (d) $n = 1.000$.

Figura 2 - Valores gerados de cada componente para o conjunto de dados A_3.

Estes resultados mostram uma complementaridade entre os critérios. Para $n \in \{100, 200\}$ temos uma melhor performance do AIC e do $AICc$ em relação ao BIC; sendo que o $AICc$ apresenta melhores resultado do que o AIC em todos os casos simulados. Para $n \in \{500, 1.000\}$ o BIC apresenta melhores resultados do que o AIC e o $AICc$.

Porém, os resultados também mostram que estes critérios devem ser utilizados com certo cuidado. Pois apresentam proporção de acerto inferior a 71% para todos os casos simulados. Somente para o conjunto de dados A_3, com duas componentes e médias bem separadas, e $n = 100$ o critério BIC apresentou proporção de acerto igual a $70,5\%$; para todos os outros casos simulados a proporção de acerto foi

inferior. Destacamos que os critérios AIC e $AICc$ em nenhum caso simulado apresentou proporção sob o verdadeiro modelo maior do que 50%. Já o BIC apresentou proporção sob o verdadeiro modelo maior do que 50% somente nos casos com componentes bem separadas ($\Delta > 2$) e tamanhos amostrais grandes, $n = \{500, 1.000\}$.

4 Aplicação

Nesta seção, aplicamos o método de estimação e os critérios de seleção de modelos a três conjuntos de dados reais. Como resultado, apresentamos para cada conjunto de dados, o número de componentes do modelo de mistura e as estimativas dos parâmetros.

4.1 Velocidade de Galáxias

Nesta primeira aplicação utilizamos o conjunto de dados referente ao estudo de galáxias. Os dados consistem de 82 observações das velocidades das galáxias. Este conjunto de dados foi previamente descrito e análisado por Roeder e Wasserman (1990), Richardson e Green (1997), Stephens (2000), Saraiva *et al.*, (2014), entre outros. Seguindo estes autores, consideramos que as velocidades são realizações de variáveis aleatórias independentes e identicamente distribuídas de acordo com um modelo de mistura de distribuições normais.

Aplicamos os critérios AIC, $AICc$ e BIC considerando $k \in \{2, \ldots, 10\}$. A Tabela 3 mostra o valor de cada critério para cada valor de k considerado. Os critérios $AICc$ e BIC indicam o melhor modelo como sendo um modelo com $k = 5$ componentes, enquanto que o AIC indica um modelo com $k = 7$ componentes.

Tabela 3 - Valores de AIC, $AICc$ e BIC

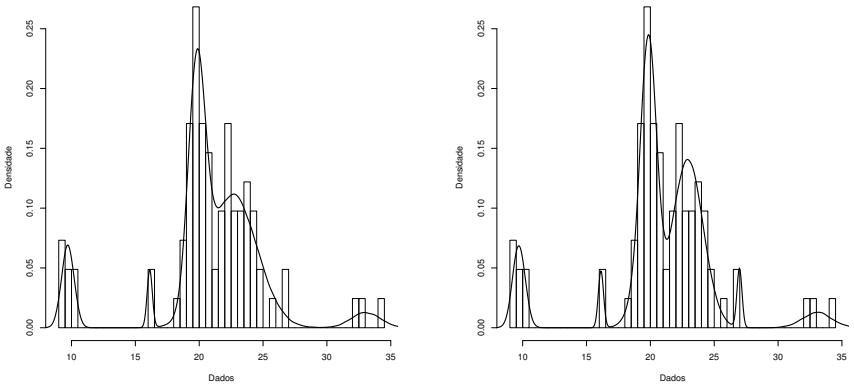
Critério	Número de componentes.								
	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$	$k = 7$	$k = 8$	$k = 9$	$k = 10$
AIC	450,9299	427,1429	427,4262	410,3159	415,6604	**406, 8655**	408,5378	417,2381	410,7293
$AICc$	451,7193	429,1155	431,1977	**416, 5846**	425,2229	420,6360	427,5723	442,7654	444,1909
BIC	462,9635	446,3967	453,9002	**444, 0100**	456,5746	454,9999	463,8923	479,8128	480,5242

Na Tabela 4 apresentamos as estimativas dos parâmetros para o modelo com $k = 5$ e $k = 7$ componentes. A Figura 3 mostra o histograma dos dados observados e a densidade dos modelos estimados.

Para uma comparação com outros resultados descritos na literatura, as estimativas para o número de componentes k para este conjunto de dados varia entre 3 e 4 em Roeder e Wasserman (1990) e entre 5 e 7 em Richardson e Green (1997) e Saraiva *et al.*, (2014).

Tabela 4 - Estimativas dos parâmetros para $k = 5$ e $k = 7$

Parâmetro	Estimativa		Parâmetro	Estimativa		Parâmetro	Estimativa	
	$k = 5$	$k = 7$		$k = 5$	$k = 7$		$k = 5$	$k = 7$
μ_1	9,7101	9,7101	σ_1	0,4564	0,4564	w_1	0,0853	0,0853
μ_2	16,1270	16,1270	σ_2	0,0608	0,0608	w_2	0,0244	0,0244
μ_3	19,8137	19,3353	σ_3	0,6391	0,8609	w_3	0,3293	0,0732
μ_4	22,6477	19,8670	σ_4	1,8807	0,5875	w_4	0,5244	0,3293
μ_5	33,0443	22,9270	σ_5	1,2743	1,0624	w_5	0,0366	0,4268
μ_6	—	26,9775	σ_6	—	0,0345	w_6	—	0,0244
μ_7	—	33,0443	σ_7	—	1,1288	w_7	—	0,0366



(a) $k = 5$. (b) $k = 7$.

Figura 3 - Histograma dos dados observados e densidades dos modelos estimados.

4.2 Dados de Expressão Gênica

Considere agora os dados de expressão gênica descrito em Baldi e Long (2001). O conjunto de dados é composto por $n = 4.290$ genes, com 4 medidas de níveis de expressão para cada gene em uma situação de controle e 4 medidas de níveis de expressão em uma situação de tratamento.

Seja $y_i = \overline{x}_{ic} - \overline{x}_{it}$ o efeito médio devido ao tratamento para o gene i, em que $\overline{x}_{ic}$ e $\overline{x}_{it}$ são, respectivamente, as médias dos níveis de expressão observadas na situação de controle e tratamento, para $i = 1, \ldots, n$.

Assumimos que genes com efeitos de tratamento similares definem um cluster. Além disso, assumimos que os efeitos de tratamento dos genes pertencentes a um mesmo cluster são uma amostra aleatória proveniente de uma mesma distribuição normal. Porém, esta distribuição normal é diferente em termos de valores de parâmetros da distribuição normal associada a outro cluster.

Assim, considerando k (desconhecido) como sendo o número de clusters e $w_1, \ldots, w_k$ como sendo a frequência relativa de cada cluster em relação a todos os genes, temos um modelo de mistura como em (1). Determinamos o número de componentes k e as estimativas dos parâmetros utilizando os critérios de seleção de modelos e o algoritmo SEM.

A Tabela 5 mostra o valor de cada critério. Para este conjunto de dados os três critérios indicam um modelo com $k = 3$ componentes como sendo o modelo que melhor ajusta os dados observados. As estimativas dos parâmetros para o modelo com $k = 3$ são apresentadas na Tabela 6. A Figura 4 mostra o histograma dos dados e a densidade do modelo estimado.

Tabela 5 - Valores de AIC, $AICc$ e BIC

Critério	Número de componentes.								
	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$	$k = 7$	$k = 8$	$k = 9$	$k = 10$
AIC	5.711,098	5.671,810	5.714,596	5.695,249	5.690,242	5.692,093	5.688,871	5.696,414	5.699,399
$AICc$	5.711,121	5.671,865	5.714,698	5.695,412	5.690,480	5.692,419	5.689,301	5.696,961	5.700,078
BIC	5.740,401	5.718,694	5.779,062	5.777,297	5.789,872	5.809,304	5.823,665	5.848,789	5.869,356

Tabela 6 - Estimativas dos parâmetros para $k = 3$

Parâmetro	Estimativa	Parâmetro	Estimativa	Parâmetro	Estimativa
μ_1	-0,0503	σ_1	0,2725	w_1	0,4828
μ_2	0,0947	σ_2	0,8584	w_2	0,3617
μ_3	0,3868	σ_3	1,6504	w_3	0,1554

4.3 Dados de Enzima

Considere agora os dados de Enzima obtido no website https://people.maths.bris.ac.uk/∼mapjg/mixdata. Este conjunto de dados refere-se à atividade enzimática no sangue, por uma enzima envolvida no metabolismo de substâncias carcinogênicas, em um grupo de 245 indivíduos.

A Tabela 7 mostra o valor de cada critério para este conjunto de dados. Os critérios AIC e AICc indicam um modelo com $k = 4$ componentes, enquanto que o critério BIC indica um modelo com $k = 3$ componentes. Para efeito de comparação, a estimativa para k dada em Richardson e Green (1997) através do uso do algoritmo reversible-jump varia entre $k = 3$ e $k = 5$.

As estimativas dos parâmetros para os modelos com $k = 3$ e $k = 4$ são apresentadas na Tabela 8. A Figura 5 mostra o histograma dos dados e a densidade dos modelos estimados.

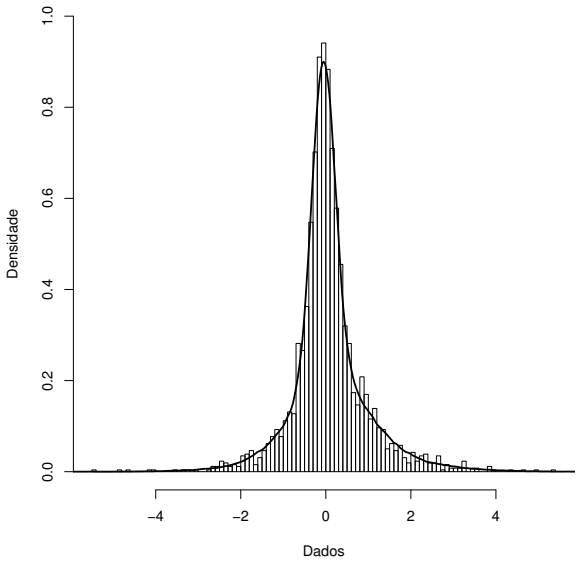


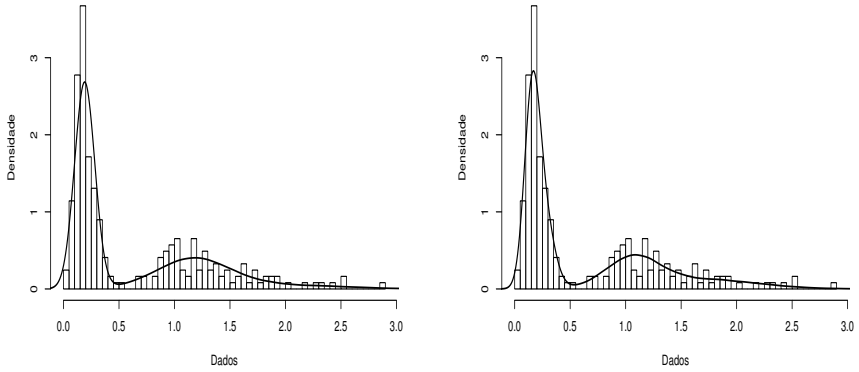
Figura 4 - Histograma dos dados observados e densidades dos modelos estimados.

Tabela 7 - Valores de AIC, $AICc$ e BIC

Critério	Número de componentes.					
	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$
AIC	119.5514	114.3209	**105.0733**	112.8112	114.8654	120.5753
$AICc$	119.8025	114.9311	**106.2064**	114.6372	117.4617	124.3253
BIC	147.0577	**142.3310**	143.5872	161.8288	174.3870	190.6005

Tabela 8 - Estimativas dos parâmetros para $k=3$ e $k=4$

Parâmetro	Estimativa		Parâmetro	Estimativa		Parâmetro	Estimativa	
	$k=3$	$k=4$		$k=3$	$k=4$		$k=3$	$k=4$
μ_1	0,1906	0,1562	σ_1	0,0063	0,0029	w_1	0,6122	0,3959
μ_2	1,1691	0,2565	σ_2	0,1098	0,0068	w_2	0,3265	0,2204
μ_3	2,0223	1,0654	σ_3	0,2479	0,0575	w_3	0,0613	0,2327
μ_4	-	1,6953	σ_4	-	0,2336	w_4	-	0,1510



(a) $k = 3$. (b) $k = 4$.

Figura 5 - Histograma dos dados observados e densidades dos modelos estimados.

5 Considerações finais

Neste artigo, apresentamos uma revisão sobre modelos de mistura e o procedimento de estimação via máxima verossimilhança. Desenvolvemos um estudo de simulação para verificar a performance dos critérios de seleção de modelos para determinação de k. Comparamos os critérios AIC, AICc e BIC utilizando a proporção de acerto de cada critério para dados simulados de diferentes modelos de distribuições normais com diferentes tamanhos amostrais.

Os resultados mostram que estes critérios devem ser utilizados com certa cautela. Para situações com componentes próximas os resultados não são satisfatórios. Por exemplo, para os conjuntos de dados A_1 e A_4, temos que as proporções sob o verdadeiro modelo é inferior a 40%. Mesmo para os casos com médias bem separadas, $\Delta > 2$, os três critérios apresentam proporção de acerto inferior a 70%. O BIC apresentou proporção sob o verdadeiro modelo maior do que 50% somente nos casos com componentes bem separadas ($\Delta > 2$) e tamanhos amostrais $n = \{500, 1.000\}$. Já o AIC e o $AICc$ em nenhum dos casos simulados apresentou proporção sob o verdadeiro modelo superior a 50%.

No Apêndice A do material suplementar apresentamos o estudo de simulação para o caso normal com variâncias diferentes de 1. Nos Apêndices B, C e D apresentamos o estudo de simulação considerando um modelo de mistura de distribuições Poisson e Gama. Similar aos resultados apresentados para o caso normal com variância igual a 1 os resultados obtidos não foram satisfatórios. Os três critérios apresentaram baixa proporção de escolha do verdadeiro modelo. Mesmo para caso com médias bem separadas a proporção de acerto ficou abaixo de 0.60.

Todas as simulações foram desenvolvidas utilizando o software R. Os códigos podem ser obtidos via solicitação por email aos autores. Como trabalho futuro

estamos desenvolvimento o procedimento Bayesiano de estimação de um modelo de mistura. Seguindo a mesma linha deste artigo, vamos verificar a performance dos critérios de seleção de modelos DIC e ICL (BIERNACKI *et al.*, 2003) para escolha do número de componentes e comparar com os resultados apresentados neste artigo.

Agradecimentos

Os autores agradecem o apoio finaceiro dado pelo Conselho Nacional de Desenvolvimento Científico e Tecnologico (CNPq) e aos revisores e editores pelas sugestões apresentadas.

ZUCARELI, L. R.; SARAIVA, E. F.; SUZUKI, A. K.; A Tutorial on estimation in mixture model. *Rev. Bras. Biom.,* Lavras, v.36, n.4, p.968-997, 2018.

- *ABSTRACT: In the last years, there has been a growing interest in modeling using mixture models due its flexibility. One of the main issues in the estimation procedure of its parameters is the determination of the number of components. In this paper, through a Monte Carlo simulation study, we compared the performance of the model selection criteria AIC, AICc and BIC for determining the number of components of a mixture model. The results obtained show a complementarity between the criteria and that these should be used with some care, since they present percentages of success less than 70% in most of the studied cases.*

- *KEYWORDS: Mixture model; likelihood function; EM algorithm; model selection; simulation.*

Referências

ALLENBY, G. M.; ARORA, N.; GINTER, J. L. On the heterogeneity of demand. *Journal of Marketing Research*, v.35, p.384-389, 1998.

BALDI P.; LONG, D. A. Bayesian framework for the analysis of microarray expression data: Regularized t-test and statistical inferences of gene changes. *Bioinformatics*, v.17, p.509-519, 2001.

BEHBOODIAN, A. N. On the Modes of a Mixture of Two Normal Distributions. *Technometrics*, v.12, p.131-139, 1970.

BIERNACKI, C.; CELEUX, G.; GOVAERT, G. Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Computational Statistics and Data Analysis*, v.41, p.561-575, 2003.

BOZDOGAN, H. Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrica*, v.52, p.345-370, 1987.

BROËT, P.; RICHARDSON, S.; RADVANYI, F. Bayesian hierarchical model for identifying changes in gene expression from microarray experiments. *Journal of Computational Biology*, v.9, p.671-683, 2008.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B*, v.39, p.1-22, 1977.

DIEBOLT, J.; ROBERT, C. Estimation of finite mixture distributions by Bayesian sampling. *Journal of the Royal Statistical society B*, v.56, p.363-375, 1994.

FINCH, S.; MENDELL, N.; THODE, H. Probabilistic measures of adequacy of a numerical search for a global maximum. *Journal of the American Statistical Association*, v.84, p.1020-1023, 1989.

FRÜHWIRTH-SCHNATTER, S. *Finite Mixture and Markov Switching Models*. London: Springer Series in Statistics, 2006.

GREEN, P. J.; RICHARDSON, S. Hidden Markov models and disease mapping. *Journal of the American Statistical Association*, v.97, p.1-16, 2002.

JEDIDI, K.; JAGPAL, H. S.; DESARBO, W. S. Finite-mixture structural equation models for response-based segmentation and unobserved heterogeneity. *Marketing Science*, v.16, p.39-59, 1997.

KARLIS, D.; XEKALAKI, W. D. Choosing initial values for the EM algorithm for Finite mixtures. *Computational Statistics & Data Analysis*, v.41, p.577-590, 2003.

LAMOREUX, C. G.; LASTSRAPES, W. D. Endogenous trading volume and momentum in stock return volatility. *Journal of Business and Economic Statistics*, v.12, p.253-260, 1994.

MCLACHLAN, G. J.; PEEL, D. *Finite Mixture Models*. Wiley Series in Probability and Statistics. New York: Wiley, 2000. 419p.

MACQUEEN, J. *Some methods for classification and analysis of multivariate observations*, Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, 281-297, University of California Press, Berkeley, Calif., 1967.

MENG, X. L. The EM algorithm and medical studies: A historical link. *Statistical Methods in Medical Research*, v.6, p.3-23, 1997.

REDNER, R. A.; WALKER, H. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, v.26, p.195-239, 1984.

RICHARDSON, S.; GREEN, P. J. On Bayesian analysis of mixture with unknown number of components. *Journal of the Royal Statistical Society*, v.59, p.731-792, 1997.

ROBERT, C. P.; RYDEN, T.; TITTERINGTON, D. M. Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method. *Journal of the Royal Statistical Society B*, v.62, p.57-75, 2000.

ROEDER, K.; WASSERMAN, L. Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *Journal of the American Statistical Association*, v.85, p.617-624, 1990.

SARAIVA, E. F.; MILAN, L. A. Clustering Gene Expression Data using a Posterior Split-Merge-Birth Procedure. *Scandinavian Journal of Statistics*, v.39, p.399-415, 2012.

SARAIVA, E. F.; LOUZADA, F.; MILAN, L. A. Mixture models with an unknown number of components via a new posterior split-merge MCMC algorithm. *Applied Mathematics and Computation*, v.244, p.959-975, 2014.

SARAIVA, E. F.; MILAN, L. A. Partitioning gene expression data by data-driven Markov chain Monte Carlo. *Journal of Applied Statistics*, v.43, p.1155-1173, 2016.

STEPHENS, M. Bayesian Analysis of mixture models with an unknown number of components - an alternative to reversible jump method. *The Annals of Statistics*, v.34, p.187–220, 2000a.

STEPHENS, M. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society B*, v.62, p.795–809, 2000b.

ZHOU, Y,; ZHU, H. Image Segmentation Using a Trimmed Likelihood Estimator in the Asymmetric Mixture Model Based on Generalized Gamma and Gaussian Distributions. *Mathematical Problems in Engineering*, Article ID 3468967, 17 pages.

Recebido em 08.08.2017.

Aprovado após revisão em 04.07.2018.

Apêndice A: Estudo de simulação - Caso normal

Neste apêndice apresentamos os resultado do estudo de simulação considerando o modelo de mistura de distribuições normais com variâncias diferentes de 1. Consideramos três conjuntos de dados simulados, denotados por B_1, B_2 e B_3. Os dados foram gerados de acordo com os valores dos parâmetros descritos na Tabela 9. Fixamos os tamanhos amostrais em $n = 100, 200, 500, 1.000$.

Para o conjunto de dados B_1, temos

$$\Delta_1 = (\mu_2 - \mu_1) = 6 > 2min(\sigma_1, \sigma_2) = 2 \text{ e } \Delta_2 = (\mu_3 - \mu_2) = 7 > 2min(\sigma_2, \sigma_3) = 4,$$

i.e., as componentes são duas a duas bimodais. De forma similar, também temos que as componentes para os conjuntos de dados B_2 e B_3 também são duas a duas bimodais. Ou seja, temos três conjuntos de dados com componentes bem separadas.

Cada conjunto de dados foi replicado $M = 1.000$ vezes. Para cada um dos M conjuntos de dados calculamos o valor AIC, $AICc$ e BIC. A Tabela 10 mostra a propoção de vezes que um modelo com k componentes foi selecionado como sendo o melhor modelo dentre os considerados, para B_1, B_2 e B_3.

Para os conjuntos de dados B_1 e B_2 somente o BIC apresentou proporção de acerto maior do que 0.5; porém esta proporção de acerto é ≤ 0.65. Para o conjunto de dados B_3 a proporção de acerto dos três critérios é inferior a 0.5. Para este caso, o BIC apresentou melhor desempenho, *i.e.*, maior proporção de acerto em relação ao AIC e AICc. Ou seja, como discutido no artigo, os três critérios devem ser utilizados com certa cautela para seleção do número de componentes de um modelo de mistura, devido a baixa taxa de acerto.

Tabela 9 - Número de componentes e valores dos parâmetros usados para gerar os dados simulados

Dados simulados	Número de componentes	Valores dos Parâmetros
B_1	$k_{real} = 3$	$\mu_1 = -6, \quad \mu_2 = 0, \quad \mu_3 = 7$ $\sigma_1 = 1, \quad \sigma_2 = 2, \quad \sigma_3 = 3$ $w_1 = 0.20, \quad w_2 = 0.30, \quad w_3 = 0.50$
B_2	$k_{real} = 3$	$\mu_1 = -7, \quad \mu_2 = 0, \quad \mu_3 = 8$ $\sigma_1 = 2, \quad \sigma_2 = 2, \quad \sigma_3 = 2$ $w_1 = 0.30, \quad w_2 = 0.40, \quad w_3 = 0.30$
B_3	$k_{real} = 4$	$\mu_1 = -6, \quad \mu_2 = 0, \quad \mu_3 = 7, \quad \mu_4 = 14$ $\sigma_1 = 1, \quad \sigma_2 = 2, \quad \sigma_3 = 2, \quad \sigma_4 = 1$ $w_1 = 0.10, \quad w_2 = 0.40, \quad w_3 = 0.40, \quad w_4 = 0.10$

Tabela 10 - Proporção de valores de k selecionados para B_1, B_2 e B_3

Dados	k_{real}	k	Tamanho da amostra											
			100			200			500			1.000		
			AIC	*AICc*	*BIC*	*AIC*	*AICc*	*BIC*	*AIC*	*AICc*	*BIC*	*AIC*	*AICc*	*BIC*
B_1	3	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		2	0.019	0.030	0.112	0.000	0.002	0.015	0.000	0.000	0.000	0.000	0.000	0.000
		3	0.323	0.381	0.553	0.338	0.364	0.603	0.376	0.384	0.598	0.355	0.356	0.535
		4	0.277	0.287	0.242	0.315	0.324	0.274	0.244	0.246	0.248	0.249	0.253	0.269
		5	0.214	0.192	0.076	0.195	0.186	0.083	0.215	0.213	0.122	0.225	0.222	0.142
		6	0.167	0.110	0.017	0.152	0.124	0.025	0.165	0.157	0.032	0.171	0.169	0.054
B_2	3	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		2	0.000	0.000	0.009	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		3	0.367	0.431	0.650	0.358	0.376	0.607	0.350	0.356	0.577	0.343	0.345	0.518
		4	0.265	0.290	0.236	0.284	0.292	0.257	0.272	0.273	0.263	0.260	0.260	0.267
		5	0.215	0.177	0.082	0.196	0.194	0.110	0.204	0.204	0.108	0.222	0.221	0.141
		6	0.153	0.102	0.023	0.162	0.138	0.026	0.174	0.167	0.052	0.175	0.174	0.074
B_3	4	1	0.001	0.001	0.043	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		2	0.002	0.002	0.025	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		3	0.047	0.071	0.188	0.007	0.009	0.057	0.001	0.001	0.007	0.000	0.000	0.000
		4	0.247	0.323	0.380	0.257	0.280	0.473	0.260	0.267	0.484	0.282	0.284	0.470
		5	0.240	0.264	0.217	0.244	0.263	0.294	0.226	0.234	0.274	0.237	0.238	0.283
		6	0.198	0.167	0.101	0.207	0.203	0.126	0.195	0.197	0.141	0.207	0.207	0.137
		7	0.151	0.108	0.032	0.150	0.135	0.040	0.175	0.170	0.074	0.140	0.141	0.073
		8	0.114	0.064	0.014	0.135	0.110	0.010	0.143	0.131	0.020	0.134	0.130	0.037

Apêndice B: Modelo de mistura de distribuições Poisson

Considere o modelo com mistura da equação (2) do artigo com $f(y_i|\theta_j)$ sendo a densidade da distribuição Poisson com média θ_j, *i.e.*,

$$f(y_i|\theta_j) = \frac{\theta_j^{y_i} e^{-\theta_j}}{y_i!}$$

para $i = 1, \ldots, n$ e $j = 1, \ldots, k$.

A função log-verossimilhança para uma componente j é dada por

$$l(\theta_j|\mathbf{y}) \propto log(\theta_j) \sum_{i;c_i=j} y_i - n_j\theta_j,$$

para $j = 1, \ldots, k$.

Derivando $l(\theta_j|\mathbf{y})$, com relação a θ_j e igualando o resultado zero, obtemos como estimador

$$\hat{\theta}_j = \bar{y}_j = \frac{1}{n_j} \sum_{\{y_i;c_i=j\}} y_i,$$

i.e., a média das observações alocadas na componente j, para $j = 1, \ldots, k$.

Algoritmo SEM para mistura de distribuições Poisson. Inicie o algoritmo com um configuração $\mathbf{c}^{(0)} = \left(c_1^{(0)}, \ldots, c_n^{(0)}\right)$. Para s-ésima iteração do algoritmo, $s = 1, \ldots$:

(i) obtenha as estimativas de máxima verossimilhança $\hat{\theta}_j^{(s)}$ e $\hat{w}_j^{(s)}$ condicional em $\mathbf{c}^{(s-1)}$;

(ii) se $\left|\frac{l\left(\hat{\boldsymbol{\theta}}^{(s)}|\mathbf{y}\right) - l\left(\hat{\boldsymbol{\theta}}^{(s-1)}|\mathbf{y}\right)}{l\left(\hat{\boldsymbol{\theta}}^{(s-1)}|\mathbf{y}\right)}\right| < \epsilon$, pare o algoritmo.

(iii) caso contrário, gere $\mathbf{z}_i \sim Multinomial(1, 1, \mathbf{w}_i^*)$, em que $\mathbf{w}_i^* = (w_{i1}^*, \ldots, w_{ik}^*)$ e

$$w_{ij}^* \propto \hat{w}_j^{(s)} \hat{\theta}_j^{y_i} exp\left\{-\hat{\theta}_j\right\}$$

para $i = 1, \ldots, n$ e $j = 1, \ldots, k$. Se $z_{ij} = 1$, faça $c_i^{(s)} = j$, para $i = 1, \ldots, n$ e $j = 1, \ldots, k$. Faça $s = s + 1$ e retorne ao passo (i).

Apêndice B_1: Estudo de simulação - caso Poisson

Similar ao estudo de simulação para o caso normal apresentado no Apêndice A, desenvolvemos um estudo de simulação para verificar a performance dos critérios de seleção de modelos na determinação do número de componentes de um modelo de mistura de distribuições Poisson.

Para este caso, consideramos dois conjunto de dados, B_1 e B_2, com duas e três componentes, respectivamente. Os dado foram gerados de acordo com os valores decritos na Tabela 11.

Tabela 11 - Número de componentes e valores dos parâmetros usados para gerar os dados simulados

Dados	Número de componentes	Valores dos parâmetros
B_1	$k_{real} = 2$	$\theta_1 = 2, \quad \theta_2 = 5$ $w_1 = 0.40, \quad w_2 = 0.60$
B_2	$k_{real} = 3$	$\theta_1 = 2, \quad \theta_2 = 6, \quad \theta_3 = 9$ $w_1 = 0.50, \quad w_2 = 0.30, \quad w_3 = 0.20$

A Tabela 12 mostra a propoção de vezes que um modelo com k componentes foi selecionado como sendo o melhor modelo dentre os considerados, para B_1 e B_2.

Para os dois caso simulados, os critérios apresentaram proporção de acerto inferior a 0.5; exceto o BIC para o B_1 com $n = 200$. Para B_2 as proporções de acerto foram inferiores a 0.3. Ou seja, como para o caso normal, os critérios AIC, $AICc$ e BIC devem ser utilizados com muita cautela para determinação do número de componentes em um modelo de mistura Poisson. Neste caso, os resultados indicam que os critérios não deveriam ser utilizados para determinação do número de componentes devido sua baixa taxa de acerto, especialmente para o caso B_2.

Tabela 12 - Proporção de valores de k selecionados para B_1 e B_2

Dados	k_{real}	k	Tamanho da amostra											
			100			200			500			1.000		
			AIC	*AICc*	*BIC*	*AIC*	*AICc*	*BIC*	*AIC*	*AICc*	*BIC*	*AIC*	*AICc*	*BIC*
B_1	2	1	0.196	0.203	0.302	0.106	0.109	0.128	0.040	0.040	0.069	0.010	0.010	0.016
		2	0.352	0.369	0.468	0.322	0.330	0.643	0.292	0.293	0.443	0.309	0.310	0.459
		3	0.218	0.219	0.165	0.266	0.265	0.159	0.245	0.245	0.247	0.254	0.254	0.264
		4	0.149	0.140	0.050	0.177	0.172	0.057	0.235	0.235	0.162	0.226	0.225	0.169
		5	0.085	0.069	0.015	0.129	0.124	0.013	0.188	0.187	0.079	0.201	0.201	0.092
B_2	3	1	0.004	0.004	0.005	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		2	0.381	0.398	0.544	0.310	0.319	0.462	0.280	0.281	0.535	0.232	0.234	0.331
		3	0.264	0.268	0.262	0.278	0.278	0.288	0.225	0.225	0.180	0.227	0.227	0.264
		4	0.163	0.158	0.092	0.178	0.178	0.115	0.203	0.203	0.109	0.218	0.218	0.171
		5	0.112	0.097	0.026	0.141	0.133	0.048	0.157	0.156	0.053	0.188	0.186	0.109

Apêndice C: Modelo de mistura de distribuições Gama

Considere o modelo com mistura em (2) do artigo com $f(y_i|\theta_j)$ sendo a densidade da distribuição Gama de parâmetros α_j e β_j, *i.e.*,

$$f(y_i|\theta_j) = \frac{\beta_j^{\alpha_j}}{\Gamma(\alpha_j)} y_i^{\alpha_j - 1} e^{-\beta_j y_i}$$

onde $\theta_j = (\alpha_j, \beta_j)$, para $i = 1, \ldots, n$ e $j = 1, \ldots, k$.

A função log-verossimilhança para uma componente j é dada por

$$l(\theta_j|\mathbf{y}) = n_j \alpha_j log(\beta_j) - n_j log\left(\Gamma(\alpha_j)\right) + (\alpha_j - 1) \sum_{\{i;c_i=j\}} log(y_i) - \beta_j \sum_{i;c_i=j} y_i$$

para $j = 1, \ldots, k$.

Derivando $l(\theta_j|\mathbf{y})$ com relação a α_j e β_j obtemos

$$\frac{dl(\theta_j|\mathbf{y})}{d\alpha_j} = n_j log(\beta_j) - nD(\alpha_j) + \sum_{\{i;c_i=j\}} log(y_i) \quad (11)$$

$$\frac{dl(\theta_j|\mathbf{y})}{d\beta_j} = \frac{n_j \alpha_j}{\beta_j} + \sum_{\{i;c_i=j\}} y_i. \quad (12)$$

onde $D(\alpha) = \frac{d}{d\alpha_j} log\left(\Gamma(\alpha_j)\right)$.

Igualando (11) e (12) a zero, obtemos os estimadores de máxima verossimilhança para α_j e β_j. Porém, não conseguiremos obter os estimadores de forma analítica "fechada". Os estimadores de máxima verossimilhança são dados pela solução das seguintes equações

$$\beta_j = \frac{\overline{y}_j}{\alpha_j} \quad (13)$$

$$nD(\alpha) - n_j log(\beta_j) = \sum_{i;ci=j} log(y_i), \quad (14)$$

onde $\overline{y}_j$ é a média das observações alocadas na componente j, para $j = 1, \ldots, k$.

Dessa forma, devemos resolver as equações em (13) e (14) numericamente. Neste ponto, poderíamos utilizar, por exemplo, o método de Newton-Raphson. Porém, optamos por obter os estimadores de máxima verossimilhança para α_j e β_j usando o software *R* e o comando *optim*.

As probabilidades de alocação são dadas por

$$w_{ij}^* \propto \hat{w}_j^{(l)} y_i^{\hat{\alpha}_j - 1} exp\left\{-\hat{\beta}_j y_i\right\},$$

para $i = 1, \ldots, n$ e $j = 1, \ldots, k$.

O algoritmo SEM para mistura de distribuições gama é similar ao descrito para o modelo de mistura de distribuições normais e Poisson.

Apêndice C_1: Estudo de simulação - caso Gama

Para este caso, consideramos dois conjunto de dados, B_1 e B_2, com duas e três componentes, respectivamente. Os dado foram gerados de acordo com os valores decritos na Tabela 13.

A Tabela 14 mostra a proporção de vezes que um modelo com k componentes foi selecionado como sendo o melhor modelo dentre os considerados, para B_1 e B_2.

Tabela 13 - Número de componentes e valores para os parâmetros usados para geração dos dados simulados

Dados	Número de componentes	Valores dos parâmetros
B_1	$k_{real} = 2$	$\alpha_1 = 2, \quad \alpha_2 = 16$ $\beta_1 = 2, \quad \beta_2 = 4$ $w_1 = 0.50, \quad w_2 = 0.50$
B_2	$k_{real} = 3$	$\alpha_1 = 2, \quad \alpha_2 = 15, \quad \alpha_3 = 44$ $\beta_1 = 2, \quad \beta_2 = 3, \quad \beta_3 = 4$ $w_1 = 0.40, \quad w_2 = 0.20, \quad w_3 = 0.40$

Para os dois casos simulados, os critérios AIC e $AICc$ apresentaram proporção de acerto inferior a 0.4. Já o BIC, apresenta proporção de acerto maior do que 0.5 para B_1 com $n \in \{200, 500, 1.000\}$; para os outros casos simulados a proporção de acerto é inferior a 0.5. Ou seja, como para o caso normal e Poisson, os critérios AIC, $AICc$ e BIC devem ser utilizados com muita cautela para determinação do número de componentes em um modelo de mistura Gama. Em geral, os resultados mostram uma baixa taxa de acertação dos critérios na seleção do correto número de componentes.

Tabela 14 - Proporção de valores de k selecionados para B_1 e B_2

Dados	k_{real}	k	Tamanho da amostra											
			100			200			500			1.000		
			AIC	*AICc*	*BIC*	*AIC*	*AICc*	*BIC*	*AIC*	*AICc*	*BIC*	*AIC*	*AICc*	*BIC*
B_1	2	1	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		2	0.176	0.217	0.439	0.339	0.360	0.610	0.343	0.345	0.589	0.331	0.332	0.559
		3	0.196	0.207	0.232	0.247	0.251	0.244	0.270	0.271	0.266	0.272	0.273	0.268
		4	0.286	0.278	0.170	0.210	0.205	0.107	0.220	0.218	0.106	0.222	0.221	0.123
		5	0.342	0.298	0.158	0.204	0.184	0.039	0.167	0.166	0.039	0.175	0.174	0.050
B_2	3	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		2	0.064	0.087	0.222	0.047	0.050	0.169	0.017	0.017	0.058	0.001	0.001	0.008
		3	0.170	0.197	0.252	0.297	0.301	0.359	0.307	0.306	0.402	0.299	0.299	0.396
		4	0.201	0.205	0.149	0.234	0.235	0.197	0.293	0.295	0.260	0.274	0.274	0.268
		5	0.258	0.222	0.120	0.205	0.201	0.086	0.204	0.204	0.120	0.239	0.239	0.158
		6	0.307	0.289	0.257	0.217	0.213	0.189	0.179	0.178	0.160	0.187	0.187	0.170