REVISTA DE LA FACULTAD DE INGENIERIA
www.revistaingenieria.uda.cl

Um Novo Modelo de Sobrevivência Bayesiano Bivariado: Modelagem, Inferência e Análise de Influência

Natan Hilário da Silva[1] and Adriano Kamimura Suzuki[2]

[1]Programa Interinstitucional de Pós-Graduação em Estatística UFSCar-USP, São Carlos, SP, Brasil; *natan.hilario@usp.br*
[2]Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, SP, Brasil; *suzuki@icmc.usp.br*

RESUMO

Um interesse frequente na análise de sobrevivência é a modelagem de múltiplos tempos de ocorrência para diferentes eventos e as relações entre eles. Neste trabalho, é proposto um novo modelo para tempos de vida bivariados através da cópula Farlie-Gumbel-Morgenstern com marginais seguindo a distribuição A. É feita uma breve introdução da recente distribuição A e o processo inferencial Bayesiano para a detecção de observações influentes. É feito um estudo de simulação com perturbação artificial dos dados, seguido pela aplicação em um contexto real de tempos até a realização da cirurgia de apendicectomia em gêmeos australianos. Também é considerada a análise do efeito de um medicamento no tempo de reação de ratos de laboratório com diferentes níveis de chumbo no organismo. Os resultados observados mostram que o modelo é robusto a valores *outliers*, podendo ser um forte candidato em contextos de dados com forte assimetria à direita. Também, a distribuição A se mostra flexível a tempos de vida distantes de zero, diferentemente de distribuições uniparamétricas usuais como Exponencial, Lindley e Rayleigh.

Palabras claves: Análise de sobrevivência; inferência Bayesiana; análise de influência; cópulas; dados bivariados.

ABSTRACT

A frequent interest in survival analysis is the modeling of multiple occurrence times for different events and the relationships between them. In this work, a new model for bivariate lifetimes is proposed using the Farlie-Gumbel-Morgenstern copula with marginals following the A distribution. A brief introduction to the recent A distribution and the Bayesian inferential process for detecting influential observations is provided. A simulation study with artificially perturbed data is conducted, followed by application in a real context of times until appendectomy surgery in Australian twins. It is also considered an analysis of the effects of a medication in the reaction times of lab mice with different lead concentrations in their bodies. The observed results demonstrate that the model is robust to outliers, making it a strong candidate in data contexts with strong right skewness. Additionally, distribution A proves to be flexible for lifetimes far from zero, unlike usual uniparametric distributions such as Exponential, Lindley, and Rayleigh.

Keywords: Survival analysis; Bayesian inference; influence analysis; copulas; bivariate data.

1 Introdução

A modelagem de dados de sobrevivência bivariados e multivariados no geral é um tópico muito desenvolvido na literatura e de crescente relevância no estudo da Estatística. Tempos de vida bivariados correspondem à presença de dois tempos de ocorrência distintos para um mesmo indivíduo. Tais problemas são muito comuns na área médica, envolvendo estudos relacionados a diversas categorias, como o estudo de pacientes com HIV positivo [28, 24], retinopatia diabética [32, 23], problemas renais [21] e ocorrência de cáries em crianças [9]. Em particular, neste trabalho exploramos os tempos até a realização da cirurgia de apendicectomia de gêmeos australianos, apresentado por [13], que enfatiza a presença de efeitos genéticos na doença. Ainda, consideramos a extensão do modelo proposto para o contexto de regressão, analisando tempos de reação de ratos de laboratório com diferentes níveis de concentração de chumbo no organismo. Este conjunto de dados foi analisado por [8], que considerou um modelo multivariado com marginais com distribuição Weibull. A partir destes exemplos podemos perceber a grande variedade de aplicações da teoria de análise de sobrevivência bivariada.

Além de ser muito relevante na área médica, esta área se estende a contextos mais amplos. [5] aplicam um modelo de fragilidade bivariado com fração de cura para a modelagem de tempos associados ao escore de crédito de clientes em uma empresa brasileira. [11] também passam a aplicar métodos baseados em cópulas para a modelagem de tempos relacionados a seguros de vida. O interesse recente na aplicação de modelos de sobrevivência para outros contextos além da saúde pode caracterizar uma nova grande expansão desta área do conhecimento para inúmeras outras áreas nos próximos anos.

Como pode ser percebido na breve revisão literária acima, duas grandes abordagens se sobressaem ao tratarmos dados de sobrevivência bivariados. A primeira consiste na indução de um modelo multivariado a partir da inserção de coeficientes de fragilidade ao modelo, o qual passa a ser responsável pela dependência entre as variáveis. A segunda abordagem, muito empregada em trabalhos recentes, é a suposição de que a dependência entre as variáveis modeladas pode ser descrita por uma família de funções cópulas. Esta abordagem possibilita uma grande flexibilidade na escolha das distribuições marginais dos dados, já que a escolha da função cópula pode ser feita de maneira completamente independente.

Este trabalho apresenta um novo modelo, baseado na cópula FGM, em que os modelos marginais são caracterizadas pela recente distribuição A. O artigo é dividido da seguinte forma: na seção 2 é apresentada a distribuição A e algumas propriedades de interesse, na seção 3 é formulado o modelo bivariado e estipulado o processo inferencial Bayesiano. Também introduzimos a metodologia de análise de observações influentes e os métodos de comparação entre diferentes modelos aplicados aos conjuntos de dados para a verificação de bondade de ajuste. A seção 4 conta com estudos de simulação do novo modelo e com duas aplicações em dados reais, correspondendo à idade em que gêmeos australianos passaram por cirurgias de apendicectomia e aos dados referentes a ratos de laboratório, mencionados acima. Na seção 5 é feita uma breve conclusão com os principais resultados encontrados com esta pesquisa.

2 Distribuição A: Propriedades

A distribuição A foi introduzida por [1] com o intuito de desenvolver um novo modelo que possa competir com outras alternativas uniparamétricas positivas, como a distribuição Exponencial e Rayleigh, por exemplo. Dizemos que a variável aleatória $X > 0$ segue uma distribuição da família A, se sua função densidade pode ser expressa por

$$f(x) = \frac{1}{x^2} \exp\left\{\frac{1}{\beta}\left(1 - \exp\left(\frac{\beta}{x}\right)\right) + \frac{\beta}{x}\right\},$$

com $x > 0$, $\beta > 0$. De fato, a família de distribuições A corresponde a um caso particular da família Gompertz inversa, com parâmetro de forma α igual a 1 (veja [14]). A partir desta definição, obtemos que a sua respectiva função de sobrevivência pode ser expressa como

$$S(x) = 1 - \exp\left\{\frac{1}{\beta}\left(1 - \exp\left(\frac{\beta}{x}\right)\right)\right\}, \qquad (1)$$

em que $x > 0$, $\beta > 0$. Na Figura 1 são apresentados os gráficos correspondentes às funções de

densidade, sobrevivência e risco. A seguir são dispostas algumas propriedades desta distribuição, as quais são mencionadas no artigo original de [1]:

- A sua função de risco aumenta até um ponto de máximo global e decresce a zero para valores de x superiores (forma de banheira invertida). A função de risco pode ser expressa por

$$h(x) = \frac{f(x)}{S(x)} = \frac{\exp\left(\frac{\beta}{x}\right)}{x^2\left[\exp\left(\frac{1}{\beta}\left(\exp\left(\frac{\beta}{x}\right)-1\right)\right)-1\right]}.$$

- O quantil q da distribuição A pode ser obtido por

$$x_q = \frac{\beta}{\log\left(1-\beta\log q\right)}, \quad 0 < q < 1.$$

- A moda da distribuição A pode ser obtida resolvendo a equação

$$\exp\left(\frac{\beta}{x}\right) - 2x - \beta = 0 \tag{2}$$

numericamente.

3 O modelo A bivariado baseado na cópula FGM

3.1 Definição do modelo

Uma função cópula C pode ser interpretada como uma distribuição multivariada de um vetor aleatório $\boldsymbol{U} = (U_1, \ldots, U_p)$, cujas marginais U_i, $i = 1, \ldots, p$ possuem distribuição $U(0,1)$, ou seja, a expressão explícita de uma cópula representa

$$C(u_1, \ldots, u_p) = P(U_1 \leq u_1, \ldots, U_p \leq u_p). \tag{3}$$

Com base nesta distribuição e no fato de que a função de distribuição $F(x) \sim U(0,1)$ para qualquer variável aleatória X com distribuição F, uma cópula pode ser utilizada diretamente para a construção de qualquer distribuição multivariada com marginais definidas, sendo a responsável pela atribuição da relação entre as distribuições marginais. No caso particular deste artigo, estaremos utilizando a cópula FGM, que caracteriza relações de dependência fraca entre as variáveis do modelo. A família de cópulas bivariadas, FGM, depende de um parâmetro de dependência $\phi \in [-1, 1]$ e pode ser expressa como

$$C_\phi(u_1, u_2) = u_1 u_2 \left[1 + \phi(1-u_1)(1-u_2)\right].$$

A baixa dependência entre as variáveis pode ser visualizada pela medida de concordância τ de Kendall (como descrito em [25]). Para o caso particular das cópulas FGM, temos $\tau_\phi = \frac{2\phi}{9}$, ou seja, a dependência entre as duas variáveis resposta do modelo varia no intervalo $\left[-\frac{2}{9}, \frac{2}{9}\right]$.

De forma mais formal, definiremos o modelo bivariado com base em sua função de sobrevivência. Definimos o vetor aleatório bidimensional $T = (X, Y)$, com distribuições marginais seguindo a distribuição A, a partir da função de sobrevivência

$$\begin{aligned} S(x, y|\boldsymbol{\theta}) &= P(X > x, Y > y) \\ &= C_\phi(S_x(x), S_y(y)), \end{aligned} \tag{4}$$

em que S_x e S_y representam a função de sobrevivência (1) com os parâmetros β_x e β_y, respectivamente. Equivalentemente, escrevemos a função densidade do novo modelo como

$$f(x, y|\boldsymbol{\theta}) = f_x(x) f_y(y) \left[1 + \phi(1 - 2S_x(x))(1 - 2S_y(y))\right].$$

De forma a flexibilizar o modelo para o contexto da análise de sobrevivência, consideramos a presença de dados censurados à direita. Considere que cada observação $t_i = (x_i, y_i)$ é acompanhada de um vetor indicador de censura $\delta_i = (\delta_{xi}, \delta_{yi})$ de modo que δ_{xi} e δ_{yi} são iguais a um, respectivamente se a primeira ou a segunda variável são censuradas para o i-ésimo indivíduo, sendo iguais a zero caso contrário. Considere também que o vetor de parâmetros a ser estimado é dado por $\boldsymbol{\theta} = (\beta_x, \beta_y, \phi)$. Neste caso, o logaritmo da função de verossimilhança pode ser expresso como

$$\begin{aligned} \ell(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y}) &= \sum_{i=1}^{n} \ell_i(x_i, y_i|\boldsymbol{\theta}) \\ &= \sum_{i=1}^{n} \Big\{ \delta_{xi}\delta_{yi} \log f_\phi(x, y) \\ &+ (1-\delta_{xi})(1-\delta_{yi}) \log S_\phi(x_i, y_i) \\ &+ \delta_{xi}(1-\delta_{yi}) \log\left(-\frac{\partial S_\phi(x_i, y_i)}{\partial x}\right) \\ &+ (1-\delta_{xi})\delta_{yi} \log\left(-\frac{\partial S_\phi(x_i, y_i)}{\partial y}\right) \Big\}. \end{aligned} \tag{5}$$

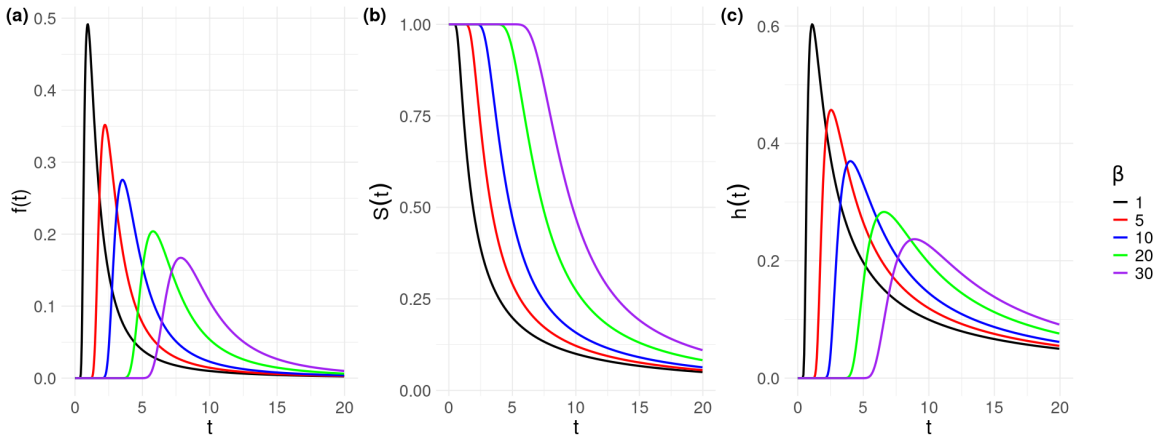

Figura 1: Gráficos da densidade **(a)**, função de sobrevivência **(b)** e risco **(c)** associadas à distribuição A para diferentes valores do parâmetro β.

3.2 Processo inferencial

Para as seções seguintes, a metodologia selecionada para a estimação de parâmetros é dada por métodos Bayesianos baseados no algoritmo Monte Carlo via Cadeias de Markov (MCMC), dada a complexidade matemática envolvida neste tipo de processo. Para a aplicação da metodologia Bayesiana, o primeiro passo é a definição das distribuições a priori selecionadas para o modelo. Para isso temos três parâmetros $\boldsymbol{\theta} = (\beta_x, \beta_y, \phi)$, definidos nos intervalos $(0, \infty)$, $(0, \infty)$ e $[-1, 1]$, respectivamente. Assumiremos que os parâmetros possuem as seguintes distribuições a priori:

$$\begin{aligned} \beta_x &\sim \text{Gama}(0{,}001; 0{,}001), \\ \beta_y &\sim \text{Gama}(0{,}001; 0{,}001), \\ \phi &\sim U(-1; 1). \end{aligned}$$

Estas distribuições a priori foram escolhidas com o propósito de reduzir o grau de informação atribuído a diferentes valores do espaço paramétrico. Embora a atribuição de uma priori não informativa seja um processo nada trivial e de grande polêmica, esta atribuição traz uma solução rápida e facilmente aplicável no contexto prático, além de ser muito comum e já reconhecida na literatura de trabalhos Bayesianos aplicados. O processo inferencial foi realizado utilizando o pacote *R2jags*, da linguagem R.

3.3 Análise de influência caso a caso

Uma vez definidas as distribuições a priori, um último ponto importante a ser destacado é a identificação de pontos influentes e possíveis *outliers* presentes no conjunto de dados. Uma forma geral de identificação de pontos influentes no contexto Bayesiano se dá na comparação entre distribuições a posteriori dos parâmetros em diferentes conjuntos de pontos utilizados para a estimação dos parâmetros do modelo. Para este trabalho utilizamos a métrica de divergência ψ introduzida por [26]. Seja $\mathcal{D}$ um conjunto de n observações e $\mathcal{D}_{(r)}$ o mesmo conjunto, porém omitindo a observação de índice r. Sendo $\pi_{\mathcal{D}}$ e $\pi_{\mathcal{D}_{(r)}}$ as distribuições a posteriori dos parâmetros correspondentes a ambos os conjuntos definidos, a medida de divergência ψ para a observação r é definida como

$$D_{\psi}(\pi_{\mathcal{D}}, \pi_{\mathcal{D}_{(r)}}) = \int \psi\left(\frac{\pi_{\mathcal{D}}(\boldsymbol{\theta}|\mathcal{D}_{(r)})}{\pi_{\mathcal{D}}(\boldsymbol{\theta}|\mathcal{D})}\right) \pi_{\mathcal{D}}(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta},$$

em que ψ representa uma função convexa qualquer, contanto que satisfaça $\psi(1) = 0$. Neste trabalho utilizaremos quatro métricas candidatas, sendo a divergência de Kullback-Leibler $(\psi(x) = -\log x)$, distância J $(\psi(x) = (x-1)\log x)$, divergência χ^2 $(\psi(x) = (x-1)^2)$ e norma L_1 $(\psi(x) = \frac{1}{2}|x-1|)$. A escolha destas funções é baseada nos trabalhos de [12] e [33].

Pode ser mostrado que a divergência ψ pode ser escrita como

$$D_{\psi}(\pi_{\mathcal{D}}, \pi_{\mathcal{D}_{(r)}}) = E^{\boldsymbol{\theta}|\mathcal{D}}\left[\psi\left(\frac{CPO_r}{f(y_r|\boldsymbol{\theta})}\right)\right],$$

em que y_r é a observação omitida em $\mathcal{D}_{(r)}$ e CPO_r representa a r-ésima ordenada preditiva, definida

inicialmente por [17] e escrita como

$$\begin{aligned}\text{CPO}_r &= \int f(y_r|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathcal{D}_{(r)})d\boldsymbol{\theta} \\ &= \left[\int \frac{\pi(\boldsymbol{\theta}|\mathcal{D})}{f(y_r|\boldsymbol{\theta})}d\boldsymbol{\theta}\right]^{-1}\end{aligned}$$

e f representa a função densidade do modelo. Veja que como consideramos a presença de dados censurados, podemos estender f para a contribuição que a r-ésima observação tem sobre a verossimilhança, ou seja, no contexto do modelo definido em (4), $f(y_r|\boldsymbol{\theta}) = \exp\{\ell_r(y_r|\boldsymbol{\theta})\}$, sendo ℓ_r o mesmo da expressão (5). Finalmente, obtendo uma amostra Monte Carlo de tamanho Q da distribuição a posteriori do vetor de parâmetros, $\boldsymbol{\theta}$, podemos estimar a divergência ψ da r-ésima observação como

$$\widehat{D}_{\psi}(\pi_{\mathcal{D}}, \pi_{\mathcal{D}_{(r)}}) = \frac{1}{Q}\sum_{q=1}^{Q}\psi\left(\frac{\widehat{CPO}_r}{f(\boldsymbol{z}_i|\boldsymbol{\theta}^{(q)})}\right), \quad (6)$$

$$\widehat{CPO}_i = \left[\frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{f(\boldsymbol{z}_r|\boldsymbol{\theta}^{(q)})}\right]^{-1}.$$

Uma vez calculados os valores das divergências para cada observação, estabelecemos que uma determinada observação do conjunto de dados é um ponto influente com base no método sugerido pelos autores da divergência ψ, em que o limiar para a divergência é definido analogamente ao viés, p, de uma moeda, sendo igual a

$$\tau = \frac{\psi(2p) + \psi(2(1-p))}{2}. \quad (7)$$

Ao longo deste trabalho, consideramos que $p = 0.9$. Consequentemente, cada uma das quatro métricas ψ selecionadas apresentará um limiar correspondente, sendo possível a detecção de um ponto como influente por uma métrica e não influente por outra.

3.4 Métricas de comparação de modelos

De forma a verificarmos a superioridade do modelo proposto em relação a outras alternativas de distribuições uniparamétricas, utilizaremos métricas usuais de avaliação de performance de modelos Bayesianos, baseadas na função *deviance*. As métricas escolhidas foram o LPML (*Log pseudo marginal likelihood*), DIC (*Deviance information criterion*) [29], EAIC (*Expected Akaike Information Criterion*) [4] e EBIC (*Expected Bayesian Information Criterion*) [6].

Todos os demais modelos comparados ao modelo A bivariado são construídos de forma completamente análoga a (4), com a modificação da especificação das distribuições marginais, i.e. consideramos os modelos bivariados Exponencial , Lindley e Rayleigh. O uso do modelo exponencial na literatura de sobrevivência é extremamente vasto, tendo suas origens em trabalhos como [10, 15]. A distribuição Lindley foi estudada para o contexto de sobrevivência por [19] e tem inúmeras modificações, como abordado em [18, 2, 16]. Finalmente, a distribuição Rayleigh também é muito empregada na literatura em trabalhos como [22, 7]. Mais trabalhos sobre o uso dessas distribuições são referenciados por [1].

4 Resultados

4.1 Estudo de simulação

Foram realizadas 1000 réplicas do modelo A bivariado considerando como parâmetros reais os valores $\beta_x = 2$, $\beta_y = 5$ e $\phi = 0{,}6$. Foram realizados estudos considerando duas diferentes configurações de censura para as observações do vetor (X, Y), sendo respectivamente, $(0\%, 0\%)$ e $(20\%, 15\%)$. Em cada configuração foram avaliadas a média das estimativas pontuais (média a posteriori) dos parâmetros em conjunto com o viés, erro quadrático médio e probabilidade de cobertura do intervalo de credibilidade 95% desses estimadores. Note que consideramos implicitamente uma função perda quadrática para a estimação dos parâmetros do modelo neste caso.

As Tabelas 1 e 2 representam os resultados simulados das duas diferentes configurações de censura propostas. Como pode ser verificado, todos os resultados são esperados, com a diminuição do erro quadrático médio (EQM) à medida que o tamanho amostral aumenta e uma variação da probabilidade de cobertura em torno do coeficiente de credibilidade escolhido.

Para o estudo de simulação referente à análise de diagnóstico de observações influentes, é muito comum a utilização de perturbações em tempos individuais a partir da soma do desvio-padrão da amostra multiplicado por um fator fixo, usual-

Tabela 1: Resultados dos estudos de simulação Bayesiana considerando apenas observações não censuradas.

n	Parâmetro	Média	Vício	EQM	Probabilidade de Cobertura
50	β_x	2,0012	0,0012	0,0735	0,941
	β_y	4,9900	-0,0100	0,1854	0,949
	ϕ	0,4234	-0,1766	0,0944	0,977
100	β_x	1,9979	-0.0021	0,0380	0,930
	β_y	4,9951	-0,0049	0,0938	0,94
	ϕ	0,4900	-0,1100	0,0596	0,967
200	β_x	1,9976	-0,0024	0,0157	0,960
	β_y	5,0028	0,0028	0,0438	0,949
	ϕ	0,5640	-0,0360	0,0312	0,961
400	β_x	1,9943	-0,0057	0,0086	0,940
	β_y	4,9969	-0,0031	0,0235	0,937
	ϕ	0,5884	-0,0116	0,0179	0,944

Tabela 2: Resultados dos estudos de simulação Bayesiana considerando censuras de 20% para X e 15% para Y.

n	Parâmetro	Média	Vício	EQM	Probabilidade de Cobertura
50	β_x	1,9971	-0,0029	0,0736	0,952
	β_y	5,0239	0,0239	0,2021	0,955
	ϕ	0,3863	-0,2137	0,1254	0,958
100	β_x	2,0099	0,0099	0,0370	0,950
	β_y	5,0007	0,0007	0,0934	0,952
	ϕ	0,4885	-0,1115	0,0644	0,965
200	β_x	1,9977	-0,0023	0,0172	0,951
	β_y	5,0006	0,0006	0,0447	0,949
	ϕ	0,5574	-0,0426	0,0328	0,961
400	β_x	1,9990	-0,0010	0,0083	0,961
	β_y	5,0027	0,0027	0,0231	0,941
	ϕ	0,5798	-0,0202	0,0217	0,938

mente igual a 5 (veja [31, 30, 27, 3]). Considerando a distribuição A, entretanto, essa perturbação apenas revela a robustez do modelo a *outliers* muito altos. Para verificarmos essa robustez, consideramos uma amostra de tamanho $n = 400$ da distribuição A bivariada considerando os mesmos parâmetros do estudo acima. A observação de índice 305 foi selecionada, tendo ambos os seus tempos de vida perturbados segundo as transformações $\widetilde{x}_{305} = x_{305} + 25S_x$ e $\widetilde{y}_{305} = y_{305} + 25S_y$, em que S_x e S_y representam os desvios-padrão amostrais dos tempos de X e Y.

Para a grande maioria dos modelos, a transformação proposta acima deveria resultar em valores de divergência altíssimas para a observação de índice 305. Pelo contrário, como evidenciado pela Tabela 3, percebemos que a divergência estimada segundo todas as funções ψ candidatas resultante é menor que a divergência para a observação original.

Assim, temos indícios de que o uso da distribuição A leva a modelos extremamente robustos a tempos de vida muito altos. Em contrapartida, percebe-se uma sensibilidade considerável para tempos de vida inferiores. Consideramos agora uma simulação em que observações selecionadas ao acaso são perturbadas para serem ligeiramente inferiores às demais. Para isso, utilizamos a perturbação representada por $\widetilde{x}_i = 0{,}3x_{(1)}F_x(x_i) + 0{,}65x_{(1)}$ e $\widetilde{y}_i = 0{,}3y_{(1)}F_x(y_i) + 0{,}65y_{(1)}$, em que F_x e F_y representam as funções de distribuição de ambas as variáveis X e Y, $x_{(1)}$ e $y_{(1)}$ representam as primeiras estatísticas de ordem amostrais. Essas transformações mapeiam os pontos para um espaço ligeiramente inferior à menor observação amostrada. Dessa vez, de forma a simplificar o intervalo de amostragem utilizamos os parâmetros $\beta_x = 150$, $\beta_y = 300$ e $\phi = 0{,}6$ sem censuras. Foram verificadas todas as diferentes combinações de perturbação para os pontos de índice 150, 305 e 331. Como mostrado na Tabela 4, todas as observações perturbadas foram identificadas como pontos influentes. A Figura 2 mostra um gráfico de influências para as divergências de todas as observações no caso h.

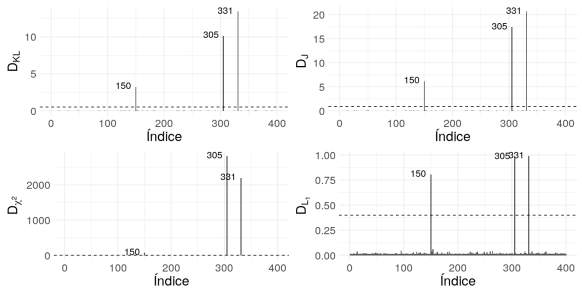

Figura 2: Gráficos das divergências ψ para todas as observações do modelo (h). Linhas pontilhadas representam os limites obtidos em (7).

4.2 Apendicectomia em gêmeos australianos

Para a aplicação do novo modelo proposto a um contexto real, temos um conjunto de dados com as idades em que gêmeos australianos realizaram a apendicectomia, a operação de remoção do apêndice. Os dados foram analisados inicialmente por [13] e uma amostra do conjunto de dados em questão foi apresentada por [20]. Essa amostra foi utilizada para a aplicação do modelo A bivariado. Temos um total de 174 observações, das quais 127 são não censuradas e 47 envolvem censuras à esquerda de um ou ambos os tempos. Uma vez que o modelo proposto admite apenas censuras à direita as observações censuradas à esquerda foram desconsideradas na análise. Como mencionado na seção 3.4, comparamos o ajuste do novo modelo proposto a distribuições uniparamétricas já conhecidas. A Tabela 5 mostra o desempenho estimado de cada modelo segundo as amostras MCMC das suas respectivas distribuições a posteriori.

Vemos que o modelo A bivariado supera os demais candidatos segundo todas as métricas de desempenho propostas. Uma comparação visual dos modelos pode ser feita observando-se as curvas de sobrevivência sobrepostas em conjunto com o estimador não paramétrico de Kaplan-Meier, na Figura 3. Uma tendência dos modelos uniparamétricos usuais ao lidar com tempos de vida em tempos distantes de zero é o decrescimento precoce da curva de sobrevivência, o que evidentemente não ocorre ao considerarmos a distribuição A.

A Figura 4 mostra divergências estimadas para cada dupla de gêmeos do conjunto de dados. Vemos que apenas uma observação (índice 16) foi detectada como um ponto potencialmente influente

Tabela 3: Medidas de divergência para a observação 305.

Modelo	D_{KL}	D_J	D_{χ^2}	D_{L_1}
original	0,0318	0,0643	0,069	0,1006
perturbado	0,0029	0,0058	0,006	0,0305

Tabela 4: Medidas de divergência estimadas para os pontos perturbados escolhidos em diferentes cenários de perturbação.

Modelo	Casos perturbados	Índice do caso	D_{KL}	D_J	D_{χ^2}	D_{L_1}
		150	0.0036	0.0071	0.0072	0.0339
a	nenhum	305	0,0249	0,0501	0,0520	0,0893
		331	0,0265	0,0533	0,0554	0,0920
		150	**7,6147**	**12,6734**	**389,8565**	**0,9426**
b	150	305	0,0226	0,0455	0,0477	0,0848
		331	0,0240	0,0484	0,0509	0,0875
		150	0,0005	0,0010	0,0010	0,0125
c	305	**305**	**21.1428**	**28,0934**	**1731,7250**	**0,9915**
		331	0,0030	0,0060	0,0061	0,0307
		150	0,0002	0,0005	0,0005	0,0085
d	331	305	0,0022	0,0044	0,0044	0,0265
		331	**22,7214**	**30,2258**	**2226,7870**	**0,9935**
		150	**6,2813**	**10,7481**	**273,5020**	**0,9146**
e	{150; 305}	**305**	**17,9534**	**23,9332**	**605,2464**	**0,9887**
		331	0,0022	0,0044	0,0044	0,0264
		150	**4,1776**	**8,1690**	**268,5590**	**0,8620**
f	{150; 331}	305	0,0022	0,0045	0,0045	0,0268
		331	**21,1165**	**28,6304**	**2993,7000**	**0,9934**
		150	0,0003	0,0006	0,0006	0,0094
g	{305; 331}	**305**	**9,2180**	**14,8946**	**523,7004**	**0,9661**
		331	**16,1159**	**24,2938**	**4509,5210**	**0,9926**
		150	**3,2323**	**6,1541**	**70,3742**	**0,8050**
h	{150; 305; 331}	**305**	**10,1450**	**17,3943**	**2816,3690**	**0,9790**
		331	**13,4196**	**20,6473**	**2194,5410**	**0,9889**

Tabela 5: Informações dos ajustes dos quatro modelos propostos segundo a abordagem Bayesiana.

Modelo Bivariado	LPML	EAIC	EBIC	DIC
A	-916,4672	1834,3590	1842,8910	1831,1770
Exponencial	-1042,5330	2090,4930	2099,0250	2086,4800
Lindley	-979,2999	1963,6270	1972,1590	1959,7180
Rayleigh	-944,3970	1892,1580	1900,6900	1888,4300

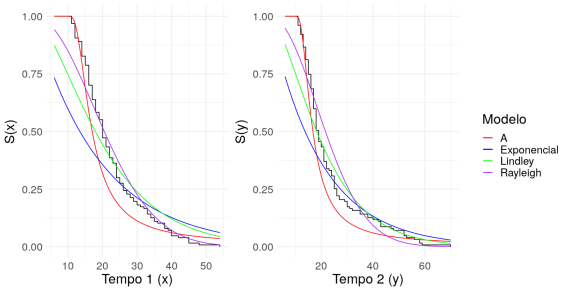

Figura 3: Curvas de Sobrevivência dos modelos ajustados e estimador de Kaplan-Meier.

segundo o critério de divergência χ^2. De fato, este par corresponde a uma dupla de gêmeos que foram submetidos à cirurgia ambos com 11 anos de idade. Condizendo com o verificado na Seção 4.1, este ponto representa os menores tempos observados em todo o conjunto de dados.

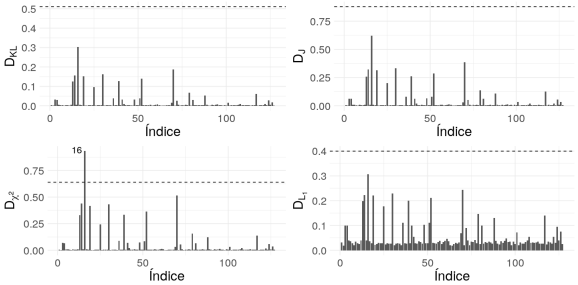

Figura 4: Gráficos das divergências ψ para todas as observações do modelo A bivariado ajustado.

Ao removermos a observação de índice 16 do conjunto de dados e ajustando o modelo novamente, obtemos uma melhora geral no ajuste, como sugerido pela Tabela 6. A Figura 5 exibe as divergências estimadas para as demais observações, após retirada a observação de índice 16 do conjunto de dados. De fato, as observações com maior divergência todas correspondem a gêmeos que fizeram a cirurgia de apendicectomia quando muito jovens. Entretanto, suas divergências não são o suficiente para serem identificados como pontos anormais, segundo o critério proposto neste trabalho. Vale destacar que do ponto de vista médico, a simples remoção de uma observação do conjunto de dados exige uma justificativa aceitável, mas no contexto puramente de modelagem, a retirada da observação de índice 16 levou a um desempenho levemente superior.

Finalmente, a Tabela 7 apresenta as estimativas pontuais a posteriori de cada um dos três

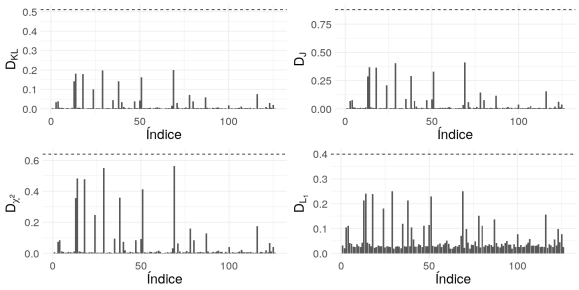

Figura 5: Gráficos das divergências ψ para todas as observações do modelo A bivariado após a remoção da observação 16.

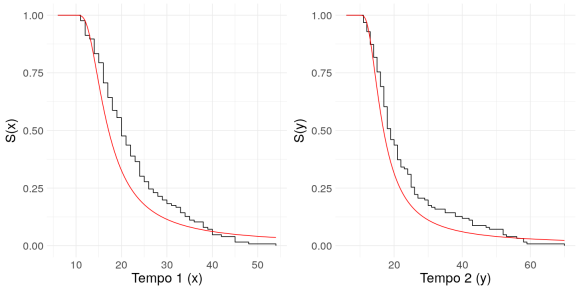

Figura 6: Curvas de Sobrevivência dos modelos ajustados e estimador de Kaplan-Meier após a remoção da observação 16.

parâmetros do modelo A bivariado. Concluímos, a partir do valor positivo de ϕ, que temos indícios de que exista uma associação fraca, porém positiva e significativa entre o tempo necessário de realização da cirurgia de apendicectomia dos irmãos gêmeos analisados. Essa associação pode indicar, por exemplo, que dada a necessidade de cirurgia de um indivíduo em uma determinada idade, o seu irmão gêmeo pode ter o mesmo problema em algum tempo próximo futuro, o que nos permite a manutenção de possíveis tratamentos do indivíduo com antecedência.

A Figura 6 apresenta a curva de sobrevivência estimada final para este conjunto de dados. Vale destacar que o modelo A bivariado pode não ser capaz de capturar completamente o comportamento dos tempos de vida, subestimando levemente os anos da realização das cirurgias. Este problema possivelmente está associado ao fato do modelo proposto apresentar apenas três parâmetros, ou seja, modelos com mais parâmetros neste contexto poderiam apresentar um desempenho mais satisfatório. De certa forma, ainda assim obtivemos um desempenho superior com relação às

Tabela 6: Comparação do modelo A bivariado com e sem a observação 16

Modelo Bivariado	LPML	EAIC	EBIC	DIC
Com a observação 16	-916,4672	1834,3590	1842,8910	1831,1770
Sem a observação 16	-907,8039	1817,1518	1825,6606	1813,9403

Tabela 7: Resumo a *posteriori* dos parâmetros do modelo final.

Parâmetro	Média	Desvio Padrão	Limite Inferior (2,5%)	Limite Superior (97,5%)
β_x	65,9641	1,5612	62,7909	68,8904
β_y	64,7366	1,5766	61,5535	67,6774
ϕ	0,6352	0,1823	0,2499	0,9505

distribuições uniparamétricas usuais, o que pode ser uma grande vantagem em contextos em que o desempenho computacional deve ser priorizado, como em grandes sistemas automatizados de monitoramento de tempos de confiabilidade, por exemplo.

4.3 Níveis de chumbo em ratos de laboratório

Na maioria das aplicações em dados reais, existirão covariáveis de interesse associadas aos dados, as quais podem influenciar diretamente o comportamento das variáveis de interesse. Um exemplo é o conjunto de dados utilizado por [8], que conta com o estudo de tempos de reação a um estímulo tátil em ratos antes e após a aplicação de uma dose de um medicamento específico. Os ratos foram separados em três grupos segundo diferentes níveis de chumbo presente em seus organismos. De forma a considerar o grupo dos ratos, assim como o nível da dose administrada, consideramos a seguinte reparametrização dos parâmetros β_x e β_y:

$$\beta_x = \exp\left\{\gamma_{0x} + \gamma_{2x} w_2 + \gamma_{3x} w_3\right\},$$
$$\beta_y = \exp\{\gamma_{0y} + \gamma_{11y} w_1 + \gamma_{12y} w_1^2 + \gamma_{13y} w_1^3 + \gamma_{2y} w_2 + \gamma_{3y} w_3\}.$$

Neste caso, w_1 representa a dose do medicamento aplicado e w_2 e w_3 representam variáveis binárias indicadoras do grupo ao qual cada rato pertence, sendo ambas iguais a zero para o primeiro grupo. Neste caso, os tempos (X, Y) representam, respectivamente, o tempo de reação do rato antes e após a aplicação do medicamento. Por esta razão o parâmetro β_x não depende da dose administrada. Os mesmos parâmetros β_x e β_y foram considerados para outros modelos bivariados com marginais uniparamétricas. A Tabela 8 apresenta a comparação dos ajustes para este conjunto de dados. O modelo A obteve o melhor ajuste segundo os critérios de desempenho escolhidos.

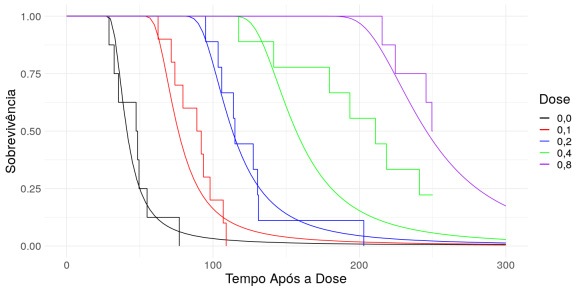

Figura 7: Curvas de sobrevivência para diferentes doses e ajuste do modelo A bivariado para a regressão nos ratos do grupo 1.

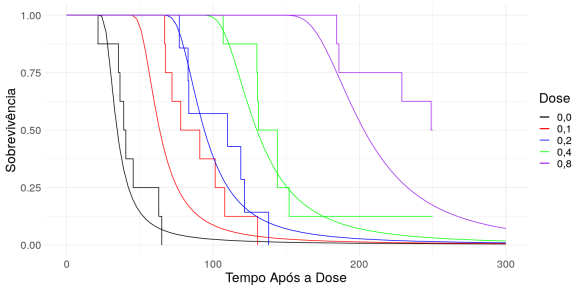

Figura 8: Curvas de sobrevivência para diferentes doses e ajuste do modelo A bivariado para a regressão nos ratos do grupo 2.

A Tabela 9 apresenta o resumo a posteriori dos parâmetros estimados. Veja que, com exceção do parâmetro γ_{3x}, nenhum outro intervalo de credibilidade contém o zero, dando indícios de que os valores obtidos são estatisticamente significativos. Veja, assim que γ_{2y} e γ_{3y}, sendo negativos, indica uma leve redução do tempo de vida para os ratos do grupo 2 e 3, o que indica que o efeito do medicamento diminui para maiores níveis de chumbo nos ratos, assim como concluído por [8] em seu artigo original, já que a droga tem efeito de retardar

Tabela 8: Informações dos ajustes dos quatro modelos de regressão propostos.

Modelo Bivariado	LPML	EAIC	EBIC	DIC
A	-1071,487	2122,957	2151,320	2118,850
Exponencial	-1235,302	2488,554	2516,917	2477,988
Lindley	-1158,362	2334,493	2362,856	2322,824
Rayleigh	-1098,648	2212,758	2241,121	2202,432

Tabela 9: Resumo a *posteriori* dos parâmetros do modelo de regressão A bivariado.

Parâmetro	Média	Desvio Padrão	Limite Inferior (2,5%)	Limite Superior (97,5%)
γ_{0x}	5,0972	0,4717	3,1052	5,2439
γ_{2x}	0,2614	0,4724	0,0802	2,2919
γ_{3x}	0,0737	0,4718	-0,1097	2,0726
γ_{0y}	5,3194	0,0443	5,2273	5,4035
γ_{11y}	9,2395	1,2174	6,4174	10,7704
γ_{12y}	-17,7668	4,5445	-22,9529	-7,0870
γ_{13y}	11,9604	3,8532	2,9643	16,3180
γ_{2y}	-0,2221	0,0403	-0,3009	-0,1413
γ_{3y}	-0,2345	0,0402	-0,3128	-0,1532
ϕ	0,6494	0,1833	0,2502	0,9520

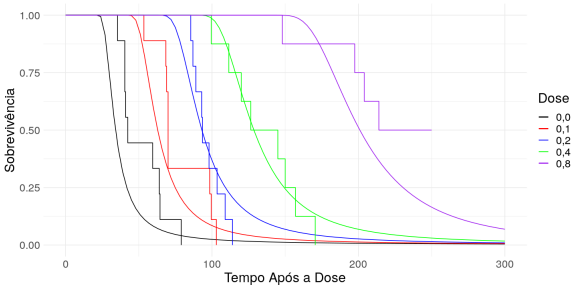

Figura 9: Curvas de sobrevivência para diferentes doses e ajuste do modelo A bivariado para a regressão nos ratos do grupo 3.

a resposta dos espécimes ao estímulo. Podemos visualizar esse fato ao comparar as curvas de sobrevivência entre as diferentes doses para os três grupos de ratos, dispostas nas Figuras 7, 8 e 9.

Veja que de fato a distância entre as curvas é menor entre os grupos 2 e 3, indicando que o medicamento se tornou menos eficaz nesses grupos. Outro ponto destacado pelo artigo original indica uma não surpresa com a concordância positiva entre o tempo antes e após a aplicação da dose do medicamento, que de fato é verificada pelo valor positivo da média a posteriori $\widehat{\phi} = 0{,}6494$, indicando que o tempo de reação antes e depois da aplicação do medicamento tendem a serem concordantes entre si, o que é esperado já que cada dupla de tempos corresponde a um mesmo indivíduo. Um ponto a ser destacado é que em certos casos a curva de sobrevivência proposta pelo modelo se localiza relativamente distante da curva de Kaplan-Meier dos dados (Dose 0,1 para o grupo 2 dos ratos, por exemplo). Isso poderia ser resolvido aumentando o grau do polinômio ajustado à dose, entretanto, optou-se pela definição de um modelo mais simples de forma a evitar um possível sobreajuste aos dados, neste caso.

5 Conclusões

Neste trabalho foi apresentado o desenvolvimento de um novo modelo com o propósito de modelar tempos de vida bivariados. Foi feita uma breve introdução da recente distribuição uniparamétrica entitulada distribuição A, proposta por [1], e foi apresentada a metodologia de identificação de pontos influentes baseada na divergência ψ de [26].

Diferente da grande maioria dos modelos de sobrevivência usualmente empregados, foi possível mostrar através de estudos de simulação, que a distribuição A demonstra grande robustez para tempos de vida muito altos, o que pode indicar boas propriedades para o uso desta distribuição em estudos de valores extremos ou em conjuntos de dados com uma concentração de *outliers*.

Uma aplicação de relevância médica foi apresentada, chegando à conclusão de que de fato existe uma associação entre as idades em que gêmeos ne-

cessitam de realizar a cirurgia de apendicectomia. A distribuição A supera, neste caso, outras distribuições uniparamétricas usualmente empregadas, mostrando uma flexibilidade no ponto crítico em que a curva de sobrevivência começa a decrescer. Também foi possível a reprodução dos resultados obtidos por [8] no estudo dos efeitos de um medicamento no tempo de reação de ratos com diferentes níveis de chumbo no corpo. A conclusão de que a concentração de chumbo no organismo de ratos reduziu a eficácia do medicamento proporciona a reflexão da extensão dos efeitos que o chumbo pode causar na eficácia de medicamentos em seres humanos, principalmente no contexto de medicamentos anestésicos.

Este trabalho apresenta o grande potencial da modelagem de dados de sobrevivência utilizando funções cópulas como ferramentas de construção de modelos multivariados, como também atesta pontos positivos para o uso da nova distribuição A em contextos aplicados da área de sobrevivência. De fato, temos indícios de que esta distribuição seja promissora ao tratarmos cenários de observações extremas, dada a sua grande robustez a observações *outliers* muito altas no conjunto de dados, o que é um ponto em aberto para possíveis trabalhos futuros.

Agradecimentos

Realizamos um agradecimento especial à Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), que financiou o desenvolvimento deste projeto.

Referências

[1] Alshenawy, R. (2020). A new one parameter distribution: properties and estimation with applications to complete and type ii censored data. *Journal of Taibah University for Science*, 14(1):11–18.

[2] Barco, K. V. P., Mazucheli, J., and Janeiro, V. (2017). The inverse power lindley distribution. *Communications in Statistics-Simulation and Computation*, 46(8):6308–6323.

[3] Biondo, T. R. (2020). *Modelos de sobrevivência bivariados baseados na cópula PVF*. PhD thesis, Universidade Federal de São Carlos.

[4] Brooks, S., Smith, J., Vehtari, A., Plummer, M., Stone, M., Robert, C. P., Titterington, D., Nelder, J., Atkinson, A., Dawid, A., et al. (2002). Discussion on the paper by spiegelhalter, best, carlin and van der linde. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 64(4):616–639.

[5] Cancho, V. G., Suzuki, A. K., Barriga, G. D., and Louzada, F. (2016). A non-default fraction bivariate regression model for credit scoring: An application to brazilian customer data. *Communications in Statistics: Case Studies, Data Analysis and Applications*, 2(1-2):1–12.

[6] Carlin, B. P. (2000). *Bayes and empirical Bayes methods for data analysis*. Chapman & Hall.

[7] Cordeiro, G. M., Cristino, C. T., Hashimoto, E. M., and Ortega, E. M. (2013). The beta generalized rayleigh distribution with applications to lifetime data. *Statistical papers*, 54:133–161.

[8] Crowder, M. (1989). A multivariate distribution with weibull connections. *Journal of the Royal Statistical Society: Series B (Methodological)*, 51(1):93–107.

[9] Cruz, J. N. d., Ortega, E. M., Cordeiro, G. M., Suzuki, A. K., and Mialhe, F. L. (2017). Bivariate odd-log-logistic-weibull regression model for oral health-related quality of life. *Communications for Statistical Applications and Methods*, 24(3):271–290.

[10] Davis, D. (1952). An analysis of some failure data. *Journal of the American Statistical Association*, 47(258):113–150.

[11] Deresa, N. W., Van Keilegom, I., and Antonio, K. (2022). Copula-based inference for bivariate survival data with left truncation and dependent censoring. *Insurance: Mathematics and Economics*, 107:1–21.

[12] Dey, D. K. and Birmiwal, L. R. (1994). Robust bayesian analysis using divergence measures. *Statistics & Probability Letters*, 20(4):287–294.

[13] Duffy, D. L., Martin, N. G., and Mathews, J. D. (1990). Appendectomy in australian twins. *American Journal of Human Genetics*, 47(3):590.

[14] Eliwa, M., El-Morshedy, M., and Ibrahim, M. (2019). Inverse gompertz distribution: properties and different estimation methods with ap-

plication to complete and censored data. *Annals of data science*, 6:321–339.

[15] Epstein, B. and Sobel, M. (1953). Life testing. *Journal of the American Statistical Association*, 48(263):486–502.

[16] Fernandes, P. G., Suzuki, A. K., and Saraiva, E. F. (2018). O modelo lindley-weibull com proporçao de cura: uma abordagem bayesiana. *Brazilian Journal of Biometrics*, 36(4):998–1022.

[17] Geisser, S. and Box, G. (1980). Discussion on sampling and bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society: Series A*, 143:416–417.

[18] Ghitany, M. E., Al-Mutairi, D. K., Balakrishnan, N., and Al-Enezi, L. (2013). Power lindley distribution and associated inference. *Computational Statistics & Data Analysis*, 64:20–33.

[19] Ghitany, M. E., Atieh, B., and Nadarajah, S. (2008). Lindley distribution and its application. *Mathematics and computers in simulation*, 78(4):493–506.

[20] Gleeja, V. and Sankaran, P. (2008). *Modelling and analysis of bivariate lifetime data using reversed hazard rates*. PhD thesis, Cochin University of Science & Technology.

[21] Hanagal, D. D. and Pandey, A. (2015). Gamma frailty models for bivariate survival data. *Journal of Statistical Computation and Simulation*, 85(15):3172–3189.

[22] Kundu, D. and Raqab, M. Z. (2005). Generalized rayleigh distribution: different methods of estimations. *Computational statistics & data analysis*, 49(1):187–200.

[23] Louzada, F., Suzuki, A., and Cancho, V. (2013). The fgm long-term bivariate survival copula model: modeling. *Bayesian estimation, and case infuence diagnostics*.

[24] Louzada, F., Suzuki, A. K., Cancho, V. G., Prince, F. L., Pereira, G. A., et al. (2012). The long-term bivariate survival fgm copula model: an application to a brazilian hiv data. *Journal of Data Science*, 10(3):511–535.

[25] Nelsen, R. B. (2006). *An introduction to copulas*. Springer.

[26] Peng, F. and Dey, D. K. (1995). Bayesian analysis of outlier problems using divergence measures. *Canadian Journal of Statistics*, 23(2):199–213.

[27] Ribeiro, T. R., Suzuki, A. K., and Saraiva, E. F. (2017). Uma abordagem bayesiana para o modelo de sobrevivência bivariado derivado da cópula amh. *Revista da Estatística da Universidade Federal de Ouro Preto*, 6.

[28] Shih, J. H. and Louis, T. A. (1995). Inferences on the association parameter in copula models for bivariate survival data. *Biometrics*, pages 1384–1399.

[29] Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology)*, 64(4):583–639.

[30] Suzuki, A. K., Barriga, G. D., Louzada, F., and Cancho, V. G. (2017). A general long-term aging model with different underlying activation mechanisms: Modeling, bayesian estimation, and case influence diagnostics. *Communications in Statistics-Theory and Methods*, 46(6):3080–3098.

[31] Suzuki, A. K., Cancho, V. G., and Louzada, F. (2016). The poisson–inverse-gaussian regression model with cure rate: a bayesian approach and its case influence diagnostics. *Statistical Papers*, 57(1):133–159.

[32] Suzuki, A. K., Louzada-Neto, F., Cancho, V. G., and Barriga, G. (2011). The fgm bivariate lifetime copula model: A bayesian approach. *Advances and Applications in Statistics*, 21(1):55–76.

[33] Weiss, R. (1996). An approach to bayesian sensitivity analysis. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(4):739–750.