University of São Paulo
Ribeirão Preto Medical School

Marcos Vinicius de Oliveira Peres

Bivariate distributions based on copulas functions: developments and applications in medical studies

Distribuições bivariadas baseadas em funções cópulas: desenvolvimento e aplicações em estudos médicos

Ribeirão Preto
2020

Marcos Vinicius de Oliveira Peres

Bivariate distributions based on copulas functions: developments and applications in medical studies

Thesis presented to the Public Health Program of the University of São Paulo at Ribeirão Preto Medical School as part of the requirements for obtaining the title of Doctor in Science.

Advisor: Edson Zangiacomi Martinez

Co-advisor: Jorge Alberto Achcar

Ribeirão Preto

2020

Marcos Vinicius de Oliveira Peres

Distribuições bivariadas baseadas em funções cópulas: desenvolvimento e aplicações em estudos médicos

Tese apresentada ao programa em Saúde Pública da Universidade de São Paulo da Faculdade de Medicina de Ribeirão Preto como parte dos requisitos para a obtenção do título de Doutor em Ciências.

Orientador: Edson Zangiacomi Martinez

Co-orientador: Jorge Alberto Achcar

Ribeirão Preto

2020

Catalog record electronically elaborated by the author, using the LaTeXtemplate of USPSC class from University of São Paulo.

Peres, Marcos Vinicius de Oliveira Peres
Bivariate distributions based on copulas functions: developments and applications in medical studies / Marcos Vinicius de Oliveira Peres. – Ribeirão Preto, 2020.
81 p. : il. (some color).

Thesis (Ph.D.) – Ribeirão Preto Medical School. University of São Paulo - FMRP/USP. Advisor: Edson Zangiacomi Martinez. Co-advisor: Jorge Alberto Achcar

1. Biostatistics. 2. Bayesian approach. 3. Bivariate models. 4. Cancer studies. 5. Cure rate. 6. Copula functions. 7. Medical studies. 8. Survival analysis. I. Martinez, Edson Zangiacomi. II. University of São Paulo. III. Ribeirão Preto Medical School. IV. Bivariate distributions based on copulas functions: developments and applications in medical studies.

Ficha catalográfica elaborada eletronicamente pelo autor, utilizando o modelo LaTeX da classe USPSC da Universidade de São Paulo.

Peres, Marcos Vinicius de Oliveira Peres

Distribuições bivariadas baseadas em funções cópulas: desenvolvimento e aplicações em estudos médicos / Marcos Vinicius de Oliveira Peres.– Ribeirão Preto, 2020.

81 p. : il. (algumas cores).

Tese (Doutorado) – Faculdade de Medicina de Ribeirão Preto. Universidade de São Paulo - FMRP/USP. Orientador: Edson Zangiacomi Martinez. Coorientador: Jorge Alberto Achcar

1. Bioestatística. 2. Abordagem Bayesiana. 3. Modelos Bivariados. 4. Estudos de câncer. 5. Fração de cura. 6. Funções cópulas. 7. Estudos médicos. 8. Análise de sobrevida. I. Martinez, Edson Zangiacomi. II. Universidade de São Paulo. III. Faculdade de Medicina de Ribeirão Preto. IV. Distribuições bivariadas baseadas em funções cópulas: desenvolvimento e aplicações em estudos médicos.

Peres, M. V. O. **Bivariate distributions based on copulas functions: developments and applications in medical studies**. 2021. 81p. Thesis. Ribeirão Preto: Ribeirão Preto Medical School, University of São Paulo; 2021.

Examing Board

Edson Zangiacomi Martinez
Advisor/President

Jair Licio Santos
Examiner 1

Edwin Moises Marcos Ortega
Examiner 2

Diego Ignacio Gallardo Mateluna
Examiner 3

Ribeirão Preto

2020

"This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001."

"O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001."

"I dedicate this Doctoral Thesis to everyone who supported me on my journey."

Acknowledgements

I would like to express my sincerest gratitude to my advisor and friend Edson Zangiacomi Martinez, who gave me advice, ideas, suggestions, and introduced me to this field of study. I am utterly grateful to him for sharing his knowledge and proposing interesting work themes along my doctorate, and I also thank him for his patience and dedication.

I would also like to thank the co-advisor Jorge Alberto Achcar for his knowledge and support, especially in the manuscripts derived from this thesis. I am also thankful to the Ribeirão Preto Medical School.

Lastly, I would like to thank my family, especially my parents, for their love, advice, support and for being by my side on this journey encouraging me. I am also grateful to my friends for all support and help, especially those now are my research partners.

Abstract

Peres, M. V. O. **Bivariate distributions based on copulas functions: developments and applications in medical studies**. 2021. 81p. Thesis. Ribeirão Preto: Ribeirão Preto Medical School, University of São Paulo; 2021.

Multivariate survival data are found in several studies, in particular, studies where there are two observed lifetimes associated to the same individual, and in some cases there exists a dependence structure between the two lifetimes. In addition, with the recent advances of medicine and improvement of treatments, there is an increasing of fraction of individuals not expecting to experience the event of interest. These individuals are immune to the event or cured for the disease during the study and known as long-term survivors or cured individuals. In these situations, the usual existing lifetime distributions are not appropriate to model data sets with long-term survivors and dependent bivariate lifetimes. For the modeling of bivariate data the use of copula survival functions is an alternative explored in this study, also assuming individuals in the presence of cure fractions modeled with standard mixture models, non-mixture models and also defective distributions. Motived by this, in this study it was introduced some continuous lifetime bivariate distributions considering copula functions in presence of censored data and lifetime data with long-term survivors. The proposed models are useful in medical situations to study the dependence structure of pair of lifetimes and in presence of cure rates. This work also proposed to compare the bivariate Kaplan-Meier estimator with the surface estimated from copulas by means of simple calculations of the distance between matrices. This methodology presented efficient results to compare bivariate models estimated by copulas with empirical survival estimates obtained using the bivariate Kaplan-Meier non-parametric estimator. Another interesting result obtained in this study is that the use bivariate distributions in presence of censoring and cure rate have better computational performance to get the inferences of interest under a Bayesian approach. According to the results obtained in our study, another interesting point is that the selected models lead to accurate estimation of the cure rate using Markov Chain Monte Carlo (MCMC) simulation algorithms with good stability in the generation of Gibbs samples of interest in the applications. Finally, it is possible to emphasize that the programing routine to get Bayesian inference of interest can be easily executed by free and open source softwares as OpenBUGS, JAGS or R with low computational costs.

Keywords: Biostatistics; Bayesian approach; bivariate models; cancer studies; cure rate; copula functions; medical studies; survival analysis.

Resumo

Peres, M. V. O. **Distribuições bivariadas baseadas em funções cópulas: desenvolvimento e aplicações em estudos médicos**. 2021. 81p. Tese. Ribeirão Preto: Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo; 2021.

Dados multivariados de sobrevida são encontrados em diversos estudos, em particular, é comum observar dois tempos de vida associadas ao mesmo indivíduo, e em alguns casos existe uma estrutura de dependência entre os dois tempos. Além disso, atualmente com o avanço da medicina e aprimoramento dos tratamentos, aumenta a presença de uma fração de indivíduos que não esperam vivenciar o evento de interesse, esses indivíduos são imunes ao evento ou curados da doença durante o estudo, conhecidos como sobreviventes de longo prazo ou indivíduos curados. Nessas situações, as técnicas usuais de analise de sobrevida existentes não são apropriadas para modelar os conjuntos de dados com sobreviventes de longo prazo e tempos de vida bivariados dependentes. Para modelar dados bivariados de sobrevivência podemos considerar o uso de funções cópulas, e no estudo de indivíduos com fração de cura é comum considerar o modelo de mistura padrão, modelos de não mistura e as distribuições defectivas. Motivado por isso, neste estudo foram introduzidas algumas distribuições bivariadas de sobrevidas contínuas baseadas em funções cópulas na presença de dados de censura e fração de cura. Os modelos propostos são úteis em situações médicas para estudar a estrutura de dependência entre os tempos de vida e as frações de cura. Este trabalho também se propôs comparar o estimador Kaplan-Meier bivariado com a superfície bivariada de sobrevida estimada a partir de cópulas, considerando cálculos simples da distância entre matrizes. Esta metodologia apresentou resultados eficientes para comparar modelos bivariados estimados por cópulas com as estimativas empíricas de sobrevivência do estimador de Kaplan-Meier bivariado. Outro resultado interessante apresentado neste estudo é que o uso de distribuições bivariadas na presença de censura e taxa de cura tem melhor desempenho computacional para obter as inferências de interesse sob uma abordagem Bayesiana. De acordo com os resultados obtidos em nosso estudo, outro ponto interessante é que os modelos selecionados levam a estimativa precisa da taxa de cura utilizando algoritmos de simulação de Markov Chain Monte Carlo (MCMC) com boa estabilidade na geração de amostras de Gibbs de interesse nas aplicações. Por fim, é possível enfatizar que a rotina de programação para obtenção da inferência Bayesiana de interesse pode ser facilmente executada por softwares livres e de código aberto como OpenBUGS, JAGS ou R e com baixo custo computacional.

Keywords: Bioestatística; Abordagem Bayesiana; modelos bivariados; estudos de câncer; fração de cura; função cópula; estudos médicos; análise de sobrevida.

List of Figures

Contents

Chapter 1

Introduction

In the lifetime data analysis, researchers commonly use standard non-parametric and parametric techniques as, for example, the Kaplan–Meier estimators for the survival function, the log-rank test or semi-parametrical Cox proportional hazard models, and traditional parametric survival distributions. However, there are situations where more complex models are needed. For example, it is possible, in some cases, to observe two time-to-event variables for each patient, in this situation, the variables related to the time to the events need to be modeled by a better approach, like a bivariate probability distribution or frailty models. As an alternative, the bivariate structure and your respective dependence can be specified by using a Copula function due to its simplicity. The copula is a multivariate distribution function with standard uniform univariate marginals. Different copula functions introduce different dependence structures between random variables. There are several copulas functions in literature, therefore, it is important to choose a suitable copula to describe the dependence structure between the variables of interest in each application.

Other common situation in the analyzes of time-to-event data, particularly in cancer research, occurs when it is expected that a fraction of subjects will not experience the event of interest. The usual methods assume that all individuals are susceptible to the event of interest. However, for example in clinical studies, there may be patients who will not experience the event under investigation, that is, these patients are immune to the event or they were cured during the research. This situation is suggested when a Kaplan-Meier estimator plot for the survival function describes a behavior with stable plateau and large censored data at the right of the curve. In this way, the use of models that incorporate this plateau, named cure rate models, could be a better alternative to predict or to identify prognostics factors that affects the survival probability. In this situation, usually it is assumed that the population is a mixture of susceptible individuals who experience the event of interest and non-susceptible individuals that supposedly will never experience it. In additional, the use of defective distributions is an alternative to these models, in this case, a correspondent survival function $S(t)$ based on a defective distribution converges to a plateau at a value ρ, where ρ is associated to the cured fraction.

The present thesis is organized as follows: in Chapter 2, presents the objectives of the thesis and of each prepared article. In the Chapter 3, we present a brief literature review about the methods used in the thesis, highlighting a review based on 15 different copulas functions. Chapter 4, we present the medical data sets considered in the articles derived from the thesis. In Chapter 5, Chapter 6 and Chapter 7, we display a brief description on the articles produced, also presenting the respective first page. In Chapter 8, it is discussed the general conclusions of this Doctoral Thesis.

Chapter 2

Objectives

The main objective of this work is to develop and to study some models to bivariate lifetime with long-term survivors based on copula functions and their applications to real medical data.

This general objective is divided into three specific objectives, as follow:

- To present a review on the copula functions that can be useful for the construction of bivariate distributions applied to the lifetime data analyses, including the presence of long-term survivors and censored data.
- To introduce and to explore the use of the Clayton copula in the analysis of bivariate lifetime data assuming a bivariate defective Gompertz distribution to estimate the cure rate, with a brief comparison among maximum likelihood and Bayesian methods.
- To present a simulation study of a proposed bivariate modified Weibull distribution derived from a Farlie–Gumbel–Morgenstern copula, in the presence of different sample sizes, percentage of censored data and different correlations between the time-to-event variables, with a brief comparison among the maximum-likelihood and Bayesian approaches.

Each specific objective has been developed in the form of different scientific articles, as presented in the next three chapters.

Chapter 3

Theoretical framework

3.1 Survival analysis

Expressions such as survival data, lifetime data, failure time data, or time to event data are terms used to describe data that measure the time to the occurrence of a given event of interest. These data are found in several research fields, such as medicine, biology, public health, engineering, reliability research, business, criminology, epidemiology, social and behavioral sciences and among many other fields. In the medical field, the event can be the development or remission of a disease, response to a treatment, relapse or death. In this way, the lifetime can be the overall survival time, time of metastasis and outcome cancer, tumor-free time, the time from the start of treatment to response, duration of remission, and time to death. Besides that, survival data can include covariates, as type of treatment, patient characteristics, tumor grade, hospitalization time, and the development of a disease. Once the time of occurrence of an event is defined, it is possible to use non-parametric, semi-parametric or parametric modeling approaches in data analysis. In the literature several authors have been described the traditional parametric and non-parametric techniques of survival analysis, and we can cite as examples Meeker and Escobar (1998), Kalbfleisch and Prentice (2002), Lee and Wang (2003), Rausand and Arnljot (2004), Colosimo and Giolo (2006), Selvin (2008), Lawless (2011), Liu (2012) andCollett (2015).

The main aspect that differs survival analysis from other statistical analyzes is the presence of censored data. The survival time of a patient is censored when it is known a partial individual information about the time of occurrence of the event of interest, however it is not known the exact time of occurrence of the event, that is, the observed time is different of the real time of occurrence. In general, the censored data can occur for a variety of reasons, as the loss of monitoring of the patient over time, that is, the follow-up of the patient was interrupted, the non-occurrence of the event of interest until the end of the experiment or the patient died of a cause different than the one studied. Colosimo and Giolo (2006) quote two important points that justify the use of censored data in statistical analysis: (I) to provide information about the patient's lifetime, although the observation is not complete, and (II) the omission of the censored observations can lead to biased estimates. It is important to point out that disregarding the presence of censored observations, the classical statistics techniques, such as regression analysis or analysis of variance, could be used in survival data analysis.

In the literature, there are different types of censoring schemes, and three of them are the most considered: right-censoring, left-censoring and interval censoring. The right-censoring occurs when the observed time to event is less than the real time, divided into three ways: type I, type II and random censoring.

- Type I censoring, occurs to patients who after a predetermined period of time, at the end of the study, the event of interest was not observed.
- Type II censoring, happens for the patients who after a predetermined number of observed event occurrences, at the end of the study, the event of interest was not experienced.
- Random censoring, occurs when an individual is removed during the study for any reason, without the event of interest has been experienced, or if the individual dies for a reason unrelated to outcome of interest. This type of censoring is the more common.
- Left-censoring, defined when what is known is that the individual has experienced the event of interest before to the start of the study.
- Interval censoring, it occurs where the only available information is that the event occurred within some interval, in this case, it is not known exactly when the event occurred, it is only know that the event occurred within a certain time interval.

In general, in survival data an individual under study is represented by an ordered pair (t, δ), where t is the failure or censoring time of the patient and δ is a binary indicator of failure or censoring. That is, let T be a random variable related to the failure time of a patient and C a random variable independent of T related to the censoring time, thus,

$$t = \min(T, C) \quad \text{and} \quad \delta = \begin{cases} 1, & \text{if } T \leq C \\ 0, & \text{if } T > C \end{cases}$$

In the lifetime data analysis, researchers commonly use standard non-parametric techniques as, for example, the Kaplan–Meier estimators for the survival function (Kaplan and Meier, 1958), the log-rank test (Mantel, 1966) or semi-parametrical Cox proportional hazard models (Cox, 1972) and parametric survival models. Parametric models are more flexible than Cox proportional hazards model, especially when there is no proportionality of risks between groups. In the parametric context, there are two functions which are the great interest in the survival data analysis: the survival function and hazard function. The survival function can be interpreted and defined as the probability that an observation will not fail until a certain time t. In probabilistic terms,

$$S(t) = P(T \geq t) = 1 - F(t)$$

where $F(t)$ denotes the cumulative distribution function and it is the probability of a randomly selected subject dying before time t. The survival function is a decreasing and left continuous function, such that, $S(0) = 1$ and $\lim_{t\to\infty} S(t) = 0$. On other hand, the hazard function $\lambda(t)$ represents the instantaneous failure rate at time t conditional on survival time t. The hazard function can be defined as the probability of dying at time t given that the individual survived until the time t, that is,

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(T \in [t, t + \Delta t) \mid T \geq t)}{\Delta t}$$

where Δt is an infinitesimal increment of time. The hazard function is a important tool in the modeling of survival data and relevant to describe the pattern of failures of individuals. It can also be obtained by the expression

$$\lambda(t) = \frac{f(t)}{S(t)}, \tag{3.1}$$

where $f(t)$ is the probability density function, such that,

$$f(t) = \frac{\mathrm{d}F(t)}{\mathrm{d}t}. \tag{3.2}$$

Some of the survival distributions commonly used in survival analysis are the following: Weibull distribution and their extensions, exponential distribution, Gamma distribution and log-normal distribution. However, in some situations more complex models are needed. A very common example in survival analysis is given when for each patient are observed two lifetimes, that is, for each sample unit, two lifetimes are measured. This situation can be found when the event of interest is to observe the lifetimes of paired human organs, such as kidneys, eyes or breasts, the time of recurrence of the cancer in two different organs, the time until and the first and second infection or hospitalization, among others.

3.2 Survival distributions

3.2.1 Univariate Kaplan-Meier estimator

The Kaplan-Meier estimator (or product-limit estimator) is the most common non-parametric estimator of the survival function (Kaplan and Meier, 1958). Assume that n individuals have lifetimes represented by a random variable T which are subject to random right-censoring and let a variable C be the corresponding censoring lifetime. The observed data are thus denoted by (t_i, δ_i), such that $t_i = \min(T_i, C_i)$ and $\delta_i = I(T_i = t_i), i = 1, \ldots, n$ with $I(\cdot)$ defined as the usual binary indicator function. Suppose that there are k distinct ordered times $y_1 < \ldots < y_k$ at which death occur, where $(k \leq n)$, and $d_j = \sum_{i=1}^{n} I(t_i = y_j, \delta_j = 1)$ represents the number of deaths at y_j. In this way, the Kaplan-Meier estimator of $S(t)$ is expressed by,

$$\widehat{S}(t) = \prod_{j:y_j \leq t} \left(1 - \frac{d_j}{n_j}\right)$$

where $n_j = \sum_{i=1}^{n} I(t_j \geq y_j)$ is the number of individuals at risk in lifetime y_j, which is the number of individuals alive and uncensored in t_j.

3.2.2 Nonparametric hazard function

The classic approach adopted to estimate $\lambda(t)$ is estimating the cumulative hazard $\Lambda(t)$ which provides the accumulated failure rate of the individual until time t using the Kaplan-Meier estimator. In this way, the estimator of the cumulative hazard at the ordered failure time, is given by,

$$\widehat{\Lambda}(t) = -\sum_{j:y_j \leq t} \log\left(1 - \frac{d_j}{n_j}\right)$$

where n_j is the number of individuals at risk in lifetime y_j and d_j is the number of censored individuals, which are less than t_j.

Several procedures are found in the literature to obtain an $\lambda(t)$ curve by means of $\Lambda(t)$ estimates. In particular Jarjoura (1988) proposed the smoothing hazard rates with cubic splines and Rebora et al. (2014) used B-splines to estimate the shape of the hazard within the generalized linear mixed models framework and Smoothness is controlled by imposing an autoregressive structure on the baseline hazard coefficients.

In this thesis for the estimation of the nonparametric hazard function it was considered the algorithm proposed by Rebora et al. (2014) available in package "bshazard".

3.2.3 The standard Weibull distribution

Let T be a random variable representing the time to some event of interest. The survival function of the standard Weibull distribution with two parameters, introduced by Fréchet (1927) and studied in detail by Weibull et al. (1951), is given by,

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

where $t \geq 0$, $\alpha > 0$ and $\beta > 0$. The corresponding probability density function, is expressed by,

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

with hazard function given by,

$$h(t) = \alpha\beta t^{\beta-1},$$

where the hazard function may be of increasing or decreasing shapes depending on $\beta > 1$ or $\beta < 1$, respectively.

The standard Weibull distribution presents two particular forms:

(1) When $\beta = 1$, it is obtained the exponential distribution, where the survival, probability density and hazard function are given respectively by,

$$S(t) = \exp(-\alpha t), \quad f(t) = \alpha \exp(-\alpha t) \quad \text{and} \quad h(t) = \alpha.$$

Note that the exponential distribution have hazard function with constant form.

(2) For $\beta = 2$, the standard Weibull distribution is reduced to a Rayleigh distribution (Rayleigh and Strutt, 1919; Siddiqui, 1962), such that the survival, probability density and hazard function are given respectively by,

$$S(t) = \exp\left(-\alpha t^2\right), \quad f(t) = 2\alpha t \exp\left(-\alpha t^2\right) \quad \text{and} \quad h(t) = 2\alpha t.$$

It is obvious that Rayleigh distribution have hazard function with increasing form.

Figure (1) shows the probability density function, the survival function and the hazard function of the standard Weibull distribution, considering different values for the parameters α and β.

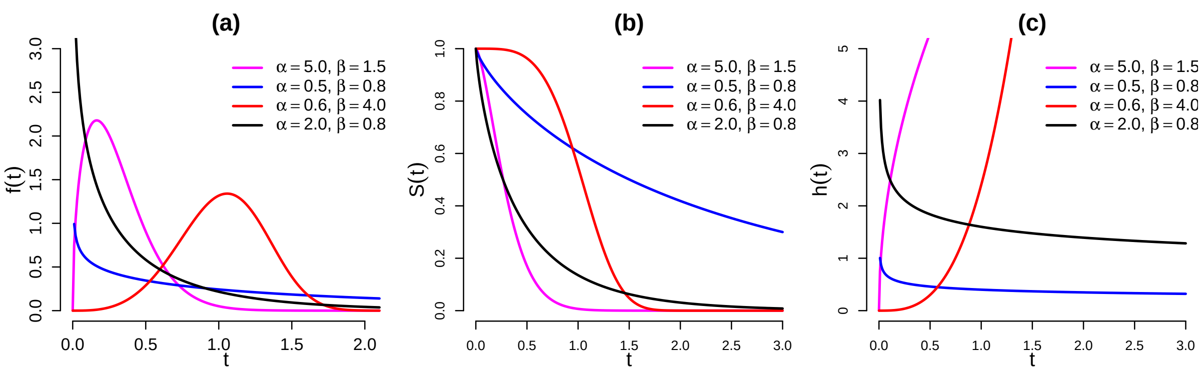

Figure 1 – The probability density function (a), survival function (b) and hazard function (c) of the standard Weibull distributions for some values of α and β.

3.2.4 The modified Weibull distribution

Let T be a random variable denoting the time to the occurrence of some event of interest. The survival function of modified Weibull distribution with three parameters, introduced by Lai et al. (2003), is given by,

$$S(t) = \exp(-\alpha t^{\beta} e^{\lambda t}),$$

where $t \geq 0$, $\alpha > 0$, $\beta > 0$ and $\lambda > 0$. The corresponding probability density, is given by,

$$f(t) = \alpha t^{\beta-1}(\beta + \lambda t) \exp(\lambda t - \alpha t^{\beta} e^{\lambda t}),$$

and the hazard function is expressed by,

$$h(t) = \alpha t^{\beta-1}(\beta + \lambda t) e^{\lambda t}.$$

The modified Weibull distribution includes as special sub-models three well-known distributions: (I) when $\lambda = 0$, it is obtained the standard Weibull distribution, (II) for $\lambda = 0$ and $\beta = 1$ the modified Weibull distribution is reduced to a exponential distribution, and (III) if $\lambda = 0$ and $\beta = 2$, it is obtained the Rayleigh distribution.

An important characteristic of the MW distribution is the flexibility of its hazard function. As observed by Lai et al. (2003), if $\beta > 1$ in the hazard function of modified Weibull distribution has an increasing form. However, if $0 < \beta < 1$ the hazard function assumes a bathtub shape. Figure (2) illustrates the probability density function, the survival function and the hazard function of the modified Weibull distribution, considering different values for the parameters α, β and λ.

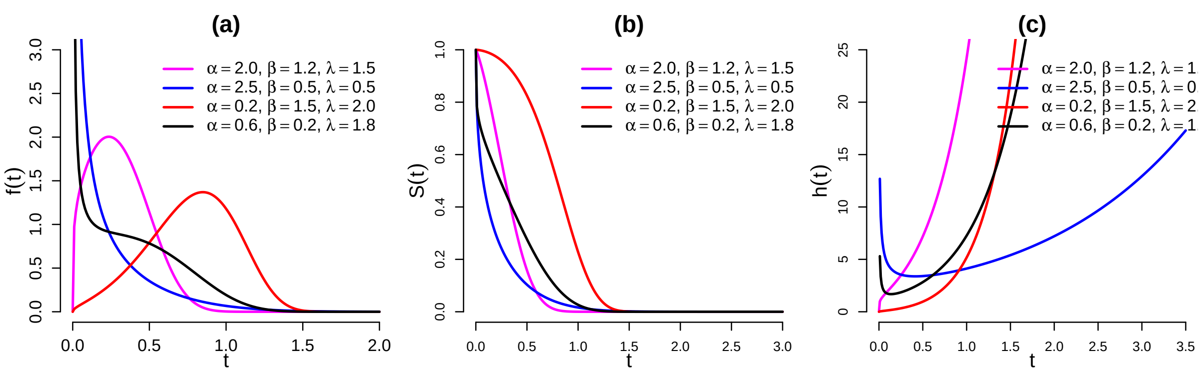

Figure 2 – The probability density function (a), survival function (b) and hazard function (c) of the modified Weibull distributions for some values of α, β and λ.

3.2.5 Bivariate Kaplan-Meier estimator

Several authors suggested different non-parametric estimators for the bivariate survival curve, such as Cambell and Földes (1980), Hanley and Parnes (1983), Tsai et al.

(1986), Dabrowska et al. (1988), Prentice and Cai (1992), Van Der Laan (1997), Akritas and Keilegom (2003), Bae et al. (2005),Prentice (2014) and Meira-Machado et al. (2016). However, these estimators are very complicated, need complex algorithms and difficult for practical use. In this thesis, it was considered the Lin and Ying (1993) estimator, due to its simple form and good numerical performance (Bae et al., 2005).

Let (T^0_{1i}, T^0_{2i}), $i = 1, ..., n$, be pairs of n independent and identically distributed lifetimes with joint survival function $S(t_1, t_2) = P(T^0_1 \geq t_1, T^0_2 \geq t_2)$ and let C_i, $i = 1, ..., n$, be random variables from n independent and identically distributed censoring times following an univariate survival function $G(t) = P(C \geq t)$. It is assumed that C_i are independent of the pairs of lifetimes (T^0_{1i}, T^0_{2i}), $i = 1, ..., n$. Thus, the observed bivariate lifetime is $(T_{1i}, T_{2i}, \delta_{1i}, \delta_{2i})$, $i = 1, ..., n$, where $T_{1i} = T^0_{1i} \wedge C_i$, $T_{2i} = T^0_{2i} \wedge C_i$, $\delta_{1i} = I(T^0_{1i} \leq C_i)$, $\delta_{2i} = I(T^0_{2i} \leq C_i)$, $I(.)$ denotes the indicator function, $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$. Thus, Lin and Ying (1993) introduced an estimator of the joint survival function $S(t_1, t_2)$, given by the following expression,

$$\widehat{S}(t_1, t_2) = \frac{\sum_{i=1}^{n} I(T_{i1} \geq t_1, T_{i2} \geq t_2)}{n \prod_{i:T_{1i} \wedge T_{2i} < t_1 \wedge t_2} \left(\frac{n^c_i - \delta^c_i}{n^c_i}\right)}, \tag{3.3}$$

where, $\delta^c_i = 1 - \delta_{1i}\delta_{2i}$ and $n^c_i = \sum_{j=1}^{n} I(T_{1j} \wedge T_{2j} \geq T_{1i} \wedge T_{2i})$.

3.3 Cure rate models

In the analysis of time-to-event data, particularly in cancer research, can to be expected that a fraction of subjects will not experience the event of interest. In these situations, there are a fraction of individuals not expecting the occurrence of the event of interest, that is, these individuals are not at risk ("long-term survivors" or "cured individuals"). This fraction of individuals is called cure rate or cured fraction. Following Vahidpour (2016), in the literature two approaches are very commonly used for cure models: the mixture cure rate models, and non-mixture rate models. The frequent way to model survival data with cured fraction is to assume that the studied population is a mixture of susceptible individuals who experience the event of interest and non-susceptible individuals that supposedly will never experience it. Statistical methods have been developed to analyze such data, like as Yu et al. (2004), Lambert et al. (2006) and Lambert (2007). Alternatively, a non-mixture formulation has been suggested by several authors like as Tsodikov et al. (2003) and Achcar et al. (2012). This model defines an asymptote for the cumulative hazard and hence for the cure fraction.

The use of defective distributions is an alternative to these models, in this case, a correspondent survival function $S(t)$ based on a defective distribution converges to a plateau at a value ρ, where ρ is associated to the cured fraction. Applications of defective distributions in analysis of lifetime data with a cure fraction can be seen in Cancho and

Bolfarine (2001), Balka et al. (2011), da Rocha et al. (2014), dos Santos et al. (2017), Rocha et al. (2017) and Martinez and Achcar (2017), among others.

Other approaches to model data with cure rate can be found in Peng et al. (1998) Rodrigues et al. (2009), Rodrigues et al. (2011), Gallardo et al. (2018), Pescim et al. (2019), Amico et al. (2019) and Leão et al. (2020).

3.3.1 Mixture and non-mixture cure rate models

Following Maller and Zhou (1996), a common approach to model the cure rate is to consider that the study population is a mixture of individuals susceptible to the event of interest and non-susceptible individuals. In this model, there is a parameter to measure the proportion of non-susceptible individuals. Let T be a random variable that represents the lifetimes, and $t > 0$ is an observation of T. Thus, the standard cure rate model (mixture cure rate model) has the survival function, given by,

$$S(t) = \rho + (1-\rho)S_0(t) \tag{3.4}$$

where $\rho \in (0,1)$ is the parameter that represents the proportion of individuals which are not susceptible to the event of interest. It represents the proportion of "long-term survivors", "non-susceptible" or "cured patients", and $S_0(t)$ denotes the basal survival function for the non-cured or susceptible group in the population, which can be any parametric survival function, like a standard Weibull or modified Weibull distributions. Note that $\lim_{t\to\infty} S(t) = \rho$, that is, the survival function has an asymptote at ρ. The probability density and the hazard functions corresponding to mixture cure rate model are given by,

$$f(t) = (1-\rho)f_0(t) \tag{3.5}$$

and,

$$h(t) = \frac{(1-\rho)f_0(t)}{\rho + (1-\rho)S_0(t)} \tag{3.6}$$

Alternatively, the non-mixture model has been proposed in the literature defining an asymptote for the cumulative hazard and thus for the cure rate (see, Tsodikov et al., 2003). In this case, the survival function for the non-mixture cure rate model is expressed by,

$$S(t) = \rho^{F_0(t)} = \exp\{\ln(\rho)F_0(t)\} \tag{3.7}$$

where $\rho \in (0,1)$ is the probability of cured patients and $F_0(t) = 1 - S_0(t)$ denotes a basal distribution function for the non-cured or susceptible individuals in the population. The probability density and the hazard functions corresponding to (3.7) are given respectively, by,

$$f(t) = -\ln(\rho)f_0(t)\exp\{\ln(\rho)F_0(t)\} \tag{3.8}$$

and,

$$h(t) = -\ln(\rho) f_0(t), \tag{3.9}$$

where, $f_0(t)$ is the basal density function for the non-cured or susceptible individuals in the population.

3.3.2 The defective Gompertz distribution

Models based on defective distributions are alternatives to the mixture model and non-mixture cure rate models. A defective distribution is defined as a distribution that is not integrated to one, integrates less than 1 for some values of their parameters, and thus it can be fitted to data including susceptible and non-susceptible individuals without directly include the parameter ρ.

The defective Gompertz distribution was introduced by Cantor and Shuster (1992) and subsequently studied by Gieser et al. (1998) and dos Santos et al. (2017). This model presents the following survival function

$$S(t) = \exp\left\{-\frac{\alpha}{\beta}\left[1 - \exp(-\beta t)\right]\right\}, \tag{3.10}$$

where $t, \alpha, \beta > 0$, being α the shape parameter and β the scale parameter. For this distribution a cure fraction ρ is given by

$$\rho = \lim_{t\to\infty} S(t) = \exp\left\{-\frac{\alpha}{\beta}\right\}. \tag{3.11}$$

We also have that the respective probability density and hazard function are given by

$$f(t) = \alpha \exp(-\beta t) \exp\left\{-\frac{\alpha}{\beta}\left[1 - \exp(-\beta t)\right]\right\} \tag{3.12}$$

and,

$$h(t) = \alpha \exp(-\beta t). \tag{3.13}$$

Note that the hazard function has decreasing shape. Figure (3) illustrates the probability density function, the survival function and the hazard function of the defective Gompertz distribution, considering different values for the parameters α and β. Note, for example, that when $\alpha = 2.0$ and $\beta = 1.5$, it is obtained $\rho = \exp\{-2.0/1.5\} = 0.2636$. This is the asymptote value of $S(t)$ for these parameters values.

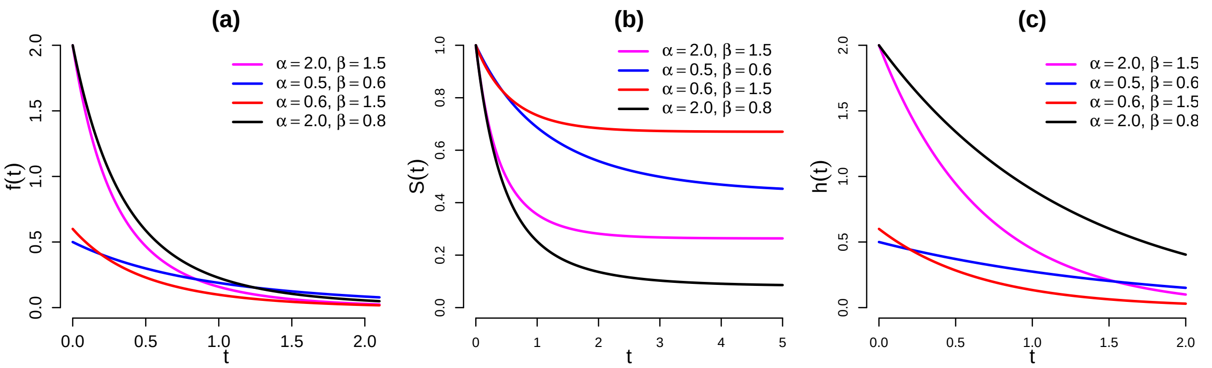

Figure 3 – The probability density function (a), survival function (b) and hazard function (c) of the defective Gompertz distribution for some values of α and β.

3.4 Copula Functions

Copula functions are used to describe the dependence structure between continuous random variables. Copula functions combines different univariate probability marginal distributions to generate multivariate distributions. The theoretical basis for the application of copulas is provided by the Sklar's theorem (Sklar, 1959, 1996; Trivedi and Zimmer, 2007), that states that a m-dimensional copula is a function C from $[0,1]^m$ to the interval $[0,1]$, satisfying the following conditions:

(I) $C(1,...,1,a_n,1,...,1) = a_n$ for every $n < m$ and all a_n in $[0,1]$;

(II) $C(a_1,...,a_m) = 0$ if $a_n = 0$ for any $n \leq m$;

(III) C is m-increasing

Considering a m-variate function F, the respective copula is a function $C : [0,1]^m \rightarrow [0,1]$ that satisfies

$$F(y_1,...,y_m) = C(F_1(y_1),...,F_m(y_m);\phi) = C_\phi(F_1(y_1),...,F_m(y_m)), \tag{3.14}$$

where ϕ is a parameter which measures the dependence between univariate probability marginal distributions. When the random variables are independent, we have

$$C_\phi(F_1(y_1),...,F_m(y_m)) = \prod_{i=1}^{m} F_i\,(y_i)\,. \tag{3.15}$$

In the special bivariate case ($m = 2$) and in the context of lifetime data, let (T_1,T_2) be the paired failure times with observations given by (t_1,t_2). In addition, $S_j(t_j) = P(T_j > t_j)$ and $f_j(t_j)$ are respectively the marginal survival functions and the marginal density functions of T_j, $j = 1, 2$. The joint distribution function of the lifetime is expressed by,

$$\begin{aligned} F(t_1, t_2) &= P(T_1 \leq t_1, T_2 \leq t_2) \\ F(t_1, t_2) &= 1 - P(T_1 > t_1) - P(T_2 > t_2) + P(T_1 > t_1, T_2 > t_2) \\ F(t_1, t_2) &= 1 - S_1(t_1) - S_2(t_2) + S(t_1, t_2), \end{aligned} \tag{3.16}$$

where $S(t_1, t_2) = P(T_1 > t_1, T_2 > t_2)$ is the joint survival function of (T_1, T_2). From the equation (3.16), we have,

$$S(t_1, t_2) = S_1(t_1) + S_2(t_1) + F(t_1, t_2) - 1. \tag{3.17}$$

In order to simplify the notation, let us consider that $u = F_1(t_1)$ and $v = F_2(t_2)$. From (3.14), the joint distribution function of (T_1, T_2) is given by

$$F(t_1, t_2) = C_\phi(F_1(t_1), F_2(t_2)) = C_\phi(u, v).$$

Thus, the joint density functions is given by,

$$f(t_1, t_2) = f_1(t_1) f_2(t_2) c_\phi(u, v), \tag{3.18}$$

where $f_1(t_1)$ and $f_2(t_2)$ are the marginal densities functions and $c_\phi(u, v)$ is the copula density function expressed by,

$$c_\phi(u, v) = \frac{\partial^2 C_\phi(u, v)}{\partial v \partial u}.$$

Copula functions allow measuring the correlation between lifetimes T_1 and T_2 by means of the parameter ϕ. Usually, it is considered two usual correlation measures between two random variables, given by the Kendall's tau (τ_k) and Spearman's rho (τ_s), which can be expressed respectively by the equations

$$\begin{aligned} \tau_k &= 4E[C_\phi(U, V)] - 1 \\ &= 4 \int_0^1 \int_0^1 C_\phi(u, v) dC_\phi(u, v) - 1 \\ &= 4 \int_0^1 \int_0^1 C_\phi(u, v) c_\phi(u, v) \mathrm{d}u \mathrm{d}v - 1 \\ &= 1 - 4 \int_0^1 \int_0^1 \frac{\partial C_\phi(u, v)}{\partial u} \frac{\partial C_\phi(u, v)}{\partial v} \mathrm{d}u \mathrm{d}v \end{aligned} \tag{3.19}$$

and

$$\begin{aligned} \tau_s &= 12E(UV) - 3 \\ &= 12 \int_0^1 \int_0^1 uv dC_\phi(u, v) - 3 \\ &= 12 \int_0^1 \int_0^1 C_\phi(u, v) \mathrm{d}u \mathrm{d}v - 3 \\ &= 3 - 12 \int_0^1 \int_0^1 u \frac{\partial C_\phi(u, v)}{\partial u} \mathrm{d}u \mathrm{d}v. \end{aligned} \tag{3.20}$$

For more details on these expressions, we can refer to Schweizer et al. (1981), Joe (1997), Embrechts et al. (2001), Nelsen (2007), Joe (2014) and Durante and Sempi (2015). At every choice of the $C_\phi(u, v)$ the expressions above can sometimes be calculated in closed form. In other cases, it may be necessary to compute the measures τ_k and τ_s by numerical integration or Monte Carlo simulation methods.

The literature presents several copula functions divided into some families. The families of copula functions differ in terms of dependence structure they represent. In this thesis, several copula functions are considered, the main families of these copulas are briefly described as follows:

(1) Archimedean Copulas: The copula that can be written in the form

$$C(u, v) = \varphi^{-1}(\varphi(u) + \varphi(v)),$$

are called Archimedean copula with generating function $\varphi(x)$, such that, $\varphi(x)$: $[0, 1] \rightarrow [0, \infty)$ is a real valued function satisfying the following conditions:

- $\varphi(0) = 0$;
- $\lim_{x\to 0} \varphi(x) = \infty$;
- $\dfrac{d\varphi}{dx} < 0$ for all $x \in (0, 1)$ and
- $\dfrac{d^2\varphi}{dx^2} > 0$ for all $x \in (0, 1)$.

The Archimedean copulas were extensively studied and popularized by Genest and MacKay (1986).

(2) Extreme-Value Copulas: A copula defined by the expression

$$C(u, v) = \exp\left[\log(uv)\varphi\left(\frac{\log(u)}{\log(uv)}\right)\right],$$

is an extreme-value copula and $\varphi(x) : [0, 1] \rightarrow [1/2, 1]$ is a convex function satisfying $\max(x, 1 - x) \leq \varphi(x) \leq 1$ for all $x \in [0, 1]$ (Pickands, 1981).

(3) FGM Copulas: this copula family consists of the generalizations of the Farlie-Gumbel-Morgenstern (FGM) copula. Some of these generalizations are presented in the following sections.

3.4.1 Generalized Farlie-Gumbel-Morgenstern (GFGM) copula

The generalized Farlie-Gumbel-Morgenstern (GFGM) copula was introduced by Bairamov and Kotz (2002). Its expression is given by

$$C_\phi(u, v) = uv\left[1 + \phi\left(1 - u^p\right)^q\left(1 - v^p\right)^q\right],$$

where $p, q \geq 1$ are additional parameters (Shih and Emura, 2016). The possible range of the dependence parameter ϕ is given by,

$$-\min\left\{1, \frac{1}{p^{2q}}\left(\frac{1+pq}{q-1}\right)^{2(q-1)}\right\} \leq \phi \leq \frac{1}{p^q}\left(\frac{1+pq}{q-1}\right)^{q-1}$$

and the respective copula density function is given by

$$\begin{aligned} c_\phi(u,v) &= 1+\phi\,(1-u^p)^{q-1}\left[1-(1+pq)\,u^p\right]\times \\ &\times\ (1-v^p)^{q-1}\left[1-(1+pq)\,v^p\right]. \end{aligned}$$

The relations between the parameter ϕ, and the Kendall and Spearman correlation measures are given respectively by,

$$\tau_k = 8\phi\left[\frac{q}{2+pq}B\left(\frac{2}{p}, q\right)\right]^2$$

and

$$\tau_s = 12\phi\left[\frac{q}{2+pq}B\left(\frac{2}{p}, q\right)\right]^2,$$

where $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ is the beta function and $\Gamma(a)$ is the gamma function.

In Figure (4),it is exhibited the contour plots considering the GFGM copula for different values of ϕ, p and q.

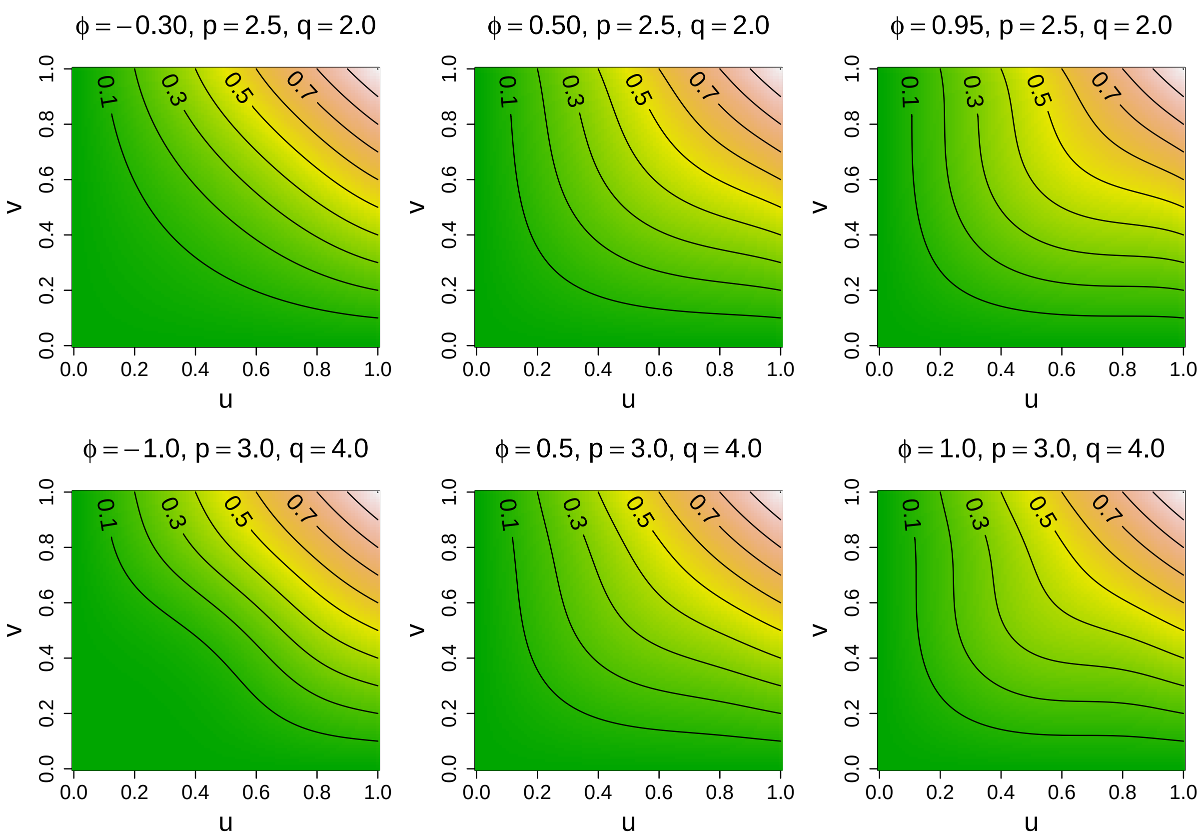

Figure 4 – Contour plots of the GFGM copula considering some values for ϕ, p and q.

3.4.2 Type 1 Huang–Kotz FGM (HKFGM1) copula

The HKFGM1 copula is a special case of the GFGM copula, where $p = 1$ in equation (3.4.1) (Huang and Kotz, 1999). In this case, the possible range of the dependence parameter ϕ is

$$-1 \leq \phi \leq \left(\frac{q+1}{q-1}\right)^{q-1}.$$

The relations between the parameter ϕ, and the Kendall and Spearman correlation measures are given respectively by,

$$\tau_k = 8\phi\left[\frac{1}{(2+q)(1+q)}\right]^2$$

and

$$\tau_s = 12\phi\left[\frac{1}{(2+q)(1+q)}\right]^2.$$

Considering the HKFGM1 copula, Figure (5) shows the contour plots for some values of ϕ and q.

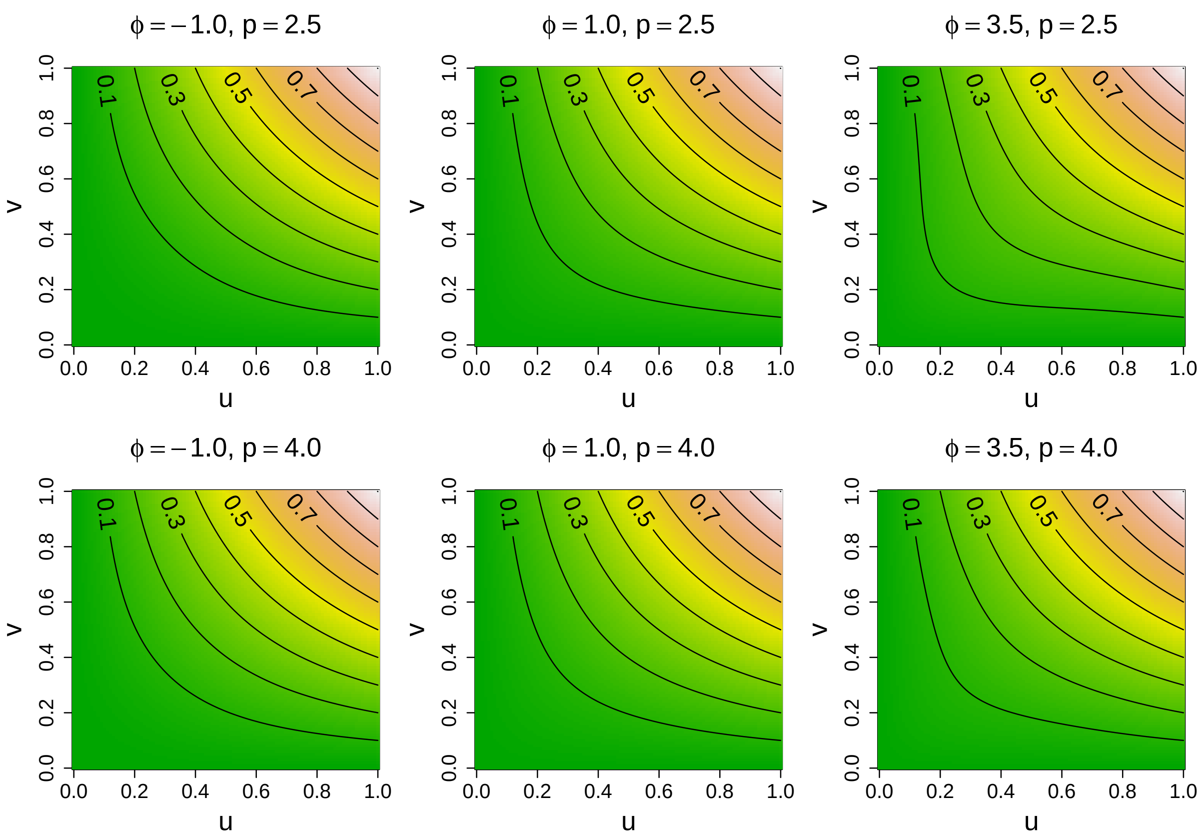

Figure 5 – Contour plots of the HKFGM1 copula considering different values for ϕ and q.

3.4.3 Type 2 Huang–Kotz FGM (HKFGM2) copula

Like the HKFGM1, the HKFGM2 copula is a special case of the GFGM copula, but in this case $q = 1$ in the equation (3.4.1) (Huang and Kotz, 1999). For this copula, the possible range of the dependence parameter ϕ is

$$-\left(\max\{1, p\}\right)^{-2} \leq \phi \leq p^{-1}.$$

The relations between the parameter ϕ, and the Kendall and Spearman correlation measures are given respectively by,

$$\tau_k = 8\phi \left[\frac{p}{2(2+p)}\right]^2$$

and

$$\tau_s = 12\phi \left[\frac{p}{2(2+p)}\right]^2.$$

Figure (6) shows the contour plots considering the HKFGM2 copula for different values of ϕ and fix $p = 2.5$.

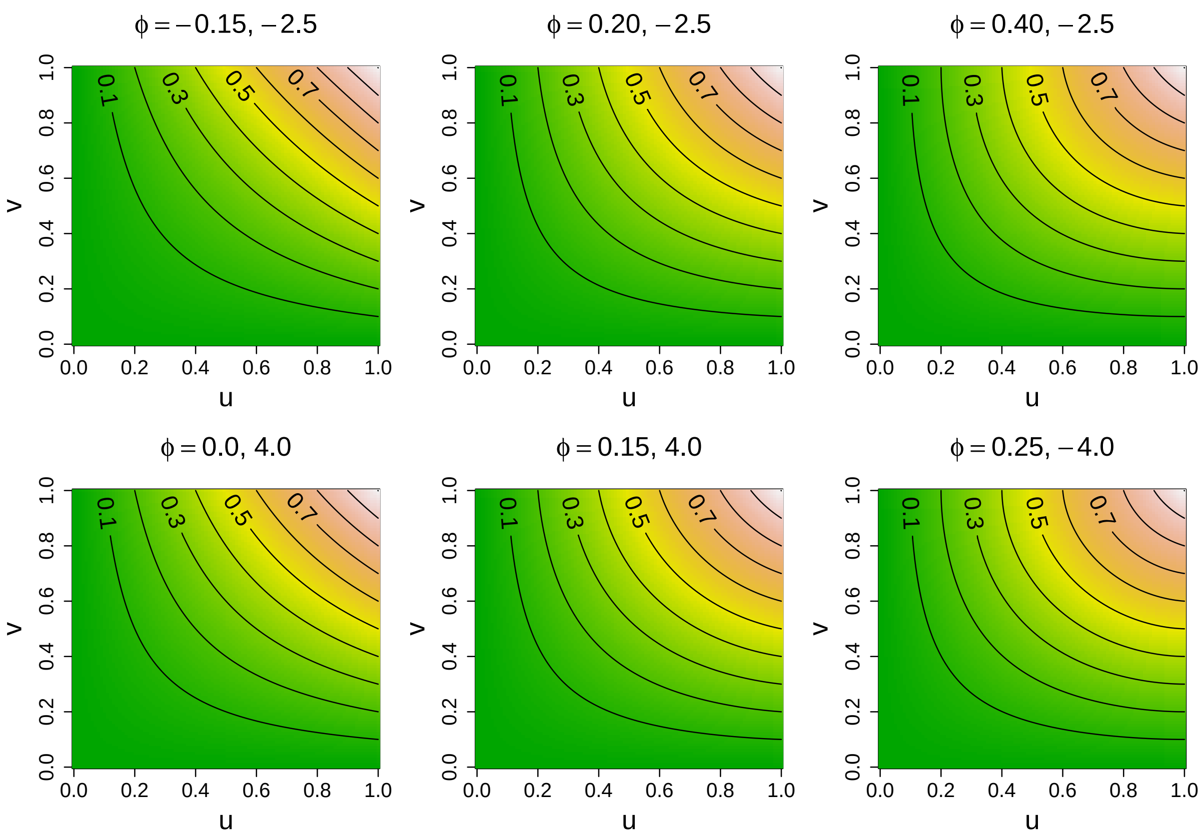

Figure 6 – Contour plots of the HKFGM2 copula considering different values for ϕ, with $p = 2.5$.

3.4.4 Farlie-Gumbel-Morgenstern (FGM) copula

The Farlie–Gumbel–Morgenstern copula was proposed by Morgenstern (1956) and further studied by Gumbel (1960a) and Farlie (1960). This copula is a special case of GFGM, obtained considering $p = q = 1$ in (3.4.1); in this way,

$$C_\phi(u, v) = uv\left[1 + \phi\,(1 - u)\,(1 - v)\right],$$

where $-1 \leq \phi \leq 1$. When $\phi = 0$, the joint survival function is reduced to $C_\phi(u, v) = uv = F_{01}(t_1)F_{02}(t_2)$, suggesting independence between T_1 and T_2. For the FGM copula, the Kendall and Spearman correlation measures are respectively given by,

$$\tau_k = \frac{2}{9}\phi$$

and

$$\tau_s = \frac{1}{3}\phi.$$

This copula measures only low dependence between T_1 and T_2, due $-2/9 \leq \tau_k \leq 2/9$ and $-1/3 \leq \tau_s \leq 1/3$, see (Nelsen, 1994). The respective copula density function is given by

$$c_\phi(u, v) = 1 + \phi\,(1 - 2u)\,(1 - 2v)\,.$$

Figure (7) shows the contour plots considering the FGM copula for different values of ϕ.

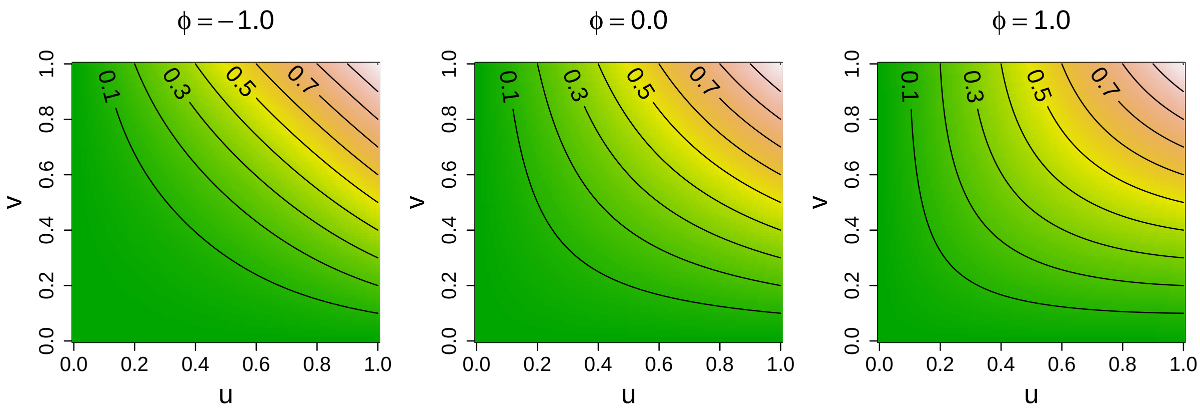

Figure 7 – Contour plots of the FGM copula for some values for ϕ.

3.4.5 Fischer and Köck FGM (FKFGM) copula

Fischer and Köck (2012) proposed some extensions of the FGM copula. One of them is expressed by,

$$C_\phi(u, v) = uv\left[1 + \phi\left(1 - u^{\frac{1}{p}}\right)\left(1 - v^{\frac{1}{p}}\right)\right]^p,$$

where $p \geq 1$ and $-1 \leq \phi \leq 1$. When $\phi = 0$ there is indication of independence between T_1 and T_2. Its copula density function is given by,

$$\begin{aligned} c_\phi(u,v) &= [k_\phi(u,v)]^p + [k_\phi(u,v)]^{p-1}\,\phi \left\{ \frac{u^{\frac{1}{p}} v^{\frac{1}{p}}}{p} - \left(1-u^{\frac{1}{p}}\right) v^{\frac{1}{p}} - \right. \\ &- \left. \left(1-v^{\frac{1}{p}}\right) u^{\frac{1}{p}} + \phi \left(1-u^{\frac{1}{p}}\right)\left(1-v^{\frac{1}{p}}\right) \frac{u^{\frac{1}{p}} v^{\frac{1}{p}}}{k_\phi(u,v)} \left[1 - \frac{1}{p}\right] \right\}, \end{aligned}$$

where

$$k_\phi(u,v) = 1 + \phi\left(1-u^{\frac{1}{p}}\right)\left(1-v^{\frac{1}{p}}\right).$$

The Kendall and Spearman correlation measures obtained by the relation with the parameter ϕ, do not have closed form. In Figure (8) it is exhibited the contour plots considering the FKFGM copula for different values of ϕ and p.

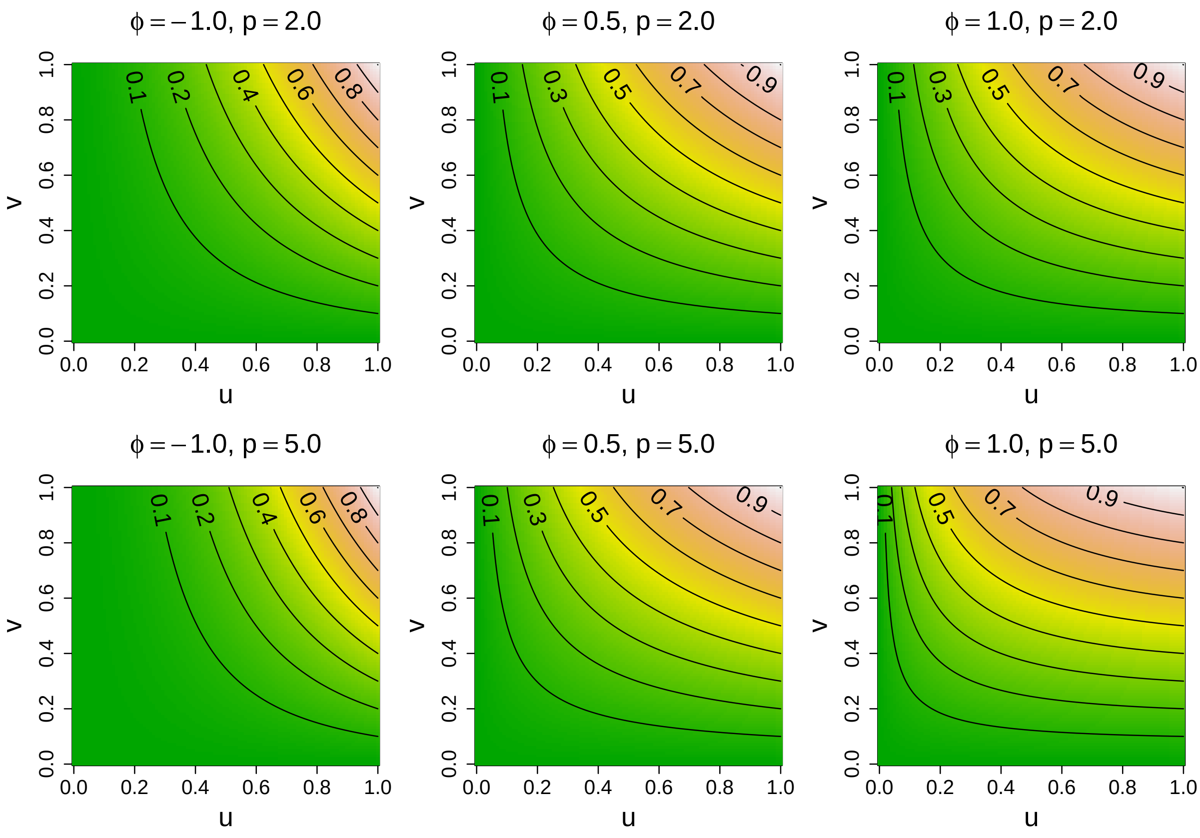

Figure 8 – Contour plots of the FKFGM copula considering different values for ϕ and p.

3.4.6 Clayton copula

The Clayton copula was first introduced by Clayton (1978) and subsequently studied by Cook and Johnson (1981) and Oakes (1982). This copula has form

$$C_\phi(u,v) = \left(u^{-\phi} + v^{-\phi} - 1\right)^{-\frac{1}{\phi}},$$

where $\phi > 0$. In this case, T_1 and T_2 become independent when ϕ tends to zero. For this copula function, the relationship between the dependence copula parameter ϕ and the Kendall tau is given by,

$$\tau_k = \frac{\phi}{\phi + 2}.$$

Note that, if ϕ tends to infinite, then τ tends to 1, indicating perfect positive dependence between T_1 and T_2. The expression for τ_s assuming the Clayton copula is very complicated. The respective copula density function is given by

$$c_\phi(u, v) = (1 + \phi)\,(uv)^{-(1+\phi)}\left(u^{-\phi} + v^{-\phi} - 1\right)^{-\left(\frac{1}{\phi}+2\right)}.$$

This copula belongs to the Archimedean family, with

$$\varphi(x) = \frac{1}{\phi}(x^{-\phi} - 1).$$

Figure (9) shows the contour plots considering the Clayton copula for different values of ϕ.

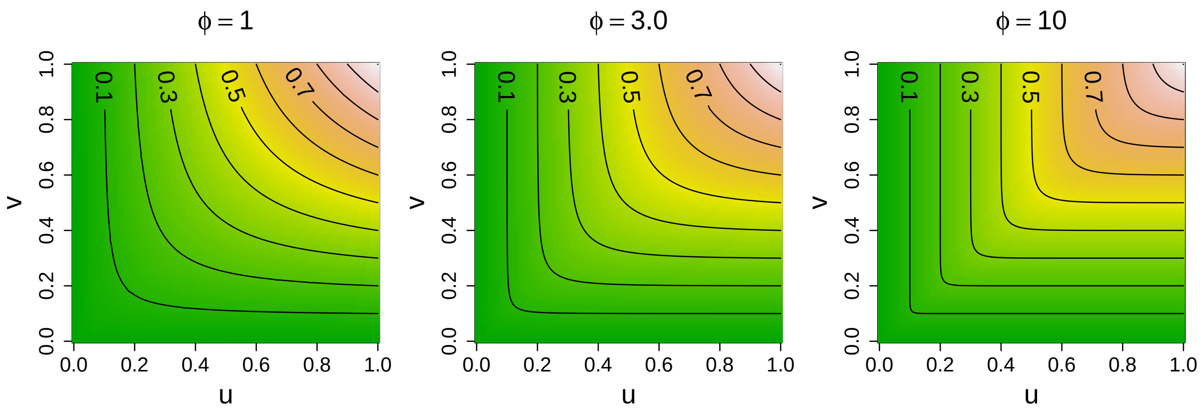

Figure 9 – Contour plots of the Clayton copula considering some values for ϕ.

3.4.7 Burr copula

The Burr copula introduced by Frees and Valdez (1998), is expressed by,

$$C_\phi(u, v) = u + v - 1 + \left[(1 - u)^{-\frac{1}{\phi}} + (1 - v)^{-\frac{1}{\phi}} - 1\right]^{-\phi},$$

where $\phi > 0$. The relationship between the dependence copula parameter ϕ and the Kendall correlation is

$$\tau_k = \frac{1}{2\phi + 1}.$$

When $\phi \to 0$ it is indicated a total dependence between T_1 and T_2. The copula density function for the Burr copula is given by

$$c_\phi(u,v) = \frac{(\phi+1)\,(1-u)^{-\left(1+\frac{1}{\phi}\right)}\,(1-v)^{-\left(1+\frac{1}{\phi}\right)}}{\phi\left[(1-u)^{-\frac{1}{\phi}} + (1-v)^{-\frac{1}{\phi}} - 1\right]^{2+\phi}}.$$

This copula does not belong to the families mentioned in this thesis. Figure (10) displays the contour plots considering the Burr copula for different values of ϕ.

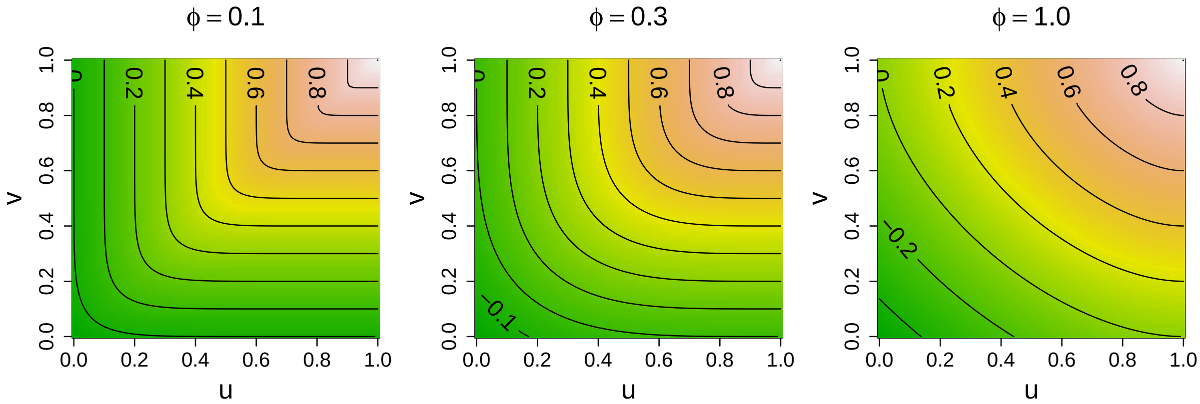

Figure 10 – Contour plots of the Burr copula considering some values for ϕ.

3.4.8 Gumbel-Hougaard (GH) copula

The Gumbel-Hougaard (GH) copula (Gumbel, 1960b; Hougaard, 1986), is defined as,

$$C_\phi(u,v) = \exp\left\{-\left[(-\log u)^\phi + (-\log v)^\phi\right]^{\frac{1}{\phi}}\right\},$$

where $\phi \geq 1$. When ϕ tends to 1, there is an indication of independence between T_1 and T_2, since $C_\phi(u,v) = uv$. This copula function is restricted for positive dependence. The relationship between ϕ and the Kendall correlation is given by,

$$\tau_k = 1 - \frac{1}{\phi}.$$

The density function for this copula, is given by,

$$\begin{aligned} c_\phi(u,v) &= \frac{C_\phi(u,v)}{uv\,(\log u)\,(\log v)}\left[(-\log u)^\phi + (-\log v)^\phi\right]^{-2+\frac{1}{\phi}} \times \\ &\times \left[(-\log u)\,(-\log v)\right]^\phi \left\{(\phi-1) + \left[(-\log u)^\phi + (-\log v)^\phi\right]^{\frac{1}{\phi}}\right\}. \end{aligned}$$

In Figure (11) it is showed the contour plots considering the GH copula for different values of ϕ.

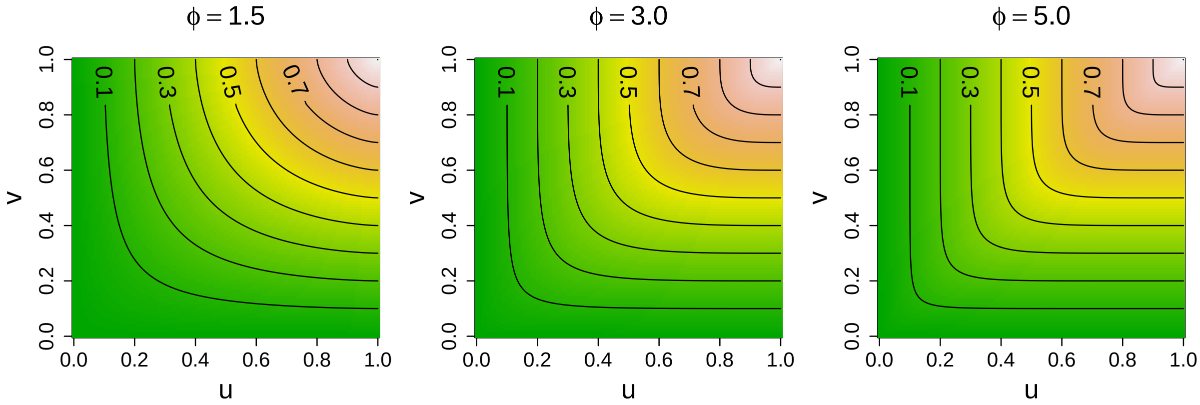

Figure 11 – Contour plots of the GH copula considering some values for ϕ.

This copula belongs to the Archimedean family, with

$$\varphi(x) = [-\log(x)]^{\phi}.$$

It is interesting to note that this copula is at the same time Archimedean and extreme-value, see Genest and Rivest (1989).

3.4.9 Gumbel–Barnett (GB) Copula

The Gumbel–Barnett (GB) copula studied by Gumbel (1960a) and Barnett (1980) is expressed by,

$$C_{\phi}(u, v) = uv \exp\left\{-\phi(\log u)(\log v)\right\},$$

where $0 < \phi \leq 1$. When ϕ tends to zero value, there is an indication of independence between the lifetime times T_1 and T_2. Following Fredricks and Nelsen (2007) the GB copula measure only negative dependence, and the relationship between the dependence copula parameter ϕ and the Kendall and Spearman correlations, are given respectively by

$$\tau_k = \exp\left(\frac{2}{\phi}\right) \mathrm{Ei}\left(-\frac{2}{\phi}\right),$$

and

$$\tau_s = -3 - \frac{12}{\phi}\left[\exp\left(\frac{4}{\phi}\right) \mathrm{Ei}\left(-\frac{4}{\phi}\right)\right],$$

where Ei(.) is the exponential integral given by

$$\mathrm{Ei}(x) = \int_{-\infty}^{x} \frac{\exp(t)}{t} \mathrm{d}t.$$

The copula density function for the Gumbel–Barnett copula is given by

$$c_{\phi}(u, v) = \frac{C_{\phi}(u, v)}{uv}\left\{\left[\phi^2(\log v) - \phi\right](\log u) - \phi\left[(\log v) + 1\right] + 1\right\}.$$

This copula belongs to the Archimedean family, such that

$$\varphi(x) = \log\left[1 - \phi \log(x)\right]$$

Figure (12) shows the contour plots considering the GB copula for different values of ϕ.

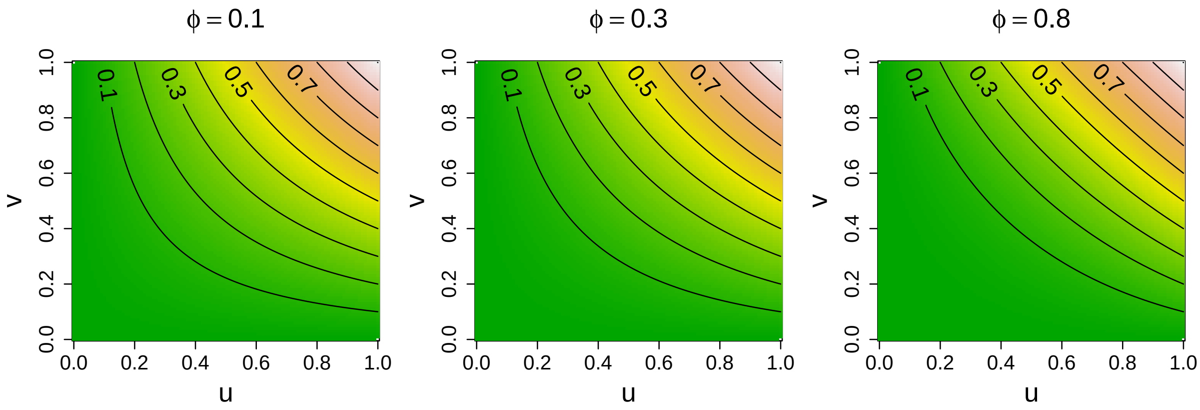

Figure 12 – Contour plots of the GB copula considering different some for ϕ.

3.4.10 Galambos Copula

The Galambos copula defined by Galambos (1975), is given by

$$C_\phi(u,v) = uv \exp\left(\left[(-\log u)^{-\phi} + (-\log v)^{-\phi}\right]^{-\frac{1}{\phi}}\right),$$

where $\phi > 0$. When $\phi \to 0$ we have independence between T_1 and T_2, and if $\phi \to \infty$ it is suggested high dependence between T_1 and T_2. The relationship between ϕ and the Spearman correlation is given by

$$\rho_s = 12 \int_0^1 \left\{2 - \left[t^{-\phi} + (1-t)^{-\phi}\right]^{-\frac{1}{\phi}}\right\}^{-2} \mathrm{d}t - 3,$$

and the relationship between ϕ and Kendall correlation is very complicated (Joe, 2014). The respective copula density function is given by

$$\begin{aligned}
c_\phi(u,v) &= \frac{\exp\left\{[h_\phi(u,v)]^{-\frac{1}{\phi}}\right\}}{(\log u)(\log v)[h_\phi(u,v)]^2} \left\{[h_\phi(u,v)]^{-\frac{2}{\phi}} (-\log u)^{-\phi} (-\log v)^{-\phi} + \right. \\
&+ \left[(\log v)(-\log u)^{-2\phi} + (-\log u)^{-\phi}(-\log v)^{-\phi} \times \right. \\
&\times \left. (\phi + (\log u) + (\log v) + 1) + (\log u)(-\log v)^{-2\phi}\right] \times \\
&\times \left. [h_\phi(u,v)]^{-\frac{1}{\phi}} + (\log u)(\log v)[h_\phi(u,v)]^2\right\}
\end{aligned}$$

where

$$h_\phi(u,v) = (-\log u)^{-\phi} + (-\log v\ )^{-\phi}.$$

The Galambos copula belongs to the extreme-value copula family, where

$$\varphi(x) = 1 - \left[x^{-\phi} + (1-x)^{-\phi}\right]^{-1/\phi}$$

Figure (13) exhibits the contour plots considering the Galambos copula for different values of ϕ.

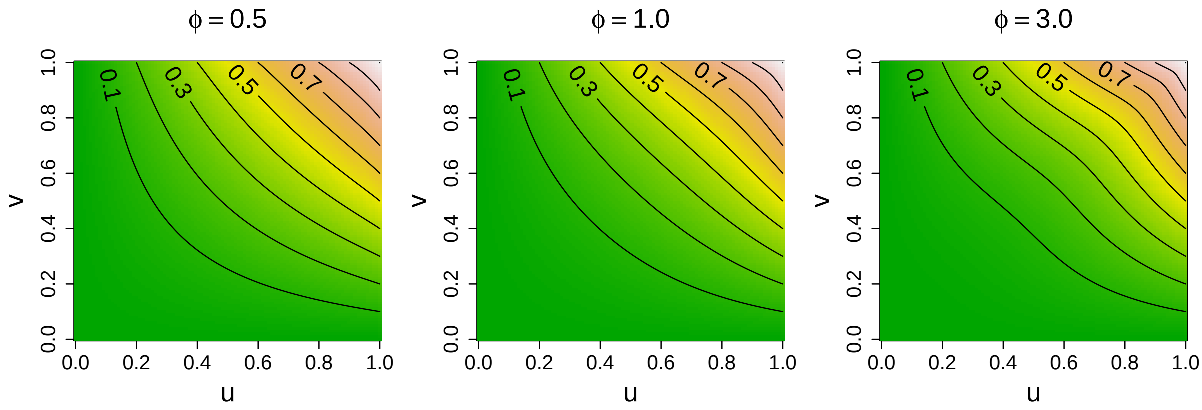

Figure 13 – Contour plots of the Galambos copula considering different values for ϕ.

3.4.11 Frank copula

The Frank copula (Frank, 1979) is given by

$$C_\phi(u,v) = -\frac{1}{\phi}\log\left\{1 + \frac{[\exp(-\phi u) - 1]\,[\exp(-\phi v) - 1]}{\exp(-\phi) - 1}\right\},$$

where $\phi > 0$. When ϕ tends to zero, there is an indication of independence between T_1 and T_2, and if ϕ tend to infinity there is an indication that T_1 and T_2 are correlated. According to Nelsen (1986) and Genest (1987) the relationship between the dependence copula parameter ϕ and the Kendall and Spearman correlation measures are given by

$$\tau_k = 1 + \frac{4}{\phi}\left[D_1(\phi) - 1\right],$$

and

$$\tau_s = 1 + \frac{12}{\phi}\left[D_2(\phi) - D_1(\phi)\right],$$

where $D_k(.)$ is the Debye function given by

$$D_k(\alpha) = \frac{k}{\alpha^k}\int_0^\alpha \frac{t^k}{\exp(t) - 1}\,\mathrm{d}t.$$

The copula density function for the Frank copula is given by

$$c_\phi(u,v) = \frac{\phi\exp(-\phi u - \phi v)\,(1 - \exp(-\phi))}{\left(\exp(-\phi) - \exp(-\phi u) - \exp(-\phi v) + \exp(-\phi u - \phi v)\right)^2}.$$

The Frank copula belongs to the Archimedean family, such that,

$$\varphi(x) = \log\left[\frac{\exp(-\phi x) - 1}{\exp(-\phi) - 1}\right].$$

Figure (14) shows the contour plots considering the Frank copula for different values of ϕ.

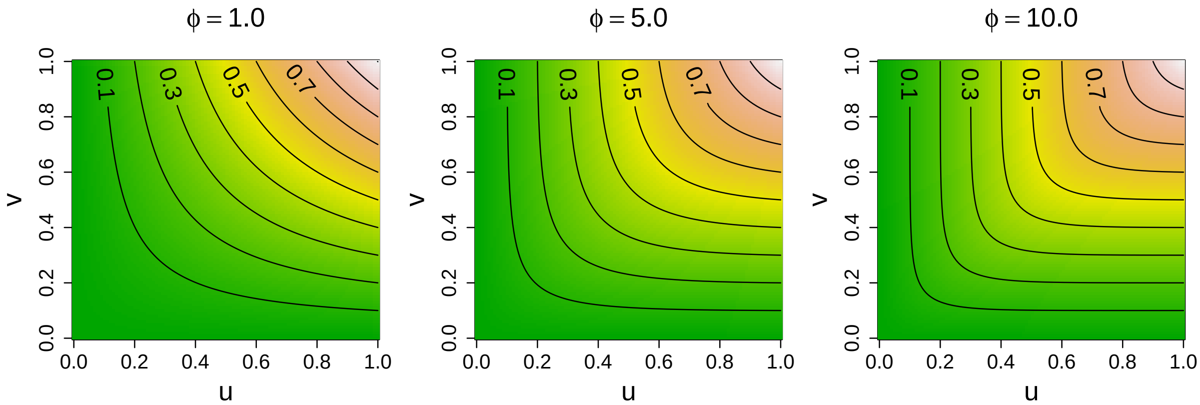

Figure 14 – Contour plots of the Frank copula considering some values for ϕ.

3.4.12 Ali-Mikhail-Haq (AMH) copula

The Ali-Mikhail-Haq (AMH) copula (Ali et al., 1978; Kumar, 2010) is given by

$$C_\phi(u, v) = \frac{uv}{1 - \phi\,(1-u)\,(1-v)},$$

where $-1 \leq \phi < 1$. Note that, the AMH copula measures both positive and negative dependence. In addition, $\phi = 0$ indicates total independence between T_1 and T_2, and in this case, we have $C_\phi(u, v) = uv$. The relationship between ϕ, and the Kendall and Spearman correlations are expressed respectively by

$$\tau_k = 1 - \frac{2}{3\phi} - \frac{2}{3\phi^2}\,(1-\phi)^2 \log(1-\phi)$$

and

$$\tau_s = \frac{12}{\phi^2}\text{dilog}(1-\phi)(1+\phi) - \frac{24}{\phi^2}\log(1-\phi)(1-\phi) - \frac{3}{\phi}(12+\phi),$$

where dilog(.) is the dilogarithm function defined by

$$\text{dilog}(x) = \int_1^x \frac{\log t}{1-t}\text{d}t.$$

Thus, it is noted that $-0.1817 \leq \tau_k < 0.3333$ and $-0.2710 \leq \rho < 0.4784$. The copula density function for the AMH copula function is expressed by

$$c_\phi(u, v) = \frac{1 - \phi + 2\phi C_\phi(u, v)}{[1 - \phi\,(1-u)\,(1-v)]^2}.$$

The AMH copula belongs to the Archimedean family, where,

$$\varphi(x) = -\log\left[\frac{1-\phi(1-x)}{x}\right].$$

In Figure (15) it is showed the contour plots considering the AMH copula for different values of ϕ.

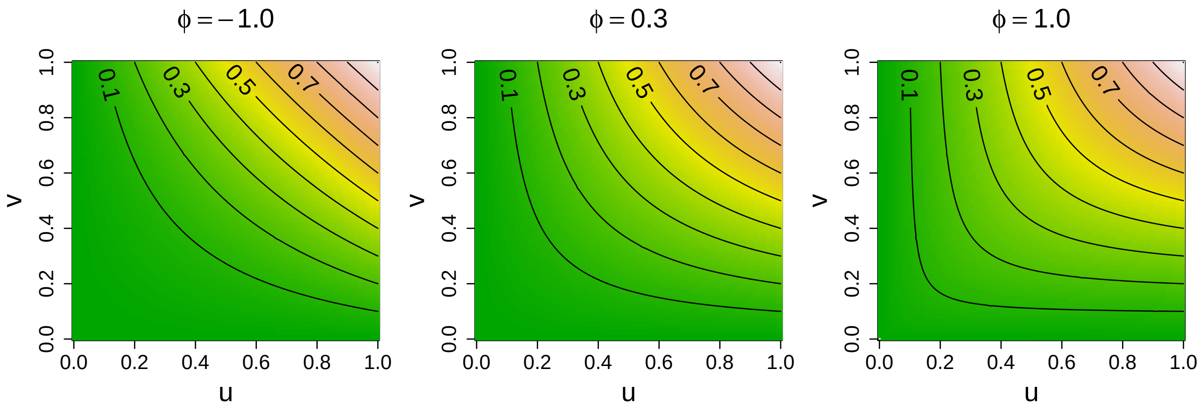

Figure 15 – Contour plots of the AMH copula considering some values for ϕ.

3.4.13 A12 copula

The A12 copula function is exhibited by Nelsen (2007) in Table 4.1 presented in page 116 of the book "An Introduction to Copulas". This copula is given by,

$$C_\phi(u,v) = \frac{1}{1+\left[\left(\frac{1}{u}-1\right)^\phi + \left(\frac{1}{v}-1\right)^\phi\right]^{\frac{1}{\phi}}},$$

where $\phi \geq 1$. In the A12 copula the relationship between the dependence copula parameter ϕ and the Kendall correlation is given by

$$\tau_k = 1 - \frac{2}{3\phi}$$

When ϕ tends to infinite, the τ_k tends to 1, indicating perfect positive dependence between the lifetimes T_1 and T_2 . In this copula it is observed that $\tau_k \in [1/3, 1]$. The copula density function for the A12 copula function is represented by,

$$c_\phi(u,v) = \frac{\left[g_\phi\left(u,v\right)\right]^{\frac{1}{\phi}-2}\left(\frac{1}{u}-1\right)^\phi\left(\frac{1}{v}-1\right)^\phi\left\{\left[g_\phi\left(u,v\right)\right]^{\frac{1}{\phi}}\left(\phi+1\right)+\phi-1\right\}}{vu\left\{1+\left[g_\phi\left(u,v\right)\right]^{\frac{1}{\phi}}\right\}^3\left(1-u\right)\left(1-v\right)}$$

where

$$g_\phi(u,v) = \left(\frac{1}{u}-1\right)^\phi + \left(\frac{1}{v}-1\right)^\phi. \tag{3.21}$$

The A12 copula belongs to the Archimedean family, with

$$\varphi(x) = \left(\frac{1}{x} - 1\right)^{\phi}.$$

In Figure (16) it is exhibited the contour plots considering the A12 copula for different values of ϕ.

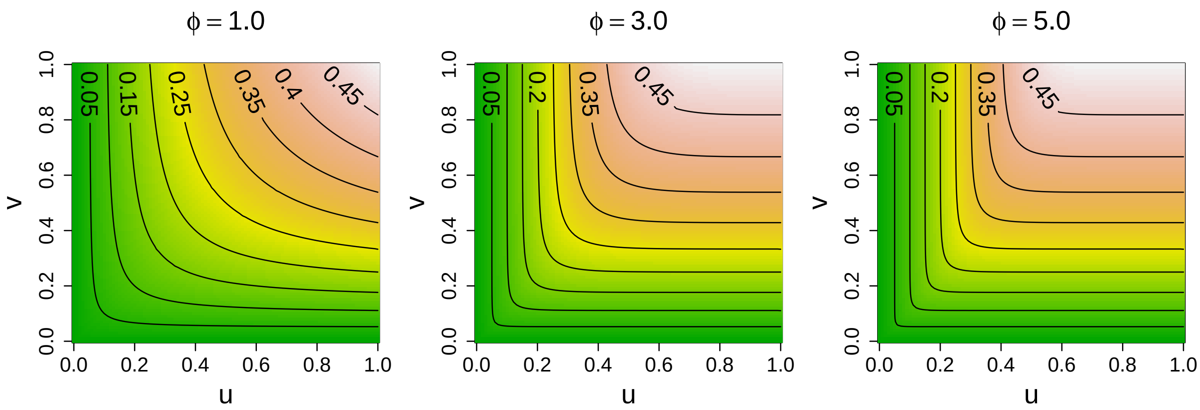

Figure 16 – Contour plots of the A12 copula considering some values for ϕ.

3.4.14 Joe copula

The Joe copula function (Joe, 1993) is represented by,

$$C_{\phi}(u,v) = 1 - \left[(1-u)^{\phi} + (1-v)^{\phi} - (1-u)^{\phi}(1-v)^{\phi}\right]^{\frac{1}{\phi}},$$

where $\phi > 1$. Joe (2014) shows that the Kendall's correlation measure for this copula function is expressed by

$$\tau_k = 1 + \frac{2}{2-\phi}\left[\text{digamma}(2) - \text{digamma}\left(\frac{2}{\phi} + 1\right)\right].$$

In the case where $\phi = 2$ we have that

$$\tau_k = 1 - \text{trigamma}(2),$$

where the digamma and trigamma functions are respectively given by, $\text{digamma}(x) = \frac{d}{dx}\log\Gamma(x)$, $\text{trigamma}(x) = \frac{d^2}{dx^2}\log\Gamma(x)$ where $\Gamma(x)$ is the gamma function. The copula density function for the Joe copula function is written by,

$$\begin{aligned} c_{\phi}(u,v) &= \left\{\left[1-(1-v)^{\phi}\right](1-u)^{\phi} + (1-v)^{\phi} + \phi - 1\right\} \times \\ &\times \left\{\left[1-(1-u)^{\phi}\right](1-v)^{\phi} + (1-u)^{\phi}\right\}^{\frac{1}{\phi}} \times \\ &\times [(1-u)(1-v)]^{\phi-1}\left\{\left[(1-v)^{\phi} - 1\right](1-u)^{\phi} - (1-v)^{\phi}\right\}^{-2}. \end{aligned}$$

The Joe copula belongs to the Archimedean family, where,

$$\varphi(x) = -\log\left[1-(1-x)^{\phi}\right].$$

In Figure (17) it is showed the contour plots considering the Joe copula for different values of ϕ.

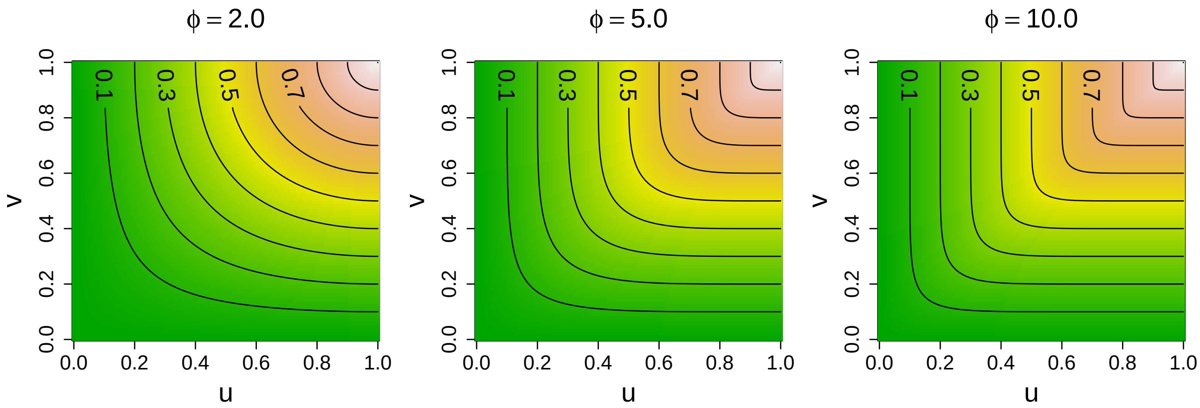

Figure 17 – Contour plots of the Joe copula considering some values for ϕ.

3.4.15 Plackett Copula

The Plackett copula (Plackett, 1965), is defined by,

$$C_{\phi}(u,v) = \frac{\left[1+(\phi-1)\left(u+v\right)\right]-\sqrt{\left[1+(\phi-1)\left(u+v\right)\right]^{2}-4\phi\left(\phi-1\right)uv}}{2(\phi-1)},$$

where $\phi > 0$. When $\phi = 1$ there is an indication that there is independence between T_1 and T_2, while if $0 < \phi < 1$ indicates negative correlation between the lifetimes. The relationship between ϕ and the Spearman correlation coefficient is given by,

$$\tau_s = \frac{\phi+1}{\phi-1} - \frac{2\phi\log\phi}{(\phi-1)^2}.$$

Note that, when $\phi \to 0$ then $\tau_s = -1$ and in case of $\phi \to \infty$ obtain $\tau_s = -0$. The copula density function for the Plackett copula function is expressed by,

$$c_{\phi}(u,v) = -\frac{2\phi\left[(\phi-1)(2u-1)v+u(1-\phi)-1\right]}{\left\{\left[1+(\phi-1)\left(u+v\right)\right]^{2}-4uv\phi\left(\phi-1\right)\right\}^{\frac{3}{2}}}.$$

This copula does not belong to the copula families mentioned in this thesis. In fact, it is considered in the literature that this copula function belongs to an own copula family (Nelsen, 2007; Balakrishnan and Lai, 2009). Figure (18) shows the contour plots considering the Plackett copula for different values of ϕ.

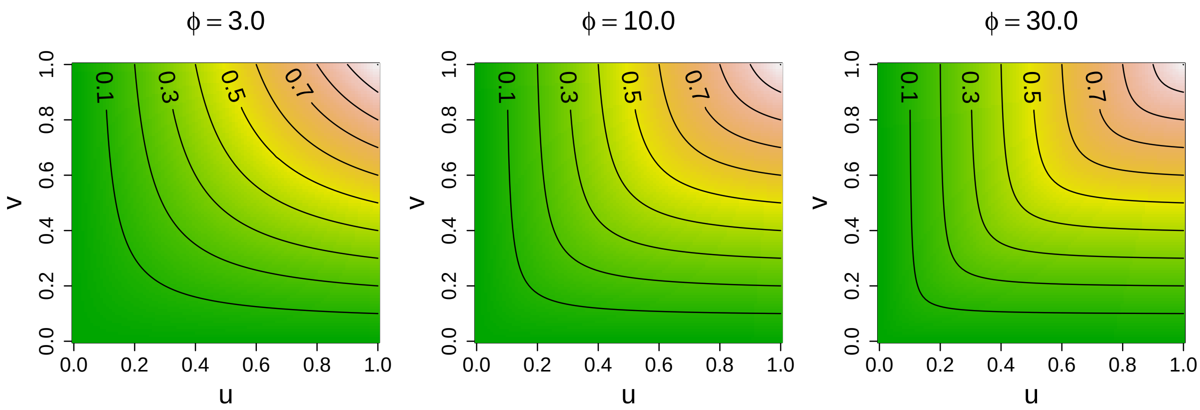

Figure 18 – Contour plots of the Plackett copula considering some values for ϕ.

3.5 Bivariate models

In the statistical literature, several authors introduced different strategies for modeling bivariate time-to-event data. As examples, it is possible to mention Hougaard (1987), Liang et al. (1995), Price and Manatunga (2001), Parner (2001) and Hougaard (2012). When modeling survival data, it is common to assume a dependence structure between the observed lifetimes and the modeling of this dependence has been the goal of many researchers. The use of frailty models proposed by Vaupel et al. (1979) to model this dependence is a very popular technique. In frailty models one or more random effects are included to model the dependence between the lifetimes. Another frequently used approach is the construction and application of bivariate parametric distributions in order to introduce specific parameters that capture the dependence between the lifetimes in these distributions. The following are some examples: Gumbel bivariate exponential (Gumbel, 1960a), Marshall-Olkin bivariate exponential (Marshall and Olkin, 1967), Block and Basu bivariate exponential (Block and Basu, 1974), bivariate exponential and geometric distributions (Muraleedharan Nair and Unnikrishnan Nair, 1988) and Basu-Dhar Bivariate geometric distributions (Basu and Dhar, 1995; de Oliveira and Achcar, 2018).

Another alternative method to model bivariate data, including the dependence between the lifetime analysis, is the use of copula functions introduced by Sklar (1959), and described for example by Nelsen (2007), Balakrishnan and Lai (2009), Jaworski et al. (2010) and Joe (2014). Copula functions in bivariate models for the analysis of survival data have been introduced by a number of authors, such as Viswanathan and Manatunga (2001), Suzuki et al. (2011), Meyer and Romeo (2015), Shih and Emura (2016), Peres et al. (2018) and Emura and Chen (2018). In the context where bivariate lifetimes and the presence of a cure rate are observed, we can mention the studies of Louzada et al. (2013), Martinez and Achcar (2014), Achcar et al. (2016), Coelho-Barros et al. (2016), Cordeiro et al. (2016), Huang (2019) and Peres et al. (2020).

3.5.1 Bivariate mixture cure rate model

Considering the equation (3.4), the mixture formulations for the survival functions in the bivariate lifetime case are given by

$$S_j(t_j) = \rho_j + (1-\rho_j)\, S_{0j}(t_j), \tag{3.22}$$

such that, $j = 1, 2$, where $\rho_j \in (0, 1)$ and $S_{0j}(t_j)$ denotes the baseline survival function for the susceptible individuals in the population. From (3.17) and (3.22), the joint survival function for T_1 and T_2 considering the mixture formulation is expressed by

$$S(t_1, t_2) = \rho_1 + \rho_2 + (1-\rho_1)\, S_{01}(t_1) + (1-\rho_2)\, S_{02}(t_2) + F(t_1, t_2) - 1.$$

Let us define two indicator variables denoted by V_1 and V_2, such that,

$$V_j = \begin{cases} 1 & \text{if the individual is susceptible in } T_j \\ 0 & \text{if the individual is cured or immune in } T_j, \end{cases} \tag{3.23}$$

where, $j = 1, 2$. Thus, $P(V_j = 0) = \rho_j$ and $P(V_j = 1) = 1 - \rho_j$. The observed lifetime data for any individual satisfies one of the following cases:

(a) The individual is not susceptible to both events of interest. So, $\psi_{00} = P(V_1 = 0, V_2 = 0) = P(V_1 = 0)P(V_2 = 0) + cov(V_1, V_2) = \rho_1\rho_2 + \omega$.

(b) The individual is not susceptible to event 1 but is susceptible to event 2. So, $\psi_{10} = P(V_1 = 1, V_2 = 0) = P(V_1 = 1)P(V_2 = 0) - cov(V_1, V_2) = (1-\rho_1)\rho_2 - \omega$.

(c) The individual is susceptible to event 1 but not to event 2. Then, $\psi_{01} = P(V_1 = 0, V_2 = 1) = P(V_1 = 0)P(V_2 = 1) - cov(V_1, V_2) = \rho_1(1-\rho_2) - \omega$.

(d) The individual is susceptible to both events. So, $\psi_{11} = P(V_1 = 1, V_2 = 1) = P(V_1 = 1)P(V_2 = 1) + cov(V_1, V_2) = (1-\rho_1)(1-\rho_2) + \omega$.

Note that $\omega = cov(V_1, V_2)$ denotes the covariance between the binary random variables V_1 and V_2, where $0 \leq \omega \leq min(p_1, p_2) - p_1p_2$. In addition, note that $\psi_{00} + \psi_{10} + \psi_{01} + \psi_{11} = 1$, $\psi_{00} + \psi_{01} = p_2$ and $\psi_{00} + \psi_{10} = p_1$. If the covariance ω takes the zero value, there is independence between the probabilities of cure associated to the lifetimes T_1 and T_2. Following Wienke et al. (2006), the joint mixture cure rate function for the bivariate lifetimes T_1 and T_2 is given by,

$$S(t_1, t_2) = \psi_{00} + \psi_{10}S_{01}(t_1) + \psi_{01}S_{02}(t_2) + \psi_{11}S_0(t_1, t_2), \tag{3.24}$$

where $S_0(t_1, t_2)$ is the joint survival function for T_1 and T_2 for the susceptible individuals.

3.6 Likelihood function under random censoring

The study of survival data analysis involves censoring data, so it is necessary to be carefully to build likelihood functions. An important assumption considered in this thesis, is that the lifetimes and censoring times are independent, and scheme of non-informative censoring.

3.6.1 Univariate Case

Let T be a positive random variable denoting the lifetime of a individual and δ indicating whether the lifetime T is observed ($\delta = 1$) or not ($\delta = 0$) and assuming that the observed data is given by the expression $t_i = \min(T_i, C_i), i = 1, 2, \ldots, n$. Thus under the random right-censoring scheme, the likelihood function is given by:

$$L(\theta) = \prod_{i=1}^{n} [f(t_i)]^{\delta_i} [S(t_i)]^{1-\delta_i} \tag{3.25}$$

where θ is the model vector of parameters, where $f(t_i)$ and $S(t_i)$ denotes, respectively, the probability density and survival function associated to the i^{th} subject.

3.6.2 Bivariate Case

For the bivariate case, we also consider the random right-censoring scheme. In this case, let $(T_{11}, T_{21}), (T_{12}, T_{22}), \ldots, (T_{1n}, T_{2n})$ be a random sample of size n from a bivariate lifetime distribution and define the following indicator variables:

$$\begin{cases} \delta_{1i} = 1 \quad \text{if} \quad T_{1i} < C_{1i} \quad \text{and 0, otherwise.} \\ \delta_{2i} = 1 \quad \text{if} \quad T_{2i} < C_{2i} \quad \text{and 0, otherwise.} \end{cases} \tag{3.26}$$

where $i = 1, 2, \ldots, n; (C_{1i}, C_{2i})$ are the right censoring times. In this way, there are four possible situations:

C_1: Both, T_{1i} and T_{2i}, are complete observations ($\delta_{1i} = 1, \delta_{2i} = 1$),

C_2: T_{1i} are complete and T_{2i} are censored observations ($\delta_{1i} = 1, \delta_{2i} = 0$),

C_3: T_{1i} are censored and T_{2i} are complete observations ($\delta_{1i} = 0, \delta_{2i} = 1$),

C_4: Both, T_{1i} and T_{2i}, are censored observations ($\delta_{1i} = 0, \delta_{2i} = 0$).

In this way,considering the information of a random sample with size n, the likelihood function is given by,

$$L \propto \prod_{i \in C_1} f(t_{1i}, t_{2i}) \prod_{i \in C_2} \left[-\frac{\partial S(t_{1i}, t_{2i})}{\partial t_{1i}} \right] \prod_{i \in C_3} \left[-\frac{\partial S(t_{1i}, t_{2i})}{\partial t_{2i}} \right] \prod_{i \in C_4} S(t_{1i}, t_{2i}) \tag{3.27}$$

From Equation (3.27) is observed that the contribution of the i-th subject from a random sample $(t_{1i}, t_{2i}, \delta_{1i}, \delta_{2i})$, $i = 1, ..., n$, for the likelihood function is given by

$$\begin{aligned} L_i &\propto \left[\frac{\partial^2 S(t_1,t_2)}{\partial t_{1i}\partial t_{2i}}\right]^{\delta_{1i}\delta_{2i}} \left[-\frac{\partial S(t_1,t_2)}{\partial t_{1i}}\right]^{\delta_{1i}(1-\delta_{2i})} \times \\ &\times \left[-\frac{\partial S(t_1,t_2)}{\partial t_{2i}}\right]^{(1-\delta_{1i})\delta_{2i}} [S(t_1,t_2)]^{(1-\delta_{1i})(1-\delta_{2i})}. \end{aligned} \tag{3.28}$$

In this expression, considering the equation (3.24), we have,

$$\begin{aligned} \frac{\partial^2 S(t_1,t_2)}{\partial t_1 \partial t_2} &= f_{01}(t_1) f_{02}(t_2)\psi_{11} c_\phi(u,v), \\ -\frac{\partial S(t_1,t_2)}{\partial t_1} &= f_{01}(t_1)\left[\psi_{10} + \psi_{11}\frac{\partial}{\partial u}C_\phi(u,v)\right] \text{ and} \\ -\frac{\partial S(t_1,t_2)}{\partial t_2} &= f_{02}(t_2)\left[\psi_{01} + \psi_{11}\frac{\partial}{\partial v}C_\phi(u,v)\right], \end{aligned}$$

The derivatives $\frac{\partial}{\partial u}C_\phi(u,v)$ and $\frac{\partial}{\partial v}C_\phi(u,v)$ for each studied copula function are given in Peres et al. (2020).

3.7 An introduction to maximum likelihood inference

The most used method to estimate the parameters of a probability distribution is the maximum likelihood estimate method, introduced by Fisher (1912). The maximum likelihood method find the values of the parameters that maximize the likelihood function when a sample of data is observed. Considering $\boldsymbol{\theta}$ as a vector of parameters of a probability distribution and $L(\boldsymbol{\theta})$ the likelihood function, the maximum likelihood estimator of $\boldsymbol{\theta}$ is the value of $\hat{\boldsymbol{\theta}} \in \Theta$ that maximizes the likelihood function, where Θ is the parametric space of $\boldsymbol{\theta}$.

It is often easier algebraically and computationally to find the value that maximizes the logarithm of the likelihood function, that is, $\ell(\boldsymbol{\theta}) = \log[L(\boldsymbol{\theta})]$. Note that to maximize $\ell(\theta; \boldsymbol{x})$ is the same that to maximize $L(\boldsymbol{\theta})$, since the logarithm function is strictly increasing. In some cases, the maximum value of $\ell(\boldsymbol{\theta})$ can be obtained explicitly. However, in general, for the most complex situations the solution is obtained by numerical approaches.

3.7.1 Asymptotic confidence interval

The confidence intervals for the parameters with point estimators obtained using the maximum likelihood method can be obtained in several ways. The most common approach is the asymptotic method. Confidence intervals for a parameter vector $\boldsymbol{\theta}$ can be constructed using any test statistics (Cordeiro and Demétrio, 2013). The asymptotic method is built from the inversion of the Z-test (Wald test) (Casella and Berger, 2002;

Cox, 2006; Millar, 2011). Regarding certain conditions the maximum likelihood estimator for $\boldsymbol{\theta}$ has an asymptotic normal distribution, from where we get the asymptotic interval for a parameter θ_k with (1-α)100% confidence given by,

$$(\hat{\theta}_k - Z_{1-\frac{\alpha}{2}}\sqrt{\mathrm{Var}(\hat{\theta}_k)}, \hat{\theta}_k + Z_{1-\frac{\alpha}{2}}\sqrt{\mathrm{Var}(\hat{\theta}_k)}),$$

where $Z_{1-\frac{\alpha}{2}}$ is the quantile of a standard Normal distribution and $\mathrm{Var}(\hat{\theta}_k)$ is the variance of $\hat{\theta}_k$, obtained from the variance - covariance matrix $\Sigma(\hat{\boldsymbol{\theta}})$, given by,

$$\Sigma(\hat{\boldsymbol{\theta}}) = -\left[E\left(H(\hat{\boldsymbol{\theta}})\right)\right]^{-1} = \begin{bmatrix} \mathrm{Var}(\theta_1) & \mathrm{Cov}(\theta_1,\theta_2) & \cdots & \mathrm{Cov}(\theta_1,\theta_k) \\ \mathrm{Cov}(\theta_1,\theta_2) & \mathrm{Var}(\theta_2) & \cdots & \mathrm{Cov}(\theta_2,\theta_k) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(\theta_k,\theta_1) & \mathrm{Cov}(\theta_2,\theta_k) & \ldots & \mathrm{Var}(\theta_k) \end{bmatrix},$$

where $H(\hat{\boldsymbol{\theta}})$ is the hessian matrix of the maximum likelihood estimators. The variance-covariance matrix is usually obtained from the inverse of the observed Fisher information matrix.

3.7.2 The delta method

The delta method is a technique commonly used to obtain confidence intervals for functions of parameters that have the parameters estimated by the maximum likelihood method. The variance of the estimator of a function of the parameters that we want to get a confidence interval, in general, is analytically difficult to obtain. The delta method obtains a linear approximation for this function, expanding it to the first order of a Taylor series obtaining a good approximation for large sample sizes. Observe that from the invariance property of the maximum likelihood estimators the estimate of one-to-one functions of parameters are also maximum likelihood estimators.

If $h(\theta)$ is a function of the a parameter θ, so the variance of $h(\hat{\theta})$ by the method delta, is given by,

$$\mathrm{Var}\left[h(\hat{\theta})\right] \approx \left(\frac{\mathrm{d}h(\theta)}{\mathrm{d}\theta}\right)^2\bigg|_{\theta=\hat{\theta}} \mathrm{Var}(\hat{\theta}).$$

The delta method for the multivariate case is simply the extension of the above specifications using matrix formulation. Suppose that $h(\boldsymbol{\theta})$ is a function of the parameter vector $\boldsymbol{\theta}$, so the variance of $h(\hat{\boldsymbol{\theta}})$ by the delta method, is expressed by,

$$\mathrm{Var}\left[h(\hat{\boldsymbol{\theta}})\right] \approx \frac{\partial h(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \Sigma(\hat{\boldsymbol{\theta}}) \frac{\partial h(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}},$$

where $\Sigma(\hat{\boldsymbol{\theta}})$ is the variance – covariance matrix. More details about delta method can be seen in Oehlert (1992), Cox (2005) and Ver Hoef (2012). In the case of a function with two parameters, $\boldsymbol{\theta} = (\theta_1, \theta_2)$, the Var $\left[h(\hat{\boldsymbol{\theta}})\right]$ is given by,

$$\begin{aligned}\text{Var}\left[h(\hat{\boldsymbol{\theta}})\right] &\approx \text{Var}\left(\hat{\theta}_1\right)\left(\frac{\partial h(\boldsymbol{\theta})}{\partial \theta_1}\right)^2\Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} + \text{Var}\left(\hat{\theta}_2\right)\left(\frac{\partial h(\boldsymbol{\theta})}{\partial \theta_2}\right)^2\Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \\ &+ 2\text{Cov}\left(\hat{\theta}_1, \hat{\theta}_2\right)\frac{\partial^2 h(\boldsymbol{\theta})}{\partial \theta_2 \partial \theta_1}\Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\end{aligned}$$

3.8 A brief introduction to Bayesian inference

In the last decades it has become evident that the of Bayesian methods is becoming increasingly important as a tool in data analysis. These alternative inference methods are based on a basic result in probability theory: the Bayes formula also know as Bayes theorem. In 1763, the English reverend and mathematician Thomas Bayes, in his posthumous publication "An Essay Towards Solving a Problem in the Doctrine of Chances", presented his theoretical foundations and methodological procedures, based on conditional probabilities.

The Bayesian approach consists on specifying a probability model for the observed data, given a vector of unknown parameters $\boldsymbol{\theta}$, such that, $\boldsymbol{\theta}$ is considered as a random variable under the Bayesian framework, and provides a rational method for updating the new information using the Bayes' theorem and prior distributions for measure the uncertainty about $\boldsymbol{\theta}$ (Ibrahim et al., 2005). That is, according to Gelman et al. (1995) the Bayesian paradigm is the process of adjusting a probability model to a data set, assuming a prior distribution for the parameters of the proposed model and summarizing the results from a joint posterior probability distribution for the parameters where it is also possible to find predictions of non-observed quantities in a direct way.

A prior distribution, plays a significant role in Bayesian analysis, which supposedly represents what is known about unknown parameters before the data is available. This distribution can be used to represent prior knowledge or relative ignorance about the vector parameter $\boldsymbol{\theta}$ (Box and Tiao, 2011). Different prior distributions can be considered to deal with uncertainty about $\boldsymbol{\theta}$, where in some cases, there is information of experts on the vector parameter $\boldsymbol{\theta}$, that is, we have informative priors, but in general this information is not available, where non-informative priors should be used.

3.8.1 Bayes Theorem

The reverend Bayes proposed in his work the following idea:

$$\text{Initial viewpoint } \times \text{ New data } \rightarrow \text{Improved viewpoint.}$$

In the context of statistical inference, this can be interpreted as

$$\text{Prior distribution} \times \text{Likelihood function} \rightarrow \text{Posterior distribution.}$$

Formalizing this result, consider a vector $\mathbf{y}$ of n observations, which probability distribution $\pi(\mathbf{y} \mid \boldsymbol{\theta})$ depending on a parameter vector $\boldsymbol{\theta}$. Suppose that $\boldsymbol{\theta}$ has a prior distribution denoted by $\pi(\boldsymbol{\theta})$. In this way, the Bayes' theorem is expressed by,

$$\pi(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{\pi(\mathbf{y} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\mathbf{y})}.$$

Note that $1/\pi(\mathbf{y})$ in this expression is independent from θ; it is a normalizing constant to have a proper probability distribution. Omitting the denominator from the Bayes' formula, the simplified form of the Bayes' theorem is obtained, useful in the estimation of parameters, given by,

$$\pi(\boldsymbol{\theta} \mid \mathbf{y}) \propto \pi(\mathbf{y} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta}).$$

where $\pi(\boldsymbol{\theta} \mid \mathbf{y})$ is the posterior distribution of $\boldsymbol{\theta}$ given the knowledge of the data. Remember that the parameter $\boldsymbol{\theta}$ is considered as a random value under the Bayesian viewpoint (Berger, 2013).

For a fixed value of y, the function $L(\theta \mid y) = \pi(y \mid \boldsymbol{\theta})$ measures the likelihood of each possible value of $\boldsymbol{\theta}$, while $\pi(\boldsymbol{\theta})$ is a priori distribution of $\boldsymbol{\theta}$. These two sources of information, prior and likelihood, are combined to obtain the joint posterior distribution of $\boldsymbol{\theta}$, $\pi(\boldsymbol{\theta} \mid \mathbf{y})$. Thus, the usual form of the Bayes' theorem is described by,

$$\pi(\boldsymbol{\theta} \mid y) \propto L(\boldsymbol{\theta} \mid y)\pi(\boldsymbol{\theta}).$$

To get the inferences from the joint posterior distribution $\pi(\boldsymbol{\theta} \mid y)$, it is needed to find a distribution for each component parameter in $\boldsymbol{\theta}$, called the marginal distribution of $\boldsymbol{\theta}$. In the continuous case, the marginal distribution of $\boldsymbol{\theta}$ is calculated by integration of the joint posterior distribution in relation to the other parameters of the model, that is, to solve the following multiple integral,

$$\pi(\theta_i|y) = \int \dots \int \pi(\theta_i, \theta_{-i}|y)\, \mathrm{d}\theta_{-i}$$

where θ_i is the parameter that is being obtained the inference and θ_{-i} is the complementary set of parameters of $\boldsymbol{\theta}$, that is, they are all parameters excluding the parameter θ_i (Rosa, 1998).

Frequently, the analytical form of the joint posterior distribution is multiparametric and complex. In this case, there are no analytical form for the marginal distributions, and to get the solution to this integral, approximations and iterative numerical methods must be used. The most common method is the Markov chain Monte Carlo (MCMC), that is used to generate values from all conditional posterior distributions for each parameter, that is equivalent to get samples from the joint posterior distribution for all parameters (Sorensen, 1996). Among the simulation methods using the MCMC, we have the Gibbs sampling, as described by many authors, such as Gelfand and Smith (1990); Meyn and Tweedie (2012); Kijima (2013) and the Metropolis-Hastings algorithm described by Chib and Greenberg (1995); Roberts and Smith (1994); Robert and Casella (1999).

According to Zeger and Karim (1991), the motivation for using Gibbs sampling is that the true posterior distribution can be approximated by an empirical distribution of B values, such that B should be large enough for the Gibbs sampling algorithm to achieve convergence and stationarity. The determination of the convergence and the stationarity of the process can be obtained by graphical techniques and by statistical methods. Among the statistical methods, it is possible to mention Heidelberger and Welch (1983), Gelman et al. (1992) and Geweke et al. (1991).Other methods to check convergence and discussions about MCMC are presented by Cowles and Carlin (1996), Gilks (2005) e Liang et al. (2011). Casella and George (1992), showed that in a sufficiently large sample, $B \longrightarrow \infty$, and using the Gibbs sampling, it is possible to obtain an empirical distribution that is sufficiently close to the real marginal distribution for each parameter.

3.8.2 Interval Estimation Method

In the point estimate, in general, all the information present in the posterior distribution is summarized by means of the mean or median of the estimates obtained by the Monte Carlo estimates based on the MCMC generated samples depending on the assumed loss function for the Bayesian estimation procedure. According to Box and Tiao (2011) two methods are commonly used to obtain Bayesian inference intervals, the credibility intervals and high posterior density interval.

The credibility interval C with probability $100(1-\alpha)\%$ for θ is such that, $P(\theta \in C) \geq 1-\alpha$. The credibility intervals are invariant under one-to-one transformations, that is, if $C = [a, b]$ is a $100(1-\alpha)\%$ credibility interval for θ, then $[h(a), h(b)]$ is a $100(1-\alpha)\%$ credibility interval for $h(\theta)$.

A high posterior density interval is a $100(1-\alpha)\%$ C credibility interval for θ, such that, $C = \{\theta \in \Theta : P(\theta \mid x) \geq k(\alpha)\}$ where $k(\alpha)$ is the higher constant such that $P(\theta \in C) \geq 1-\alpha$.

3.9 Model comparison criteria

In the maximum likelihood estimation and Bayesian estimation, there are different selection criteria. The selection criterion commonly considered to compare models in maximum likelihood estimation are: AIC (Akaike's information criterion), introduced

Akaike (1974), defined by,

$$AIC = -2\ell\left(\boldsymbol{\theta}\right) + 2k,$$

where $\ell\left(\boldsymbol{\theta}\right)$ is the log-likelihood of the adjusted model and k is the number of parameters estimated in the model. Other criterion used is the $AICc$ (Akaike information criterion corrected), where following Sugiura (1978), this criterion is expressed by,

$$AICc = AIC + \frac{2k(k+1)}{n-k-1},$$

where n is the sample size. Also it is used the BIC (Bayesian information criterion), proposed by Schwarz et al. (1978), given by

$$BIC = -2\ell\left(\boldsymbol{\theta}\right) + k\log(n).$$

Under a Bayesian approach, it is commonly considered for comparison among different models the use of the DIC (deviance information criterion) introduced by Spiegelhalter et al. (2002). The DIC value is calculated by

$$\text{DIC} = D(\hat{\boldsymbol{\theta}}) + 2p_D = 2\overline{D(\boldsymbol{\theta})} - D(\hat{\boldsymbol{\theta}}),$$

where $D(\hat{\boldsymbol{\theta}})$ is the deviance calculated in the posterior mean of the parameter of interest obtained using MCMC simulation methods and p_D is the effective number of parameters in the model, such that, $p_D = \overline{D(\boldsymbol{\theta})} - D(\hat{\boldsymbol{\theta}})$, where $\overline{D(\boldsymbol{\theta})} = E[D(\boldsymbol{\theta})]$ is the posterior mean of the deviance and can be approximated using a MCMC simulation by considering the sample mean of the simulated values of $D(\boldsymbol{\theta})$.

Other Bayesian model comparison criterion is the $EAIC$ (expected Akaike information criterion) proposed by Brooks et al. (2002), expressed by,

$$EAIC = \overline{D(\boldsymbol{\theta})} + 2k$$

where k is the number of parameters estimated in the model. Also there is the $EBIC$ (expected Bayesian information criterion), introduced by Carlin and Louis (2000), defined by,

$$EBIC = \overline{D(\boldsymbol{\theta})} + k\log(n)$$

where n is the sample size.

In all of these criteria, the lower values indicate better fit of the model. It can be especially said that when comparing models, a significant difference between models is needed to decide by the best model; some authors point out that the difference between the obtained criterion index values should be greater than 5 (Box and Tiao, 2011).

Another criterion for comparison among Bayesian models is given by the $LPML$ (logarithms of pseudo marginal likelihood functions). The $LPML$ is obtained from the conditional predictive ordinates (CPO) (Gelfand et al., 1992). For the i observation the CPO_i is given by,

$$CPO_i = \int \pi\left(\mathbf{D}_i|\boldsymbol{\theta}\right)\pi\left(\boldsymbol{\theta}|\mathbf{D}_{[-i]}\right)\,\mathrm{d}\boldsymbol{\theta},$$

where $\boldsymbol{\theta}$ is the complete vector of parameter, $\mathbf{D}_{[-i]}$ is the sample without the observation i and $\pi\left(\boldsymbol{\theta}|\mathbf{D}_{[-i]}\right)$ is the posterior density of $\boldsymbol{\theta}$ given $\mathbf{D}_{[-i]}$, $i = 1, \ldots, n$. In general the CPO_i does not have a closed form and it is very complicated to be calculated. However (Dey et al., 1997) presented an approximation based on MCMC methods for the CPO_i, expressed by,

$$\widehat{CPO}_i = \left[\frac{1}{B}\sum_{n=1}^{B}\frac{1}{\pi\left(\mathbf{D}_i|\boldsymbol{\theta}_n\right)}\right]^{-1}, i = 1, \ldots, n,$$

where B is the number of iterations during the implementation of the MCMC procedure after a burn-in period used to eliminate the effect of the started values of the iterative algorithm and where $\boldsymbol{\theta}_n$ is the vector of the samples obtained at the n-th iteration (Chen et al., 2012). In this form, the LPML value is obtained by,

$$\widehat{LPML} = \sum_{i=1}^{n}\ln\widehat{CPO}_i.$$

Following Geisser and Eddy (1979), larger values of LPML indicate better fit of the model.

Chapter 4

Survival data sets

In this chapter, it is presented the bivariate survival data related to medical studies used in the this Thesis. These data were considered, in order to evaluate the proposed methodologies. For each bivariate survival data it is exhibited the survival plot estimates by Kaplan-Meier method (see, Section 3.2.1), the empirical hazard function (obtained using the package "bshazard" in R software, see Paola Rebora and Reilly (2018)) and survival bivariate Kaplan-Meier, presented in Section 3.2.5.

4.1 Diabetic retinopathy data set

In the statistical literature the diabetic retinopathy data were considered by several authors. These data were introduced by Group et al. (1976), being related to the time to visual loss of 197 diabetic patients under 60 years of age that were followed-up for a fixed period of time. The main purpose of the study is to assess the efficacy of photocoagulation treatment for proliferative retinopathy. In the study, each patient had one eye randomized for laser treatment and the other eye receiving no treatment. It was considered for the bivariate analyzes that T_1 is the time up to visual loss for the treated eye, while T_2 is the time up to visual loss for the not treated or control eye. There are 43% censored data of treated eyes and 73% censored data of not treated eyes. Figure (19), shows the survival functions estimated by Kaplan-Meier method and hazard functions for T_1 and T_2. The bivariate Kaplan-Meier surface is displayed in Figure (20).

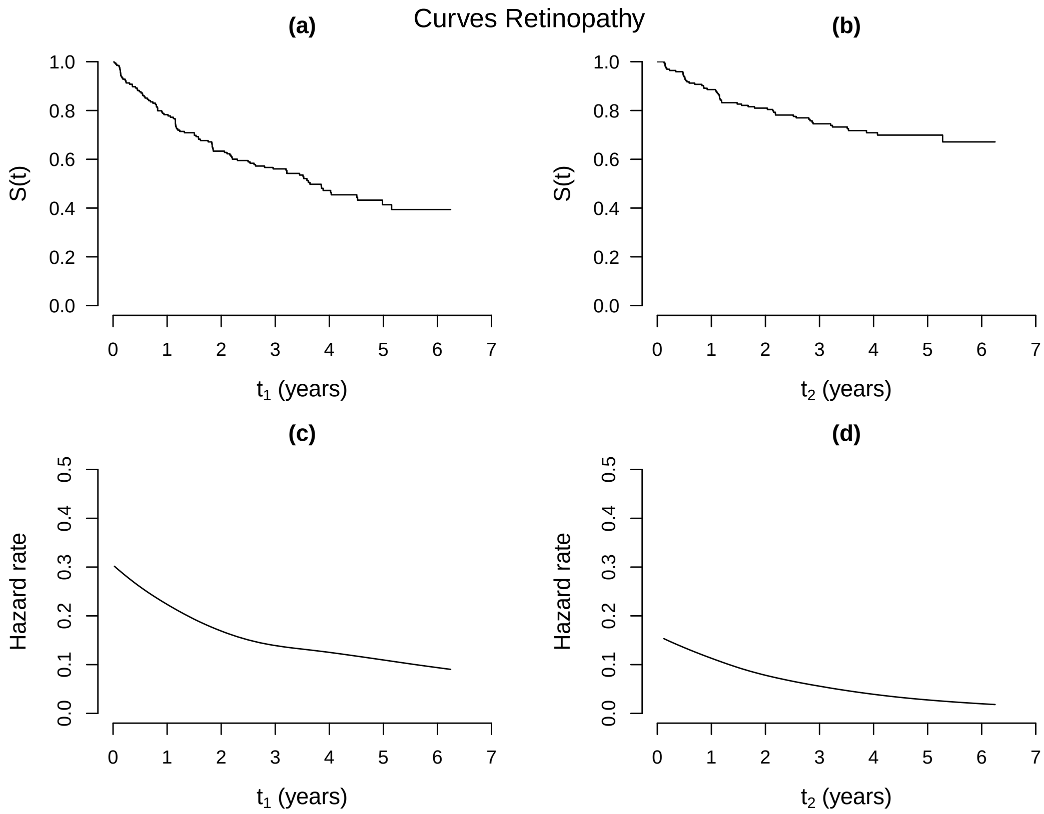

Figure 19 – Plots of the survival functions estimated by Kaplan-Meier method (upper panels) and respective hazard functions (lower panels) for treatment eye (panels (a) and (c)) and control eye (panels (b) and (d)).

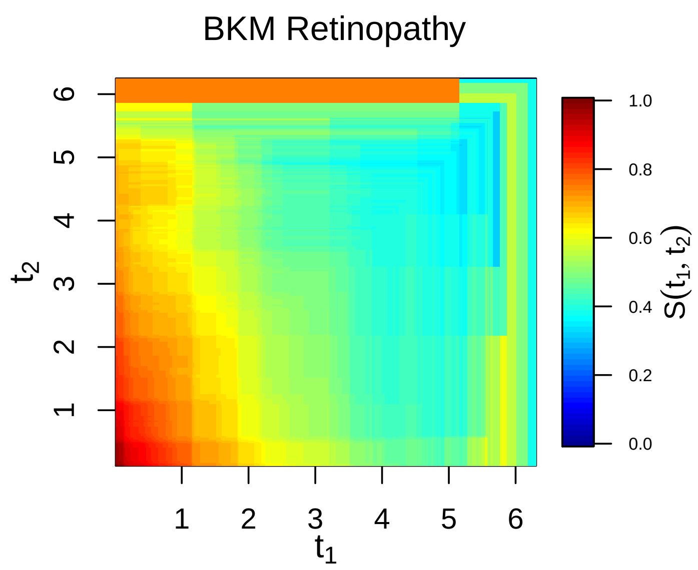

Figure 20 – Bivariate Kaplan-Meier surface plot, considering diabetic retinopathy data.

4.2 The cervical cancer data set

The cervical cancer data considered in this thesis was introduced by Brenna et al. (2004). This study is related to invasive cervical cancer in 118 women who received the standard treatment recommended by the International Federation of Gynecology and Obstetrics (FIGO) for invasive cervical carcinoma between 1992 and 2002. For the bivariate analysis T_1 is the disease-free survival (DFS) defined as the time from the date of surgery to the first event of disease recurrence, while T_2 is the overall survival (OS) defined as the time from the date of surgery to the death. There are 48% censored data in T_1 and 53% censored data in T_2. Figure (21), shows the survival functions estimated by Kaplan-Meier method and hazard functions for T_1 and T_2. Figure (22) shows the bivariate Kaplan-Meier surface based on cervical cancer data.

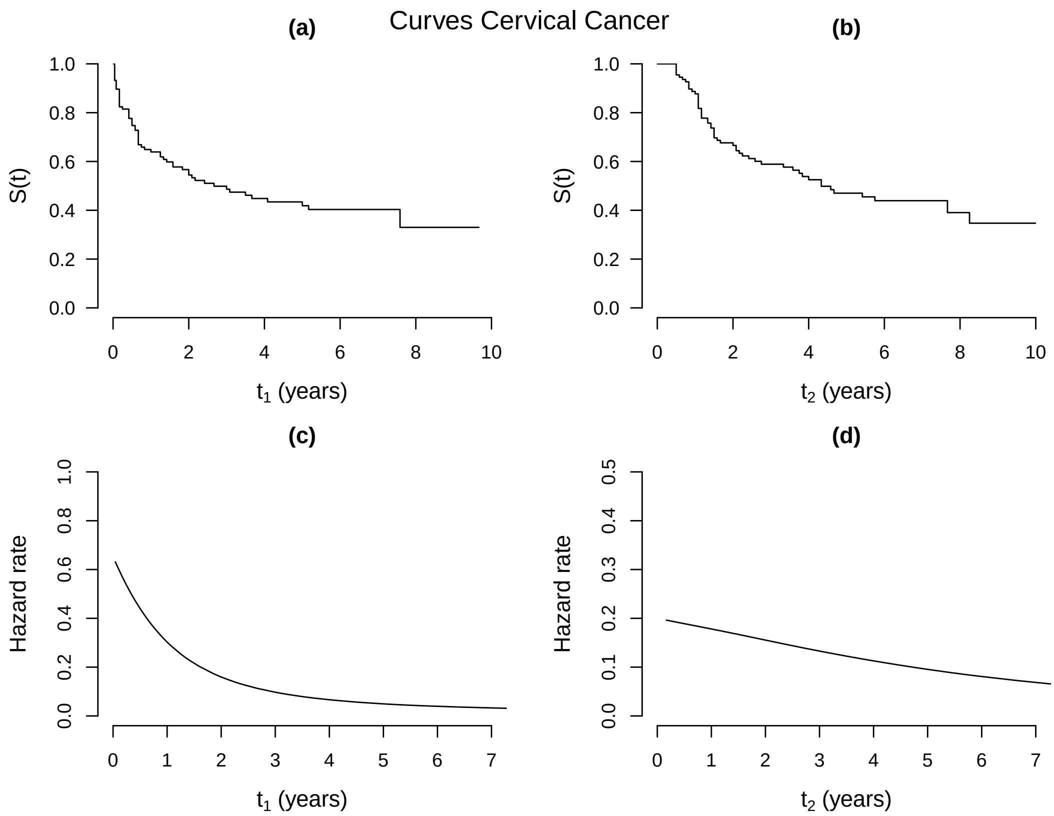

Figure 21 – Plots of the survival functions estimated by Kaplan-Meier method (upper panels) and respective hazard functions (lower panels) for DFS time (panels (a) and (c)) and OS time (panels (b) and (d)).

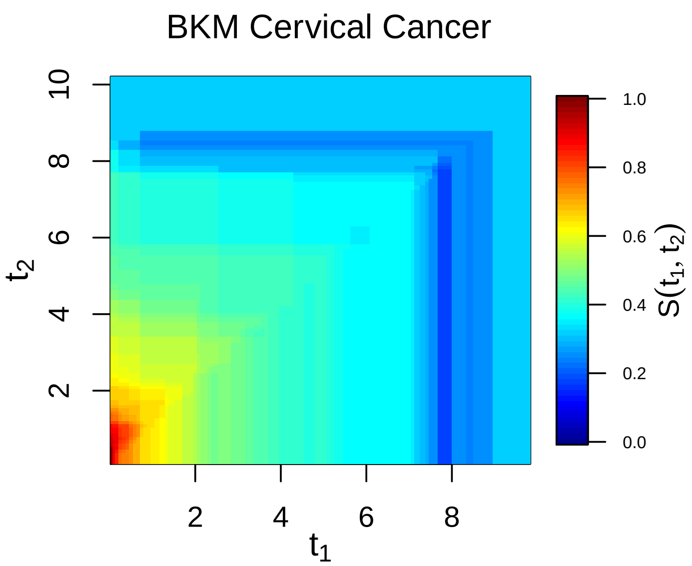

Figure 22 – Bivariate Kaplan-Meier surface plot based on cervical cancer data.

4.3 Breast cancer data set

The breast cancer data introduced by Shigemizu et al. (2017) is relative to a cohort study, where 97 patients underwent surgical treatment for breast cancer followed for the period ranging from the years 2000 to 2011. For bivariate lifetime application we consider as T_1 the disease-free survival time (DFS) and T_2 is considered the overall survival time (OS). In the dataset, there are 75% censored observations in the disease-free survival time (T_1) and 80% censored data in the overall survival time (T_2). Figure (23), shows the survival functions estimated by Kaplan-Meier method and hazard functions for T_1 and T_2. Figure (22) shows the bivariate Kaplan-Meier surface from breast cancer data.

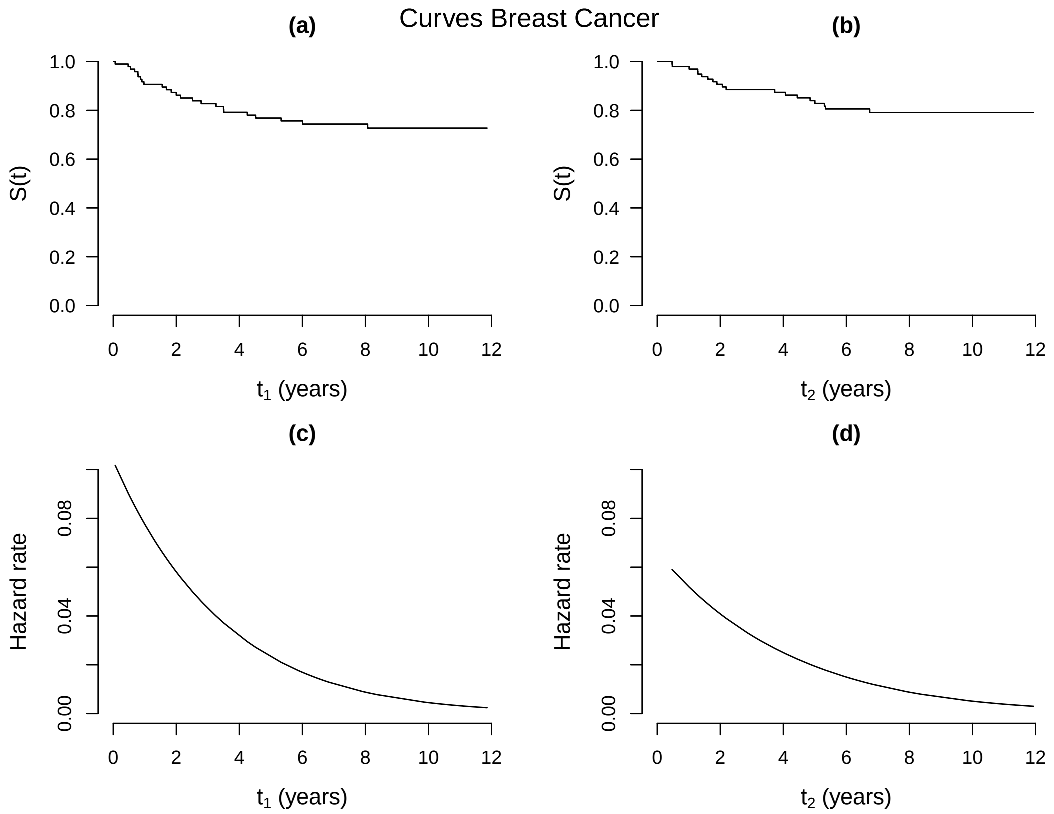

Figure 23 – Plots of the survival functions estimated by Kaplan-Meier method (upper panels) and respective hazard functions (lower panels) for DFS time (panels (a) and (c)) and OS time (panels (b) and (d)).

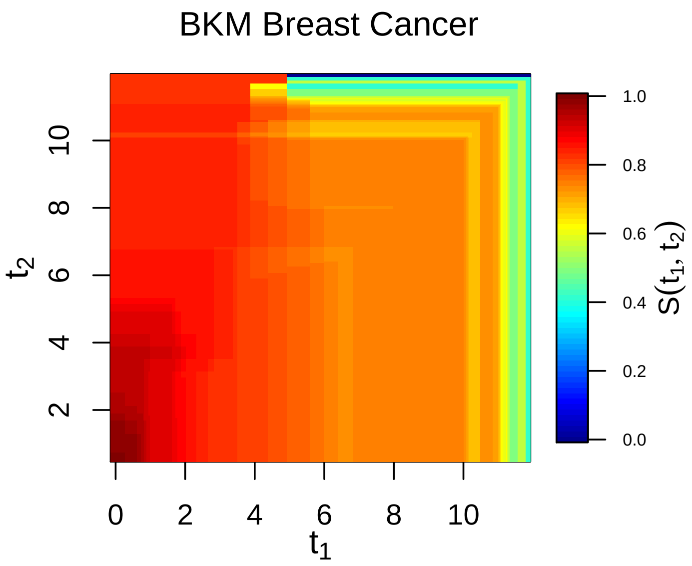

Figure 24 – Bivariate Kaplan-Meier surface plot based on breast cancer data.

4.4 Tobacco-stained fingers data set

Tobacco-stained fingers data were obtained from a retrospective cohort study on a population of 143 smokers screened between March 2006 and January 2010 in a 180-bed community hospital in La Chaux-de-Fonds, Switzerland. Data on death and hospital admission were collected to the June 2014 (see John et al. (2015)). In this bivariate study, it is considered as T_1 the time before the first hospital readmission in smokers with stains on their fingers which were censored in case of death before the closure date; and as lifetime T_2 the overall survival (OS) time of the patient with tobacco-stained on their fingers. There are 26% censored cases in T_1 and 48% censored cases in T_2. In the Figure (25) it is presented the survival functions estimated by Kaplan-Meier method and hazard functions for T_1 and T_2. Figure (26) shows the bivariate Kaplan-Meier surface from tobacco-stained fingers data.

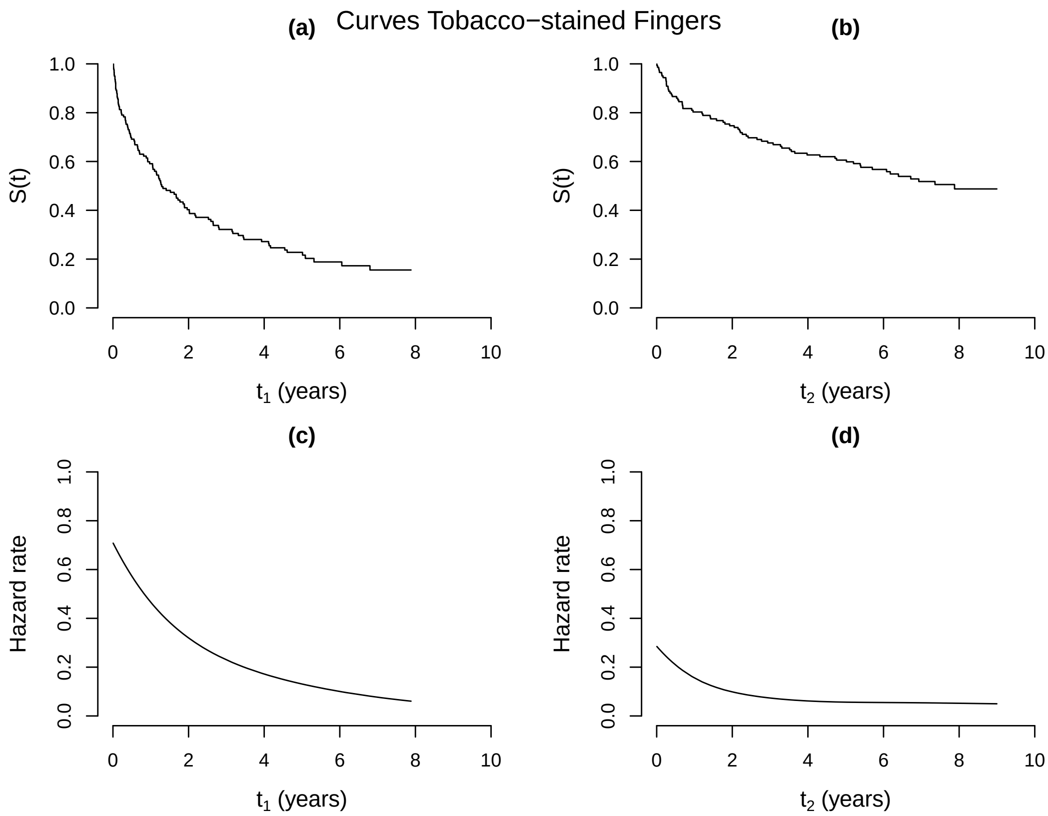

Figure 25 – Plots of the survival functions estimated by Kaplan-Meier method (upper panels) and respective hazard functions (lower panels) for first time hospital readmission (panels (a) and (c)) and OS time (panels (b) and (d)).

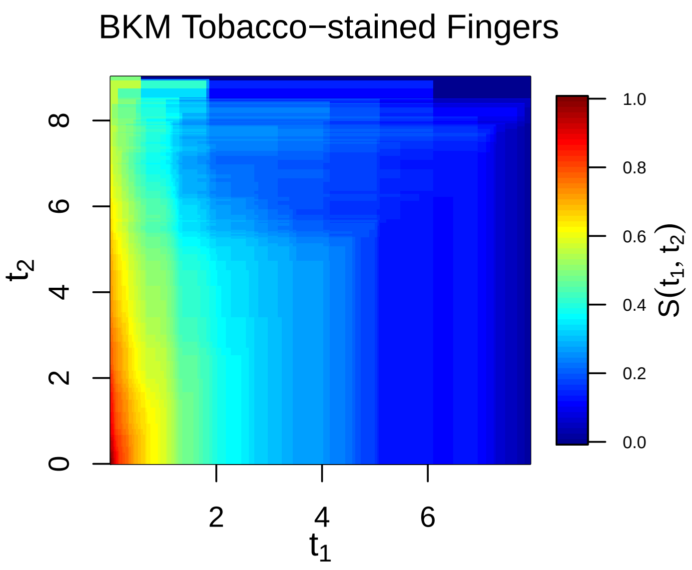

Figure 26 – Bivariate Kaplan-Meier surface plot from tobacco-stained fingers data.

Chapter 5

Bivariate lifetime models in presence of cure fraction: a comparative study with many different copula functions

The developments presented in this Chapter were published in the form of an article in the journal Heliyon, which can be accessed at https://doi.org/10.1016/j.heliyon.2020.e03961. Heliyon is an all-science, open access journal. Every article published in Heliyon is indexed by PubMed, Scopus, and Web of Science Emerging Sources Citation Index (ESCI).

In this article, we revised the literature to describe the copula functions that can be useful to the construction of bivariate distributions to analyse lifetime data, and it was introduced a Bayesian approach to fit our proposed model based on fifteen copula functions including the presence of long-term survivors and censored data. It was used three real data sets to illustrate our methodology.

Heliyon 6 (2020) e03961

Contents lists available at ScienceDirect

Heliyon

www.elsevier.com/locate/heliyon

Research article

Bivariate lifetime models in presence of cure fraction: a comparative study with many different copula functions

Marcos Vinicius de Oliveira Peres *, Jorge Alberto Achcar, Edson Zangiacomi Martinez

Department of Social Medicine, University of São Paulo, Ribeirão Preto Medical School, Brazil

ARTICLE INFO

Keywords:
Applied mathematics
Computational mathematics
Mathematical biosciences
Bayesian estimates
Cure fraction
Markov chain Monte Carlo (MCMC)
Survival analysis

ABSTRACT

In time-to-event studies it is common the presence of a fraction of individuals not expecting to experience the event of interest; these individuals who are immune to the event or cured for the disease during the study are known as long-term survivors. In addition, in many studies it is observed two lifetimes associated to the same individual, and in some cases there exists a dependence structure between them. In these situations, the usual existing lifetime distributions are not appropriate to model data sets with long-term survivors and dependent bivariate lifetimes. In this study, it is proposed a bivariate model based on a Weibull standard distribution with a dependence structure based on fifteen different copula functions. We assumed the Weibull distribution due to its wide use in survival data analysis and its greater flexibility and simplicity, but the presented methods can be adapted to other continuous survival distributions. Three examples, considering real data sets are introduced to illustrate the proposed methodology. A Bayesian approach is assumed to get the inferences for the parameters of the model where the posterior summaries of interest are obtained using Markov Chain Monte Carlo simulation methods and the Openbugs software. For the data analysis considering different real data sets it was assumed fifteen different copula models from which is was possible to find models with satisfactory fit for the bivariate lifetimes in presence of long-term survivors.

1. Introduction

In medical research, usual parametric and non-parametric tools are widely used for the data analysis of time-to-event data. These tools are useful when some observations are censored and the event of interest has not occurred for all patients at the follow-up time. The procedures most commonly used include the life-table method, the Kaplan-Meier estimator for the survival function, the Cox proportional hazards model, and parametric survival models. These techniques are described in textbooks such as Klein and Moeschberger [1] and Kalbfleisch and Prentice [2].

A common situation in the data analysis of time-to-event data, particularly in cancer research, occurs when it is expected that a fraction of subjects will not experience the event of interest. In this situation, usually it is considered frailty models [3], or it is assumed that the population is a mixture of susceptible individuals who experience the event of interest and non-susceptible individuals that supposedly will never experience it. Statistical methods have been developed to analysis such data, see Lambert et al. [4] and Yu et al. [5]. Following Maller and Zhou [6], a mixture model for these data assumes that the probability that the time-to-event is larger than some specified time t is given by the survival function

$$S(t) = P(T > t) = p + (1 - p)\, S_0(t), \tag{1}$$

where T is a nonnegative random variable denoting the lifetime of an individual, p is a parameter denoting the proportion of "long-term survivors" or "cured patients" $(0 < p < 1)$ and $S_0(t)$ is the baseline survival function for the susceptible individuals. Usual choices for $S_0(t)$ are based on the Weibull, gamma, Rayleigh and lognormal distributions, among many others. In the expression (1) it is observed that $S(t)$ converges to p as t tends to infinite, given that $S_0(t)$ converges to 0 as t tends to infinite. The correspondent probability density function for the lifetime T is given by

$$f(t) = -\frac{dS(t)}{dt} = (1 - p)\, f_0(t),$$

where $f_0(t)$ is the density function for the susceptible individuals.

In the last decades a large number of studies have been developed with the purpose of dealing with bivariate time-to-event data, such as the ones from Hanagal and Bhambure [7] and Emura and Chen [8].

* Corresponding author.
E-mail address: mvperes1991@usp.br (M.V. de Oliveira Peres).

https://doi.org/10.1016/j.heliyon.2020.e03961
Received 30 December 2019; Received in revised form 13 April 2020; Accepted 6 May 2020

Chapter 6

Bivariate modified Weibull distribution derived from Farlie–Gumbel–Morgenstern copula: a simulation study

The developments presented in this Chapter were published in the form of an article in the Electronic Journal of Applied Statistical Analysis, which can be accessed at in https://doi.org/10.1285/i20705948v11n2p463. The Electronic Journal of Applied Statistical Analysis (EJASA) is an open access international journal that publishes articles that contribute new research ideas in all areas of statistical science. The articles published in EJASA are indexed by Scopus, Web of Science ESCI, and other abstracting databases.

In this paper, we introduced a bivariate model for lifetime data analysis based on the modified Weibull distribution derived from the Farlie-Gumbel-Morgenstern (FGM). An extensive simulation study was developed, a comparison between maximum likelihood and Bayesian approaches is presented and the performance of the model was evaluated using the real data. The example showed satisfactory performance of the method.

Electronic Journal of Applied Statistical Analysis
Vol. 11, Issue 02, October 2018, 463-488
DOI: 10.1285/i20705948v11n2p463

Bivariate modified Weibull distribution derived from Farlie-Gumbel-Morgenstern copula: a simulation study

Marcos Vinicius de Oliveira Peres*, Jorge Alberto Achcar, and Edson Zangiacomi Martinez

Department of Social Medicine, University of São Paulo, Ribeirão Preto Medical School, Brazil

Published: 14 October 2018

In recent years, the use of copulas has grown rapidly, especially in survival analysis. In this paper, we introduce a bivariate modified Weibull distribution derived from the Farlie–Gumbel–Morgenstern (FGM), a copula function commonly used to model very weak linear dependences. Considering the presence of non censored data and censored data, an extensive simulation study was developed to check the performance of the maximum likelihood method in estimating the parameters of the proposed model. Maximum likelihood and Bayesian approaches for the estimation of the model parameters are presented. In the Bayesian analysis, the posterior distributions of the parameters are estimated using Markov chain Monte Carlo (MCMC) methodology. An example, considering a real data set, is introduced to illustrate the proposed methodology.

keywords: Bayesian estimates, bivariate data, copula function, simulation study, survival analysis.

1 Introduction

In the lifetime data analysis, researchers commonly use standard non-parametric techniques, as for example, the Kaplan–Meier estimators for the survival function, the log-rank test or semi-parametrical Cox proportional hazard models (Kleinbaum and Klein,

*Corresponding author: mvperes1991@usp.br

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

Chapter 7

The Bivariate Defective Gompertz Distribution Based on Clayton Copula with Applications to Medical Data

The developments presented in this Chapter were published in the form of an e-print archive in the arXiv, which can be accessed at in http://arxiv.org/abs/2012.07824. The arXiv is a free distribution service and an open-access archive in the diverse scientific areas.

In this paper, we introduced the bivariate model based on the defective Gompertz distribution derived from the Clayton Copula. We carry out an extensive simulation study, in order to evaluate the biases and the mean squared errors for the maximum likelihood estimators of the parameters associated to the proposed model. Some applications using medical data are presented to show the usefulness of the proposed model. The simulation and examples showed satisfactory performance of the method.

.

The Bivariate Defective Gompertz Distribution Based on Clayton Copula with Applications to Medical Data

Marcos Vinicius de Oliveira Peres(1) **Ricardo Puziol de Oliveira**(2)
Jorge Alberto Achcar(1) **Edson Zangiacomi Martinez**(1)

(1) Ribeirão Preto Medical School, University of Sao Paulo (USP), Ribeirao Preto, SP, Brazil
(2) Department of the Environment, State University of Maringá (UEM), Umuarama, PR, Brazil.

Abstract

In medical studies, it is common the presence of a fraction of patients who do not experience the event of interest. These patients are people who are not at risk of the event or are patients who were cured during the research. The proportion of immune or cured patients is known in the literature as cure rate. In general, the traditional existing lifetime statistical models are not appropriate to model data sets with cure rate, including bivariate lifetimes. In this paper, it is proposed a bivariate model based on a defective Gompertz distribution and also using a Clayton copula function to capture the possible dependence structure between the lifetimes. An extensive simulation study was carried out in order to evaluate the biases and the mean squared errors for the maximum likelihood estimators of the parameters associated to the proposed distribution. Some applications using medical data are presented to show the usefulness of the proposed model.

Keywords: Clayton copula, cure rate, defective Gompertz distribution, survival analysis.

1 Introduction

The use of survival statistical models for time-to-event data is common in several areas of study, especially in medical research. Traditional parametric and non-parametric tools, such as, Kaplan-Meier estimator for the survival function, log-rank and Wilcoxon tests and the semi-parametric Cox proportional hazard model, are widely used in medical data analysis (see, e.g., Kleinbaum and Klein, 2012). These methods assume that all individuals are susceptible to the event of interest. However, for example in clinical studies, there may be patients who will not experience the event under investigation, that is, these patients are immune to the event or they

Chapter 8

General conclusions

The main contributions of this Thesis consist of three scientific articles. In the first article, described in Chapter 5, it was considered fifteen different copula functions to construct bivariate standard Weibull lifetime distributions in presence of long-term survivors and applied for three different real data sets. The models considered in this study showed reasonable fit for the data in all considered applications. The estimated survival curves were close to the correspondent empirical Kaplan and Meier survival curves. For the data sets assumed in this article, it was also observed good estimators for the percentages of cure fraction in all applications. The considered models were able to measure the correlation between the both lifetimes also satisfactory measuring the covariance between the cure fractions associated to the lifetimes T_1 and T_2. The Weibull distribution was chosen in this article due to its versatility and relative simplicity, but the proposed methodology introduced in this study could be adapted to other continuous distributions assumed for the lifetime data. The use of the Bayesian approach and MCMC methods was considered due to its flexibility and effectiveness in the estimation of the parameters of more complex models. It is believed that there is also merit for the possibility of using the prior distributions for the model parameters to restrict the estimates with the corresponding parametric space.

After reviewing various copula functions that can be used for bivariate survival data, the article Peres et al. (2020) shows that the choice of the most suitable copula function usually is complicated, varying according to the data set. Among the various copula functions presented in the literature, it results that some can produce equivalent results; therefore, it is recommended in each application to use several selection procedures to see which copula is the best as considered in this study. The empirical Kaplan-Meier bivariate lifetime function estimator considered in this study is of easy interpretation and implementation. However, it was observed that this estimator in some lifetimes (especially the rightmost times) has a tendency to estimate the fraction of survivals near zero, discarding these points of a possible cure fraction; this fact is evidenced when observing the values of the absolute maximum distances where these maximum distances were close to the final lifetimes. However, the comparison between the distances of the empirical and estimated matrices gives an additional form to compare the fit of the proposed models.

The second article introduces a bivariate modified Weibull (BMW) distribution derived from the Farlie-Gumbel-Morgenstern (FGM), a copula function commonly used to model very weak linear dependences (Peres et al., 2018). This article is described in Chapter 6. Under the new BMW model, it was developed an extensive simulation study showing the performance of the obtained inference results under classical maximum likelihood and Bayeasian approaches. The simulation study showed that the maximum likelihood and Bayesian method are suitable approaches to estimate the parameters of the BMW distribution. However, in the situations where there is a high proportion of censored data (higher than 70%) and small sample sizes we do not recommend the use of this distribution. We also observed that the estimates are more easily obtained if the lifetimes are smaller values than 10. The applications with simulated and real data showed that BWM distribution can be satisfactorily.

In the third article, for now published in the form of an e-print archive in the arXiv, it is proposed a bivariate model based on a defective Gompertz (BDGD) distribution and using a Clayton copula function to capture the possible dependence structure between the lifetimes. The maximum likelihood method using existing numerical optimization algorithms was considered to get the inferences of interest under a frequentist approach and MCMC (Markov Chain Monte Carlo) simulation methods, as the popular Gibbs sampling and Metropolis-Hastings algorithms, were used to get the posterior summaries of interest under a Bayesian approach. Considering this new model, we performed a comprehensive simulation study to describe the performance of the inference results under the maximum likelihood approach. This simulation study suggested that the model is efficient to fit data with weak and strong correlation between lifetimes the T_1 and T_2 in several scenarios. However, in the situations where there is a high proportion of cure fraction and small sample sizes ($n < 100$) a careful use of this model is required. It was observed that the model estimates are more easily obtained if the lifetime variables have values lower than 20, which demands some transformation in the data in some applications. In the application studies, it was verified that both the maximum likelihood and Bayesian methods are suitable approaches to estimate the parameters of the BDGD model. It is important to point out that a suitable choice for the initial values in the maximum likelihood iterative estimation procedure is required, as well as the Bayesian method depends on adequate hyperparameter values for the prior probability distributions for the parameters of the BDGD model. Applications in simulated and real data evidenced that the BDGD model can be satisfactorily fitted in most cases, considering both the maximum likelihood and Bayesian approaches.

In view of all the results obtained, we can draw as general conclusions that the copula functions can be very useful tools in clinical and epidemiological studies that use bivariate survival data. However, the proper choice of a copula function for data modeling is a task that requires care, since improper choice of a copula can lead to spurious results. In this way, the articles presented in this Thesis are a useful guidance to researchers who intend to use copula functions in medical studies based on bivariate time-to-event data. In addition to the appropriate choice of a copula function, these articles show that both the maximum likelihood and Bayesian approaches are useful to estimate the parameters of a bivariate distribution based on copula functions. In the first article, the Bayesian approach and MCMC methods was considered due to its flexibility and effectiveness in estimating the parameters of more complex models, but in the two other articles, it is showed that the maximum likelihood method can also provide good fit to the data. These articles take into account important considerations on the use of small samples, according to the results of the simulation studies. As a final note, we can observe that the increasing availability of faster computers as well as the development of statistical packages such as R, *JAGS* and *OpenBugs* has greatly facilitated our applications of copula functions to real data.

Bibliography

Achcar, J. A., Coelho-Barros, E. A., and Mazucheli, J. (2012). Cure fraction models using mixture and non-mixture models. *Tatra Mountains Mathematical Publications*, 51(1):1–9.

Achcar, J. A., Martinez, E. Z., and Tovar Cuevas, J. R. (2016). Bivariate lifetime modelling using copula functions in presence of mixture and non-mixture cure fraction models, censored data and covariates. *Model Assisted Statistics and Applications*, 11(4):261–276.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

Akritas, M. G. and Keilegom, I. V. (2003). Estimation of bivariate and marginal distributions with censored data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):457–471.

Ali, M. M., Mikhail, N., and Haq, M. S. (1978). A class of bivariate distributions including the bivariate logistic. *Journal of Multivariate Analysis*, 8(3):405–412.

Amico, M., Van Keilegom, I., and Legrand, C. (2019). The single-index/Cox mixture cure model. *Biometrics*, 75(2):452–462.

Bae, W., Choi, H., Park, B. U., and Kim, C. (2005). Smoothing techniques for the bivariate Kaplan–Meier estimator. *Communications in Statistics-Theory and Methods*, 34(7):1659–1674.

Bairamov, I. and Kotz, S. (2002). Dependence structure and symmetry of Huang-Kotz FGM distributions and their extensions. *Metrika*, 56(1):55–72.

Balakrishnan, N. and Lai, C. (2009). *Continuous Bivariate Distributions*. Springer, New York.

Balka, J., Desmond, A. F., and McNicholas, P. D. (2011). Bayesian and likelihood inference for cure rates based on defective inverse gaussian regression models. *Journal of Applied Statistics*, 38(1):127–144.

Barnett, V. (1980). Some bivariate uniform distributions. *Communications in Statistics-Theory and Methods*, 9(4):453–461.

Basu, A. P. and Dhar, S. (1995). Bivariate geometric distribution. *Journal of Applied Statistical Science*, 2(12):33–34.

Berger, J. O. (2013). *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media.

Block, H. W. and Basu, A. (1974). A continuous bivariate exponential extension. *Journal of the American Statistical Association*, 69(348):1031–1037.

Box, G. E. and Tiao, G. C. (2011). *Bayesian inference in statistical analysis*, volume 40. John Wiley & Sons.

Brenna, S. M. F., Silva, I. D., Zeferino, L. C., Pereira, J. S., Martinez, E. Z., and Syrjänen, K. J. (2004). Prognostic value of p53 codon 72 polymorphism in invasive cervical cancer in Brazil. *Gynecologic Oncology*, 93(2):374–380.

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

Cambell, G. and Földes, A. (1980). *Large-sample properties of nonparametric bivariate estimators with censored data*. Purdue University. Department of Statistics.

Cancho, V. G. and Bolfarine, H. (2001). Modeling the presence of immunes by using the exponentiated-Weibull model. *Journal of Applied Statistics*, 28(6):659–671.

Cantor, A. B. and Shuster, J. J. (1992). Parametric versus non-parametric methods for estimating cure rates based on censored survival data. *Statistics in Medicine*, 11(7):931–937.

Carlin, B. P. and Louis, T. A. (2000). *Bayes and empirical Bayes methods for data analysis*. Chapman & Hall/CRC Boca Raton.

Casella, G. and Berger, R. L. (2002). *Statistical inference*, volume 2. Duxbury Pacific Grove, CA.

Casella, G. and George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, 46(3):167–174.

Chen, M.-H., Shao, Q.-M., and Ibrahim, J. G. (2012). *Monte Carlo methods in Bayesian computation*. Springer Science & Business Media.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The american statistician*, 49(4):327–335.

Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65(1):141–151.

Coelho-Barros, E. A., Achcar, J. A., and Mazucheli, J. (2016). Bivariate Weibull distributions derived from copula functions in the presence of cure fraction and censored data. *Journal of Data Science*, 14(2).

Collett, D. (2015). *Modelling survival data in medical research*. CRC press.

Colosimo, E. A. and Giolo, S. R. (2006). *Análise de sobrevivência aplicada*. Editora Blucher.

Cook, R. D. and Johnson, M. E. (1981). A family of distributions for modelling non-elliptically symmetric multivariate data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 43(2):210–218.

Cordeiro, G. M., Cancho, V. G., Ortega, E. M., and Barriga, G. D. (2016). A model with long-term survivors: negative binomial Birnbaum-Saunders. *Communications in Statistics-Theory and Methods*, 45(5):1370–1387.

Cordeiro, G. M. and Demétrio, C. G. (2013). Modelos lineares generalizados e extensoes. *Recife: UFRPE*.

Cowles, M. K. and Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, 91(434):883–904.

Cox, C. (2005). Delta method. *Encyclopedia of biostatistics*, 2.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B*, 34:187–220.

Cox, D. R. (2006). *Principles of statistical inference*. Cambridge University Press.

da Rocha, R. F., Tomazella, L. D., and Louzada, F. (2014). Bayesian and classic inference for the defective Gompertz cure rate model.

Dabrowska, D. M. et al. (1988). Kaplan-Meier estimate on the plane. *The Annals of Statistics*, 16(4):1475–1489.

de Oliveira, R. P. and Achcar, J. A. (2018). Basu-Dhar's bivariate geometric distribution in presence of censored data and covariates: some computational aspects. *Electronic Journal of Applied Statistical Analysis*, 11(1):108–136.

Dey, D. K., Chen, M.-H., and Chang, H. (1997). Bayesian approach for nonlinear random effects models. *Biometrics*, 53(4):1239–1252.

dos Santos, M. R., Achcar, J. A., and Zangiacomi Martinez, E. (2017). Bayesian and maximum likelihood inference for the defective Gompertz cure rate model with covariates: an application to the cervical carcinoma study. *Ciência e Natura*, 39(2).

Durante, F. and Sempi, C. (2015). *Principles of copula theory*. CRC press.

Embrechts, P., Lindskog, F., and McNeil, A. (2001). Modelling dependence with copulas. *Rapport technique, Département de mathématiques, Institut Fédéral de Technologie de Zurich, Zurich*, 14.

Emura, T. and Chen, Y.-H. (2018). *Analysis of Survival Data with Dependent Censoring*. Springer.

Farlie, D. J. (1960). The performance of some correlation coefficients for a general bivariate distribution. *Biometrika*, 47(3/4):307–323.

Fischer, M. and Köck, C. (2012). Constructing and generalizing given multivariate copulas: A unifying approach. *Statistics*, 46(1):1–12.

Fisher, R. A. (1912). On an absolute criterion for fitting frequency curves. *Messenger of Mathematics*, 41:155–156.

Frank, M. J. (1979). On the simultaneous associativity of F(x, y) and x+y-F(x, y). *Aequationes mathematicae*, 19(1):194–226.

Fréchet, M. (1927). Sur la loi de probabilité de l'écart maximum. *Annales de la Société Polonaise de Mathématique*, 6:93–116.

Fredricks, G. A. and Nelsen, R. B. (2007). On the relationship between Spearman's rho and Kendall's tau for pairs of continuous random variables. *Journal of Statistical Planning and Inference*, 137(7):2143 – 2150.

Frees, E. W. and Valdez, E. A. (1998). Understanding relationships using copulas. *North American Actuarial Journal*, 2(1):1–25.

Galambos, J. (1975). Order statistics of samples from multivariate distributions. *Journal of the American Statistical Association*, 70(351a):674–680.

Gallardo, D. I., Gómez, Y. M., and de Castro, M. (2018). A flexible cure rate model based on the polylogarithm distribution. *Journal of Statistical Computation and Simulation*, 88(11):2137–2149.

Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, 74(365):153–160.

Gelfand, A. E., Dey, D. K., and Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. Technical report, DTIC Document.

Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian data analysis*. Chapman and Hall/CRC.

Gelman, A., Rubin, D. B., et al. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472.

Genest, C. (1987). Frank's family of bivariate distributions. *Biometrika*, 74(3):549–555.

Genest, C. and MacKay, R. J. (1986). Copules archimédiennes et families de lois bidimensionnelles dont les marges sont données. *Canadian Journal of Statistics*, 14(2):145–159.

Genest, C. and Rivest, L.-P. (1989). A characterization of Gumbel's family of extreme value distributions. *Statistics & Probability Letters*, 8(3):207–211.

Geweke, J. et al. (1991). *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*, volume 196. Federal Reserve Bank of Minneapolis, Research Department Minneapolis, MN, USA.

Gieser, P. W., Chang, M. N., Rao, P., Shuster, J. J., and Pullen, J. (1998). Modelling cure rates using the Gompertz model with covariate information. *Statistics in medicine*, 17(8):831–839.

Gilks, W. R. (2005). *Markov chain Monte Carlo*. Wiley Online Library.

Group, D. R. S. R. et al. (1976). Preliminary report on effects of photocoagulation therapy. *American Journal of Ophthalmology*, 81(4):383–396.

Gumbel, E. J. (1960a). Bivariate exponential distributions. *Journal of the American Statistical Association*, 55(292):698–707.

Gumbel, E. J. (1960b). Distributions des valeurs extremes en plusiers dimensions. *Publ. Inst. Statist. Univ. Paris*, 9:171–173.

Hanley, J. A. and Parnes, M. N. (1983). Nonparametric estimation of a multivariate distribution in the presence of censoring. *Biometrics*, pages 129–139.

Heidelberger, P. and Welch, P. D. (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, 31(6):1109–1144.

Hougaard, P. (1986). A class of multivanate failure time distributions. *Biometrika*, 73(3):671–678.

Hougaard, P. (1987). Modelling multivariate survival. *Scandinavian Journal of Statistics*, 14(4):291–304.

Hougaard, P. (2012). *Analysis of multivariate survival data*. Springer Science & Business Media.

Huang, J. (2019). *Bivariate Cure Rate Model Using Copula Functions in Presence of Censored Data and Covariates*. Northern Illinois University.

Huang, J. S. and Kotz, S. (1999). Modifications of the Farlie-Gumbel-Morgenstern distributions. a tough hill to climb. *Metrika*, 49(2):135–145.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2005). *Bayesian survival analysis*. Wiley Online Library.

Jarjoura, D. (1988). Smoothing hazard rates with cubic splines. *Communications in Statistics-Simulation and Computation*, 17(2):377–392.

Jaworski, P., Durante, F., Hardle, W. K., and Rychlik, T. (2010). *Copula theory and its applications*. Springer.

Joe, H. (1993). Parametric families of multivariate distributions with given margins. *Journal of multivariate analysis*, 46(2):262–282.

Joe, H. (1997). *Multivariate Models and Multivariate Dependence Concepts*. Chapman and Hall/CRC Press.

Joe, H. (2014). *Dependence Modeling with Copulas*. Chapman and Hall/CRC Press.

John, G., Louis, C., Berner, A., and Genné, D. (2015). Tobacco stained fingers and its association with death and hospital admission: A retrospective cohort study. *PloS one*, 10(9):e0138211.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. Wiley, New York, NY, 2nd edition.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481.

Kijima, M. (2013). *Markov processes for stochastic modeling.* Springer.

Kumar, P. (2010). Probability distributions and estimation of Ali-Mikhail-Haq copula. *Applied Mathematical Sciences*, 4(14):657–666.

Lai, C., Xie, M., and Murthy, D. (2003). A modified Weibull distribution. *IEEE Transactions on Reliability*, 52(1):33–37.

Lambert, P. C. (2007). Modeling of the cure fraction in survival studies. *Stata Journal*, 7(3):351.

Lambert, P. C., Thompson, J. R., Weston, C. L., and Dickman, P. W. (2006). Estimating and modeling the cure fraction in population-based cancer survival analysis. *Biostatistics*, 8(3):576–594.

Lawless, J. F. (2011). *Statistical models and methods for lifetime data.* John Wiley & Sons.

Leão, J., Bourguignon, M., Gallardo, D. I., Rocha, R., and Tomazella, V. (2020). A new cure rate model with flexible competing causes with applications to melanoma and transplantation data. *Statistics in Medicine*, 39(24):3272–3284.

Lee, E. T. and Wang, J. W. (2003). *Statistical methods for survival data analysis.* Wiley Series in Probability and Statistics. Hoboken, NJ, third edition.

Liang, F., Liu, C., and Carroll, R. (2011). *Advanced Markov chain Monte Carlo methods: learning from past samples*, volume 714. John Wiley & Sons.

Liang, K.-Y., Self, S. G., Bandeen-Roche, K. J., and Zeger, S. L. (1995). Some recent developments for regression analysis of multivariate failure time data. *Lifetime Data Analysis*, 1(4):403–415.

Lin, D. and Ying, Z. (1993). A simple nonparametric estimator of the bivariate survival function under univariate censoring. *Biometrika*, 80(3):573–581.

Liu, X. (2012). *Survival analysis: models and applications.* John Wiley & Sons.

Louzada, F., Suzuki, A., and Cancho, V. (2013). The fgm long-term bivariate survival copula model: modeling, bayesian estimation, and case influence diagnostics. *Communications in Statistics-Theory and Methods*, 42(4):673–691.

Maller, R. A. and Zhou, X. (1996). *Survival analysis with long-term survivors.* Wiley New York.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, 50:163–170.

Marshall, A. W. and Olkin, I. (1967). A generalized bivariate exponential distribution. *Journal of Applied Probability*, 4(02):291–302.

Martinez, E. Z. and Achcar, J. A. (2014). Bayesian bivariate generalized Lindley model for survival data with a cure fraction. *Computer methods and programs in biomedicine*, 117(2):145–157.

Martinez, E. Z. and Achcar, J. A. (2017). The defective generalized Gompertz distribution and its use in the analysis of lifetime data in presence of cure fraction, censored data and covariates. *Electronic Journal of Applied Statistical Analysis*, 10(2):463–484.

Meeker, W. Q. and Escobar, L. A. (1998). *Statistical Methods for Reliability Data*. John Wiley & Sons, New York.

Meira-Machado, L., Sestelo, M., and Gonçalves, A. (2016). Nonparametric estimation of the survival function for ordered multivariate failure time data: A comparative study. *Biometrical Journal*, 58(3):623–634.

Meyer, R. and Romeo, J. S. (2015). Bayesian semiparametric analysis of recurrent failure time data using copulas. *Biometrical Journal*, 57(6):982–1001.

Meyn, S. P. and Tweedie, R. L. (2012). *Markov chains and stochastic stability*. Springer Science & Business Media.

Millar, R. B. (2011). *Maximum likelihood estimation and inference: with examples in R, SAS and ADMB*. John Wiley & Sons.

Morgenstern, D. (1956). Einfache beispiele zweidimensionaler verteilungen. *Mitteilingsblatt fur Mathematische Statistik*, 8:234–235.

Muraleedharan Nair, K. and Unnikrishnan Nair, N. (1988). On characterizing the bivariate exponential and geometric distributions. *Annals of the institute of Statistical Mathematics*, 40(2):267–271.

Nelsen, R. B. (1986). Properties of a one-parameter family of bivariate distributions with specified marginals. *Communications in Statistics-Theory and Methods*, 15(11):3277–3285.

Nelsen, R. B. (1994). A characterization of Farlie-Gumbel-Morgenstern distributions via Spearman's rho and chi-square divergence. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 476–479.

Nelsen, R. B. (2007). *An introduction to copulas*. Springer Science & Business Media.

Oakes, D. (1982). A model for association in bivariate survival data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 414–422.

Oehlert, G. W. (1992). A note on the delta method. *The American Statistician*, 46(1):27–29.

Paola Rebora, A. S. and Reilly, M. (2018). *bshazard: Nonparametric Smoothing of the Hazard Function*. R package version 1.1.

Parner, E. (2001). A composite likelihood approach to multivariate survival data. *Scandinavian Journal of Statistics*, 28(2):295–302.

Peng, Y., Dear, K. B., and Denham, J. (1998). A generalized F mixture model for cure rate estimation. *Statistics in Medicine*, 17(8):813–830.

Peres, M. V. d. O., Achcar, J. A., and Martinez, E. Z. (2018). Bivariate modified Weibull distribution derived from Farlie-Gumbel-Morgenstern copula: a simulation study. *Electronic Journal of Applied Statistical Analysis*, 11(2):463–488.

Peres, M. V. d. O., Achcar, J. A., and Martinez, E. Z. (2020). Bivariate lifetime models in presence of cure fraction: a comparative study with many different copula functions. *Heliyon*, 6(6):e03961.

Pescim, R. R., Ortega, E. M., Suzuki, A. K., Cancho, V. G., and Cordeiro, G. M. (2019). A new destructive Poisson odd log-logistic generalized half-normal cure rate model. *Communications in Statistics-Theory and Methods*, 48(9):2113–2128.

Pickands, J. (1981). Multivariate extreme value distributions (with discussion). *Bulletin of the International Statistical Institute*, pages 859–878.

Plackett, R. L. (1965). A class of bivariate distributions. *Journal of the American Statistical Association*, 60(310):516–522.

Prentice, R. (2014). Self-consistent nonparametric maximum likelihood estimator of the bivariate survivor function. *Biometrika*, 101(3):505–518.

Prentice, R. L. and Cai, J. (1992). Covariance and survivor function estimation using censored multivariate failure time data. *Biometrika*, 79(3):495–512.

Price, D. L. and Manatunga, A. K. (2001). Modelling survival data with a cured fraction using frailty models. *Statistics in medicine*, 20(9-10):1515–1527.

Rausand, M. and Arnljot, H. (2004). *System reliability theory: models, statistical methods, and applications*, volume 396. John Wiley & Sons.

Rayleigh, L. and Strutt, J. W. (1919). On the problem of random vibrations, and of random flights in one, two, or three dimensions. *Philosophical Magazine*, 37:321–347.

Rebora, P., Salim, A., and Reilly, M. (2014). bshazard: A Flexible Tool for Nonparametric Smoothing of the Hazard Function. *The R Journal*, 6(2):114–122.

Robert, C. P. and Casella, G. (1999). The Metropolis-Hastings algorithm. In *Monte Carlo Statistical Methods*, pages 231–283. Springer.

Roberts, G. O. and Smith, A. F. (1994). Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. *Stochastic Processes and their Applications*, 49(2):207–216.

Rocha, R., Nadarajah, S., Tomazella, V., Louzada, F., and Eudes, A. (2017). New defective models based on the Kumaraswamy family of distributions with application to cancer data sets. *Statistical methods in medical research*, 26(4):1737–1755.

Rodrigues, J., de Castro, M., Balakrishnan, N., and Cancho, V. G. (2011). Destructive weighted poisson cure rate models. *Lifetime data analysis*, 17(3):333–346.

Rodrigues, J., de Castro, M., Cancho, V. G., and Balakrishnan, N. (2009). Com-Poisson cure rate survival models and an application to a cutaneous melanoma data. *Journal of Statistical Planning and Inference*, 139(10):3605–3611.

Rosa, G. J. M. (1998). *Análise Bayesiana de modelos lineares mistos robustos via Amostrador de Gibbs*. Doutorado em estatística e experimentação agronômica, Escola Superior de Agricultura Luiz de Queiroz da Universidade de São Paulo, Piracicaba.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Schweizer, B., Wolff, E. F., et al. (1981). On nonparametric measures of dependence for random variables. *The Annals of Statistics*, 9(4):879–885.

Selvin, S. (2008). *Survival analysis for epidemiologic and medical research*. Cambridge University Press.

Shigemizu, D., Iwase, T., Yoshimoto, M., Suzuki, Y., Miya, F., Boroevich, K. A., Katagiri, T., Zembutsu, H., and Tsunoda, T. (2017). The prediction models for postoperative overall survival and disease-free survival in patients with breast cancer. *Cancer medicine*, 6(7):1627–1638.

Shih, J.-H. and Emura, T. (2016). Bivariate dependence measures and bivariate competing risks models under the generalized FGM copula. *Statistical Papers*, pages 1–18.

Siddiqui, M. (1962). Some problems connected with Rayleigh distributions. *Journal of Research of the National Bureau of Standards D*, 66:167–174.

Sklar, A. (1996). Random variables, distribution functions, and copulas: a personal look backward and forward. *Lecture notes-monograph series*, pages 1–14.

Sklar, M. (1959). Fonctions de repartition an dimensions et leurs marges. *Publications de l'Institut de statistique de l'Université de Paris*, 8:229–231.

Sorensen, D. (1996). Gibbs sampling in quantitative genetics. *Intern Rapport. Statens Husdyrbrugsforsoeg (Denmark). no. 82.*

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

Sugiura, N. (1978). Further analysts of the data by Akaike's information criterion and the finite corrections: Further analysts of the data by Akaike's. *Communications in Statistics-Theory and Methods*, 7(1):13–26.

Suzuki, A. K., Louzada-Neto, F., Cancho, V. G., and Barriga, G. D. (2011). The FGM bivariate lifetime copula model: a Bayesian approach. *Advances and Applications in Statistics*, 21(1):55–76.

Trivedi, P. K. and Zimmer, D. M. (2007). Copula modeling: An introduction for practitioners. *Foundations and Trends in Econometrics*, 1(1):1–111.

Tsai, W.-Y., Leurgans, S., Crowley, J., et al. (1986). Nonparametric estimation of a bivariate survival function in the presence of censoring. *The Annals of Statistics*, 14(4):1351–1365.

Tsodikov, A., Ibrahim, J., and Yakovlev, A. (2003). Estimating cure rates from survival data: an alternative to two-component mixture models. *Journal of the American Statistical Association*, 98(464):1063–1078.

Vahidpour, M. (2016). *Cure Rate Models*. PhD thesis, École Polytechnique de Montréal.

Van Der Laan, M. J. (1997). Nonparametric estimators of the bivariate survival function under random censoring. *Statistica Neerlandica*, 51(2):178–200.

Vaupel, J. W., Manton, K. G., and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16(3):439–454.

Ver Hoef, J. M. (2012). Who invented the delta method? *The American Statistician*, 66(2):124–127.

Viswanathan, B. and Manatunga, A. K. (2001). Diagnostic plots for assessing the frailty distribution in multivariate survival data. *Lifetime Data Analysis*, 7(2):143–155.

Weibull, W. et al. (1951). A statistical distribution function of wide applicability. *Journal of Applied Mechanics*, 18(3):293–297.

Wienke, A., Locatelli, I., and Yashin, A. I. (2006). The modelling of a cure fraction in bivariate time-to-event data. *Austrian Journal of Statistics*, 35(1):67–76.

Yu, B., Tiwari, R. C., Cronin, K. A., and Feuer, E. J. (2004). Cure fraction estimation from the mixture cure models for grouped survival data. *Statistics in medicine*, 23(11):1733–1747.

Zeger, S. L. and Karim, M. R. (1991). Generalized linear models with random effects; a Gibbs sampling approach. *Journal of the American Statistical Association*, 86(413):79–86.