NTERNATIONAL JOURNAL OF MATHEMATICS AND COMPUTERS IN SIMULATION
DOI: 10.46300/9102.2020.14.5 Volume 14, 2020

Some of the most common copulas for simulating complex survival data

Gustavo Soutinho
EPIUnit - University of Porto
Rua das Taipas n.º 135, 4050-600 Porto
Portugal

Luís Meira-Machado
University of Minho
Campus de Azurém, 4800-050 Guimarães
Portugal

Received: June 23, 2020. Revised: August 31, 2020. 2nd Revised: September 21, 2020.
Accepted: September 22, 2020. Published: September 23, 2020.

Abstract- Simulation studies play an important role in the evaluation of the performance of a variety of statistical methods. Such assessment is performed under computer intensive procedures and cannot be achieved with studies of real data alone. These studies are increasingly employed in evaluating the properties of the proposed methods being the generation of data the most fundamental and important component. However, only a few of published studies provide sufficient details to allow readers to understand fully all the processes to generate the data. In this paper we present a collection of practical algorithms for simulating multivariate data from a wide class of multivariate copulas. This paper also details important considerations necessary when generating the survival data in a variety of scenarios. A software application for R was developed in which we implement all the methods.

Keywords- Copulas, Archimedean copulas, Random number generation, Inversion of Laplace transforms, Survival data.

I. Introduction

Recent advances in computer and software technology have allowed simulation studies to be more accessible. However, performing simulations is not a simple issue. Important guidelines to achieve a good quality simulation study are given by [6]. Data generation is probably the most important step to achieve a good quality simulation study and require a rigorous planning. Unfortunately, only few published articles provide sufficient details to assess the integrity of the study design or to allow readers to understand fully all the processes required when designing their own simulation study. In addition, it is important to obtain simple and high-quality simulations that reflect the complex situations seen in practice, such as, for example, for survival data.

Longitudinal survival data often require the joint modeling of two or more random variables. For example, to model the relationship between survival time of a patient and the hemoglobin level; to model the relationship between two consecutive events of the same nature (recurrent events) or to model different stages in the evolution of an illness (multi-state models). Simulating data for such studies is a challenging issue that can be performed using copulas. Copulas provide a useful method for deriving joint distributions given the marginal distributions, especially when the variables are non-normal as in the case of time-to-event variables. In addition, in a bivariate context, copulas can be used to easily control the measures of dependence for the pairs of random variables.

A copula C is a multivariate distribution function that links a univariate marginal distribution to their full multivariate distribution. Copulas were first introduced by [36] and its terminology is derived from the Latin word copulare, to connect or to join. Our objective in this paper is to present algorithms to generate 2-dimensional random vectors (X, Y) whose distribution is $H(x, y) = C(F(x), G(y))$ where F and G denote the marginal distribution function and C is a copula. We will illustrate the usefulness of these methods to generate survival data in a variety of scenarios.

This paper explores the topic of random generation in several families of copulas. In addition, we describe basic properties of copulas, their relationships to measures of dependence, and some of the most known families of copulas that have appeared in the literature. One important aim of this work is to present several algorithms for the generation of multivariate survival data from several copulas. These algorithms are based on three of the most used techniques for generating multivariate data from copulas: the conditional distribution method; based on the bivariate distribution of the copula or sampling algorithms based on numerical inversion of Laplace transforms. A conceptual framework of these algorithms is presented in Figure 1.

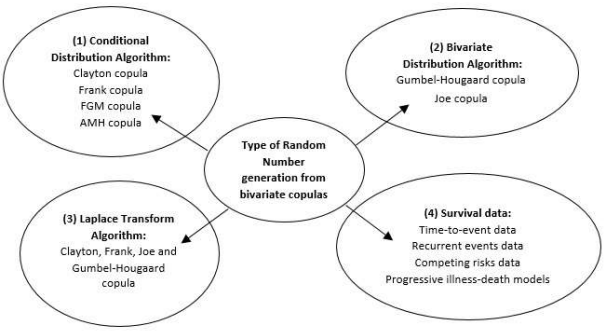

Fig. 1: Copulas and Random Number Generation

The purpose of this paper is to introduce copulas, their characteristics and properties, and their applicability to simulate multivariate survival data. Section 2 discusses properties and characteristics of copulas. Section 3 provides practical algorithms for simulating data from a wide class of multivariate copulas. Sampling algorithms are also given to simulate multivariate survival data in a variety of scenarios. Section 4 briefly presents an R package to implement all the methods. A discussion of the main conclusions of this work and some future research are reported in Section 5.

II. Most common bivariate copulas

A. Definitions and properties of copulas

Copulas are functions that link multivariate distributions to their one-dimensional margins. These functions are restrictions to $[0,1]^2$ of bivariate distribution functions whose margins are uniform in $[0,1]$. [36] showed that if H is a bivariate distribution function with margins $F(x)$ and $G(y)$, then there exists a copula C such that $H(x,y) = C(F(x), G(y))$. Sklar also showed that if the marginal distributions are continuous, then there is a unique copula representation. In the multivariable case, if H is an p-dimensional cumulative distribution function with univariate margins $F_1, ..., F_p$, then there exists an p-dimensional copula C such that $F(x_1, ..., x_p) = C(F_1(x_1), ..., F_p(x_p))$. The case $p = 2$ has attracted special attention and will be considered from now on.

A function $\varphi : [0,1] \to [0,\infty]$ is called a *generator* if it is convex, decreasing and $\varphi(1) = 0$. The generalized inverse of φ (also known as pseudo-inverse) is denoted by $\varphi^{[-1]} = \inf\{u \in [0,1] \mid \varphi(u) \leq t\}$, $t \in [0,\infty]$.

A copula C is called Archimedean if there exists a generator φ such that $C(u,v) = \varphi^{]-1[}(\varphi(u) + \varphi(v)), (u,v) \in [0,1]^2$. The copula C determines the generator φ uniquely up to a multiplicative constant. In Table 1 we present the different choices of generator for several important families of Archimedean copulas.

Archimedean copulas are popular because they are easily derived and are capable of capturing wide ranges of dependence. Given a pair of variables (X,Y) whose distribution is H, and C the associated copula, this dependence can be measured by Kendall's tau τ or Spearman's ρ. Kendall's tau can be defined as the difference between the probabilities of concordance and discordance for any two independent pairs. In terms of copulas, Kendall's τ is defined by

$$\tau_C = 4 \int_0^1 \int_0^1 C(u,v) dC(u,v) - 1.$$

The Spearman's ρ coefficient is defined as

$$\rho_C = 12 \int_0^1 \int_0^1 (C(u,v) - uv) du dv.$$

Table 2 illustrates the calculation of these correlation measures.

Modeling the multivariate dependence also involves quantifying tail-dependence. Tail-dependence describes the concordance between extreme values of the random

Family	Space Parameter	Generator $\varphi(t)$	Generator inverse $\varphi^{-1}(s)$	Bivariate copula $C(u,v)$
[7]	$\theta \in (0,\infty]$	$\frac{1}{\theta}(t^{-\theta}-1)$	$(1+\theta s)^{-1/\theta}$	$(u^{-\theta}+v^{-\theta}-1)^{-1/\theta}$
[13]	$\theta \in \mathbb{R} \setminus \{0\}$	$-\ln\left[\frac{e^{-\theta t}-1}{e^{-\theta}-1}\right]$	$-\frac{1}{\theta}\ln(1+e^{-s}(e^{-\theta}-1))$	$-\theta^{-1}\ln\left[1+\frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1}\right]$
[16]	$\theta \in [1,\infty)$	$(-\ln t)^{\theta}$	$e^{-s^{1/\theta}}$	$exp\{-[(-\ln u)^{\theta}+(-\ln v)^{\theta}]^{1/\theta}\}$
[1]	$\theta \in [-1,1)$	$\ln\left[\frac{1-\theta(1-t)}{t}\right]$	$\frac{1-\theta}{e^s-\theta}$	$\frac{uv}{1-\theta(1-u)(1-v)}$
[20]	$\theta \in [1,\infty)$	$-\ln[1-(1-t)^{\theta}]$	$1-(1-e^{-s})^{1/\theta}$	$1-[(1-u)^{\theta}+(1-v)^{\theta}+(1-u)^{\theta}(1-v)^{\theta}]^{1/\theta}$

Table 1: Generators and their inverses for one-parameter Archimedean copulas.

variables X and Y. Lower tail-dependence λ_L and the upper tail-dependence λ_U can also be expressed in terms of bivariate copulas

$$\lambda_L = \lim_{u\to 0^+} \frac{C(u,u)}{u} \text{and} \lambda_U = \lim_{u\to 1^-} \frac{1-C(u,u)}{1-u}.$$

One of the most popular families of copulas, that were studied by [12], [16] and [32], is the Farlie-Gumbel-Morgenstern (FGM) family that is defined by

$$C(u,v) = uv(1+\theta(1-u)(1-v)), -1 \leq \theta \leq 1.$$

The FGM copula can be seen as a perturbation of the product copula which is obtained for $\theta = 0$. This copula is attractive because of its simplicity but is restrictive since is only useful when dependence between the two marginals is small. A maximum correlation of 33% is attained for the Spearman's coefficient while this correlation is limited to the interval $[-\frac{2}{9}, \frac{2}{9}]$ for Kendall's τ correlation.

To demonstrate the dependence properties of different copulas we simulate 500 pairs of exponential random variables (with rate 1) from the Clayton, Frank, Gumbel, AMH, Joe, and FGM copulas using the approaches outlined in next section. This is illustrated in Figure 2.

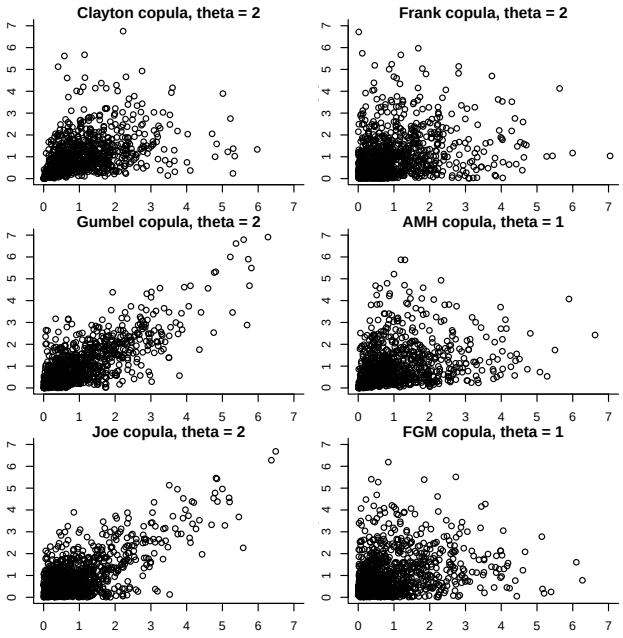

Fig. 2: Simulated samples from copulas (cut at a level of 7).

The pairs of exponential variables are plotted in order to illustrate dependence properties of the copulas. For four of the six copulas, the dependence parameter θ is set to 2. For the remaining copulas the dependence parameter was set to 1. Note that the dependence parameter in FGM, is set such that the dependence between the two variables are maximized (the FGM is unable to accommodate larger dependencies).

Family	Kendall's τ	$\tau \in \Omega$	Spearman's ρ
[7]	$\frac{\theta}{\theta+2}$	$[0,1)$	No simple form
[13]	$1-\frac{4}{\theta}\{D_1(-\theta)-1\}$	$[-1,1]\setminus\{0\}$	$1-\frac{12}{\theta}\{D_2(-\theta)-D_1(-\theta)\}$
[16]	$\frac{\theta-1}{\theta}$	$[0,1)$	No simple form
[1]	$1-\frac{2}{3\theta}-\frac{2}{3\theta^2}(\theta-1)^2 ln(1-\theta)$	$[-0.181726, \frac{1}{3}]$	a^*
[20]	$1+\frac{4}{\theta}E_J(\theta)$	$[0,1)$	No simple form
FGM	$\frac{2}{9}\theta$	$[-\frac{2}{9}, \frac{2}{9}]$	$\frac{\theta}{3}$

Table 2: Copulas and their measures of dependence. $D_k(x) = \frac{k}{x^k}\int_0^x \frac{t^k}{e^t-1}dt$ denotes the "Debye" function; $a^* = \frac{12(1+\theta)dilog(1-\theta)-24(1-\theta)\ln(1-\theta)}{\theta^2} - \frac{3(\theta+12)}{\theta}$; $dilog(x) = \int_1^x \frac{\ln t}{1-t}dt$; $E_J(\theta) = \int_0^1 \frac{(1-t^\theta)\ln(1-t^\theta)}{t^{\theta-1}}dt$.

III. Copulas and Random Number Generation

Simulations have an important role in statistical inference. They are particularly useful to investigate properties of estimators and to study the quality of a model. Moreover, they are also necessary to understand the underlying multivariate distribution. The copula construction allows us to simulate outcomes from many multivariate distributions easily.

The goal of this section is to present practical algorithms to simulate bivariate random variables for all copulas mentioned in the previous section. Assume that (X, Y) is a 2-dimensional random vector whose distribution is

$$H(x, y) = C(F(x), G(y))$$

where F denotes the marginal distribution of X, G the marginal distribution of Y and C is a copula.

A. Conditional distribution algorithm

One popular algorithm for simulating random variables is based on the conditional distribution approach. This approach separates the copula into several univariate components, each of which can be easily sampled. This method can be used in many copulas (Clayton, Frank, FGM, AMH). Assume that (X, Y) has a bivariate distribution function based on the two-dimensional Archimedean copula (Clayton, Frank, FGM or AMH). To generate data from a bivariate distribution function (X, Y) we first sample (u_1, u_2) from the copula-based distribution $C(u_1, u_2)$ with uniform margins and then we have to invert each u_i using the marginal distributions to obtain the data for the (X, Y). The procedure is to generate the observation of one margin, say U_1, and then to generate an observation for U_2 from its distribution given U_1. Consider two uniform random variables U_1 and U_2 with known copula C. Assuming sufficient regularity conditions, we obtain the conditional cumulative distribution function (cdf)

$$C_{2|1}(u_2 \mid u_1) = P(U_2 \leq u_2 \mid U_1 \leq u_1) = \frac{\partial C(u_1, u_2)}{\partial u_1}$$

Thus, the procedure to sample (u_1, u_2) from a copula-based distribution $C(u_1, u_2)$ is based on the algorithm 1 shown below.

Algorithm 1
(1) Simulate two independent uniform $U(0, 1)$ random variables, say (v_1, v_2).
(2) Set $u_1 = v_1$.
(3) Find the conditional distribution $C_{2|1}(v_2 \mid v_1)$ and its quasi-inverse $C^{-1}_{2|1}(v_2 \mid v_1)$. Set $u_2 = C^{-1}_{2|1}(v_2 \mid v_1)$.
Then, the pairs (u_1, u_2) are uniformly distributed variables drawn from the respective copula $C(u_1, u_2)$.
(4) The desired simulated values are $x = F^{-1}(u_1)$ and $y = G^{-1}(u_2)$.

Algorithm 1.1: Generating bivariate outcomes from Clayton copula
(1) Simulate two independent uniform $U(0, 1)$ random variables, say (v_1, v_2).
(2) Set $u_1 = v_1$.
(3) Set $u_2 = [v_1^{-\theta}(v_2^{-\theta/(1+\theta)} - 1) + 1]^{-1/\theta}$.
(4) The desired simulated values are $x = F^{-1}(u_1)$ and $y = G^{-1}(u_2)$.

Algorithm 1.2: Generating bivariate outcomes from Frank's copula
(1) Simulate two independent uniform $U(0, 1)$ random variables, say (v_1, v_2).
(2) Set $u_1 = v_1$.
(3) Set $u_2 = -\frac{1}{\theta} \ln \left(1 + \frac{v_2(1-e^{-\theta})}{v_2(e^{-\theta v_1} - 1) - e^{-\theta v_1}}\right)$.
(4) The desired simulated values are $x = F^{-1}(u_1)$ and $y = G^{-1}(u_2)$.

Algorithm 1.3: Generating bivariate outcomes from FGM copula
(1) Simulate two independent uniform $U(0, 1)$ random variables, say (v_1, v_2).
(2) Set $u_1 = v_1$.
(3) Set $a = 1 + \theta(1 - 2v_1)$; $b = \sqrt{a^2 - 4(a-1)v_2}$.
(4) Set $u_2 = 2v_2/(a + b)$.
(5) The desired simulated values are $x = F^{-1}(u_1)$ and $y = G^{-1}(u_2)$.

Algorithm 1.4: Generating bivariate outcomes from AMH copula
(1) Simulate two independent uniform $U(0, 1)$ random variables, say (v_1, v_2).
(2) Set $a = 1 - v_1$; $b = 1 - \theta(1 + 2av_2) + 2\theta^2 a^2 v_2$; $c = 1 + \theta(2 - 4a + 4av_2) + \theta^2(1 - 4av_2 + 4a^2 v_2)$.
(3) Set $u_2 = (2t(a\theta - 1)^2)/(b + \sqrt{c})$.
(4) The desired simulated values are $x = F^{-1}(u_1)$ and $y = G^{-1}(u_2)$.

The conditional distribution algorithm can be extended to the general case of p variables. In higher dimensions, the full distribution of $(X_1, ..., X_p)$ is simulated by recursively simulating the conditional distribution of X_k given $X_1, ..., X_{k-1}$ for $k = 2, ..., p$ ([5]).

B. Bivariate distribution algorithm

For some copulas the conditional distribution is not directly invertible and so different algorithms are necessary. This is the case of the Gumbel-Hougaard copula and the Joe copula. One alternative and popular algorithm that can be used to simulate random variables from an Archimedean copula is based on the following Theorem.

Theorem Let U_1 and U_2 be uniform $U(0, 1)$ random variables and let its bivariate distribution function be defined by the Archimedean copula generated by φ. Then,

the function $K_C(t) = t - \varphi(t)/(\varphi^{'}(t))$ is the distribution function of $C(U_1, U_2)$. Furthermore, the joint distribution of the random variables $X = \varphi(U_1)/[\varphi(U_1)+\varphi(U_2)]$ and $Y = C(U_1, U_2)$ is characterized by $H(x, y) = x \times K_C(y)$, for all $(x, y) \in I^2$ with X and Y independent, and X uniformly distributed on $(0, 1)$. Following, we present a proof in case of copula C to be absolutely continuous. A proof for the general case can be found in [15].

The joint density $h(x,y) = \dfrac{\partial^2}{\partial u_1 \partial u_2} C(u_1, u_2) \cdot \left| \dfrac{\partial(u_1, u_2)}{\partial(x, y)} \right|$ in terms of x and y, where $\partial^2 C(u_1, u_2)$ is given as follow and $\partial(u_1, u_2)/\partial(x, y)$ correspond to the Jacobian of the transformation $\varphi(u_1) = x\varphi(y)$, $\varphi(u_2) = (1 - x)\varphi(y)$. Since

$$\frac{\partial(u_1, u_2)}{\partial(x, y)} = \frac{\varphi(y)\varphi'(y)}{\varphi'(u_1)\varphi'(u_2)}$$

and consequently

$$h(x,y) = \left(- \frac{\varphi''(y)\varphi'(u_1)\varphi'(u_2)}{\left[\varphi'(y)\right]^3} \right) \cdot \left(- \frac{\varphi(y)\varphi'(y)}{\varphi'(u_1)\varphi'(u_2)} \right) = \frac{\varphi''(y)\varphi'(y)}{\left[\varphi'(y)\right]^2}$$

Thus

$$H(x,y) = \int_0^x \int_0^y \frac{\varphi''(z)\varphi(z)}{\left[\varphi'(z)\right]^2} dzdw = x \cdot \left[z - \frac{\varphi(z)}{\varphi'(z)} \right]_0^y = x \cdot K_c(y)$$

and the conclusion follows.

The resulting simulation procedure follows algorithm 2.

Algorithm 2
(1) Simulate two independent uniform $U(0, 1)$ random variables, say (v_1, v_2).
(2) Set $t = K_C^{-1}(v_2)$ where $K_C(t) = t - \varphi(t)/(\varphi^{'}(t))$
(3) Set $u_1 = \varphi^{-1}(v_1\varphi(t))$ and $u_2 = \varphi^{-1}((1 - v_1)\varphi(t))$
Then, the pairs (u_1, u_2) are uniformly distributed variables drawn from the respective copula C.
(4) The desired simulated values are $x = F^{-1}(u_1)$ and $y = G^{-1}(u_2)$.

Extensions of the results shown in the above theorem can be used to provide the corresponding simulation algorithm to the multi-dimensional case ([39]). The main challenge for the practical implementation of this algorithm is to find the inverse function of K_C.

Algorithm 2.1: Generating bivariate outcomes from the Gumbel-Hougaard copula
(1) Simulate two independent uniform $U(0, 1)$ random variables, say (v_1, v_2).
(2) Set $K_C(t) = t \times (1 - \ln(t)/\theta) = v_2$, and solve numerically for $0 < t < 1$.
(3) Set $u_1 = \exp[v_1^{1/\theta} \ln(t)]$ and $u_2 = \exp[(1 - v_1)^{1/\theta} \ln(t)]$.
Then, the pairs (u_1, u_2) are uniformly distributed variables drawn from the respective copula C.
(4) The desired simulated values are $x = F^{-1}(u_1)$ and $y = G^{-1}(u_2)$.

Algorithm 2.2: Generating bivariate outcomes from Joe copula
(1) Simulate two independent uniform $U(0, 1)$ random variables, say (v_1, v_2).
(2) Set $K_C(t) = t - \frac{\left[\ln(1-(1-t)^\theta)\right]\left[1-(1-t)^\theta\right]}{\left[\theta(1-t)^{\theta-1}\right]} = v_2$, and solve numerically for $0 < t < 1$.
(3) Set $u_1 = 1 - \{1 - [1 - (1 - t)^\theta]^{v_1}\}^{1/\theta}$ and $u_2 = 1 - \{1 - [1 - (1 - t)^\theta]^{1-v_1}\}^{1/\theta}$.
Then, the pairs (u_1, u_2) are uniformly distributed variables drawn from the respective copula C.
(4) The desired simulated values are $x = F^{-1}(u_1)$ and $y = G^{-1}(u_2)$.

C. Laplace Transform algorithm

Clayton, Frank, Joe and Gumbel-Hougaard copulas fall into the class of the so-called Laplace (Stieltjes) Transform Archimedean copulas (LT-Archimedean copulas). This LT representation leads to a useful way of simulating such copulas ([25]; [20]; [18]). For such copulas, the inverse of the generator function φ has a nice representation on a Laplace Transform of some function G. The algorithm 3, based on the LT representation is given below:

Algorithm 3
(1) Generate a variable V with distribution function G with $\psi(t) = \int_0^{+\infty} e^{tx} dG(x)$, $t \geq 0$, the Laplace-Stieltjes transform of G.
(2) Generate independent standard uniform random variables v_1, v_2.
(3) Set $u_i = \psi(-ln(v_i)/V)$.

Then, the vector (u_1, u_2) has the desired Archimedean copula dependence structure with generator $\varphi = \psi^{-1}$.

- For a Clayton copula, V is gamma distributed $Ga(1/\theta, 1)$ and $\psi(t) = (1 + t)^{-1/\theta}$.
- For a Gumbel-Hougaard copula, V is stable distributed $St(1/\theta, 1, (cos(\Pi/(2\theta)))^\theta, 0; 1)$ (see [33]) and $\psi(t) = exp(-t^{1/\theta})$.
- For a Frank copula, V is discrete with $P(V = k) = (1 - e^{-\theta})^k/(k\theta)$ and $\psi(t) = -\frac{1}{\theta} ln[1 + e^{-t}(e^{-\theta} - 1)]$, $k \in \mathbb{N}$.
- For the AMH copula, V is discrete with $P(V = k) = (1 - \theta)\theta^{k-1}$ and $\psi(t) = \frac{1-\theta}{e^t-\theta}$, $k \in \mathbb{N}$.
- For Joe copula, V is discrete with $P(V = k) = (-1)^{k+1}\binom{1/\theta}{k}$ and $\psi(t) = 1 - (1 - e^{-t})^{1/\theta}$, $k \in \mathbb{N}$.

Unfortunately, it is not known how to find G explicitly. If we know how to sample G, this algorithm provides a powerful tool for sampling these copulas with large dimensions.

D. Survival data and Random Number Generation

Copulas have been widely studied in the last decades. Their first applications were mainly in actuarial sciences and finances but their use has spread to other areas such as survival analysis. The copula construction allows the selection of different marginal distributions for each outcome while accounting for the dependence between the random variables. They can be used to model and understand explanatory variables in survival analysis. The copula structure can also be used to study different survival models such as the bivariate survival. For example, suppose we are considering to examine the survival of twins. There is strong empirical evidence that supports the dependence of their lifetimes. Another problem that often appear in survival analysis and that can be modeled with copulas is the issue of competing risks. Though in many cases the outcomes (competing risks; see Figure 3) are assumed to be statistically independent there is strong evidence that this assumption is not realistic. To account for this dependence, one general approach is to apply copulas ([11]; [21]).

In many longitudinal studies subjects can experience several events across a follow-up period. The events of concern may be of the same nature (e.g., cancer patients may experience recurrent disease episodes) or represent different states in the disease process (e.g., alive and disease-free, alive with recurrence and dead). If the events are of the same nature these are usually referred as recurrent event, whereas if they represent different states (i.e. multi-state models) they are usually modeled through their intensity functions ([37]; [17]; [23]). The dependence between the different outcomes can also be modeled using copulas ([8]; [19]; [24]; [35]).

The algorithms shown above can be used to generate survival data that can be used in many of these situations. One can use them to generate survival data observed subject to random right-censoring ([22]; [19]), arising from censored gap times ([10]; [30]), competing-risks ([34]) and multi-state models ([3]; [26]; [28]; [29]). Below we present the algorithms to generate data for the four models.

Time-to-event data

Standard survival data measure the time from some particular time origin until the occurrence of one type of event. The main feature of survival data is censoring. Right-censoring is the most common type of censoring and can occur because of insufficient follow-up, loss to follow-up or failure unrelated to the study. We denote the random variable survival time by Y. Next, we denote the random censoring variable by Z, which we assume to be independent of Y; and $\Delta = I(Y \leq Z)$ the indicator status indicating either a failure (i.e., $\Delta = 1$) or censorship occurred. Because of censoring rather than Y we ob-

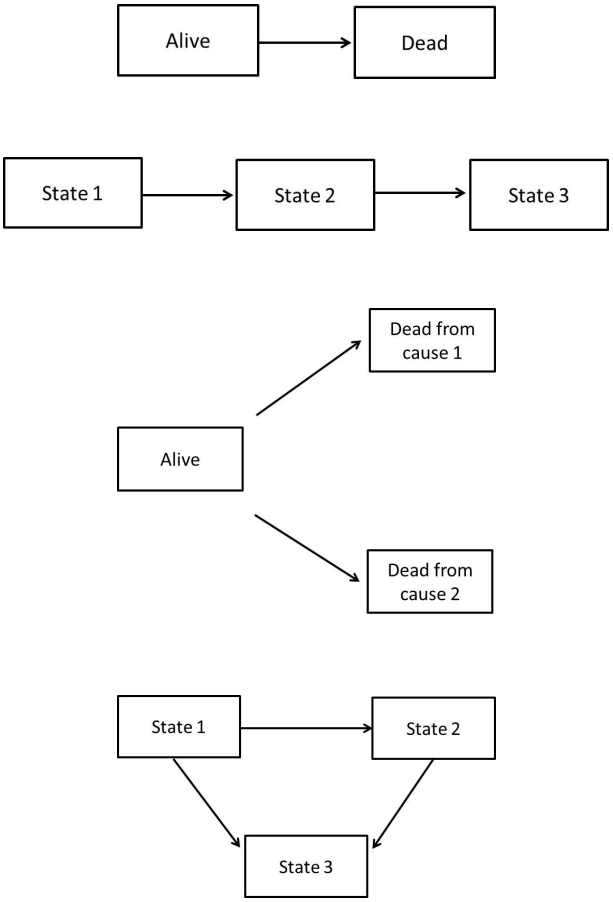

Fig. 3: Schematic representation of some common multi-state models. Mortality model for survival analysis (top); recurrent events model (second row); competing risks model (third row) and progressive illness-death model (bottom).

serve (T, Δ) where $T = \min(Y, Z)$ is the observed time. If covariables, X, are present, the observed data consists of the triplets (T_i, Δ_i, X_i) $(i = 1, \ldots, n)$ of independent and identically distributed replicates of (T, Δ, X).

The procedure to generate such data is as follows:
(1) Generate (X, Y) from a bivariate distribution function based on some known two-dimensional copula.
(2) An independent censoring time Z is generated, according to some particular model (e.g., Uniform or Exponential).
(3) Set $T = \min(Y, Z)$ and $\Delta = I(Y \leq Z)$.

Recurrent events data
Recurrent events involve repeat occurrences of the same type of event over time ([8]). Recurrent events in longitudinal studies include recurrent leukaemia episodes, tumor recurrences in cancer patients (e.g. bladder cancer) or heart failure hospitalizations. Let (X, Y) be gap times corresponding to two consecutive events, which are observed subject to random right-censoring. The fact that the variables X and Y are recorded successively, rather than simultaneously, is important when the variables are subject to censoring. Again, we consider here random right censoring (denoted by Z). In the present context of successive events, we only observe the second gap time if the first failure time is uncensored. More precisely, the observable variables are given by $(T_1, T_2, \Delta_1, \Delta_2)$ where $T_1 = \min(X, Z)$, $\Delta_1 = I(X \leq Z)$, $T_2 = \min(Y, Z_2)$ and $\Delta_2 = I(Y \leq Z_2)$, where $Z_2 = (Z - X)I(X \leq Z)$ is the censoring variable for the second gap time.

The procedure to generate such data is as follows:
(1) Generate (X, Y) from a bivariate distribution function based on some known two-dimensional copula.
(2) An independent censoring time Z is generated, according to some particular model (e.g., Uniform or Exponential).
(3) Set $T_1 = \min(X, Z)$; $\Delta_1 = I(X \leq Z)$; $T_2 = \min(Y, Z - X) \times I(X \leq Z)$; $\Delta_2 = I(X + Y \leq Z)$.

Competing risks data
Competing risks data (Figure 3, third row) are encountered in many medical studies where the subjects under study are at risk for more than one mutually exclusive event. The observable data in these models is represented by the failure time T and the indicator status variable Δ, which in this case will take the value 0 if the competing risk process does not move from the initial state at the survival time T, or the value 1 and 2 for the possible causes of death 1 and 2. The observable data may also include a possibly covariable vector, which we shall ignore for the moment. The survival time and cause of death may be modeled as arising from the minimum of latent failure times corresponding to the different causes. The procedure to generate such data is as follows:
(1) Generate (X, Y) from a bivariate distribution function based on some known two-dimensional copula.
(2) An independent censoring time Z is generated, according to some particular model (e.g., Uniform or Exponential).
(3) If $X \leq Y$ then $D = 1$; otherwise $D = 2$.
(4) Set $T = \min(X, Y, Z)$; $\Delta = I(\min(X, Y) \leq Z) \times D$.

Alternative simulation designs for competing risks data are given by [4].

Progressive illness-death multi-state model
In some cases the events of concern may not be of the same nature, representing different stages in the disease process. In biomedical applications these stages or states may represent health conditions (e.g., healthy, diseased, dead), disease stages (e.g., stages of cancer or HIV infection) or a nonfatal complication in the course of the illness (e.g., cancer recurrence, transplantation, etc.)[38]. These are known as multi-state models and are usually modeled through their intensity functions ([3]; [26]; [29]; [14]). Consider for example a cancer study, where X represents the time between tumor resection and recurrence (local or distant), and Y represents the time between development of a recurrence and death of the patient. Some individuals may die without observing a recurrence. The progressive illness-death model, also known as disability model, is probably the most popular one in the medical literature. The irreversible version of this model (Figure 3, bottom), describes the pathway from an initial state to an absorbing state either directly or through an intermediate state. Many event-history data sets from biomedical studies with multiple endpoints can be reduced to this generic structure.

To simulate the data in the progressive illness-death model, we separately consider the subjects passing through State 2 at some time, and those who directly go to the absorbing State 3. For the first subgroup of individuals, the successive gap times can be simulated using a two-dimensional copula, whereas those in the second group can be simulated from any continuous distribution.

The procedure to generate such data is as follows:
(1) Draw $\rho \sim Ber(p)$ where p is the proportion of subjects passing through State 2.
(2) If $\rho = 1$ then generate (X, Y) from a bivariate distribution function based on some known two-dimensional copula.
(3) If $\rho = 0$, one particular model (e.g., Uniform, Exponential or Weibull) is used to generate the transition time, W, from State 1 to State 3.
(4) An independent censoring time Z is generated, according to some particular model (e.g., Uniform or Exponential).
(5) If $\rho = 1$ then set $T_1 = \min(X, Z)$ and $\Delta_1 = I(X \leq Z)$. Set also $T = \min(X + Y, Z)$ and $\Delta = I(X + Y \leq Z)$.
(6) If $\rho = 0$ then set $T_1 = \min(W, Z)$ and $\Delta_1 = I(W \leq Z)$. Set also $T = T_1$ and $\Delta = \Delta_1$.

The stochastic behavior of the process in this model is characterized by the vector of random variables

$(T_1, T, \Delta_1, \Delta)$, where T_1 is the sojourn in State 1, T the total time and Δ_1 and Δ the corresponding indicator statuses.

The general (and usual) censoring distributions assumed to model censoring are uniform and exponential. The parameters in these distributions can be determined by iterative algorithms to control the censoring percentage one wishes to obtain.

IV. An R package to generate complex survival data

In R, several packages provide functions for simulating survival data. A comprehensive list of these packages can be seen in the CRAN task view 'Survival Analysis' ([2]). Some of them can be used to simulate data from complex processes, such as the genSurv ([27]) package that permits to generate data with one binary time-dependent covariable and data stemming from a progressive illness-death model. Univariate and semi-competing risks data can be generated using the `SimSCRPiecewise` package. The `survsim` package ([9]; [31]) can also be used to simulate simple and complex survival data such as recurrent event data and competing risks data. Complex multi-state models data with possibly nonlinear baseline hazards and nonlinear covariable effects can be simulated using functions available as part of the `simMSM` package.

To provide researchers with an easy-to-use tool for simulating complex survival data we develop an R package called `survCopula`. This package is composed by a set of functions which allow the user to simulate a cohort with the objective of studying its behavior in a variety of scenarios including survival, competing risks, recurrent events and some multi-state models. The main feature of the package is its ability for using different copulas for simulating correlated multivariate survival data in a variety of scenarios as discussed in Section 3. Copulas are a useful tool to model multivariate distributions. They allow us to control the dependence between time variables with knowledge of the marginal distributions. This software and source code are all available at the GitHub repository at `https://github.com/gsoutinho/survCopula`. Details on the usage of its functions can be obtained with the corresponding help pages after the package is installed.

For illustration purposes, suppose we are interested to simulate survival data for the mortality model. One possibility would be using a bivariate copula with marginal functions uniformly distributed on (0; 5), where the survival (denoted by T) could be for instances the survival time (in years) of lung cancer since diagnosis, and tumor size (in cm) is a covariate value measured for each individual. Individuals alive at the end of the follow-up have right censored observations (i.e., Delta = 0). Such data can be obtained using the **dgCopula** function of the `survCopula` package through the following input commands:

```
> library(survCopula)
> setseed(2345)
> sim.data<-dgCopula(typeCopula ='clayton',
  theta=1, typeX='Unif', num1_X=0, num2_X=5,
  typeY='Unif',  num1_Y=0, num2_Y=5,
  typeCens='Unif', num1_Cens=0, num2_Cens=7,
  nsim=250,typeSurvData='time-to-event')

> head(sim.data)

        T         Z       Delta
  1 3.3786641 5.3939928     1
  2 4.5925602 6.3436964     1
  3 1.9646380 1.9646380     0
  4 0.5421364 5.1900408     1
  5 0.4418575 0.5881083     1
  6 2.1502214 2.1502214     0
```

As arguments, `typeCopula` correspond to the bivariate copula used to generate data and `theta` to a numeric value for the space parameter. In the arguments `typeX`, `typeY` and `typeCens` are indicated the marginal distributions for the bivariate copula and the censuring time distribution, respectively. The parameters of the distributions are given by the arguments *num1* and *num2*.

V. Discussion and Future research

Copulas have become a popular tool to create distributions that model correlated multivariate data. In this paper a review of the most common copulas is presented with the goal to introduce the generators functions of some important families of Archimedean copulas as well as their dependence that can be measured by Kendall's tau τ or Spearman's ρ. Due to the important role of the simulation studies in statistical inference this article also describes several algorithms to generate bivariate data from several copulas, and explored the use of these correlated data for generating multivariate survival data in a variety of scenarios. In fact, the use of copulas is suitable for this purpose since they can be used to introduce dependence between time and covariates, or between times of different transitions in more complex survival systems. In case of the Conditional distribution algorithm this can be applied in many copulas such as Clayton, Frank, FGM or AMH. Since some copulas are not directly invertible for the Gumbel-Hougaard and the Joe copulas was also discussed an alternative algorithm making use of the function the function K_C. A Laplace Transform algorithm is also described for some copulas and finally, four types of survival data and random number generation are presented covering different situations, including recurrent events, competing risks and models with multiple events of different types. In this paper we also demonstrated the application of these methods of copulas to the biomedical statistics namely in simulation studies involving different models in survival analysis or multistate models who have the advantage to take in consideration the dependence of variables. In order to be used on biomedical practices a user-friendly

software in the form of an R package is provided too. The package provides several functions that can be used to generate survival data in a variety of scenarios including competing risks, recurrent event and multi-state models. Users can choose the marginal distributions as well as the dependence between the correlated data which is induced in the joint distribution by means of copulas. As a future field of research we are interested to use copulas to simulate longitudinal and survival data. This type of data is particularly relevant in cancer studies in which longitudinal biomarkers may be associated to the survival time.

Acknowledgment

This research was financed by Portuguese Funds through FCT - "Fundação para a Ciência e a Tecnologia", within the research grant PD/BD/142887/2018. Luís Meira-Machado acknowledges financial support from the Spanish Ministry of Economy and Competitiveness MINECO through project MTM2017-82379-R funded by (AEI/FEDER, UE) and acronym "AFTERAM".

References

[1] Ali, M.M., Mikhail, N.N. and Haq, M.S. (1978). A Class of Bivariate Distributions Including the Bivariate Logistic. *Journal of Mulvariate Analysis*, **8**, 405–412.

[2] Allignol, A. and Latouche, A. (2019). CRAN Task View: Survival Analysis. Version 2019-09-01, URL http://CRAN.Rproject.org/view=Survival

[3] Andersen, P.K., Borgan, O., Gill, R.D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

[4] Beyersmann, J., Latouche, A., Buchholz, A. and Schumacher, M. (2009). Simulating competing risks data in survival analysis. *Statistics in Medicine*, **28**, 956–971.

[5] Bouyè, E., Durrleman, V., Bikeghbali, A., Riboulet, G. and Roncall, T. (2000). Copulas for finance - A reading guide and some applications. *Working paper, Group de Rechercher Opérationnelle, Crédit Lyonnais*.

[6] Burton, A., Altman, D.G., Royston, P. and Holder, R.L (2006). The design of simulation studies in medical statistics. *Statistics in Medicine*, **25**, 4279–4292.

[7] 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**, 141–152.

[8] Cook, R.J. and Lawless, J. (2007). *The Statistical Analysis of Recurrent Events*, Springer-Verlag, New York.

[9] Crowther, M.J. and Lambert, P.C. (2013). Simulating biologically plausible complex survival data. *Statistics in Medicine*, **32(23)**, 4118–4134.

[10] de Uña-Álvarez, J. and Meira-Machado, L.F. (2008). A simple estimator of the bivariate distribution function for censored gap times. *Statistics & Probability Letters*, **78**, 2440–2445.

[11] Escarela, G. and Carrière, J.F. (2003). Fitting Competing Risks with an Assumed Copula. *Statistical Methods in Medical Research*, **12(4)**, 333–349.

[12] Farlie, D.G.J. (1960). The performance of some correlation coefficients for a general Bivariate distribution *Biometrika* **47**, 307-323.

[13] Frank, M.J. (1979). On the simultaneous Associativity of F(x,y) and x+y-F(x,y) *Aequationes Mathematicae*, **19**, 124–226.

[14] Loïc Ferrer, Virginie Rondeau, James J. Dignam, Tom Pickles, Hélène Jacqmin-Gadda, Cécile Proust-Lima (2016). Joint modelling of longitudinal and multi-state processes: application to clinical progressions in prostate cancer. *Statistics in Medicine*, **35(22)**: 3933–3948.

[15] Genest C, Rivest L-P (1993). Statistical inference procedures for bivariate Archimedean copulas. *J Amer Statist Assoc*, **36(88)**, 1034–1043

[16] Gumbel, E.J. (1960). Bivariate Exponential Distributions. *Journal of the American Statistical Association*, **55**, 698–707.

[17] Harden, J. J. and Kropko, J. (2018). Simulating Duration Data for the Cox Model. *Political Science Research and Methods*, **7(4)**, 921–928.

[18] Hofert, M. (2008). Sampling Archimedean copulas. *Computational Statistics & Data Analysis*, **52(12)**, 5163–5174.

[19] Hougaard, P. (2000). *Analysis of Multivariate Survival Data*, Springer-Verlag, New York.

[20] Joe, H. (1997). *Multivariate Models and Dependence Concepts*. Chapman & Hall Ltd.

[21] Kaishev, V.K., Dimitrova, D.S., and Haberman, S. (2007). Modelling the joint distribution of competing risks survival times using copula functions. *Insurance: Mathematics and Economics*, **41(3)**, 339–361.

[22] Kalbfleisch, J. D. and Prentice R. L. (1980). *The statistical analysis of failure time data*. John Wiley & Sons, New York.

[23] Kropko, J. and Harden, J. (2018). Beyond the Hazard Ratio: Generating Expected Durations from the Cox Proportional Hazards Model. *British Journal of Political Science*, **50(1)**, 303–320.

[24] Malehi, A.S., Hajizadeh, E., Ahmadi, K.A., and Mansouri, P. (2015). Joint modelling of longitudinal biomarker and gap time between recurrent events: copula-based dependence. *Journal of Applied Statistics*, **42(9)**, 1931–1945.

[25] Marshall, A.W. and Olkin, I. (1988). Families of multivariate distributions. *Journal of the American Statistical Association*, **83**, 834–841.

[26] Meira-Machado, L., de Uña-Álvarez, J., Cadarso-Suárez, C. and Andersen, P.K. (2009). Multi-state models for the analysis of time to event data. *Statistical Methods in Medical Research*, **18**, 195–222.

[27] Meira-Machado, L. and Faria, S. (2014). A simu-

lation study comparing modeling approaches in an illness-death model. *Communications in Statistics - Simulation and Computation*, **43(5)**, 929–946.

[28] Meira-Machado, L. and Sestelo, M. (2016). condSURV: An R Package for the Estimation of the Conditional Survival Function for Ordered Multivariate Failure Time Data. *The R Journal*, **8(2)**, 460–473.

[29] Meira-Machado, L. and Sestelo M. (2019). Estimation in the progressive illness-death model: a nonexhaustive review. *Biometrical Journal*, **61(2)**, 245–263.

[30] Moreira, A., Meira Machado, L. (2012). survivalBIV: Estimation of the Bivariate Distribution Function for Sequentially Ordered Events Under Univariate Censoring. *Journal of Statistical Software*, **46(13)**, 1–16.

[31] Morina, D. and Navarro, A. (2014). The R Package survsim for the Simulation of Simple and Complex Survival Data. *Journal of statistical software*, **59(2)**, 1–20.

[32] Morgenstern, D. (1956). Einfache beispiele zweidimensionaler verteilungen. *Mitteilungsblatt fürMathematische Statistik*, **8**, 234–235.

[33] Nolan, J.P. (2007). *Stable distributions - models for heavy tailed data*. Birkhauser, Boston, MA.

[34] Putter, H., Fiocco, M. and Geskus R.B. (2007). Tutorial in biostatatistics: Competing risks and multistate models. *Statistics in Medicine*, **26**, 2389–2430.

[35] Rotolo, F., Legrand, C. and Van Keilegom, I. (2013). A Simulation Procedure Based on Copulas to Generate Clustered Multi-State Survival Data. *Computer Methods and Programs in Biomedicine*, **109(3)**, 305–312.

[36] Sklar, A.W. (1959). Fonctions de répartition à n dimension et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, **8**, 229–231.

[37] Soutinho, G., Meira-Machado, L. and Oliveira, P. (2020). A comparison of presmoothing methods in the estimation of transition probabilities. *Communications in Statistics - Simulation and Computation*. DOI: 10.1080/03610918.2020.1762895

[38] Soutinho, G. and Meira-Machado, L. (2020). Estimation of the Transition Probabilities in Multi-state Survival Data: New Developments and Practical Recommendations. *WSEAS Transactions on Mathematics*, **19**, 353–366.

[39] Wu, F., Valdez, E. A. and Sherris, M. (2006). Simulating Exchangeable Multivariate Archimedean Copulas and its Applications. *Communications in Statistics - Simulation and Computation*, **36(5)**, 1019–1034.