Departamento de Estatística - Universidade de Brasília

Estimação de função discriminante para mistura de duas distribuições Kumaraswamy

Carlos Eduardo Linhares Levicoy

Brasília
2016

Carlos Eduardo Linhares Levicoy

Estimação de função discriminante para mistura de duas distribuições Kumaraswamy

Projeto apresentado para obtenção do título de Bacharel em Estatística ao Departamento de Estatística da Universidade de Brasília

Orientador: Profa. Dra. Cira Etheowalda Guevara Otiniano

Brasília
2016

Resumo

Nos modelos de mistura de distribuições é importante discutir de qual componente da mistura cada observação provém. Neste trabalho, encontramos as estimativas de máxima verossimilhança da mistura de duas distribuições Kumaraswamy quando temos observações classificadas e não classificadas. Posteriormente, foi calculada a função de discriminante não linear do modelo para avaliar a qual componente um conjunto de elementos pertence. Além disso, foi apresentada a taxa de erros associada a essa função. Para avaliar o desempenho do discriminante fizemos análises através de dados simulados. Por último, a análise de discriminante de um conjunto de dados reais foi adicionado.

Sumário

1 Introdução

A mistura de densidades tem como função modelar os dados de uma população que é constituída ou por uma densidade assimétrica ou por uma distribuição heterogênea de subpopulações. Campos em que modelos de mistura foram aplicados com sucesso incluem astronomia, biologia, genética, medicina, psiquiatria, economia, engenharia, entre outros (McLachlan, 2000). A mistura de k populações é a combinação convexa de k densidades de uma certa família, sendo cada densidade denominada de componente da mistura e os coeficientes chamados de pesos de mistura. O vetor dos parâmetros do modelo de mistura tem no minimo 2k-1 parâmetros a serem estimados, em que k-1 são referentes aos pesos e k as componentes. Esses parâmetros podem ser estimados por máxima verossimilhança, algoritmo EM ou por inferência bayesiana.

Em modelos de mistura um problema interessante é sobre o estudo do discriminante, pois dada uma observação podemos classificar de qual subpopulação ela provém. O estudo do discriminante para pequenas amostras da mistura de duas componentes da mesma família tem sido bastante tratada por diversos autores. Dentre os trabalhos mais recentes podemos citar os de Sultan (2011), Ahmad (2010), Ahmad e Elrahman (1994), Mahmoud e Moustafa (1993) e Amoh (1991) que trataram o discriminante para componentes Weibull inversa, Gumbel, Weibull, Gamma e Normal inversa, respectivamente. Sendo assim, no trabalho a seguir será realizado o estudo do discriminante da mistura de duas distribuições Kumaraswamy.

Neste trabalho, calculamos as estimativas dos parâmetros da mistura de duas distribuições Kumaraswamy e sua função do discriminante não linear, com finalidade de classificar as observações em dois grupos distintos. Porém, como a coleta de dados tem influência no cálculo do discriminante teremos diferentes tipos de classificador. Quando temos toda população de um conjunto de dados, a função de discriminante será obtida utilizando os verdadeiros parâmetros de algum modelo, com isso teremos a função de discriminante chamada ótima. Como normalmente não se tem acesso a toda uma população, é preciso retirar amostras para estimar os parâmetros do modelo. Se no processo de amostragem não é conhecido de qual subpopulação provém cada observação (amostra não classificada), então os parâmetros da mistura podem ser estimados via Algoritmo EM e a função de discriminante encontrada será baseada nessa estimativa. Se conhecemos de onde cada observação foi gerada (amostra classificada), então as estimativas dos parâmetros podem ser obtidas por outro método, com isso teríamos outra função de discriminante.

Como forma de avaliar a qualidade das estimativas dos parâmetros e os classificadores serão feitas simulações para diferentes tamanho de amostra. A performance dos classificadores serão avaliadas de acordo com as proporções de observações mal classificadas que cada um produziu. Os cálculos feitos ao decorrer do trabalho serão realizados através do

software R.

O trabalho está organizado em três capítulos. No Capítulo 2, apresentamos o modelo de mistura e em quais situações ele pode ser aplicado. Em seguida, descrevemos o conceito geral de discriminante não linear que utilizamos neste trabalho para classificar observações em diferentes componentes (clusters). Ainda neste capítulo, calculamos as estimativas dos parâmetros da mistura de duas distribuições Kumaraswamy para observações classificadas e não classificadas, em seguida calculamos a função discriminante da mistura bem como os erros de classificação. No Capítulo 3, são mostrados os resultados das estimativas dos parâmetros de nosso modelo para amostras classificadas e não classificadas. Além disso, calculamos as probabilidades de cometer um erro de classificação qualquer e em relação a cada componente. Para ilustrar a aplicabilidade de nosso modelo com a análise de discriminante, usamos um banco de dados reais. No Capítulo 4, apresentamos as conclusões do trabalho.

2 Metodologia

Neste capítulo estudamos a função discriminante da mistura de duas densidades Kumaraswamys. Na seção 2.1 é explicado o que é uma mistura e situações que utilizamos este tipo de modelo. Na seção 2.2 é explicado como funciona uma análise de discriminante em casos de mistura. Na seção 2.3 é descrita a mistura de duas Kumaraswamys. Como na prática geralmente os parâmetros desse modelo são desconhecidos, eles devem ser estimados por algum método. Sendo assim, na seção 2.4 descrevemos os estimadores de máxima verossimilhança dos parâmetros da mistura, que serão utilizados na seção 2.5 para obter os estimadores do algoritmo EM. Na seção 2.6 são obtidas as estimativas dos parâmetros para amostra classificada. Na seção 2.7 e 2.8 apresentamos a função discriminante para mistura de duas Kumaraswamys. Finalmente, na seção 2.9 serão calculados os erros de classificação obtidos através das observações mal classificadas pela função de discriminante.

2.1 Mistura de Distribuições

Um modelo de mistura é uma combinação linear convexa de funções densidades de probabilidade. Formalmente, isso seria dizer que se f é uma mistura então sua densidade é da forma

$$f(x;\theta) = \sum_{k=1}^{K} p_k f_k(x;\theta_k), \qquad \text{em que} \quad \sum_{k=1}^{K} p_k = 1.$$

Por propriedade da convexidade, cada função $f_k(x/\theta_k)$ também representa uma função densidade, chamada de componente da mistura. Pelo modelo de mistura ser escrito em função destas densidades ele é bastante flexível, o que faz com que seja explorado em diversas áreas.

Para ilustrar uma aplicação de mistura, obtivemos dados referentes aos intervalos de tempo em minutos que o geyser Old Faithful (Yellowstone National Park, Wyoming, USA) demora para entrar em erupção. Estes dados são do ano de 1991 e podem ser obtidos através do pacote mixtools do *software* R. Pela Figura 1 é possível notar que o histograma dos dados possui uma forma bimodal. Sendo assim, é possível imaginar que as subpopulações dos dados poderiam ser modeladas por diferentes densidades. Porém, isso seria inviável porque não se sabe a qual subpopulação cada dado pertence, logo uma solução seria modelar os dados por uma mistura. Para esse exemplo foi ajustado uma mistura de duas densidades Normais, representada pela curva pontilhada. As curvas em vermelho e verde são dadas por $p_1 f_1(x;\theta_1)$ e $p_2 f_2(x;\theta_2)$, respectivamente.

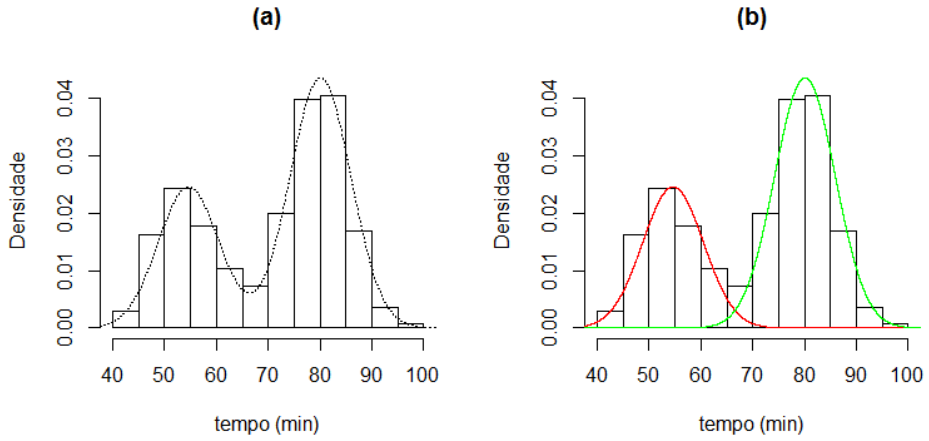

Figura 1 – (a) - Mistura de duas densidades Normais para os dados de intervalo de tempo entre as erupções do geyser Old Faithful, (b) - Componentes $p_1 f_1(x;\theta_1)$ (curva vermelha) e $p_2 f_2(x;\theta_2)$ (curva verde) da mistura de (a).

De acordo com a Figura 1, é possível notar que misturas de distribuições são úteis para modelar densidades bimodais. Além disso, modelos de mistura também podem ser utilizados para modelar densidades multimodais até distribuições assimétricas. Para mostrar esta flexibilidade serão apresentados exemplos de misturas de Normais retiradas do livro McLachlan (2000), página 12. Foram pegos quatro exemplos de misturas de Normais, em que seus gráficos representam uma densidade bimodal, trimodal, multimodal e assimétrica, respectivamente.

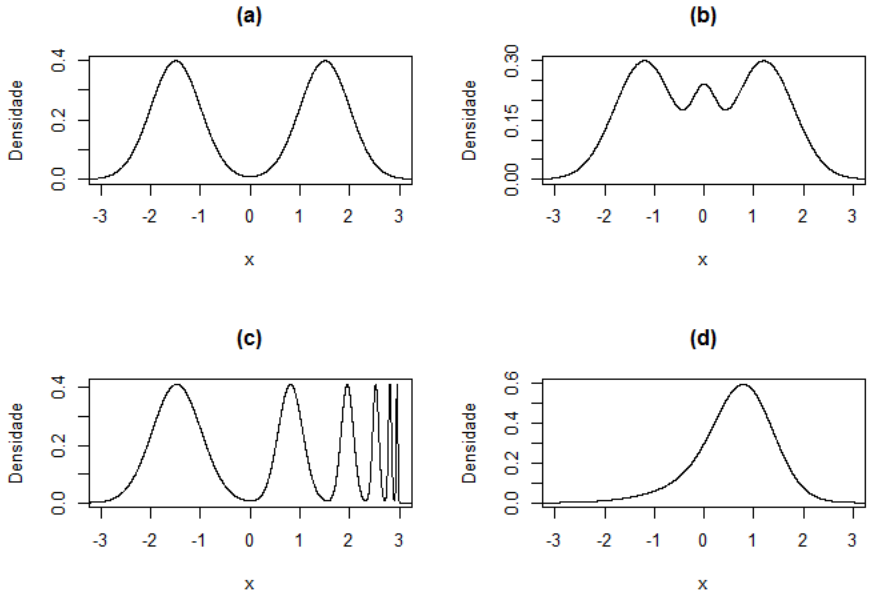

Figura 2 – (a) - Mistura de duas Normais com curva bimodal, (b) - Mistura de três Normais com curva trimodal, (c) - Mistura de cinco Normais com curva multimodal e (d) - Mistura de três Normais com curva unimodal assimétrica.

A Tabela 1 apresenta as misturas de densidades Normais utilizadas para geração da Figura 2.

Tabela 1 – Funções densidades das Misturas de Normais da Figura 2.

Densidade	f(x)
(a) - Mistura bimodal	$\frac{1}{2}N(-\frac{3}{2},(\frac{1}{2})^2)+\frac{1}{2}N(\frac{3}{2},(\frac{1}{2})^2)$
(b) - Mistura trimodal	$\frac{9}{20}N(-\frac{6}{5},(\frac{3}{5})^2)+\frac{9}{20}N(\frac{6}{5},(\frac{1}{2})^2)+\frac{1}{10}N(0,(\frac{1}{4})^2)$
(c) - Mistura multimodal	$\sum_{i=0}^{5}(2^{5-i}/63)N((65-96(\frac{1}{2})^i/21,(\frac{32}{63})^2/2^{2i}))$
(d) - Mistura unimodal	$\frac{1}{5}N(0,1)+\frac{1}{5}N(\frac{1}{2},(\frac{2}{3})^2)+\frac{13}{15}N(0,(\frac{5}{9})^2)$

2.2 Análise de Discriminante

Segundo KHATTREE & NAIK (2000) a análise de discriminante é uma técnica estatística multivariada que estuda a separação de objetos de uma população em duas ou mais classes. Atualmente existem vários métodos de classificação como LDA (Linear discriminant analysis), QDA (Quadratic discriminant analysis), classificação logística, k-means, entre outros. Geralmente o principal objetivo destas técnicas é classificar novos objetos que desconhecemos a origem em suas verdadeiras classes. Para isso, é criada alguma função discriminante baseada em dados passados, que de acordo com alguma regra de decisão classificará o objeto em algum cluster. Se esta função depender linearmente das observações então temos um discriminante linear, caso contrário ele é não linear.

Como queremos prever com uma alta precisão o verdadeiro cluster de um novo objeto, é de suma importância criar classificadores que minimizem o número de observações mal classificadas. Se conhecemos a priori o grupo que um elemento pertence, então saberemos se ele foi classificado corretamente. Observando todos elementos mal classificados é possível verificar a proporção das observações mal classificadas. Dessa maneira, a ideia seria criar um discriminante que diminuísse esta proporção sem causar um overfitting (super ajuste) em relação aos dados, para que ele também seja bom para discriminar novos elementos.

Neste trabalho serão analisados discriminantes para amostras classificadas e não classificadas de misturas de duas distribuições Kumaraswamys. Usualmente, o ideal seria que ao realizar a coleta dos dados, fossem pegas informações de que cluster foi gerada cada observação. Porém, em casos de mistura muitas vezes estes clusters não são definidos previamente, omitindo a variável do grupo que cada elemento provém. Com isso, em vez de realizar um novo processo de amostragem, uma possibilidade seria analisar os dados não classificados. Por exemplo, imagine que um pesquisador decida medir dados referentes a alturas de girafas. Ao coletar os dados ele se depara com um gráfico bimodal pelo fato de não estratificar os dados por sexo. Dessa maneira, para descobrir se uma

girafa desconhecida é macho ou fêmea, ele teria a opção de criar um classificador com base nos dados não classificados que ele ja possui ou fazer uma nova amostragem com a informação do sexo de cada girafa para tentar criar um melhor classificador. Logo, se nessa situação o discriminante para amostras não classificadas produzir resultados tão bons como o para amostras classificadas, ele poderia economizar o trabalho de realizar uma nova amostragem.

A maioria dos métodos usados para criar uma função de discriminante leva em consideração a informação de qual classe cada uma das observações é associada, pois se temos um número alto de clusters ou se as observações são parecidas essa informação se torna de suma importância para se ter um bom classificador. Como o assunto abordado neste trabalho se refere a dois clusters, pode ser que a falta desta informação não prejudique tanto a qualidade do discriminante. O estudo de discriminante para misturas foi introduzido por O'Neil (1969), que obteve a função discriminante para mistura de densidades Normais para dados não classificados. O modelo utilizado baseava-se em uma mistura de duas densidades

$$f(x;\theta) = p_1 f_1(x;\theta_1) + p_2 f_2(x;\theta_2), \quad 0 < p_1 < 1, \;\; p_2 = 1 - p_1,$$

cujos gráficos obtidos são do tipo da normal bimodal visto na Figura 2 (a).

Temos que se π_1 é o grupo de elementos gerados pela densidade f_1 e π_2 o grupo de elementos gerados pela densidade f_2, uma observação x_j deve pertencer a π_1 ou π_2 com probabilidades $P(x_j \in \pi_1)$ ou $P(x_j \in \pi_2)$. Como essas probabilidades podem ser obtidas pela fórmula de bayes e elas assumem valores entre 0 e 1, consideramos a função logística para obter estas probabilidades,

$$Pr(x_j \in \pi_1) = \frac{p_1 f_1(x_j;\theta_1)}{p_1 f_1(x_j;\theta_1) + p_2 f_2(x_j;\theta_2)} = \frac{1}{(1 + exp(\alpha + \beta z))} \quad \text{e}$$
$$Pr(x_j \in \pi_2) = \frac{p_2 f_2(x_j;\theta_2)}{p_1 f_1(x_j;\theta_1) + p_2 f_2(x_j;\theta_2)} = \frac{exp(\alpha + \beta z)}{(1 + exp(\alpha + \beta z))},$$

com

$z = h(x_j;\theta_1,\theta_2), \;\; \alpha = \alpha(\theta_1,\theta_2) \;\; e \;\; \beta = \beta(\theta_1,\theta_2).$

Para classificar uma observação em π_1 teríamos que ter $Pr(x_j \in \pi_1) > Pr(x_j \in \pi_2)$, que ao simplificarmos resultaria em $\alpha + \beta z < 0$, como pode ser visto a seguir.

$$\begin{aligned} Pr(x_j \in \pi_1) &> Pr(x_j \in \pi_2) \\ \frac{1}{(1 + exp(\alpha + \beta z))} &> \frac{exp(\alpha + \beta z)}{(1 + exp(\alpha + \beta z))} \\ exp(\alpha + \beta z) &< 1 \\ \alpha + \beta z &< 0. \end{aligned}$$

Analogamente, para classificar uma observação em π_2 seria preciso que $\alpha + \beta z > 0$. Desse modo, a função de discriminante para mistura de duas densidades seria dado por $\alpha + \beta z$ porque podemos tomar decisões de classificação a partir dela. Se z é uma função não linear, então temos uma função de discriminante não linear que é dada por:

$$NLD(x_j) = \alpha + \beta z.$$

Deve-se notar que se para uma observação x_j temos o $NLD(x_j) = 0$, então $Pr(x_j \in \pi_1) = Pr(x_j \in \pi_2) = 1/2$, logo não teríamos evidências para classificá-la em qualquer uma das populações. Resumindo, caso $NLD(x_j) > 0$ classificaríamos a observação como pertencente a segunda população π_2, já para $NLD(x_j) < 0$ classificaríamos ela em π_1. Quando implementarmos computacionalmente o algoritmo de classificação, no caso em que $NLD(x_j) = 0$ classificamos essa observação como pertencente a componente π_1, somente para não deixar essa observação sem classificação. Como dificilmente $NLD(x_j) = 0$ não temos muitos problema em adotar essa estratégia.

2.3 Mistura de Kumaraswamys

A mistura de distribuições Beta é amplamente utilizada em análise de risco de crédito, bioinformática e genética. Uma das desvantagens desse modelo é devido a distribuição acumulada da Beta não possuir forma fechada simples, dificultando o cálculo dos quantis, função de sobrevivência e de risco. Uma distribuição alternativa para substitui-la é a Kumaraswamy. A distribuição Kumaraswamy possui propriedades semelhantes a Beta e tem função de distribuição acumulada com forma fechada simples, dentre outras vantagens em relação a Beta que podem ser vistas em Jones(2009).

Neste trabalho estudamos uma mistura com duas componentes Kumaraswamy. A função densidade de probabilidade (fdp) dessa mistura é dada por

$$f(x;\theta) = p_1 f_1(x;\theta_1) + p_2 f_2(x;\theta_2), \quad 0 < p_1 < 1, \;\; p_2 = 1 - p_1, \tag{2.1}$$

em que $\theta = (p, a_1, a_2, b_1, b_2)$, $\theta_i = (a_i, b_i)$, i=1,2, e as densidades $f_i(x;\theta_i)$ associadas as componentes são

$$f_i(x;\theta_i) = a_i b_i x^{a_i - 1}(1 - x^{a_i})^{b_i - 1}, \quad 0 < x < 1, \quad a_i, b_i > 0, \quad i = 1, 2. \tag{2.2}$$

A função de distribuição acumulada da mistura de Kumaraswamys é dada por

$$F(x;\theta) = p_1 F_1(x;\theta_1) + p_2 F_2(x;\theta_2), \quad 0 < p_1 < 1, \;\; p_2 = 1 - p_1, \tag{2.3}$$

sendo

$$F_i(x;\theta_i) = 1 - (1 - x^{a_i})^{b_i}, \quad 0 < x < 1, \quad a_i, b_i > 0, \quad i = 1, 2. \tag{2.4}$$

2.4 Estimador de Máxima Verossimilhança

Supondo que temos uma amostra $x_1, x_2, \ldots, x_n$ com as observações independentes e identicamente distribuídas, em que cada x_i segue a função densidade dada por 2.1 com $a_1 = a_2$. Então a verossimilhança de $x_1, x_2, \ldots, x_n$ é dada por

$$L(\boldsymbol{\theta}) = \prod_{j=1}^{n} \left[p_1 \left(ab_1 x_j^{a-1} (1 - x_j^a)^{b_1 - 1} \right) + p_2 \left(ab_2 x_j^{a-1} (1 - x_j^a)^{b_2 - 1} \right) \right], \quad 0 < p_1 < 1, \;\; p_2 = 1 - p_1.$$

Por conveniência escreveremos a verossimilhança como:

$$L(\boldsymbol{\theta}) = a^n \prod_{j=1}^{n} x_j^{a-1} Q_j, \tag{2.5}$$

sendo

$$Q_j = p_1 g_1(x_j; \theta_1) + p_2 g_2(x_j; \theta_2) \tag{2.6}$$

e

$$g_i(x_j; \theta_i) = b_i (1 - x_j^a)^{b_i - 1}, \;\; i = 1, 2. \tag{2.7}$$

Para simplificar os cálculos foi aplicado o logaritmo na verossimilhança. Tal procedimento pode ser efetuado porque o logaritmo é uma função monótona crescente, com isso o argumento que maximizará a função será o mesmo.

$$Log(L(\boldsymbol{\theta})) = n \; log \; a + (a - 1) \sum_{j=1}^{n} log \; x_j + \sum_{j=1}^{n} log \; Q_j. \tag{2.8}$$

Em seguida foi calculado o gradiente em relação ao vetor de parâmetros $\theta = (p_1, a, b_1, b_2)$ e cada derivada parcial foi igualada a zero.

$$\frac{\partial Log(L(\boldsymbol{\theta}))}{\partial p_1} = \sum_{j=1}^{n} \frac{g_1(x_j; \theta_1) - g_2(x_j; \theta_2)}{Q_j} = 0, \tag{2.9}$$

$$\begin{aligned} \frac{\partial Log(L(\boldsymbol{\theta}))}{\partial b_1} &= \sum_{j=1}^{n} \frac{p_1 (1 - x_j^a)^{b_1 - 1} + p_1 b_1 (1 - x_j^a)^{b_1 - 1} \; log(1 - x_j^a)}{Q_j} \\ &= \sum_{j=1}^{n} \frac{p_1 (1 - x_j^a)^{b_1 - 1} + p_1 g_1(x_j; \theta_1) \; log(1 - x_j^a)}{Q_j} \\ &= \sum_{j=1}^{n} \frac{p_1 g_1(x_j; \theta_1)(b_1^{-1} + log(1 - x_j^a))}{Q_j} \\ &= \sum_{j=1}^{n} \frac{b_1^{-1} p_1 g_1(x_j; \theta_1)(1 + b_1 \; log(1 - x_j^a))}{Q_j} = 0, \end{aligned} \tag{2.10}$$

$$\frac{\partial Log(L(\boldsymbol{\theta}))}{\partial b_2} = \sum_{j=1}^{n} \frac{b_2^{-1} p_2 g_2(x_j;\theta_2)(1 + b_2\ log(1 - x_j^a))}{Q_j} = 0, \tag{2.11}$$

$$\begin{aligned}\frac{\partial Log(L(\boldsymbol{\theta}))}{\partial a} &= \frac{n}{a} + \sum_{j=1}^{n} log(x_j) + \sum_{j=1}^{n}\sum_{i=1}^{2} \frac{p_i b_i (1 - x_j^a)^{b_i - 1}(b_i - 1)(-x_j^a\ log(x_j))(1 - x_j^a)^{-1}}{Q_j} \\ &= \frac{n}{a} + \sum_{j=1}^{n} log(x_j) - \sum_{j=1}^{n}\sum_{i=1}^{2} \frac{p_i g_i(x_j;\theta_i)(b_i - 1)(x_j^a\ log(x_j))(1 - x_j^a)^{-1}}{Q_j} = 0.\end{aligned} \tag{2.12}$$

Os estimadores de máxima verossimilhança dos parâmetros $\theta = (p_1, a, b_1, b_2)$ são obtidos resolvendo as equações 2.9, 2.10, 2.11 e 2.12, utilizando algum método numérico do tipo Newton Raphson.

2.5 Algoritmo EM

O algoritmo EM é um método que consiste em encontrar estimativas de máxima verossimilhança para um conjunto de parâmetros θ, geralmente utilizado quando temos variáveis não observáveis. No caso de trabalharmos com variáveis observáveis, dados completos, na maioria das vezes utilizamos o EMV (estimador de máxima verossimilhança), pois nosso problema tem em geral apenas um ponto de ótimo. Para as variáveis não observáveis, pode ser que nosso problema possua vários ótimos locais, então o EMV pode se confundir e basear a estimação dos parâmetros em um máximo local. O algoritmo EM acaba por ser uma generalização do EMV, pois ele também parte da maximização da log-verossimilhança, porém ao realizar sua otimização divide o problema em resolver várias subfunções separadamente em cada uma delas tem apenas um máximo global.

Uma área que o método EM é bastante utilizado é na estimação de parâmetros para mistura de distribuições. Isso ocorre porque as funções de verossimilhança baseadas em modelos de mistura podem ter vários pontos de máximo, então o algoritmo EM acaba tendo melhor desempenho em relação a outras técnicas de estimação. O algoritmo EM consiste em resolver os dois seguintes passos:

• Expectation: Calcular o valor esperado da log-verossimilhança dos dados faltantes, levando em conta chutes iniciais para os parâmetros.
• Maximization: Maximizar a esperança da log-verossimilhança achando novas estimativas provisórias para cada um dos parâmetros. Esses dois passos são repetidos iterativamente até atingir um critério de parada estabelecido no problema.

Para utilização do algoritmo EM é preciso reescrever as derivadas em relação a cada parâmetro encontradas pelo EMV em função de W_{ij}, que é a probabilidade de uma

observação x_j pertencer a i-ésima componente. Se as componentes assumem proporções fixas, então W_{ij} pode ser escrito de acordo com as seguintes equações:

$$W_{1j} = \frac{p_1 g_1(x_j;\theta_1)}{Q_j} \quad e \quad W_{2j} = 1 - W_{1j} = \frac{p_2 g_2(x_j;\theta_2)}{Q_j}. \tag{2.13}$$

Outra representação de W_{1j}, que será útil na análise de discriminante, é obtida manipulando as expressões de 2.13.

$$\begin{aligned}
W_{1j} &= \frac{p_1 g_1(x_j;\theta_1)}{Q_j} \\
&= \frac{p_1 g_1(x_j;\theta_1)}{p_1 g_1(x_j;\theta_1) + p_2 g_2(x_j;\theta_2)} \\
&= \frac{1}{1 + \frac{p_2 g_2(x_j;\theta_2)}{p_1 g_1(x_j;\theta_1)}} \\
&= \frac{1}{1 + \frac{p_2 b_2 (1-x_j^a)^{b_2-1}}{p_1 b_1 (1-x_j^a)^{b_1-1}}} \\
&= \frac{1}{1 + exp(log\frac{p_2 b_2}{p_1 b_1} + (b_2 - b_1) log(1 - x_j^a))} \\
&= (1 + exp(\alpha + \beta z))^{-1},
\end{aligned} \tag{2.14}$$

sendo

$$\alpha = log\frac{p_2 b_2}{p_1 b_1},\ \beta = (b_2 - b_1)\ e\ z = log(1 - x_j^a). \tag{2.15}$$

Para determinar o estimador de p_1 em função de W_{ij}, substituímos as expressões de 2.13 na equação 2.9.

$$\begin{aligned}
&\sum_{j=1}^{n} \frac{g_1(x_j;\theta_1) - g_2(x_j;\theta_2)}{Q_j} = 0 \\
&\sum_{j=1}^{n} \frac{p_1 g_1(x_j;\theta_1) - p_1 g_2(x_j;\theta_2)}{Q_j} = 0 \\
&\sum_{j=1}^{n} (W_{1j} - \frac{p_1 g_2(x_j;\theta_2)}{Q_j}) = 0 \\
&\sum_{j=1}^{n} (W_{1j} - \frac{p_1 p_2 g_2(x_j;\theta_2)}{p_2 Q_j}) = 0 \\
&\sum_{j=1}^{n} (W_{1j} - \frac{p_1 W_{2j}}{p_2}) = 0 \\
&\sum_{j=1}^{n} W_{1j} = \sum_{j=1}^{n} \frac{p_1 (1 - W_{1j})}{1 - p_1}
\end{aligned}$$

$$\frac{1-p_1}{p_1} = \frac{\sum_{j=1}^{n}(1-W_{1j})}{\sum_{j=1}^{n} W_{1j}}$$

$$\frac{1}{p_1} - 1 = \frac{n}{\sum_{j=1}^{n} W_{1j}} - 1$$

$$\hat{p}_1 = \frac{\sum_{j=1}^{n} W_{1j}}{n}. \tag{2.16}$$

Para obter o estimador de b_1, substituímos as expressões de 2.13 em 2.10 e isolamos $\hat{b_1}$.

$$\sum_{j=1}^{n} \frac{b_1^{-1} p_1 g_1(x_j;\theta_1)(1 + b_1\ log(1-x_j^a))}{Q_j} = 0$$

$$\sum_{j=1}^{n} b_1^{-1} W_{1j} b_1\ log(1-x_j^a) = -\sum_{j=1}^{n} b_1^{-1} W_{1j}$$

$$\hat{b}_1 = -\frac{\sum_{j=1}^{n} W_{1j}}{\sum_{j=1}^{n} W_{1j}\ log(1-x_j^a)}. \tag{2.17}$$

O estimador de b_2 é obtido da mesma forma, porém utilizando a equação 2.11.

$$\hat{b}_2 = -\frac{\sum_{j=1}^{n} W_{2j}}{\sum_{j=1}^{n} W_{2j}\ log(1-x_j^a)}. \tag{2.18}$$

Obtemos o estimador para a substituindo as expressões de 2.13 na equação 2.12 e procurando um $\hat{a}$ de tal forma que a equação seja valida utilizando $\hat{p}, \hat{b}_1, \hat{b}_2$ encontrados anteriormente.

$$\frac{n}{a} + \sum_{j=1}^{n} log(x_j) - \sum_{j=1}^{n}\sum_{i=1}^{2} \frac{p_i g_i(x_j;\theta_i)(b_1-1)(x_j^a\ log(x_j))(1-x_j^a)^{-1}}{Q_j} = 0$$

$$\hat{a}^{-1} = \frac{1}{n}\{\sum_{j=1}^{n} log(x_j)^{-1} + \sum_{j=1}^{n}\sum_{i=1}^{2} W_{ij}(b_i-1)(x_j^{\hat{a}}\ log(x_j))(1-x_j^{\hat{a}})^{-1}\}. \tag{2.19}$$

Os estimadores de $\theta = (p_1, b_1, b_2, a)$ são obtidos calculando as expressões de 2.16 até 2.19 iterativamente. Para isto, é necessário escolher valores iniciais para os parâmetros $(p_1^0, b_1^0, b_2^0, a^0)$. Foram obtidas expressões fechadas para os estimadores de p_1, b_1, b_2, enquanto para o estimador de a não foi possível. Sendo assim, na realização do algoritmo EM serão primeiramente resolvidas as equações (2.15,2.16,2.17) para encontrar $\hat{p}_1, \hat{b}_1, \hat{b}_2$ que serão utilizados na equação 2.19 para encontrar o valor de $\hat{a}$ através de algum método iterativo.

A matriz de covariância dos parâmetros $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4) = (p_1, b_1, b_2, a)$ pode ser obtida atráves do valor esperado das segundas derivadas do logaritmo da verossimilhança. Note que essa matriz é a inversa da matriz da informação de Fisher.

$$\sum = I^{-1}(\boldsymbol{\theta}) = \left[-E\left(\frac{\partial Log(L(\boldsymbol{\theta}))}{\partial\theta_i \partial\theta_j}\right)\right]^{-1}, \quad i, j = 1, 2, 3, 4. \tag{2.20}$$

2.6 Amostra Classificada

Se temos uma amostra de tamanho n da mistura na qual sabemos que n_1 observações são provenientes da primeira componente π_1 e n_2 da segunda π_2, em que $n = n_1 + n_2$, então conhecemos o grupo no qual cada elemento da amostra proveio. Nesta situação, o procedimento de estimação paramétrica tratará a função de verossimilhança da mistura como o produto de duas verossimilhanças, em que cada uma conterá somente os elementos associados a cada densidade. Se não tivéssemos o parâmetro a comum nas duas componentes então o processo de estimação poderia ser separado em resolver o problema de maximização para cada densidade Kumaraswamy separadamente. Porém, pelo fato de termos esta restrição as estimativas de máxima verossimilhança terão que ser obtidas utilizando a função de verossimilhança referente a mistura. A função de verossimilhança para esse tipo de amostragem é dada por

$$\begin{aligned}
L(\boldsymbol{\theta}) &= L_1(\boldsymbol{\theta}_1)L_1(\boldsymbol{\theta}_2) \\
&= \prod_{j=1}^{n_1} p_1 f_1(x_j;\theta_1) \prod_{j=1}^{n_2} p_2 f_2(x_j;\theta_2) \\
&= \prod_{j=1}^{n_1} p_1 a b_1 x_j^{a-1}(1-x_j^a)^{b_1-1} \prod_{j=1}^{n_2} p_2 a b_2 x_j^{a-1}(1-x_j^a)^{b_2-1} \\
&= a^{n_1+n_2}(p_1 b_1)^{n_1}((1-p_1)b_2)^{n_2} \prod_{j=1}^{n_1} x_j^{a-1}(1-x_j^a)^{b_1-1} \prod_{j=1}^{n_2} x_j^{a-1}(1-x_j^a)^{b_2-1} \\
&= a^n p_1^{n_1} b_1^{n_1}(1-p_1)^{n_2} b_2^{n_2} \prod_{j=1}^{n_1} x_j^{a-1}(1-x_j^a)^{b_1-1} \prod_{j=1}^{n_2} x_j^{a-1}(1-x_j^a)^{b_2-1}.
\end{aligned} \tag{2.21}$$

Sendo assim, o logaritmo da verossimilhança é

$$\begin{aligned}
l(\boldsymbol{\theta}) = nlog(a) + n_1 log(p_1) + n_1 log(b_1) + n_2 log(1-p_1) + n_2 log(b_2) + (a-1)\sum_{j=1}^{n_1} log(x_{1j}) \\
+ (a-1)\sum_{j=1}^{n_2} log(x_{2j}) + (b_1-1)\sum_{j=1}^{n_1} log(1-x_{1j}^a) + (b_2-1)\sum_{j=1}^{n_2} log(1-x_{2j}^a).
\end{aligned} \tag{2.22}$$

A seguir foram calculadas as derivadas parciais em relação a $\boldsymbol{\theta} = (p_1, a, b_1, b_2)$ e cada derivada parcial foi igualada a zero.

$$\frac{\partial Log(L(\boldsymbol{\theta}))}{\partial p_1} = \frac{n_1}{p_1} - \frac{n_2}{(1-p_1)} = 0$$
$$\frac{n_1}{p_1} = \frac{n_2}{(1-p_1)}$$
$$\frac{1-p_1}{p_1} = \frac{n-n_1}{n_1}$$
$$\frac{1}{p_1} - 1 = \frac{n}{n_1} - 1$$
$$\tilde{p_1} = \frac{n_1}{n}. \tag{2.23}$$

$$\frac{\partial Log(L(\boldsymbol{\theta}))}{\partial b_1} = \frac{n_1}{b_1} + \sum_{j=1}^{n_1} log(1-x_{1j}^a) = 0$$
$$\frac{n_1}{b_1} = -\sum_{j=1}^{n_1} log(1-x_{1j}^a)$$
$$\tilde{b}_1 = \frac{-n_1}{\sum\limits_{j=1}^{n_1} log(1-x_{1j}^a)}. \tag{2.24}$$
$$\tilde{b}_2 = \frac{-n_2}{\sum\limits_{j=1}^{n_2} log(1-x_{2j}^a)}. \tag{2.25}$$

$$\frac{\partial Log(L(\boldsymbol{\theta}))}{\partial a} = \frac{n}{a} + \sum_{i=1}^{2}\sum_{j=1}^{n_i}\left[log(x_{ij}) - (b_i-1)\frac{x_{ij}^a log(x_{ij})}{1-x_{ij}^a}\right] = 0$$
$$\frac{n}{a} + \sum_{i=1}^{2}\sum_{j=1}^{n_i}\left[log(x_{ij}) + \left(\frac{n_i}{\sum\limits_{j=1}^{n_i} log(1-x_{ij}^a)} + 1\right)\frac{x_{ij}^a log(x_{ij})}{1-x_{ij}^a}\right] = 0$$
$$\tilde{a} = n\left(\sum_{i=1}^{2}\sum_{j=1}^{n_i}\left[log(x_{ij}) + \frac{[n_i + \sum\limits_{j=1}^{n_i} log(1-x_{ij}^{\tilde{a}})]x_{ij}^{\tilde{a}} log(x_{ij})}{(1-x_{ij}^{\tilde{a}})\sum\limits_{j=1}^{n_i} log(1-x_{ij}^{\tilde{a}})}\right]\right)^{-1}. \tag{2.26}$$

Podemos obter a estimativa de p_1 de forma direta através da equação 2.23 porque conhecemos o número de observações da primeira componente n_1 e o tamanho da amostra n. Para achar a estimativa de a é preciso resolver a equação 2.26 através de algum algoritmo numérico. Para encontrar as estimativas de b_1 e b_2 basta resolver as equações 2.24 e 2.25 utilizando o valor de a obtido através da equação 2.26.

2.7 Função de Discriminante Ótimo

Como foi retratado na seção 2.2, atualmente existem diversas técnicas para discriminar um conjunto de dados em diferentes clusters. Dessa maneira, a função de discriminante

para mistura de duas Kumaraswamys será utilizada para classificar uma observação em uma das duas subpopulações presentes na mistura. Para introduzir a noção de discriminante com misturas bimodais vamos retratar inicialmente a função de discriminante ótima, que é usada quando conhecemos todos os parâmetros da densidade estudada.

A regra de classificação adotada será classificar a observação no cluster que apresentar maior verossimilhança para esse dado, ou seja, se a probabilidade de uma dada observação pertencer a primeira componente π_1 é maior do que pertencer a π_2 classificaremos ela como sendo da primeira componente, caso contrário classificaremos na segunda. Para definir a probabilidade de um elemento qualquer pertencer a primeira subpopulação usaremos a equação 2.14, pelo fato de acharmos uma forma simples de obter a função de discriminante através dela. Vale lembrar que a probabilidade de uma observação pertencer a segunda subpopulação é a probabilidade complementar de pertencer a primeira. As probabilidades de uma observação ser gerada por π_1 e π_2 são dadas por:

$$Pr(x_j \in \pi_1) = \frac{1}{(1+exp(\alpha+\beta z))} \quad \text{e} \quad Pr(x_j \in \pi_2) = \frac{exp(\alpha+\beta z)}{(1+exp(\alpha+\beta z))}, \tag{2.27}$$

sendo

$$\alpha = log\frac{p_2 b_2}{p_1 b_1},\ \beta = (b_2 - b_1)\ e\ z = log(1 - x_j^a). \tag{2.28}$$

A obtenção da função discriminante é obtida de acordo com a explicação da seção 2.2. Como z não depende linearmente de x_j, temos uma função do discriminante não linear ótimo que é dada por:

$$NLD_o(x_j) = \alpha + \beta z. \tag{2.29}$$

Assim, uma observação x_j é classificada como pertencente a π_1 se $NLD_o(x_j) \leq 0$ e pertencente a π_2 se $NLD_o(x_j) > 0$.

2.8 Função Discriminante para amostras classificadas e não classificadas

Quando os parâmetros da distribuição são desconhecidos, a função de discriminante utilizada dependerá dos parâmetros estimados da densidade. A estimativa dos parâmetros é feita através do método da máxima verossimilhança, porém dependendo de como os dados são amostrados a função de verossimilhança acaba tendo formas diferentes e consequentemente as estimativas dos parâmetros ficam diferentes. Desse modo, vamos considerar dois tipos de amostragem para obter as estimativas dos parâmetros, que são quando os dados são obtidos por amostras classificadas e amostras misturadas (não classificadas).

A função de discriminante não linear para amostras classificadas é obtida pela mesma expressão do discriminante não linear ótimo, porém nesse caso em vez de termos o valor conhecido dos parâmetros $\theta = (p, a, b_1, b_2)$ foi preciso encontrar suas estimativas através das equações 2.23, 2.24, 2.25 e 2.26. Sendo assim, a função de discriminante não linear para amostras classificadas é dada por:

$$NLD_c(x_j) = \tilde{\alpha} + \tilde{\beta}\tilde{z}, \tag{2.30}$$

sendo

$$\tilde{\alpha} = log\frac{\tilde{p}_2\tilde{b}_2}{\tilde{p}_1\tilde{b}_1}, \tilde{\beta} = (\tilde{b}_2 - \tilde{b}_1) \; e \; \tilde{z} = log(1 - x_j^{\tilde{a}}). \tag{2.31}$$

Se para uma observação x_j temos $NLD_c(x_j) > 0$ classificaremos ela como sendo pertencente a segunda população π_2 e caso contrário, $NLD_c(x_j) \leq 0$, então classificaremos a observação como pertencente a primeira população π_1.

Quando temos uma amostra de n elementos e as observações são não classificadas, então não sabemos de qual grupo cada elemento veio, sendo assim esse tipo de amostra é chamada de não classificada. Para cada observação da amostra não classificada sabemos somente o valor da densidade da mistura que ela assume, diferentemente da amostra classificada em que é conhecida a subpopulação que a observação foi gerada.

A função de discriminante não linear para amostra misturada (não classificada) é dada pela mesma expressão dos anteriores, mas com os parâmetros estimados pelas equações 2.16, 2.17, 2.18 e 2.19, que são provenientes das derivadas parciais do logaritmo da verossimilhança dado pela expressão 2.5. Desse modo, o $NLD_m(x_j)$ é obtido pelas equações abaixo.

$$NLD_m(x_j) = \hat{\alpha} + \hat{\beta}\hat{z}, \tag{2.32}$$

sendo

$$\hat{\alpha} = log\frac{\hat{p}_2\hat{b}_2}{\hat{p}_1\hat{b}_1}, \hat{\beta} = (\hat{b}_2 - \hat{b}_1) \; e \; \hat{z} = log(1 - x_j^{\hat{a}}). \tag{2.33}$$

Classificaremos uma observação como pertencente a segunda população π_2 se temos $NLD_c(x_j) > 0$ e em π_1 se $NLD_c(x_j) \leq 0$.

2.9 Erros de Classificação

Anteriormente foram encontradas as funções do discriminante não linear quando os parâmetros da mistura são conhecidos (discriminante ótimo) e quando são estimados (discriminante para amostras classificadas e mistas). De acordo com o valor dessas funções

classificamos a observação no grupo em que ela apresenta maior probabilidade de pertencer, portanto por se tratar de probabilidades podemos cometer erros no momento de classificar cada elemento. Para um conjunto de dados podemos formar várias funções para discriminar as observações, porém não existe a função de discriminante correta, simplismente cada uma parte de uma ideia diferente para tentar reduzir os erros de classificação.

Uma dada observação é classificada em π_1 se $NLD_k(x_j) \leq 0$, para k=o(ótimo), c(classificada), m(mista) e em π_2 caso contrário. Dessa maneira, quando cometemos um erro significa que classificamos uma observação como π_1 ($NLD_k(x_j) \leq 0$) dado que a ela era pertencente a π_2 ou classificamos em π_2 ($NLD_k(x_j) > 0$) dado que ela pertence a π_1. As probabilidades de cometermos esses erros são dadas por $e_{1k} = Pr(NLD_k(x_j) > 0|\pi_1)$ e $e_{2k} = Pr(NLD_k(x_j) \leq 0|\pi_2)$, para k=o(ótimo), c(classificada), m(mista).

Como sabemos que $NLD_k(x_j) = \alpha_k + \beta_k z_k$, sendo $\alpha_o = \alpha$ e $\beta_o = \beta$, $\alpha_c = \hat{\alpha}$ e $\beta_c = \hat{\beta}$, $\alpha_m = \tilde{\alpha}$ e $\beta_m = \tilde{\beta}$, então podemos definir e_{1k} como:

$$\begin{aligned}
e_{1k} &= Pr(NLD_k(x_j) > 0 \,|\pi_1) \\
&= Pr(\alpha_k + \beta_k z_k > 0 \,|\pi_1) \\
&= Pr\left(log(\frac{p_{2k}b_{2k}}{p_{1k}b_{1k}}) + (b_{2k} - b_{1k})log(1 - x_j^{a_k}) > 0 \,|\pi_1\right) \\
&= Pr\left((b_{2k} - b_{1k})log(1 - x_j^{a_k}) > log(\frac{p_{1k}b_{1k}}{p_{2k}b_{2k}}) \,|\pi_1\right) \\
&= Pr\left(1 - x_j^{a_k} > \left(\frac{p_{1k}b_{1k}}{p_{2k}b_{2k}}\right)^{1/(b_{2k}-b_{1k})} \,|\pi_1\right) \\
&= Pr\left(x_j^{a_k} < 1 - \left(\frac{p_{1k}b_{1k}}{p_{2k}b_{2k}}\right)^{1/(b_{2k}-b_{1k})} \,|\pi_1\right) \\
&= Pr\left(x_j < \left(1 - \left(\frac{p_{1k}b_{1k}}{p_{2k}b_{2k}}\right)^{1/(b_{2k}-b_{1k})}\right)^{1/a_k} \,|\pi_1\right). \qquad (2.34)
\end{aligned}$$

Para que a probabilidade tratada convirja é preciso que $\left(1 - \left(\frac{p_{1k}b_{1k}}{p_{2k}b_{2k}}\right)^{1/(b_{2k}-b_{1k})}\right)$ seja positivo, porque dependendo da escolha de a_k chegaríamos na acumulada de um número complexo. Além disso, se a expressão falada não for positiva também temos a possibilidade de ter uma acumulada de um valor negativo, que não faria sentido pois a distribuição Kumaswamy tem valores não nulos somente no intervalo (0,1). Outra maneira de reescrever a restrição está expressa abaixo:

$$1 > \left(\frac{p_{1k}b_{1k}}{p_{2k}b_{2k}}\right)^{1/(b_{2k}-b_{1k})},$$

implicando nos dois casos a seguir:

$$\begin{cases} p_{2k}b_{2k} > p_{1k}b_{1k} \;\; e \;\; b_{2k} > b_{1k} \\ p_{2k}b_{2k} < p_{1k}b_{1k} \;\; e \;\; b_{2k} < b_{1k} \end{cases}$$

$$= \begin{cases} b_{2k} - \dfrac{p_{1k}b_{1k}}{p_{2k}} > 0 \;\; e \;\; b_{2k} - b_{1k} > 0 \\ b_{2k} - \dfrac{p_{1k}b_{1k}}{p_{2k}} < 0 \;\; e \;\; b_{2k} - b_{1k} < 0 \end{cases}$$

Se as condições acima forem satisfeitas, então e_{1k} é dado pela expressão abaixo.

$$e_{1k} = Pr\left(x_j < \left(1 - \left(\frac{p_{1k}b_{1k}}{p_{2k}b_{2k}}\right)^{1/(b_{2k}-b_{1k})}\right)^{1/a_k} |\pi_1\right) \tag{2.35}$$

$$e_{1k} = F(w_k, b_{1k}, a_k), \tag{2.36}$$

$$\text{sendo} \;\; w_k = \left(1 - \left(\frac{p_{1k}b_{1k}}{p_{2k}b_{2k}}\right)^{1/(b_{2k}-b_{1k})}\right)^{1/a_k}.$$

Da mesma maneira podemos obter e_{2k}, que é a probabilidade de classificarmos uma observação em π_1 ($NLD_k(x_j) \leq 0$) dado que a ela pertencente a π_2. Para que e_{2k} esteja bem definido precisamos das mesma condições que definimos para e_{1k}. Dessa forma, e_{2k} é dado pela expressão abaixo.

$$e_{2k} = 1 - F(w_k, b_{2k}, a_k). \tag{2.37}$$

A função $F(w_k, b_i, a)$ pode ser obtida através da distribuição acumulada das componentes da Kumaraswamy. Escrevendo ela em função de w_j temos que:

$$F(w_k, b_i, a) = 1 - (1 - w_k^{a_k})^{b_{ik}}, \;\; \text{para} \;\; i = 1, 2 \;\; \text{e} \;\; k = o, c, m. \tag{2.38}$$

Como na prática dificilmente saberemos de qual subpopulação cada observação provém, calcular somente e_{1k} e e_{2k} não ajudariam no problema de classificação. Dessa maneira, seria útil calcular a probabilidade de classificar uma observação como pertencendo a uma subpopulação e ela pertencer a outra. Para calcular esse tipo de erro definiremos como ε_{1k} a probabilidade de classificarmos uma observação em π_2 e ela pertencer a π_1 e como ε_{2k} a probabilidade de classificar em π_1 e ela ser de π_2. A seguir, apresentamos como obter ε_{1k} em função de e_{1k}.

$$\begin{aligned} \varepsilon_{1k} &= Pr(NLD_k(x_j) > 0 \cap x_j \in \pi_1) \\ \varepsilon_{1k} &= Pr(x_j \in \pi_1)\; Pr(NLD_k(x_j) > 0 \,|x_j \in \pi_1) \\ \varepsilon_{1k} &= p_{1k}e_{1k}. \end{aligned} \tag{2.39}$$

Da mesma forma podemos obter ε_{2k}, que é dado pela expressão abaixo.

$$\varepsilon_{2k} = (1 - p_{1k})e_{2k}. \tag{2.40}$$

A probabilidade total de cometer um erro de classificação pode ser definida como classificar uma observação de π_1 em π_2 ou o contrário.

$$\begin{aligned} \varepsilon_k =& \varepsilon_{1k} + \varepsilon_{2k} \\ \varepsilon_k =& p_{1k}e_{1k} + (1 - p_{1k})e_{2k}. \end{aligned} \tag{2.41}$$

3 Resultados Numéricos e aplicação

3.1 Simulação

No intuito de avaliar as técnicas propostas no trabalho, foram realizados experimentos com dados simulados. A simulação tem como objetivo averiguar a qualidade das estimativas dos parâmetros da mistura quando temos amostras classificadas e não classificadas, além de verificar as taxas de erros de classificação obtidas através das diferentes funções de discriminante $NLD_o(x), NLD_m(x)$ e $NLD_c(x)$.

O experimento foi realizado para seis diferentes combinações de parâmetros, variando o tamanho de amostras geradas entre n=100 e n=500. A seguir estão os gráficos das misturas para cada diferente combinação paramétrica utilizada.

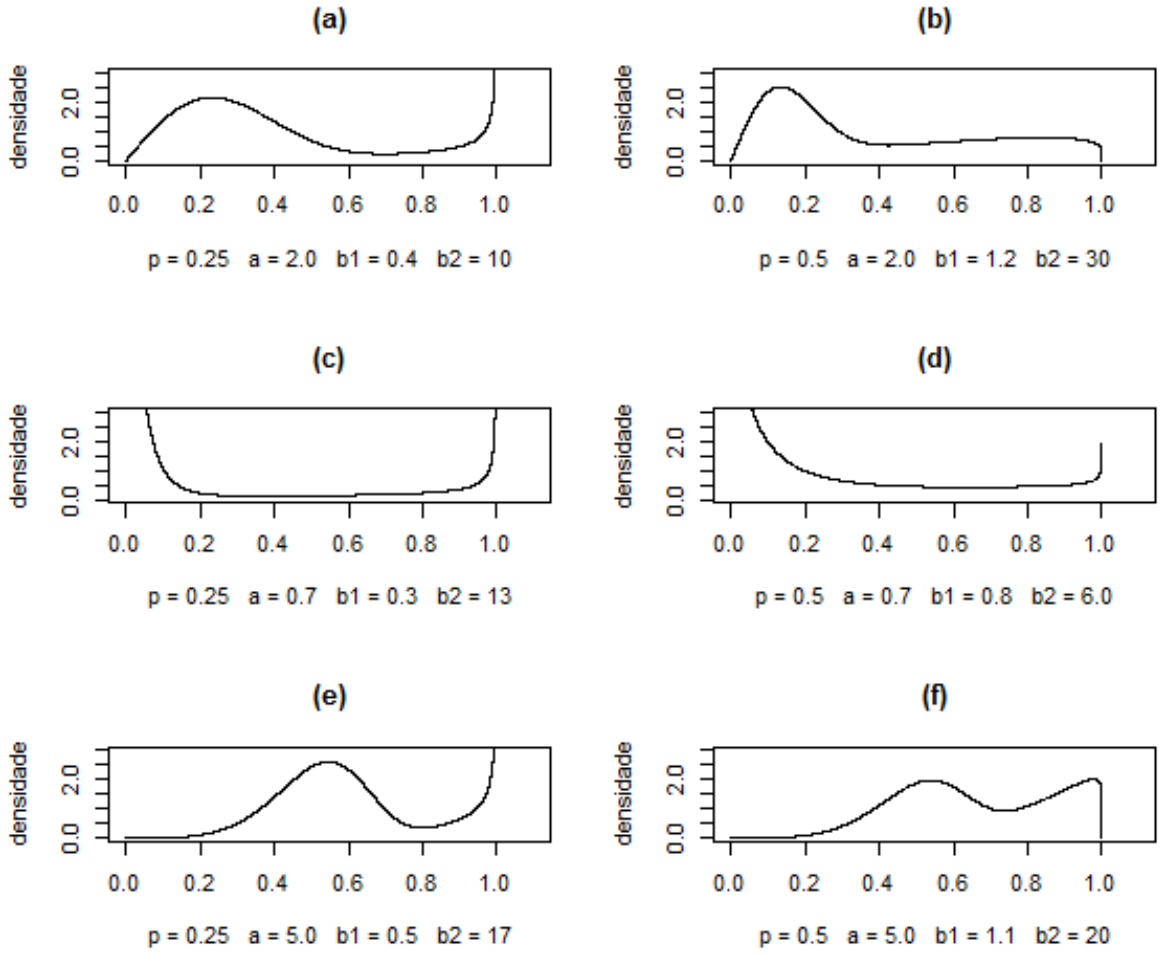

Figura 3 – Misturas de duas densidades Kumaraswamys com curvas bimodais representadas por diferentes combinações de parâmetros.

No conjunto de parâmetros foram utilizadas proporções p=(0.25, 0.50), o parâmetro a variando como a=(2.0, 0.7, 5.0) e os parâmetros b_1 e b_2 foram escolhidos de modo fossem obtidos gráficos bimodais.

É possível notar que os gráficos com $a = 0.7$ apresentam descontinuidade no ponto $x = 0$. Essa descontinuidade ocorre para qualquer valor de a escolhido no intervalo de (0,1), pois o termo x^{a-1} da densidade Kumaraswamy acaba por deixar x no denominador quando a é menor do que um. Também é possível notar que temos descontinuidade no ponto $x = 1$ quando b_1 ou b_2 estão no intervalo (0,1). Isso acontece porque o termo $(1 - x^a)^{b_i-1}$ faz com que $(1 - x^a)$ fique no denominador da densidade para b_i menor que 1.

Algoritmo para gerar uma observação de mistura de Kumaraswamys.

• Gerar duas observações U_1 e U_2 que seguem distribuição U(0,1).
• Se U_1<p, então é gerada uma observação da primeira componente pelo método da inversão. Para gerar a observação calcula-se $F^{-1}(U_2, a, b_1)$.
• Se $U_1 \geq p$, então geramos uma observação da segunda componente calculando $F^{-1}(U_2, a, b_2)$.

Estes passos são repetidos 100 vezes quando queremos gerar 100 observações e 500 vezes para 500. No final do processo as observações que foram geradas pela primeira componente formam um vetor que chamaremos de $mist_1$ e as observações geradas pela segunda em outro vetor que chamaremos $mist_2$. Ao juntarmos $mist_1$ com $mist_2$ criamos um vetor de observações $mist$ no qual não sabemos de qual componente cada observação provém. O vetor $mist$ será usado para o desenvolvimento do algoritmo EM porque estamos trabalhando com uma amostra não classificada, enquanto os vetores $mist_1$ e $mist_2$ serão usados para encontrar as estimativas via amostras classificadas.

Passos para estimação pelo algoritmo EM

Para realização da simulação com algoritmo EM são geradas observações da mistura e vamos supor que não se sabe de qual componente cada observação foi gerada (amostra não classificada). Sendo assim, basta escolher valores para os parâmetros p_1, a, b_1, b_2 para gerar um vetor de observações ($mist$) que será utilizado para estimar os verdadeiros valores de cada parâmetro. Utilizaremos as estimativas obtidas em 2.20, 2.27, 2.28 e 2.29 para o desenvolvimento do algoritmo EM para mistura de duas Kumaraswamys que está descrito a seguir.

O procedimento do algoritmo EM para mistura de duas distribuições Kumaraswamy está descrito a seguir:

• Escolher valores iniciais para os parâmetros p_1^0, b_1^0, b_2^0, a^0.

• E-step: Calcular os valores esperados das observações virem da primeira componente na k-ésima atualização.

$$\hat{W}_{1j}^{(k)} = \frac{\hat{p}_1^{(k)} g_1(x_j; \hat{\theta}_1^{(k)})}{\hat{p}_1^{(k)} g_1(x_j; \hat{\theta}_1^{(k)}) + \hat{p}_2^{(k)} g_2(x_j; \hat{\theta}_2^{(k)})}.$$

• M-step: São calculadas as estimativas dos parâmetros p_1, b_1, b_2 na (k+1)-ésima atualização.

$$\hat{p}_1^{(k+1)} = \frac{\sum_{j=1}^{n} W_{1j}^{(k)}}{n}$$

$$\hat{b}_1^{(k+1)} = -\frac{\sum_{j=1}^{n} W_{1j}^{(k)}}{\sum_{j=1}^{n} W_{1j}^{(k)}\ log(1 - x_j^{a^{(k)}})}.$$

$$\hat{b}_2^{(k+1)} = -\frac{\sum_{j=1}^{n} (1 - W_{1j}^{(k)})}{\sum_{j=1}^{n} (1 - W_{1j}^{(k)})\ log(1 - x_j^{a^{(k)}})}.$$

Calcular a estimativa do parâmetro a utilizando $\hat{b}_1$ e $\hat{b}_2$ calculados anteriormente. Pode-se obter o valor de $\hat{a}$ por um método do tipo Newton Rhapson de tal forma que resolva a equação abaixo.

$$\frac{1}{\hat{a}^{(k+1)}} = \frac{1}{n}\{\sum_{j=1}^{n}\ log(x_j)^{-1} + \sum_{j=1}^{n}\sum_{i=1}^{2} W_{ij}^{(k)}(b_i^{(k+1)} - 1)(x_j^{\hat{a}^{(k+1)}}\ log(x_j))(1 - x_j^{\hat{a}^{(k+1)}})^{-1}\}.$$

• Calcular novamente $\hat{W}_{1j}$ com um novo vetor de estimativas $\hat{\theta} = (\hat{p}, \hat{a}, \hat{b}_1, \hat{b}_2)$ e em seguida os passos E e M são repetidos. O algoritmo EM chegará ao fim quando $|\hat{\theta}^k - \hat{\theta}^{k-1}| < 0.001$, sendo k o número de iterações realizadas.

Passos para estimação de amostras classificadas

Na estimação dos parâmetros com amostras classificadas os dados gerados pela mistura serão separados em dois vetores. O primeiro vetor $mist_1$ será composto pelas observações que foram geradas pela primeira componente x_{1j} e o vetor $mist_2$ as observações geradas pela segunda x_{2j}. Dessa maneira, como utilizamos uma informação a mais sobre os dados em relação ao algoritmo EM são esperadas melhores estimativas dos parâmetros. A seguir está o processo para encontrar as estimativas dos parâmetros com dados classificados via máxima verossimilhança através das estimativas obtidas em 2.30, 2.31, 2.32 e 2.33.

• Primeiramente encontraremos as estimativas $\tilde{p}$ e $\tilde{a}$ porque suas expressões não dependem de outros parâmetros. Calculamos $\tilde{a}$ iterativamente porque não temos uma expressão fechada para ele, enquanto $\tilde{p}$ foi obtido pela proporção de elementos que vieram da primeira subpopulação.

$$\tilde{p}_1 = \frac{n_1}{n},$$
$$\tilde{a} = n\left(\sum_{i=1}^{2}\sum_{j=1}^{n_i}\left[log(x_{ij}) + \frac{[n_i + \sum_{j=1}^{n_i} log(1-x_{ij}^{\tilde{a}})]x_{ij}^{\tilde{a}} log(x_{ij})}{(1-x_{ij}^{\tilde{a}})\sum_{j=1}^{n_i} log(1-x_{ij}^{\tilde{a}})}\right]\right)^{-1}.$$

• Utilizamos o valor de $\tilde{a}$ obtido para encontrarmos $\tilde{b}_1$ e $\tilde{b}_2$.

$$\tilde{b}_1 = \frac{-n_1}{\sum_{j=1}^{n_1} log(1-x_{1j}^{a})}.$$
$$\tilde{b}_2 = \frac{-n_2}{\sum_{j=1}^{n_2} log(1-x_{2j}^{a})}.$$

3.1.1 Comparação dos métodos propostos

No intuito de comparar a estimação dos parâmetros para amostras classificadas e não classificadas foram escolhidas seis combinações diferentes de parâmetros para realização da simulação. Para cada caso o processo foi repetido cem vezes com novas amostras e foi calculado o valor médio de cada um dos quatro parâmetros.

A seguir temos as Tabelas das estimativas utilizando amostras não classificadas (Mistura) e classificadas (Classificada) para amostras de tamanho n=100 e n=500. É possível notar que na Tabela 2 (n=100) as estimativas médias obtidas para b_2 geram resultados mais precisos nas amostras classificadas, enquanto para os outros parâmetros não se observa muita diferença entre os dois procedimentos. Na Tabela 3 o tamanho das amostras foi aumentado para n=500, com isso os resultados obtidos nas estimativas ficaram praticamente iguais. Além disso, quando o tamanho da amostra foi aumentado em ambos os casos as estimativas melhoraram bastante.

Tabela 2 – Estimativa média dos parâmetros com n=100.

Valores para os parâmetros				Mistura				Classificada			
p	a	b_1	b_2	$\hat{p}$	$\hat{a}$	$\hat{b}_1$	$\hat{b}_2$	$\tilde{p}$	$\tilde{a}$	$\tilde{b}_1$	$\tilde{b}_2$
0.25	2.0	0.4	10	0.254	2.037	0.430	11.268	0.252	2.037	0.425	10.982
0.50	2.0	1.2	30	0.497	2.054	1.277	37.210	0.498	2.050	1.281	34.271
0.25	0.7	0.3	13	0.252	0.728	0.306	15.282	0.250	0.725	0.304	14.627
0.50	0.7	0.8	6	0.480	0.721	0.806	7.166	0.501	0.710	0.823	6.277
0.25	5.0	0.5	17	0.251	5.085	0.539	18.857	0.255	5.075	0.544	18.018
0.50	5.0	1.1	20	0.504	5.199	1.168	25.855	0.505	5.169	1.166	23.845

Tabela 3 – Estimativa média dos parâmetros com n=500.

Valores para os parâmetros				Mistura				Classificada			
p	a	b_1	b_2	$\hat{p}$	$\hat{a}$	$\hat{b}_1$	$\hat{b}_2$	$\tilde{p}$	$\tilde{a}$	$\tilde{b}_1$	$\tilde{b}_2$
0.25	2.0	0.4	10	0.248	2.010	0.409	10.293	0.248	2.009	0.408	10.209
0.50	2.0	1.2	30	0.503	2.030	1.209	32.328	0.504	2.025	1.210	31.525
0.25	0.7	0.3	13	0.252	0.701	0.300	13.099	0.252	0.701	0.301	13.080
0.50	0.7	0.8	6	0.497	0.704	0.796	6.036	0.502	0.706	0.804	6.058
0.25	5.0	0.5	17	0.253	4.988	0.510	17.190	0.252	4.987	0.509	17.092
0.50	5.0	1.1	20	0.501	5.080	1.120	21.396	0.500	5.075	1.118	20.960

A fim de avaliar a variabilidade das estimativas obtidas foi calculado o erro quadrático médio em relação aos parâmetros propostos. O erro quadrático médio (EQM) dos estimadores de θ para amostras classificadas e não classificadas são definidos respectivamente por:

$$EQM_m = \sum_{k=1}^{100} \frac{(\hat{\theta}_k - \theta)^2}{100}$$

e

$$EQM_c = \sum_{k=1}^{100} \frac{(\tilde{\theta}_k - \theta)^2}{100}.$$

Nas Tabelas 4 e 5 são apresentados o EQM de cada parâmetro para amostras de tamanho n=100 e n=500, respectivamente. Nas duas Tabelas as estimativas obtidas utilizando amostras classificadas tiveram menor variabilidade em relação aos verdadeiros parâmetros. Tanto no caso de amostras misturadas quanto nas classificadas o EQM diminuiu quando a amostra foi aumentada de n=100 para n=500.

Tabela 4 – Estimativa do Erro Quadrático Médio dos parâmetros com n=100.

Valores para os parâmetros				Mistura				Classificada			
p	a	b_1	b_2	$\hat{p}$	$\hat{a}$	$\hat{b}_1$	$\hat{b}_2$	$\tilde{p}$	$\tilde{a}$	$\tilde{b}_1$	$\tilde{b}_2$
0.25	2.0	0.4	10	0.002	0.058	0.020	17.393	0.002	0.040	0.014	9.311
0.50	2.0	1.2	30	0.004	0.073	0.060	426.243	0.002	0.040	0.057	157.337
0.25	0.7	0.3	13	0.002	0.008	0.004	39.904	0.001	0.006	0.003	21.812
0.50	0.7	0.8	6	0.010	0.015	0.024	22.713	0.003	0.006	0.014	3.135
0.25	5.0	0.5	17	0.002	0.306	0.022	90.367	0.002	0.195	0.018	27.452
0.50	5.0	1.1	20	0.004	0.467	0.049	275.196	0.002	0.308	0.039	188.022

Tabela 5 – Estimativa do Erro Quadrático Médio dos parâmetros com n=500.

Valores para os parâmetros				Mistura				Classificada			
p	a	b_1	b_2	$\hat{p}$	$\hat{a}$	$\hat{b}_1$	$\hat{b}_2$	$\tilde{p}$	$\tilde{a}$	$\tilde{b}_1$	$\tilde{b}_2$
0.25	2.0	0.4	10	0.000	0.009	0.002	1.890	0.000	0.006	0.002	1.034
0.50	2.0	1.2	30	0.001	0.015	0.009	69.744	0.000	0.009	0.008	27.366
0.25	0.7	0.3	13	0.000	0.001	0.001	3.382	0.000	0.001	0.001	2.236
0.50	0.7	0.8	6	0.002	0.002	0.006	1.782	0.000	0.001	0.004	0.384
0.25	5.0	0.5	17	0.000	0.055	0.002	5.806	0.000	0.041	0.002	4.466
0.50	5.0	1.1	20	0.001	0.103	0.008	19.299	0.001	0.055	0.007	8.082

Para visualizar o quanto as estimativas obtidas diferiram em relação a cada parâmetro foi calculado o viés para cada estimativa. O calculo do viés para amostras misturadas e classificadas é obtido pelas expressões abaixo.

$$Viés_m = \sum_{k=1}^{100} \frac{(\hat{\theta}_k - \theta)}{100}$$
$$Viés_c = \sum_{k=1}^{100} \frac{(\tilde{\theta}_k - \theta)}{100}$$

Nas Tabelas 6 e 7 são apresentados os valores dos vieses de cada estimativa. Na Tabela 6 é possível perceber que somente para o parâmetro b_2 o viés para amostras classificadas foi sempre menor que para amostras misturadas. Na Tabela 7 os vieses não tiveram diferença significativa para amostras classificadas em relação as misturadas, pelo fato do tamanho amostral ter aumentado.

Tabela 6 – Estimativa do Viés para os parâmetros com n=100.

Valores para os parâmetros				Mistura				Classificada			
p	a	b_1	b_2	$\hat{p}$	$\hat{a}$	$\hat{b}_1$	$\hat{b}_2$	$\tilde{p}$	$\tilde{a}$	$\tilde{b}_1$	$\tilde{b}_2$
0.25	2.0	0.4	10	0.004	0.037	0.030	1.268	0.002	0.037	0.025	0.982
0.50	2.0	1.2	30	-0.003	0.054	0.077	7.210	-0.002	0.050	0.081	4.271
0.25	0.7	0.3	13	0.002	0.028	0.006	2.282	0.000	0.025	0.004	1.627
0.50	0.7	0.8	6	-0.020	0.021	0.006	1.166	0.001	0.010	0.023	0.277
0.25	5.0	0.5	17	0.001	0.085	0.039	1.857	0.005	0.075	0.044	1.018
0.50	5.0	1.1	20	0.004	0.199	0.068	5.855	0.005	0.169	0.066	3.845

Tabela 7 – Estimativa do Viés para os parâmetros com n=500.

Valores para os parâmetros				Mistura				Classificada			
p	a	b_1	b_2	$\hat{p}$	$\hat{a}$	$\hat{b}_1$	$\hat{b}_2$	$\tilde{p}$	$\tilde{a}$	$\tilde{b}_1$	$\tilde{b}_2$
0.25	2.0	0.4	10	-0.002	0.010	0.009	0.293	-0.002	0.009	0.008	0.209
0.50	2.0	1.2	30	0.003	0.030	0.009	2.328	0.004	0.024	0.010	1.525
0.25	0.7	0.3	13	0.002	0.001	0.000	0.099	0.002	0.001	0.001	0.080
0.50	0.7	0.8	6	-0.003	0.004	-0.004	0.036	0.002	0.006	0.004	0.058
0.25	5.0	0.5	17	0.003	-0.012	0.010	0.190	0.002	-0.013	0.009	0.092
0.50	5.0	1.1	20	0.001	0.080	0.020	1.396	-0.000	0.075	0.018	0.960

Nas Tabelas 8 e 9 foram calculadas as probabilidades de classificar uma observação errada dado que sabemos sua origem. Para calcular essas probabilidades usaremos as fórmulas obtidas em 2.36 e 2.37.

$$
\begin{aligned}
&e_{1k} = F(w_j, b_{1k}, a_k),\\
&e_{2k} = 1 - F(w_j, b_{2k}, a_k),\\
&\text{sendo } \; w_k = \left(1 - \left(\frac{p_{1k}b_{1k}}{p_{2k}b_{2k}}\right)^{1/(b_{2k}-b_{1k})}\right)^{1/a_k}.
\end{aligned}
$$

No caso ótimo foram calculadas as probabilidades de cada combinação paramétrica somente uma vez, pois conhecemos os parâmetros do modelo. Já para amostras não classificadas e classificadas foram calculadas as estimativas médias dos erros para as 100 vezes que os algoritmos eram repetidos com novos banco de dados. Podemos notar que essas estimativas não foram afetadas com o aumento da amostra de 100 para 500.

Tabela 8 – Probabilidade de cometer erros de classificação dado que se conhece a origem de cada observação para n=100.

Valores para os parâmetros				Ótimo		Mistura		Classificada	
p	a	b1	b2	$\bar{e}_{1o}$	$\bar{e}_{2o}$	$\bar{e}_{1m}$	$\bar{e}_{2m}$	$\bar{e}_{1c}$	$\bar{e}_{2c}$
0.25	2.0	0.4	10	0.165	0.011	0.165	0.012	0.165	0.011
0.50	2.0	1.2	30	0.126	0.035	0.128	0.038	0.126	0.036
0.25	0.7	0.3	13	0.109	0.007	0.105	0.007	0.105	0.007
0.50	0.7	0.8	6	0.267	0.098	0.278	0.099	0.268	0.104
0.25	5.0	0.5	17	0.131	0.008	0.138	0.009	0.137	0.010
0.50	5.0	1.1	20	0.155	0.046	0.151	0.048	0.152	0.047

Tabela 9 – Probabilidade de cometer erros de classificação dado que se conhece a origem de cada observação para n=500.

Valores para os parâmetros				Ótimo		Mistura		Classificada	
p	a	b1	b2	$\bar{e}_{1o}$	$\bar{e}_{2o}$	$\bar{e}_{1m}$	$\bar{e}_{2m}$	$\bar{e}_{1c}$	$\bar{e}_{2c}$
0.25	2.0	0.4	10	0.165	0.011	0.166	0.011	0.166	0.011
0.50	2.0	1.2	30	0.126	0.035	0.122	0.035	0.123	0.035
0.25	0.7	0.3	13	0.109	0.007	0.109	0.007	0.109	0.007
0.50	0.7	0.8	6	0.267	0.098	0.270	0.100	0.265	0.099
0.25	5.0	0.5	17	0.131	0.008	0.133	0.009	0.133	0.009
0.50	5.0	1.1	20	0.155	0.046	0.152	0.046	0.153	0.046

Nas Tabelas 10 e 11 foram calculadas as probabilidades totais de cometer um erro de classificação, ou seja, discriminar um objeto pertencente a primeira componente na segunda ou o oposto. Foi obtida a estimativa média das probabilidades totais calculando o valor médio destas probabilidades para cada 100 vezes que o processo foi repetido. Para obtenção da probabilidade total utilizamos a fórmula 2.41.

Os cálculos do EQM e Viés para amostras misturadas e classificadas foram obtidos utilizando a diferença das 100 estimativas obtidas para cada conjunto de parâmetros em relação a probabilidade total obtida utilizando os verdadeiros parâmetros da distribuição. A seguir estão as fórmulas utilizadas para os cálculos do viés e EQM em relação ao erro total ótimo.

$$Viés_k = \sum_{k=1}^{100} \frac{(\hat{\varepsilon}_k - \varepsilon_o)}{100}, \quad k = m, c.$$

$$EQM_k = \sum_{k=1}^{100} \frac{(\hat{\varepsilon}_k - \varepsilon_o)^2}{100}, \quad k = m, c.$$

Tabela 10 – Probabilidade total dos erros de classificação com seus respectivos erros quadráticos médios e vieses para n=100.

Valores para os parâmetros				Ótimo	Mistura			Classificada		
p	a	b1	b2	$\bar{\varepsilon}_o$	$\bar{\varepsilon}_m$	MSE	Vies	$\bar{\varepsilon}_c$	MSE	Vies
0.25	2.0	0.4	10	0.050	0.050	0.000	0.001	0.050	0.000	0.000
0.50	2.0	1.2	30	0.080	0.081	0.001	0.001	0.080	0.000	-0.000
0.25	0.7	0.3	13	0.032	0.031	0.000	-0.001	0.031	0.000	-0.001
0.50	0.7	0.8	6	0.182	0.175	0.003	-0.007	0.183	0.001	0.001
0.25	5.0	0.5	17	0.039	0.042	0.000	0.002	0.042	0.000	0.003
0.50	5.0	1.1	20	0.101	0.098	0.001	-0.003	0.099	0.000	-0.002

Tabela 11 – Probabilidade total dos erros de classificação com seus respectivos erros quadráticos médios e vieses para n=500.

Valores para os parâmetros				Ótimo	Mistura			Classificada		
p	a	b1	b2	$\bar{\varepsilon}_o$	$\bar{\varepsilon}_m$	MSE	Vies	$\bar{\varepsilon}_c$	MSE	Vies
0.25	2.0	0.4	10	0.050	0.049	0.000	-0.000	0.049	0.000	-0.000
0.50	2.0	1.2	30	0.080	0.079	0.000	-0.002	0.079	0.000	-0.001
0.25	0.7	0.3	13	0.032	0.033	0.000	0.000	0.033	0.000	0.000
0.50	0.7	0.8	6	0.182	0.182	0.000	-0.000	0.182	0.000	0.000
0.25	5.0	0.5	17	0.039	0.040	0.000	0.001	0.040	0.000	0.001
0.50	5.0	1.1	20	0.101	0.099	0.000	-0.002	0.099	0.000	-0.002

As probabilidades totais dos erros de classificação foram calculadas utilizando as distribuições acumuladas das componentes Kumaraswamy. Sendo assim, no intuito de averiguar se os erros são condizentes para as amostras geradas, foram calculadas as proporções dos erros encontrados utilizando a função de discriminante para cada observação das amostras simuladas. A seguir segue o procedimento para calcular tais proporções:

• Utilizando as estimativas dos parâmetros obtidas pelos métodos propostos é criada a função do discriminante.

$$NLD_k(x_j) = log(\frac{p_{2k}b_{2k}}{p_{1k}b_{1k}}) + (b_{2k} - b_{1k})log(1 - x_j^{a_k}).$$

• Cada observação x_j gerada pela simulação entra na função discriminante, que retorna em qual grupo ela foi classificada. Com isso, é criado um vetor que indica o grupo em que cada observação foi classificada.

• Como sabemos de qual componente cada observação foi gerada, calculamos a proporção

de observações mal classificadas pela função discriminante. Os erros são contabilizados quando classificamos uma observação pertencente ao primeiro grupo no segundo ou caso contrário.

$$p_k = \frac{\text{número de observações mal classificadas}}{n}.$$

As Tabelas 12 e 13 foram construídas com intenção de comparar os erros $\bar{\varepsilon}_m$ e $\bar{\varepsilon}_c$ obtidos nas Tabelas 10 e 11 com a proporção de erros encontradas utilizando a função do discriminante para cada observação das amostras simuladas. Cada combinação paramétrica foi simulada 100 vezes e no final foi calculada o valor médio para proporção de erros.

Tabela 12 – Proporção total das observações mal classificadas em comparação com a probabilidade total dos erros de classificação para n=100.

Valores para os parâmetros				Mistura		Classificada	
p	a	b1	b2	$\bar{p}_m$	$\bar{\varepsilon}_m$	$\bar{p}_c$	$\bar{\varepsilon}_c$
0.25	2.0	0.4	10	0.053	0.050	0.048	0.050
0.50	2.0	1.2	30	0.088	0.081	0.083	0.080
0.25	0.7	0.3	13	0.031	0.031	0.027	0.031
0.50	0.7	0.8	6	0.206	0.175	0.184	0.183
0.25	5.0	0.5	17	0.045	0.042	0.042	0.042
0.50	5.0	1.1	20	0.106	0.098	0.099	0.099

Tabela 13 – Proporção total das observações mal classificadas em comparação com a probabilidade total dos erros de classificação para n=500.

Valores para os parâmetros				Mistura		Classificada	
p	a	b1	b2	$\bar{p}_m$	$\bar{\varepsilon}_m$	$\bar{p}_c$	$\bar{\varepsilon}_c$
0.25	2.0	0.4	10	0.051	0.049	0.049	0.049
0.50	2.0	1.2	30	0.081	0.079	0.080	0.079
0.25	0.7	0.3	13	0.033	0.033	0.032	0.033
0.50	0.7	0.8	6	0.190	0.182	0.184	0.182
0.25	5.0	0.5	17	0.040	0.040	0.040	0.040
0.50	5.0	1.1	20	0.101	0.099	0.099	0.099

Assim como nas Tabelas 12 e 13, as Tabelas 14 e 15 foram criadas com objetivo de averiguar se os erros encontrados através das distribuições de cada componente estão de acordo com a proporção de erros obtidas pela função discriminante. A diferença é que agora estão sendo analisados os erros específicos para cada grupo. As proporções p_{1k} e p_{2k} são obtidos pelas razões do número total de observações mal classificadas no grupo 1 e 2 em relação ao número total de observações, respectivamente.

$$p_{1k} = \frac{\text{número de observações mal classificadas do grupo 1}}{n},$$
$$p_{2k} = \frac{\text{número de observações mal classificadas do grupo 2}}{n}.$$

As probabilidades de cometer erros de classificação podem ser obtidas pelo produto dos erros condicionais das Tabelas 8 e 9 com as estimativas das proporções encontradas nas Tabelas 2 e 3.

$$\begin{aligned}\varepsilon_{1k} &= p_k e_{1k} \\ \varepsilon_{2k} &= (1 - p_k) e_{2k}.\end{aligned}$$

Sendo assim, notamos que nas Tabelas 14 e 15 as proporções p_{ik} assumem valores praticamente iguais aos erros ε_{ik}, para i=1,2.

Tabela 14 – Proporção das observações mal classificadas de cada grupo e a probabilidade de cometer erros de classificação em cada grupo para n=100.

Valores para os parâmetros				Mistura				Classificada			
p	a	b1	b2	$\bar{p}_{1m}$	$\bar{\varepsilon}_{1m}$	$\bar{p}_{2m}$	$\bar{\varepsilon}_{2m}$	$\bar{p}_{1c}$	$\bar{\varepsilon}_{1c}$	$\bar{p}_{2c}$	$\bar{\varepsilon}_{2c}$
0.25	2.0	0.4	10	0.043	0.042	0.010	0.009	0.041	0.042	0.007	0.009
0.50	2.0	1.2	30	0.067	0.063	0.021	0.019	0.063	0.063	0.020	0.018
0.25	0.7	0.3	13	0.025	0.026	0.005	0.005	0.023	0.026	0.004	0.005
0.50	0.7	0.8	6	0.155	0.134	0.051	0.052	0.133	0.134	0.052	0.052
0.25	5.0	0.5	17	0.039	0.035	0.006	0.007	0.036	0.035	0.005	0.007
0.50	5.0	1.1	20	0.080	0.076	0.026	0.024	0.075	0.077	0.024	0.024

Tabela 15 – Proporção das observações mal classificadas de cada grupo e a probabilidade de cometer erros de classificação em cada grupo para n=500.

Valores para os parâmetros				Mistura				Classificada			
p	a	b1	b2	$\bar{p}_{1m}$	$\bar{\varepsilon}_{1m}$	$\bar{p}_{2m}$	$\bar{\varepsilon}_{2m}$	$\bar{p}_{1c}$	$\bar{\varepsilon}_{1c}$	$\bar{p}_{2c}$	$\bar{\varepsilon}_{2c}$
0.25	2.0	0.4	10	0.042	0.041	0.009	0.008	0.041	0.041	0.008	0.008
0.50	2.0	1.2	30	0.063	0.061	0.018	0.017	0.062	0.062	0.017	0.017
0.25	0.7	0.3	13	0.028	0.028	0.005	0.005	0.027	0.027	0.005	0.005
0.50	0.7	0.8	6	0.140	0.134	0.050	0.050	0.135	0.133	0.049	0.050
0.25	5.0	0.5	17	0.033	0.034	0.007	0.007	0.033	0.034	0.007	0.007
0.50	5.0	1.1	20	0.076	0.076	0.024	0.023	0.076	0.077	0.023	0.023

3.2 Exemplo Ilustrativo

Como o algoritmo EM é um método iterativo, será realizada uma simulação para uma amostra de tamanho n=100, na qual mostraremos as estimativas obtidas para diferentes números de iterações. Em seguida, a estimativa encontrada no final do processo será comparada com a que foi obtida pelo método para amostras classificadas. Além disso, será feito um gráfico comparando as distribuições resultantes de cada técnica com o histograma gerado pelos dados simulados.

No exemplo a amostra de tamanho n=100 será gerada por uma mistura de Kumaraswamy com a combinação paramétrica $\theta = (0.4, 2.0, 0.5, 15)$. A seguir vamos apresentar as estimativas obtidas pelo algoritmo EM após 5, 10, 20 e 43 iterações. Usaremos 43 iterações como resultado final porque foi quando o critério de parada $|\hat{\theta}^k - \hat{\theta}^{k-1}| < 0.001$ chegou ao fim. Os valores iniciais escolhidos para os parâmetros foram $\theta^0 = (0.5, 5.0, 5.0, 20)$.

Abaixo são apresentados os gráficos das estimativas obtidas para diferentes números de iterações do algoritmo EM:

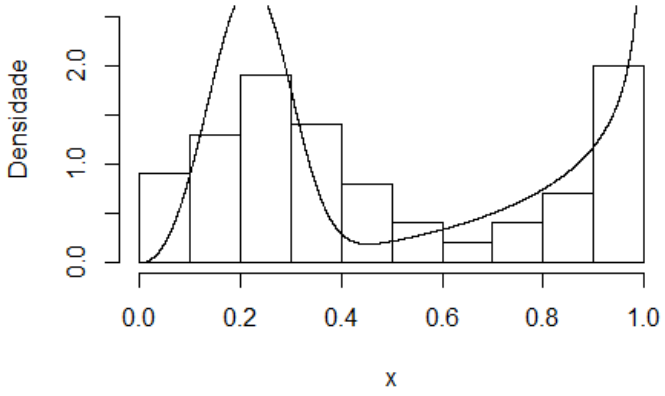

Figura 4 – Densidade obtida via algoritmo EM após 5 iterações.

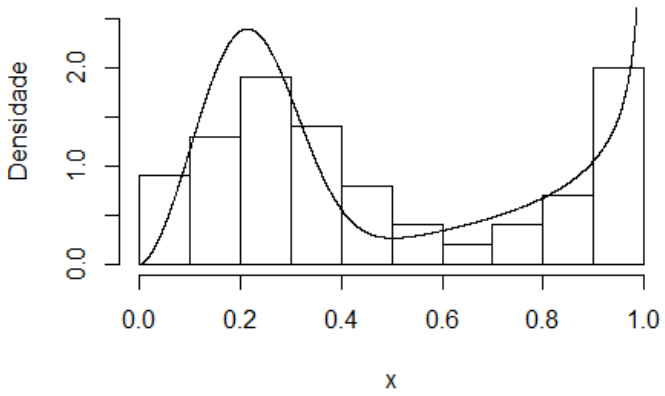

Figura 5 – Densidade obtida via algoritmo EM após 10 iterações.

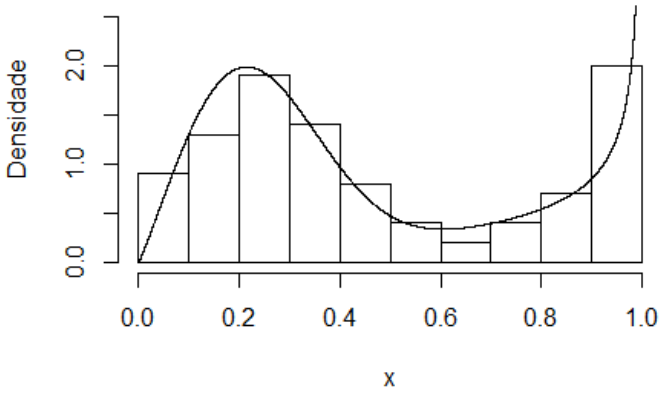

Figura 6 – Densidade obtida via algoritmo EM após 20 iterações.

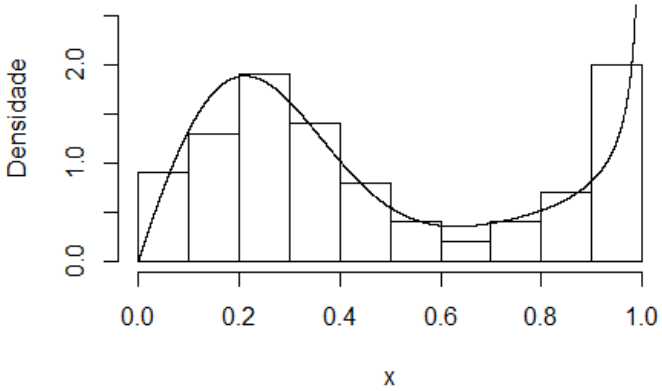

Figura 7 – Densidade obtida via algoritmo EM após critério de parada (43 iterações).

A seguir temos os valores estimados dos parâmetros $\theta = (0.4, 2.0, 0.5, 15)$ em relação a cada número de iterações do algoritmo EM.

Tabela 16 – Estimativas obtidas pelo algoritmo EM em função do número de iterações.

Iterações	$\hat{p}$	$\hat{a}$	$\hat{b}_1$	$\hat{b}_2$
5	0.456	3.240	0.664	93.318
10	0.447	2.641	0.595	37.934
20	0.407	2.088	0.493	14.019
43	0.402	1.948	0.472	11.067

Diferentemente do algoritmo EM, a estimação dos parâmetros para amostras classificadas não é resolvida iterativamente. Como as equações para obtermos $\tilde{p}$ e $\tilde{a}$ dependem só dos dados amostrais podemos resolve-las primeiramente, em seguida encontramos $\tilde{b_1}$ e $\tilde{b_2}$ que precisam do valor que foi encontrado para $\tilde{a}$ e dos valores gerados das observações.

A seguir temos o gráfico da mistura de Kumaraswamys e as estimativas que obtemos para a estimação por amostras classificadas.

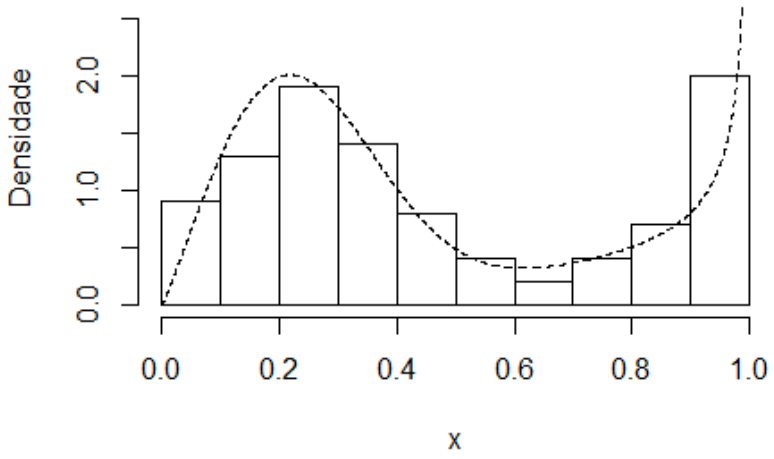

Figura 8 – Densidade obtida por estimação com amostra classificada.

As estimativas dos parâmetros obtidos para amostra classificada são apresentados pela Tabela 17.

Tabela 17 – Estimativas obtidas para amostra classificada.

$\tilde{p}$	$\tilde{a}$	$\tilde{b}_1$	$\tilde{b}_2$
0.390	2.086	0.470	13.527

De acordo com as estimativas obtidas, é possível notar que neste caso não usarmos a informação adicional de qual componente cada observação provém, trouxe resultados semelhantes para quando usamos está informação. Em seguida, é apresentado a comparação do gráfico e das estimativas obtidas por cada método.

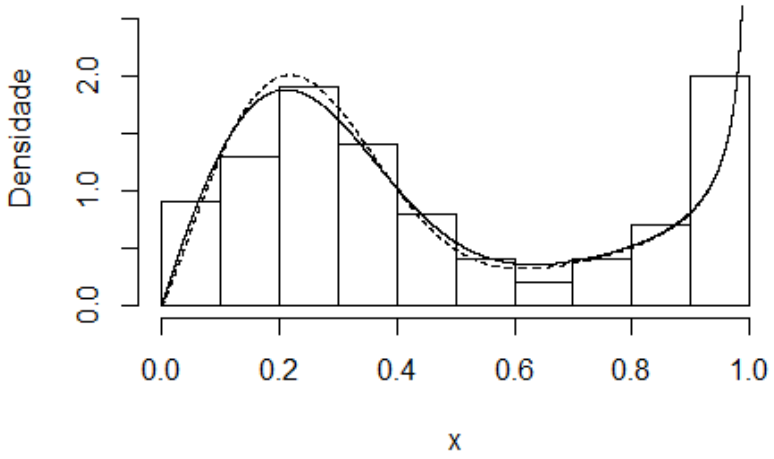

Figura 9 – Densidade obtida por EM (curva contínua) em comparação com a densidade obtida para amostra classificada (curva pontilhada).

Na Tabela 18 apresentamos as estimativas obtidas pelo algoritmo EM após 43 iterações e as estimativas obtidas por amostras classificadas para n=100.

Tabela 18 – Estimativas obtidas por amostra classificada e via algoritmo EM.

Valores para os parâmetros				Mistura				Classificada			
p	a	b_1	b_2	$\hat{p}$	$\hat{a}$	$\hat{b}_1$	$\hat{b}_2$	$\tilde{p}$	$\tilde{a}$	$\tilde{b}_1$	$\tilde{b}_2$
0.4	2	0.5	15	0.402	1.948	0.472	11.067	0.390	2.086	0.470	13.527

As estimativas dos parâmetros obtidos permitem que calculemos os erros de classificação para cada um dos métodos propostos. Logo, apresentaremos através das Tabelas 19 e 20 a probabilidade de cometer erros de classificação em cada grupo e a probabilidade total de cometer erros de classificação, respectivamente.

Tabela 19 – Probabilidade de cometer erros de classificação em cada grupo para n=100.

Valores para os parâmetros				Mistura		Classificada	
p	a	b1	b2	$\hat{\varepsilon}_{1m}$	$\hat{\varepsilon}_{2m}$	$\tilde{\varepsilon}_{1c}$	$\tilde{\varepsilon}_{2c}$
0.4	2	0.5	15	0.058	0.014	0.049	0.011

Tabela 20 – Probabilidade total dos erros de classificação para n=100.

Valores para os parâmetros				Mistura	Classificada
p	a	b1	b2	$\hat{\varepsilon}_m$	$\tilde{\varepsilon}_c$
0.4	2	0.5	15	0.073	0.061

É possível notar que mesmo com as estimativas para amostras classificadas e misturadas serem relativamente distantes, as probabilidades de cometer erros são muito parecidas para ambos os casos.

3.3 Análise com dados reais

Na análise para dados simulados as estimativas obtidas ficaram próximas dos verdadeiros parâmetros, porém era de se esperar um bom ajuste porque os dados gerados eram de uma mistura de duas Kumaraswamys. No intuito de analisar a qualidade do ajuste deste modelo de maneira mais geral serão obtidas as estimativas desta mistura para dados não simulados. O banco de dados a ser trabalhado foi apresentado na seção sobre Mistura de Distribuições

e se refere a 272 intervalos de tempo em minutos que o geyser Old Faithful (Yellowstone National Park, Wyoming, USA) demora para entrar em erupção. Como estes dados não estão no intervalo (0,1), correspondente ao suporte da distribuição Kumaraswamy, então será aplicada uma transformação nos dados para deixa-los neste intervalo. Sendo assim, optamos por dividir o tempo de cada observação por 100, tratando as observações em centésimos de minuto.

A principio não é conhecido o fenômeno que fez com que os dados gerados apresentassem dois grupos distintos. Sendo assim, para a análise referente aos erros de classificação trataremos as componentes da mistura como grupo 1 e grupo 2. Através da função discriminante vamos estimar quantas observações foram geradas pelo grupo 1 e 2 e calcular as probabilidades referentes a classificar uma observação qualquer no grupo errado. Como saberíamos a qual grupo cada uma das observações está associada, isso poderia ajudar pessoas da área a identificar o fenômeno que fez com que estes dados se comportassem como pertencentes a dois grupos diferentes.

A seguir apresentamos o histograma dos dados transformados e a curva ajustada para eles através da mistura de duas Kumaraswamys. Também segue os valores referentes a estimativa de cada parâmetro da mistura.

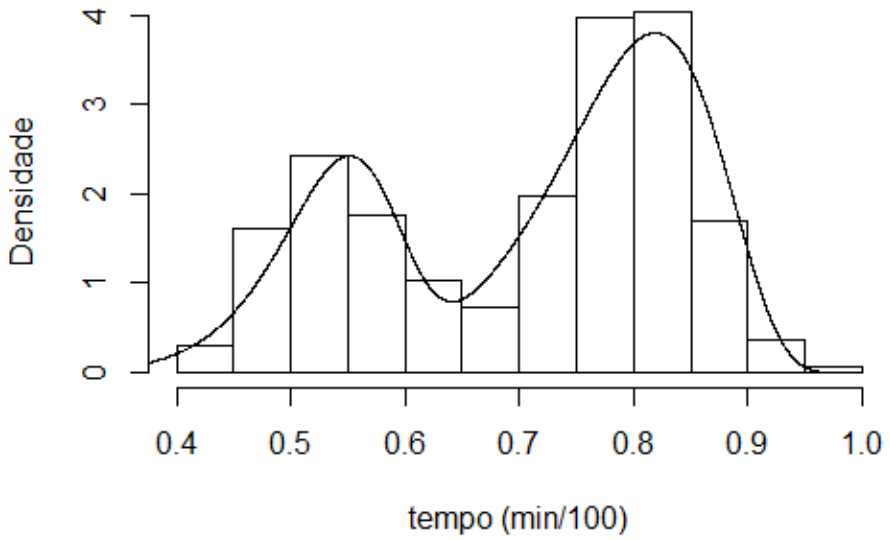

Figura 10 – Densidade de uma mistura de duas Kumaraswamys para os intervalos de tempo em (minutos/100) entre erupções do geyser Old Faithful.

As estimativas obtidas para o ajuste da mistura de duas densidades Kumaraswamys são descritas na Tabela 21.

Tabela 21 – Estimativas obtidas para os intervalos de tempo em (minutos/100) entre erupções do geyser Old Faithful.

$\hat{p}$	$\hat{a}$	$\hat{b}_1$	$\hat{b}_2$
0.701	11.443	9.071	878.083

A função de discriminante não linear pode ser encontrada pela expressão 2.30 utilizando as estimativas obtidas na Tabela 21.

$$NLD_m(x_j) = 3.719 + (869.012)log(1 - x_j^{11.443}). \tag{3.1}$$

Para cada observação x_j foi calculado o valor da função discriminante correspondente, se temos que $NLD_o(x_j) \leq 0$ então classificamos a observação em π_1 e se $NLD_o(x_j) > 0$ em π_2. Com isso obtemos que das 271 observações 185 foram classificadas como pertencentes a componente 1 e 87 a componente 2.

A seguir apresentamos os gráficos das curvas correspondentes a cada uma das densidades Kumaraswamys com suas respectivas proporções. Com isso, identificamos que a primeira componente $f_1(x_j;\theta_1)$ está associada com a curva a direita do histograma e a segunda $f_2(x_j;\theta_2)$ a curva a esquerda. Ou seja, as 185 observações que classificamos como geradas pela primeira componente estão representadas pela parte direita do histograma e as 87 classificadas na segunda como parte da esquerda.

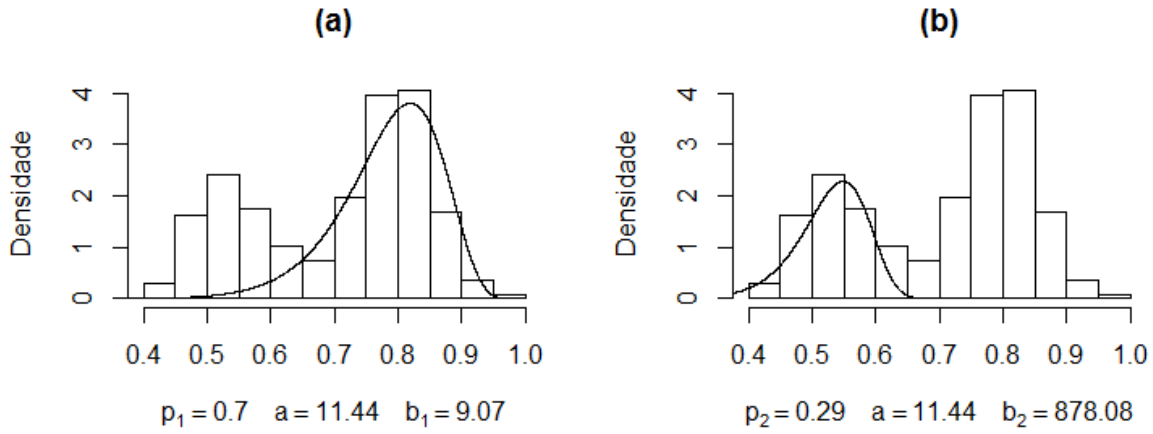

Figura 11 – (a) - Densidade da primeira componente da mistura associada ao seu peso da mistura e (b) - Densidade da segunda componente da mistura associada ao seu peso da mistura.

Através da função discriminante também é possível calcular a probabilidade de classificarmos uma observação qualquer em um dos grupos sendo que ela pertence ao outro. Com isso, a probabilidade de cometer o erro de classificar uma observação no grupo 2 e ela pertencer ao grupo 1 é $\hat{\varepsilon}_{1m} = 0.0267$ e a probabilidade de classificar no grupo 1 e ela pertencer ao grupo 2 é $\hat{\varepsilon}_{2m} = 0.0069$. Somando as duas probabilidades temos a probabilidade de se cometer um erro qualquer que é igual a $\hat{\varepsilon}_m = 0.0336$.

4 Conclusão

O trabalho abordado retrata como são realizados os métodos para estimação dos parâmetros de uma mistura de duas distribuições Kumaraswamys a suas respectivas funções de discriminante. São analisados os comportamentos das estimativas e dos discriminantes quando temos amostras classificadas e não classificadas para diferentes tamanhos de amostra.

O resultado esperado era que a análise feita com dados classificados fosse mais precisa pelo fato de serem usadas mais informações referentes as observações. De fato os Vieses e os EQM's das estimativas dos parâmetros foram sempre menores quando utilizadas amostras classificadas, ou seja, para cada uma das 100 diferentes amostras geradas as estimativas ficam em geral mais próximas dos verdadeiros parâmetros. Porém, ao observarmos as estimativas médias para cada parâmetro os resultados para amostras misturadas e classificadas foram parecidos para os tamanhos amostrais como 100 e 500, valendo ressaltar que quanto maior a amostra, as estimativas para amostras não classificadas ficam cada vez mais parecidas com as das amostras classificadas. Em suma, usar amostras classificadas trazem estimativas um pouco mais precisas, porém a análise para amostras misturadas também apresentam bons resultados, melhorando com o aumento do tamanho amostral.

Para analisar o desempenho da função discriminante foram observadas as probabilidades de se cometer erros de classificação para cada diferente combinação paramétrica. Note que a melhor estimativa do discriminante não é a que apresenta menor probabilidade de classificar uma observação errado e sim o que a probabilidade se aproxima mais em relação ao discriminante ótimo. Sendo assim, ao analisarmos as probabilidades de erro de classificação é possível perceber que elas apresentam minimas variações quando mudamos o tamanho das amostras ou quando temos amostras misturadas ou classificadas. Logo, a função discriminante utilizada se mostrou invariante em relação ao tipo de amostragem e ao tamanho amostral.

Referências

[1] McLachlan, G; Peel, D. (2000) *Finite Mixture Models*, New York. Willey Series In Probability And Statistics.

[2] Jones, M.C. (2009). *Kumaraswamy's distribution: A beta-type distribution with some tractability advantages*. Statistical Methodology, p. 70–81.

[3] Mahmoud, M. A. W., Moustafa, H. M. (1993). *Estimation of a discriminant function from a mixture of two gamma distributions when the sample size is small*, Mathl. Comput. Modelling, p. 87-95.

[4] Ahmad, K. E.,Abd-Elrahman, A.M. (1994) *Updating a nonlinear discriminant function estimated from a mixture of two Weibull distributions*, Math. Comput. Modelling, p. 41–51.

[5] Ahmad, K.E., Jaheen, Z.F. and Modhesh,A.A. (2010), *Estimation of a discriminant function based on small sample size from a mixture of two Gumbel distributions*, Comm. Statist. Simulation Comput, p. 713–725.

[6] Amoh, R.K. (1985), *Estimation of a discriminant function from a mixture of two inverse Gaussian distributions when the sample size is small*, J. Statist. Comput. Simul, p. 275-286.

[7] O'Neill, T. J. (1978). Normal discrimination with unclassified observations. J. Am. Stat. Assoc. 73:821–826.

[8] Khattree, R. & Naik, D.N. (2000). Multivariate data reduction and discrimination with SAS software. Cary, NC, USA: SAS Institute Inc., 558 p.

[9] Casella, G., Berger, R. L. Inferência estatística - segunda edição. Centage Learning, 2010.

[10] Tatiana Benaglia, Didier Chauveau, David R. Hunter, Derek Young (2009). mixtools: An R Package for Analyzing Finite Mixture Models. Journal of Statistical Software, 32(6), 1-29. URL http://www.jstatsoft.org/v32/i06/.

[11] R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.