UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Modelos de Sobrevivência Bivariados Baseados na Cópula PVF

Thiago Ramos Biondo

Dissertação de Mestrado do Programa Interinstitucional de Pós-Graduação em Estatística (PIPGEs)

ICMC USP SÃO CARLOS

Thiago Ramos Biondo

Modelos de Sobrevivência Bivariados Baseados na Cópula PVF

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP e ao Departamento de Estatística – DEs-UFSCar, como parte dos requisitos para obtenção do título de Mestre em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística. *VERSÃO REVISADA*

Área de Concentração: Estatística

Orientador: Prof. Dr. Adriano Kamimura Suzuki

USP – São Carlos
Março de 2020

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

Biondo, Thiago Ramos
B615m Modelos de Sobrevivência Bivariados Baseados na Cópula PVF / Thiago Ramos Biondo; orientador Adriano Kamimura Suzuki. -- São Carlos, 2020.
106 p.

Dissertação (Mestrado - Programa Interinstitucional de Pós-graduação em Estatística) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2020.

1. Análise de Sobrevivência. 2. Funções Cópulas. 3. Cópula PVF. 4. Inferência Bayesiana. 5. Simulação. I. Suzuki, Adriano Kamimura, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Thiago Ramos Biondo

Bivariate Survival Models Based on PVF Copula

Dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP and to the Department of Statistics – DEs-UFSCar – in accordance with the requirements of the Statistics Interagency Graduate Program, for the degree of Master in Statistics. *FINAL VERSION*

Concentration Area: Statistics

Advisor: Prof. Dr. Adriano Kamimura Suzuki

USP – São Carlos
March 2020

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa Interinstitucional de Pós-Graduação em Estatística

Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Dissertação de Mestrado do candidato Thiago Ramos Biondo, realizada em 13/03/2020:

Prof. Dr. Adriano Kamimura Suzuki
USP

Prof. Dr. Paulo Henrique Ferreira da Silva
UFBA

Prof. Dr. Erlandson Ferreira Saraiva
UFMS

Certifico que a defesa realizou-se com a participação à distância do(s) membro(s) Erlandson Ferreira Saraiva e, depois das arguições e deliberações realizadas, o(s) participante(s) à distância está(ao) de acordo com o conteúdo do parecer da banca examinadora redigido neste relatório de defesa.

Prof. Dr. Adriano Kamimura Suzuki

Dedico esse trabalho em memória de minha mãe Alice, que sempre me incentivou e me deu as melhores condições possíveis para, a cada dia, me tornar uma pessoa melhor e alcançar tudo o que já conquistei e que ainda sonho em conquistar.

AGRADECIMENTOS

À minha família, por sempre me apoiar;

Aos meus amigos, em especial aos da minha turma de graduação e de mestrado, pelos estudos e aprendizados que, assim como suas amizades, levarei para toda a vida;

Ao programa Interinstitucional de Pós-Graduação em Estatística e a cada um de seus docentes, por colaborarem na minha formação;

Ao Prof. Dr. Pablo Martin Rodriguez que foi o primeiro a me incentivar e mostrar a importância da pesquisa acadêmica, e ao Prof. Dr. Mário de Castro Andrade Filho por ser um exemplo de pessoa e profissional que muito me inspira;

Finalmente, ao meu orientador e amigo Prof. Dr. Adriano Kamimura Suzuki pela sua orientação, que sempre foi fundamental para o meu crescimento, e por ser exemplo de comprometimento e dedicação profissional.

“As invenções são, sobretudo,
o resultado de um trabalho de teimoso.”
(Santos Dumont)

RESUMO

BIONDO, T. R. **Modelos de Sobrevivência Bivariados Baseados na Cópula PVF**. 2020. 106 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2020.

Uma alternativa desenvolvida para estudar associações entre os tempos de sobrevivência multivariados é o uso dos modelos baseados em funções cópulas.

Neste trabalho, utilizamos o modelo de sobrevivência derivado da cópula PVF, baseada na distribuição *Power Variance Function*, para modelar a dependência de dados bivariados 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 são 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, ilustramos a aplicabilidade dos modelos propostos a conjuntos de dados simulados e reais.

Palavras-chave: Análise de Sobrevivência, Funções cópulas, Cópula PVF, Inferência Bayesiana, Simulação.

ABSTRACT

BIONDO, T. R. **Bivariate Survival Models Based on PVF Copula**. 2020. 106 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2020.

An alternative developed to study associations among multivariate survival times is the use of models based on copula functions.

In this work, we use the survival model derived from the PVF copula, based on the Power Variance Function distribution, to model the dependence of bivariate data in the presence of covariates and censored observations. For inferential purposes, we perform a Bayesian approach using Monte Carlo Markov Chain (MCMC) methods. Some discussions about model selection criteria are presented. In order to detect influential observations, we used the Bayesian method of deletion influence analysis of cases based on divergence ψ. Finally, we show the applicability of the proposed models to simulated and real datasets.

Keywords: Survival analysis, Copula functions, PVF Copula, Bayesian Inference, Simulation.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

A Análise de Sobrevivência se caracteriza como um conjunto de técnicas que visam estudar os tempos até a ocorrência de um evento de interesse, como por exemplo, em estudos médicos temos o tempo até a morte, início de uma doença ou o tempo de sobrevida dos pacientes a partir do início de um tratamento. Em aplicações de engenharia, o tempo até falha(s) de sistemas mecânicos. Em análise de risco, o tempo transcorrido até uma pessoa se tornar inadimplente. E em análise de confiabilidade, a capacidade de um produto funcionar bem durante um período de tempo especificado. Em Lawless (2003), Louzada et al. (2002a) e Rodrigues et al. (2008) encontramos uma base teórica geral para tratar de análise de sobrevivência e confiabilidade. Freitas & Colosimo (1997) fazem uma análise de tempo de falha em testes de vida acelerados. Nelson (1990) oferece aos engenheiros, cientistas e estatísticos um recurso confiável sobre o uso eficaz de testes de vida acelerados para medir e melhorar a confiabilidade de produtos. Meeker & Escobar (1998) apresentam métodos estatísticos computacionais para análise de dados de confiabilidade e planejamento de testes para produtos industriais. Carvalho et al. (2011) trazem teoria e aplicações na área de saúde. César (2005) e Gyõrffy et al. (2013) apresentam aplicações de modelos de sobrevivência para dados de câncer de mama e de pulmão, respectivamente. Louzada et al. (2002b) analisam dados da vida útil associados às funções em forma de banheira e de risco multimodal em modelos de mistura. Santos & Achcar (2011) introduzem uma análise Bayesiana para dados multivariados de sobrevivência na presença de um vetor de covariáveis e observações censuradas.

Como o tempo até o evento de interesse é uma variável aleatória, muitas vezes deseja-se estimar a função de sobrevivência utilizando a Tabela de Vida e o estimador de Kaplan-Meier (Souza, 2015). Além disso, uma das principais características dos dados de sobrevivência é a presença de censura, que é a observação parcial da resposta acarretada por uma interrupção no acompanhamento do paciente, seja porque o paciente abandonou o estudo, ou o estudo terminou para a análise dos dados ou porque o indivíduo morreu de causa diferente da estudada.

Com o surgimento de problemas cada vez mais complexos, novas metodologias e variações à tradicional análise de sobrevivência foram sendo propostas, como, por exemplo, modelos de riscos proporcionais ou testes de sobrevivência acelerados. Porém, essas metodologias atendem a casos em que o objetivo é analisar os dados de sobrevivência quando cada unidade experimental sofre apenas uma vez o evento de interesse. Em geral, baseiam-se no pressuposto de que os tempos de falha observados são independentes, no entanto, essa suposição de independência não é satisfeita para dados de sobrevivência multivariados, isto é, situações em que cada indivíduo pode experimentar recorrências do mesmo tipo de eventos, por exemplo, ataques recorrentes de asma. Os pressupostos de independência também são violados quando os dados consistem em eventos paralelos, como o aparecimento de retinopatia no olho esquerdo e direito, bem como para tempos de falha agrupados, como o tempo de sobrevida de gêmeos ou pacientes no mesmo leito de hospital. Nesses casos, mais de um tempo de sobrevivência é observado para cada indivíduo em estudo e, desse modo, supõe-se que exista associação entre os tempos de um mesmo indivíduo (Colosimo & Giolo, 2006).

Na literatura, esta provável dependência entre os tempos de sobrevivência é frequentemente modelada 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.

Uma alternativa para modelar a dependência entre dados multivariados são os modelos baseados em funções cópulas. Estes modelos vêm sendo cada vez mais utilizados atualmente, como por exemplo, nas áreas biológicas, ciências atuariais e finanças, pois esta teoria permite a criação de distribuições multivariadas sem a necessidade de se supor qualquer tipo de restrição às distribuições marginais e muito menos às multivariadas.

As funções cópulas são descritas em Nelsen (2006), Mikosch (2006), Viola (2009) e Joe (2014), dentre outras fontes, e conceituam-se como funções que ligam (conectam) a função de distribuição conjunta com suas funções de distribuição marginais univariadas. De acordo com Fischer (1997), cópulas são de interesse para estatísticos por duas razões: primeiro, é uma forma de estudar medidas de dependência e, segundo, a partir delas se constroem famílias de distribuições bivariadas.

Funções cópulas são usadas, 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 às 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, 2005). Outros exemplos são 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, 2007); para a determinação de uma estrutura de dependência entre as taxas de resgate dos seguros de capital diferido e as taxas de juro (Leal, 2010). Há inúmeras outras situações nas quais o uso de funções cópulas é adequado e eficiente, tais como nas ciências atuariais, em que cópulas são utilizadas na modelagem de mortalidade e perdas (Frees & Wang, 2005), em finanças (Cherubini et al., 2004; Cherubini et al., 2011; Irene & Klaus, 2014), na classificação de crédito e modelagem de risco (Embrechts et al., 2003), assim como na modelagem do contágio financeiro (Santos, 2010), em estudos biomédicos, na modelagem de eventos correlacionados e riscos competitivos (Achcar & Boleta, 2012; Louzada et al., 2013) e, até mesmo, na política (Quiroz Flores, 2008).

Como mostrado por Oakes (1989) e exemplificado por Romeo et al. (2006), um modelo de fragilidade induz um modelo de cópula arquimediana, ou seja, a cópula arquimediana com a transformada de Laplace da densidade da fragilidade como seu gerador. Por exemplo, a distribuição da fragilidade gama leva à cópula de Clayton, a distribuição de fragilidade positiva estável à cópula de Gumbel. A transição dos modelos de fragilidade para o modelo de cópula tem uma vantagem adicional, além de fornecer uma abordagem mais geral. Embora as covariáveis possam ser incluídas em um modelo de fragilidade, a interpretação dos coeficientes de regressão está condicionada a fragilidades não observadas e elas precisam ser integradas se houver interesse nos efeitos das covariáveis sobre os tempos de falha marginal. Em um modelo baseado em cópulas, por outro lado, as covariáveis atuam diretamente nas distribuições marginais (ver também Goethals et al. (2008) para uma discussão mais aprofundada sobre semelhanças e diferenças de modelos de fragilidade e modelos de cópulas Arquimedianas).

Nesse contexto, o objetivo deste trabalho é modelar dados de sobrevivência bivariados, que são os dados que caracterizam situações em que se observam dois tempos de vida para um mesmo equipamento ou indivíduo, baseados na cópula PVF que é derivada da distribuição *Power Variance Function* (PVF) de três parâmetros, proposta por Tweedie (1984). Como será mostrado adiante, a classe de cópulas PVF engloba uma família de cópulas Arquimedianas que inclui as cópulas de Clayton, Gumbel e Gaussiana Inversa como casos especiais.

De acordo com Romeo (2017), usar uma abordagem unificada considerando modelos da grande família PVF ao invés de se adequar as cópulas individuais de Clayton, Gumbel e Gaussiana Inversa e escolhendo uma com o melhor ajuste, tem várias vantagens. Entre elas, a análise não depende mais de suposições restritivas e possivelmente não verificadas de modelos individuais, mas encontra a melhor combinação de parâmetros na família maior que engloba os modelos mais simples e especiais. Se o modelo de melhor ajuste dentro da família geral se revelar um (ou próximo a um) dos modelos especiais, isso aumentará a confiança no uso desse modelo especial. Além disso, elimina a necessidade de comparação dos modelos cópulas candidatos.

O restante do texto está organizado da seguinte maneira: no Capítulo 2 iniciamos com uma breve revisão dos conceitos básicos de Análise de Sobrevivência e funções cópulas, bem como a apresentação do modelo proposto e a sua formalização.

No capitulo 3, descrevemos a abordagem bayesiana e o procedimento de estimação dos parâmetros de interesse considerando a distribuição Wiebull.

No capitulo 4 expandimos a utilização da cópula PVF para uma distribuição marginal mais flexível: a distribuição Exponencial Generalizada.

Por fim, no Capítulo 5 apresentamos as considerações finais juntamente com nossas perspectivas futuras para pesquisa.

Modelamos a dependência de dados de sobrevivência bivariados na presença de covariáveis e observações censuradas por meio da cópula PVF. Para fins inferenciais, foram utilizados métodos Monte Carlo em Cadeias de Markov (MCMC). Com o objetivo de detectar observações influentes nos dados foi utilizado o método Bayesiano de análise de influência de deleção de caso baseado na divergência ψ (Cook Weisberg, 1982). E ainda aplicamos os modelos estudados em um conjunto de dados reais de retinopatia diabética (The Diabetic Retinopathy Study Research Group, 1976) e em um conjunto de dados reais de apendicectomia em gêmeos adultos (Australian NH MRC Twin Registry dado por Duffy et al. (1990)).

CAPÍTULO

2

ANÁLISE DE SOBREVIVÊNCIA E FUNÇÕES CÓPULAS

Neste capítulo apresentamos alguns conceitos básicos de Análise de Sobrevivência que são essenciais para a compreensão do trabalho, assim como algumas propriedades e teoremas envolvendo funções cópulas.

2.1 Análise de Sobrevivência

Esta seção é subdividida em conceitos em Análise de Sobrevivência e Análise de Sobrevivência Bivariada, na qual definimos como foi feito o trabalho com dados de sobrevivência bivariados.

2.1.1 Alguns Conceitos em Análise de Sobrevivência

Segundo Colosimo & Giolo (2006), os principais componentes constituintes de um conjunto de dados de sobrevivência são os tempos de falha e os tempos de censura.

2.1.1.1 Tempo de Falha

Tempo de falha é o tempo decorrido, a partir de um instante inicial, até a ocorrência de um evento de interesse (morte, falha, recorrência de uma doença, etc.).

O tempo de falha é definido pelo tempo inicial, a escala de medida e o evento de interesse (Colosimo & Giolo, 2006). O tempo inicial do estudo deve ser definido com precisão. Na área médica, por exemplo, pode-se considerar a data do início do tratamento ou do diagnóstico da doença como possíveis escolhas.

A escala de medida depende do problema em estudo. Para dados médicos, a escala de medida é o tempo real (em horas, dias, etc.). Nas engenharias, a escala pode ser dada, por

exemplo, pelo número de ciclos, de quilometragem de um carro, entre outros.

Quanto ao evento de interesse, também chamado de falha, precisa ser definido de forma clara e precisa. Em algumas situações o evento de interesse é simples de ser diagnosticado, tais como morte ou recidiva de uma doença, mas às vezes pode ser mais complexo de ser demarcado, como por exemplo, para os fabricantes de produtos alimentícios que desejam saber quando seu produto fica inapropriado para o consumo.

Outro tipo de evento de interesse bem comum em Análise de Sobrevivência são os eventos recorrentes, que acontecem mais de uma vez para um mesmo indivíduo ou equipamento, dentre os quais podemos citar gestações, internações, cáries, infartos do miocárdio, fraturas e danificação de máquinas que podem ser reparadas. Por fim, temos também os diferentes tipos de eventos decorrentes de um mesmo fator de risco em estudo, como efeitos adversos de medicamentos e doenças oportunistas da AIDS.

2.1.1.2 Censura

A censura ocorre quando o tempo de falha de um indivíduo (ou peça) não é observado, ou seja, o paciente deixa de ser observado (ou o experimento deve ser encerrado e ainda existem itens em funcionamento). Outro fator de censura é quando a falha acontece por outras causas que não é a esperada no estudo.

Segundo Colosimo & Giolo (2006), podemos classificar a censura em:

- **Tipo I:** o estudo é conduzido até um tempo limite L pré-fixado e os indivíduos que ainda não experimentaram o evento são censurados;
- **Tipo II:** é aquela em que o teste será terminado após ter ocorrido a falha em um número pré-estabelecido de elementos sob teste;
- **Aleatória:** semelhante à censura do Tipo I, porém com os indivíduos sendo incorporados de maneira aleatória.

Quanto aos mecanismos, a censura é subdividida em:

- **À direita:** não se observa o desfecho e sabe-se que o tempo entre o início do estudo e o evento é maior do que o tempo observado. Por exemplo, um paciente não respondeu ao tratamento até ao final do experimento;
- **À esquerda:** acontece quando não conhecemos o momento da ocorrência do evento, mas sabemos que ocorreu antes do tempo observado. Por exemplo, em um estudo para determinar a idade em que crianças aprendem a ler, o pesquisador pode encontrar crianças que já sabiam ler e não se lembravam com que idade isto tinha acontecido, caracterizando, desta forma, observações censuradas à esquerda;

- **Intervalar:** ocorrência do evento entre tempos conhecidos. Pode ocorrer, por exemplo, em estudos que pacientes são acompanhados em visitas periódicas e é de conhecimento que o evento de interesse ocorreu em um certo intervalo de tempo.

A Figura 1 apresenta a ilustração de alguns mecanismos de censura, em que • representa a falha e ∘ a censura. No caso (a) todos os pacientes experimentaram o evento antes do final do estudo, não havendo assim censura. No caso (b) temos uma censura do tipo I, com o tempo limite L pré-fixado em 15 anos e o paciente 1 não experimentou o evento de interesse até o final do estudo e sofreu censura à direita. Em (c) temos a censura do tipo II em que o estudo foi finalizado após a ocorrência de um número pré-estabelecido de duas falhas, e os pacientes 1 e 4 sofreram censura à direita por não apresentarem o evento de interesse. Em (d) o acompanhamento de alguns pacientes foi interrompido por alguma razão e alguns pacientes não experimentaram o evento até o final do estudo, sofrendo assim censura à direita.

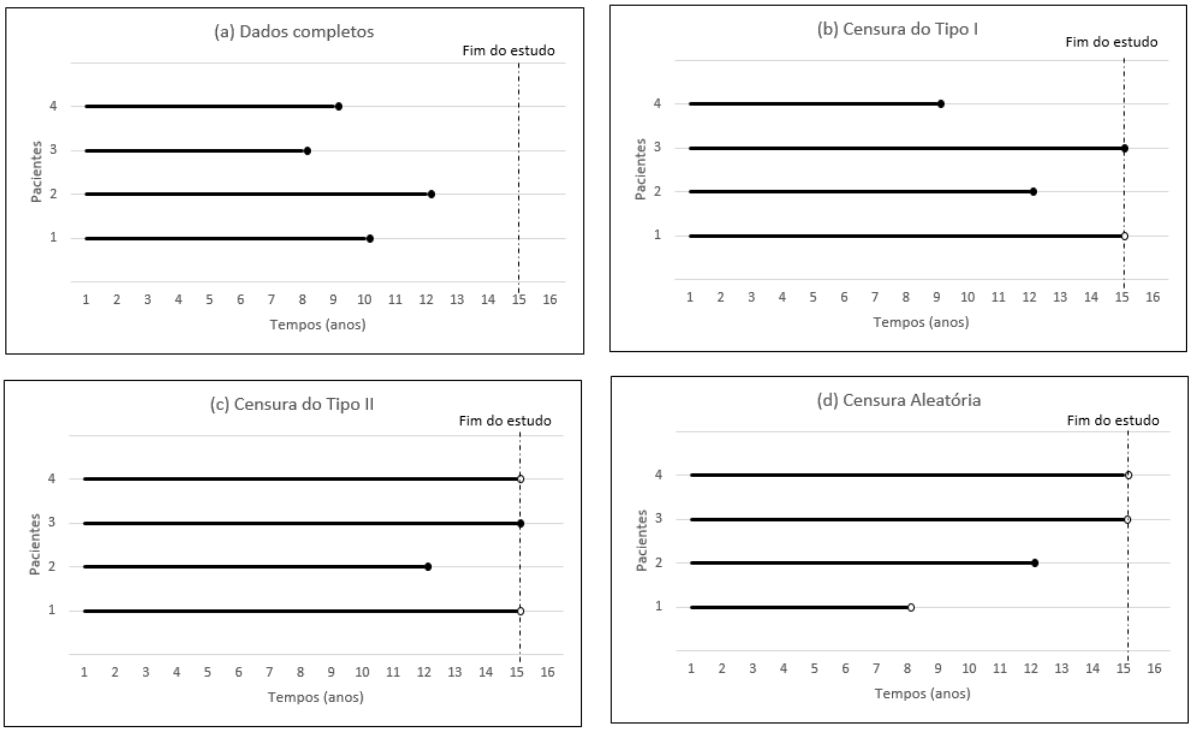

Figura 1 – Mecanismos de censura. (Fonte: Elaborado pelo autor)

2.1.1.3 Função de Sobrevivência e Função de Taxa de Falha

Seja T uma variável aleatória contínua não-negativa com função densidade de probabilidade $f(t \mid \theta)$, descrevendo os tempos de vida de uma população, em que θ é o parâmetro de interesse. Em Análise de Sobrevivência a variável aleatória T é geralmente especificada pela sua função de sobrevivência ou pela função taxa de falha. Em termos probabilísticos, a função de sobrevivência é dada por (Colosimo & Giolo, 2006):

$$S(t \mid \theta) = P(T > t \mid \theta). \tag{2.1}$$

Em consequência, a função de distribuição acumulada é definida como a probabilidade de uma observação não sobreviver até o tempo t, isto é, $F(t \mid \theta) = 1 - S(t \mid \theta)$.

Assim, a probabilidade de ocorrer uma falha no intervalo $[t_1, t_2)$ é dada por:

$$P(t_1 \leq T < t_2 \mid \theta) = F(t_2 \mid \theta) - F(t_1 \mid \theta) = (1 - S(t_2 \mid \theta)) - (1 - S(t_1 \mid \theta)) = S(t_1 \mid \theta) - S(t_2 \mid \theta). \tag{2.2}$$

A taxa de falha no intervalo $[t_1, t_2)$ é definida como sendo a probabilidade de que a falha ocorra neste intervalo, dado que não ocorreu antes de t_1, dividida pelo comprimento do intervalo (Colosimo & Giolo, 2006), isto é,

$$\frac{P(T \in [t_1, t_2) | T \geq t_1)}{(t_2 - t_1)} = \frac{P(T \in [t_1, t_2))}{(t_2 - t_1) P(T \geq t_1 \mid \theta)} = \frac{S(t_1 \mid \theta) - S(t_2 \mid \theta)}{\Delta(t) S(t_1 \mid \theta)}, \tag{2.3}$$

em que $\Delta(t) = t_2 - t_1$.

De forma geral, redefinindo o intervalo como $[t, t + \Delta t)$, a expressão (2.3) se torna:

$$\frac{S(t \mid \theta) - S(t + \Delta t \mid \theta)}{\Delta t S(t \mid \theta)}. \tag{2.4}$$

Assumindo Δt bem pequeno, $\lambda(t)$ representa a taxa de falha instantânea no tempo t condicional à sobrevivência até o tempo t.

Então, a função de taxa de falha de T é definida como:

$$\begin{aligned} \lambda(t \mid \theta) &= \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t, T \geq t)}{\Delta t P(T \geq t \mid \theta)} \\ &= \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t S(t \mid \theta)} = \frac{1}{S(t \mid \theta)} \lim_{\Delta t \to 0} \frac{F(t + \Delta t \mid \theta) - F(t \mid \theta)}{\Delta t} \\ &= \frac{F'(t \mid \theta)}{S(t \mid \theta)} = \frac{f(t \mid \theta)}{S(t \mid \theta)}. \end{aligned} \tag{2.5}$$

Além disso, como $F(t) = 1 - S(t \mid \theta)$, temos que:

$$\lambda(t \mid \theta) = \frac{F'(t \mid \theta)}{S(t \mid \theta)} = \frac{1}{S(t \mid \theta)} \frac{dF(t \mid \theta)}{dt} = \frac{1}{S(t \mid \theta)} \frac{d(1 - S(t \mid \theta))}{dt} = -\frac{1}{S(t \mid \theta)} \frac{dS(t \mid \theta)}{dt}.$$

Portanto,

$$\lambda(t \mid \theta) = -\frac{d \log(S(t \mid \theta))}{dt}. \tag{2.6}$$

As taxas de falha são números positivos sem limite superior. Este fato é útil para descrever a distribuição do tempo de vida de indivíduos (ou produtos, máquinas, entre outros) que pode ser: crescente, decrescente, constante ou em forma de banheira, como ilustrado na Figura 2.

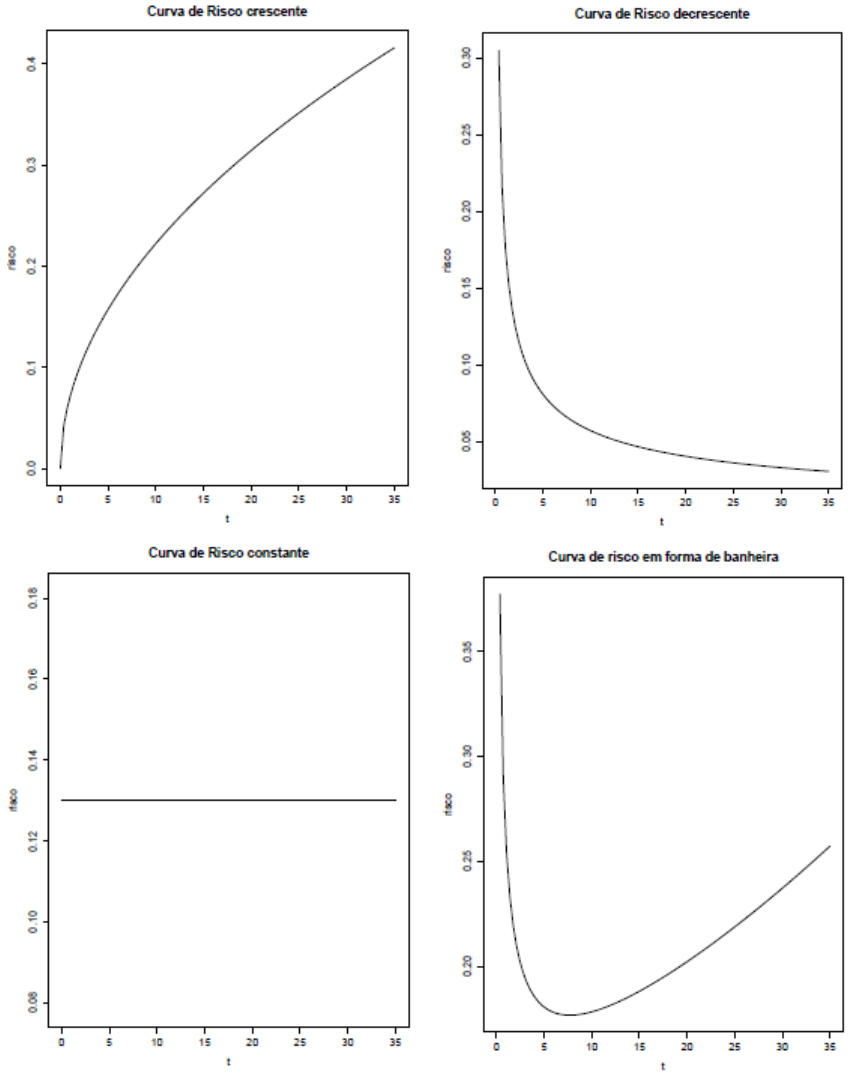

Figura 2 – Formas da curva de risco. Crescente (superior esquerdo), decrescente (superior direito), constante (inferior esquerdo), banheira (inferior direito). (Fonte: Colosimo & Giolo, 2006)

2.1.1.4 O Estimador de Kaplan-Meier

O estimador não-paramétrico de Kaplan-Meier (Kaplan & Meier, 1958) para estimar a função de sobrevivência, é também chamado de estimador limite-produto. É uma adaptação da função de sobrevivência empírica que, na ausência de censuras, é definida como (Colosimo & Giolo, 2006):

$$\widehat{S}(t) = \frac{\text{nº de observações que não falharam até o tempo } t}{\text{nº total de observações no estudo}}. \tag{2.7}$$

A função $\widehat{S}(t)$ é uma função do tipo escada com degraus nos tempos observados de falha de tamanho $1/n$, em que n é o tamanho da amostra. Se existirem empates em um certo tempo t, o tamanho do degrau fica multiplicado pelo número de empates.

Para obtermos a expressão geral do estimador de Kaplan-Meier, considere:

- $t_1 < t_2 < ... < t_k$, os k tempos distintos e ordenados de falha;
- d_j o número de falhas em $t_j,\ j = 1, ..., k$;
- n_j o número de indivíduos sob risco em t_j, ou seja, os indivíduos que sobreviveram e não foram censurados até o instante imediatamente anterior a t_j.

O estimador de Kaplan-Meier é, então, definido como:

$$\widehat{S}(t) = \prod_{j:t_j<t} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{j:t_j<t} \left(1 - \frac{d_j}{n_j} \right). \tag{2.8}$$

2.1.2 Alguns Conceitos em Análise de Sobrevivência para Dados Bivariados

Como já destacado ao longo do trabalho, há muitos casos em que se observa dois tempos de vida para um mesmo paciente ou equipamento. Por exemplo, o tempo entre a primeira e a segunda internação de um paciente decorrente de infecções oportunistas em pessoas com HIV, que são importantes causas de mortalidade em pessoas com o vírus. Pessoas infectadas com HIV estão sujeitas a inúmeras infecções e complicações neoplásicas relacionadas à sua doença.

Através da Figura 3, podemos ver que o comportamento da função de sobrevivência estimada pelo Kaplan-Meier tem comportamentos bem diferentes para os tempos de primeira (T_1) e segunda (T_2) internação.

Como nossa proposta de pesquisa é trabalhar com dados de sobrevivência bivariados, apresentamos alguns conceitos essenciais para a realização do estudo.

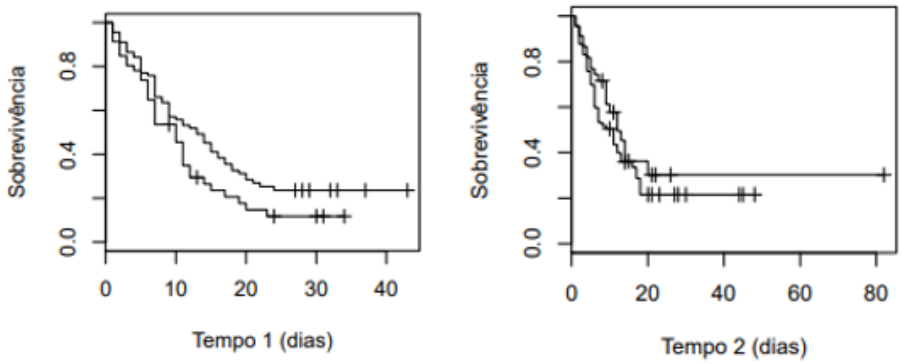

Figura 3 – Estimativas de Kaplan-Meier da função de sobrevivência para os dados de HIV. (Fonte: Pereira et al. (2012))

2.1.2.1 *Gráfico do Tempo Total sob Teste*

Com o objetivo de verificarmos o comportamento da função de risco dos tempos observados, utilizamos um método gráfico baseado no Tempo Total sob Teste (TTT *plot*), em que podemos encontrar mais detalhes em Aarset (1985). A versão empírica do gráfico do Teste do Tempo Total é dada por:

$$G(r/n) = \frac{\sum_{i=1}^{r} Y_{i:n} - (n-r)Y_{r:n}}{\sum_{i=1}^{r} Y_{i:n}}, \tag{2.9}$$

em que $r = 1, \ldots, n$ e $Y_{i:n}$ representam as estatísticas de ordem da amostra, com n sendo o tamanho da amostra.

Temos que a função de risco cresce (decresce) se o gráfico do Teste do Tempo Total é côncavo (convexo). Se o gráfico aproxima de uma linha diagonal temos função de risco constante e, se a curvatura é côncava e depois convexa a função de risco tem forma unimodal. Se o gráfico apresentar curvatura convexa e depois côncava a função de risco é em forma de banheira. O gráfico do Teste do Tempo Total é apenas uma condição suficiente e não necessária para indicar a forma da função de risco. A Figura 4 representa exemplos dos gráficos TTT aplicados a um conjunto de dados simulados, indicando função de risco crescente e decrescente.

Para o exemplo de tempos entre as internações para pacientes com HIV, temos, pela Figura 5, que a função de risco é crescente.

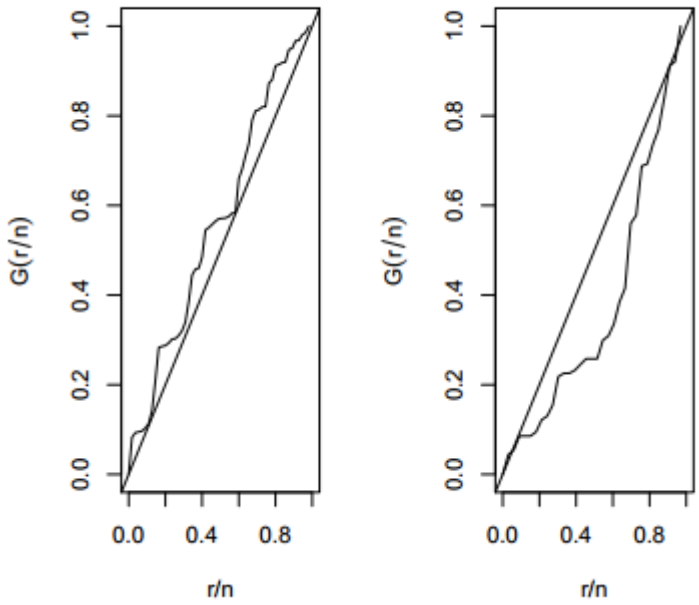

Figura 4 – Gráfico do Tempo Total sob Teste para dados com função de risco crescente (esquerda) e função de risco decrescente (direita). (Fonte: Dados simulados pelo autor).

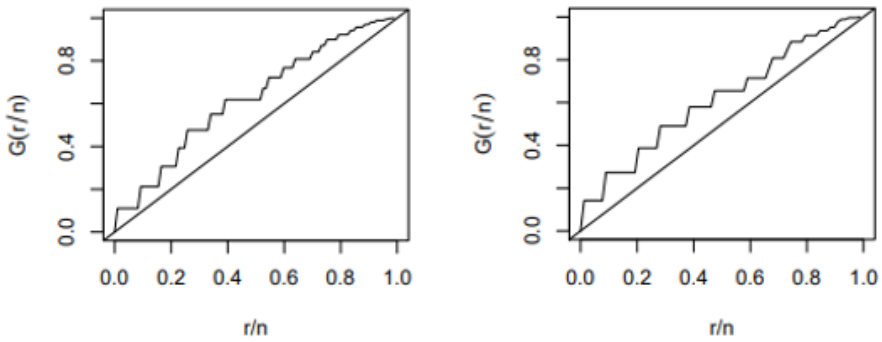

Figura 5 – Gráfico TTT para os dados de HIV, tempo da primeira internação (esquerda) e tempo da segunda internação (direita). (Fonte: Pereira et al. (2012))

2.1.2.2 Função de Verossimilhança

Vamos supor que T_1 e T_2 são duas variáveis aleatórias relativas aos tempos até a ocorrência de falha, e t_{1i} e t_{2i} observações amostrais de T_1 e T_2, respectivamente, para o i-ésimo indivíduo, $i = 1,\ldots,n$. Ao classificar os n pares de observações (t_{1i},t_{2i}) em classes, tem-se:

- C_1: t_{1i} e t_{2i} são tempos de sobrevida observados;
- C_2: t_{1i} é o tempo de sobrevida observado e t_{2i} é o tempo de censura;
- C_3: t_{1i} é o tempo de censura e t_{2i} é o tempo de sobrevida;

- C_4: t_{1i} e t_{2i} são os tempos de censura.

Considere θ_1 e θ_2 vetores de parâmetros desconhecidos para as distribuições marginais de T_1 e T_2, respectivamente, e ϕ o vetor de parâmetros de dependência. Então, a função de verossimilhança para θ_1, θ_2 e ϕ é dada por (Lawless, 2003):

$$L(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \phi | t_1, t_2) = \prod_{i \in C_1} f(t_{1i}, t_{2i} \mid \boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \phi) \prod_{i \in C_2} S'_{(t_1)} \prod_{i \in C_3} S'_{(t_2)} \prod_{i \in C_4} S(t_{1i}, t_{2i} \mid \boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \phi), \quad (2.10)$$

em que $f(t_{1i}, t_{2i} \mid \theta_1, \theta_2, \phi)$ é a função densidade de probabilidade conjunta para T_{1i} e T_{2i}, $S(t_{1i}, t_{2i} \mid \theta_1, \theta_2, \phi)$ é a função de sobrevivência conjunta e $S'_{(t_1)} = -\frac{\partial S(t_{1i}, t_{2i} | \theta_1, \theta_2, \phi)}{\partial t_{1i}}$ e $S'_{(t_2)} = -\frac{\partial S(t_{1i}, t_{2i} | \theta_1, \theta_2, \phi)}{\partial t_{2i}}$.

Por outro lado, considere:

$$\delta_{ji} = \begin{cases} 0, \text{ se } t_{ji} \text{ é uma observação censurada} \\ 1, \text{ se } t_{ji} \text{ é o tempo de sobrevida observado} \end{cases}$$

para $j = 1, 2$ e $i = 1, ..., n$, em que n é o número de observações, como uma função indicadora de censura e seja $\mathbf{D} = \{x_1, ..., x_n\}$, em que $\mathbf{x}_i = (t_{1i}, t_{2i}, \delta_{1i}, \delta_{2i})$, os dados observados. Assim, a função de verossimilhança em (2.10) pode ser reescrita como:

$$\begin{aligned} L(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \phi | \mathbf{D}) &= \prod_{i=1}^{n} \Big[f(t_{1i}, t_{2i} \mid \boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \phi)^{\delta_{1i}\delta_{2i}} S'_{(t_1)}{}^{\delta_{1i}(1-\delta_{2i})} \cdot \\ &\times S'_{(t_2)})^{\delta_{2i}(1-\delta_{1i})} S(t_{1i}, t_{2i} \mid \boldsymbol{\theta_1}, \theta_2, \phi)^{(1-\delta_{1i})(1-\delta_{2i})} \Big]. \end{aligned} \quad (2.11)$$

Observe que, se não houver dados censurados, a função verossimilhança (2.11) se reduz a:

$$L(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \phi | \mathbf{D}) = \prod_{i=1}^{n} f(t_{1i}, t_{2i} \mid \boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \phi), \quad (2.12)$$

que é produto das funções densidade de probabilidade conjunta.

2.1.2.3 Distribuições Marginais

A seguir, apresentamos algumas características que envolvem as distribuições marginais mais comumente usadas em Análise de Sobrevivência: Exponencial, Weibull e, a alternativa, Exponencial Generalizada.

- **Distribuição Exponencial**

A distribuição Exponencial se caracteriza por uma função taxa de falha constante, sendo a única distribuição absolutamente contínua com essa propriedade. É considerada uma das mais simples em termos matemáticos e tem sido extensivamente utilizada para modelar o tempo de vida de certos produtos e materiais, tais como óleos isolantes, dielétricos, entre outros (ver Walpole et al, 2016).

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

$$f(t \mid \alpha) = \alpha e^{-\alpha t}, \tag{2.13}$$

em que $t \geq 0$ e $\alpha > 0$.

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

$$S(t \mid \alpha) = P(T > t) = \exp(-\alpha t) \tag{2.14}$$

e

$$\lambda(t \mid \alpha) = \alpha. \tag{2.15}$$

Como a função de risco dada em (2.15) é constante, temos que tanto uma unidade que está em operação há 20 horas quanto uma unidade que está em operação há 40 horas tem a mesma chance de falharem em um intervalo futuro de mesmo comprimento. Esta propriedade é chamada de falta de memória da distribuição.

Na Figura 6 apresentamos os gráficos da função densidade, da função de sobrevivência e da função de risco da distribuição Exponencial. O objetivo destes gráficos é verificar o comportamento desta distribuição para diferentes valores de seu parâmetro α.

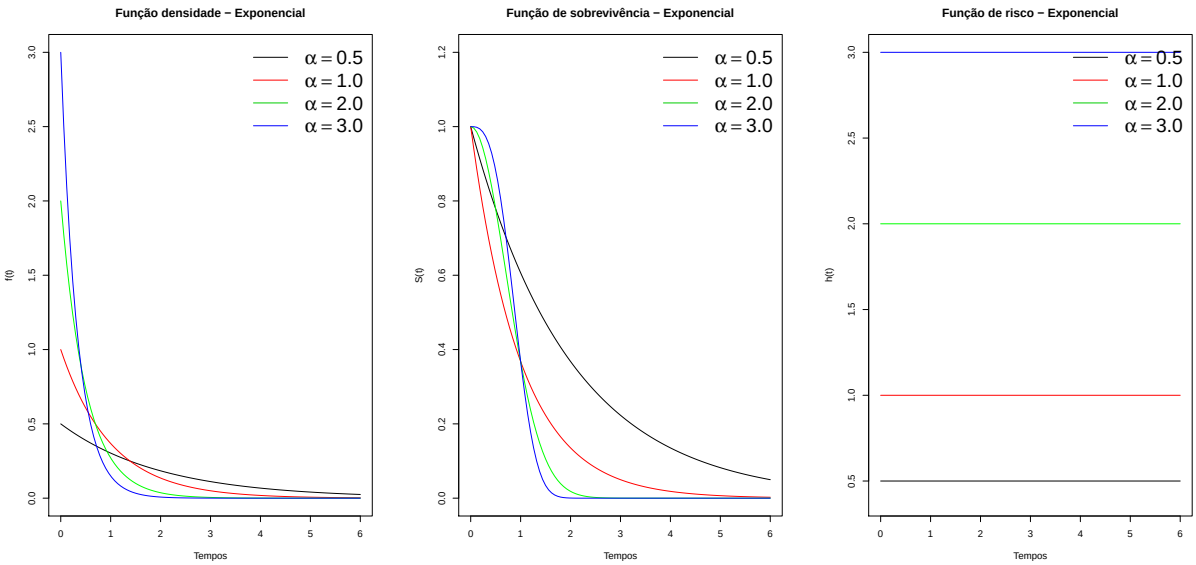

Figura 6 – Gráfico da função densidade de probabilidade (esquerda), da função de sobrevivência (centro) e da função de risco (direita) para diferentes valores do parâmetro α do modelo Exponencial. (Fonte: Elaborado pelo autor).

- **Distribuição Weibull**

  A distribuição Weibull (Weibull, 1939), nomeada devido a *Waloddi Weibull* que, em 1951, lançou um artigo descrevendo a distribuição em detalhes e propondo diversas aplicações, é frequentemente usada em estudos biomédicos e industriais e é amplamente conhecida em virtude de sua simplicidade e flexibilidade em acomodar diferentes formas de função de risco.

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

$$f(t \mid \alpha, \lambda) = \frac{\alpha}{\lambda^{\alpha}} t^{\alpha-1} \exp\left(-\left(\frac{t}{\lambda}\right)^{\alpha}\right), \tag{2.16}$$

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

  A função de sobrevivência do modelo Weibull é dada por:

$$S(t \mid \alpha, \lambda) = \exp\left(-\left(\frac{t}{\lambda}\right)^{\alpha}\right) \tag{2.17}$$

  e a função de risco por:

$$\lambda(t \mid \alpha, \lambda) = \frac{\alpha}{\lambda^{\alpha}} t^{\alpha-1}. \tag{2.18}$$

  Esta distribuição possui riscos crescentes para $\alpha > 1$, decrescentes para $\alpha < 1$ e constantes para $\alpha = 1$, em que o modelo se reduz à distribuição Exponencial com parâmetro $\lambda^* = \frac{1}{\lambda}$, neste caso.

Os gráficos da função densidade, da função de sobrevivência e da função de risco da distribuição Weibull são apresentados na Figura 7, em que verifica-se o comportamento desta distribuição para diferentes valores de seus parâmetros α e λ.

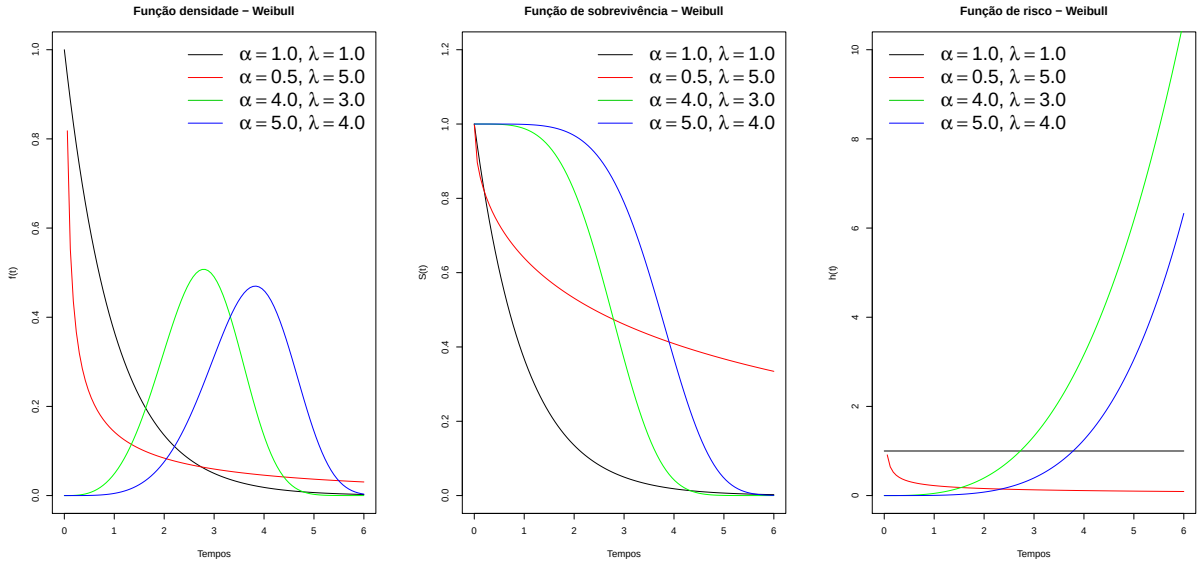

Figura 7 – Gráfico da função densidade de probabilidade (esquerda), da função de sobrevivência (centro) e da função de risco (direita) para diferentes valores dos parâmetros α e λ do modelo Weibull. (Fonte: Elaborado pelo autor).

- **Distribuição Exponencial Generalizada**

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

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

$$f(t \mid \alpha,\lambda) = \alpha\lambda\,(1-\exp(-\lambda t))^{\alpha-1}\exp(-\lambda t), \tag{2.19}$$

  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 a essa densidade são dadas, respectivamente, por:

$$S(t \mid \alpha,\lambda) = P(T > t) = 1-(1-\exp(-\lambda t))^{\alpha} \tag{2.20}$$

  e

$$\lambda = \frac{f(t \mid \alpha,\lambda)}{S(t \mid \alpha,\lambda)} = \frac{\alpha\lambda\,(1-\exp(-\lambda t))^{\alpha-1}\exp(-\lambda t)}{1-(1-\exp(-\lambda t))^{\alpha}}. \tag{2.21}$$

Na Figura 8 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 o comportamento desta distribuição para diferentes valores de seus parâmetros α e λ.

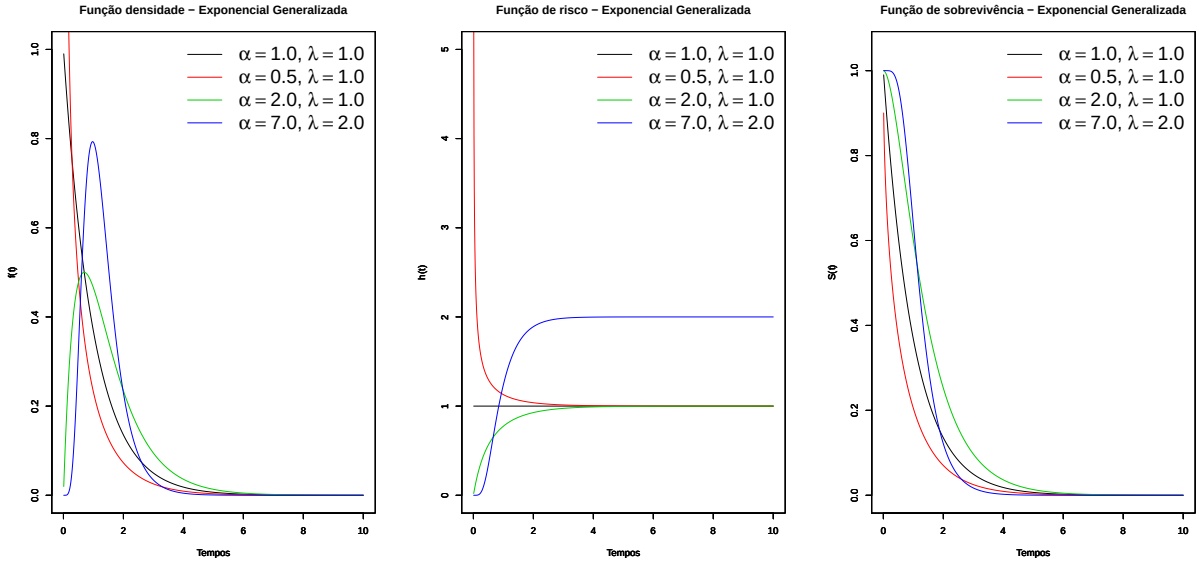

Figura 8 – Gráfico da função densidade de probabilidade (esquerda), da função de sobrevivência (centro) e da função de risco (direita) para diferentes valores dos parâmetros α e λ do modelo Exponencial Generalizada. (Fonte: Elaborado pelo autor).

2.1.2.4 Análise Bayesiana

Utilizando a metodologia Bayesiana para inferirmos sobre os parâmetros do modelo, e assumindo que não há conhecimentos prévios sobre eles, utilizamos distribuições *a priori* pouco informativas.

Consideramos que as marginais T_j têm distribuição Weibull com parâmetros α_j e $\lambda_{ij} = \exp(\beta_{0j} + \beta_{1j} x_i)$, em que x_i representa a covariável dicotômica do modelo, $i = 1, ..., n$ e $j = 1, 2$

Com o objetivo de garantir que a distribuição *a posteriori* conjunta seja própria, consideramos uma distribuição *a priori* conjunta própria para os parâmetros do modelo e assumimos distribuições *a priori* independentes. A densidade *a priori* conjunta de $\theta = (\phi, \alpha_1, \alpha_2, \beta_1, \beta_2)$, em que $\beta_1 = (\beta_{01}, \beta_{11})$ e $\beta_2 = (\beta_{02}, \beta_{12})$, e ainda, ϕ é o vetor de parâmetros de dependência da função cópula vista com detalhes na Seção 2.2, é dada por:

$$\pi(\theta) = \pi(\phi) \prod_{j=1}^{2} \pi(\alpha_j) \prod_{j=1}^{2} \pi(\beta_{\mathbf{j}}), \tag{2.22}$$

em que

$$\pi(\beta_{\mathbf{j}}) = \pi(\beta_{0j}, \beta_{1j}, ..., \beta_{qj}) = \prod_{k=0}^{q} \prod_{j=1}^{2} \pi(\beta_{kj}), \text{ e} \tag{2.23}$$

q representa o número de covariáveis, $j = 1, 2$ e $k = 0, ..., q$.

Combinando as distribuições *a priori* independentes (2.22) com a função de verossimilhança (2.11), obtemos a distribuição conjunta *a posteriori* do vetor de parâmetros θ, $\pi(\theta|\mathbf{D})$, em que $\mathbf{D}$ é o conjunto de dados observados. As estimativas dos parâmetros são dadas pelas médias da distribuição *a posteriori*.

Esta densidade *a posteriori* conjunta é analiticamente intratável e suas condicionais não são conhecidas. Assim, para inferência utilizamos métodos MCMC. Realizamos aplicações em conjuntos de dados simulados e reais. Todas as implementações computacionais foram realizadas utilizando os sistemas JAGS - *Just Another Gibbs Sampler* (Plummer, 2003) e R (R Core Team, 2019) por meio do pacote *rjags* (Denwood et al., 2016).

2.1.2.5 Critérios de Comparação de Modelos

Por diversas vezes pode não ser intuitivo avaliar a adequabilidade de um modelo a certo conjunto de dados e/ou propor um melhor modelo para seus dados, por isso uma etapa importante no processo de modelagem se deve ao estudo de metodologias para seleção e comparação de modelos, a fim de, dentro de uma coleção de modelos, escolher o que mais se adequa ao seu problema em estudo.

Na literatura encontramos diversas metodologias que se propõem a analisar essa adequabilidade. Neste trabalho, assim como foi feito em Louzada et al. (2013), utilizamos quatro critérios Bayesianos 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 (*Logarithm of the pseudomarginal likelihood*).

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

$$\bar{D} = \frac{1}{M} \sum_{m=1}^{M} D(\theta_m), \tag{2.24}$$

sendo m indicando a m-ésima realização de um total de M 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 e DIC podem ser calculados, respectivamente, por $\widehat{\text{EAIC}} = \bar{D} + 2q$, $\widehat{\text{EBIC}} = \bar{D} + q\ln(n)$ e $\widehat{\text{DIC}} = 2\bar{D} - \widehat{D}$, em que q é o número de parâmetros no

modelo e $\widehat{D} = D\left(\frac{1}{M}\sum_{q=1}^{M}\theta_q\right)$, que é um estimador para $D\{E(\theta)\}$.

Comparando modelos alternativos, o modelo preferido é aquele com menor valor desses critérios.

Um outro critério que será utilizado nesse trabalho é derivado das ordenadas da densidade preditiva condicional (CPO) (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 por meio de uma simples amostra MCMC a partir da distribuição *a posteriori* $\pi(\theta|\mathbf{D})$. Considere $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(M)}$ uma amostra de tamanho M de $\pi(\theta|\mathbf{D})$ após o *burn-in*. Uma aproximação Monte Carlo de CPO é dada por:

$$\widehat{\text{CPO}_i} = \left(\frac{1}{M}\sum_{q=1}^{M}\frac{1}{f(t_{1i}, t_{2i}|\theta^{(q)})}\right)^{-1}. \tag{2.25}$$

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

2.1.2.6 Diagnóstico de Observações Influentes

Com início na análise de resíduos (Cox & Snell, 1968), várias metodologias de avaliar a influência de uma observação no ajuste de um modelo foram sendo discutidas na literatura. Cox & Weisberg (1982) propuseram uma forma de avaliação da influência de uma observação no ajuste de um modelo por meio da exclusão de casos.

Sob pequenas perturbações no modelo e/ou nos dados, Cook (1986) propôs a avaliação da influência conjunta das observações, ao invés da retirada individual ou conjunta de pontos. Caso essas perturbações causem efeitos desproporcionais, pode-se ter evidências de que há um mau ajuste do modelo, ou problemas com as suposições adotadas.

Técnicas de influência local têm sido amplamente utilizadas, por exemplo em Cancho et al. (2010), Vidal & Castro (2010), Suzuki et al. (2012), Louzada et al. (2013) e Suzuki et al. (2016).

Neste trabalho, vamos considerar a análise de influência de deleção de casos baseada 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 θ para os dados completos e, $P_{(-i)}$ a distribuição *a posteriori* sem o i-ésimo caso. Especificamente,

$$D_\psi(P; P_{(-i)}) = \int \psi\left(\frac{\pi(\theta|D^{(-i)})}{\pi(\theta|D)}\right)\pi(\theta|D)\,d\theta, \tag{2.26}$$

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.

Temos que $D_\psi(P;P_{(-i)})$ pode ser calculado considerando uma amostra da distribuição *a posteriori* de θ via métodos MCMC. Considere $\theta^{(1)},\ldots,\theta^{(M)}$ uma amostra de tamanho M de $\pi(\theta|\mathbf{D})$. Então, uma estimativa Monte Carlo é dada por:

$$\widehat{D_\psi}(P;P_{(-i)}) = \frac{1}{M}\sum_{q=1}^{M}\psi\left(\frac{\pi(\theta^{(q)}|D^{(-i)})}{\pi(\theta^{(q)}|D)}\right). \tag{2.27}$$

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 θ.

Para um profissional na área médica, 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), em que uma moeda viesada com probabilidade de sucesso p é considerada. Então, a divergência ψ entre a moeda viesada e a não viesada é:

$$D_\psi(f_0;f_1) = \int \psi\left(\frac{f_0(x)}{f_1(x)}\right) f_1(x)dx, \tag{2.28}$$

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 pode ser facilmente verificado que d_ψ satisfaz a seguinte equação:

$$d_\psi(p) = \frac{\psi(2p)+\psi(2(1-p))}{2}. \tag{2.29}$$

Observa-se 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, uma observação na qual $d_J > 0,416$ pode ser considerada influente.

2.2 Funções Cópulas

Nesta seção apresentamos uma breve introdução sobre as funções cópulas, bem como alguns de seus resultados básicos. O uso destas funções permite a construção da distribuição conjunta de uma variável multivariada com as distribuições marginais conhecidas.

Há duas maneiras básicas de dizer o que são cópulas. Primeiramente, cópulas são funções que ligam (conectam) a função distribuição conjunta com suas funções distribuição marginais

univariadas. Alternativamente, cópulas são funções distribuição multivariadas cujas marginais unidimensionais são Uniformes em $[0,1]$. As cópulas representam uma abordagem útil para modelar e entender o fenômeno de dependência, ressaltando sua estrutura.

As referências básicas para o estudo destas funções são os livros de Cherubini et al. (2004), Nelsen (2006), Kolev et al. (2006) e Jaworski et al. (2010).

A seguir apresentamos alguns conceitos em relação as cópulas e suas propriedades.

Definição 2.2.1. *Uma cópula é uma distribuição multivariada cujas marginais são Uniforme (0,1). Considerando o vetor aleatório $U = (U_1, \ldots, U_n) \in I^n$ com cópula n-dimensional C, temos:*

$$C(u_1, \ldots, u_n \mid \phi) = P(U_1 \leq u_1, \ldots, U_n \leq u_n \mid \phi), \tag{2.30}$$

em que ϕ é o parâmetro, ou vetor de parâmetros, associado à função cópula.

O Teorema de Sklar (Sklar, 1959) enunciado a seguir, é um dos resultados mais importantes referente a teoria e aplicações de cópulas, em que a 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 2.2.1. *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:*

$$H(t_1, \ldots, t_n \mid \phi) = C(F_1(t_1), \ldots, F_n(t_n) \mid \phi). \tag{2.31}$$

Se $F_1(t_1), \ldots, F_n(t_n)$ são todas absolutamente 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.

Atualmente, a classe de cópulas Arquimedianas é a mais utilizada na prática, pois sua representação 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. Alguns exemplos dessas cópulas são a de Clayton (Clayton, 1978), a de Frank (Frank, 1979), a de Gumbel-Hougard (Hougaard, 1986) e a de Ali-Mikhail-Haq (AMH) (Ali, Mikhail and Haq, 1978).

Nesse contexto, uma distribuição bivariada pertence à família de cópulas Arquimedianas se tem a seguinte representação:

$$C_\phi(u,v) = \varphi(\varphi(u)^{-1} + \varphi(v)^{-1}), \;\; 0 \leq u, v \leq 1, \tag{2.32}$$

em que $\varphi(t)$ assume valores entre 0 e 1, $\lim_{t\to 0}\varphi(t)=1$, $\varphi'(t)$ é estritamente decrescente, $\varphi''(t)$ é estritamente crescente e ϕ é o parâmetro de dependência da cópula.

Todas as cópulas Arquimedianas usualmente encontradas possuem expressões com forma fechada. O Teorema 2.2.2 é um importante resultado em relação às cópulas Arquimedianas, pois apresenta algumas propriedades que este tipo de cópula possui.

Teorema 2.2.2. *Seja C uma cópula Arquimediana com gerador φ. Então:*

1. *C é simétrica, isto é, $C(u,v)=C(v,u)$ para todo $u,v\in \boldsymbol{I}$;*
2. *C é associativa, isto é, $C(C(u,v),w)=C(u,C(v,w))$ para todo $u,v,w\in \boldsymbol{I}$;*
3. *Se $c>0$ é uma constante qualquer, então $c\varphi$ é também um gerador de C.*

2.2.1 Dependência

Aqui, serão exploradas formas com as quais as funções cópulas podem ser usadas no estudo de dependência ou associação entre variáveis aleatórias.

2.2.1.1 Concordância

Seja (x_i,y_i) e (x_j,y_j) duas observações de um vetor (X,Y) de variáveis aleatórias contínuas. Dizemos que (x_i,y_i) e (x_j,y_j) são *concordantes* se $x_i<x_j$ e $y_i<y_j$, ou se $x_i>x_j$ e $y_i>y_j$. Por outro lado, dizemos que (x_i,y_i) e (x_j,y_j) são *discordantes* se $x_i<x_j$ e $y_i>y_j$ ou se $x_i>x_j$ e $y_i<y_j$. De maneira análoga, (x_i,y_i) e (x_j,y_j) são concordantes se $(x_i-x_j)(y_i-y_j)>0$ e discordantes se $(x_i-x_j)(y_i-y_j)<0$, (Nelsen, 2006). As principais medidas de concordância são o tau de Kendall e o rho de Spearman, dadas a seguir.

- **Tau de Kendall:** De acordo com Kruskal (1958), Hollander & Wolfe (1973) e Lehmann (1975), o tau de Kendall é dado em termos de concordância da seguinte forma: considere uma amostra aleatória $\{(x_1,y_1),(x_2,y_2),...,(x_n,y_n)\}$ de n observações de um vetor (X,Y) de variáveis aleatórias contínuas. Há $\binom{n}{2}$ pares distintos (x_i,y_i) e (x_j,y_j) de observações na amostra, sendo cada par concordante ou discordante. Considere c o número de pares concordantes e d o número de pares discordantes. Então, o tau de Kendall para a amostra é definido como:

$$t=\frac{c-d}{c+d}. \tag{2.33}$$

  Equivalentemente, t é a probabilidade da concordância menos a probabilidade da discordância para um par de observações (x_i,y_i) e (x_j,y_j) que é escolhido aleatoriamente da amostra. A versão populacional do tau de Kendall para o vetor (X,Y) de variáveis

aleatórias contínuas com função de distribuição conjunta H é definida de maneira análoga. Sejam (X_1, Y_1) e (X_2, Y_2) vetores aleatórios independentes e identicamente distribuídos, cada um com função de distribuição conjunta H. Então a versão populacional do tau de Kendall é definida como a probabilidade da concordância menos a probabilidade da discordância:

$$\tau = \tau_{X,Y} = P[(X_1 - X_2)(Y_1 - Y_2) > 0] - P[(X_1 - X_2)(Y_1 - Y_2) < 0]. \tag{2.34}$$

Apresentamos, no Teorema 2.2.3, a expressão da versão populacional do tau de Kendall para X e Y.

Teorema 2.2.3. *Sejam X e Y variáveis aleatórias contínuas cuja cópula é C. Então a versão populacional do tau de Kendall para X e Y (denotado por $\tau_{X,Y}$ ou τ_C) é dada por:*

$$\tau_{X,Y} = \tau_C = 4 \int \int_{I^2} C(u,v) dC(u,v) - 1. \tag{2.35}$$

- **Rho de Spearman:** Para obtermos a versão populacional da medida do rho de Spearman (Kruskal, 1958; Lehmann, 1966), considere $(X_1, Y_1), (X_2, Y_2)$ e (X_3, Y_3) três vetores aleatórios independentes com uma função de distribuição conjunta H comum (cujas marginais são F e G) e cópula C. A versão populacional $\rho_{X,Y}$ do rho de Spearman é definida como sendo proporcional a probabilidade da concordância menos a probabilidade da discordância para os dois vetores (X_1, Y_1) e (X_2, Y_3), isto é, um par de vetores com as mesmas marginais, mas um vetor tem função de distribuição H, enquanto os componentes do outro são independentes:

$$\rho = \rho_{X,Y} = 3(P[(X_1 - X_2)(Y_1 - Y_3) > 0] - P[(X_1 - X_2)(Y_1 - Y_3) < 0]). \tag{2.36}$$

Note que, enquanto a função de distribuição conjunta de (X_1, Y_1) é $H(x,y)$, a função de distribuição conjunta de (X_2, Y_3) é $F(x)G(y)$, já que X_2 e Y_3 são independentes.

No Teorema 2.2.4 temos a expressão da versão populacional do rho de Spearman para X e Y.

Teorema 2.2.4. *Sejam X e Y variáveis aleatórias contínuas cuja cópula é C. Então, a versão populacional do rho de Spearman para X e Y (denotado por $\rho_{X,Y}$ ou ρ_C) é dada por:*

$$\rho_{X,Y} = \rho_C = 12 \int \int_{I^2} C(u,v) du dv - 3. \tag{2.37}$$

2.2.2 Alguns Exemplos de Cópulas Arquimedianas

Nesta seção apresentamos algumas das principais cópulas Arquimedianas existentes.

- **Cópula de Clayton:** A cópula Arquimediana bivariada de Clayton (Clayton, 1978), tem a seguinte forma:

$$C_\phi(u,v) = (u^{-\phi} + v^{-\phi} - 1)^{-\frac{1}{\phi}}, \phi \in \mathbb{R}^+. \tag{2.38}$$

Para ϕ tendendo a 0 temos $C_\phi(u,v) = uv$, que denota independência. Além disso, a sua função geradora é dada por:

$$\varphi(t) = \frac{1}{\phi}(t^{-\phi} - 1), \tag{2.39}$$

e sua medida de concordância tau de Kendal é $\tau_\phi = \frac{\phi}{\phi+2}$.

Na Figura 9, observamos o comportamento da cópula de Clayton a partir dos gráficos de superfície da densidade, do contorno e de dispersão para um valor de parâmetro de dependência $\phi = 0,8$.

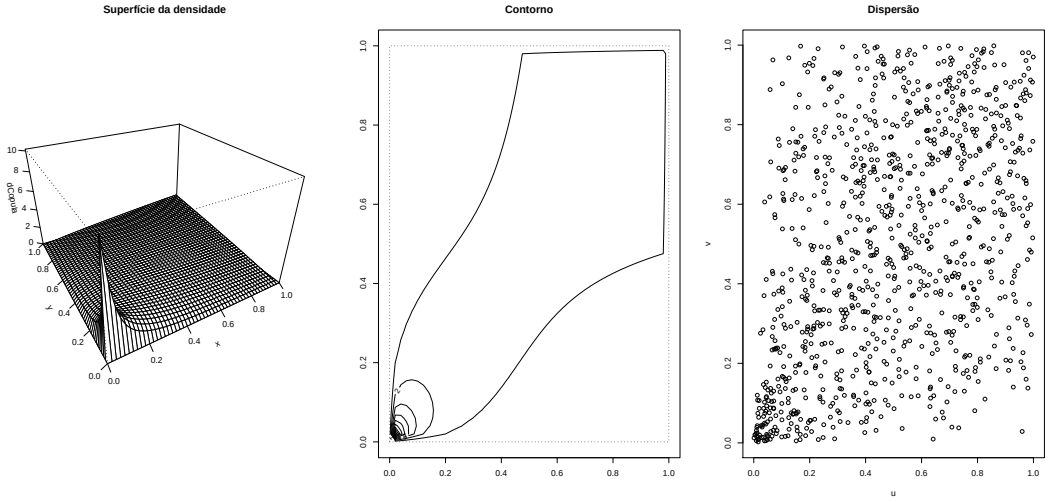

Figura 9 – Gráfico de superfície da densidade, contorno e dispersão para a cópula de Clayton. (Fonte: Elaborado pelo autor).

- **Cópula de Gumbel:** A cópula bivariada de Gumbel (Hougaard, 1986), tem a forma:

$$C_\alpha(u,v) = exp\{-\left[(-logu)^{1/\alpha} + (-logv)^{1/\alpha}\right]^\alpha\}, \alpha \geq 1. \tag{2.40}$$

Para α, parâmetro de dependência da cópula, tendendo a 1 temos $C_\alpha(u,v) = uv$, que denota independência. Além disso, a sua função geradora é dada por:

$$\varphi(t) = (-log(t))^\alpha, \tag{2.41}$$

e sua medida de concordância tau de Kendal é $\tau_\alpha = 1 - \alpha$.

Na Figura 10, observamos o comportamento da cópula de Gumbel a partir dos gráficos de superfície da densidade, do contorno e de dispersão para um valor de parâmetro de dependência $\alpha = 1,5$.

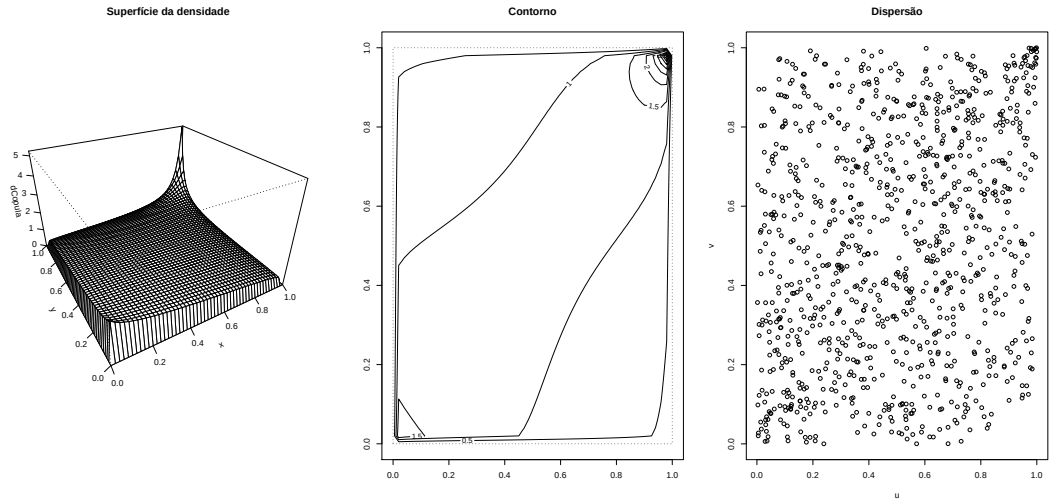

Figura 10 – Gráfico de superfície da densidade, contorno e dispersão para a cópula de Gumbel. (Fonte: Elaborado pelo autor).

- **Cópula Gaussiana Inversa:** A cópula bivariada Gaussiana Inversa (Schlösser A., 2011), tem a forma:

$$C_\eta(u,v) = exp\{-2\left[\eta^{1/2}\left(b(u)^2 + b(v)^2 - \eta\right)^{1/2} - \eta\right]\} \tag{2.42}$$

em que $b(s) = \eta^{1/2} - (1/2)\eta^{-1/2}log(s)$, sendo η o parâmetro de dependência.

O valor $\eta \to \infty$ representa independência, ou seja, $C_\eta(u,v) = uv$.

Além disso, a sua função geradora é dada por:

$$\varphi(t) = \frac{1}{4}\eta^{-1}(log(t))^2 - log(t). \tag{2.43}$$

O tau de Kendall para a cópula Gaussiana Inversa não tem forma fechada e pode ser obtido numericamente mas, sabe-se que $\tau_\eta \in (0;0,5)$.

2.3 O Modelo de Sobrevivência PVF Bivariado

A distribuição *Power Variance Function* (PVF) univariada de três parâmetros foi sugerida pela primeira vez por Tweedie (1984) e derivada, independentemente, como uma extensão da distribuição estável positiva por Hougaard (1986). Mallick & Ravishanker (2006) descreveram uma abordagem Bayesiana ao modelo de fragilidade PVF assumindo distribuições Exponenciais para o risco base. Abordagens clássicas foram exploradas em Hougaard (2000), Duchateau & Janssen (2008) e Wienke (2010). Uma mistura univariada da distribuição de fragilidade Weibull e PVF foi aplicada em Wasinrat et al. (2013) e uma mistura Weibull-PVF bivariada foi considerada em Hanagal (2009).

Uma variável aleatória W possui distribuição PVF de três parâmetros se sua função densidade de probabilidade for dada por:

$$f(w \mid \alpha,\delta,\theta) = -\frac{1}{\pi w} exp(-\theta w + \delta\theta^{\alpha}/\alpha) \sum_{k=1}^{\infty} \frac{\Gamma(k\alpha+1)}{\Gamma(k+1)} (\frac{-w^{-\alpha}\delta}{\alpha})^{k} sin(k\pi), \;\; 0<\alpha<1, \delta>0, \theta\geq 0. \tag{2.44}$$

A transformada de Laplace $L_W(s) = E[exp(-sW)]$ de W é dada por:

$$L_W(s) = exp\Big\{-\frac{\delta}{\alpha}[(\theta+s)^{\alpha}-\theta^{\alpha}]\Big\}.$$

No caso p-variado, sejam $T_1,\ldots,T_p$ variáveis aleatórias contínuas não negativas com funções de sobrevivência $S_1,\ldots,S_p$. Se $T_1,\ldots,T_p$ são condicionalmente independentes dado uma variável aleatória W, não negativa, em um modelo de fragilidade, isto é, $S(t_1,\ldots,t_p \mid w) = \prod_{j=1}^{p} S_j(t_j \mid w) = \prod_{j=1}^{p} \bar{G}_j^{w}$, em que $\bar{G}_j$ são as funções de sobrevivências basais, então a função de sobrevivência conjunta tem a seguinte representação de cópula Arquimediana (Oakes, 1989):

$$S(t_1,\ldots,t_p) = C_{\phi}(S_1(t_1),\ldots,S_p(t_p)) = \phi^{-1}\Big[\sum_{j=1}^{p}\phi(S_j(t_j))\Big],$$

com gerador da cópula C_{ϕ} dado por $\phi(.) = L_W^{-1}(.)$, a inversa da transformada de Laplace de W.

Isso é visto a partir da representação da cópula de sobrevivência:

$$C(u_1,\ldots,u_p) = S\Big(S_1^{-1}(u_1),\ldots,S_p^{-1}(u_p)\Big),$$

em que $S_j^{-1} = \bar{G}_j^{-1}\Big(exp(-L_W^{-1}(u_j))\Big), 0<u_j<1, j=1,\ldots,p.$

As distribuições de fragilidade comumente usadas são a distribuição estável positiva, Gaussiana Inversa e Gama. Ao considerar uma distribuição de fragilidade da família da distribuição PVF, englobamos essas distribuições de fragilidade comuns.

Como apontado por Romeo (2017), para tornar os parâmetros identificáveis, estabelecemos $\delta = \eta^{1-\alpha}$ e $\theta = \eta$ com $\alpha \in (0,1)$ e $\eta > 0$. Assim, a transformada de Laplace de

$W \sim PVF(\alpha, \eta)$ é dada por:

$$\phi^{-1}(s) = exp\Big\{ -\frac{1}{\alpha}\Big[\eta^{1-\alpha}(\eta+s)^{\alpha} - \eta\Big]\Big\}.$$

Se $\eta = 0$ obtemos a transformada de Laplace da distribuição estável positiva com parâmetro $\alpha \in (0,1)$, isto é, $L_W(s) = exp(-s^{\alpha})$. A distribuição Gaussiana Inversa é obtida fazendo $\alpha = 0.5$ com $L_W(s) = exp\{2\eta[1-(1+s\eta^{-1})1/2]\}$. E, no caso quando $\alpha \to 0$ a transformada de Laplace se reduz à transformada de Laplace de uma distribuição Gama com média 1 e variância $\eta > 0$, isto é, $L_W(s) = (1+s\eta)^{-1/n}$ (Duchateau & Janssen, 2008).

Observação 2.3.1. *A transformada de Laplace é $L_W(s) = E[exp(-sW)]$ e caracteriza a distribuição. Basta tomar $-s = t$ que temos a função geradora de momentos $M_W(t) = E[exp(tW)]$.*

Observação 2.3.2. *O gerador da cópula C_ϕ é dado por $\phi(.) = L_W^{-1}(.)$, a inversa da transformada de Laplace de W.*

Definindo $g(s) = \eta^{\alpha} - \alpha\eta^{\alpha-1}log(s)$ como a inversa da transformada de Laplace de W, o gerador da cópula PVF é dado por $\phi(s) = g(s)^{\frac{1}{\alpha}} - \eta$. A função cópula multivariada PVF pode ser escrita como:

$$C_{\alpha,\eta}(u_1,\ldots,u_p) = exp\Big\{ -\frac{1}{\alpha}\Big[\eta^{1-\alpha}\Big(\sum_{j=1}^{p} g(u_j)^{1/\alpha} - (p-1)\eta\Big)^{\alpha} - \eta\Big]\Big\},$$

em que $u_j = S_j(t_j)$.

Quando $p = 2$, a cópula bivariada PVF, com parâmetros de dependência α e η, é dada por:

$$C_{\alpha,\eta}(u_1,u_2) = exp\Big[-\frac{1}{\alpha}\Big[\eta^{1-\alpha}(g(u_1)^{1/\alpha} + g(u_2)^{1/\alpha} - \eta)^{\alpha} - \eta\Big]\Big].$$

Essa cópula contém, como casos especiais, as cópulas de Gumbel, Gaussiana Inversa e Clayton.

O modelo de cópula PVF permite apenas a modelagem da dependência positiva. Deve-se ressaltar que os tempos de sobrevivência bivariados (T_1, T_2) são independentes se $\alpha \to 1$ para todos $\eta \in R^{+}$ ou quando $\eta \to \infty$ para todos $\alpha \in (0,1)$. A cópula de comonotonicidade, isto é, $\tau \to 1$, é obtida quando ambos os parâmetros vão para zero.

Para a cópula PVF, o coeficiente tau de Kendall não tem uma expressão de forma fechada, mas pode ser calculado numericamente e é dado por (Romeo, 2017)

$$\tau = 1 + 4\int_0^1 \frac{\varphi(t)}{\varphi'(t)}dt = 1 - \frac{4}{\eta^{\alpha-1}}\int_0^1 \frac{g(t)^{1/\alpha} - \eta}{g(t)^{1/\alpha-1}}dt,$$

em que $\varphi'(s) = -\eta^{\alpha-1}g(s)^{1/\alpha-1}s^{-1}$.

A Figura 11 mostra os gráficos de dispersão de 1000 simulações de uma cópula PVF para diferentes valores de tau de Kendall, 0,9; 0,7; 0,5 e 0,2, respectivamente.

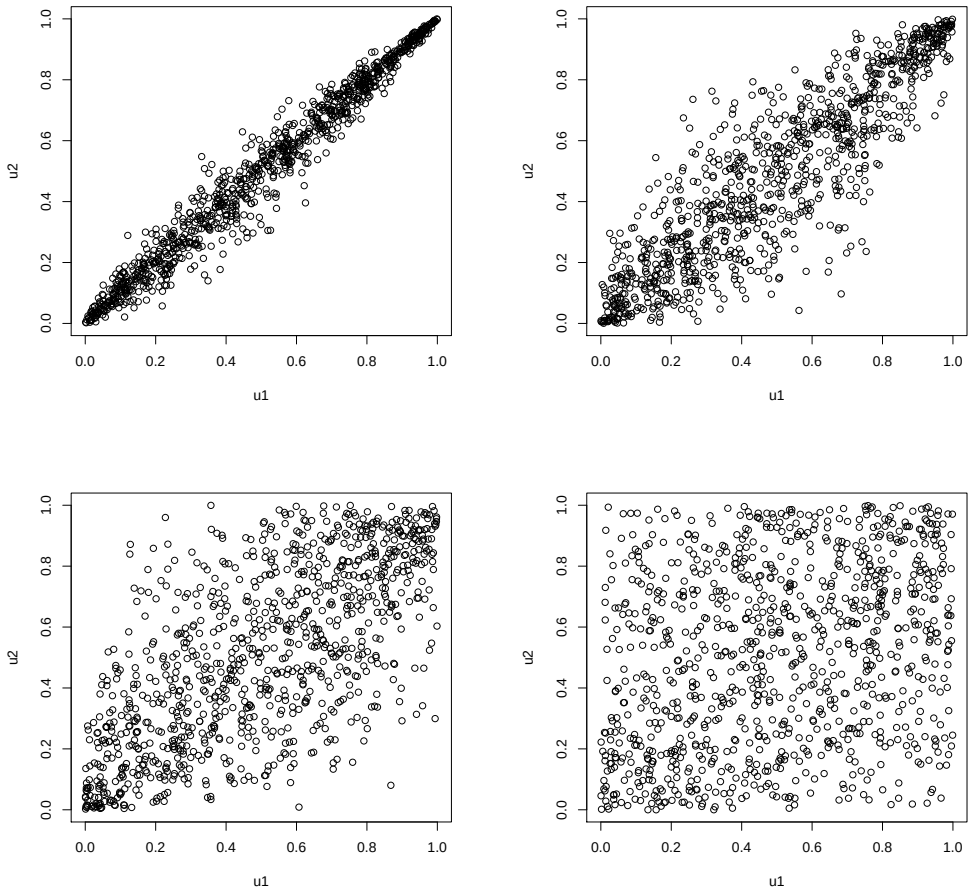

Figura 11 – Gráfico de dispersão com $\tau = 0,9$ (superior esquerdo), $\tau = 0,7$ (superior direito), $\tau = 0,5$ (inferior esquerdo) e $\tau = 0,2$ (inferior direito). (Fonte: Elaborado pelo autor).

CAPÍTULO

3

MODELO DE SOBREVIVÊNCIA PVF BIVARIADO COM MARGINAIS WEIBULL

Nesta seção, estudamos o modelo de sobrevivência PVF bivariado considerando que ambas as distribuições marginais têm distribuição Weibull.

Realizamos um estudo de simulação com a presença de covariáveis, considerando o caso com e sem censura. Por fim, mostramos a aplicabilidade do modelo a um conjunto de dados reais.

3.1 Simulação

Um resultado importante nos conceitos de cópulas e na parte de simulação é a *Transformada Integral de Probabilidade*.

Definição 3.1.1. *Seja X uma variável aleatória com função distribuição $F(x) = P(X \leq x)$ para $x \in \mathbb{R}$. Suponha que $F(x)$ seja contínua. Então, para $u \in (0,1)$, existe um valor mínimo único $x(u)$, tal que $F(x(u)) = u$. Formalmente,*

$$x(u) = F^{-1}(u) = inf\{x; F(x) \geq u\},$$

o qual define a função distribuição inversa. Então $F(x) \leq u$ se, e somente se, $x \leq F^{-1}(u)$. Uma vez que $F(x)$ é não decrescente e contínua então sua inversa $F^{-1}(u)$ também é não decrescente e contínua sobre $u \in (0,1)$. Portanto,

$$P(F(X) \leq u) = P(X \leq F^{-1}(u)) = F(F^{-1}(u)) = u$$

e

$$P(F(X) \leq u) = \begin{cases} 1 \ , & para\ u \geq 1 \\ 0 \ , & para\ u < 0 \end{cases}$$

Consequentemente, $F(X)$ tem distribuição uniforme em $[0,1]$.

A transformação $U = F(X)$ é conhecida como transformada integral de probabilidade e tem aplicações importantes em simulação computacional quando é necessário se gerar uma amostra aleatória de uma dada distribuição de probabilidade (dos Anjos et al, 2004).

Partindo do pressuposto de que os parâmetros do modelo são conhecidos, geramos conjuntos de dados para estudar as propriedades dos estimadores Bayesianos. O objetivo deste estudo de simulação é verificar o bom comportamento das estimativas Bayesianas, com base na média *a posteriori*, além de realizar comparação de modelos por meio das medidas EAIC, EBIC, DIC e LPML.

Para simular n observações (t_{i1}, t_{i2}) do modelo baseado na cópula PVF, assumindo que as marginais T_1 e T_2 têm distribuição Weibull, com parâmetros α_j, η e $\lambda_{ij} = exp(\beta_{0j} + \beta_{1j} x_i)$, $j = 1,2$, realizamos os seguintes passos:

Passo 1: Faça $i = 1$ e fixe o tamanho da amostra n.

Passo 2: Gerar as covariáveis x_i de uma distribuição de Bernoulli com parâmetro 0,5.

Passo 3: 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 4: Gerar $u_1 \sim U(0,1)$ para obter T_{i1} e calcular t_{i1} da seguinte forma: $T_{i1} = (-log(1-u_1)/\lambda_1)^{1/\alpha_1}$. 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 5: Gerar $u_2 \sim U(0,1)$ e obter $w \in (0,1)$, em que w é a solução da equação não linear $u_2 - \frac{\partial C(u_1,w)}{\partial u_1} = 0$.

Passo 6: Obter T_{i2} da seguinte forma: $T_{i2} = (-log(1-w)/\lambda_2)^{1/\alpha_2}$. 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 $t_{i2} = min(T_{i2}; C_{i2})$.

Passo 7: Faça $i = i+1$. Se $i = n$ pare. Caso contrário, retorne ao passo 2.

No *R*, podemos usar o pacote *rootSolve* para encontrar a solução da equação não linear.

3.1.1 Estudo de simulação para casos sem censura

Neste estudo de simulação, geramos os conjuntos de dados assumindo ausência de dados censurados, $(0\%, 0\%)$, para três diferentes tamanhos de amostras $N = 100, 200$ e 300. Para cada caso, geramos 200 conjuntos Monte Carlo (amostras) de dados.

As seguintes distribuições *a priori* independentes foram consideradas para o amostrador de Gibbs: $\alpha_j \sim Gama(0,1;0,1)$, $\beta_{ij} \sim N(0, 10^3)$, $i = 0,1$ e $j = 1,2$. Assumimos $\phi \sim Beta(1;1)$ e $\eta \sim Gama(0,1;0,1)$ para os parâmetros da cópula.

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

Para cada conjunto de dados gerados, consideramos duas cadeias de tamanho 30.000. Para eliminar o efeito dos valores iniciais, foram desconsideradas as primeiras 10.000 iterações. Para evitar problemas de autocorrelação, considerou-se um salto de tamanho 10, obtendo uma amostra efetiva de tamanho 4.000 sobre a qual a inferência *a posteriori* é baseada. Para cada amostra, a média e o desvio-padrão *a posteriori* dos parâmetros são obtidos.

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 ($\leq 1,01$).

Na Tabela 1 temos a média MC das estimativas dos parâmetros ajustando a cópula PVF com distribuição marginal Weibull para o caso sem censura, $(0\%, 0\%)$, juntamente com o percentual de cobertura (PC) para cada parâmetro e o erro quadrático médio (EQM) para três tamanhos de amostras simuladas ($N = 100$, $N = 200$ e $N = 300$). Podemos observar que o percentual de cobertura está próximo, em média, de $0,95$ e os resultados são, em média, melhores para tamanhos de amostras maiores, em que o EQM diminui com o aumento da amostra.

Tabela 1 – Média MC, percentual de cobertura (PC) e EQM das estimativas dos parâmetros ajustando o modelo de PVF bivariado com marginais Weibull para as diferentes configurações de tamanhos de amostras simuladas.

Parâmetro	Valor real	N = 100			N = 200			N = 300		
		Estimativa	PC	EQM	Estimativa	PC	EQM	Estimativa	PC	EQM
β_{01}	-1	-1,030	0,95	0,033	-1,007	0,99	0,013	-1,001	0,98	0,011
β_{02}	1,5	1,524	0,98	0,023	1,511	0,94	0,014	1,513	0,96	0,008
β_{11}	0,5	0,524	0,95	0,040	0,518	0,95	0,019	0,496	0,95	0,014
β_{12}	-0,5	-0,479	0,99	0,035	-0,507	0,97	0,020	-0,518	0,98	0,011
α_1	5	5,119	0,91	0,188	5,032	0,96	0,064	5,030	0,94	0,051
α_2	3	3,073	0,92	0,065	3,029	0,96	0,025	3,016	0,96	0,016
ϕ	0,2	0,372	0,96	0,048	0,302	0,97	0,026	0,254	0,97	0,013
η	1	0,970	0,94	0,197	0,890	0,95	0,112	0,970	0,95	0,099

3.1.1.1 Diagnóstico de Observações Influentes

Para examinar o desempenho da medida de diagnóstico, geramos uma amostra de tamanho 200 para o modelo de cópula PVF bivariado com marginais Weibull, considerando os seguintes valores para os parâmetros: $\beta_{01} = -1,0$, $\beta_{11} = 0,5$, $\alpha_1 = 1,5$, $\beta_{02} = 1,5$, $\beta_{12} = -0,5$, $\alpha_2 = 1.5$, $\eta = 1$ e $\phi = 0,2$.

Selecionamos os casos $50, 100$ e 175 para perturbação. Para criar observações artificialmente influentes no conjunto de dados, escolhemos um, dois ou três destes casos selecionados. Para cada caso, perturbamos os dois tempos de vida da seguinte forma: $\widetilde{t}_{jb} = t_{jb} + 5D_t,\ j = 1,2$ e $b \in \{50, 100, 175\}$, 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 na Seção 3.1.

Na Tabela 2 vemos que as inferências *a posteriori* são sensíveis à perturbação do(s) caso(s) selecionado(s), em que o conjunto de dados (a) denota os dados originais simulados sem nenhuma perturbação e os resultados estão mais próximos dos reais valores considerados como parâmetros, e os casos (b) a (h) denotam os conjuntos de dados com pelo menos algum caso perturbado, em que os valores simulados diferem bastante dos valores dos parâmetros reais.

Tabela 2 – Média, desvio padrão (DP) e intervalo de credibilidade para os parâmetros do modelo de sobrevivência PVF bivariado para cada conjunto de dados simulados.

		β_{01}	β_{02}	β_{11}	β_{12}	α_1	α_2	ϕ	η
Dados	Perturbados	Média (DP) (IC[0.025 ; 0.975])	Média (DP) (IC[0.025 ; 0.975])	Média (DP) (IC[0.025 ; 0.975])	Média (DP) (IC[0.025 ; 0.975])	Média (DP) (IC[0.025 ; 0.975])	Média (DP) (IC[0.025 ; 0.975])	Média (DP) (IC[0.025 ; 0.975])	Média (DP) (IC[0.025 ; 0.975])
a	Nenhum	-0,953 (0,120) (-1,180 ; -0,719)	1,505 (0,114) (1,280 ; 1,732)	0,569 (0,138) (0,293 ; 0,834)	-0,368 (0,140) (-0,644 ; -0,097)	1,219 (0,216) (0,815 ; 1,667)	1,132 (0,194) (0,794 ; 1,544)	0,280 (0,141) (0,022 ; 0,552)	0,864 (0,405) (0,269 ; 1,848)
b	50	-0,885 (0,118) (-1,121 ; -0,661)	1,276 (0,101) (1,077 ; 1,470)	0,680 (0,138) (0,412 ; 0,966)	-0,222 (0,132) (-0,478 ; 0,037)	1,028 (0,181) (0,689 ; 1,414)	1,112 (0,147) (0,843 ; 1,413)	0,105 (0,071) (0,005 ; 0,270)	0,897 (0,282) (0,444 ; 1,560)
c	100	-0,743 (0,107) (-0,953 ; -0,534)	1,468 (0,112) (1,255 ; 1,682)	0,283 (0,136) (0,017 ; 0,545)	-0,454 (0,140) (-0,739 ; -0,184)	1,069 (0,179) (0,739 ; 1,421)	0,928 (0,154) (0,652 ; 1,242)	0,249 (0,101) (0,064 ; 0,453)	0,753 (0,315) (0,305 ; 1,505)
d	175	-0,796 (0,107) (-1,007 ; -0,587)	1,423 (0,110) (1,209 ; 1,634)	0,344 (0,131) (0,085 ; 0,601)	-0,421 (0,136) (-0,683 ; -0,153)	1,138 (0,181) (0,800 ; 1,508)	1,021 (0,165) (0,728 ; 1,365)	0,147 (0,092) (0,009 ; 0,350)	0,906 (0,315) (0,408 ; 1,641)
e	50 e 100	-0,756 (0,107) (-0,971 ; -0,551)	1,228 (0,096) (1,036 ; 1,407)	0,408 (0,129) (0,155 ; 0,657)	-0,277 (0,129) (-0,533 ; -0,028)	0,926 (0,156) (0,639 ; 1,243)	1,003 (0,131) (0,766 ; 1,274)	0,120 (0,066) (0,014 ; 0,264)	0,761 (0,245) (0,385 ; 1,327)
f	50 e 175	-0,811 (0,110) (-1,027 ; -0,593)	1,227 (0,098) (1,031 ; 1,414)	0,499 (0,126) (0,255 ; 0,740)	-0,299 (0,127) (-0,544 ; -0,053)	1,039 (0,157) (0,754 ; 1,362)	1,034 (0,134) (0,788 ; 1,316)	0,070 (0,055) (0,003 ; 0,202)	0,800 (0,213) (0,447 ; 1,275)
g	100 e 175	-0,708 (0,105) (-0,914 ; -0,499)	1,382 (0,108) (1,168 ; 1,594)	0,184 (0,135) (-0,079 ; 0,453)	-0,491 (0,136) (-0,765 ; -0,222)	1,067 (0,158) (0,781 ; 1,393)	0,925 (0,146) (0,662 ; 1,230)	0,162 (0,082) (0,026 ; 0,340)	0,763 (0,272) (0,345 ; 1,405)
h	50, 100 e 175	-0,731 (0,108) (-0,950 ; -0,525)	1,179 (0,099) (0,982 ; 1,368)	0,311 (0,128) (0,064 ; 0,565)	-0,353 (0,129) (-0,607 ; -0,103)	0,960 (0,137) (0,708 ; 1,245)	0,968 (0,123) (0,740 ; 1,224)	0,095 (0,056) (0,009 ; 0,222)	0,674 (0,196) (0,363 ; 1,128)

A Tabela 3 evidência que que o conjunto de dados (a), sem nenhum caso perturbado, teve o melhor ajuste de acordo com todos os critérios Bayesianos dentre todos os diferentes casos de conjuntos de dados.

Tabela 3 – Critérios Bayesianos ajustando o modelo de sobrevivência PVF bivariado para cada conjunto de dados simulados.

Conjunto de dados	DIC	EAIC	EBIC	LPML
a	**935,559**	**901,523**	**927,909**	**-466,600**
b	993,120	976,720	1003,106	-495,848
c	1079,412	1050,788	1077,174	-542,966
d	1034,990	1013,013	1039,400	-522,678
e	1043,962	1029,007	1055,393	-520,487
f	1010,953	1001,316	1027,702	-507,263
g	1077,337	1055,734	1082,121	-540,676
h	1038,550	1025,062	1051,448	-516,611

Consideramos as amostras da distribuição *a posteriori* dos parâmetros do modelo de sobrevivência PVF bivarido para obter uma estimativa das três medidas de divergência apresentadas. Os resultados da Tabela 4 mostram que todos os casos que não foram perturbados tiveram pequenas medidas de divergência, mesmo quando fazem parte do conjunto de dados com algum caso perturbado, entretanto evidenciam que as três medidas aumentam e detectam quando algum caso é influente.

Tabela 4 – Medidas de divergência para os dados simulados.

Conjunto de dados	Casos perturbados	d_{K-L}	d_J	d_{L-1}
a	50	0,247	0,520	0,278
	100	0,083	0,172	0,162
	175	0,099	0,205	0,177
b	**50**	**2,136**	**5,021**	**0,734**
	100	0,051	0,103	0,129
	175	0,066	0,132	0,147
c	50	0,247	0,516	0,280
	100	**0,628**	**1,423**	**0,447**
	175	0,086	0,176	0,167
d	50	0,240	0,489	0,277
	100	0,068	0,136	0,149
	175	**2,224**	**6,464**	**0,780**
e	**50**	**1,341**	**3,014**	**0,615**
	100	**0,851**	**1,836**	**0,513**
	175	0,059	0,12	0,137
f	**50**	**2,442**	**6,068**	**0,784**
	100	0,026	0,051	0,090
	175	**2,462**	**5,998**	**0,796**
g	50	0,199	0,407	0,249
	100	**0,627**	**1,575**	**0,440**
	175	**1,552**	**4,214**	**0,683**
h	**50**	**1,252**	**2,655**	**0,587**
	100	**0,703**	**1,552**	**0,472**
	175	**1,367**	**2,976**	**0,634**

A Figura 19 mostra os gráficos de índices das três medidas de divergência para os casos (b) e (f). Claramente, podemos ver que as três medidas de divergência detectaram os pontos influentes.

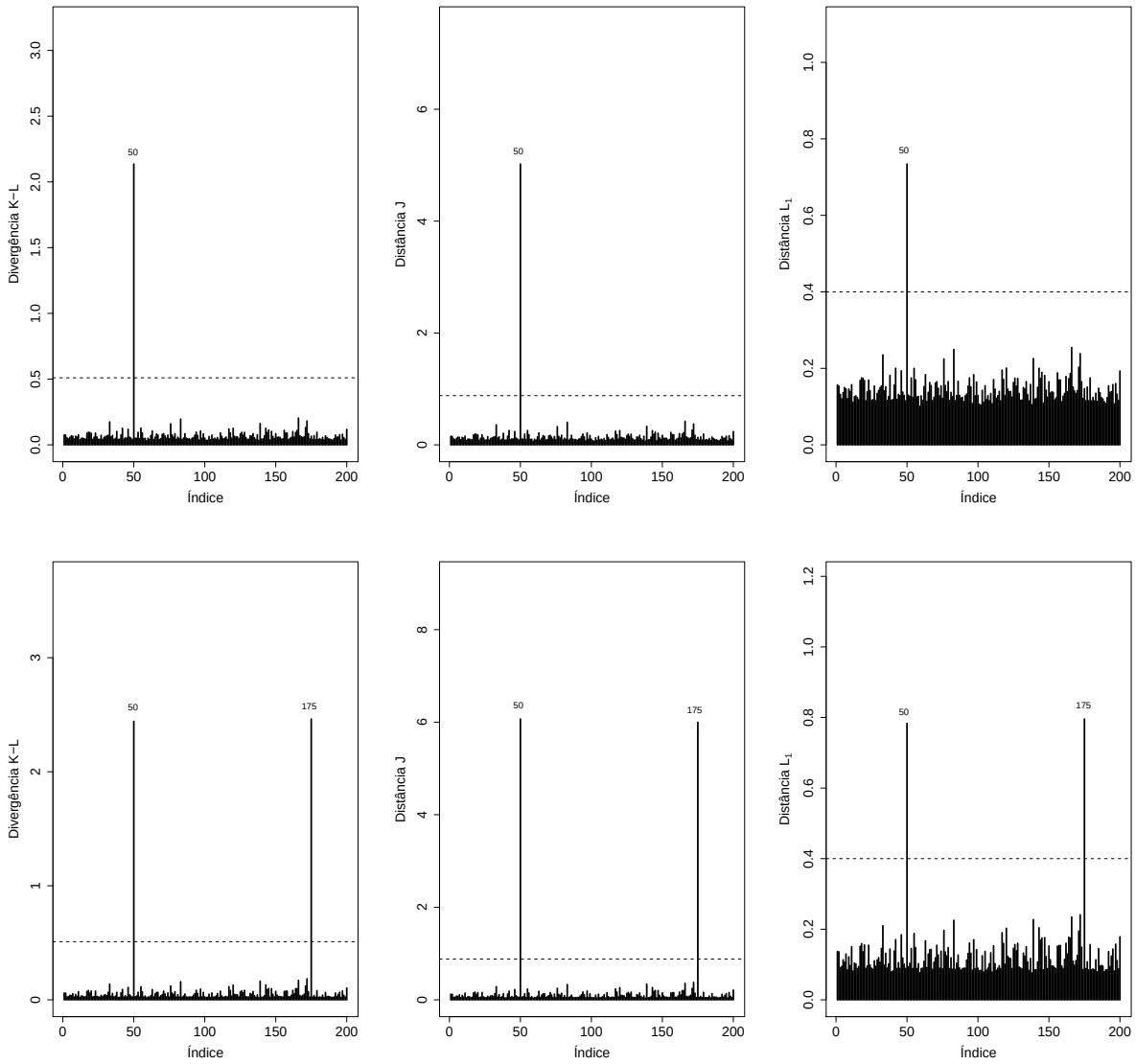

Figura 12 – Gráfico de índices das medidas de divergência para o conjunto de dados (b) (superior) e (f) (inferior). (Fonte: Elaborado pelo autor).

3.1.2 Estudo de simulação para casos com censura

Neste estudo de simulação, geramos os conjuntos de dados assumindo a presença de dados censurados, $(15\%, 15\%)$, para três diferentes tamanhos de amostras $N = 100, 200$ e 300. Para cada caso, geramos 200 conjuntos Monte Carlo (amostras) de dados.

As seguintes distribuições *a priori* independentes foram consideradas para o amostrador de Gibbs: $\alpha_j \sim Gama(0,1;0,1)$, $\beta_{ij} \sim N(0, 10^3)$, $i = 0,1$ e $j = 1,2$. Assumimos $\phi \sim Beta(1;1)$ e $\eta \sim Gama(0,1;0,1)$ para os parâmetros da cópula.

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

Para cada conjunto de dados gerados, consideramos duas cadeias de tamanho 30.000.

Para eliminar o efeito dos valores iniciais, foram desconsideradas as primeiras 10.000 iterações. De forma análoga ao caso anterior, para evitar problemas de autocorrelação, considerou-se um salto de tamanho 10, obtendo uma amostra efetiva de tamanho 4.000 sobre a qual a inferência *a posteriori* é baseada. Para cada amostra, a média e o desvio-padrão *a posteriori* dos parâmetros são obtidos.

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 ($\leq 1,01$).

Na Tabela 5 temos a média MC das estimativas dos parâmetros ajustando a cópula PVF com distribuição marginal Weibull para o caso considerando censura, $(15\%, 15\%)$, e três tamanhos de amostras simuladas ($N = 100$, $N = 200$ e $N = 300$). Apesar de individualmente algumas estimativas dos parâmetros não ficarem mais próximas conforme aumenta o tamanho da amostra, o EQM dessas estimativas diminuem com o aumento do tamanho da amostra, indicando que o aumento da amostra faz com que as estimativas sejam melhores avaliando-se o erro médio.

Tabela 5 – Média MC, percentual de cobertura (PC) e EQM das estimativas dos parâmetros ajustando o modelo de PVF bivariado com marginais Weibull para as diferentes configurações de tamanhos de amostras simuladas para o caso com censura.

Parâmetro	Valor real	N = 100			N = 200			N = 300		
		Estimativa	PC	EQM	Estimativa	PC	EQM	Estimativa	PC	EQM
β_{01}	-0,5	-0,463	0,98	0,042	-0,417	0,98	0,025	-0,416	1,00	0,021
β_{02}	-1	-0,867	0,99	0,065	-0,854	0,99	0,047	-0,856	0,99	0,039
β_{11}	1	1,015	0,98	0,055	0,975	0,94	0,035	0,973	0,94	0,022
β_{12}	2	1,941	0,93	0,109	1,929	0,93	0,052	1,921	0,90	0,037
α_1	5	5,131	0,96	0,267	5,038	0,95	0,106	5,022	0,98	0,065
α_2	3	3,023	0,90	0,126	2,996	0,93	0,044	2,979	0,95	0,024
ϕ	0,2	0,445	0,95	0,083	0,358	0,95	0,044	0,295	0,98	0,023
η	1	1,145	0,97	0,656	0,968	0,97	0,204	1,021	0,98	0,122

3.1.2.1 *Diagnóstico de Observações Influentes*

Para examinar o desempenho da medida de diagnóstico, geramos uma amostra de tamanho 200 para o modelo de cópula PVF bivariado com marginais Weibull, considerando os seguintes valores para os parâmetros: $\beta_{01} = -1,0$, $\beta_{11} = 0,5$, $\alpha_1 = 2$, $\beta_{02} = 1,5$, $\beta_{12} = -0,5$, $\alpha_2 = 0.5$, $\eta = 1$ e $\phi = 0,2$.

Selecionamos os casos $45, 90$ e 160 para perturbação, em que nos casos 45 e 90 são observados ambos tempos de falha e o caso 160 apresenta os dois tempo censurados. Para criar observações artificialmente influentes no conjunto de dados, escolhemos um, dois ou três destes casos selecionados. Para cada caso, perturbamos os dois tempos de vida da seguinte forma: $\widetilde{t}_{jb} = t_{jb} + 5D_t,\ j = 1,2$ e $b \in \{45, 190, 160\}$, 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 na Seção 3.1.

Na Tabela 6 vemos que as inferências *a posteriori* são sensíveis à perturbação do(s) caso(s) selecionado(s), em que o conjunto de dados (a) denota os dados originais simulados sem nenhuma perturbação, e os casos (b) a (h) denotam os conjuntos de dados com pelo menos algum caso perturbado. Observamos que o melhor ajuste considerando a coleção de todos os parâmetros do modelo foi dado no conjunto de dados (a), sem nenhuma perturbação, principalmente nos parâmetros β_{01} que teve estimativa igual a $-0,919$ contra $-0,582$ como estimativa para o caso em que perturbamos apenas a observação 45 no caso (b), por exemplo. Essa grande superioridade no ajuste também é observada em outros parâmetros como, por exemplo, β_{02}, β_{12} e η.

Tabela 6 – Média, desvio padrão (DP) e intervalo de credibilidade para os parâmetros do modelo de sobrevivência PVF bivariado para cada conjunto de dados simulados com censura

		β_{01}	β_{02}	β_{11}	β_{12}	α_1	α_2	ϕ	η
Dados	Perturbados	Média (DP) (IC[0.025 ; 0.975])	Média (DP) (IC[0.025 ; 0.975])	Média (DP) (IC[0.025 ; 0.975])	Média (DP) (IC[0.025 ; 0.975])	Média (DP) (IC[0.025 ; 0.975])	Média (DP) (IC[0.025 ; 0.975])	Média (DP) (IC[0.025 ; 0.975])	Média (DP) (IC[0.025 ; 0.975])
a	Nenhum	-0,919 (0,132) (-1,185 ; -0,663)	1,607 (0,127) (1,354 ; 1,855)	0,478 (0,155) (0,170 ; 0,790)	-0,585 (0,150) (-0,880 ; -0,294)	2,284 (0,278) (1,765 ; 2,849)	0,233 (0,070) (0,111 ; 0,384)	0,280 (0,166) (0,018 ; 0,640)	1,102 (0,595) (0,233 ; 2,540)
b	45	-0,582 (0,113) (-0,806 ; -0,367)	1,594 (0,124) (1,348 ; 1,835)	0,019 (0,150) (-0,274 ; 0,313)	-0,680 (0,153) (-0,986 ; -0,383)	1,971 (0,186) (1,616 ; 2,360)	0,320 (0,059) (0,213 ; 0,444)	0,332 (0,110) (0,128 ; 0,559)	0,624 (0,327) (0,193 ; 1,466)
c	90	-0,595 (0,113) (-0,820 ; -0,381)	1,574 (0,123) (1,322 ; 1,806)	0,028 (0,149) (-0,267 ; 0,323)	-0,671 (0,153) (-0,980 ; -0,371)	1,993 (0,188) (1,655 ; 2,383)	0,332 (0,059) (0,223 ; 0,454)	0,238 (0,102) (0,051 ; 0,454)	0,788 (0,371) (0,293 ; 1,676)
d	160	-0,861 (0,130) (-1,119 ; -0,609)	1,314 (0,100) (1,109 ; 1,502)	0,566 (0,158) (0,260 ; 0,876)	-0,393 (0,143) (-0,677 ; -0,125)	1,784 (0,204) (1,407 ; 2,199)	0,311 (0,061) (0,200 ; 0,437)	0,074 (0,062) (0,002 ; 0,238)	0,989 (0,314) (0,508 ; 1,752)
e	45 e 90	-0,511 (0,111) (-0,730 ; -0,292)	1,556 (0,123) (1,314 ; 1,793)	-0,140 (0,151) (-0,437 ; 0,160)	-0,727 (0,151) (-1,018 ; -0,429)	1,794 (0,163) (1,506 ; 2,132)	0,321 (0,052) (0,225 ; 0,426)	0,235 (0,097) (0,053 ; 0,429)	0,681 (0,321) (0,230 ; 1,490)
f	45 e 160	-0,680 (0,119) (-0,918 ; -0,447)	1,221 (0,104) (1,013 ; 1,416)	0,228 (0,149) (-0,065 ; 0,519)	-0,417 (0,144) (-0,698 ; -0,134)	1,664 (0,171) (1,359 ; 2,019)	0,327 (0,054) (0,231 ; 0,438)	0,180 (0,078) (0,044 ; 0,352)	0,610 (0,278) (0,243 ; 1,279)
g	90 e 160	-0,680 (0,118) (-0,912 ; -0,455)	1,221 (0,101) (1,015 ; 1,416)	0,221 (0,146) (-0,061 ; 0,504)	-0,411 (0,143) (-0,696 ; -0,136)	1,685 (0,176) (1,368 ; 2,061)	0,330 (0,054) (0,232 ; 0,439)	0,126 (0,068) (0,016 ; 0,280)	0,699 (0,263) (0,329 ; 1,341)
h	45, 90 e 160	-0,606 (0,115) (-0,845 ; -0,390)	1,178 (0,101) (0,978 ; 1,370)	0,057 (0,148) (-0,239 ; 0,342)	-0,438 (0,144) (-0,720 ; -0,153)	1,550 (0,157) (1,264 ; 1,884)	0,313 (0,050) (0,218 ; 0,411)	0,164 (0,070) (0,038 ; 0,308)	0,556 (0,234) (0,236 ; 1,129)

A Tabela 7 evidencia que o conjunto de dados (a), sem nenhum caso perturbado teve o melhor ajuste de acordo com todos os critérios Bayesianos dentre todos os diferentes conjuntos de dados.

Tabela 7 – Critérios Bayesianos ajustando o modelo de sobrevivência PVF bivariado para cada conjunto de dados simulados com censura

Conjunto de dados	DIC	EAIC	EBIC	LPML
a	**1435,822**	**1413,754**	**1440,140**	**-718,113**
b	1493,664	1465,901	1492,288	-756,026
c	1524,370	1492,797	1519,183	-775,590
d	1547,010	1522,792	1549,179	-777,673
e	1524,370	1492,797	1519,183	-775,590
f	1547,010	1522,792	1549,179	-777,673
g	1571,034	1538,979	1565,366	-796,161
h	1741,763	1691,360	1717,746	-884,941

Consideramos as amostras da distribuição *a posteriori* dos parâmetros do modelo de sobrevivência PVF bivariado para obter uma estimativa das três medidas de divergência apresentadas. Os resultados da Tabela 8 mostram que todos os casos que não foram perturbados tiveram pequenas medidas de divergência, mesmo quando fazem parte do conjunto de dados com algum caso perturbado, entretanto evidenciam que as três medidas aumentam e detectam quando algum caso é influente.

Tabela 8 – Medidas de divergência para os dados simulados com censura.

Dados	Perturbados	d_{K-L}	d_J	d_{L-1}
a	45	0,210	0,441	0,260
	90	0,183	0,384	0,244
	160	0,368	0,753	0,341
b	**45**	**1,344**	**4,786**	**0,654**
	90	0,109	0,227	0,187
	160	0,277	0,576	0,298
c	45	0,147	0,318	0,218
	90	**1,028**	**2,61**	**0,569**
	160	0,327	0,685	0,321
d	45	0,136	0,280	0,209
	90	0,129	0,265	0,204
	160	**0,699**	**1,883**	**0,476**
e	**45**	**0,595**	**1,460**	**0,434**
	90	**0,623**	**1,533**	**0,444**
	160	0,335	0,696	0,325
f	**45**	**1,061**	**2,712**	**0,592**
	90	0,105	0,220	0,184
	160	**0,651**	**1,829**	**0,447**
g	45	0,111	0,238	0,188
	90	**1,385**	**3,608**	**0,678**
	160	**1,164**	**3,290**	**0,622**
h	**45**	**0,588**	**1,567**	**0,441**
	90	**0,607**	**1,634**	**0,448**
	160	**0,605**	**1,602**	**0,438**

A Figura 13 mostra os gráficos de índices das três medidas de divergência para os casos (d) e (h). Claramente, podemos ver que as três medidas de divergência detectaram os pontos influentes, mesmo o caso (d) tendo a observação 160 com ambos os tempos censurados.

3.2 Aplicação a Dados Reais

Nesta seção, aplicamos o modelo proposto a um conjunto de dados reais de retinopatia diabética (The Diabetic Retinopathy Study Research Group, 1976).

3.2.1 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.

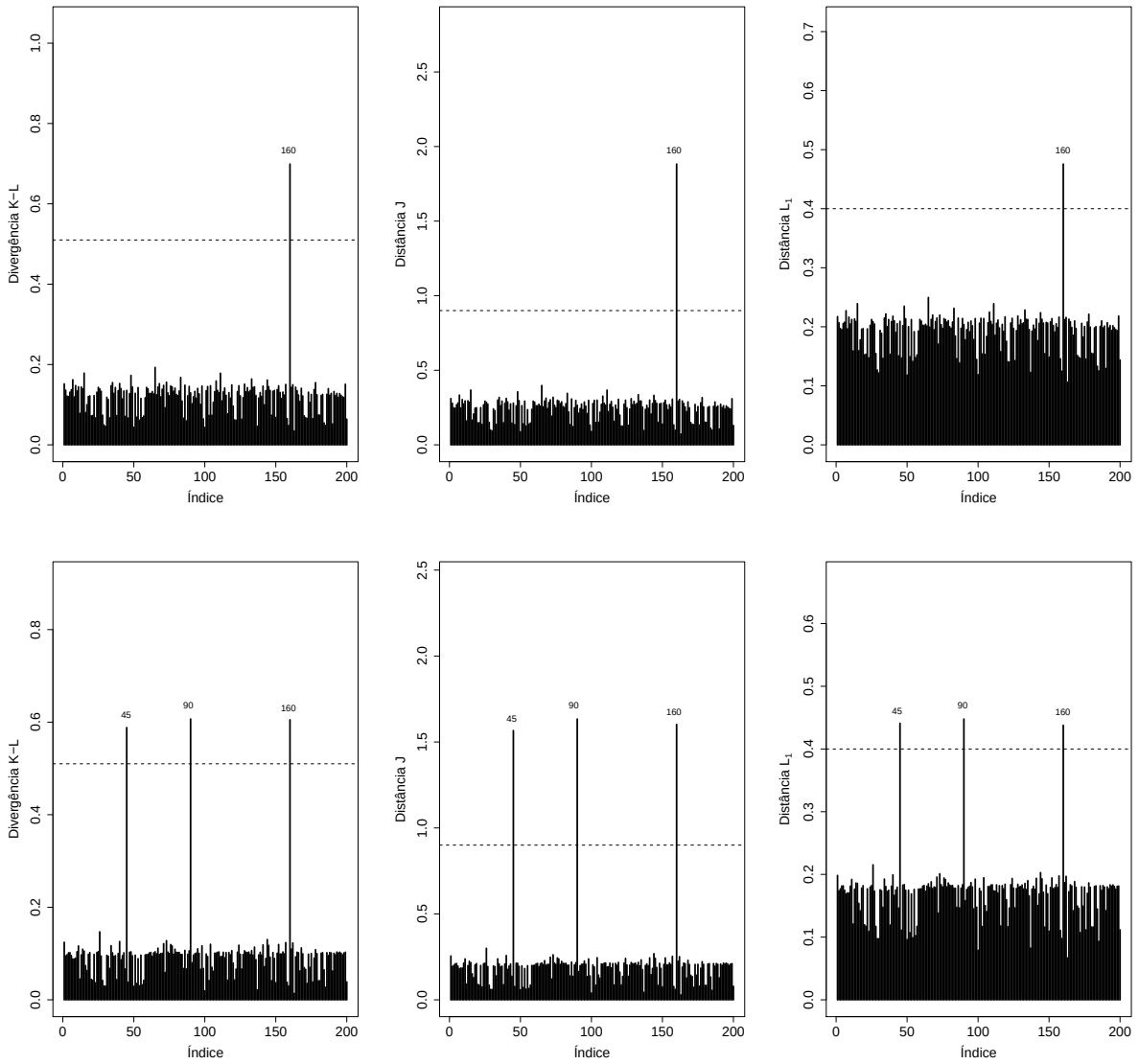

Figura 13 – Gráfico de índices das medidas de divergência para o conjunto de dados (d) (superior) e (h) (inferior). (Fonte: Elaborado pelo autor).

O sintoma mais comum é a vista 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, para a retinopatia diabética, em retardar o aparecimento da cegueira. Para esta análise do tratamento foi realizada uma pesquisa composta por 197 pacientes (The Diabetic Retinopathy Study Research Group, 1976).

O tratamento foi aleatoriamente 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 no início da diabete foi considerada como covariável e 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 visual para o olho de tratamento e T_2 como o vetor de tempos até a perda visual para o olho controle.

Com uma breve análise desse conjunto de dados, podemos ver na Figura 14 pela curva de Kaplan-Meier a existência de indivíduos que poderiam ser ditos curados, principalmente no Tempo 1, e pelo gráfico TTT, notamos que a maioria das observações estão acima da reta identidade, concluindo que a função de risco é crescente, sugerindo que a distribuição Weibull adotada é adequada para a modelagem desses dados.

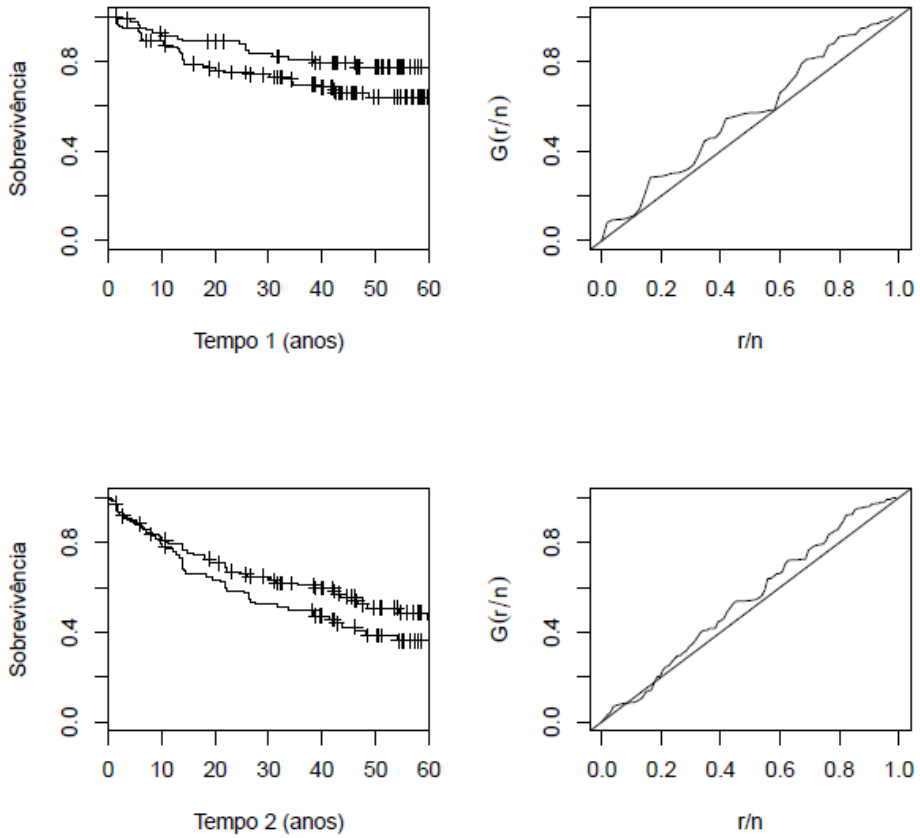

Figura 14 – Estimativas de Kaplan-Meier da função de sobrevivência e gráficos TTT para os dados de retinopatia diabética. (Fonte: Elaborado pelo autor).

Outra informação que podemos ter através de uma análise inicial da Figura 15, é que os tempos são correlacionados. O valor da correlação através da medida tau de Kendall é de $0,46$, mostrando uma associação fraca entre os tempos.

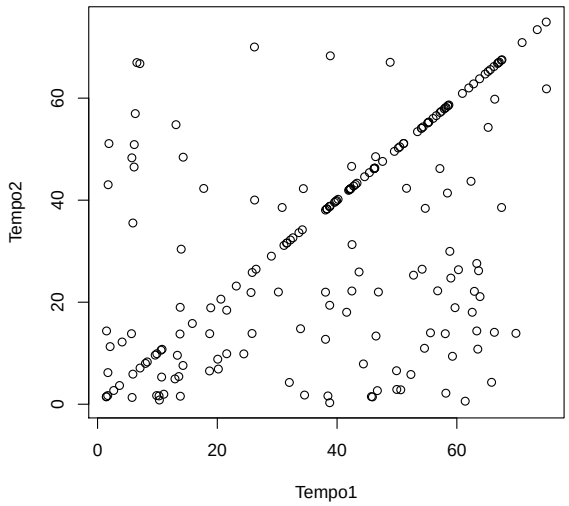

Figura 15 – Gráfico de dispersão para o conjunto de dados de retinopatia diabética. (Fonte: Elaborado pelo autor).

Ajustamos o modelo PVF bivariado com ambas distribuições marginais Weibull, considerando duas cadeias de tamanho 100.000 para cada parâmetro, desconsiderando as primeiras 60.000 iterações para eliminar o efeito dos valores iniciais e, para evitar problemas de autocorrelação, foi considerado um salto de tamanho 20, obtendo uma amostra efetiva de tamanho 4.000 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).

Para todos os parâmetros do modelo especificamos distribuições a priori não informativas. Sendo que, para os parâmetros da cópula foi especificado $\phi \sim Beta(1,1)$ e $\eta \sim Gamma(0.1,0.1)$. Para as marginais foi especificado $\beta_{ij} \sim N(0,10^3)$ e $\alpha_j \sim Gamma(0.1,0.1)$, em que $i = 0,1$ e $j = 1,2$.

Na Tabela 9 são apresentados as médias *a posteriori* para os parâmetros PVF bivariado com ambas distribuições marginais Weibull. Vale notar que as estimativas dos parâmetros ϕ e η, de dependência da cópula, não dependem das marginais utilizadas.

Tabela 9 – Média *a posteriori*, desvio padrão (DP) e intervalo de credibilidade de 95% para os parâmetros do modelo PVF bivariado com distribuições marginais Weibull.

	Parâmetro	Média	DP	IC[0.025 ; 0.975]
Tempo 1	α_1	0,802	0,099	[0,607 ; 0,987]
	β_{01}	-4,012	0,404	[-4,775 ; -3,212]
	β_{11}	-0,495	0,289	[-1,106 ; 0,021]
Tempo 2	α_2	0,826	0,071	[0,689 ; 0,963]
	β_{02}	-3,669	0,297	[-4,262 ; -3,114]
	β_{12}	0,369	0,199	[0,003 ; 0,762]
Cópula	ϕ	0,321	0,212	[0,001 ; 0,689]
	η	1,068	1,237	[0,002 ; 2,998]

Na literatura, é de conhecimento que nesse conjunto de dados não há nenhum ponto influente. Na Figura 16 vemos pelos gráficos de índices, considerando o modelo PVF com distribuição Weibull, que todas as medidas não detectaram nenhum ponto influente.

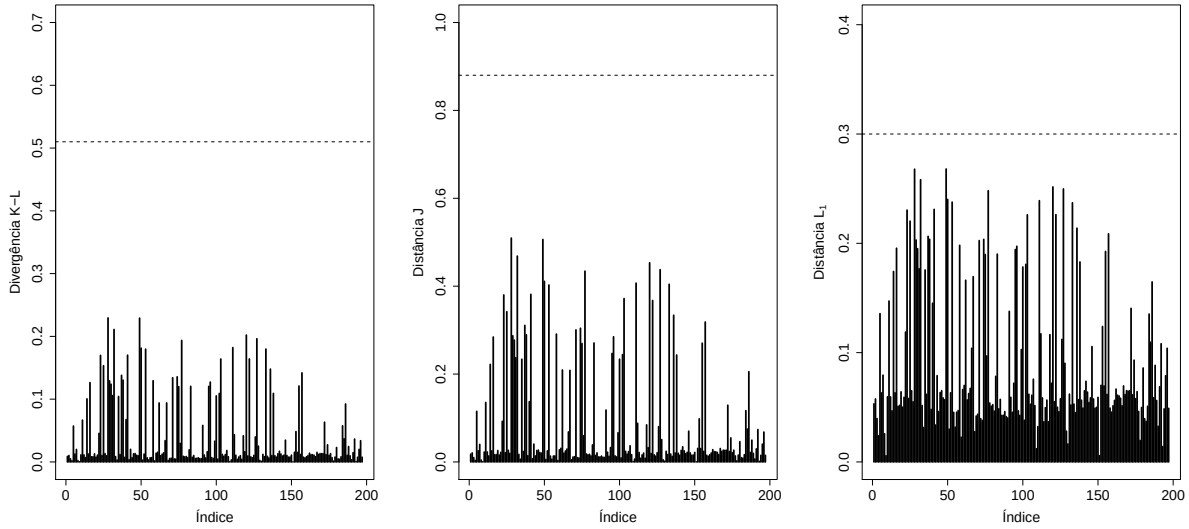

Figura 16 – Gráfico de índices das medidas de divergência para o conjunto de dados reais. (Fonte: Elaborado pelo autor).

A Figura 17 mostra 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 PVF com distribuições marginais Weibull.

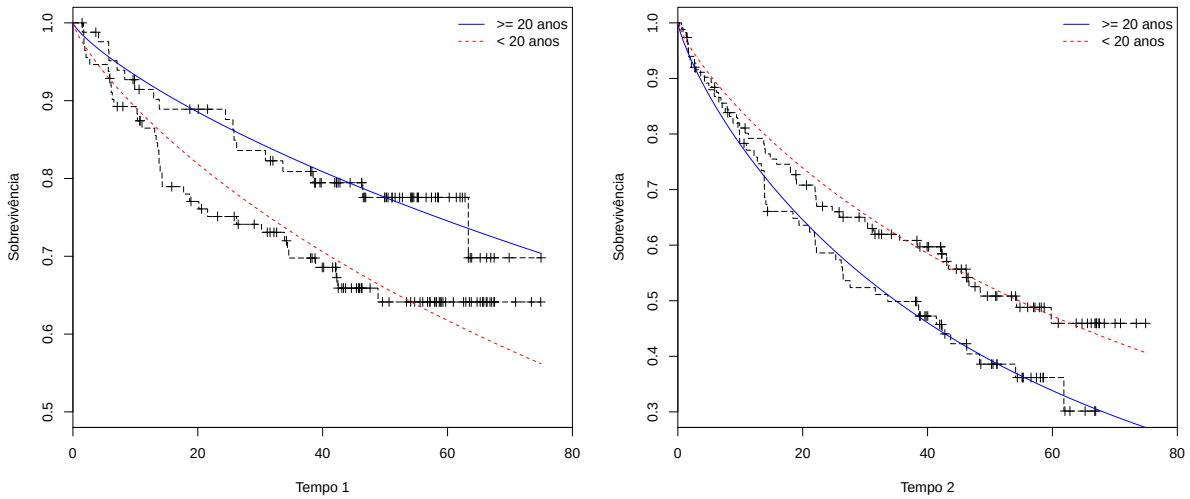

Figura 17 – Curvas de Kaplan-Meier e curvas de sobrevivências Weibull estimadas para o conjunto de dados reais. (Fonte: Elaborado pelo autor).

Podemos observar, na medida do possível, o bom ajuste do modelo bivariado PVF. Observamos também, de acordo com as curvas de Kaplan-Meier, que o tempo de sobrevivência para a variável T_1, tempo até a perde visual para o olho de tratamento, não tende a zero, isto é, pode-se considerar que há uma fração de cura e, com isso, fazendo que o ponto que há o menor ajuste do modelo proposto seja exatamente esse.

CAPÍTULO

4

MODELO DE SOBREVIVÊNCIA PVF BIVARIADO COM MARGINAIS EXPONENCIAL GENERALIZADA

Nesta seção, estudamos o modelo de sobrevivência PVF bivariado considerando que ambas as distribuições marginais têm distribuição Exponencial Generalizada.

Análogo ao capítulo anterior, realizamos um estudo de simulação com a presença de covariáveis, considerando o caso com e sem censura. Por fim, mostramos a aplicabilidade do modelo a um conjunto de dados reais e o comparamos ao modelo PVF bivariado com distribuição Weibull.

4.1 Simulação

Para simular n observações (t_{i1}, t_{i2}) do modelo baseado na cópula PVF, assumindo que as marginais T_1 e T_2 têm distribuição Exponencial Generalizada, com parâmetros α_j, η e $\lambda_{ij} = exp(\beta_{0j} + \beta_{1j} x_i)$, $j = 1,2$, realizamos os passos de 1 a 7 apresentados na Seção 3.1.

4.1.1 Estudo de simulação para casos sem censura

Neste estudo de simulação, geramos os conjuntos de dados assumindo ausência de dados censurados, $(0\%, 0\%)$, para três diferentes tamanhos de amostras $N = 100, 200$ e 300. Para cada caso, geramos 200 conjuntos MC (amostras) de dados.

As seguintes distribuições *a priori* independentes foram consideradas para o amostrador de Gibbs: $\alpha_j \sim Gama(0,1;0,1)$, $\beta_{ij} \sim N(0, 10^3)$, $i = 0,1$ e $j = 1,2$. Assumimos $\phi \sim Beta(1;1)$ e $\eta \sim Gama(0,1;0,1)$ para os parâmetros da cópula.

Para o modelo com marginais Exponencial Generalizada foram considerados os seguin-

Tabela 10 – Média MC, percentual de cobertura (PC) e EQM das estimativas dos parâmetros ajustando o modelo de PVF bivariado com marginais Exponencial Generalizada para as diferentes configurações de tamanhos de amostras simuladas.

Parâmetro	Valor real	N = 100			N = 200			N = 300		
		Estimativa	PC	EQM	Estimativa	PC	EQM	Estimativa	PC	EQM
β_{01}	1	1,012	0,93	0,049	1,008	0,96	0,020	0,996	0,97	0,012
β_{02}	1,5	1,508	0,94	0,027	1,506	0,92	0,013	1,499	0,94	0,008
β_{11}	-0,5	-0,493	0,95	0,080	-0,496	0,98	0,036	-0,503	0,99	0,020
β_{12}	-0,5	-0,472	0,99	0,036	-0,500	0,96	0,018	-0,488	0,97	0,012
α_1	0,5	0,509	0,98	0,003	0,510	0,94	0,002	0,505	0,96	0,001
α_2	1	1,032	0,94	0,021	1,016	0,96	0,008	1,020	0,94	0,006
ϕ	0,4	0,411	0,90	0,010	0,410	0,96	0,005	0,400	0,92	0,003
η	0,1	0,139	0,90	0,012	0,114	0,94	0,004	0,113	0,92	0,003

tes valores para os parâmetros: $\alpha_1 = 0,5,\ \beta_{01} = 1,\ \beta_{11} = -0,5,\ \alpha_2 = 1,\ \beta_{02} = 1,5,\ \beta_{12} = -0,5,\ \eta = 0,1$ e $\phi = 0,4$.

Para cada conjunto de dados gerados, consideramos duas cadeias de tamanho 30.000. Para eliminar o efeito dos valores iniciais, foram desconsideradas as primeiras 10.000 iterações. Para evitar problemas de autocorrelação, considerou-se um salto de tamanho 10, obtendo uma amostra efetiva de tamanho 4.000 sobre a qual a inferência *a posteriori* é baseada. Para cada amostra, a média e o desvio-padrão *a posteriori* dos parâmetros são obtidos.

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 ($\leq 1,01$).

Na Tabela 10 temos a média MC das estimativas dos parâmetros ajustando a cópula PVF com distribuição marginal Exponencial Generalizada para o caso sem censura, $(0\%, 0\%)$, juntamente com o PC para cada parâmetro e o EQM para três tamanhos de amostras simuladas ($N = 100$, $N = 200$ e $N = 300$). Podemos observar que o percentual de cobertura está próximo, em média, de $0,95$ e os resultados são, em média, melhores para tamanhos de amostras maiores, em que o EQM diminui com o aumento da amostra.

4.1.1.1 Diagnóstico de Observações Influentes

Para examinar o desempenho da medida de diagnóstico, geramos uma amostra de tamanho 300 para o modelo de cópula PVF bivariado com marginais Exponencial Generalizada, considerando os seguintes valores para os parâmetros: $\beta_{01} = 1,0$, $\beta_{11} = -0,5$, $\alpha_1 = 0,5$, $\beta_{02} = 1,5$, $\beta_{12} = -0,5$, $\alpha_2 = 1$, $\eta = 0,1$ e $\phi = 0,4$.

Selecionamos os casos $8, 63$ e 210 para perturbação. Para criar observações artificialmente influentes no conjunto de dados, escolhemos um, dois ou três destes casos selecionados. Para cada caso, perturbamos os dois tempos de vida da seguinte forma: $\widetilde{t}_{jb} = t_{jb} + 5D_t,\ j = 1,2$ e $b \in \{8,63,210\}$, 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 na Seção 3.1.

Na Tabela 11 vemos que as inferências *a posteriori* são sensíveis à perturbação do(s) caso(s) selecionado(s), em que o conjunto de dados (a) denota os dados originais simulados sem nenhuma perturbação e os resultados estão mais próximos dos reais valores considerados como parâmetros, e os casos (b) a (h) denotam os conjuntos de dados com pelo menos algum caso perturbado, em que os valores simulados diferem bastante dos valores dos parâmetros reais.

Tabela 11 – Média, desvio padrão (DP) e intervalo de credibilidade para os parâmetros do modelo de sobrevivência PVF bivariado Exp. Gen. para cada conjunto de dados simulados.

		β_{01}	β_{02}	β_{11}	β_{12}	α_1	α_2	ϕ	η
Dados	Perturbados	Média (DP) (IC[0.025 ; 0.975])	Média (DP) (IC[0.025 ; 0.975])	Média (DP) (IC[0.025 ; 0.975])	Média (DP) (IC[0.025 ; 0.975])	Média (DP) (IC[0.025 ; 0.975])	Média (DP) (IC[0.025 ; 0.975])	Média (DP) (IC[0.025 ; 0.975])	Média (DP) (IC[0.025 ; 0.975])
a	Nenhum	0,967 (0,114) (0,739 ; 1,182)	1,524 (0,090) (1,342 ; 1,686)	-0,616 (0,150) (-0,909 ; -0,336)	-0,540 (0,106) (-0,756 ; -0,328)	0,492 (0,023) (0,453 ; 0,541)	1,025 (0,070) (0,893 ; 1,167)	0,394 (0,051) (0,290 ; 0,491)	0,092 (0,042) (0,030 ; 0,193)
b	8	0,865 (0,121) (0,626 ; 1,100)	1,436 (0,094) (1,257 ; 1,614)	-0,534 (0,152) (-0,827 ; -0,229)	-0,484 (0,109) (-0,690 ; -0,275)	0,484 (0,022) (0,448 ; 0,531)	0,989 (0,070) (0,857 ; 1,129)	0,367 (0,051) (0,269 ; 0,462)	0,102 (0,042) (0,037 ; 0,197)
c	63	0,868 (0,119) (0,631 ; 1,093)	1,437 (0,094) (1,246 ; 1,615)	-0,534 (0,157) (-0,848 ; -0,235)	-0,485 (0,108) (-0,688 ; -0,274)	0,485 (0,021) (0,450 ; 0,532)	0,990 (0,070) (0,860 ; 1,131)	0,371 (0,050) (0,277 ; 0,471)	0,100 (0,042) (0,034 ; 0,196)
d	210	0,871 (0,118) (0,636 ; 1,090)	1,440 (0,093) (1,250 ; 1,609)	-0,530 (0,151) (-0,825 ; -0,240)	-0,481 (0,108) (-0,699 ; -0,277)	0,486 (0,022) (0,452 ; 0,534)	0,993 (0,069) (0,865 ; 1,136)	0,370 (0,051) (0,269 ; 0,468)	0,102 (0,045) (0,036 ; 0,212)
e	8 e 63	0,765 (0,118) (0,522 ; 0,991)	1,339 (0,095) (1,147 ; 1,520)	-0,454 (0,156) (-0,752 ; -0,143)	-0,416 (0,112) (-0,631 ; -0,203)	0,477 (0,021) (0,443 ; 0,523)	0,955 (0,065) (0,836 ; 1,090)	0,343 (0,051) (0,241 ; 0,442)	0,112 (0,046) (0,042 ; 0,220)
f	8 e 210	0,774 (0,116) (0,541 ; 0,989)	1,350 (0,095) (1,150 ; 1,523)	-0,460 (0,158) (-0,771 ; -0,151)	-0,423 (0,111) (-0,636 ; -0,203)	0,478 (0,021) (0,444 ; 0,526)	0,959 (0,067) (0,830 ; 1,096)	0,344 (0,052) (0,240 ; 0,445)	0,113 (0,049) (0,040 ; 0,236)
g	63 e 210	0,768 (0,120) (0,537 ; 1,006)	1,347 (0,095) (1,154 ; 1,527)	-0,452 (0,156) (-0,760 ; -0,156)	-0,417 (0,111) (-0,630 ; -0,201)	0,478 (0,021) (0,444 ; 0,522)	0,960 (0,067) (0,832 ; 1,096)	0,349 (0,051) (0,246 ; 0,450)	0,110 (0,044) (0,042 ; 0,210)
h	8, 63 e 210	0,721 (0,120) (0,477 ; 0,945)	1,296 (0,096) (1,098 ; 1,479)	-0,414 (0,160) (-0,736 ; -0,100)	-0,385 (0,112) (-0,596 ; -0,164)	0,474 (0,021) (0,440 ; 0,522)	0,938 (0,063) (0,820 ; 1,067)	0,333 (0,051) (0,231 ; 0,430)	0,118 (0,048) (0,045 ; 0,231)

A Tabela 12 evidencia que o conjunto de dados (a), sem nenhum caso perturbado, teve o melhor ajuste de acordo com todos os critérios Bayesianos dentre todos os diferentes casos de conjuntos de dados.

Tabela 12 – Critérios Bayesianos ajustando o modelo de sobrevivência PVF bivariado Exp. Gen. para cada conjunto de dados simulados.

Conjunto de dados	DIC	EAIC	EBIC	LPML
a	**-415,186**	**-415,261**	**-385,631**	**204,052**
b	-358,304	-355,538	-325,908	172,636
c	-356,638	-356,991	-327,361	174,303
d	-356,449	-354,916	-325,286	172,813
e	-286,195	-290,417	-260,787	139,810
f	-283,414	-288,536	-258,906	139,195
g	-290,800	-293,712	-264,082	141,312
h	-245,767	-253,627	-223,997	121,122

Consideramos as amostras da distribuição *a posteriori* dos parâmetros do modelo de sobrevivência PVF bivarido para obter uma estimativa das três medidas de divergência apresentadas. Os resultados da Tabela 13 mostram que todos os casos que não foram perturbados tiveram pequenas medidas de divergência, mesmo quando fazem parte do conjunto de dados com algum caso perturbado, entretanto evidenciam que as três medidas aumentam e detectam quando algum caso é influente.

Tabela 13 – Medidas de divergência para os dados simulados com marginais Exp. Gen..

Conjunto de dados	Casos perturbados	K-L	J	L-1
a	8	0,118	0,244	0,193
	63	0,03	0,060	0,097
	210	0,002	0,004	0,024
b	**8**	**0,682**	**1,397**	**0,447**
	63	0,03	0,061	0,100
	210	0,002	0,004	0,025
c	8	0,119	0,247	0,193
	63	**0,623**	**1,274**	**0,424**
	210	0,002	0,004	0,025
d	8	0,123	0,257	0,198
	63	0,028	0,057	0,095
	210	**0,545**	**1,176**	**0,404**
e	**8**	**0,658**	**1,350**	**0,442**
	63	**0,655**	**1,331**	**0,441**
	210	0,002	0,004	0,024
f	**8**	**0,679**	**1,477**	**0,441**
	63	0,028	0,056	0,094
	210	**0,586**	**1,255**	**0,412**
g	8	0,137	0,288	0,203
	63	**0,61**	**1,322**	**0,421**
	210	**0,591**	**1,28**	**0,415**
h	**8**	**0,637**	**1,377**	**0,438**
	63	**0,610**	**1,321**	**0,428**
	210	**0,564**	**1,220**	**0,413**

A Figura 18 mostra os gráficos de índices das três medidas de divergência para os casos (a) e (g). Podemos ver que para os dados originais, isto é, sem nenhuma perturbação, as três medidas de divergência não detectaram os pontos influentes. E, para caso em que houve perturbação, as três medidas de divergência apontaram corretamente os pontos perturbados.

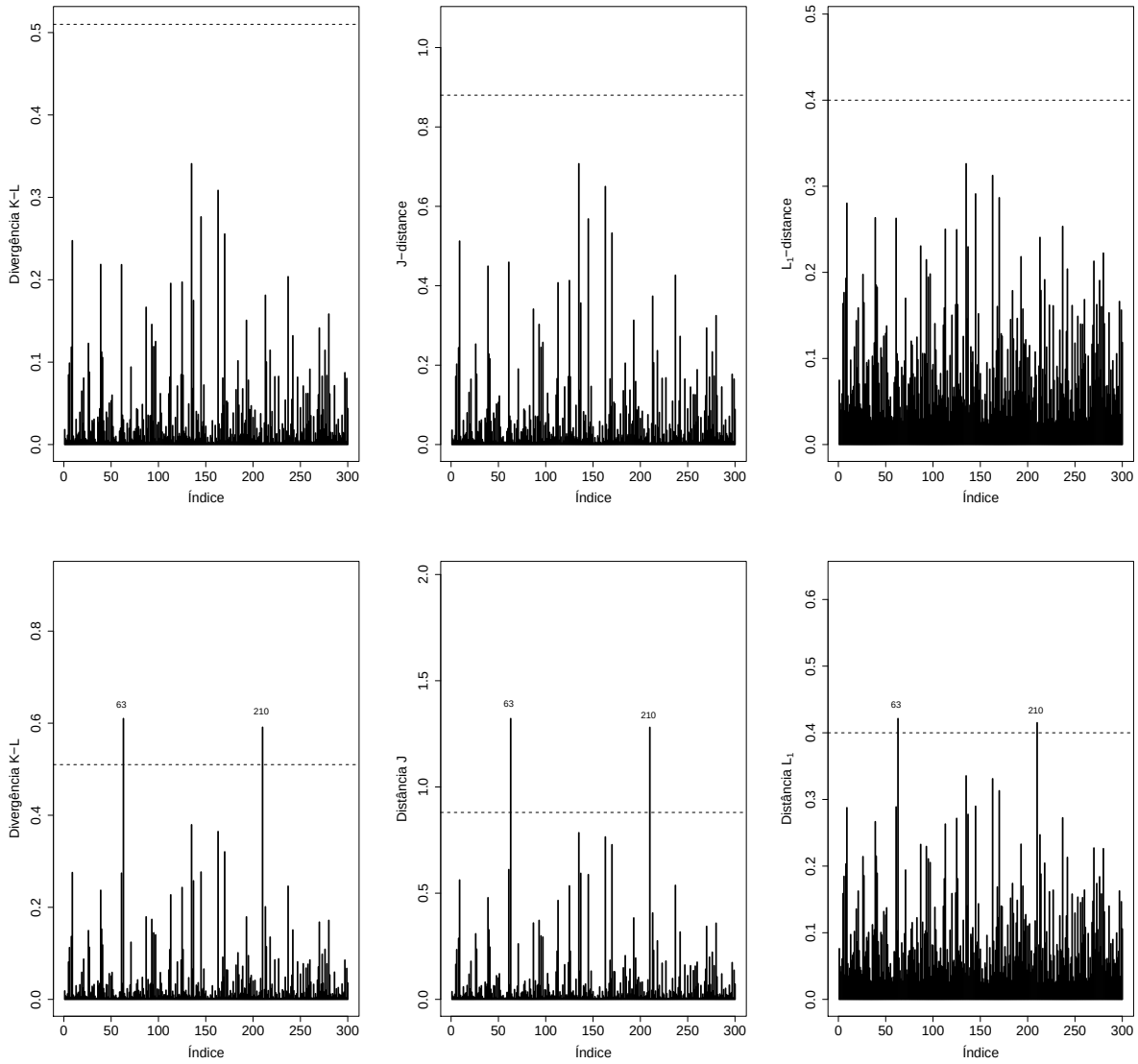

Figura 18 – Gráfico de índices das medidas de divergência para o conjunto de dados (a) (superior) e (g) (inferior). para o modelo com marginais Exp. Gen. (Fonte: Elaborado pelo autor).

4.1.2 Estudo de simulação para casos com censura

Neste estudo de simulação, geramos os conjuntos de dados assumindo a presença de dados censurados, $(15\%, 10\%)$, para três diferentes tamanhos de amostras $N = 100, 200$ e 300. Para cada caso, geramos 200 conjuntos Monte Carlo (amostras) de dados.

As seguintes distribuições *a priori* independentes foram consideradas para o amostrador de Gibbs: $\alpha_j \sim Gama(0,1;0,1)$, $\beta_{ij} \sim N(0,10^3)$, $i = 0,1$ e $j = 1,2$. Assumimos $\phi \sim Beta(1;1)$ e $\eta \sim Gama(0,1;0,1)$ para os parâmetros da cópula.

Para o modelo com marginais Exponencial Generalizada foram considerados os seguintes valores para os parâmetros: $\alpha_1 = 0,5,\ \beta_{01} = 1,\ \beta_{11} = -0,5,\ \alpha_2 = 1,\ \beta_{02} = 1,5,\ \beta_{12} = -0,5,\ \eta = 0,1$ e $\phi = 0,4$.

Tabela 14 – Média MC, percentual de cobertura (PC) e EQM das estimativas dos parâmetros ajustando o modelo de PVF bivariado com marginais Exponencial Generalizada para as diferentes configurações de tamanhos de amostras simuladas para o caso com censura.

Parâmetro	Valor real	N = 100			N = 200			N = 300		
		Estimativa	PC	EQM	Estimativa	PC	EQM	Estimativa	PC	EQM
β_{01}	1	1,143	0,86	0,076	1,171	0,86	0,051	1,127	0,90	0,034
β_{02}	1,5	1,532	0,94	0,030	1,524	0,95	0,013	1,521	0,96	0,008
β_{11}	-0,5	-0,472	0,96	0,097	-0,464	0,98	0,044	-0,493	0,94	0,031
β_{12}	-0,5	-0,494	0,98	0,044	-0,505	0,96	0,021	-0,500	0,96	0,014
α_1	0,5	0,572	0,79	0,010	0,582	0,82	0,009	0,558	0,86	0,005
α_2	1	1,059	0,96	0,018	1,043	0,97	0,008	1,036	0,96	0,004
ϕ	0,4	0,438	0,92	0,012	0,432	0,94	0,007	0,419	0,94	0,003
η	0,1	0,140	0,90	0,014	0,113	0,95	0,005	0,107	0,96	0,002

Para cada conjunto de dados gerados, consideramos duas cadeias de tamanho 30.000. Para eliminar o efeito dos valores iniciais, foram desconsideradas as primeiras 10.000 iterações. De forma análoga ao caso anterior, para evitar problemas de autocorrelação, considerou-se um salto de tamanho 10, obtendo uma amostra efetiva de tamanho 4.000 sobre a qual a inferência *a posteriori* é baseada. Para cada amostra, a média e o desvio-padrão *a posteriori* dos parâmetros são obtidos.

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 ($\leq 1,01$).

Na Tabela 14 temos a média MC das estimativas dos parâmetros ajustando a cópula PVF com distribuição marginal Exponencial Generalizada para o caso considerando censura, $(15\%, 10\%)$, e três tamanhos de amostras simuladas ($N = 100$, $N = 200$ e $N = 300$). As estimativas dos parâmetros estão próximas do verdadeiro valor e, com o aumento do tamanho da amostra, chegamos a resultados mais satisfatórios. Avaliando o percentual de cobertura, saímos de uma média de $0,91$ para o caso com $N = 100$, para $0,95$ no caso com $N = 300$. Quando avaliamos o erro quadrático médio, também temos que o aumento da amostra faz com que as estimativas sejam melhores neste aspecto.

4.1.2.1 Diagnóstico de Observações Influentes

Para examinar o desempenho da medida de diagnóstico, geramos uma amostra de tamanho 300 para o modelo de cópula PVF bivariado com marginais Exponencial Generalizada, considerando os seguintes valores para os parâmetros: $\beta_{01} = 1,0$, $\beta_{11} = -0,5$, $\alpha_1 = 0,5$, $\beta_{02} = 1,5$, $\beta_{12} = -0,5$, $\alpha_2 = 1$, $\eta = 0,1$ e $\phi = 0,4$.

Selecionamos os casos $50, 116$ e 225 para perturbação, em que são observados ambos tempos de falha para os três casos. Para criar observações artificialmente influentes no conjunto de dados, escolhemos um, dois ou três destes casos selecionados. Para cada caso, perturbamos os dois tempos de vida da seguinte forma: $\widetilde{t}_{jb} = t_{jb} + 5D_t$, $j = 1,2$ e $b \in \{50, 116, 225\}$, 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 na Seção 3.1.

Na Tabela 15 vemos que as inferências *a posteriori* são sensíveis a perturbação do(s) caso(s) selecionado(s), em que o conjunto de dados (a) denota os dados originais simulados sem nenhuma perturbação, e os casos (b) a (h) denotam os conjuntos de dados com pelo menos algum caso perturbado. Podemos observar também que os maiores impactos são dados marginalmente, em que há as maiores diferenças das estimativas dos conjuntos de dados sem perturbação para os com pelo menos alguma obervação perturbada.

Tabela 15 – Média, desvio padrão (DP) e intervalo de credibilidade para os parâmetros do modelo de sobrevivência PVF bivariado Exp. Gen. para cada conjunto de dados simulados com censura

		β_{01}	β_{02}	β_{11}	β_{12}	α_1	α_2	ϕ	η
Dados	Perturbados	Média (DP) (IC[0.025 ; 0.975])	Média (DP) (IC[0.925 ; 0.975])	Média (DP) (IC[0.025 ; 0.975])	Média (DP) (IC[0.025 ; 0.975])	Média (DP) (IC[0.025 ; 0.975])	Média (DP) (IC[0.025 ; 0.975])	Média (DP) (IC[0.025 ; 0.975])	Média (DP) (IC[0.025 ; 0.975])
a	Nenhum	0,932 (0,129) (0,663 ; 1,174)	1,377 (0,096) (1,181 ; 1,559)	-0,774 (0,183) (-1,139 ; -0,424)	-0,331 (0,113) (-0,554 ; -0,114)	0,460 (0,028) (0,407 ; 0,517)	1,128 (0,080) (0,981 ; 1,290)	0,509 (0,053) (0,402 ; 0,611)	0,045 (0,033) (0,001 ; 0,131)
b	50	0,826 (0,124) (0,582 ; 1,065)	1,304 (0,095) (1,114 ; 1,490)	-0,684 (0,180) (-1,024 ; -0,342)	-0,274 (0,114) (-0,500 ; -0,062)	0,452 (0,028) (0,401 ; 0,507)	1,098 (0,077) (0,954 ; 1,254)	0,482 (0,055) (0,370 ; 0,589)	0,054 (0,038) (0,004 ; 0,152)
c	116	0,838 (0,123) (0,588 ; 1,072)	1,317 (0,097) (1,116 ; 1,495)	-0,696 (0,179) (-1,046 ; -0,358)	-0,284 (0,116) (-0,508 ; -0,056)	0,453 (0,028) (0,401 ; 0,510)	1,104 (0,076) (0,966 ; 1,259)	0,479 (0,054) (0,368 ; 0,582)	0,056 (0,039) (0,004 ; 0,152)
d	225	0,821 (0,123) (0,558 ; 1,051)	1,302 (0,096) (1,106 ; 1,490)	-0,678 (0,177) (-1,126 ; -0,329)	-0,270 (0,112) (-0,492 ; -0,056)	0,452 (0,027) (0,399 ; 0,508)	1,101 (0,078) (0,954 ; 1,272)	0,477 (0,054) (0,365 ; 0,576)	0,056 (0,150) (0,007 ; 0,150)
e	50 e 116	0,726 (0,120) (0,491 ; 0,946)	1,237 (0,097) (1,036 ; 1,416)	-0,601 (0,184) (-0,981 ; -0,266)	-0,225 (0,114) (-0,441 ; -0,001)	0,445 (0,027) (0,393 ; 0,498)	1,074 (0,073) (0,938 ; 1,220)	0,449 (0,057) (0,327 ; 0,552)	0,068 (0,044) (0,008 ; 0,183)
f	50 e 225	0,730 (0,122) (0,492 ; 0,959)	1,240 (0,097) (1,042 ; 1,418)	-0,604 (0,175) (-0,948 ; -0,277)	-0,228 (0,116) (-0,446 ; -0,004)	0,445 (0,027) (0,396 ; 0,499)	1,078 (0,074) (0,940 ; 1,227)	0,447 (0,058) (0,319 ; 0,552)	0,069 (0,045) (0,011 ; 0,186)
g	116 e 225	0,746 (0,123) (0,508 ; 0,979)	1,240 (0,097) (1,044 ; 1,425)	-0,618 (0,178) (-0,969 ; -0,267)	-0,228 (0,118) (-0,454 ; -0,003)	0,447 (0,027) (0,369 ; 0,501)	1,076 (0,074) (0,935 ; 1,227)	0,447 (0,058) (0,331 ; 0,553)	0,068 (0,045) (0,011 ; 0,178)
h	50, 116 e 225	0,641 (0,119) (0,400 ; 0,867)	1,166 (0,095) (0,972 ; 1,351)	-0,529 (0,176) (-0,884 ; -0,190)	-0,174 (0,114) (-0,389 ; -0,048)	0,439 (0,027) (0,387 ; 0,493)	1,048 (0,071) (0,911 ; 1,197)	0,414 (0,064) (0,275 ; 0,528)	0,085 (0,057) (0,016 ; 0,234)

A Tabela 16 evidencia que o conjunto de dados (a), sem nenhum caso perturbado, teve o melhor ajuste de acordo com todos os critérios Bayesianos dentre todos os diferentes conjuntos de dados.

Tabela 16 – Critérios Bayesianos ajustando o modelo de sobrevivência PVF bivariado Exp. Gen. para cada conjunto de dados simulados com censura

Conjunto de dados	DIC	EAIC	EBIC	LPML
a	**-577,241**	**-530,804**	**-501,174**	**254,674**
b	-493,266	-474,346	-444,716	233,250
c	-516,447	-485,538	-455,908	235,101
d	-476,564	-465,072	-435,442	230,402
e	-449,199	-436,923	-407,292	215,289
f	-435,578	-430,218	-400,588	212,986
g	-439,598	-432,449	-402,818	214,128
h	-401,070	-397,429	-367,799	195,410

Consideramos as amostras da distribuição *a posteriori* dos parâmetros do modelo de sobrevivência PVF bivariado para obter uma estimativa das três medidas de divergência apresentadas. Os resultados da Tabela 17 mostram que todos os casos que não foram perturbados tiveram pequenas medidas de divergência, mesmo quando fazem parte do conjunto de dados com algum caso perturbado, entretanto evidenciam que as três medidas aumentam e detectam quando algum caso é influente, principalmente quando o conjunto de dados é composto por três observações perturbadas. Em negrito é destacado o ponto perturbado e suas medidas de divergência.

Tabela 17 – Medidas de divergência para os dados simulados com censura com marginais Exp. Gen..

Dados	Perturbados	d_{K-L}	d_J	d_{L-1}
a	50	0,306	0,596	0,294
	116	0,180	0,326	0,208
	225	0,265	0,504	0,26
b	**50**	**1,025**	**2,175**	**0,542**
	116	0,158	0,313	0,215
	225	0,243	0,498	0,269
c	50	0,302	0,607	0,291
	116	**0,636**	**1,314**	**0,424**
	225	0,269	0,525	0,269
d	50	0,245	0,501	0,278
	116	0,127	0,252	0,197
	225	**0,611**	**1,229**	**0,430**
e	**50**	**0,963**	**2,407**	**0,525**
	116	**0,633**	**1,310**	**0,431**
	225	0,261	0,545	0,281
f	**50**	**0,812**	**1,718**	**0,488**
	116	0,150	0,311	0,218
	225	**0,684**	**1,487**	**0,451**
g	50	0,336	0,760	0,324
	116	**0,571**	**1,204**	**0,415**
	225	**0,688**	**1,517**	**0,461**
h	**50**	**0,936**	**2,252**	**0,529**
	116	**0,980**	**2,832**	**0,547**
	225	**0,866**	**2,089**	**0,511**

A Figura 19 mostra os gráficos de índices das três medidas de divergência para os casos (b) e (f). Claramente, podemos ver que que as três medidas de divergência detectaram os pontos influentes. Ainda na Figura 19, podemos observar também algumas medidas que não foram perturbadas com alto valor de divergência, mas que não foram erroneamente apontadas.

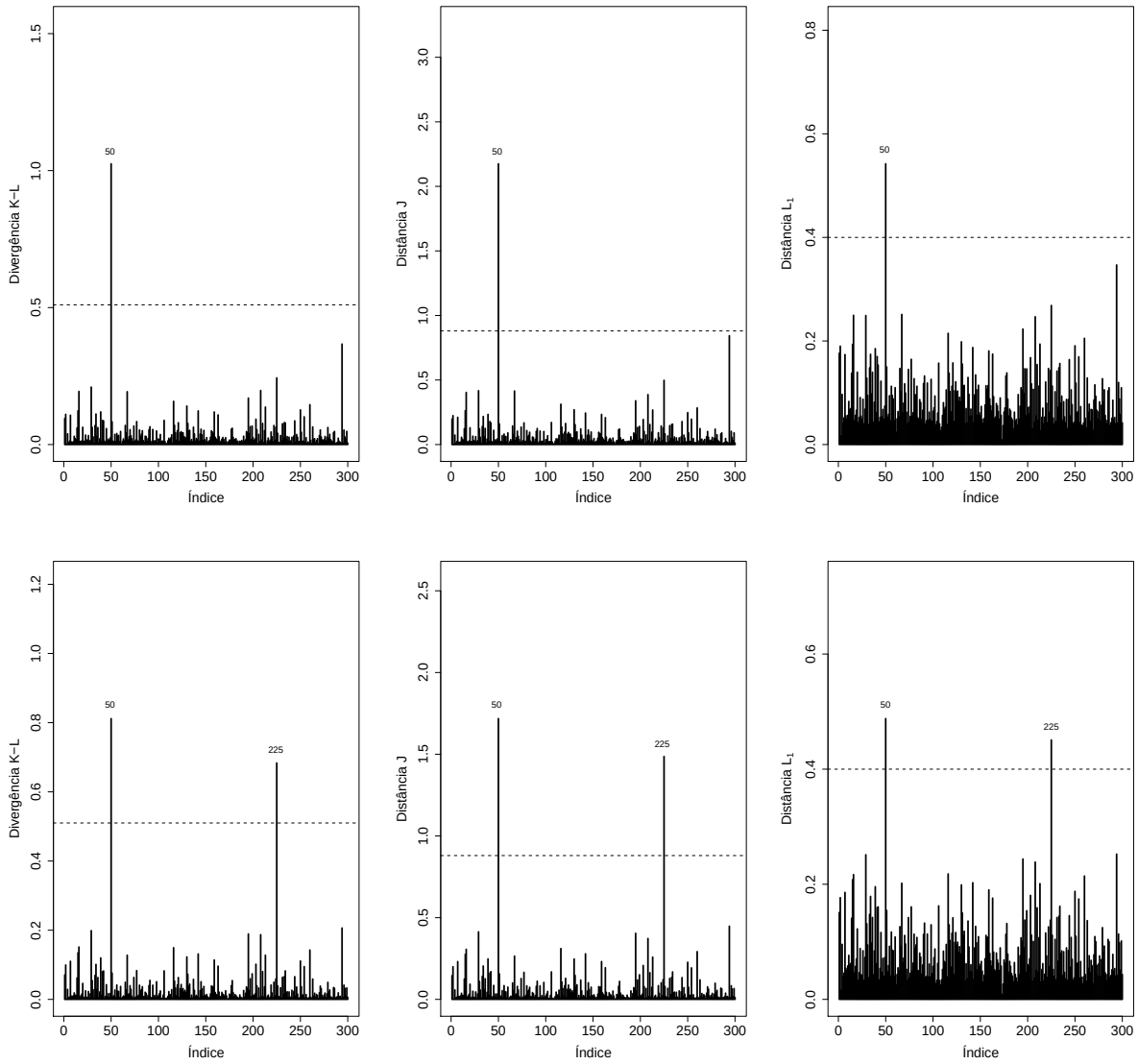

Figura 19 – Gráfico de índices das medidas de divergência para o conjunto de dados (b) (superior) e (f) (inferior) com marginais Exp. Gen. (Fonte: Elaborado pelo autor).

4.2 Aplicação a Dados Reais

Nesta seção, aplicamos os modelos propostos para analisar dados sobre o tempo de apendicectomia para gêmeos adultos no Australian NH & MRC Twin Registry dado por Duffy et al. (1990).

4.2.1 Dados Reais de Apendicectomia para gêmeos adultos

O Australian Twin Registry é um registro voluntário de gêmeos que existe desde 1981 e foi formado pela fusão de vários registros baseados em instituições - principalmente os do Dr. John Mathews baseado na Universidade de Melbourne, e dos Drs Nicholas Martin e John Gibson, da Universidade Nacional Australiana.

Este estudo foi realizado para investigar se a força da dependência entre pares gêmeos quanto ao risco de aparecimento de várias doenças, incluindo apendicite aguda, é diferente para gêmeos monozigóticos (MZ) e dizigóticos (DZ).

Os gêmeos podem ser de dois tipos: monozigóticos e dizigóticos. Os gêmeos monozigóticos são também chamados de idênticos ou univitelinos. Eles se originam de um único zigoto (célula-ovo), ou seja, um único óvulo fecundado por um único espermatozóide. Do ponto de vista genético, esses gêmeos são idênticos e, obviamente, pertencem sempre ao mesmo sexo. Cerca de 25% dos gêmeos são monozigóticos. Os gêmeos dizigóticos ou bivitelinos são o resultado de fertilização de dois óvulos e dois espermatozóides. Eles compartilham até 50% de informação genética, podem ou não ser do mesmo sexo e ter ou não o mesmo fator sanguíneo. E não se assemelham mais do que dois irmãos com a mesma idade. (Departamento de Embriologia, UFMG).

Como seria de esperar que qualquer efeito potencial de um ambiente compartilhado fosse muito semelhante para os gêmeos MZ e DZ, uma dependência mais forte nos riscos de apendicite entre os membros de dois pares de MZ seria indicativo de um efeito genético no risco de apendicite aguda e evidências do papel da hereditariedade no início da apendicite.

O banco de dados é composto por 1798 gêmeos MZ (1231 mulheres e 567 homens), 1098 gêmeos DZ do mesmo sexo (748 mulheres e 350 homens) e 912 gêmeos DZ de sexo diferentes, totalizando 3808 pares de gêmeos. Os sujeitos que não foram submetidos a apendicectomia antes da pesquisa foram censurados (aproximadamente 73% para ambos os tipos de zigotos).

Com uma breve análise desse conjunto de dados, podemos ver na Figura 20 pela curva de Kaplan-Meier a existência de indivíduos que poderiam ser ditos "curados".

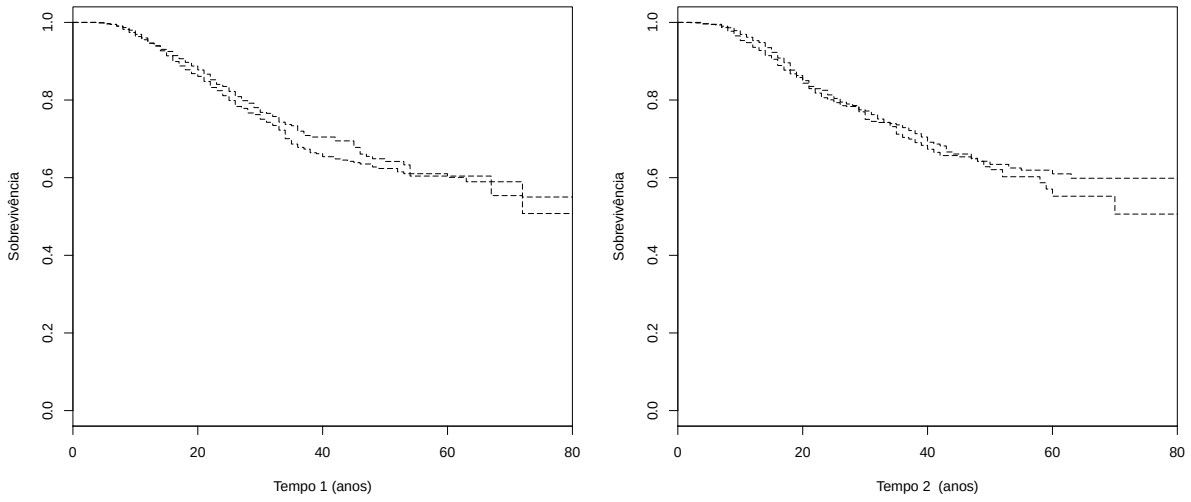

Figura 20 – Estimativas de Kaplan-Meier da função de sobrevivência para os dados de apendicectomia para gêmeos adultos. (Fonte: Elaborado pelo autor).

Tabela 18 – Média *a posteriori*, desvio padrão (DP) e intervalo de credibilidade de 95% para os parâmetros do modelo PVF bivariado.

	Parâmetro	Weibull			Exponencial Generalizada		
		Média	DP	IC[0.025 ; 0.975]	Média	DP	IC[0.025 ; 0.975]
Tempo 1	α_1	1,179	0,094	[1,600 ; 1,965]	5,339	0,058	[5,210 ; 5,436]
	β_{01}	-5,985	0,350	[-6,666 ; -5,298]	-2.350	0,058	[-2,461 ; -2234]
	β_{11}	0,080	0,141	[-0,216 ; 0,329]	0,061	0,071	[-0,080 ; 0,197]
Tempo 2	α_2	1,751	0,091	[1,570 ; 1,923]	5,365	0,066	[5,234 ; 5,497]
	β_{02}	-5,854	0,344	[-6,584 ; -5,217]	-2,341	0,057	[-2,456 ; -2,231]
	β_{12}	0,023	0,140	[-0,262 ; 0,278]	0,031	0,071	[-0,111 ; 0,165]
Cópula	ϕ	0,544	0,035	[0,483 ; 0,617]	0,643	0,059	[0,517 ; 0,744]
	η	0,000	0,001	[0,000 ; 0,003]	0,028	0,050	[0,000 ; 0,126]

Como o objetivo do estudo é avaliar evidências de possível hereditariedade no início da apendicite, e não estamos trabalhando com fração de cura, vamos seguir o que foi feito em Romeo (2017) e retirar uma amostra desse conjunto de dados. Nossa amostra será uma amostra aleatória de 560 gêmeos com ambos os sexo feminino (366 DZ e 194 MZ), tal que o percentual de censura fique em torno de 10%, valor esse que está próximo dos estudos de simulação que realizamos.

O tipo de zigosidade (DZ ou MZ) foi considerada como covariável. Considerou-se T_1 o tempo (em anos) até o primeiro gêmeo ser submetido a cirurgia de apendicite e T_2 o tempo até o segundo gêmeo ser submetido a cirurgia de apendicite.

Ajustamos o modelo PVF bivariado com ambas as distribuições marginais Weibull ou ambas Exponencial Generalizada, para os dois casos consideramos duas cadeias de tamanho 70.000 para cada parâmetro, desconsiderando as primeiras 10.000 iterações para eliminar o efeito dos valores iniciais e, para evitar problemas de autocorrelação, foi considerado um salto de tamanho 20, obtendo uma amostra final de tamanho 3.000 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).

Para todos os parâmetros do modelo especificamos distribuições a priori não informativas. Sendo que, para os parâmetros da cópula foi especificado $\phi \sim Beta(1,1)$ e $\eta \sim Gamma(0.1,0.1)$. Para as marginais foi especificado $\beta_{ij} \sim N(0,10^3)$ e $\alpha_j \sim Gamma(0.1,0.1)$, em que $i = 0,1$ e $j = 1,2$.

Na Tabela 18 são apresentadas as médias *a posteriori* para os parâmetros PVF bivariado considerando ambas as distribuições estudadas.

Tabela 19 – Valores dos critérios para comparação dos modelos.

Marginais	Critérios Bayesianos			
	DIC	EAIC	EBIC	LPML
Weibull	2481,220	2731,919	2760,090	-1431,084
Exp. Gen.	**2325,470**	**2662,687**	**2690,859**	**-1415,220**

Para comparar e avaliar qual das distribuições teve o melhor ajuste, iremos novamente nos basear nos quatro critérios que discutimos ao longo do trabalho.

Na Tabela 19 destacamos que os quatro critérios apontam que o modelo que obteve o melhor ajuste foi o modelo considerando ambas as marginais Exponencial Generalizada.

Podemos ver, pela Figura 21, que nenhum ponto influente foi detectado pelos gráficos de índices considerando o modelo PVF com distribuição Exponencial Generalizada.

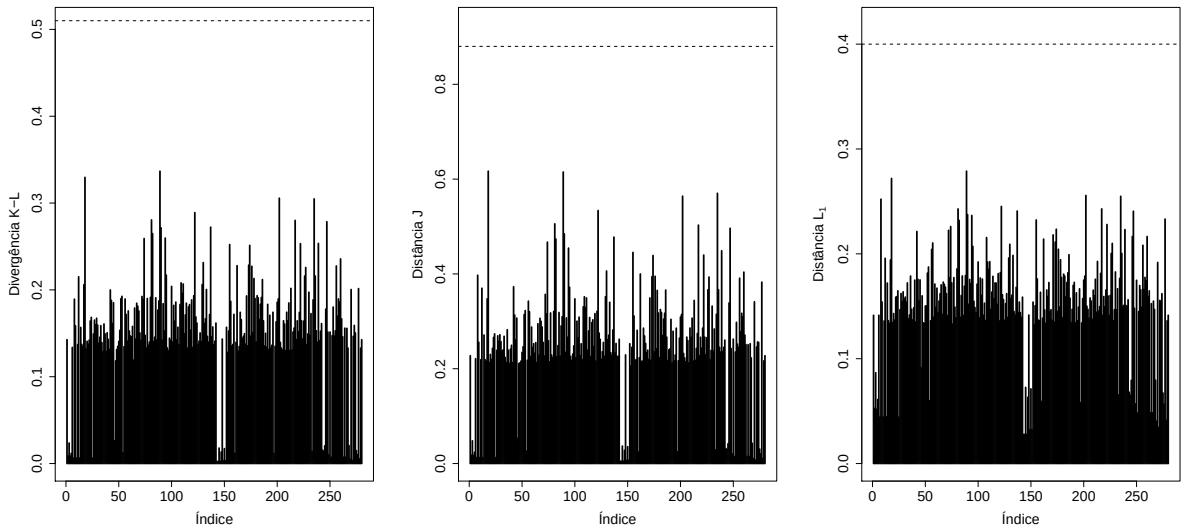

Figura 21 – Gráfico de índices das medidas de divergência para o conjunto de dados reais considerando a distribuição Exp. Gen. (Fonte: Elaborado pelo autor).

A Figura 22 mostra as curvas de Kaplan-Meier para as variáveis T_1 e T_2 dicotomizadas pelo tipo de zigosidade do paciente, juntamente com os ajustes do modelo de sobrevivência bivariado baseado na cópula PVF com distribuições marginais Exponencial Generalizada.

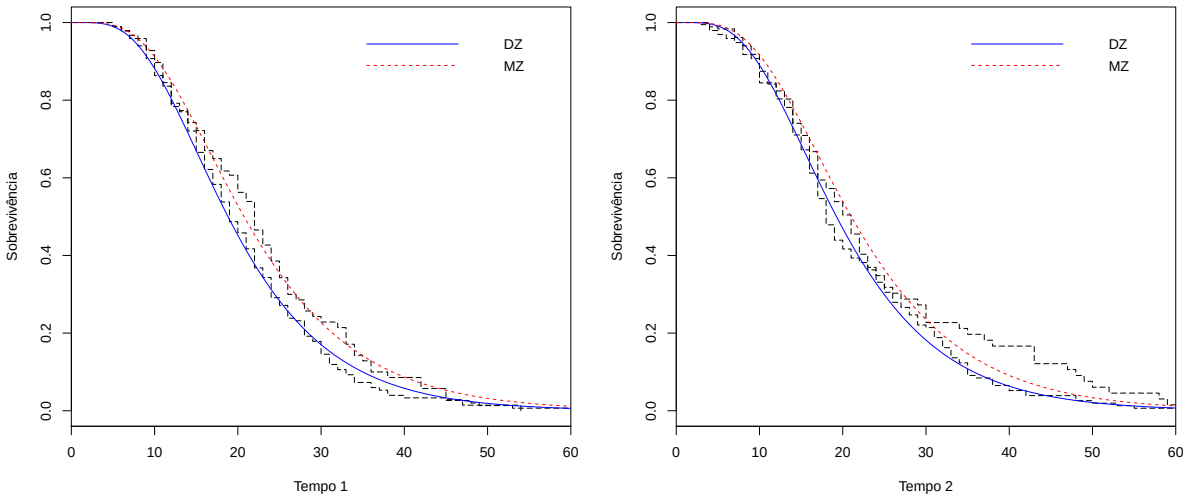

Figura 22 – Curvas de Kaplan-Meier e curvas de sobrevivências Expg. Gen. estimadas para o conjunto de dados reais. (Fonte: Elaborado pelo autor).

Podemos observar o bom ajuste do modelo bivariado PVF, e que ele se molda bem às curvas das estimativas de Kaplan-Meier, devido à flexibilidade que a inclusão de um parâmetro a mais de dependência ao modelo traz. Observamos também, de acordo com as curvas de Kaplan-Meier, que os tempos de sobrevivência para as variáveis T_1 e T_2 são muito próximas, indicando não haver uma relação entre hereditariedade e tipo de zigosidade com a apendicite.

CAPÍTULO

5

CONSIDERAÇÕES FINAIS E PERSPECTIVAS FUTURAS

Neste trabalho foram apresentados alguns conceitos referentes à Análise de Sobrevivência e às funções cópulas, em especial as cópulas Arquimedianas, como eficientes ferramentas para modelar dados de sobrevivência.

Quanto ao procedimento inferencial, foi realizado sob uma abordagem Bayesiana assumindo ausência de informação *a priori*. Foi feito todo um estudo de simulação com o objetivo de mostrar o bom comportamento das estimativas Bayesianas com base no Erro Quadrático Médio. 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, para ambos os modelos de cópulas. O código utilizado pode ser encontrado no Apêndice A.

A comparação de modelos foi realizada via critérios Bayesianos EAIC, EBIC, DIC e LPML. Simulamos amostras a partir dos modelos de PVF, que tem como casos particulares os modelos de Clayton, Gumbel e Gaussiana Inversa, com marginais Weibull e Exponencial Generalizada.

Além disso, aplicamos o método Bayesiano de análise de influência de deleção de casos baseado na divergência ψ, cujo objetivo é detectar possível(is) observação(ões) influente(s) nos dados analisados. Para isso, foram assumidas três particulares escolhas para a função ψ nas quais resultaram a divergência de Kullback-Leibler (K-L), a distância J e a distância variacional ou norma L_1. Para uma amostra simulada do modelo, perturbamos uma, duas ou três observações e, a partir disso, conseguimos averiguar que as três medidas de divergência detectaram os pontos perturbados. Todos os casos considerando a distribuição Weibull, além dos que já foram apresentados no decorrer do texto, estão disponíveis nos Anexos A e B.

Por fim, aplicamos os modelos propostos a dois conjuntos de dados reais.

Como continuidade do trabalho, pretendemos flexibilizar ainda mais o modelo proposto

utilizando distribuições marginais mais flexíveis, assumir modelos com proporção de cura e incluir mais do que uma covariável ao modelo final. Além disso, também é possível a comparação com outras cópulas, e explorar aspectos da dependência caudal.

REFERÊNCIAS

AARSET, M. The null distribution for a test of constante versus “bathtub” failure rate. **Scandinavian Journal of Statistics**, v. 12, n. 1, p. 55-61, 1985.

ACHCAR, J. A.; BOLETA, J. Distribuição Exponencial Generalizada bivariada derivada de funções cópulas: Uma aplicação a dados de câncer gástrico. **Revista Brasileira de Biometria**, v. 30, n. 4, p. 401-414, 2012.

ACHCAR, J. A.; LOUZADA F. A Bayesian approach for accelerated life tests considering the Weibull distribution. **Computational Statistics Quarterly**, v. 7, p. 355-355, 1992.

ACHCAR, J. A.; MOALA, F. A. Use of copula funcions for the reliability of series systems. **International Journal of Quality and Reliability Management**, v. 32, p. 617-634, 2015.

ACHCAR, J. A.; MOALA, F. A.; TARUMOTO, M. H.; COLLADELO, L. F. A bivariate Generalized Exponential distribution derived from copula functions in the presence of censored data and covariates. **Pesquisa Operacional** (Online), v. 35, p. 165-186, 2015.

ALI, M. M.; MIKHAIL, N. N.; HAQ, M. S. A class of bivariate distributions including the bivariate Logistic. **Journal of Multivariate Analysis**, v. 8, p. 405-412, 1978.

BOLETA, J. **Distribuição Exponencial Generalizada: Uma Análise Bayesiana Aplicada a Dados de Câncer**. Dissertação de mestrado. Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo, 94 pages, 2012.

BROOKS, S. P. Discussion on the paper by Spiegelhalter, Best, Carlin, and van der Linde, v. 64, p. 616-618, 2002.

CANCHO, V.; BOLFARINE, H.; ACHCAR, J. A. A Bayesian analysis for the exponentiated-Weibull distribution. **Journal Applied Statistical Science**, v. 8, n. 4, p. 227-42, 1999.

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**, v. 140, p. 2486-2496, 2010.

CANCHO, V. G.; SUZUKI, A. K.; BARRIGA, G. D. C.; LOUZADA, F. 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**. To appear, 2016.

CARLIN, B. P.; LOUIS, T. A. **Bayes and Empirical Bayes Methods for Data Analysis**. ISBN: 9781584881704, 440 pages, 2ºEdição, 2001.

CARVALHO, M. S.; ANDREOZZI, V. L.; CODEÇO, C. T.; CAMPOS, D. P.; BARBOSA, M. T. S.; SHIMAKURA, S. E. **Análise de Sobrevivência: Teoria e Aplicações em Saúde**. ISBN: 9788575412169, 432 pages, 2º Edição, 2011.

CÉSAR, K. A. **Análise Estatística de Sobrevivência: Um Estudo com Pacientes com Câncer de Mama**. Monografia (Graduação), Universidade Católica de Brasília, 2005.

CHARPENTIER, A.; FERMANIAN, J.; SCAILLET, O. **The estimation of copulas : theory and practice**. Jörn Rank. Copulas: from theory to application in finance. London: Risk Books., p. 35-64, 2007.

CHEN, M. H.; IBRAHIM, J. G.; SINHA, D. Bayesian inference for multivariate survival data with a cure fraction. **Journal of Multivariate Analysis**, v. 80, p. 101-126, 2002.

CHERUBINI, U.; LUCIANO, E.; VECCHIATO, W. **Copula Methods in Finance**. John Wiley & Sons, Ltd. ISBN: 9780470863442, 310 pages, 1ºEdição, 2004.

CHERUBINI, U.; MULINACCI, S.; GOBBI, F.; ROMAGNOLI, S. **Dynamic Copula Methods in Finance**. Wiley finance. ISBN: 9780470683071, 288 pages, 1ºEdição, 2011.

CHO, H.; IBRAHIM, J. G.; SINHA, D.; SHU, H. Bayesian case influence diagnostics for survival models. **Biometrics**, v. 65, p. 116-124, 2009.

CLAYTON, G. A model for association in bivariate life tables and its application in epidemiological studies in familial tendency in chronic disease incidence. **Biometrika**, v. 65, p. 141-151, 1978.

COOK, R. D.; WEISBERG, S. **Residuals and Influence in Regression**. ISBN: 0412242800, 229 pages, 1ºEdição, 1982.

COLOSIMO, E. A. & GIOLO, S. R. **Análise de Sobrevivência Aplicada**. ISBN: 9788521203841, 367 pages, 1º Edição, 2006.

COWLESS, M. K.; CARLIN, B. P. Markov chain Monte Carlo convergence diagnostics: a comparative review. **Journal of the American Statistical Association**, v. 91, p. 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**, v. 71, p. 584, 2016.

DEY, D.; BIRMIWAL, L. Robust Bayesian analysis using divergence measures. **Statistics and Probability Letters**, v. 20, p. 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.

FISCHER, N. I. Copulas. **In: Encyclopedia of Statistical Sciences**, Update v. 1, p. 159-163. John Wiley Sons, New York, 1997.

FLEMING, T. R.; HARRINGTON, D. P. **Counting Processes and Survival Analysis**. Wiley-Interscience Paperback Series. ISBN: 9781118150665, 448 pages, 1ºEdição, 2011.

FRANK, M. J. On the simultaneous associativity of $F(x,y)$ and $x+y-F(x,y)$. **Aequations Mathematicae**, v.19, p.194-226, 1979.

FREES, E.; WANG, P. Credibility using copulas. **North American Actuarial Journal**, v. 9, p. 31-48, 2005.

FREITAS, M.; COLOSIMO, E. A. **Confiabilidade: Análise de Tempo de Falha e Testes de Vida Acelerados**. Belo Horizonte: Fundação Cristiano Ottoni - UFMG. ISBN: 9788585447571, 309 pages, 1ºEdição, 1997.

GELMAN, A.; RUBIN, D. B. Inference from iterative simulation using multiple sequences. **Statistical Science**, v. 7, p. 457-511, 1992.

GUPTA, R. D.; KUNDU, D. Generalized Exponential distributions. **Australian** & **New Zealand Journal of Statistics**, v. 41, p. 173-188, 1999.

GUPTA, R. D.; KUNDU, D. On bivariate inverse Weibull distribution. **Brazilian Journal of Probability and Statistics**, 2016.

GYÕRFFY, B.; SUROWIAK, P.; BUDCZIES, J.; LÁNCZKY, A. Online survival analysis software to assess the prognostic value of biomarkers using transcriptomic data in non-small-cell lung cancer. PLoS ONE 8(12): e82241, 2013.

HARRELL, F. **Regression Modeling Strategies: With Applications to Linear Models, Logisticand Ordinal Regression, and Survival Analysis**. Springer Series in Statistics. ISBN: 9783319194257, 582 pages, 2ºEdição, 2015.

HOLLANDER, M.; WOLFE, D. A. **Nonparametric statistical methods**. Wiley, New York, 1973.

HOUGAARD, P. A class of ultivariate failure time distributions. **Biometrika**. v. 73, p. 671-678, 1986.

IBRAHIM, J. G.; CHEN, M.; SINHA, D. **Bayesian Survival Analysis**. New York: Springer-Verlag. ISBN: 9781441929334, 480 pages, 1ºEdição, 2001.

IRENE, G.; KLAUS, H. On the distribution of sums of random variables with copula-induced dependence. **Insurence: Mathematics and Economics**, v. 59, p. 27-44, 2014.

JAWORSKI, P.; DURANTE, F.; HARDLE, W. K.; RYCHLIK, T. **Copula Theory and Its Applications**. ISBN: 9783642124648, 198 pages, 2010.

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. **Journal of the American Statistical Association**, v. 53, n. 282, p. 457-481, 1958.

KLEIN, J. P.; MOESCHBERGER, M. L. **Survival Analysis: Techniques for Censored and Truncated Data**. Statistics for biology and health. ISBN: 038795399X, 542 pages, 2ºEdição, 2003.

KLEINBAUM, D. G.; KLEIN, M. **Survival Analysis: A Self-Learning Text**. Statistics for biology and health. ISBN: 9781441966469, 700 pages, 3ºEdição, 2012.

KOLEV, N.; DOS ANJOS, U.; MENDES, B. V. M. Copulas: a review and recent developments. **Stochastic Models**, v. 22, n. 4, p. 617-660, 2006.

KRUSKAL, W. H. Ordinal measures of association. **Journal of the American Statistical Association**, v. 53, p. 814-861, 1958.

LAWLESS, J. F. **Statistical Models and Methods for Lifetime Data**. ISBN: 0471372153, 67 pages, 2ºEdição, 2003.

LEAL, D. M. B. **Aplicação de Cópulas ao Ramo Vida: Risco de Resgate e Risco de Taxa de Juro**. DM - Dissertações de Mestrado. Universidade Técnica de Lisboa. Instituto Superior de Economia e Gestão, 2010.

LEHMANN, E. L. **Nonparametrics: statistical methods based on ranks**. Holden-Day, San Francisco, 1975.

LEHMANN, E. L. Some concepts of dependence. **The Annals of Mathematical Statistics**, v. 37, p. 1137-1153, 1966.

LOUZADA, F.; MAZUCHELI, J.; ACHCAR, J. A. **Análise de Sobrevivência e Confiabilidade**. Monografias del IMCA. Lima, Peru: IMCA, 2002a.

LOUZADA, F.; MAZUCHELI, J.; ACHCAR, J. A. Mixture hazard models for lifetime data. **Biometrical Journal**, Alemanha, v. 44, p. 3-14, 2002b.

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**, v. 10, p. 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**, v. 42, n. 4, p. 673-691, 2013.

MARTINEZ, E. Z.; ACHCAR, J. A. Trends in epidemiology in the 21st century: time to adopt Bayesian methods. **Cadernos de Saúde Pública**, v. 30, n. 4, p. 703-714, 2014.

MCGILCHRIST C. A.; AISBETT C. W. Regression with frailty is survival analysis. **Biometrics**, v. 47, p. 461-466, 1991.

MEEKER, W. Q.; ESCOBAR, L. A. **Statistical Methods for Reliability Data**. Wiley series in probability and statistics. ISBN: 9780471143284, 712 pages, 1ºEdição, 1998.

MIKOSCH, T. Copulas: tales and facts. **Extremes**, v. 9, p. 3-20, 2006.

NELSEN, R. Properties of a one-parametric family of bivariate distributions with specified marginals. **Communications in Statistics**, v. 15, p. 3277-3285, 1986.

NELSEN, R. **An Introduction to Copulas**. New York: Springer. ISBN: 0387286594, 269 pages, 2ºEdição, 2006.

NELSON, W. **Accelerated Life Testing: Statistical Models Data Analysis and Test Plants**. New York: John Wiley and Sons, 1990.

OAKES, D. Bivariate survival models induced by frailties. **Journal of the American Statistical Association**, v. 84, p. 487-493, 1989.

OLIVEIRA, M. A.; SUZUKI, A. K.; SARAIVA, E. F. Uma abordagem Bayesiana para modelos de sobrevivência bivariados baseados em cópulas arquimedianas. **Revista Brasileira de Biometria**, v. 32, p. 390-411, 2014.

PENG, F.; DEY, D. Bayesian analysis of outlier problems using divergence measures. **The Canadian Journal of Statistics - La Revue Canadienne de Statistique**, v. 23, p. 199-213, 1995.

PEREIRA, G. A.; LOUZADA NETO, F.; SUZUKI, AK; CANCHO, VG; PRINCE, FL. The Long-Term Bivariate Survival FGM Copula Model: An Application to a Brazilian HIV Data. **Journal of Data Science (Print)**, v. 10, p. 511-535, 2012.

PIERCE, D. A.; STEWART, W. H.; KOPECHY, K. Distribution-free regression analysis of grouped survival data. **Biometrics**, Washington, v. 35, p. 785-793, 1979.

PLUMMER, M. JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. **DSC 2003 Working Papers**, 2003.

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 - preliminaries. **International Actuarial Association: 13th EAA Conference: Bali**, Indonesia, p. 12-15, 2005.

QUIROZ 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 CORE TEAM. **R: A Language and Environment for Statistical Computing**. R Foundation for Statistical Computing: Vienna, Austria. ISBN: 3900051070. http://www.R-project.org, 2019.

RIBEIRO, T. R.; SUZUKI, A. K.; SARAIVA, E. F. Uma abordagem bayesiana para o modelo de sobrevivência bivariado derivado da cópula AMH. **Revista da estatística da Universidade Federal de Ouro Preto**, v. 6, 2017.

RODRIGUES, J.; CANCHO, V. G.; CASTRO, M. **Teoria Unificada de Análise de Sobrevivência**. 18º Sinape - São Pedro, 94 pages, 2008.

ROMEO, J. S.; TANAKA, N. I.; LIMA, A. C. P. Bivariate survival modeling: a Bayesian approach based on copulas. **Lifetime Data Analysis**, Hinghan, v. 12, p. 205-222, 2006.

SANTOS, R. P. S. **Modelando Contágio Financeiro Através de Cópulas**. Dissertação de Mestrado em Economia. Escola de Economia da Fundação Getúlio Vargas, 2010.

SANTOS, C. A.; ACHCAR, J. A. A Bayesian analysis in the presence of covariates for multivariate survival data: an example of application. **Revista Colombiana de Estadística**, v. 34, p. 111-131, 2011.

SKLAR, A. Fonctions de répartition à n dimensions et leurs marges. **Publications de l'Institut de Statistique de l'Université de Paris**, v. 8, p. 229-231, 1959.

SCHLOSSER A. Normal Inverse Gaussian Factor Copula Model. In: Pricing and Risk Management of Synthetic CDOs. **Lecture Notes in Economics and Mathematical Systems, Springer, Berlin, Heidelberg**, v. 646, 2011.

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**, v. 64, p. 583-639, 2002.

SUZUKI, A. K.; BARRIGA, G. D. C.; LOUZADA, F.; CANCHO, V. G. A general long-term aging model with different underlying activation mechanisms: modeling, Bayesian estimation and case influence diagnostics. **Communications in Statistics-Theory and Methods**, 2016. To appear.

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**, v. 21, n. 1, p. 55-76, 2011.

SUZUKI, A. K.; LOUZADA, F.; 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**, v. 10, p. 511-535, 2012.

The Diabetic Retinopathy Study Research Group, Preliminary report on the effect of photocoagulation therapy, American Journal of Ophthalmology, v. 81, p. 383-396, 1976.

when borrowers default. **European Journal of Operations Research**, v. 218, p. 132-139, 2012.

VAUPEL, J. W., MANTON, K. G.; STALLARD, E. The impact of heterogeneity in individual frailty on the dynamics of mortality. **Demography**, v. 16, p. 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**, v. 1, p. 17-34, 2010.

VIOLA, M. L. L. **Tipos de Dependência entre Variáveis Aleatórias e Teoria de Cópulas**. Instituto de Matemática, Estatística e Computação Científica, 2009.

YIN, G.; IBRAHIM, J. G. Cure rate models: a unified approach. **The Canadian Journal of Statistics**, v. 33, n. 4, p. 559-570, 2005.

ZHANG, L.; SINGH, V. P. Bivariate rainfall fequency distributions using Archimedean copulas. **Journal of Hydrology**, Department of Biological & Agricultural Engineering, Texas A & M University, 2117 TAMU, College Station, Texas USA, v. 332, p. 93-109, 2007.

WALPOLE, Ronald E.; MYERS, Raymond H.; MYERS, Sharon L. e YE, Keying. *Probability & Statistics for Engineers & Scientists*. Pearson Education International. ISBN 0132047675, E. 9, 2016.

WEIBULL, W. A statistical theory of the strength of material. **Royal Technical University**, Stockholm, v. 151, 1939.

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

WUERTZ, D.; MAECHLER, M.; MEMBERS, R. C. T. Stable distribution functions. Density, probability and quantile functions, and random number generation for (skew) stable distributions, using the parametrizations of Nolan, **https://cran.r-project.org/web/packages/stabledist/stabledist.pdf**, 2016.

APÊNDICE

A

CÓDIGO DO MODELO PVF USANDO *JAGS*

Código-fonte 1 – Código para ajuste do modelo bayesiano PVF

```
cat( "model

    {

    for (i in 1:N){

    theta1[i]<-exp(beta01+beta11*x1[i])
    theta2[i]<-exp(beta02+beta12*x2[i])

    # Sobrevivência marginal
    S1[i]<-exp(-theta1[i]*pow(t1[i],r1))
    S2[i]<-exp(-theta2[i]*pow(t2[i],r2))

    f1[i]<- r1*theta1[i]*pow(t1[i],r1-1)
    f2[i]<-r2*theta2[i]*pow(t2[i],r2-1)


    A1[i]<-pow(eta,alpha)-alpha*pow(eta,alpha-1)*log(S1[i])
    A2[i]<-pow(eta,alpha)-alpha*pow(eta,alpha-1)*log(S2[i])

    A1A2eta[i]<-pow(A1[i],1/alpha)+pow(A2[i],1/alpha)-eta

    Cs1s2[i]<-exp(-(1/alpha)*(pow(eta,1-alpha)*pow(A1A2eta[i],
  alpha)-eta))

    parte1[i]<-log(
```

```
26:       pow(A1[i]*A2[i],1/alpha-1)*f1[i]*f2[i]*Cs1s2[i]*pow(
      A1A2eta[i],alpha-2)*
27:       (pow(A1A2eta[i],alpha)+(1-alpha)*pow(eta,alpha-1)) )
28:
29:       parte2[i]<-log(Cs1s2[i]*pow(A1A2eta[i],alpha-1)*pow(A1[i
      ],1/alpha-1)*
30:       r1*theta1[i]*pow(t1[i],r1-1))
31:
32:       parte3[i]<-log(Cs1s2[i]*pow(A1A2eta[i],alpha-1)*pow(A2[i
      ],1/alpha-1)*
33:       r2*theta2[i]*pow(t2[i],r2-1))
34:
35:       parte4[i]<-log(Cs1s2[i])
36:
37:       llike[i]<-d1[i]*d2[i]*parte1[i]+d1[i]*(1-d2[i])*parte2[i]
38:       +(1-d1[i])*d2[i]*parte3[i]+(1-d1[i])*(1-d2[i])*parte4[i]
39:
40:       L[i]<-exp(llike[i])
41:
42:       phi[i]<-L[i]/C
43:
44:       zeros[i]~dbern(phi[i])
45:
46:       }
47:
48:       alpha~dbeta(1,1)
49:       eta~dgamma(0.1,0.1)
50:       r1~dgamma(0.1,0.1)
51:       r2~dgamma(0.1,0.1)
52:       beta11~dnorm(0,0.001)
53:       beta12~dnorm(0,0.001)
54:       beta01~dnorm(0,0.001)
55:       beta02~dnorm(0,0.001)
56:
57:       C<-100000
58:
59:       }",
60:
61: file="PVFR.jags" )
```

ANEXO

A

MEDIDAS DE DIAGNÓSTICO PARA O CASO SEM CENSURA.

A seguir apresentamos todos os gráficos de medidas de diagnóstico para os dados sem censura. Na Figura 23 temos a amostra sem nenhum caso perturbado. Nas Figuras 24, 25 e 26 temos a amostra com uma das observações perturbadas. Nas Figuras 27, 28 e 29 temos a amostra com duas observações perturbadas e, por fim, na Figura 30 temos a amostra com três observações perturbadas.

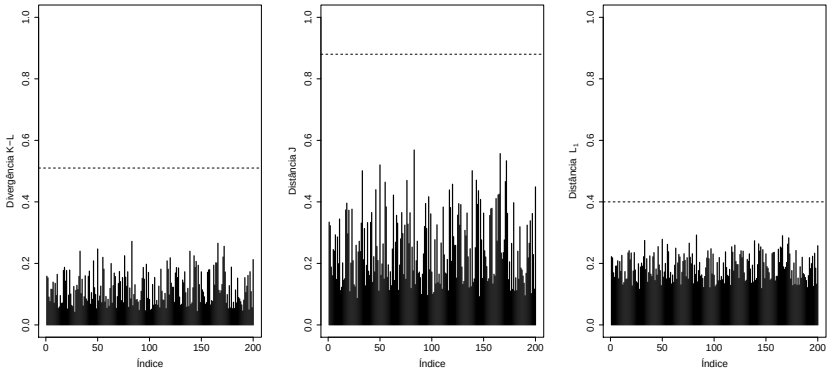

Figura 23 – Gráfico de índices das medidas de divergência para o conjunto de dados (a).

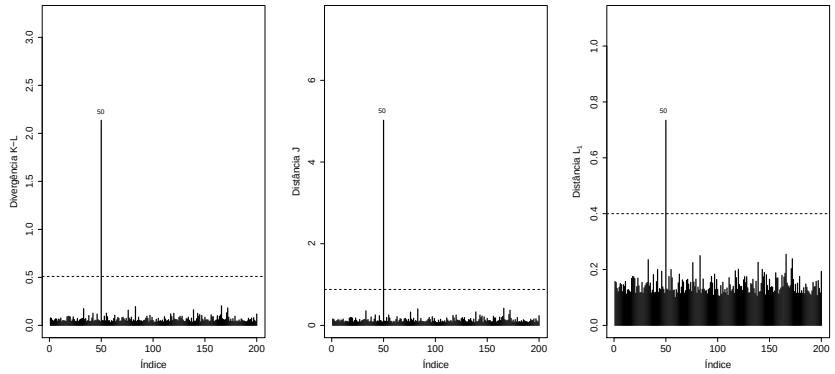

Figura 24 – Gráfico de índices das medidas de divergência para o conjunto de dados (b).

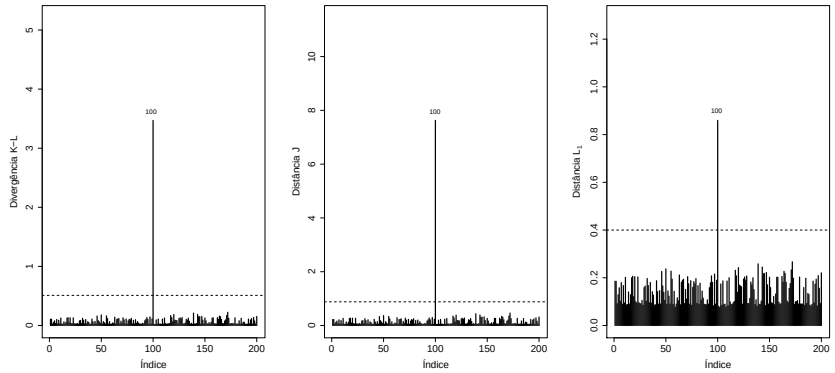

Figura 25 – Gráfico de índices das medidas de divergência para o conjunto de dados (c).

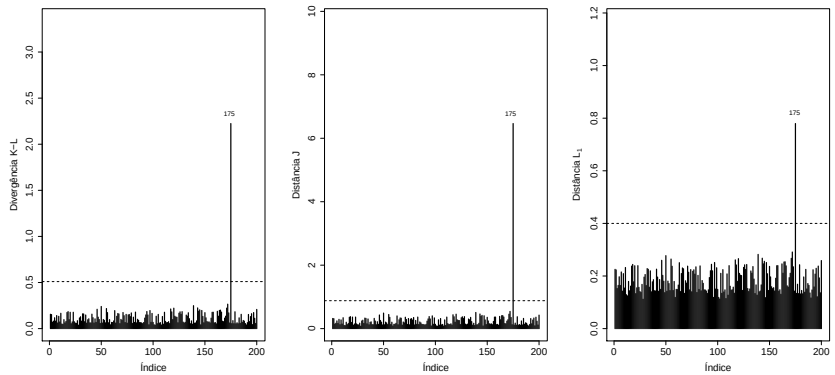

Figura 26 – Gráfico de índices das medidas de divergência para o conjunto de dados (d).

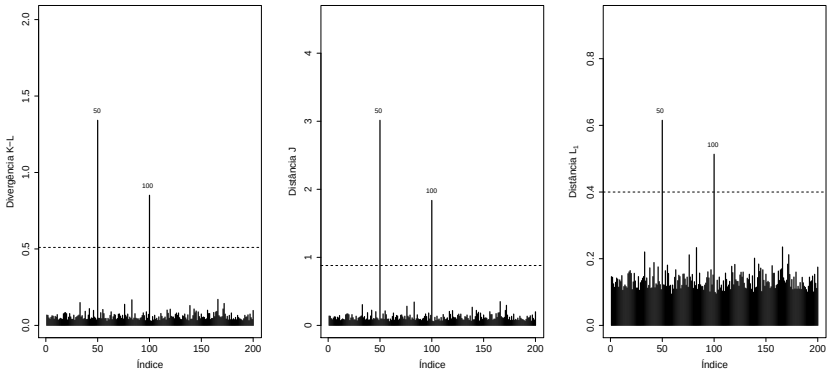

Figura 27 – Gráfico de índices das medidas de divergência para o conjunto de dados (e).

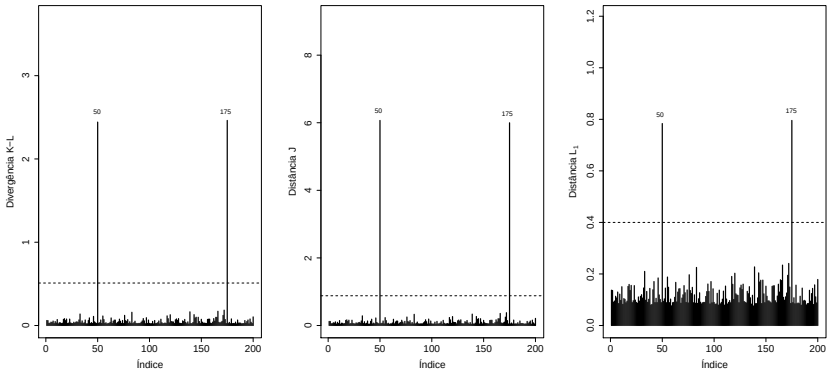

Figura 28 – Gráfico de índices das medidas de divergência para o conjunto de dados (f).

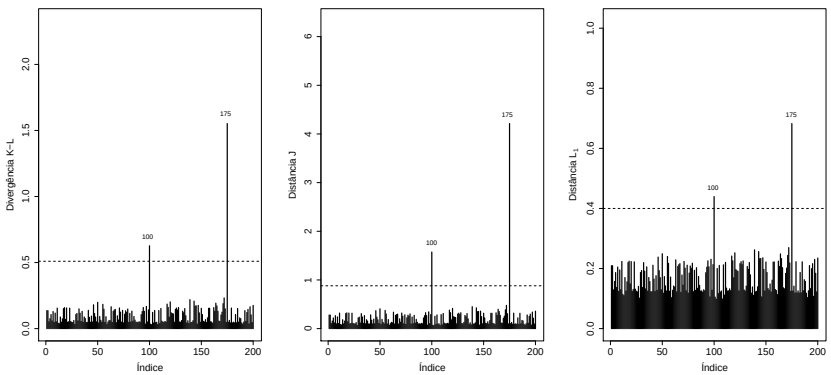

Figura 29 – Gráfico de índices das medidas de divergência para o conjunto de dados (g).

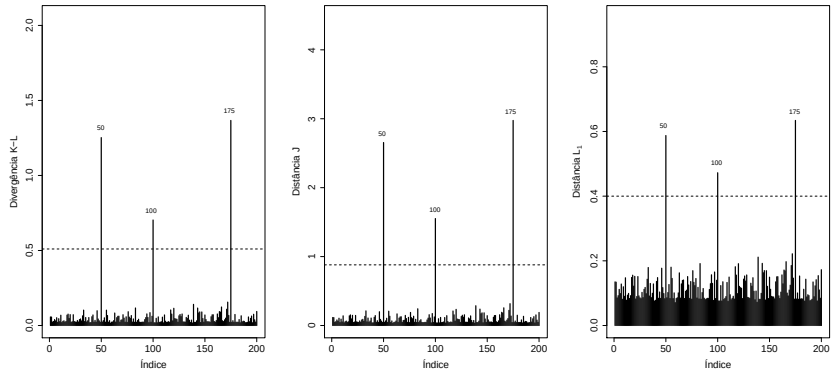

Figura 30 – Gráfico de índices das medidas de divergência para o conjunto de dados (h).

ANEXO

B

MEDIDAS DE DIAGNÓSTICO PARA O CASO COM CENSURA.

A seguir apresentamos todos os gráficos de medidas de diagnóstico para os dados com censura. Na Figura 31 temos a amostra sem nenhum caso perturbado. Nas Figuras 32, 33 e 34 temos a amostra com uma das observações perturbadas. Nas Figuras 35, 36 e 37 temos a amostra com duas observações perturbadas e, por fim, na Figura 38 temos a amostra com três observações perturbadas.

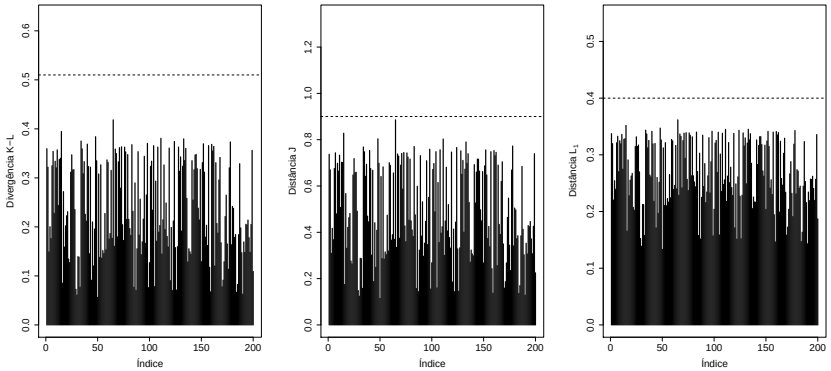

Figura 31 – Gráfico de índices das medidas de divergência para o conjunto de dados (a), considerando censura.

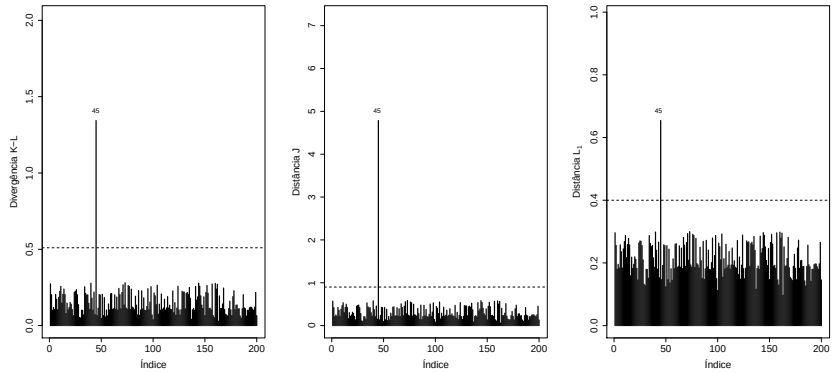

Figura 32 – Gráfico de índices das medidas de divergência para o conjunto de dados (b), considerando censura

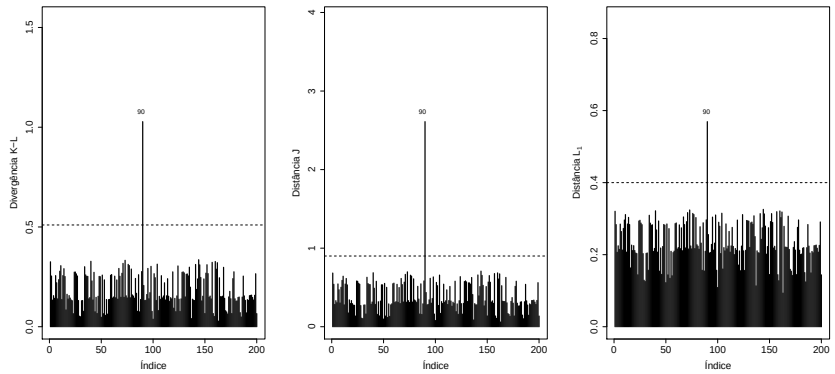

Figura 33 – Gráfico de índices das medidas de divergência para o conjunto de dados (c), considerando censura

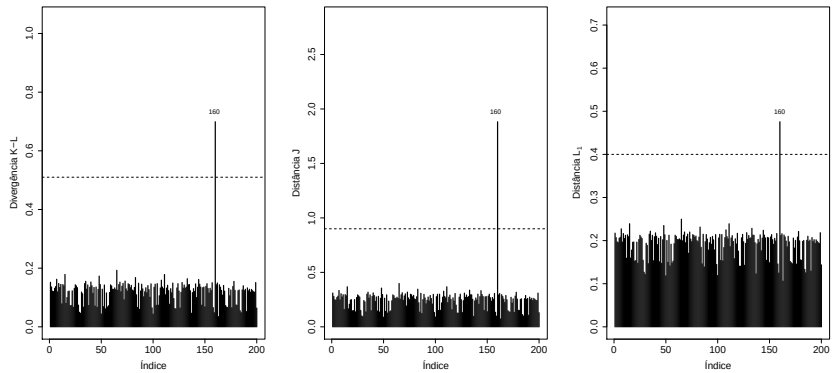

Figura 34 – Gráfico de índices das medidas de divergência para o conjunto de dados (d), considerando censura

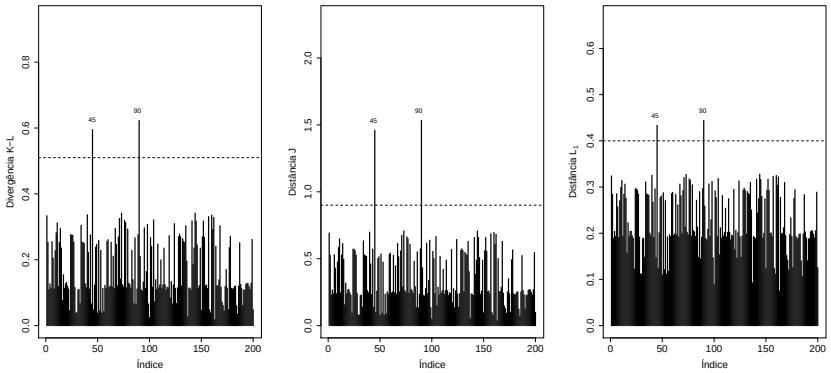

Figura 35 – Gráfico de índices das medidas de divergência para o conjunto de dados (e), considerando censura

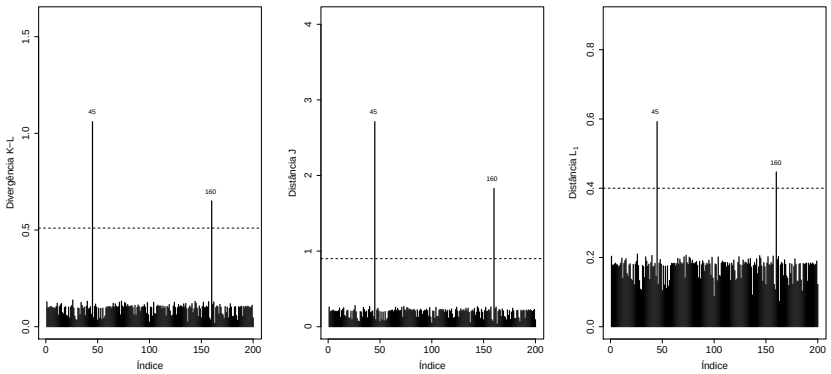

Figura 36 – Gráfico de índices das medidas de divergência para o conjunto de dados (f), considerando censura

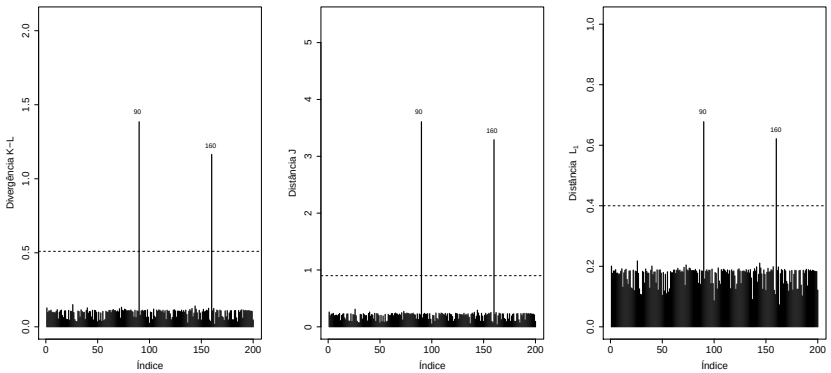

Figura 37 – Gráfico de índices das medidas de divergência para o conjunto de dados (g), considerando censura

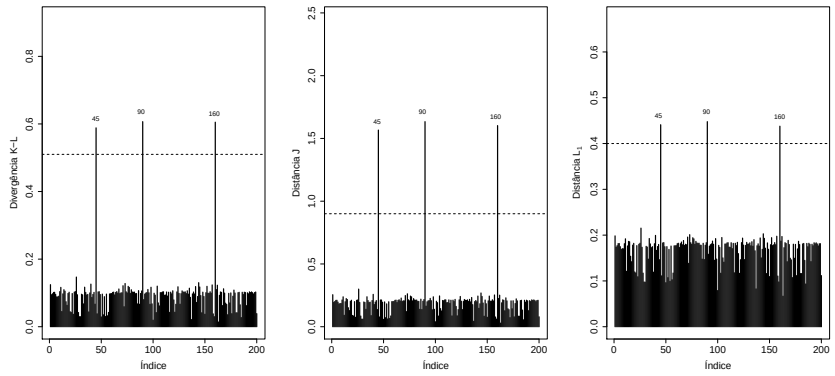

Figura 38 – Gráfico de índices das medidas de divergência para o conjunto de dados (h), considerando censura


ICMC
USP
SÃO CARLOS