Tema

Tendências em Matemática Aplicada e Computacional, **15**, N. 1 (2014), 59-71

www.scielo.br/tema
doi: 10.5540/tema.2014.015.01.0059

Estimação dos Parâmetros da Mistura de Duas Componentes GEV via Algoritmo EM*

C.E.G. OTINIANO** e E.C.M. TEIXEIRA

Recebido em 9 julho, 2013 / Aceito em 26 março, 2014

RESUMO. Modelos probabilísticos de misturas finitas de densidades cujas componentes são as densidades da distribuição de Valor Extremal Generalizado tem uma ampla aplicabilidade em diversas áreas como finanças e hidrologia. Neste trabalho, os estimadores dos parâmetros da mistura de duas componentes GEV são obtidos via o algoritmo EM. Apresentamos também ilustrações numéricas do comportamento dos estimadores obtidos através de simulação, bem como uma aplicação a dados reias.

Palavras-chave: GEV 1, mistura finita 2, algoritmo EM 3.

1 INTRODUÇÃO

Dadas as funções de densidade de probabilidade (f.d.p.) d-dimensionais $f_1, \ldots, f_k$ de uma família de densidades

$$\mathcal{F} = \{f(x;\theta), \quad x \in R^d, \quad \theta \in \Theta\},$$

onde Θ é o espaço paramétrico dessa família. Uma mistura finita de $f_1, \ldots, f_k \in \mathcal{F}$ é uma f.d.p.

$$h = p_1 f_1 + \cdots + p_k f_k, \quad \sum_{j=1}^{k} p_j = 1, \tag{1.1}$$

onde os números $p_1, \ldots, p_k$ são chamados de pesos da mistura e $f_1, \ldots, f_k$ chamadas de componentes da mistura.

Devido a sua definição, as misturas finitas de distribuições ou de densidades de probabilidade são frequentemente utilizadas para modelar dados de populações heterogêneas. Tais misturas permitem construir modelos probabilísticos de uma ampla variedade de fenômenos, em por exemplo, engenharia, economia e hidrologia, entre outros. Teoria e aplicações de misturas podem ser encontradas em Titterington et al. (1985), McLachlan & Basford (1998), e McLachlan & Peel (2000).

*Pesquisa parcialmente financiada por CAPES/PROCAD.
**Autor correspondente: Cira Etheowalda Guevara Otiniano
Departamento de Estatística, Universidade de Brasília, 70910-900 Brasília, DF, Brasil. E-mail: cira@unb.br

Neste trabalho abordamos a estimação do modelo (1.1) quando $d = 1$ e $k = 2$, onde as componentes da mistura são densidades da distribuição de Valor Extremal Generalizado univariadas.

A distribuição de Valor Extremal Generalizado do inglês *generalized extreme value (GEV)* distribution, G_γ, é a distribuição limite de máximos (ou mínimos) normalizados de sequências de variáveis aleatórias (v.a.'s) independentes e identicamente distribuídas (i.i.d.). Isto é, seja $\{X_n\}$ tal sequência. Se existirem sequências de números reais $\{c_n\},\ c_n > 0$ e $\{d_n\}$ tais que

$$\lim_{n\to\infty} P\left(\frac{\max\{X_1, X_2, \ldots, X_n\} - d_n}{c_n} \leq x\right) = G_\gamma(x),$$

então a distribuição GEV introduzida por Jenkinson (1955) é definida por

$$G_\gamma(x;\sigma,\mu) = \begin{cases} \exp\left\{-\left[1+\dfrac{\gamma}{\sigma}(x-\mu)\right]^{-\frac{1}{\gamma}}\right\}, & \gamma \neq 0 \\ \exp\left[-\exp\left(-\dfrac{x-\mu}{\sigma}\right)\right], & \gamma = 0, \end{cases} \tag{1.2}$$

onde $\gamma,\ \mu \in R$ e $\sigma > 0$ são parâmetros de forma (índice caudal), locação e escala, respectivamente. Aqui $1 + \frac{\gamma}{\sigma}(x-\mu) \geq 0$.

A correspondente f.d.p. de G_γ, é dada por

$$g(x;\gamma,\sigma,\mu) = \begin{cases} \dfrac{1}{\sigma}\left[1+\dfrac{\gamma}{\sigma}(x-\mu)\right]^{-\frac{1}{\gamma}-1} \exp\left\{-\left[1+\dfrac{\gamma}{\sigma}(x-\mu)\right]^{-\frac{1}{\gamma}}\right\}, & \gamma \neq 0 \\ \dfrac{1}{\sigma}\exp\left(-\dfrac{x-\mu}{\sigma}\right)\exp\left[-\exp\left(-\dfrac{x-\mu}{\sigma}\right)\right], & \gamma = 0. \end{cases} \tag{1.3}$$

Estas distribuições são utilizadas em hidrologia para analisar a frequência de fluídos e em finanças para calcular o valor em risco (VaR) de retornos, perdas ou ganhos extremos.

Escalante-Sandoval (2007) utilizou misturas de duas componentes GEV para avaliar dados sobre a frequência de fluídos de rios de uma região do Noroeste do México. Neste caso, os parâmetros foram estimados por máxima verossimilhança utilizando vários métodos computacionais intensivos e as estimativas obtidas por esse procedimento, para vários casos, não convergiram. Outro trabalho que envolve mistura de distribuições com alguma componente GEV é o de Kollu et al. (2012), no qual, com o objetivo de descrever características da velocidade de ventos, os autores apresentam modelos de mistura de duas densidades cujas componentes são Weibull e GEV, Weibull e lognormal, e GEV e lognormal. Neste caso, a estimação também foi feita por máxima verossimilhança. Em geral, o método de máxima verossimilhança, o algoritmo EM, e os métodos bayesianos são os métodos alternativos de estimação dos parâmetros de misturas finitas de distribuições.

Com o objetivo de apresentar um método de estimação alternativo ao de Escalante-Sandoval (2007), dos parâmetros de misturas de duas componentes GEV, neste trabalho descrevemos as expressões a serem calculadas utilizando o algoritmo EM e mostramos o resultado das estimativas para alguns experimentos. Uma aplicação com dados reais também é adicionada.

Este trabalho esta organizado em cinco seções. Na Seção 2, apresentamos os principais resultados deste trabalho, expressões que devem ser calculadas para estimar os parâmetros da mistura de duas componentes GEV via algoritmo EM. Na seção 3, mostramos os resultados das estimações obtidas para algumas simulações. Já na seção 4, uma aplicação da estimação de um modelo de mistura para dados reais é apresentada. Por fim, nas Seções 5 e 6 fazemos os agradecimentos e citamos as referência bibliográficas, respectivamente.

2 ESTIMAÇÃO VIA ALGORITMO EM

O algoritmo EM do inglês "Estimation Maximization" é um algoritmo clássico da estatística usado para determinar estimativas de parâmetros em modelos de mistura, em modelos com dados faltantes, e em modelos em que a estimação dos parâmetros por máxima verossimilhança apresenta problemas. O livro McLachlan & Peel (2000) é uma boa referência para esse assunto. Nesta seção utilizamos o algoritmo EM para obter as estimativas de misturas de duas componentes GEV.

Seja X uma v.a. com f.d.p. $h(x, \Theta)$ mistura de duas componentes, definida em (1.1), e dada por

$$h(x; \Theta) = p_1 g_1(x; \theta_1) + p_2 g_2(x; \theta_2), \quad p_1 + p_2 = 1, \tag{2.1}$$

onde $\Theta = (p_1, \theta_1, \theta_2)$ com $\theta_\ell = (\gamma_\ell, \sigma_\ell, \mu_\ell)$, $\ell = 1, 2$ e $g_\ell(.; \theta_\ell)$ é uma componente de uma das famílias

$$\mathcal{G}_{\gamma^+} := \left\{ g(\gamma, \sigma, \mu) = \frac{1}{\sigma}\left[1 + \frac{\gamma}{\sigma}(x-\mu)\right]^{-\frac{1}{\gamma}-1} \exp\left\{ -\left[1 + \frac{\gamma}{\sigma}(x-\mu)\right]^{-\frac{1}{\gamma}} \right\},\ \gamma > 0 \right\},$$

$$\mathcal{G}_{\gamma^-} := \left\{ g(\gamma, \sigma, \mu) = \frac{1}{\sigma}\left[1 + \frac{\gamma}{\sigma}(x-\mu)\right]^{-\frac{1}{\gamma}-1} \exp\left\{ -\left[1 + \frac{\gamma}{\sigma}(x-\mu)\right]^{-\frac{1}{\gamma}} \right\},\ \gamma < 0 \right\},$$

$$\mathcal{G}_0 := \left\{ g(0, \sigma, \mu) = \frac{1}{\sigma} \exp\left(-\frac{x-\mu}{\sigma}\right) \exp\left[-\exp\left(-\frac{x-\mu}{\sigma}\right)\right] \right\}.$$

Nesta seção, usamos o Algoritmo EM para obter estimativas de Θ, baseado nos valores x_1, $x_2, \ldots, x_N$ de uma amostra aleatoria de tamanho N da v.a. X.

Uma breve descrição do algoritmo EM, é a seguinte. Se assumirmos que X é observado e gerado por alguma distribuição paramétrica $f(x; \Theta)$, chamamos X ($\{x_1, \ldots, x_N\}$) de dados incompletos. Estes dados são completados com $Y(\{y_1, \ldots, y_N\})$, então $Z = (X, Y)$ são os dados completos cuja densidade conjunta é

$$f(z; \Theta) = f(x, y; \Theta) = f(y/x, \Theta) f(x; \Theta).$$

Com esta nova densidade, defini-se a função de verossimilhança dos dados completos

$$L(\Theta/z) = L(\Theta/x, y) = f(X, Y; \Theta).$$

O algoritmo EM alterna o Passo E (da esperança), onde obtem-se

$$Q(\Theta, \Theta^{(k)}) = E[\ln(f(X, Y/\Theta))/X, \Theta^{(k)}],$$

com o Passo M (da maximização), onde se calcula $\Theta^{(k+1)}$ ao maximizar $Q(\Theta, \Theta^{(k)})$.

No caso de mistura, $f(x; \Theta) = h(x; \Theta)$ dada em (2.1) e Q é dada pela expressão

$$Q(\Theta, \Theta^{(k)}) = \sum_{l=1}^{2} \sum_{i=1}^{N} \log(p_l g_l(x_i; \theta_l^{(k)})) g(l/x_i, \Theta^{(k)}).$$

Em síntese, na $(k+1)$-ésima iteração do passo-E, a atualização da estimativa de p_1, é dada por

$$p_1^{(k+1)} = \frac{1}{N} \sum_{i=1}^{N} g(1/x_i, \theta_1^{(k)}), \tag{2.2}$$

onde,

$$g(\ell/x_i, \theta_1^{(k)}) = \frac{p_\ell^{(k)} g_\ell(x_i; \theta_\ell^{(k)})}{\sum_{\ell=1}^{2} p_\ell^{(k)} g_\ell(x_i; \theta_\ell^{(k)})}, \quad \ell = 1, 2, \tag{2.3}$$

e na $(k+1)$-ésima iteração do passo-M, a atualização das estimativas de $\theta_\ell^{(k+1)}$, $\ell = 1, 2$ são obtidas resolvendo as equações

$$\sum_{i=1}^{N} g(\ell/x_i, \theta_\ell^{(k)}) \frac{\partial}{\partial \theta_\ell^{(k)}} \log(g_\ell(x_i; \theta_\ell^{(k)})) = 0. \tag{2.4}$$

Na estimação de Θ do modelo (2.1), como cada componente da mistura $g_\ell(.; \theta_\ell)$ deve ser de uma das famílias, $\mathcal{G}_{\gamma^+}$, $\mathcal{G}_{\gamma^-}$, $\mathcal{G}_0$, há seis possíveis casos para aplicar. Os três primeiros casos descritos a seguir correspondem a mistura de componentes da mesma família e os outros três casos seguintes correspondem a mistura de componentes das diferentes famílias.

Caso 1. (Mistura de componentes em $\mathcal{G}_0$) As componentes $g_1(.; \theta_1)$ e $g_2(.; \theta_2)$ no modelo (2.1) são da família $\mathcal{G}_0$, onde $\theta_\ell = (\sigma_\ell, \mu_\ell)$, $\ell = 1, 2$, e $\Theta = (p_1, \theta_1, \theta_2)$.

No passo-E, as atualizações de p_1^{k+1} são obtidas conforme (2.2). As atualizações de $\sigma_\ell^{(k+1)}$, $\ell = 1, 2$ são obtidas ao resolver, utilizando o método de Newton-Raphson, a seguinte equação:

$$\sum_{i=1}^{N} g(\ell/x_i, \theta_\ell^{(k)}) \left\{ -(\sigma_\ell^{(k)})^{-1} + \frac{x_i - \mu_\ell^{(k)}}{(\sigma_\ell^{(k)})^2} - \frac{x_i - \mu_\ell^{(k)}}{(\sigma_\ell^{(k)})^2} \exp\left[-\frac{x_i - \mu^{(k)}}{\sigma_\ell^{(k)}} \right] \right\} = 0. \tag{2.5}$$

Para as atualizações de $\mu_\ell^{(k+1)}$, $\ell = 1, 2$ usamos a fórmula fechada

$$\mu_\ell^{(k+1)} = \sigma_\ell^{(k)} \log \left[\frac{\sum_{i=1}^{N} g(\ell/x_i, \theta_\ell^{(k)})}{\sum_{i=1}^{N} g(\ell/x_i, \theta_\ell^{(k)}) \exp(-x_i/\sigma_\ell^{(k)})} \right]. \tag{2.6}$$

Caso 2. (Mistura de componentes em $\mathcal{G}_+$). As componentes $g_1(.;\theta_1)$ e $g_2(.;\theta_2)$ no modelo (2.1) são da família $\mathcal{G}_+$, onde $\theta_\ell = (\gamma_\ell, \sigma_\ell, \mu_\ell)$, $\ell = 1, 2$, e $\Theta = (p_1, \theta_1, \theta_2)$.

Quando $\mu_1 - \frac{\sigma_1}{\gamma_1} = \mu_2 - \frac{\sigma_2}{\gamma_2}$, no passo-E, as atualizações de p_1^{k+1} são obtidas conforme (2.2) e (2.3). Porém, quando $\mu_1 - \frac{\sigma_1}{\gamma_1} < \mu_2 - \frac{\sigma_2}{\gamma_2}$, as atualizações de p_1^{k+1} são obtidas de (2.2), onde (2.3) é substituido por

$$g(1/x_i, \theta^{(k)}) = \begin{cases} \dfrac{p_1^{(k)} g_1(x; \theta_1^{(k)})}{\sum_{\ell=1}^{2} p_\ell^{(k)} g_\ell(x; \theta_\ell^{(k)})}, & x_i \geq \mu_2 - \dfrac{\sigma_2}{\gamma_2} \\ 1, & \mu_1 - \dfrac{\sigma_1}{\gamma_1} \leq x_i < \mu_2 - \dfrac{\sigma_2}{\gamma_2} \end{cases} \tag{2.7}$$

e

$$g(2/x_i, \theta^{(k)}) = \begin{cases} \dfrac{p_2^{(k)} g_2(x; \theta_2^{(k)})}{\sum_{\ell=1}^{2} p_\ell^{(k)} g_\ell(x; \theta_\ell^{(k)})}, & x_i \geq \mu_2 - \dfrac{\sigma_2}{\gamma_2} \\ 0, & \mu_1 - \dfrac{\sigma_1}{\gamma_1} \leq x_i < \mu_2 - \dfrac{\sigma_2}{\gamma_2}. \end{cases} \tag{2.8}$$

No passo-M, as atualizações de $\gamma_\ell^{(k+1)}$, $\sigma_\ell^{(k+1)}$, e $\mu_\ell^{(k+1)}$, $\ell = 1, 2$ são obtidas ao resolver as equações

$$\sum_{i=1}^{N} g(\ell/x_i, \theta_\ell^{(k)}) \times (\gamma_\ell^{(k)})^{-2} \log\left[1 + \frac{\gamma_\ell^{(k)}}{\sigma_\ell^{(k)}}(x_i - \mu_\ell^{(k)})\right] \times \left\{1 - \left[1 + \frac{\gamma_\ell^{(k)}}{\sigma_\ell^{(k)}}(x_i - \mu_\ell^{(k)})\right]^{-1/\gamma_\ell^{(k)}}\right\} = 0, \tag{2.9}$$

$$\sum_{i=1}^{N} g(\ell/x_i, \theta_\ell^{(k)}) \times (x_i - \mu_\ell^{(k)})\left[1 + \frac{\gamma_\ell^{(k)}}{\sigma_\ell^{(k)}}(x_i - \mu_\ell^{(k)})\right]^{-1} \times \left\{1 + \gamma_\ell^{(k)} - \left[1 + \frac{\gamma_\ell^{(k)}}{\sigma_\ell^{(k)}}(x_i - \mu_\ell^{(k)})\right]^{-1/\gamma_\ell^{(k)}}\right\} = 0, \tag{2.10}$$

e

$$\sum_{i=1}^{N} g(\ell/x_i, \theta_\ell^{(k)}) \times \left[1 + \frac{\gamma_\ell^{(k)}}{\sigma_\ell^{(k)}}(x_i - \mu_\ell^{(k)})\right]^{-1} \times \left\{1 + \gamma_\ell^{(k)} - \left[1 + \frac{\gamma_\ell^{(k)}}{\sigma_\ell^{(k)}}(x_i - \mu_\ell^{(k)})\right]^{-1/\gamma_\ell^{(k)}}\right\} = 0, \tag{2.11}$$

respectivamente.

Caso 3. (Mistura de componentes em $\mathcal{G}_-$). As componentes $g_1(.;\theta_1)$ e $g_2(.;\theta_2)$ no modelo (2.1) são da família $\mathcal{G}_-$, onde $\theta_\ell = (\gamma_\ell, \sigma_\ell, \mu_\ell)$, $\ell = 1, 2$, e $\Theta = (p_1, \theta_1, \theta_2)$.

Quando $\mu_1 - \frac{\sigma_1}{\gamma_1} = \mu_2 - \frac{\sigma_2}{\gamma_2}$, no passo-E, as atualizações de p_1^{k+1} são obtidas conforme (2.2) e (2.3). Porém, quando $\mu_1 - \frac{\sigma_1}{\gamma_1} < \mu_2 - \frac{\sigma_2}{\gamma_2}$, em vez de (2.3) utilizamos

$$g(1/x_i, \theta^{(k)}) = \begin{cases} \dfrac{p_1^{(k)} g_1(x;\theta_1^{(k)})}{\sum_{\ell=1}^{2} p_\ell^{(k)} g_\ell(x;\theta_\ell^{(k)})}, & x_i \le \mu_1 - \dfrac{\sigma_1}{\gamma_1} \\ 1, & \mu_1 - \dfrac{\sigma_1}{\gamma_1} < x_i \le \mu_2 - \dfrac{\sigma_2}{\gamma_2} \end{cases} \tag{2.12}$$

e

$$g(2/x_i, \theta^{(k)}) = \begin{cases} \dfrac{p_2^{(k)} g_2(x;\theta_2^{(k)})}{\sum_{\ell=1}^{2} p_\ell^{(k)} g_\ell(x;\theta_\ell^{(k)})}, & x_i \le \mu_1 - \dfrac{\sigma_1}{\gamma_1} \\ 0, & \mu_1 - \dfrac{\sigma_1}{\gamma_1} < x_i \le \mu_2 - \dfrac{\sigma_2}{\gamma_2}. \end{cases} \tag{2.13}$$

No passo-M, as atualizações de $\gamma_\ell^{(k+1)}$, $\sigma_\ell^{(k+1)}$, e $\mu_\ell^{(k+1)}$, $\ell = 1, 2$ são obtidas ao resolver as equações (2.9), (2.10), e (2.11), respectivamente.

Caso 4. (Mistura de componentes em $\mathcal{G}_+ \bigcup \mathcal{G}_0$). As componentes no modelo (2.1) são $g_1(x;\theta_1) \in \mathcal{G}_+$ e $g_2(x;\theta_2) \in \mathcal{G}_0$, onde $\theta_1 = (\gamma_1, \sigma_1, \mu_1)$, e $\theta_1 = (\sigma_2, \mu_2)$ $\ell = 1, 2$. Então $\Theta = (p_1, \gamma_1, \sigma_1, \mu_1, \sigma_2, \mu_2)$.

No passo-E, as atualizações de p_1^{k+1} são obtidas conforme (2.2), onde

$$g(1/x_i, \theta^{(k)}) = \begin{cases} \dfrac{p_1^{(k)} g_1(x;\theta_1^{(k)})}{\sum_{\ell=1}^{2} p_\ell^{(k)} g_\ell(x;\theta_\ell^{(k)})}, & x_i \ge \mu_1 - \dfrac{\sigma_1}{\gamma_1} \\ 0, & x_i < \mu_1 - \dfrac{\sigma_1}{\gamma_1} \end{cases} \tag{2.14}$$

e

$$g(2/x_i, \theta^{(k)}) = \begin{cases} \dfrac{p_2^{(k)} g_2(x;\theta_2^{(k)})}{\sum_{\ell=1}^{2} p_\ell^{(k)} g_\ell(x;\theta_\ell^{(k)})}, & x_i \ge \mu_1 - \dfrac{\sigma_1}{\gamma_1} \\ 1, & x_i < \mu_1 - \dfrac{\sigma_1}{\gamma_1}. \end{cases} \tag{2.15}$$

No passo-M, as atualizações de $\gamma_1^{(k+1)}$, $\sigma_1^{(k+1)}$, e $\mu_1^{(k+1)}$ são obtidas ao resolver as equações (2.9), (2.10), e (2.11), respectivamente. Para as atualizações de $\sigma_2^{(k+1)}$ e $\mu_2^{(k+1)}$ utilizamos (2.5) e (2.6), respectivamente.

Caso 5. (Mistura de componentes em $\mathcal{G}_- \bigcup \mathcal{G}_0$). As componentes no modelo (2.1) são $g_1(x;\theta_1) \in \mathcal{G}_-$ e $g_2(x;\theta_2) \in \mathcal{G}_0$, onde $\theta_1 = (\gamma_1, \sigma_1, \mu_1)$, e $\theta_1 = (\sigma_2, \mu_2)$ $\ell = 1, 2$. Então $\Theta = (p_1, \gamma_1, \sigma_1, \mu_1, \sigma_2, \mu_2)$.

No passo-E, as atualizações de p_1^{k+1} são obtidas conforme (2.2), onde

$$g(1/x_i, \theta^{(k)}) = \begin{cases} \dfrac{p_1^{(k)} g_1(x;\theta_1^{(k)})}{\sum_{\ell=1}^{2} p_\ell^{(k)} g_\ell(x;\theta_\ell^{(k)})}, & x_i \leq \mu_1 - \dfrac{\sigma_1}{\gamma_1} \\ 0, & x_i > \mu_1 - \dfrac{\sigma_1}{\gamma_1} \end{cases} \tag{2.16}$$

e

$$g(2/x_i, \theta^{(k)}) = \begin{cases} \dfrac{p_2^{(k)} g_2(x;\theta_2^{(k)})}{\sum_{\ell=1}^{2} p_\ell^{(k)} g_\ell(x;\theta_\ell^{(k)})}, & x_i \leq \mu_1 - \dfrac{\sigma_1}{\gamma_1} \\ 1, & x_i > \mu_1 - \dfrac{\sigma_1}{\gamma_1}. \end{cases} \tag{2.17}$$

No passo-M, as atualizações de $\gamma_1^{(k+1)}$, $\sigma_1^{(k+1)}$, $\mu_1^{(k+1)}$ $\sigma_2^{(k+1)}$, e $\mu_2^{(k+1)}$ são obtidas como no Caso 4.

Caso 6. (Mistura de componentes em $\mathcal{G}_- \bigcup \mathcal{G}_+$). Neste caso, $g_1(x;\theta_1) \in \mathcal{G}_+$ e $g_2(x;\theta_2) \in \mathcal{G}_-$ onde $\theta_\ell = (\gamma_\ell, \sigma_\ell, \mu_\ell)$, $\ell = 1, 2$ and $\Theta = (p_1, \gamma_1, \sigma_1, \mu_1, \gamma_2, \sigma_2, \mu_2)$.

Consideramos os dois possíveis sub casos:

Quando $\mu_1 - \frac{\sigma_1}{\gamma_1} < \mu_2 - \frac{\sigma_2}{\gamma_2}$

$$g(1/x_i, \theta^{(k)}) = \begin{cases} \dfrac{p_1^{(k)} g_1(x;\theta_1^{(k)})}{\sum_{\ell=1}^{2} p_\ell^{(k)} g_\ell(x;\theta_\ell^{(k)})}, & \mu_1 - \dfrac{\sigma_1}{\gamma_1} \leq x_i \leq \mu_2 - \dfrac{\sigma_2}{\gamma_2} \\ 0, & x_i < \mu_1 - \dfrac{\sigma_1}{\gamma_1} \\ 1, & x_i > \mu_2 - \dfrac{\sigma_2}{\gamma_2} \end{cases} \tag{2.18}$$

e

$$g(2/x_i, \theta^{(k)}) = \begin{cases} \dfrac{p_2^{(k)} g_2(x;\theta_2^{(k)})}{\sum_{\ell=1}^{2} p_\ell^{(k)} g_\ell(x;\theta_\ell^{(k)})}, & \mu_1 - \dfrac{\sigma_1}{\gamma_1} \leq x_i \leq \mu_2 - \dfrac{\sigma_2}{\gamma_2} \\ 1, & x_i < \mu_1 - \dfrac{\sigma_1}{\gamma_1} \\ 0, & x_i > \mu_2 - \dfrac{\sigma_2}{\gamma_2}. \end{cases} \tag{2.19}$$

E quando $\mu_1 - \frac{\sigma_1}{\gamma_1} \geq \mu_2 - \frac{\sigma_2}{\gamma_2}$

$$g(1/x_i, \theta^{(k)}) = \begin{cases} 1, & x_i > \mu_1 - \dfrac{\sigma_1}{\gamma_1} \\ 0, & \text{c.c.} \end{cases} \tag{2.20}$$

and

$$g(2/x_i, \theta^{(k)}) = \begin{cases} 1, & x_i \leq \mu_2 - \dfrac{\sigma_2}{\gamma_2} \\ 0, & \text{c.c..} \end{cases} \tag{2.21}$$

Neste caso as atualizações de p_1^{k+1}, são obtidas conforme $\gamma_\ell^{(k+1)}$, $\sigma_\ell^{(k+1)}$, e $\mu_\ell^{(k+1)}$, $\ell = 1, 2$ são obtidas ao resolver as equações (2.2), (2.9), (2.10), e (2.11), respectivamente.

3 SIMULAÇÃO

Nesta seção, testamos via simulação as estimativas de Θ do modelo (2.1) para a mistura de componetes de diferentes famílias, o que corresponde trabalhar com os casos (4)-(6), pois os casos (1)-(3) tratam de componentes da mesma família que são mais facilmente calculados. Para isto, seguimos o seguinte procedimento:

1. Geramos amostras aleatórias de tamanho $n = 100$ para cada escolha do vetor Θ.
2. A amostra aleatória da variável aleatória X cuja densidade é a mistura (2.1) é gerada como segue:
   (a) Gerar duas variáveis uniformes $u1$ e $u2$.
   (b) Se $u_1 < p_1$, então usamos u_2 para gerar um valor x da v.a. X de (2.1), onde $x = G_1^{-1}(u_2)$ e G_1 é a distribuição acumulada g_1.
   (c) Se $u_1 \geq p_1$, então usamos u_2 para gerar um valor x da v.a. X, onde $x = G_2^{-1}(u_2)$ e G_2 é a distribuição acumulada g_2.
3. Calculamos iterativamente os estimadores de Θ utilizando as expressões de $g(\ell/x_i, \theta^{(k)})$, $\ell = 1, 2$ mostradas nos casos (4)-(6).
4. A regra de parada que utilizamos no Algoritmo EM é $\log(L[\Theta^{(k+1)}]) - \log(L[\Theta^{(k)}]) < n10^{-5}$.
5. Obtemos estimativas de Θ utilizando 100 amostras de tamanho 100, então calculamos a média e o erro quadrático médio de Θ por Monte Carlo. Tais resultados são mostrados nas Tabelas 2 e 3.

Definimos vários experimentos, valores de Θ, para testar o comportamento dos estimadores do algoritmo EM. Esses experimentos mostrados na Tabela 1 foram escolhidos de tal maneira a contemplar os casos (4)-(6).

Os experimentos $\Theta_{4,i}$, $\Theta_{5,i}$, $\Theta_{6,ia}$, e $\Theta_{6,ib}$, $i = 1, 2, 3$, correspondem aos casos (4), (5), e (6), respectivamente.

Nas Figuras 1-4, comparamos a densidades de uma amostra $h(x; \Theta)$ com a densidade $h(x; \hat{\Theta})$, onde os valores de $\hat{\Theta}$ são as médias de $\hat{\Theta}$, mostradas na Tabela 2. Por essas figuras podemos concluir que o algoritmo EM teve um bom desempenho, pois a densidade $h(x; \hat{\Theta})$ teve um bom ajuste, em quase todos os experimentos exceto no caso $\Theta_{6,2b}$. Os resultados do erro quadrático médio das estimativas são mostrados na Tabela 3, os quais confirmam que os resultados da Tabela 1 são bons estimadores dos parâmetros em análise. Novamente apenas o caso $\Theta_{6,2b}$ não foi muito bom.

Tabela 1: Experimentos.

Θ	p	γ_1	γ_2	σ_1	σ_2	μ_1	μ_2
$\Theta_{4,1}$	0.4	1	0	1	2	1	0
$\Theta_{4,2}$	0.4	0.5	0	0.5	3	–1	2
$\Theta_{4,3}$	0.6	2	0	2	1	1	0
$\Theta_{5,1}$	0.5	–0.5	0	1	1	–2	0
$\Theta_{5,2}$	0.3	–0.5	0	0.5	2.5	2	0
$\Theta_{5,3}$	0.4	–0.4	0	0.5	2.5	1	3
$\Theta_{6,1a}$	0.1	0.5	–0.5	1	1	–1	1
$\Theta_{6,2a}$	0.8	2	–0.5	3	1	0	1
$\Theta_{6,3a}$	0.4	1	–0.5	1	3	1	0
$\Theta_{6,1b}$	0.8	2	–0.5	3	1	7	1
$\Theta_{6,2b}$	0.1	0.5	–0.5	1	1	–1	1
$\Theta_{6,3b}$	0.4	0.5	–0.5	0.5	0.5	4	–1

Tabela 2: Média de $\hat{\Theta}$.

$\hat{\Theta}$	n	$\hat{p}$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\sigma}_1$	$\hat{\sigma}_2$	$\hat{\mu}_1$	$\hat{\mu}_2$
$\hat{\Theta}_{4,1}$	100	0.3993	1.0779	–	0.9979	1.9997	0.5626	–0.0032
$\hat{\Theta}_{4,2}$	100	0.3992	0.9209	–	0.4962	2.9999	–1.5485	2.0073
$\hat{\Theta}_{4,3}$	100	0.5996	2.0006	–	1.9647	0.9992	0.8902	–0.0046
$\hat{\Theta}_{5,1}$	100	0.4926	–0.5165	–	1.1033	0.9993	–2.0015	0.0317
$\hat{\Theta}_{5,2}$	100	0.2986	–0.5315	–	0.4936	2.4999	1.9978	0.0411
$\hat{\Theta}_{5,3}$	100	0.3926	–0.5293	–	0.4995	2.4999	0.9991	3.0634
$\hat{\Theta}_{6,1a}$	100	0.0974	1.3900	–0.5449	0.9551	1.0775	–1.3823	0.9976
$\hat{\Theta}_{6,2a}$	100	0.2993	2.0003	–0.5370	2.9750	1.0706	–0.1539	–0.0021
$\hat{\Theta}_{6,3a}$	100	0.3993	1.0955	–0.5325	0.9980	4.1617	0.5645	–0.0006
$\hat{\Theta}_{6,1b}$	100	0.8053	2.0008	–0.5842	2.9542	1.0001	6.8273	0.9994
$\hat{\Theta}_{6,2b}$	100	0.4958	0.9584	–0.5475	0.9869	1.0342	3.9027	0.9986
$\hat{\Theta}_{6,3b}$	100	0.3963	0.9900	–0.5419	0.4930	0.4993	3.4549	–1.0027

Tabela 3: Erro quadrático médio de $\hat{\Theta}$.

$\hat{\Theta}$	n	$\hat{p}$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\sigma}_1$	$\hat{\sigma}_2$	$\hat{\mu}_1$	$\hat{\mu}_2$
$\hat{\Theta}_{4,1}$	100	0.0006	0.0468	–	0.0000	0.0000	0.1933	0.0455
$\hat{\Theta}_{4,2}$	100	0.0012	0.2394	–	0.0010	0.0000	0.3037	0.0886
$\hat{\Theta}_{4,3}$	100	0.0009	0.0000	–	0.0023	0.0000	0.0127	0.0153
$\hat{\Theta}_{5,1}$	100	0.0017	0.0066	–	0.0105	0.0000	0.0000	0.0070
$\hat{\Theta}_{5,2}$	100	0.0011	0.0031	–	0.0000	0.0000	0.0000	0.0671
$\hat{\Theta}_{5,3}$	100	0.0013	0.0089	–	0.0000	0.0000	0.0000	0.0643
$\hat{\Theta}_{6,1a}$	100	0.0004	0.3536	0.0093	0.0197	0.0098	0.1675	0.0000
$\hat{\Theta}_{6,2a}$	100	0.0017	0.0000	0.0210	0.0049	0.0449	0.0410	0.0000
$\hat{\Theta}_{6,3a}$	100	0.0008	0.0460	0.0032	0.0000	0.2501	0.1919	0.0000
$\hat{\Theta}_{6,1b}$	100	0.0017	0.0000	0.0210	0.0049	0.0449	0.0410	0.0000
$\hat{\Theta}_{6,2b}$	100	0.0025	0.2527	0.0060	0.0146	0.0351	1.2141	0.0000
$\hat{\Theta}_{6,3b}$	100	0.0023	0.2966	0.0044	0.0021	0.0000	0.3004	0.0000

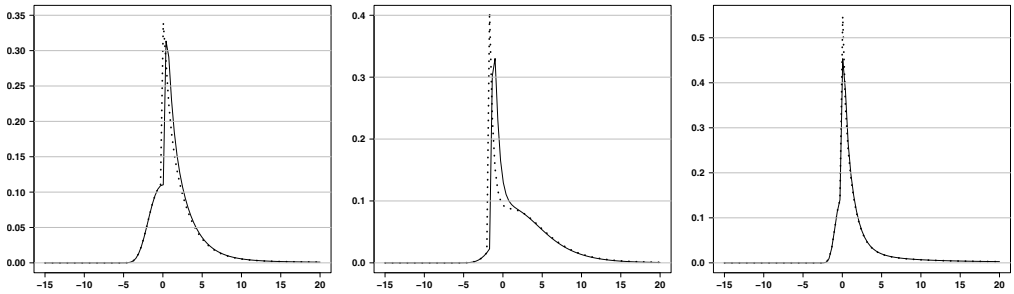

Figura 1: Gráfico das densidades $h(x;\Theta)$ (curva contínua) e $h(x;\hat{\Theta})$ (curva pontilhada) para $\Theta_{4,1}$, $\Theta_{4,2}$ and $\Theta_{4,3}$, de esquerda para direita.

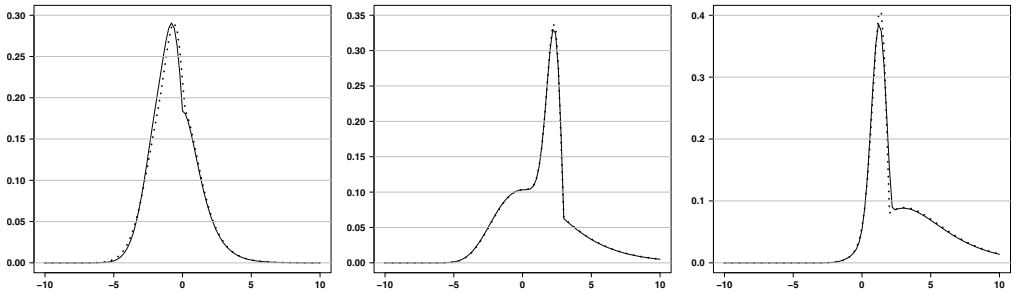

Figura 2: Gráfico das densidades $h(x;\Theta)$ (curva contínua) e $h(x;\hat{\Theta})$ (curva pontilhada) para $\Theta_{5,1}$, $\Theta_{5,2}$ e $\Theta_{5,3}$.

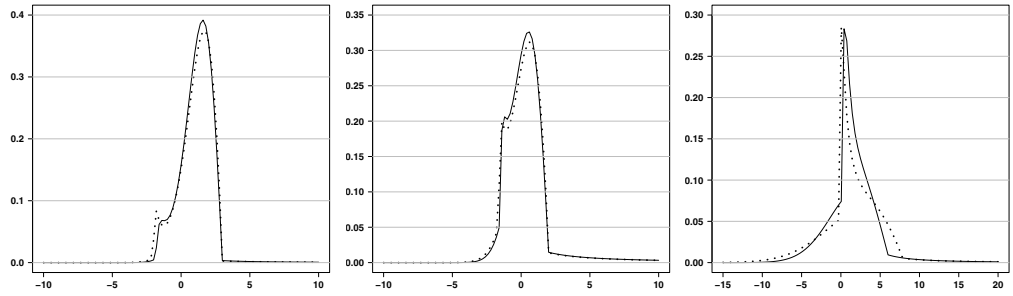

Figura 3: Gráfico das densidades $h(x;\Theta)$ (curva contínua) e $h(x;\hat{\Theta})$ (curva pontilhada) para $\Theta_{6,1a}$, $\Theta_{6,2a}$ e $\Theta_{6,3a}$.

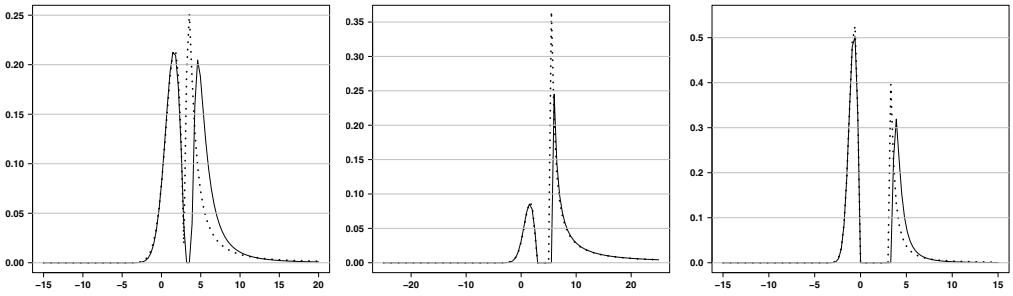

Figura 4: Gráfico das densidades $h(x;\Theta)$ (curva contínua) e $h(x;\hat{\Theta})$ (curva pontilhada) para $\Theta_{6,1b}$, $\Theta_{6,2b}$ e $\Theta_{6,3b}$.

4 APLICAÇÃO

Para demonstrar a aplicabilidade da mistura (2.1), ajustamos o logaritmo do consumo per capita de petróleo em 135 países no ano de 2001. Os dados foram obtidos no endereço `http://data.worldbank.org/indicator/EG.USE.PCAP.KG.OE`.

Para esses dados, os valores dos parâmetros estimados do modelo (2.1), pelo algoritmo EM, foram $p = 0.5$, $\sigma_1 = 0.5155666$, $\sigma_2 = 0.5985117$, $\mu_1 = 6.414$, $\mu_2 = 8.235$. O que indica a mistura de duas componentes da família $\mathcal{G}_0$. Na Figura 5 mostra-se a densidade ajustada aos dados.

O QQ-Plot da Figura 6, mostra que o ajustamento da mistura de duas distribuições com densidade Gumbel é adequado.

5 CONCLUSÕES

Neste artigo, os estimadores dos parâmetros do modelo (2.1) são dados em seis possíveis casos, usando o Algoritmo EM. O método de estimação é testado com doze experimentos. Apenas para um desses experimentos o erro quadrático médio não foi suficientemente pequeno. Isso pode ter

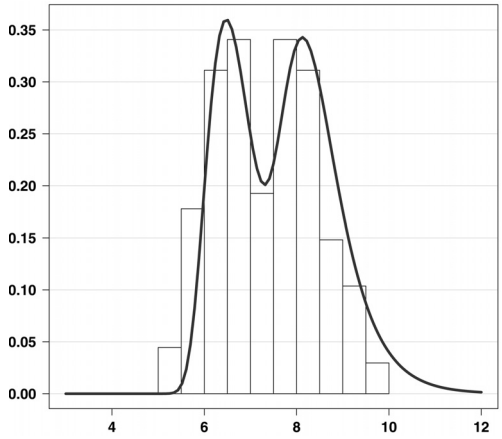

Figura 5: Histograma e função densidade de probabilidade ajustada aos dados de petróleo para mistura de duas componentes Gumbel com $\hat{\Theta} = (0.5016, 0.5123, 0.5954, 6.4758, 8.2016)$.

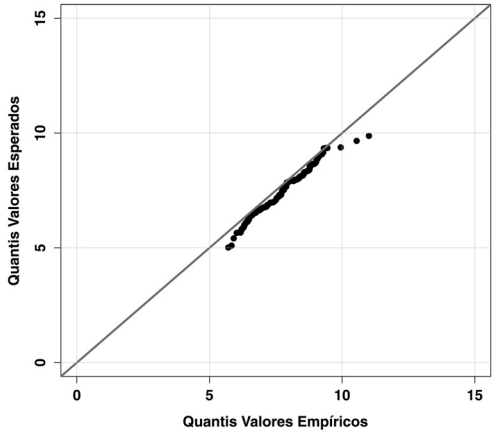

Figura 6: QQ-Plot dados de petróleo para mistura de duas componentes Gumbel com com $\hat{\Theta} = (0.5016, 0.5123, 0.5954, 6.4758, 8.2016)$.

ocorrido pelo fato dos parâmetros de locação estarem muito distantes. Finalmente, utilizamos dados reais para mostrar uma aplicação do algoritmo EM. Não fizemos comparação de nossas estimativas com as de por exemplo Atienza-Sandoval (2007), pois os algoritmos por eles utilizados demandam um custo computacional intensivo. O tempo computacional para o cálculo de nossas estimativas foi inferior a um minuto.

AGRADECIMENTOS

Agradecemos aos revisores deste artigo por seus valiosos comentários para melhorar a primeira versão. Agradecemos também as agências de fomento à pesquisa CAPES/PROCAD e DPP/UnB pelo financiamento parcial para este trabalho.

ABSTRACT. Probabilistic models of finite mixtures with components Generalized Extremal Value (GEV) distributions are widely applied in diverse areas such as finance and hydrology. In this work, the estimators of the parameters of the GEV mixture of two components are obtained via the EM algorithm. We also present numerical examples of the behavior of the estimates obtained through of simulation, well as an application to real data.

Keywords: GEV, finite mixture, EM algorithm.

REFERÊNCIAS

[1] C. Escalante-Sandoval. A Mixed distribution with EV1 and GEV components for analyzing heterogeneous samples. *Ingeniería Investigación y Tecnología 8*, **3** (2007), 123–133.

[2] R. Kollu, S.R. Rayapudi, S. Narasimham & K.M. Pakkurthi. Mixture probability distribution functions to model wind speed distributions. *International Journal of Energy and Environmental Engineering*, (2012), 1–10.

[3] G.J. McLachlan & K.E. Basford. "Mixture models". Marcel Dekker, NewYork, (1988).

[4] G.J. McLachlan & D. Peel. "Finite mixture models". Wiley, New York, (2000).

[5] D.M. Titterington, A.M.F. Smith & U.E. Makov. "Statistical analysis of finite mixture distributions". Wiley, NewYork, (1985).