Uma abordagem bayesiana para o modelo de sobrevivência bivariado derivado da cópula AMH

Resumo

Neste trabalho, propomos um modelo derivado da cópula arquimediana de Ali-Mikhail-Haq (AMH) para modelar a dependência de dados bivariados de sobrevivência na presença de covariáveis e observações censuradas. Para fins inferenciais, realizamos uma abordagem bayesiana usando métodos Monte Carlo em Cadeias de Markov (MCMC). Algumas discussões sobre os critérios de seleção de modelos foram apresentadas. Com o objetivo de detectar observações influentes utilizamos o método bayesiano de análise de influência de deleção de casos baseado na divergência ψ. Por fim, mostramos a aplicabilidade dos modelos propostos a conjuntos de dados simulados e reais.

1 Introdução

É cada vez mais comum nos depararmos com situações em que a suposição de independência entre os tempos de sobrevivência pode não ser válida. Por exemplo, imagine que indivíduos de um estudo estão sujeitos a múltiplos eventos semelhantes, conhecidos por *eventos recorrentes* (Colosimo & Giolo, 2006), tais como ataques epiléticos ou ataques cardíacos, dentre outros. Nesses casos, mais de um tempo de sobrevivência é observado para cada indivíduo em estudo e, desse modo, é também válida a suposição que exista associação entre os tempos de um mesmo indivíduo. Além disso, eventos de tipos diferentes, tais como múltiplas sequelas em pacientes com doenças crônicas, descrevem outras situações em que a suposição de independência entre os tempos pode não ser válida.

Essas prováveis associações entre os tempos de sobrevivência são frequentemente ajustadas por meio de modelos de fragilidade, que foram propostos por Vaupel et al. (1979), em que um ou mais efeitos aleatórios, denominado fragilidade, são introduzidos na função de risco para descrever essa possível heterogeneidade entre as unidades em estudo. Além disso, nestes tipos de modelos, os tempos marginais são condicionalmente independentes dada a variável fragilidade.

No entanto, uma outra alternativa que vem sendo cada vez mais desenvolvida ultimamente para modelar a dependência entre dados multivariados, como, por exemplo, nas áreas biológicas, ciências atuariais e finanças, é o uso dos modelos de cópulas, descritas, por exemplo, em Nelsen (2006) e Joe (2014). Cópulas são funções que ligam

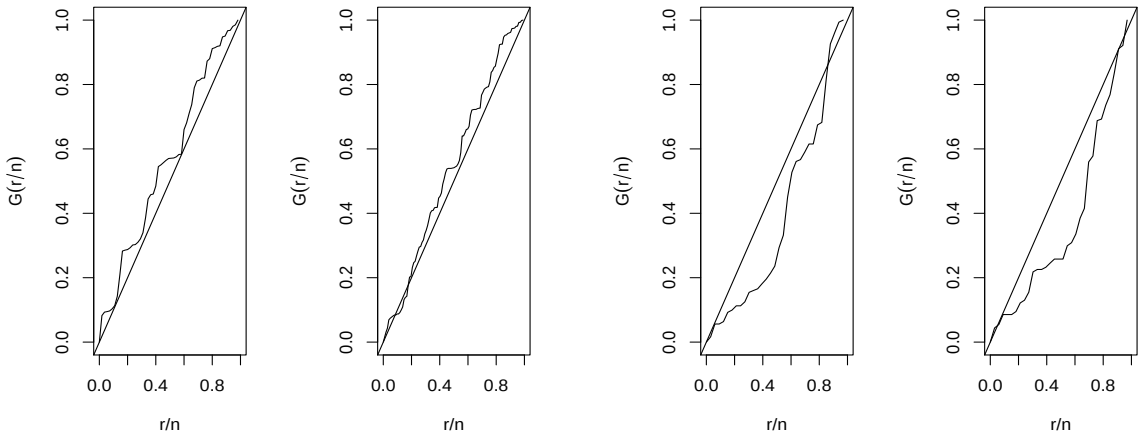
(conectam) a função distribuição conjunta com suas funções distribuição marginais univariadas. Diferentes funções cópulas representam diferentes estruturas de dependência entre as variáveis (Nelsen, 2006). O uso de cópulas permite que possamos estimar os parâmetros das distribuições marginais e, em seguida, estimar o parâmetro da função cópula até mesmo no caso em que seja difícil especificar a distribuição conjunta.

Dessa forma, os modelos de cópulas apresentam um relevante leque de aplicações, como, por exemplo, no estudo dos tempos de vida de pessoas "associadas", tais como os cônjuges, em que, de acordo com estudos, estes tempos podem apresentar "dependência" devido a condições como desastre comum, estilo de vida comum, ou a chamada "síndrome do coração partido", sendo este tipo de estudo extremamente importante para empresas de seguro de vida (Purwono, 2001). Temos também o caso em que utilizamos os modelos de cópulas para a construção da distribuição conjunta de intensidade e profundidade de precipitação, intensidade e duração da chuva, ou profundidade e duração de chuvas, que são elementos fundamentais na elaboração de um projeto hidrológico (Zhang & Singh, 2006). Há inúmeras outras situações nas quais o uso de funções cópulas é empregado, por exemplo, em ciências atuariais são utilizadas na modelagem de mortalidade e perdas (Frees et al., 2005); em finanças, na classificação de crédito e modelagem de risco (Embrechts et al., 2003); em estudos biomédicos, na modelagem de eventos correlacionados e riscos competitivos (Achcar e Boleta, 2012; Suzuki et al., 2011; Louzada et al., 2013) e, até mesmo, na sobrevivência política (Quiroz Flores, 2008).

Neste atigo, o nosso enfoque é trabalhar com dados de sobrevivência bivariados, que são os dados que caracterizam situações em que se apresentam dados censurados e se observam dois tempos de vida para um mesmo equipamento ou paciente. Por exemplo, na area médica pode ocorrer o interesse em estudar os tempos de vida de orgãos humanos pareados como rins e olhos ou o tempo entre a primeira e a segunda internação por uma determinada enfermidade, dentre outras. Por outro lado, em aplicações industriais, o estudo de dados bivariados exemplifica-se na análise de um sistema cujo o tempo de duração depende da durabilidade de dois componentes, como o tempo de vida de motores de um avião bimotor.

Sob uma abordagem bayesiana, modelamos a dependência de dados de sobrevivência bivariados na presença de covariáveis e observações censuradas por meio das cópulas Arquimedianas, em particular a cópula Ali-Mikhail-Haq (AMH) e, para ambas as distribuições marginais, assumimos a distribuição Exponencial Generalizada ou a distribuição Weibull. Para fins inferenciais, foram utilizados métodos Monte Carlo em Cadeias de Markov (MCMC). Com o objetivo de detectar observações influentes nos dados usamos o método bayesiano de análise de influência caso a caso baseado na divergência ψ. Mostramos a aplicabilidade dos modelos propostos a conjuntos de dados simulados e reais.

A escolha em trabalhar com as marginais Weibull e Exponencial Generalizada foi em razão dos TTT Plot que fizemos para os dados reais de retinopatia diabética e de insuficiência renal. Nas Figuras 1(a,b) observamos que os TTT Plot indicam uma função de risco crescente no caso dos dados de retinopatia diabética e, decrescente no caso dos dados de insuficência renal. Logo, como ambas as distribuições Weibull e Exponencial Generalizada acomodam funções de risco crescentes e decrescentes, elas foram escolhidas como nossas distribuições marginais.



(a) Dados de retinopatia diabética. (b) Dados de insuficiência renal.

Figura 1: TTT Plot

2 Uma breve introdução às funções cópulas

Nesta seção, apresentamos uma breve introdução às funções cópulas bem como alguns resultados básicos referentes à elas, sendo que, o uso destas funções permite a construção da distribuição conjunta de variáveis com as distribuições marginais conhecidas.

Cópulas são funções que ligam (conectam) a função distribuição conjunta com suas funções distribuição marginais univariadas. Por outro lado, cópulas também são conceituadas como funções distribuição multivariadas cujas distribuições marginais unidimensionais são Uniformes em $(0, 1)$.

As referências básicas para o estudo destas funções são os livros de Nelsen (1999), Kolev (2006) e Joe (2014). A seguir apresentamos a definição de cópula e algumas de suas propriedades.

Definição 2.1 *Uma cópula é uma distribuição multivariada cujas marginais são Uniforme (0,1). Considere o vetor aleatório $U = (U_1, \ldots, U_n) \in I^n$ com cópula n-dimensional C, temos:*

$$C(u_1, \ldots, u_n; \phi) = P(U_1 \leq u_1, \ldots, U_n \leq u_n; \phi)$$

em que ϕ é o parâmetro associado à função cópula.

O teorema que será apresentado a seguir, conhecido como teorema de Sklar, consiste no resultado mais importante referente à teoria e aplicações de cópulas, em que, à partir deste, temos que uma cópula conecta as distribuições marginais univariadas formando uma distribuição multivariada; ou então que uma função distribuição multivariada pode ser decomposta nas marginais univariadas e na estrutura de dependência dada pela cópula.

Teorema 2.1 *Seja H uma função de distribuição conjunta com marginais $F_1(t_1), \ldots, F_n(t_n)$. Então existe uma cópula n-dimensional C tal que:*

$$H(t_1, \ldots, t_n; \phi) = C(F_1(t_1), \ldots, F_n(t_n); \phi).$$

Se $F_1(t_1), \ldots, F_n(t_n)$ são todas contínuas, então C é única.

Reciprocamente, se C é uma cópula n-dimensional e $F_1(t_1), \ldots, F_n(t_n)$ são funções de distribuição, então a função H é uma função de distribuição conjunta n-dimensional.

Logo, podemos estabelecer que a cópula C é uma função que liga a função distribuição conjunta a suas marginais. O nome cópula foi escolhido para enfatizar a maneira como a cópula une uma função distribuição conjunta às suas marginais univariadas.

Atualmente, a classe de cópulas Arquimedianas é a mais utilizada na prática. A representação da cópula Arquimediana permite reduzir o estudo de cópula multivariada ao estudo de uma função univariada φ, comumente chamada de gerador de uma cópula Arquimediana. Além disso, a classe de cópulas Arquimedianas é bastante flexível, permitindo a modelagem de diversas formas de dependência, incluindo assimetria e dependência nas extremidades. Algumas das principais cópulas Arquimedianas são a cópula de Clayton, a cópula de Frank, a cópula de Gumbel-Hougard e a cópula de Ali-Mikhail-Haq (AMH), sendo esta última a cópula com a qual trabalharemos.

Uma distribuição bivariada pertence à família de cópulas Arquimedianas se tem a seguinte representação:

$$C_\phi(u, v) = \varphi(\varphi(u)^{-1} + \varphi(v)^{-1}), \;\; 0 \leq u, v \leq 1 \tag{1}$$

em que $0 < \varphi < 1$, $\varphi(0) = 1$, $\varphi' < 0$, $\varphi'' > 0$ e ϕ é o parâmetro de dependência da cópula.

Todas as cópulas Arquimedianas usualmente encontradas possuem expressões com forma fechada.

Considere X e Y variáveis aleatórias independentes, $F_X(x)$ e $F_Y(y)$ distribuições acumuladas em X e Y, respectivamente e $H(x, y)$ como definida no Teorema 2.1, temos que:

$$\frac{1 - H(x, y)}{H(x, y)} = \frac{1 - F_X(x)}{F_X(x)} + \frac{1 - F_Y(y)}{F_Y(y)} + \frac{(1 - F_X(x))}{F_X(x)} \frac{(1 - F_Y(y))}{F_Y(y)}. \tag{2}$$

Ali et al. (1978) propuseram o estudo de distribuições bivariadas para os quais os excedentes de sobrevivência relativos a (X, Y), X e Y satisfizessem:

$$\frac{1 - H(x, y)}{H(x, y)} = \frac{1 - F_X(x)}{F_X(x)} + \frac{1 - F_Y(y)}{F_Y(y)} + (1 - \phi) \frac{(1 - F_X(x))}{F_X(x)} \frac{(1 - F_Y(y))}{F_Y(y)}, \tag{3}$$

para alguma constante ϕ.

Supondo que a igualdade (3) é sempre satisfeita para X e Y contínuas, com $-1 \leq \phi \leq 1$, segue do Teorema de Sklar o fato da família de cópulas Ali-Mikhail-Haq ser definida por:

$$C(u, v) = \frac{uv}{1 - \phi(1 - u)(1 - v)}, \; -1 \leq \phi \leq 1. \tag{4}$$

A função geradora da cópula AMH é dada por $\varphi_\phi(t) = \ln \frac{1 - \phi(1-t)}{t}$.

Após algumas manipulações algébricas, a igualdade dada em (3) pode ser escrita da seguinte forma:

$$1 + (1-\phi)\frac{1-H(x,y)}{H(x,y)} = [1 + (1-\phi)\frac{(1-F_X(x))}{F_X(x)}][1 + (1-\phi)\frac{(1-F_Y(y))}{F_Y(y)}], \tag{5}$$

ou seja, $h(H(x,y)) = h(F_X(x))h(F_Y(y))$, em que $h(t) = 1 + (1-\phi)\frac{(1-t)}{t}$, para $-1 \leq \phi < 1$. Além disso, note que podemos escrever $h(t) = \frac{1-\phi(1-t)}{t}$, para $-1 \leq \phi < 1$.

Definindo a função φ tal que $\varphi(t) = \ln h(t)$, podemos escrever H como a soma das marginais F_X e F_Y:

$$\varphi(H(x,y)) = \varphi(F_X(x)) + \varphi(F_Y(y)),\ \forall\ (x,y) \in \mathbb{R}^2.$$

Para a cópula Ali-Mikhail-Haq, C_ϕ^{AMH} em que $-1 \leq \phi < 1$, podemos escrever:

$$\varphi(C_\phi^{AMH}(u,v)) = \varphi(u) + \varphi(v),$$

demonstrando, assim, o fato da cópula AMH ser Arquimediana.

2.1 Inferência

Sejam (T_{i1}, T_{i2}) e (C_{i1}, C_{i2}) os i-ésimos tempos de vida e de censura bivariados, para $i = 1, \ldots, n$. Suponha que os tempos (T_{i1}, T_{i2}) e (C_{i1}, C_{i2}) são independentes. Para cada indivíduo i, as quantidades individuais são representadas pelas variáveis aleatórias $t_{ij} = \min(T_{ij}, C_{ij})$ e $\delta_{ij} = I(t_{ij} = T_{ij})$ que denota o indicador de falha, em que $j = 1, 2$.

Considere $S(t_1|\boldsymbol{\gamma}_1)$ e $S(t_2|\boldsymbol{\gamma}_2)$ as funções de sobrevivência de T_{i1} e T_{i2}, respectivamente, sendo $\boldsymbol{\gamma}_1$ e $\boldsymbol{\gamma}_2$ vetores de parâmetros de q_1 e q_2 elementos associados a cada uma das distribuições marginais.

Assumindo a função de sobrevivência bivariada baseada na cópula de AMH e tomando as funções de sobrevivência $u = S(t_1|\boldsymbol{\gamma}_1)$ e $v = S(t_2|\boldsymbol{\gamma}_2)$, temos que:

$$S(t_1, t_2|\phi, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2) = C_\phi(S(t_1|\boldsymbol{\gamma}_1), S(t_2|\boldsymbol{\gamma}_2)) = \frac{S(t_1|\boldsymbol{\gamma}_1)S(t_2|\boldsymbol{\gamma}_2)}{1 - \phi(1 - S(t_1|\boldsymbol{\gamma}_1))(1 - S(t_2|\boldsymbol{\gamma}_2))}. \tag{6}$$

Considerando a função de sobrevivência bivariada $S(t_1, t_2|\phi, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)$ dada em (6), a contribuição do i-ésimo indivíduo para a log-verossimilhança de $\boldsymbol{\theta} = (\phi, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)$ é dada por:

$$\begin{aligned}\ell_i(\boldsymbol{\theta}) = {} & \delta_{i1}\delta_{i2}\log\left(\frac{\partial^2 S(t_1, t_2|\boldsymbol{\theta})}{\partial t_{i1}\partial t_{i2}}\right) + \delta_{i1}(1-\delta_{i2})\log\left(\frac{-\partial S(t_1, t_2|\boldsymbol{\theta})}{\partial t_{i1}}\right) \\ & + \delta_{i2}(1-\delta_{i1})\log\left(\frac{-\partial S(t_1, t_2|\boldsymbol{\theta})}{\partial t_{i2}}\right) + (1-\delta_{i1})(1-\delta_{i2})\log S(t_1, t_2|\boldsymbol{\theta}).\end{aligned} \tag{7}$$

2.2 Funções densidades de probabilidade

Para ambas as distribuições marginais assumimos a distribuição Exponencial Generalizada ou a distribuição Weibull A seguir apresentamos algumas características que envolvem estas distribuições.

2.2.1 Distribuição Exponencial Generalizada

A distribuição Exponencial Generalizada (Gupta e Kundu, 1999) pode ser uma boa alternativa ao uso das tradicionais distribuições Gama e Weibull utilizadas na análise de dados de sobrevivência (Boleta, J. 2012).

A distribuição Exponencial Generalizada de dois parâmetros tem função densidade de probabilidade dada por:

$$f(t;\alpha,\lambda) = \alpha\lambda\left(1 - exp(-\lambda t)\right)^{\alpha-1} exp(-\lambda t),$$

em que, $t > 0$, $\alpha > 0$ e $\lambda > 0$ são os parâmetros de forma e escala, respectivamente.

As funções de sobrevivência e de risco associadas à essa densidade são dadas, respectivamente, por:

$$S(t;\alpha,\lambda) = P(T > t) = 1 - \left(1 - exp(-\lambda t)\right)^{\alpha}$$

e

$$h(t;\alpha,\lambda) = \frac{f(t;\alpha,\lambda)}{S(t;\alpha,\lambda)} = \frac{\alpha\lambda\left(1 - exp(-\lambda t)\right)^{\alpha-1} exp(-\lambda t)}{1 - \left(1 - exp(-\lambda t)\right)^{\alpha}}.$$

Abaixo apresentamos os gráficos da função densidade, da função de sobrevivência e da função de risco da distribuição Exponencial Generalizada. O objetivo destes gráficos é verificar como se comporta esta distribuição alterando-se os valores de seus parâmetros α λ.

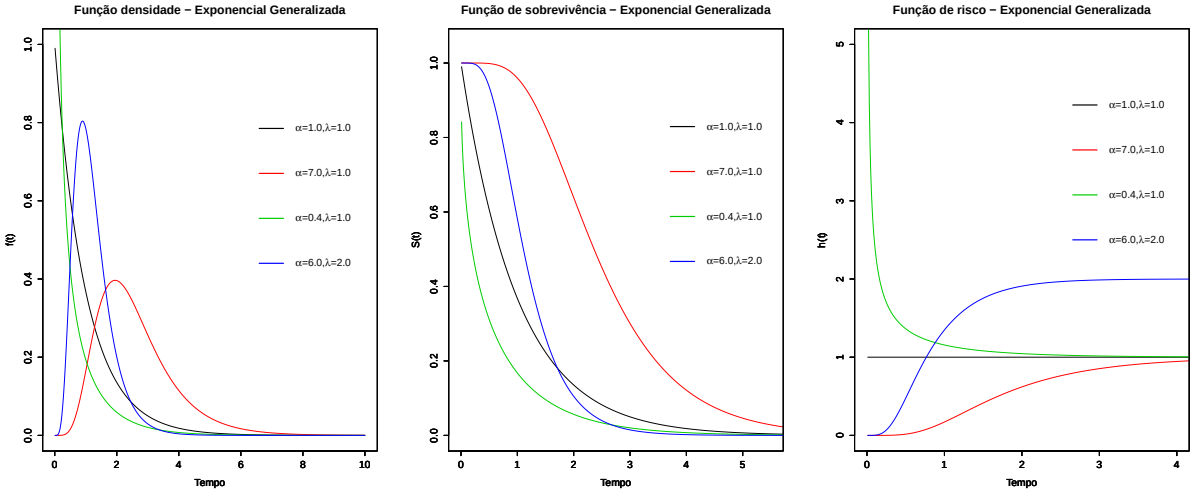


Figura 2: Gráfico da função densidade de probabilidade (esquerda), da função de sobrevivência (centro) e da função de risco (direita) para diferentes valores de α e λ.

2.2.2 Distribuição Weibull

A distribuição Weibull (Weibull, 1939) vem sendo frequentemente usada em estudos biomédicos e industriais, amplamente conhecida em virtude de sua simplicidade e flexibilidade em acomodar diferentes formas de função de risco, sendo um dos modelos mais utilizado em análise de sobrevivência.

Para uma variável aleatória T com distribuição Weibull, a função densidade de probabilidade é dada por:

$$f(t;\alpha,\lambda) = \frac{\alpha}{\lambda^{\alpha}} t^{\alpha-1} exp\left(-\left(\frac{t}{\lambda}\right)^{\alpha}\right)$$

em que $t \geq 0$, $\alpha > 0$ e $\lambda > 0$ são os parâmetros de forma e escala, respectivamente.

A função de sobrevivência do modelo Weibull é dada por:

$$S(t;\alpha,\lambda) = exp\left(-\left(\frac{t}{\lambda}\right)^{\alpha}\right)$$

e a função de risco por:

$$h(t;\alpha,\lambda) = \frac{\alpha}{\lambda^{\alpha}} t^{\alpha-1}.$$

Esta distribuição possui riscos crescentes para $\alpha > 1$, decrescentes para $\alpha < 1$ e constantes para $\alpha = 1$. Neste último caso, o modelo se reduz a distribuição Exponencial.

A Figura 3, mostra o gráfico da função densidade, da função de sobrevivência e da função de risco da distribuição Weibull para diferentes valores de α e λ.

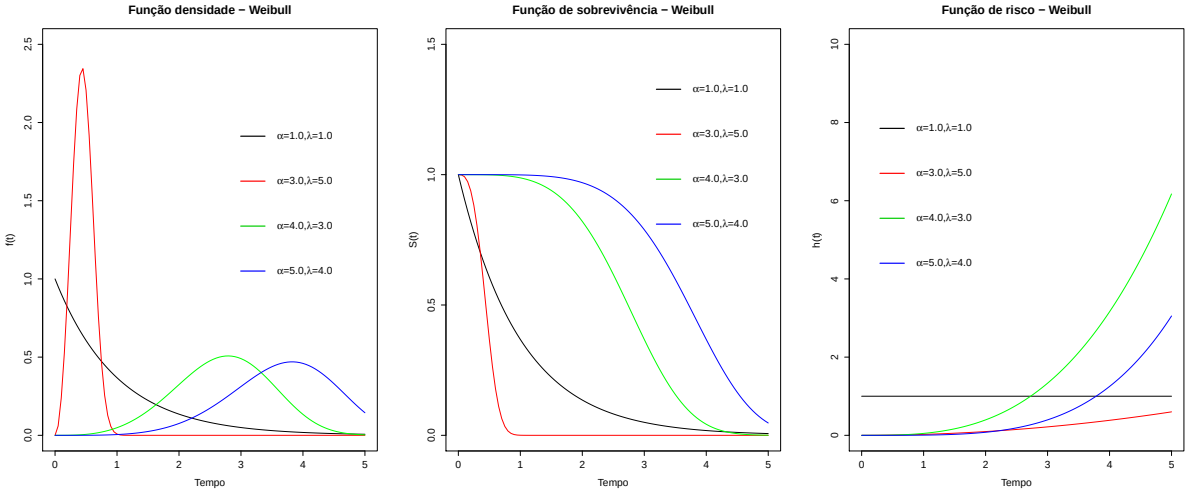


Figura 3: Gráfico da função densidade de probabilidade (esquerda), da função de sobrevivência (centro) e da função de risco (direita) para diferentes valores de α e λ.

3 Análise Bayesiana

Considere que as distribuições marginais T_j têm distribuição Exponencial Generalizada ou Weibull com parâmetros α_j e $\lambda_{ij} = exp(\beta_{0j} + \beta_{1j}x_i)$, $i = 1, ..., n$ e $j = 1, 2$.

Utilizando a metologia Bayesiana para realizarmos a inferência, assumimos que não há conhecimentos prévios dos parâmetros por meio das distribuições *a priori* não informativas. Além disso, para garantir que a distribuição *a posteriori* conjunta seja própria, consideramos uma distribuição *a priori* conjunta própria para os parâmetros do modelo.

Dessa forma, assumimos distribuições *a priori* independentes com a densidade *a priori* conjunta para $\boldsymbol{\theta} = (\phi, \alpha_1, \alpha_2, \beta_1, \beta_2)$ dada por:

$$\pi(\boldsymbol{\theta}) = \pi(\phi) \prod_{j=1}^{2} \pi(\alpha_j) \prod_{j=1}^{2} \pi(\beta_j), \tag{8}$$

em que, $\pi(\boldsymbol{\beta}_j) = \pi(\beta_{j1}, ..., \beta_{jq}) = \prod_{i=1}^{q} \pi(\beta_{ji})$.

Tanto para o caso da distribuição Weibull quanto para a distribuição Exponencial Generalizada, assumimos as seguintes distribuições *a priori* independentes para o amostrador de Gibbs: $\alpha_j \sim Gama(0,1,0,01)$ e $\beta_{ij} \sim N(0, 10^3)$, $i = 0,1$ e $j = 1,2$. Consideramos, $\phi \sim U(-1,1)$ para o parâmetro da cópula de AMH.

Combinando as distribuições *a priori* independentes com a função de verossimilhança $L(\boldsymbol{\theta}) = \exp(\sum_{i=1}^{n} \ell_i(\boldsymbol{\theta}))$, em que $\ell_i(\boldsymbol{\theta})$ é dada em (7), obtemos a distribuição *a posteriori* conjunta dos parâmetros $\boldsymbol{\theta}$, $\pi(\boldsymbol{\theta}|\mathbb{D})$, em que $\mathbb{D}$ é o conjunto de dados observados. As estimativas dos parâmetros são dados pelas médias da distribuição *a posteriori*.

Como a densidade *a posteriori* conjunta é analiticamente intratável, então o procedimento de inferência foi realizado utilizando métodos MCMC. Todas as implementações computacionais foram realizadas utilizando os sistemas JAGS - Just Another Gibbs Sampler (Plummer, 2003) e R (R Development Core Team, 2012) por meio do pacote *rjags* (Denwood et al., 2015). Para Ilustrar a aplicação do método utilizamos conjuntos de dados simulados e reais. Os resultados estão descritos nas Seções 3 e 5.

3.1 Critérios de comparação de modelos

Os critérios de comparação de modelos, têm por objetivo averiguar se um conjunto de dados foi satisfatoriamente ajustado a um determinado modelo, além de servir como uma ferramenta para a escolha do melhor modelo dentre uma coleção de modelos. Objetivando resolver estes dilemas, a literatura apresenta diversas metodologias.

Neste trabalho, assim como foi feito em Louzada et al., (2013), utilizamos quatro critérios de seleção de modelos: o DIC (*Deviance Information Criterion*), o EAIC (*Expected Akaike Information Criterion*), o EBIC (*Expected Bayesian* (ou *Schwarz*) *Information Criterion*) e o LPML, os quais especificamente são usados na metodologia bayesiana em que as amostras das distribuições *a posteriori* para os parâmetros do modelo são obtidas usando métodos MCMC.

Primeiramente, o critério DIC proposto por Spiegelhalter et al. (2002), o EAIC proposto por Brooks (2002) e o EBIC proposto por Carlin & Louis (2001) são critérios baseados na média *a posteriori* da deviância, $E[D(\theta)]$, que é uma medida de ajuste e que pode ser aproximada por:

$$\bar{D} = \frac{1}{V} \sum_{v=1}^{V} D(\theta_v),$$

sendo v o índice que indica a v-ésima realização de um total de V realizações (após o *burn-in*) e $D(\theta) = -2 \sum_{i=1}^{n} ln(f(t_{1i}, t_{2i}|\theta))$, em que $f(.)$ é a função densidade de probabilidade correspondente ao modelo.

Dessa forma, os critérios EAIC, EBIC, DIC podem ser calculados, respectivamente, por $\widehat{\text{EAIC}} = \bar{D} + 2q$, $\widehat{\text{EBIC}} = \bar{D} + qln(n)$ e $\widehat{\text{DIC}} = 2\bar{D} - \widehat{D}$, em que q é o número de parâmetros no modelo e $D\{E(\theta)\}$ pode ser estimada por $\widehat{D} = D\Big(\frac{1}{V}\sum_{v=1}^{V}\theta_q\Big)$.

Tendo como base, então, os valores obtidos através do cálculo destes critérios, temos que o modelo preferido, dentre uma coleção, é aquele com menores valores destes critérios.

Um outro critério que será usado nesse trabalho é derivado das ordenadas da densidade preditiva condicional (CPO), que é uma ferramenta de avaliação do modelo muito útil e intensamente usada na literatura estatística sob vários contextos (Ibrahim et al., 2001).

Para o modelo proposto não é possível encontrar uma forma fechada de CPO. Entretanto, uma estimativa Monte Carlo de CPO pode ser obtida através de uma simples amostra MCMC a partir da distribuição *a posteriori* $\pi(\boldsymbol{\theta}|\mathbb{D})$. Considere $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(V)}$ uma amostra de tamanho V de $\pi(\boldsymbol{\theta}|\mathbb{D})$ após o *burn-in*. Uma aproximação Monte Carlo de CPO é dada por:

$$\widehat{\text{CPO}_i} = \left(\frac{1}{q}\sum_{q=1}^{V}\frac{1}{f(t_{1i}, t_{2i}|\theta^{(q)})}\right)^{-1}.$$

Para seleção de modelos, utilizamos a estatística LPML $= \sum_{i=1}^{n} log(\widehat{\text{CPO}_i})$, em que, maiores valores de LPML indicam o melhor modelo.

3.2 Diagnóstico

O método da deleção de casos (Cook & Weisberg, 1982) é uma ferramenta muita utilizada quando se objetiva avaliar a influência de uma observação no ajuste de um modelo. Estas técnicas de influência local têm sido amplamente utilizadas, por exemplo, em Cancho et al. (2010), Vidal & Castro (2010) e Louzada et al. (2012, 2013).

Neste artigo, consideramos a análise de influência de deleção de casos baseado na divergência ψ. Seja $D_\psi(P; P_{(-i)})$ a divergência ψ entre P e $P_{(-i)}$, em que P indica a distribuição *a posteriori* de $\boldsymbol{\theta}$ para os dados completos e, $P_{(-i)}$ a distribuição *a posteriori* sem o i-ésimo caso. Especificamente,

$$D_\psi(P; P_{(-i)}) = \int \psi\left(\frac{\pi(\boldsymbol{\theta}|\mathcal{D}^{(-i)})}{\pi(\boldsymbol{\theta}|\mathcal{D})}\right)\pi(\boldsymbol{\theta}|\mathcal{D})\, d\boldsymbol{\theta}, \tag{9}$$

em que ψ é uma função convexa com $\psi(1) = 0$. Várias escolhas de ψ são dadas em Dey & Birmiwal (1994). Por exemplo, $\psi(z) = -\log(z)$ define a divergência de Kullback-Leibler (K-L), $\psi(z) = (z-1)\log(z)$ a distância J (ou a versão simétrica da divergência de K-L), $\psi(z) = 0,5|z-1|$ a distância variacional ou norma L_1 e $\psi(z) = (z-1)^2$ define a divergência χ^2.

Considere Dado $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(V)}$, uma amostra de tamanho V de $\pi(\boldsymbol{\theta}|\mathbb{D})$, então, uma estimativa Monte Carlo de (9) é dada por:

$$\widehat{D_\psi}(P; P_{(-i)}) = \frac{1}{V}\sum_{q=1}^{V}\psi\left(\frac{\pi(\boldsymbol{\theta}^{(q)}|\mathcal{D}^{(-i)})}{\pi(\boldsymbol{\theta}^{(q)}|\mathcal{D})}\right). \tag{10}$$

Dizemos que esta medida $D_\psi(P; P_{(-i)})$ define a divergência ψ do efeito da exclusão do i-ésimo caso dos dados completos na distribuição *a posteriori* de $\boldsymbol{\theta}$.

Como discutido por Peng & Dey, (1995) e Weiss (1996), é uma tarefa muito difícil tentar avaliar o ponto de corte da medida de divergência, de modo a determinar se uma observação ou um pequeno subconjunto de observações é influente ou não. Sendo assim, usaremos a proposta dada por Peng & Dey (1995) e Weiss (1996). Esta proposta é desenvolvida utilizando uma moeda viesada com probabilidade de sucesso p. A divergência ψ entre a moeda viesada e a não viesada é dada por

$$D_\psi(f_0; f_1) = \int \psi\left(\frac{f_0(x)}{f_1(x)}\right) f_1(x)dx, \tag{11}$$

em que, $f_0(x) = p^x(1-p)^{1-x}$ e $f_1(x) = 0,5$ para $x = 0, 1$. Se $D_\psi(f_0, f_1) = d_\psi(p)$, então d_ψ satisfaz a seguinte equação:

$$d_\psi(p) = \frac{\psi(2p) + \psi(2(1-p))}{2}. \tag{12}$$

Note que, para as medidas de divergência consideradas, d_ψ aumenta à medida que p afasta-se de $0,5$. Além disso, $d_\psi(p)$ é simétrica em torno de $p = 0,5$ e d_ψ atinge seu mínimo em $p = 0,5$. Neste ponto, $d_\psi(0,5) = 0$ e $f_0 = f_1$. Portanto, se considerarmos $p > 0,80$ (ou $p \leq 0,20$) como uma moeda muito viciada, então $d_{L_1}(0,80) = 0,30$. Esta relação implica que o i-ésimo caso é considerado influente quando $d_{L_1}(0,80) > 0,30$. Assim, se usarmos a divergência de Kullback-Leibler, podemos considerar que uma observação é influente quando $d_{\text{K-L}} > 0,223$. Da forma análoga, utilizando a distância J ou a divergência χ^2, uma observação na qual $d_J > 0,416$ ou $d_{\chi^2}(0,80) > 0,360$ pode ser considerada influente.

3.3 Estudo de simulação

Partindo-se do pressuposto de que os parâmetros do modelo são conhecidos, geramos conjuntos de dados para estudar as propriedades dos estimadores bayesianos. O objetivo, então, deste estudo de simulação é verificar o bom comportamento das estimativas bayesianas, com base na média frequentista e nas medidas utilizadas para comparação de modelos: EAIC, EBIC, DIC e LPML.

Para simular n observações (t_{i1}, t_{i2}) do modelo baseado na cópula de AMH, assumindo que as marginais T_j têm distribuição Exponencial Generalizada ou Weibull, com paramêtros α_j e $\lambda_{ij} = \exp(\beta_{0j} + \beta_{1j}x_i)$, $j = 1, 2$, realizamos o seguinte passos:

Passo 1: Gerar as covariáveis x_i de uma distribuição Bernoulli com parâmetro $0,5$.

Passo 2: Gerar os tempos de censura C_{ij} a partir de uma distribuição Uniforme $U(0, \tau_j)$, com τ_j controlando o percentual de observações censuradas, $j = 1, 2$.

Passo 3: Gerar $u_{i1} \sim U(0, 1)$ para obter o T_{i1} e calcular t_{i1} da seguinte forma:

- para a distribuição Weibull: Gerar $T_{i1} = (-\log(1 - u_{i1})/\lambda_{i1})^{1/\alpha_1}$, em que, $u_{i1} \sim U(0, 1)$.

- para a distribuição Exponencial Generalizada: Gerar $T_{i1} = ((-log(u_{i1})/\lambda_{i1}))^{1/\alpha_1}$, em que, $u_{i1} \sim U(0,1)$.

Comparar T_{i1} com o valor de censura C_{i1} a fim de determinar o indicador de censura δ_{i1} e o valor observado dado por $t_{i1} = \min(T_{i1}, C_{i1})$.

Passo 4: Gerar $u_{i2} \sim U(0,1)$ e obter w_i, a solução da equação não linear $u_{i2} - \frac{w_i[1-\phi(1-w_i)]}{[1-\phi(1-u_{i1})(1-w_i)]^2} = 0$. Calcular o tempo T_{i2} da seguinte forma:

- para a distribuição Weibull: $T_{i2} = (-\log(1-w_i)/\lambda_{i2})^{1/\alpha_2}$.
- para a distribuição Exponencial Generalizada: $T_{i2} = (-\log(w_i)/\lambda_{i2})^{1/\alpha_2}$.

Comparar T_{i2} com o valor de censura C_{i2} a fim de determinar o indicador de censura δ_{i2} e o valor observado dado por $t_{i2} = \min(T_{i2}, C_{i2})$.

Geramos os conjuntos de dados assumindo ausência de dados censurados (0%, 0%) e (30%, 20%) de censuras para três diferentes tamanhos de amostras $N = 50$, $N = 100$ e $N = 200$. Para cada caso, geramos 200 conjuntos Monte Carlo de dados.

Para o modelo com marginais Weibull, foram considerados os seguintes valores para os parâmetros: $\alpha_1 = 2$, $\beta_{01} = -1$, $\beta_{11} = 0,5$, $\alpha_2 = 3$, $\beta_{02} = 1,5$, $\beta_{12} = -0,5$ e $\phi = 0,5$. Para o modelo com marginais Exponencial Generalizada, foram considerados $\alpha_1 = 0,5$, $\beta_{01} = -1,5$, $\beta_{11} = 1$, $\alpha_2 = 2$, $\beta_{02} = 1,5$, $\beta_{12} = -0,5$ e $\phi = 0,5$.

Para cada conjunto de dados gerados, consideramos duas cadeias de tamanho 60.000. Para eliminar o efeito dos valores iniciais, foram desconsideradas as primeiras 10.000 iterações. Para evitar problemas de autocorrelaçao, considerou-se um espaçamento de tamanho 10, obtendo uma amostra efetiva de tamanho 10.000 sobre a qual a inferência *a posteriori* é baseada. Para cada amostra, a média e o desvio padrão *a posteriori* dos parâmetros e os valores dos critérios de seleção de modelo são gravados.

A convergência das cadeias foi monitorada de acordo com os métodos recomendados por Cowless & Carlin, (1996),por meio do pacote CODA (Plummer et al., 2006). Em todos os casos, a convergência foi verificada por meio do diagnóstico de Gelman-Rubin (Gelman & Rubin, 1992) sendo muito próximo a 1 ($\leq 1,01$).

A Tabela 1, mostra a média Monte Carlo (MC) das estimativas dos parâmetros ajustando a cópula AMH com distribuições marginais Weibull e Exponencial Generalizada para o caso sem censura (0%, 0%) e com censura (30%, 20%) e três tamanhos de amostras ($N = 50, 100, 200$). Nota-se que, nos casos em que se gera e se obtém o ajuste do mesmo modelo (com e sem a presença de censura) as estimativas obtidas estão próximas, em média, do verdadeiro valor e, para os modelos cruzados, as estimativas diferem bastante. Como exemplo, note o caso em que estamos gerando à partir do modelo com marginais Exponencial Generalizada, sendo assim, observe que, para o parâmetro α_1, em que o verdadeiro valor é $0,5$, no caso $N = 50$ e sem censura, o ajuste do modelo com marginais Exponencial Generalizada nos dá como estimativa para este mesmo parãmetro o valor $0,509$ que é mais próximo de $0,5$ do que $0,652$, que é o valor da estimativa deste parâmetro considerando o ajuste do modelo com marginais Weibull. Vale ressaltar, igualmente, que estas diferenças nas estimativas dos parâmetros para os modelos cruzados acentuam-se ainda mais no caso em que o verdadeiro modelo é Weibull e estimamos a partir da Exponencial Generalizada.

Tabela 1: Média MC das estimativas dos parâmetros ajustando o modelo AMH bivariado com marginais Weibull e Exponencial Generalizada.

Verdadeiro Modelo			N = 50		N = 100		N = 200	
			Exp.Ge	Weib.	Exp.Ge	Weib.	Exp.Ge	Weib.
Exponencial Generalizada	Sem censura	α_1 (0,5)	0,509	0,652	0,512	0,654	0,505	0,645
		β_{01} (-1,5)	-1,513	-0,483	-1,481	-0,477	-1,505	-0,477
		β_{11} (1,0)	0,997	0,654	0,991	0,647	1,025	0,665
		α_2 (2,0)	2,141	1,480	2,065	1,443	2,075	1,443
		β_{02} (1,5)	1,510	1,472	1,513	1,453	1,507	1,431
		β_{12} (-0,5)	-0,502	-0,746	-0,512	-0,738	-0,496	-0,715
		ϕ (0,5)	0,352	0,350	0,391	0,392	0,456	0,456
	Com censura	α_1 (0,5)	0,514	0,626	0,502	0,608	0,504	0,608
		β_{01} (-1,5)	-1,685	-0,571	-1,584	-0,527	-1,552	-0,522
		β_{11} (1,0)	1,137	0,703	1,035	0,660	1,037	0,662
		α_2 (2,0)	2,224	1,545	2,061	1,485	2,040	1,467
		β_{02} (1,5)	1,493	1,518	1,500	1,477	1,498	1,457
		β_{12} (-0,5)	-0,477	-0,739	-0,505	-0,726	-0,494	-0,711
		ϕ (0,5)	0,302	0,300	0,395	0,393	0,448	0,448
Weibull	Sem censura	α_1 (2,0)	3,906	2,089	3,587	2,038	3,467	2,020
		β_{01} (-1,0)	0,285	-1,073	0,277	-1,026	0,269	-1,022
		β_{11} (0,5)	0,259	0,544	0,254	0,515	0,248	0,506
		α_2 (3,0)	10,953	3,156	9,399	3,048	8,833	3,034
		β_{02} (1,5)	1,657	1,573	1,631	1,522	1,619	1,509
		β_{12} (-0,5)	-0,170	-0,537	-0,168	-0,514	-0,170	-0,511
		ϕ (0,5)	0,385	0,335	0,452	0,383	0,530	0,442
	Com censura	α_1 (2,0)	3,716	2,098	3,550	2,084	3,321	2,026
		β_{01} (-1,0)	0,223	-1,083	0,241	-1,060	0,231	-1,011
		β_{11} (0,5)	0,252	0,514	0,254	0,524	0,252	0,501
		α_2 (3,0)	10,127	3,145	9,432	3,090	8,178	3,015
		β_{02} (1,5)	1,610	1,535	1,616	1,537	1,584	1,502
		β_{12} (-0,5)	-0,168	-0,526	-0,173	-0,532	-0,165	-0,500
		ϕ (0,5)	0,340	0,318	0,399	0,352	0,503	0,438

A Tabela 2 apresenta a média Monte Carlo (MC) dos quatro critérios de comparação de modelos, com o intuito de comparar os modelos de sobrevivência bivariados baseados na cópula de AMH com marginais Weibull ou Exponencial Generalizada. Podemos observar que, para os casos com e sem censura, o verdadeiro modelo gerado supera o outro de acordo com todos os critérios considerados. Por exemplo, para o caso em que estamos gerando à partir do modelo com marginais Weibull e $N = 100$ com censura, temos que: para o ajuste do modelo com marginais Weibull DIC $= 143,945$, EAIC $= 151,170$, EBIC $= 169,406$ e LPML $= -72,282$; enquanto que, para o ajuste do modelo com marginais Exponencial Generalizada, DIC $= 156,725$, EAIC $= 164,120$, EBIC $= 182,357$ e LPML $= -78,801$. Ou seja, para o caso em que estamos gerando à partir do modelo com marginais Weibull, o ajuste do modelo com marginais Weibull obtém um melhor resultado, já que apresenta menores valores de DIC, EAIC e EBIC e um maior valor de LPML em comparação ao ajuste do modelo com marginais Exponencial Generalizada.

4 Diagnóstico de observações influentes

Para examinar o desempenho da medida de diagnóstico, geramos uma amostra de tamanho 300 para o modelo AMH bivariado com marginais Weibull, considerando os seguintes valores para os parâmetros: $\beta_{01} = -1,0$, $\beta_{11} = 0,5$, $\alpha_1 = 2$, $\beta_{02} = 1,5$, $\beta_{12} = -0,5$, $\alpha_2 = 3$ e $\phi = 0,5$. As porcentagens de observações censuradas, na amostra, para os tempos t_1 e t_2 foram, respectivamente, $33,3\%$ e $18,7\%$.

Selecionamos os casos 10, 149 e 285 para perturbação. Para criar observações artificialmente influentes no conjunto de dados, escolhemos um, dois ou três desses casos selecionados. Para cada caso, pertubamos ambos os tempos da seguinte forma: $\widetilde{t_i} = t_i + 5D_t, i = 1, 2$, em que D_t é o desvio padrão dos t_i's.

Para a implementação do algoritmo MCMC, assim como a verificação da convergência das cadeias, realizamos os mesmos procedimentos descritos anteriormente.

A Tabela 3, mostra que as inferências *a posteriori* são sensíveis à perturbação do(s) caso(s) selecionado(s). Como podemos notar nesta tabela, o conjunto de dados (a), que denota os dados originais simulados sem perturbação, tem suas médias das estimativas *a posteriori* dos parâmetros do modelo considerado muito próximas dos verdadeiros valores destes parâmetros, principalmente se comparadas com as médias das estimativas dos conjuntos de dados (b) a (h), que designam os conjuntos de dados com casos perturbados.

A Tabela 4 apresenta os critérios bayesianos do ajuste de diferentes casos de conjuntos de dados perturbados. Podemos observar que o conjunto de dados (a) teve o melhor ajuste, ou seja, menores valores de EAIC, EBIC e DIC e maior valor de LPML em comparação com os conjuntos de dados (b) a (h), o que era de se esperar, já que ele consiste no conjunto dos dados originais simulados.

Na Tabela 5 apresentamos uma estimativa das quatro medidas de divergência para o modelo AMH bivariado com marginais Weibull para cada conjunto de dados simulados. Note que, antes da perturbação, representado pelo conjunto de dados (a), todos os casos selecionados não são influentes, com pequenas medidas de divergência. Entretanto, após perturbações, representado pelos conjuntos de dados (b) a (h), as quatro medidas aumentam, indicando que os casos são influentes.

Tabela 2: Média Monte Carlo dos quatro critérios bayesianos baseados sobre as 200 amostras geradas.

Verdadeiro Modelo			N = 50		N = 100		N = 200	
			Exp.Ge	Weib.	Exp.Ge	Weib.	Exp.Ge	Weib.
Exponencial Generalizada	Sem censura	DIC	138,448	140,659	273,420	277,941	541,072	550,064
		EAIC	145,808	147,997	280,579	285,093	548,156	557,140
		EBIC	159,193	161,381	298,815	303,329	571,244	580,228
		LPML	-69,503	-70,785	-136,886	-139,277	-270,632	-275,240
	Com censura	DIC	81,585	82,258	159,886	161,693	312,709	317,536
		EAIC	89,065	89,676	167,128	168,902	319,811	324,626
		EBIC	102,449	103,060	185,364	187,138	342,900	347,715
		LPML	-41,150	-41,664	-80,144	-81,221	-156,458	-159,017
Weibull	Sem censura	DIC	95,158	86,467	188,383	170,300	376,046	336,784
		EAIC	102,779	93,814	195,771	177,454	383,311	343,853
		EBIC	116,164	107,199	214,007	195,690	406,399	366,941
		LPML	-48,205	-43,724	-94,629	-85,344	-188,549	-168,488
	Com censura	DIC	82,803	76,566	156,725	143,945	315,056	287,383
		EAIC	90,554	84,084	164,120	151,170	322,279	294,491
		EBIC	103,938	97,469	182,357	169,406	345,368	317,579
		LPML	-42,060	-38,867	-78,801	-72,282	-157,945	-143,815

Tabela 3: Média e desvio padrão (DP) das estimativas a posteriori dos parâmetros do modelo AMH bivariado com marginais Weibull para cada conjunto de dados simulados.

Nomes dos dados	Casos perturbados	α_1(2,0) Média DP	β_{01}(-1,0) Média DP	β_{11}(0,5) Média DP	α_2(3,0) Média DP	β_{02}(1,5) Média DP	β_{12}(-0,5) Média DP	ϕ(0,5) Média DP
a	Nenhum	1,962	-1,036	0,396	3,097	1,488	-0,492	0,406
		0,107	0,119	0,143	0,149	0,109	0,131	0,174
b	10	1,877	-0,985	0,282	2,832	1,381	-0,572	0,519
		0,100	0,114	0,140	0,130	0,104	0,133	0,182
c	149	1,921	-1,070	0,442	2,931	1,325	-0,367	0,502
		0,104	0,121	0,143	0,139	0,103	0,127	0,169
d	285	1,872	-0,984	0,290	3,018	1,456	-0,555	0,484
		0,099	0,117	0,144	0,144	0,109	0,131	0,176
e	{10, 149}	1,832	-1,032	0,336	2,691	1,217	-0,440	0,674
		0,097	0,115	0,140	0,124	0,105	0,129	0,178
f	{10, 285}	1,807	-0,959	0,200	2,780	1,346	-0,619	0,652
		0,096	0,116	0,140	0,129	0,111	0,136	0,181
g	{149, 285}	1,831	-1,029	0,350	2,851	1,288	-0,424	0,611
		0,097	0,116	0,14	0,132	0,103	0,131	0,172
h	{10, 149, 285}	1,769	-1,030	0,265	2,623	1,146	-0,469	0,898
		0,090	0,115	0,136	0,122	0,106	0,129	0,132

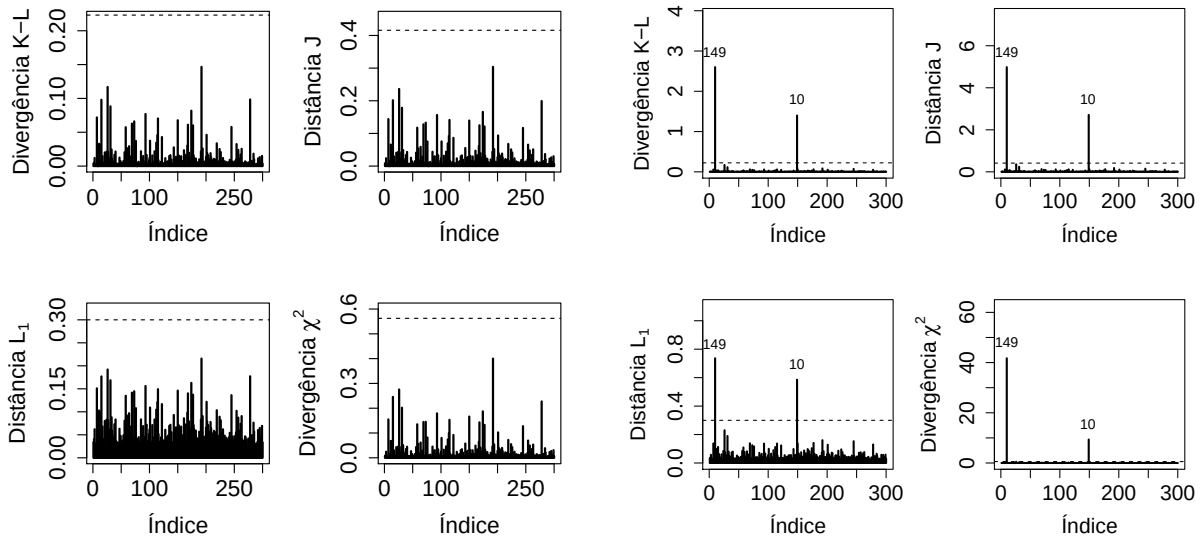
Tabela 4: Critérios bayesianos ajustando o modelo de sobrevivência AMH bivariado com marginais Weibull para cada conjunto de dados simulados.

	Critérios Bayesianos			
Nomes dos dados	EAIC	EBIC	DIC	LPML
a	443,242	469,168	436,150	-218,261
b	485,839	511,765	478,694	-241,912
c	473,508	499,435	466,322	-234,217
d	466,057	491,983	458,960	-230,505
e	513,740	539,666	506,464	-256,997
f	504,424	530,350	497,287	-251,869
g	495,786	521,713	488,572	-246,533
h	526,680	552,607	516,568	-266,260

Tabela 5: Medidas de divergência para o modelo AMH bivariado com marginais Weibull para cada conjunto de dados simulados.

Nome dos dados	Caso perturbado	medidas de divergencia			
		K-L	J	L_1	χ^2
a	10	0,008	0,016	0,050	0,016
	149	9×10^{-5}	$1,8 \times 10^{-4}$	0,005	2×10^{-4}
	285	0,010	0,020	0,057	0,021
b	10	2,854	5,512	0,785	47,600
c	149	1,221	2,423	0,569	7,449
d	285	1,146	2,454	0,563	11,816
e	10	2,600	4,982	0,736	41,711
	149	1,399	2,711	0,586	9,367
f	10	2,407	4,601	0,727	27,127
	285	1,024	1,992	0,519	4,763
g	149	1,483	2,981	0,619	12,600
	285	1,078	2,194	0,545	6,815
h	10	2,707	5,045	0,755	37,350
	149	2,410	5,199	0,752	104,256
	285	1,642	3,176	0,653	11,057

As Figuras 4(a,b) mostram os gráficos de índices das quatro medidas de divergência para o conjunto de dados (a) e (e). Em relação ao conjunto de dados (a), observemos que em nenhum dos quatro gráficos foram detectados pontos influentes, o que está de acordo, pois nenhum caso foi perturbado neste conjunto de dados. Por outro lado, como no conjunto de dados (e) os casos 10 e 149 foram perturbados, observa-se que os quatro gráficos conseguiram detectá-los como pontos influentes.



(a) caso (a). (b) caso (e).

Figura 4: Gráficos de índices das medidas de divergência.

5 Aplicação à dados reais

Nesta seção, aplicamos o modelo proposto a dois conjuntos de dados reais. O primeiro é um conjunto de dados de insuficiência renal, descrito em McGilchrist & Aisbett (1991). O segundo, é um conjnto de dados de retinopatia diabética.

5.1 Dados de insuficiência renal

Nossa primeira aplicação à dados reais que será apresentada é referente à 38 pacientes com insuficiência renal (perda das funções dos rins, podendo ser aguda ou crônica). Os tempos bivariados medidos em dias são a respeito de recorrência de infecção no local onde foi inserido o catéter nos pacientes que utilizaram um aparelho portátil de diálise, sendo dado para cada paciente dois tempos de recorrência. Como covariável, consideramos o sexo do paciente, atribuindo-se 0 para o sexo masculino e 1 para o sexo feminino.

Para realizarmos o ajuste do modelo AMH bivariado com ambas marginais Weibull ou Exponencial Generalizada, consideramos duas cadeias de tamanho 60.000, das quais, foram desconsideradas as primeiras 10.000 iterações, com o objetivo de eliminar o efeito dos valores iniciais. Para evitar problemas de autocorrelaçao, considerou-se um espaçamento de tamanho 10, adquirindo uma amostra efetiva de tamanho 10.000 sobre a qual a inferência *a posteriori* é baseada. A convergência das cadeias foi monitorada de acordo com os métodos recomendados por (Cowless & Carlin, 1996).

Na Tabela 6 apresentamos os resumos *a posteriori* para os parâmetros do modelo AMH bivariado para ambas as distribuições.

Tabela 6: Média *a posteriori*, desvio padrão (DP) e intervalo de confiança (IC) de 95% para os parâmetros do modelo AMH bivariado com marginais Weibull ou Exponencial Generalizada

	Parâmetro	Exponencial Generalizada			Weibull		
		Média	DP	HPD (95%)	Média	DP	HPD(95%)
Tempo 1	α_1	0,923	0,204	(0,530; 1,315)	0,948	0,126	(0,705; 1,191)
	β_{01}	-3,656	0,381	(-4,449; 2,952)	-3,353	0,607	(-4,52; -2,158)
	β_{11}	-1,692	0,413	(-2,521; -0,891)	-1,622	0,423	(-2,413; -0,756)
Tempo 2	α_2	0,734	0,160	(0,446; 1,055)	0,803	0,107	(0,599; 1,015)
	β_{02}	-4,802	0,421	(-5,626; -4,006)	-3,474	0,645	(-4,763; -2,258)
	β_{12}	-0,332	0,466	(-1,244; 0,569)	-0,402	0,384	(-1,141; 0,357)
Copula	ϕ	0,242	0,512	(-0,740; 1,000)	0,208	0,513	(-0,764; 1,000)

A Tabela 7 apresenta os critérios de comparação de modelos.De acordo com todos os critérios, o modelo que apresentou melhores resultados foi o modelo de sobrevivência bivariado baseado na cópula de AMH com marginais Weibull, pois ele obteve menores valores de EAIC, EBIC e DIC e um maior valor de LPML em comparação com o modelo de sobrevivência bivariado baseado na cópula de AMH com marginais Exponencial Generalizada.

Tabela 7: Critérios para comparação dos modelos

Modelo	Critérios Bayesianos			
	EAIC	EBIC	DIC	LPML
Exp Gen	743,353	754,816	735,212	-370,409
Weibull	741,891	753,354	734,020	-369,570

Na Figura 5 apresentamos os gráficos de índices considerando o modelo de AMH com marginal Weibull. Podemos observar que todas as medidas detectam a observação 21 como possível ponto influente.

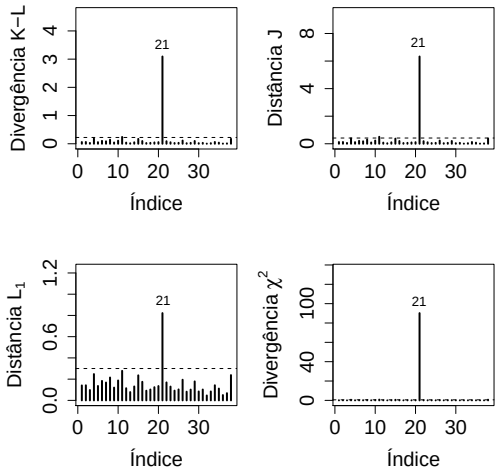


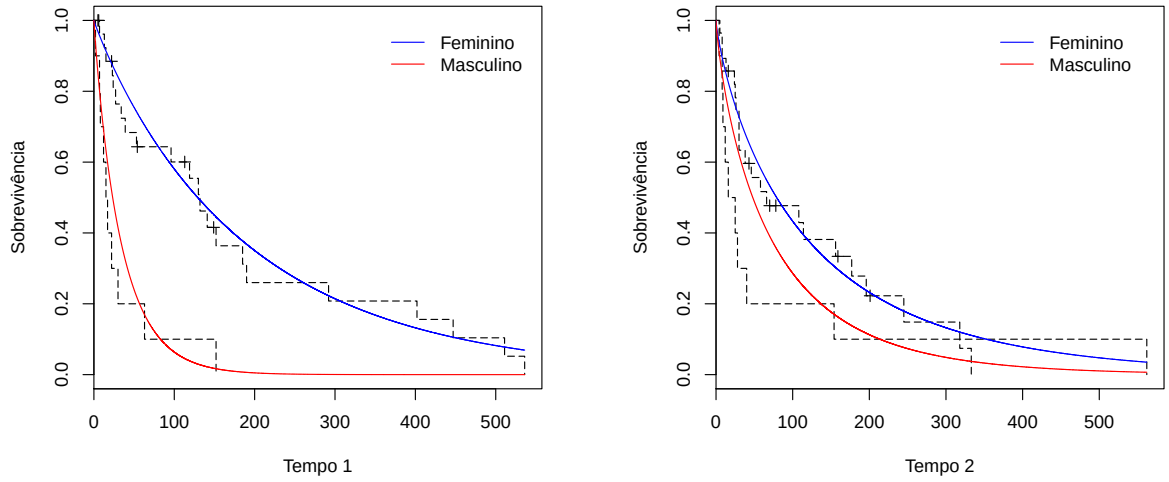
Figura 5: Gráficos de índices das medidas de divergência considerando a distribuição Weibull.

A Tabela 8 mostra as quatro medidas de divergência para a observação 21 e, como podemos notar pelos valores altos, ela é realmente uma observação influente.

Tabela 8: Medidas de divergência para a observação 21

	Medida de divergência			
Nome do dado	K-L	J	L_1	χ^2
Observação 21	3,093	6,331	0,822	90,230

As Figuras 6(a,b) mostram, respectivamente, as curvas de Kaplan-Meier para as variáveis T_1 e T_2 dicotomizadas pelo sexo do paciente juntamente com os ajustes do modelo de sobrevivência bivariado baseado na cópula de AMH com marginais Weibull e, como podemos inferir por estes gráficos, os dados foram adequadamente ajustados por este modelo.



(a) Variável T_1.

(b) Variável T_2.

Figura 6: Curvas de Kaplan-Meier e curvas de sobrevivências Weibull estimadas.

5.2 Dados reais de retinopatia diabética

A retinopatia diabética é uma complicação que ocorre quando o excesso de glicose no sangue danifica os vasos sanguíneos de dentro da retina, tornando esta doença uma das principais causas de cegueira no mundo e os diabéticos 25 vezes mais propensos de se tornarem cegos do que os não diabéticos. O sintoma mais comum é a visão embaçada, sendo que a perda visual pode ser um sintoma tardio, expressando a gravidade da situação.

O interesse do estudo é verificar a eficácia do tratamento de fotocoagulação com raio laser, em retardar o aparecimento da cegueira.

O conjunto de dados é composto por 197 pacientes. O tratamento foi aleatóriamente atribuído para um olho de cada paciente. O olho que não recebeu tratamento foi considerado como controle. A censura foi causada por morte, abandono ou término do estudo, sendo que estas observações censuradas aconteceram em 73% dos olhos tratados e 49% dos olhos não tratados.

A idade do paciente no início da diabete foi considerada como covariável. Para criar dois grupos foi considerado um ponto de corte de 20 anos (58% dos pacientes tinham menos de 20 anos de idade). Considerou-se T_1 como um vetor de tempos até a perda da visão para o olho de tratamento e T_2 como o vetor de tempos até a perda visual para o olho controle.

O procedimento de o ajuste do modelo AMH bivariado foi feito de forma similar ao descrito na Seção 5.1. Ou seja, consideramos duas cadeias de tamanho 60.000, com descarte das primeiras 10.000 e espaçamento de tamanho 10, obtendo uma amostra de tamanho 10.000 para inferência. A convergência das cadeias foi monitorada de acordo com os métodos recomendados por (Cowless & Carlin, 1996).

Na Tabela 9 apresentamos os resumos *a posteriori* para os parâmetros do modelo AMH bivariado para ambas as distribuições.

Tabela 9: Média *a posteriori*, desvio padrão (DP) e intervalo de confiança (IC) de 95% para os parâmetros do modelo AMH bivariado com marginais Weibull ou Exponencial Generalizada

	Parâmetro	Exponencial Generalizada			Weibull		
		Média	DP	HPD (95%)	Média	DP	HPD (95%)
Tempo 1	α_1	0,771	0,113	(0,544; 0,984)	0,812	0,101	(0,611; 1,003)
	β_{01}	-5,208	0,339	(-5,88 ; -4,591)	-4,039	0,415	(-4,883; -3,248)
	β_{11}	-0,608	0,37	(-1,346; 0,111)	-0,499	0,291	(-1,105; 0,034)
Tempo 2	α_2	0,789	0,091	(0,622; 0,976)	0,832	0,073	(0,693; 0,976)
	β_{02}	-4,634	0,229	(-5,07; -4,190)	-3,687	0,304	(-4,295; -3,113)
	β_{12}	0,443	0,237	(-0,014; 0,915)	0,368	0,198	(-0,029; 0,744)
Copula	ϕ	0,804	0,149	(0,521; 1,000)	0,804	0,146	(0,523; 1,000)

A Tabela 10 apresenta os critérios de comparação de modelos.De acordo com todos os critérios, o modelo que apresentou melhores resultados foi o modelo de sobrevivência bivariado baseado na cópula de AMH com marginais Weibull, pois ele obteve menores valores de EAIC, EBIC e DIC e um maior valor de LPML em comparação com o modelo de sobrevivência bivariado baseado na cópula de AMH com marginais Exponencial Generalizada.

Tabela 10: Critérios para comparação dos modelos

Modelo	Critérios Bayesianos			
	EAIC	EBIC	DIC	LPML
Exp Gen	1.673,120	1.696,102	1.665,413	-832,525
Weibull	1.671,481	1.694,464	1.663,862	-831,725

Na Figura 7, apresentamos os gráficos de índices considerando o modelo de AMH com marginal Weibull. Podemos observar que todas as medidas não detectam nenhum possível ponto influente.

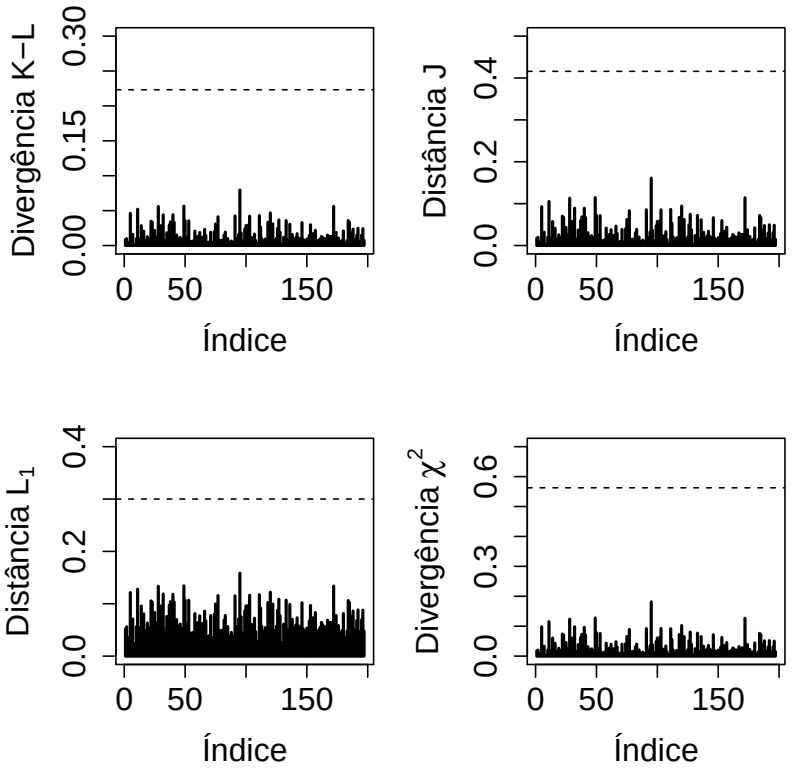
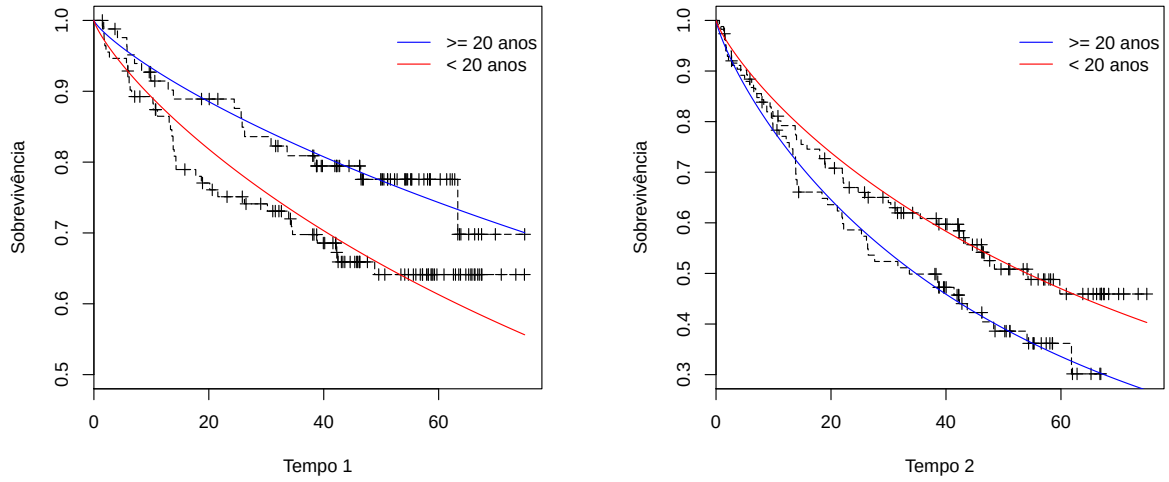


Figura 7: Gráficos de índices das medidas de divergência considerando a distribuição Weibull.

As Figuras 8(a,b) mostram, respectivamente, as curvas de Kaplan-Meier para as variáveis T_1 e T_2 dicotomizadas pela idade do paciente juntamente com os ajustes do modelo de sobrevivência bivariado baseado na cópula de AMH com marginais Weibull e, como podemos inferir por estes gráficos, os dados foram adequadamente ajustados por este modelo.



(a) Variável T_1.

(b) Variável T_2.

Figura 8: Curvas de Kaplan-Meier e curvas de sobrevivências Weibull estimadas.

6 Considerações finais

Neste artigo foram apresentadas as funções cópulas, em especial as cópulas Arquimedianas, como eficientes ferramentas para modelar dados de sobrevivência. Foi construído um novo modelo bivariado baseado na cópula AMH.

Todo o procedimento inferencial foi realizado sob uma abordagem bayesiana assumindo distribuições *a priori* não informativas. Foi realizado um estudo de simulação com o objetivo de verificar o bom comportamento das estimativas bayesianas com base na média frequentista. Por meio destas simulações também foi verificado que, com diferentes tamanhos amostrais e diferentes configurações de censura, as estimativas obtidas foram próximas do verdadeiro valor.

Também, foi realizada comparações de modelos por meio dos critérios bayesianos EAIC, EBIC, DIC e LPML. Simulamos amostras a partir do modelo de AMH com marginais ou Weibull ou Exponencial Generalizada e observamos que todos os critérios indicaram o modelo no qual as amostras foram geradas.

Para analisarmos a robustez do modelo relacionado às escolhas dos hiperparâmetros das distribuições a *priori*, foi realizado um estudo de sensibilidade no qual concluímos que as estimativas dos parâmetros a *posteriori* não apresentaram diferenças significativas nos resultados das aplicações aos dados artificiais e aos dados reais.

Além disso, aplicamos o método bayesiano de análise de influência de deleção de casos baseado na divergência ψ cujo o objetivo é detectar possível(is) observação(ões) influente(s) nos dados analisados. Para isso, foram assumidas quatro particulares escolhas para a função ψ nas quais resultaram a divergência de Kullback-Leibler (K-L), a distância J, a distância variacional ou norma L_1 e a divergência χ^2. Para uma amostra simulada de cada modelo, perturbamos uma, duas ou três observações e, à partir disso, conseguimos averiguar que as quatro medidas de divergência detectaram os pontos perturbados.

Por fim, realizamos duas aplicações aos dados reais de pacientes com infecção renal e de pacientes com retinopatia diabética, obtendo, no final, as curvas de Kaplan-Meier e curvas de sobrevivências Weibull bivariadas estimadas para ambas as aplicações aos dados reais.

Referências

- ALI, M. M.; MIKHAIL, N. N.; HAQ, M. S. *A class of bivariate distributions including the bivariate logistic*. Journal of Multivariate Analysis **8**, 405-412, 1978.
- BOLETA, J.; ACHCAR, J. A. *Distribuição Exponencial Generalizada bivariada derivada de funções cópulas: Uma aplicação a dados de câncer gástrico*. Revista Brasileira de Biometria, **30**(4), 401-414, 2012.
- BROOKS, S. P. *Discussion on the paper by Spiegelhalter, Best, Carlin, and van der Linde*, **64**, 616-618, 2002.
- CANCHO, V.; ORTEGA, E.; PAULA, G. *On estimation and influence diagnostics for log-Birnbaum-Saunders Student-t regression models: Full Bayesian analysis*. Journal of Statistical Planning and Inference, **140**, 2486-2496, 2010.
- CARLIN, B. P.; LOUIS, T. A. *Bayes and Empirical Bayes Methods for Data Analysis*. 2.ed. Boca Raton: Chapman and Hall, 2001.
- CHO, H.; IBRAHIM, J. G.; SINHA, D.; SHU, H. *Bayesian case influence diagnostics for survival models*. Biometrics, **65**, 116-124, 2009.
- COOK, R. D.; WEISBERG, S. *Residuals and Influence in Regression*. Boca Raton: Chapman and Hall, 1982.
- COLOSIMO, E. A. & GIOLO, S. R. *Análise de Sobrevivência Aplicada*, Editora Blucher, São Paulo, 2006.
- COWLESS, M. K.; CARLIN, B. P. *Markov chain Monte Carlo convergence diagnostics: a comparative review*. Journal of the American Statistical Association, **91**, 883-904, 1996.
- DENWOOD M. J.; STUKALOV A.; PLUMMER M. *runjags: An R package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC models in JAGS*. Journal of Statistical Software, 2015.
- DEY, D.; BIRMIWAL, L. *Robust Bayesian analysis using divergence measures*. Statistics and Probability Letters, **20**, 287-294, 1994.
- EMBRECHTS, P.; LINSKOG, F.; MCNIEL, A. *Modelling dependence with copulas and applications to risks management*. http://www.math.ethz.ch/baltes/ftp/papers.html, 2003.
- FREES, E., ANDWANG, P. *Credibility using copulas*. North American Actuarial Journal, 2005.

- GELMAN, A.; RUBIN, D. B. *Inference from iterative simulation using multiple sequences*. Statistical Science, **7**, 457-511, 1992.
- GUPTA, R. D.; KUNDU, D. *Generalized Exponential distributions*. Aust. N.Z. J. Stat., Oxford, **41**, 173-188, 1999.
- JOE, H. *Dependence Modeling with Copulas.* London: Chapman and Hall, 2014.
- KOLEV, N.; DOS ANJOS, U.; MENDES, B. V. M. *Copulas: A review and recent developments*. Stochastic Models, **22**(4), 617-660, 2006.
- LAWLESS, J. F. *Statistical Models and Methods for Lifetime Data*. New York: Wiley and Sons, 2003.
- LOUZADA, F.; SUZUKI, A. K.; CANCHO, V. G.; PRINCE F. L.; PEREIRA, G. A. *The Long-Term Bivariate Survival FGM Copula Model: An Application to a Brazilian HIV Data*. Journal of Data Science, **10**, 511-535, 2010.
- LOUZADA, F.; SUZUKI, A. K.; CANCHO, V. G. *The FGM Long-Term Bivariate Survival Copula Model: Model, Bayesian Estimation, and Case Influence Diagnostics*. Communications in Statistics - Theory and Methods, **42** (4), 673-691, 2013.
- MCGILCHRIST C. A.; AISBETT C. W. *Regression with frailty is survival analysis*. Biometrics, **47**, 461-466, 1991.
- NELSEN, R. *Properties of a one-parametric family of bivariate distributions with specified marginals*. Communications in Statistics, **15**, 3277-3285, 1986.
- NELSEN, R. *An Introduction to Copulas*. 2.ed. New York: Springer, 2006.
- PENG, F.; DEY, D. *Bayesian analysis of outlier problems using divergence measures*. The Canadian Journal of Statistics - La Revue Canadienne de Statistique, **23**, 199-213, 1995.
- PLUMMER, M.; BEST, N.; COWLES, K.; VINES, K.*Output analysis and diagnostics for MCMC*. http://cran.r-project.org/web/packages/coda/index.html, 2006.
- PURWONO, Y. *Copula inference for multiple lives analysis - preliminares*. Department of Management Faculty of Economics, University of Indonesia, Indonesia, 2009.
- QUEIROS FLORES, A. *Copula functions and bivariate distributions for survival analysis: An application to political survival*. Wilf Department of Politics. New York University. 19 West 4th St., Second Floor, 2008.
- R DEVELOPMENT CORE TEAM. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing: Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org, 2007.
- SPIEGELHALTER, D. J.; BEST, N. G.; CARLIN, B. P.; VAN DER LINDE, A. *Bayesian measures of model complexity and fit*. Journal of the Royal Statistical Society Series B, **64**, 583-639, 2002.

- SUZUKI, A. K.; LOUZADA-NETO, F.; CANCHO, V. G.; BARRIGA, G. D. C. *The FGM bivariate lifetime copula model: a bayesian approach*. Advances and Applications in Statistics, **21**(1), 55-76, 2011.

- VAUPEL, J. W., MANTON, K. G. & Stallard, E. *The impact of heterogeneity in individual frailty on the dynamics of mortality*. *Demography*, **16**, 439-454, 1979

- VIDAL, I.; CASTRO, L. M. *Influential observations in the independent Student-t measurement error model with weak nondifferential error*. Chilean Journal of Statistics, **1**, 17-34, 2010.

- ZHANG, L. & SINGH, V. P. *Bivariate rainfall fequency distributions using Archimedean copulas*. Department of Biological & Agricultural Engineering, Texas A & M University, 2117 TAMU, College Station, Texas USA, 2006.

- WEIBULL W. *A statistical theory of the strength of material*. Proc. Roy. Swedish Inst. Eng. Res. 151, 1939.

- WEISS, R. *An approach to Bayesian sensitivity analysis*. Journal of the Royal Statistical Society Series B, **58**, 739-750, 1996.