UCLouvain

FACULTE DES SCIENCES

Faculté des sciences

Estimation of survival function in presence of dependent censoring

Auteures : Laura Vermeren
Promotrices : Ingrid Van Keilegom
Lectrices : Catherine Legrand
Année académique 2023-2024

Declaration

Declaration

During the preparation of this dissertation, I made use of artificial intelligence such as ChatGPT and DeepL.

These tools were used to correct spelling and grammar, and sometimes to reformulate sentences. These tools were not used to generate information.

By signing this declaration, I affirm that the content of this dissertation reflects my original work, augmented by the responsible use of IA.

[Vermeren Laura], [June 3, 2024]

Abstract

This master thesis explores the estimation of the survival function under dependent censoring, a situation frequently encountered in survival analysis. Building on the methodologies proposed in Czado and Van Keilegom (2023) and Rivest and Wells (2001), we employ copulas to model the dependence without needing to specify the copula's association parameter explicitly. The survival time distribution is estimated non-parametrically using the copula graphic estimator from Rivest and Wells (2001), while the censoring time distribution is approached parametrically. Our results demonstrate that our estimator generally outperforms the Kaplan-Meier estimator. Specifically, for minor dependencies (Kendall's tau of 0.2), a larger sample size is required for our estimator to achieve lower mean absolute deviation. However, for more substantial dependencies (Kendall's tau of 0.5 and 0.7), our estimator performs better than Kaplan-Meier across most small sample sizes, albeit with some exceptions. We also find that Kendall's tau estimates can be numerically unstable, as evidenced in two out of five simulations. Additionally, accurate specification of margins proves critical for robust estimation.

Acknowledgments

I would like to thank everyone who helped during the writing process of this master thesis. I am especially grateful to my promotor, Prof. Dr. Ingrid Van Keilegom, for her guidance and insightful recommendations. She was always available to answer any questions and provided substantial help, from coding to writing. I would also like to thank Morine Delhelle, who took the time to answer my questions and assist me with the coding.

Contents

1 Introduction **1**

2 Theoretical Framework **3**

2.1 Survival analysis 3

2.1.1 Mathematical notation 4

2.1.2 Hypothesis of the Independence between T and C 7

2.2 Copulas 10

2.2.1 Properties of copulas 11

2.2.2 Some classical copulas 12

3 Methodology **16**

3.1 Resume of Czado and Van Keilegom (2023) 19

3.2 Resume of Rivest and Wells (2001) 21

3.3 Development of our dependent censoring Model 23

4 Simulation **28**

4.1 Main scenario 31

4.2 Secondary scenario 39

5 Data illustration **43**

6 Conclusion **46**

6.1 Summary of the results 46

6.2 Limits and future research 47

Appendices **51**

A Derivation of the Kaplan-Meier formula **51**

B Simulation tables of estimated parameters **53**

B.1 Scenario 1 53

B.2 Scenario 2 55

B.3 Scenario 3 57

B.4 Scenario 4 58

C Code **59**

1 Introduction

Hearing the name 'survival analysis' for the first time can be a little daunting. You expect to study only probabilities of dying. But as you learn a little more about this branch of statistics, you'll notice that it can have many diverse applications. Indeed, survival analysis takes its name from the fact that this branch of statistics was originally created to study patient survival times in the medical field and clinical trials. The development of life analysis dates back to the 17th century, with the first life table produced by John Graunt in 1662. (Camilleri, 2019).

Survival analysis refers to the study of the time until the occurrence of a specific event (Kleinbaum and Klein, 2012). In the medical field, it is common to analyze the time to death caused by a specific disease - the event, in this case, would be "dying of disease X". However, other applications in this field not directly related to patient survival can also be studied, such as the duration of optimal functioning of a hip prothesis or the period of efficacy of vaccines.

Survival analysis also has applications in several fields far removed from medical patient survival. In economics, for example, it can be used to study how long it takes for a person to find a new job; in psychology, it can be used to examine how long it takes for a person to come out of rehab; and in technology, it is used to predict how long a machine will operate before it fails.

Now that the idea that survival analysis is limited to the study of time to death has been dispelled, we can turn our attention to the key concepts of this discipline. It's not possible to carry out a study of infinite duration, nor to observe every possible event. Fortunately, survival analysis is able to deal with situations where information on the event of interest is incomplete, known as censoring. For example, in a three-year study of pancreatic cancer patients in a hospital, it is likely that some will move, die of other causes, or survive their cancer. These patients will be considered censored.

Survival analysis is often studied via a survival function, which represents the probability of an individual surviving longer than a certain time. Its best-known estimator is the Kaplan and Meier (1958) estimator. Although powerful, it assumes independence between survival time and censoring time, an assumption often difficult to verify in practice. This master thesis will therefore explore statistical tools that allow us to take into account cases where there is a dependency between survival time and censoring time.

The aim of this master's thesis is to develop a new model that integrates the advantages of two existing models developed in the paper of Rivest and Wells (2001) and Czado and Van Keilegom (2023). These models analyze the dependence between survival time and censoring time using copulas. The first model provides an estimator for the survival function; however, it requires a known parameter that specifies the degree of dependence between the survival and the censoring time, which is not always practical. The second model does not require the copula parameter to be known but assumes that the distribution functions of the survival and censoring time are parametric. This master thesis will combine the strengths of these approaches: it will adopt the flexibility of not needing to know the copula parameter from Czado and Van Keilegom (2023), and will apply the method from the first model to estimate the survival time distribution non-parametrically.

In the first section of this thesis, we will review the basics of survival analysis and copulas. After a brief literature review, we will explain in more detail the two articles we are using to build our model. We'll then explain how our estimator is constructed, and continue with some simulations to check the quality of our model. We will then conclude with an analysis of real data.

2 Theoretical Framework

In this section, we will explain in more detail the two key concepts of this thesis: survival analysis and copulas. In the first section on survival analysis, we explain in more detail the concept of censoring, introduce mathematical notations and concepts, and return to the hypothesis of independence between survival time and censoring time. In the second section, we introduce the mathematical concept of copulas and define some explicit copulas.

2.1 Survival analysis

As mentioned in the introduction, survival analysis is a statistical branch that focuses on a variable representing the time until an event of interest occurs and finds application in numerous fields (Kleinbaum and Klein, 2012). In practice, there are instances where the event of interest cannot be observed for various reasons. This may be due to the event of interest not being observed by the end of the study, another event taking place before the event of interest, or a person being lost to follow-up during the study. (Kleinbaum and Klein, 2012). This phenomenon is called censoring. Let's take a simple example from the field of engineering to illustrate this concept: a manufacturing company wishes to estimate the lifespan of its new machines to better plan their maintenance. The company decides to carry out a survival study over two years, to observe how long the machines operate before experiencing their first breakdown. If a machine has not had a breakdown by the end of the two-year observation period, it is considered censored. This means that the event of interest (the breakdown) has not been observed during the study period. There exist three major types of censoring in survival analysis: right censoring, left censoring, and interval censoring. Right, censoring occurs when the true survival time is equal to or greater than the observed survival time, while left censoring occurs when the true survival time is less than or equal to the observed one. In the interval-censored case, the true survival time is within a known time interval (Kleinbaum and Klein, 2012). For a clearer comprehension, let's illustrate these three types of censoring. Consider a scenario where a dentist investigates the duration until a child develops their first cavity. An instance of right censorship would occur if a child changes dentists before experiencing their first cavity. Consequently, we would lack precise

information regarding the timing of this initial cavity. An illustration of left censoring would involve a child attending their first dental appointment having already a cavity; we would only have information from the moment we started observing the child. The final form of censorship would manifest if we identify a cavity in a child between two distant appointments. In such a scenario, it becomes challenging to determine precisely when the decay began to develop during the interval between appointments. Censoring time is a random variable and is, generally, supposed to be independent of the survival time. We will come back to this hypothesis in more detail below. In the following, we will focus on right censoring.

2.1.1 Mathematical notation

We can now introduce the mathematical notations for survival analysis. This section is inspired by the book of Kleinbaum and Klein (Kleinbaum and Klein, 2012) and Professor Van Keilegom's course (Van Keilegom, 2023).

We denote the survival time, which is the time until the event of interest, as T. It is a non negative continuous random variable. We denote by a small t any specific value of the random variable T. The distribution function of T is denoted by $F_T(t)$, representing the probability $P(T \leq t)$, and its density function is denoted as $f_T(t)$.

In survival analysis, we are interested in the survival function, denoted as $S_T(t)$, and the hazard function, denoted by $h_T(t)$. The survival function represents the probability that a randomly selected individual will survive beyond a specific time, t. In mathematical terms, it can be expressed as,

$$S_T(t) = P(T > t) = 1 - F_T(t).$$

It is a decreasing function, taking values between zero and one. When time is zero, $t = 0$, the survival function equals one, $S_T(0) = 1$. This signifies that at the beginning of the study, no one has experienced the event, so the probability of surviving past time 0 is one. As time extends to infinity, $t = \infty$, the survival function approaches zero, $S_T(\infty) = 0$. This represents the theoretical scenario where, with an infinite study duration, no one would ultimately survive indefinitely.

Theoretically, the survival function looks like in Figure 1, but in practice, it is not a smooth function and we never have infinity study time. The survival function looks in practice like Figure 2.

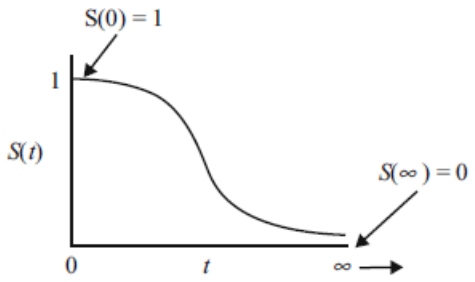


Figure 1: Theoretical survival function, (Kleinbaum and Klein, 2012)

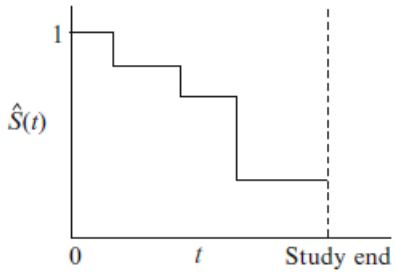


Figure 2: Survival function in practice, (Kleinbaum and Klein, 2012)

The survival function is central in survival analysis. Obtaining survival probabilities for different values of t provides crucial information and it gives us access to a second important concept: the hazard function. By taking the negative logarithm of the survival function, we derive the cumulative hazard function, denoted as $H_T(t)$. This is done using the formula

$$H_T(t) = -\log S_T(t).$$

With the cumulative hazard function $H_T(t)$ at hand, we can calculate the hazard function by taking its derivative, expressed as

$$h_T(t) = \frac{d}{dt} H_T(t). \tag{2.1}$$

In concrete terms, the hazard function quantifies the instantaneous risk of experiencing the event (e.g., death) immediately after time t, given that an individual is alive at time t. It

provides valuable insights into the time-varying risk associated with the event of interest. To delve deeper into the relation 2.1, consider the hazard function as derived by the limit definition of the derivative,

$$\begin{aligned} h_T(t) &= \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t | T > t)}{\Delta t} \\ &= \frac{1}{P(T \geq t)} \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta T} \\ &= \frac{f_T(t)}{S_T(t)} \\ &= \frac{-d}{dt} \log S_T(t) \\ &= \frac{d}{dt} H_T(t) \end{aligned}$$

This shows the relationship between the cumulative hazard function and the hazard function. In contrast to the survival function, the hazard curve can take different shapes as we can see in Figure 3. It does not have to start at one and go down to zero.

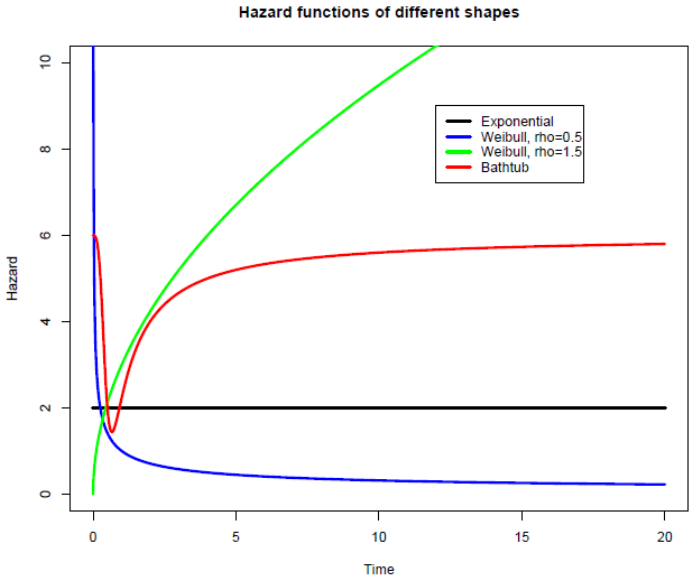


Figure 3: Hazard functions of different shapes (Van Keilegom, 2023)

We can now introduce the concept of right censoring mathematically. In cases of right-censored data, we do not always observe the true survival time (T), because some individuals

may, for instance, drop out of the study or experience other unrelated events. In this case, what we observe is the censoring time, denoted as C. Formally, in our dataset, we have observed pairs (Y, δ), where Y corresponds to the minimum of T and C (i.e., $Y = \min(T, C)$), and δ is an indicator variable defined as $\delta = \mathbb{1}(T \leq C)$. When δ equals 1, it indicates that the individual observed the event of interest, and in this case, the data is considered uncensored. Therefore, Y is equal to the true survival time T. However, when δ equals 0, it signifies that the data is censored, indicating that the individuals did not observe the event of interest. In such instances, Y is set to the censoring time C.

2.1.2 Hypothesis of the Independence between T and C

Estimators for the survival function exist in cases of right censoring, assuming the independence of survival time and censoring time. One such estimator is the Kaplan and Meier (1958) estimator, a non-parametric method based on the order of occurrence of events and censored observations.

Let n be the number of observations $Y_1, ..., Y_n$ with censoring indicators $\delta_1, ..., \delta_n$. We can order these events : $Y_{(1)}, ..., Y_{(n)}$. In this list of ordered events, there are a total of k distinct event times, denoted as $Y^*_{(1)}, ..., Y^*_{(k)}$ with $d_{(1)}, ..., d_{(k)}$ as the corresponding counts of events. We define $r_{(j)}$ as the size of the risk set at event time $Y^*_{(j)}$. The risk set represents the individuals at risk of experiencing an event at time $Y^*_{(j)}$. This consists of individuals who have not yet experienced the event and have not been censored up to that time point. For instance, in our dentist example, the risk set at a specific time point would include all children who have not yet developed a cavity and who have not experienced censoring.

The Kaplan-Meier estimator, denoted $\hat{S}_T(t)$, estimates the survival function. It is calculated using the product of the survival probabilities at each distinct event time $Y^*_{(j)}$ and is given by (Van Keilegom, 2023),

$$\hat{S}_T(t) = \prod_{j:Y^*_{(j)} \leq t} \frac{r_{(j)} - d_{(j)}}{r_{(j)}}. \tag{2.2}$$

In this formula, the numerator represents the number of individuals in the risk group at a given time ($Y^*_{(j)}$) who have not yet experienced the event of interest. The denominator represents the total individuals in the risk set at this given time. This estimator provides an estimate of the probability that an individual survives beyond time t, considering the observed data. Further details on the derivation of the Kaplan-Meier formula can be found in the Appendix 1.

We will illustrate this formula with an example for greater clarity. Again, let's take the example of the dentist. Assume that the dentist studies the time (in months) until 20 children develop their first cavity after their first visit. Obeservation time, $Y_1, ..., Y_{20}$, are given by 3, 3, 3, 5, 6, 10, 16, 24, 24, 3+, 4+, 10+, 12+, 15+, 20+, 22+, 22+, 36+, 52+, 52+, where the sign "+" indicates censoring. We can make a table of the ordered event times with the corresponding number of events and the risk.

$Y^*_{(j)}$	$r_{(j)}$	$d_{(j)}$	$\hat{S}(t)$
3	20	3	$1 \times \frac{17}{20} = 0.85$
5	15	1	$0.85 \times \frac{14}{15} = 0.79$
6	14	1	$0.79 \times \frac{13}{14} = 0.73$
10	13	1	$0.73 \times \frac{12}{13} = 0.67$
16	9	1	$0.67 \times \frac{8}{9} = 0.60$
24	5	2	$0.60 \times \frac{3}{5} = 0.36$

Table 1: Example of Kaplan-Meier estimator

In this table, the first column $Y^*_{(j)}$ is the uncensored observed event times, $r_{(j)}$ is the number of children who are at risk at event time $Y^*_{(j)}$. The last column is the Kaplan-Meier estimator given by the formula (2.2).

The Kaplan-Meier estimator is based on the assumption of independence between T and C, an assumption not always reflective of real-world scenarios. In practical studies, the survival time and censoring time are frequently intertwined. For example, in a study on cancer mortality, it may not be reasonable to assert that deaths from other diseases or the cessation of the study due to health concerns are entirely unrelated to cancer. It is

important to note that since T and C are never observed simultaneously, the independence between these variables can not be tested in a fully nonparametric way (Van Keilegom, 2023). If we neglect the dependency between T and C, even though it is present, this can result in an overestimation or underestimation of the true survival function. The extent of this bias is influenced by the proportion of subjects that are censored (Huang and Zhang, 2008). To highlight the importance of considering the dependence between survival time and censoring time, let's explore an example inspired by (Van Keilegom, 2023). Consider,

$$(\log T, \log C) \sim N_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right).$$

Let $Y = \min(T, C)$ and $\delta = \mathbb{1}(T \leq C)$. For an arbitrary sample of size 500 and different values of ρ, we calculate the true survival function $S(t)$ of $T \sim \exp(N(0, 1))$ and the Kaplan-Meier estimator $\hat{S}_T(t)$ which assumes that T and C are independent.

In Figure 4, it can be observed that a positive correlation between the survival and censoring time results in an overestimated survival function, while a negative correlation leads to an underestimated survival curve. A positive correlation between censoring time and survival time suggests that shorter survival times are often accompanied by shorter censoring times, and vice versa. This scenario might occur when individuals with a short survival time leave the study prematurely due to severe illness, for example. Such cases lead to an overestimation of the survival function because individuals who exit the study early are no longer included in the risk set, potentially skewing survival estimates upward. Conversely, a negative correlation between T and C indicates situations where individuals with longer survival times are censored early. For instance, a person might be censored due to non-health-related reasons while still healthy. This early censoring prevents the survival function from capturing the full extent of these individuals' potential survival times, resulting in an underestimation of the survival function.

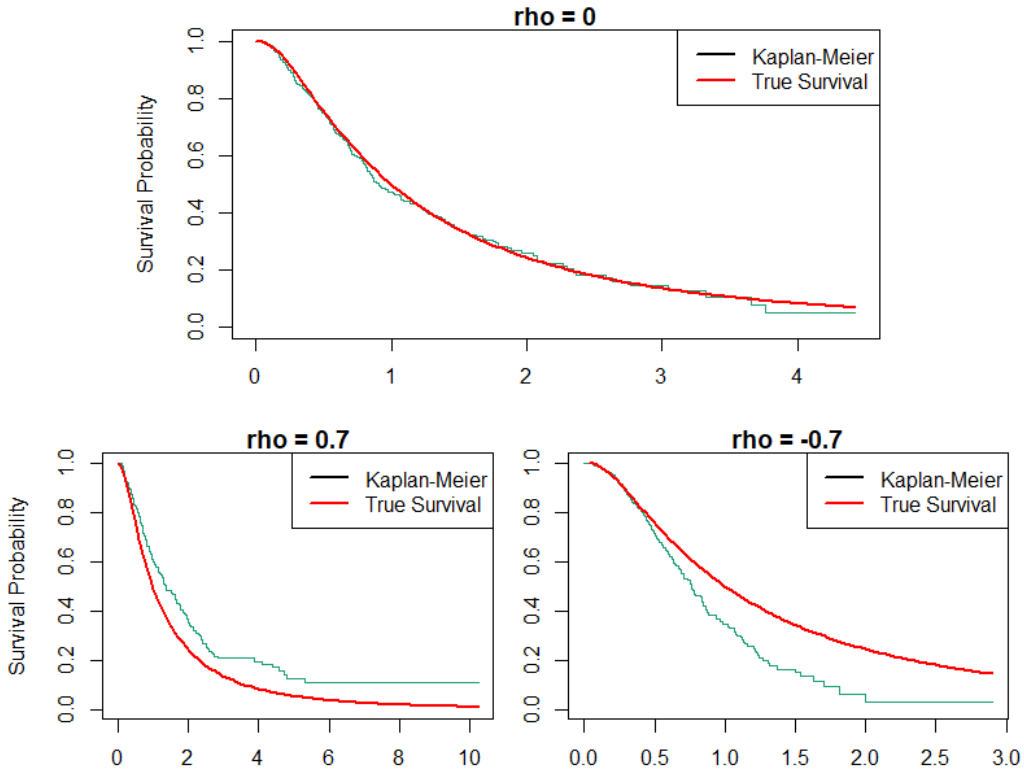


Figure 4: Wrong assumption of independence between T and C

Considering the dependence between T and C is crucial for achieving more accurate and meaningful survival analysis. To address this, we will employ copulas, which are particularly adept at modeling dependencies between variables. This approach will allow us to more precisely capture the interactions between survival time and censoring time. Further details on how we implement this methodology using copulas will be provided in the subsequent sections. In the next section, we will explore copulas and their properties in more detail.

2.2 Copulas

In this section, we will provide a brief overview of copulas based on Nelsen (2006); Schmidt (2006); Genest and Favre (2007); Emura and Chen (2018); Czado (2019); Härdle and Simar (2019). We will focus on the bivariate case since our aim is to model the dependence between two random variables, survival time and censoring time.

Copulas model the dependence between two continuous random variables, X and Y, independently of the choice of the marginal distribution of these variables, denoted by $F_X(x)$ and $F_Y(y)$, respectively. A copula, denoted as $\mathscr{C}$, is a bivariate distribution function with uniform margins, defined as $\mathscr{C} : [0,1] \times [0,1] \to [0,1]$. The joint cumulative distribution function (c.d.f), $F_{XY}(x,y)$, of (X,Y) can be expressed as:

$$F_{XY}(x,y) = \mathscr{C}\{F_X(x), F_Y(y)\}, \quad x,y \in \mathbb{R}. \tag{2.3}$$

Sklar (1959) shows that if the marginal distributions, F_X and F_Y, are continuous, there exists a unique copula for which Equation (2.1) holds. We present some properties of copulas in the following.

2.2.1 Properties of copulas

The copula function $\mathscr{C}$ satisfies several fundamental properties :

- For every $x, y \in [0,1]$,
$$\mathscr{C}(x,0) = \mathscr{C}(0,y) = 0 \tag{2.4}$$
- For every $x, y \in [0,1]$,
$$\mathscr{C}(x,1) = x \quad \text{and} \quad \mathscr{C}(1,y) = y \tag{2.5}$$
- For every $(x_1,y_1),(x_2,y_2) \in [0,1] \times [0,1]$ with $x_1 \leq y_1$ and $x_2 \leq y_2$:
$$\mathscr{C}(y_1,y_2) + \mathscr{C}(x_1,x_2) - \mathscr{C}(y_1,x_2) - \mathscr{C}(x_1,y_2) \geq 0. \tag{2.6}$$

Another interesting property of copulas is that the unique copula associated with a random pair (X,Y) remains invariant under monotone increasing transformations of the marginals. In mathematical terms, if ϕ and ψ are two increasing transformations, the copula of $(\phi(X), \psi(Y))$ is the same as the copula of (X,Y).

The degree of association specified by a copula can be conveniently measured using Kendall's tau, denoted as τ. This is a dependence measure based on the ranks, making

it invariant with respect to monotone transformations of the margins. Kendall's tau is defined as the probability of concordance minus the probability of discordance between two random variables X and Y,

$$\tau = \Pr\left((X_2 - X_1)(Y_2 - Y_1) > 0\right) - \Pr\left((X_2 - X_1)(Y_2 - Y_1) < 0\right)$$

where (X_1, X_2) and (Y_1, Y_2) have the same distribution as (X, Y). Kendall's tau can also be expressed in terms of the copula:

$$\tau = 4 \int_{[0,1]^2} \mathscr{C}(x, y)\, d\mathscr{C}(x, y) - 1.$$

Closed-form expressions for τ often exist in terms of the copula parameters.

2.2.2 Some classical copulas

Dependence and Independence

It is easy to observe that X and Y are independent continuous random variables if and only if the copula is the product copula $\mathscr{C} = \Pi$, where $\Pi(x, y) = x \times y$ for all $x, y \in [0, 1]$.

We can also introduce the Fréchet-Hoeffding bounds as follows:

$$W(x, y) = \max(0, x + y - 1) \qquad M(x, y) = \min(x, y),$$

with $x, y \in [0, 1]$. When the copula $\mathscr{C}$ is equal to W, it indicates that Y is a decreasing function of X. On the other hand, when the copula $\mathscr{C}$ is equal to M, it suggests that Y is a monotone increasing function of X. Any copula $\mathscr{C}$ models a dependence that lies between these two Fréchet-Hoeffding bounds,

$$W(x, y) \leq \mathscr{C}(x, y) \leq M(x, y)$$

for $x, y \in [0, 1]$.

Copulas obtained from distributions

We can also derive copulas from the distribution of the random variables X and Y. For example, the multivariate normal distribution leads to the Gaussian copula, while the multivariate Student-t distribution leads to the t-copula. These copulas belong to the family of

Elliptic copulas.

To better understand where these copulas come from, it is interesting to see that from the relationship 2.3, we can obtain,

$$\mathscr{C}(x, y) = F(F_X^{\leftarrow}(x), F_Y^{\leftarrow}(y)). \tag{2.7}$$

Where $F^{\leftarrow}$ is the generalized inverse of F given by,

$$F^{\leftarrow}(y) := \inf(x : F(x) \geq y)$$

Based on the relation (2.7) we can obtain copula from a bivariate distribution function. For example the Gaussian copula is given by

$$\mathscr{C}_\rho(u, v) = \Phi_\Sigma \left(\Phi^{-1}(u), \Phi^{-1}(v)\right),$$

where Φ denotes the cumulative distribution function (cdf) of a standard normal distribution, Φ_Σ is the cdf of a bivariate normal distribution with mean zero and a covariance matrix equal to the correlation matrix, $\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$, and ρ denotes the correlation coefficient. For normal distribution , independence is equivalent to zero correlation, so for $\rho = 0$, Gaussian copula is equal to the independence copula.

Similar to the Gaussian copula, the t - copula or Student copula is given by

$$\mathscr{C}^t_{\nu,\Sigma}(u, v) = t_{\nu,\Sigma} \left(t_\nu^{-1}(u), t_\nu^{-1}(v)\right)$$

where t_ν denotes the cumulative distribution function (cdf) of the one-dimensional Student distribution, $t_{\nu,\Sigma}$ is the cdf of the two-dimensional Student distribution and $\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ is the correlation matrix. More information about the two-dimensional Student distribution can be found in Schmidt (2006); Roth (2012).

Explicit copulas

In contrast to the elliptic copula, there are copulas that can be directly expressed with relatively simple forms. Archimedean copulas are one of them. Archimedean copulas are

mathematically represented as

$$\mathscr{C}_\theta(u,v) = \phi^{[-1]}\left(\phi(u) + \phi(v)\right).$$

Where, the parameter θ describes the degree of dependence, $\phi : [0,1] \to [0,\infty]$, is a continuous and strictly decreasing function from $\phi(0) > 0$ to with $\phi(1) = 0$, known as the generator of the copula. The term $\phi^{[-1]}$ denotes the pseudo-inverse of ϕ, and is defined as

$$\phi^{[-1]} = \begin{cases} \phi^{-1}(t) & 0 \leq t \leq \phi(0) \\ 0 & t \geq \phi(0). \end{cases}$$

Under these conditions, $\mathscr{C}_\theta$ satisfies the condition (2.4) and (2.5). The generator have to be a convex function to satisfies the condition (2.6). The proof verifying these three conditions can be found in Nelsen (2006), Theorem 4.1.4.

Kendall's tau have a simpler form for the Archimedean copula, it is given by,

$$\tau = 1 + 4\int_0^1 \frac{\phi(t)}{\phi'(t)}dt$$

In the following, we present examples of some Archimedean copulas.

The Gumbel copula denoted as $\mathscr{C}_\theta^{Gu}(u,v)$, is defined as:

$$\mathscr{C}_\theta^{Gu}(u,v) = \exp\left(-\left((-\log u)^\theta + (-\log v)^\theta\right)^{\frac{1}{\theta}}\right).$$

In this expression, the generator is given by $\phi_\theta(u) = (-\log(u))^\theta$ where $\theta \in [1,\infty)$. When θ is equal to 1, the Gumbel copula reduces to the product copula, indicating independence. Kendall's tau is given by $\tau = \frac{\theta-1}{\theta}$.

The Clayton copula, denoted as $\mathscr{C}_\theta^{Cl}(u,v)$, is expressed as:

$$\mathscr{C}_\theta^{Cl}(u,v) = \left(\max\left(u^{-\theta} + v^{-\theta} - 1, 0\right)\right)^{-\frac{1}{\theta}}.$$

Here, the generator is given by $\phi_\theta(u) = \theta^{-1}(u^{-\theta} - 1)$ where $\theta \in (0,\infty)$. Notably, as θ tends to zero, the Clayton copula converges to the product copula. Kendall's tau is given by $\tau = \frac{\theta}{\theta+2}$.

The Frank copula, denoted as $\mathscr{C}_\theta^{Fr}(u,v)$, is defined as:

$$\mathscr{C}_\theta^{Fr}(u,v) = -\frac{1}{\theta}\log\left(1+\frac{(e^{-\theta u}-1).(e^{-\theta v}-1)}{e^{-\theta}-1}\right).$$

Here, the generator is given by $\phi_\theta(u) = -\log\left(\frac{\exp(-\theta u)-1}{\exp(-\theta)-1}\right)$ where $\theta \in \mathbb{R}\backslash\{0\}$. Notably, as θ tends to zero, the Frank copula converges to the product copula. Here, the formula for Kendall's tau is a bit more complicated, $\tau = 1 - \frac{4}{\theta}\left(1 - \frac{1}{\theta}\int_0^\theta \frac{t}{\exp(t)-1}\,dt\right)$.
There are additional examples of Archimedean copulas, but they will not be explored in this work.

Copula based on the data

The empirical copula, a non-parametric approach to copula estimation based on the ranks, can be defined using a sample $(X_i, Y_i)_{i=1,\dots,n}$ of (X,Y). The empirical copula $\mathscr{C}_n(u,v)$ is given by:

$$\mathscr{C}_n(u,v) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left(\frac{R_i}{n+1} \leq u, \frac{S_i}{n+1} \leq v\right)$$

Here R_i represents the rank of X_i, S_i is the rank of Y_i and $\mathbb{1}(A)$ is the indicator function of the set A. More information about this copula can be found in (Deheuvels, 1979).

While there exist various other methods for estimating a copula, we won't delve into them in this discussion. For this master's thesis, we will only use the Archimedean copula. Now that we have explained survival analysis and copulas in more detail, and introduced the necessary notations, we can introduce our methodology for constructing our estimator in the next section.

3 Methodology

As highlighted in the theoretical framework, the assumption of independence between censoring time (C) and survival time (T) may not accurately reflect real-world scenarios. The literature offers diverse approaches aimed at modeling the dependence between these two critical components. In this section, we will delve into a comprehensive review of the existing methodologies proposed in the field. We will categorize these approaches into two groups: methods based on copulas and methods not based on copulas. Subsequently, we will delve deeper into the details of Czado and Van Keilegom (2023) and Rivest and Wells (2001) to lay the foundation for our methodology. This section concludes with the construction of our proposed model.

Methods based on copulas

The pioneering work in this domain is the paper by Zheng and Klein (1995). They addressed the issue of dependent censoring by introducing a copula-graphic (CG) estimator. They employed a known copula and extended the Kaplan & Meier estimator to handle dependent censoring. Although the paper provided a partial solution to the identifiability problem, it does not give a closed form of this estimator. Later, Rivest and Wells (2001) achieved an explicit form of the copula-graphic estimator proposed by Zheng and Klein for Archimedean copulas. They derived its asymptotic properties using a martingale approach and conducted a sensitivity analysis on the choice of the assumed copula.

Other researchers have explored the dependence between survival time and censoring time using a completely known copula. For example, Escarela and Carriere (2003) proposed a fully parametric model with a copula function to model the dependence structure between survival time and censoring time, finding that this model is more accurate than a simpler one. In their paper, Braekers and Veraverbeke (2005) extended the copula graphic estimator to deal with one covariate. However, this method is quite restrictive for medical studies due to its limitation to handling only one covariate. In their study, Una-Álvarez and Veraverbeke (2013) employed an Archimedean copula graphic estimator and derived an asymptotic representation of the estimator as a sum of independent and identically distributed random variables. Emura and Chen (2016) examined a copula-based estimator to assess the bias

induced by dependent censoring on gene selection.

A significant limitation of these methods is their dependence on a fully known copula with a specified association parameter. In practical applications, the association parameter is usually unknown and can have a major impact on the resulting estimator for the marginal distribution (Czado and Van Keilegom, 2023). To illustrate this point, we applied the estimator of Rivest and Wells (2001) to the "Lung" data set available in the "compound.Cox" package. This is a subset of the lung cancer data consisting of 97 gene expressions from 125 patients with non-small-cell lung cancer (Chen et al., 2007). We estimate the survival function with the copula graphic estimator proposed by Rivest and Wells with a Frank copula. Initially, we set the association parameter, θ, to 1, which corresponds to a τ value of 0.11. Subsequently, we adjust the value of θ to 10, leading to a τ value of 0.66. The results are given in Figure 5. We can clearly see an important difference in the estimation of the survival function, highlighting the importance of the choice of the association parameter.

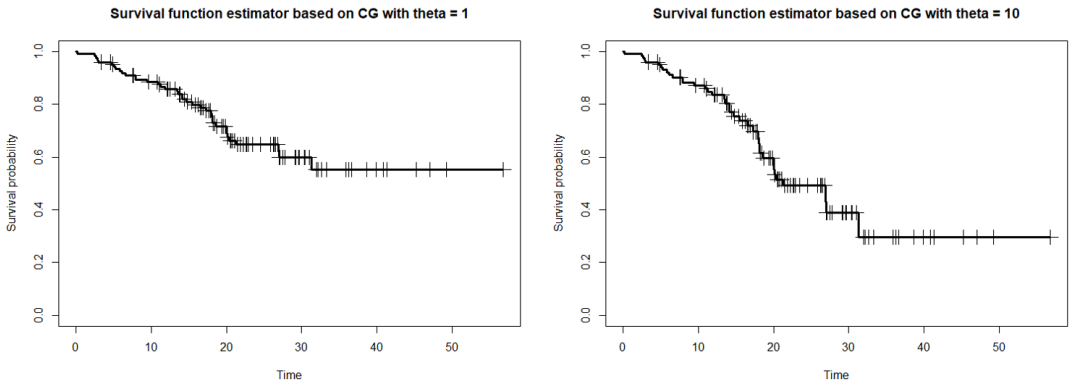


Figure 5: Impact of the association parameter

Several papers have attempted to estimate the copula parameter within a parametric model, although the identifiability of such models remains unclear. For instance, Shih et al. (2019) and Fan et al. (2019) work with a restrictive parametric model, while Emura and Chen (2016) estimate the dependence parameter using a cross-validation criterion. Recently Czado and Van Keilegom (2023) have demonstrated that the assumption of a fully known copula is not essential for identifying the joint parametric distribution of T and C. Deresa et al. (2022)

proposed a fully parametric copula-based model for analyzing bivariate survival data, accommodating dependent censoring, left truncation, and administrative censoring when only the first event time is observable. To evaluate the validity of the copula model they implemented, they also introduced a goodness-of-fit test. Delhelle and Van Keilegom (2024) also proposed a completely parametric model, where the copula parameter is unknown, in the presence of dependent censoring and a cure fraction. Although in these papers the copula parameter need not be known, their methods rely on the fact that the original distributions of survival time, T, and censoring time, C, are parametric. Deresa and Van Keilegom (2023) propose a model that utilizes a parametric copula to describe the relationship between the survival time and the censoring time. In this model, the marginal distribution of T is handled by a semiparametric Cox proportional hazards model, while C is modeled using a parametric approach. In their later work, Deresa and Van Keilegom (2024) propose a model that accounts for dependent survival and censoring times, with an additional independent administrative censoring time. This model utilizes a parametric Archimedean copula to handle the dependence between survival and censoring times, without requiring prior knowledge of the association parameter. Furthermore, it incorporates a semiparametric transformation model for survival and censoring times, which includes semiparametric proportional hazard and proportional odds models as special cases.

Methods not based on copulas

Other researchers have explored modeling the dependency between the survival time and censoring time without resorting to copulas. For instance, Emoto and Matthews (1990) employed a bivariate Weibull model for this purpose and concluded that the maximum likelihood estimates for the parameters of the joint distribution are consistent. In a continuous-time setting, Scharfstein and Robins (2002) delved into semiparametric models, while Jackson et al. (2014) developed a procedure to quantify the sensitivity of conclusions from fitted Cox proportional hazards models when the independent censoring assumption is questionable. Additionally, Slud and Rubinstein (1983) constructed an estimator that relies on knowledge of a specific dependence function, dependent on the first partial derivatives of the joint distribution and the form of the unobservable marginal distributions. We could mention other authors in this context, including Klein and Moeschberger (1988), Basu and Ghosh (1978),

Deresa and Van Keilegom (2020), Deresa and Van Keilegom (2021), among others.

However, we won't extensively cover non-copula-based methods in this master thesis. The choice of copula-based models for our study is driven by their widespread application in the analysis of survival data under dependent censoring. From a methodological perspective, copulas present a more transparent and intuitive strategy for constructing a model of dependent censoring and estimating parameters within the model (Emura and Chen, 2018).

In essence, copulas offer a flexible framework for bivariate survival models, facilitating survival analysis in the presence of dependent censoring. They place no restrictions on the choice of marginal survival models, allowing for the selection of any copula. This flexibility enhances adaptability across various types of data, making copulas a valuable tool in survival analysis, (Emura and Chen, 2018).

The primary goal of this master's thesis is to provide an estimation of the survival function in scenarios where survival time and censoring time are dependent. To achieve this goal, we will combine models presented in the two seminal works of Czado and Van Keilegom (2023) and Rivest and Wells (2001). Before clarifying our model, we provide an overview of the models presented in these two papers.

3.1 Resume of Czado and Van Keilegom (2023)

Czado and Van Keilegom (2023) demonstrate that the assumption of a completely known copula is unnecessary for identifying the joint distribution of T and C. In particular, they established the identifiability of the association parameter of the copula function. However, the trade-off for achieving this identifiability is that the marginals are no longer fully non-parametric.

The model

The authors assume that T and C are nonnegative, and the marginal distributions F_T and F_C of T and C, respectively, are continuous and belong to parametric families:

$$F_T \in \{F_{T,\theta_T} : \theta_T \in \Theta_T\}, \qquad F_C \in \{F_{C,\theta_C} : \theta_C \in \Theta_C\},$$

for certain parameter spaces Θ_T and Θ_C. The bivariate distribution of (T, C), denoted as $F_{T,C}$, is modeled by a parametric copula with an unknown parameter. Following Sklar (1959) and leveraging the continuity of the marginal distributions, it is known that there exists a unique copula $\mathscr{C}$ for which

$$P(T < t, C < c) = F_{T,C}(t, c) = \mathscr{C}\left(F_T(t), F_C(c)\right) \qquad \forall t, c \geq 0.$$

Here, the copula $\mathscr{C}$ belongs to a parametric copula set, i.e. $\mathscr{C} \in \{\mathscr{C}_\theta : \theta \in \Theta\}$ where Θ represents the parameter space.

In their methodology, the authors require expressing conditional distribution functions in terms of their corresponding copulas,

$$h_{C|T,\theta}(v|u) = \frac{\partial}{\partial u}\mathscr{C}_\theta(u, v) \qquad h_{T|C,\theta}(u|v) = \frac{\partial}{\partial v}\mathscr{C}_\theta(u, v).$$

Let $F_{Y,\delta}(y, \Delta) = \mathrm{P}(Y \leq y, \delta = \Delta)$ and $f_{Y,\delta}(y, \Delta) = \frac{\partial}{\partial y}F_{Y,\delta}(y, \Delta)$ for $\Delta = 0, 1$. Additionally, let $F_Y(y) = \sum_{\Delta=0}^{1} F_{Y,\delta}(y, \Delta)$ represent the distribution of the observable random variable Y and let $f_Y(y) = \sum_{\Delta=0}^{1} f_{Y,\delta}(y, \Delta)$ be the density of the observable random variable Y. The authors express

$$\begin{aligned}
F_{T|C}(t|c) &= h_{T|C}\left(F_T(t)|F_C(c)\right) \\
F_{C|T}(c|t) &= h_{C|T}\left(F_C(c)|F_T(t)\right) \\
F_Y(y) &= F_T(y) + F_C(y) - \mathscr{C}\left(F_T(y), F_C(y)\right) \\
f_{Y,\delta}(y, 1) &= f_T(y)[1 - h_{C|T}\left(F_C(y)|F_T(y)\right)] \\
f_{Y,\delta}(y, 0) &= f_C(y)[1 - h_{T|C}\left(F_T(y)|F_C(y)\right)]
\end{aligned}$$

The derivation of the result can be found in Appendix of paper (Czado and Van Keilegom, 2023). and the proof establishing the identifiability of this model is clearly presented in Section 3 of (Czado and Van Keilegom, 2023).

Estimator

To estimate the parameters in the joint parametric model for the survival time T and the censoring time C, Czado and Van Keilegom (2023) assume the availability of an independent and identically distributed sample $\mathcal{D} = \{(Y_i, \Delta_i), i = 1, ..., n\}$. The joint loglikelihood for the parameter vector $\alpha = (\theta, \theta_T, \theta_C)$ is

$$\begin{aligned}\ell(\alpha; \mathcal{D}) &= \sum_{i=1}^{n} \log\{f_{Y_i,\delta,\alpha}(Y_i, \Delta_i)\} \\ &= \sum_{\Delta_i=1} \log\{f_{T,\theta_T}(Y_i)[1 - h_{C|T,\theta}(F_{C,\theta_C}(Y_i)|F_{T,\theta_T}(Y_i)]\} \\ &\quad + \sum_{\Delta_i=0} \log\{f_{C,\theta_C}(Y_i)[1 - h_{T|C,\theta}(F_{T,\theta_T}(Y_i)|F_{C,\theta_C}(Y_i)]\}\end{aligned} \tag{3.1}$$

Then, they follow a maximum likelihood approach by maximizing the log-likelihood. They define parameter estimators by

$$\hat{\alpha} = (\hat{\theta}, \hat{\theta}_T, \hat{\theta}_C)^T = \arg\max_{\alpha \in A} \ell(\alpha; \mathcal{D}),$$

where $A = \Theta \times \Theta_T \times \Theta_C$. The authors use unconstrained optimization in data application to obtain the estimator. The asymptotic normality of $(\hat{\theta}, \hat{\theta}_T, \hat{\theta}_C)$ is demonstrated thanks to (White, 1982).

3.2 Resume of Rivest and Wells (2001)

This paper builds upon the proposal by Zheng and Klein (1995), who defined the joint distribution of T and C based on a copula and introduced a nonparametric estimator known as the copula-graphic estimator. However, Zheng and Klein did not provide an explicit form for this estimator. This paper addresses that gap by presenting the explicit form of the copula graphic estimator when the underlying copula is Archimedean.

The model

The model considered in this paper consists of pairs of independent finite random variables representing survival and censoring times $(T_i, C_i)_{i=1}^n$. As usual, the observable data are Y_i, and δ_i is the censoring indicator.

To simplify the presentation, the authors assume that the distribution of Y is absolutely continuous, making multiple deaths at the same time point impossible. The underlying survival function for (T_i, C_i) is given by an Archimedean copula,

$$P(T > t, C > c) = S_{T,C}(t, c) = \mathscr{C}(S_T(t), S_C(c))$$

where $S_T(\cdot)$ and $S_C(\cdot)$ are the marginal survival functions of T_i and C_i respectively. The authors consider only the twice differentiable generator functions satisfying $\phi(0) = \infty$ for their work.

Copula-Graphic estimator : a closed form

Under these conditions, the copula graphic estimator for the survival function $S_T(t)$ (resp. $S_C(c)$) is a right continuous, decreasing step function $\hat{S_T}(t)$ (resp. $\hat{S_C}(c)$). This function satisfies $\hat{S_T}(0) = 1$ (resp. $\hat{S_C}(0) = 1$) with jumps at points Y_i where $\delta_i = 1$ (resp. $\delta_i = 0$) such that

$$\phi^{-1}\left[\phi(\hat{S_T}(Y_i)) + \phi(\hat{S_C}(Y_i))\right] = \hat{\pi}(Y_i) \qquad \text{for } i = 1, .., n.$$

Here $\hat{\pi}(x) = \frac{1}{n}\sum \mathbb{1}[Y_i \geq x]$ is the standard estimate of the survival function of Y.

The closed form for $\hat{S}(.)$ is given by

$$\hat{S_T}(t) = \phi^{-1}\left[-\sum_{Y_i \leq t, \delta_i = 1} \phi(\hat{\pi}(Y_i)) - \phi(\hat{\pi}(Y_i) - 1/n)\right]. \tag{3.2}$$

The details of the procedure used by the authors can be found in Section 2.1 of (Rivest and Wells, 2001). When $\phi(t) = -\log(t)$, survival and censoring times are independent, and the copula graphic estimator reduces to the classical Kaplan-Meier estimator.

3.3 Development of our dependent censoring Model

In this section, we introduce a novel approach to estimating the survival function under conditions of dependent censoring. To achieve this goal, we will combine the methods of the two papers seen above, leveraging the strengths of both. The advantage of Rivest and Wells (2001)'s approach is that, based on the copula graphic estimator, it allows for the deduction of a non-parametric distribution for the survival time, T. The advantage of the method proposed by Czado and Van Keilegom (2023) is that we don't need to know the copula parameter in the model. In our model, the survival time will therefore follow a non-parametric distribution and we won't need to know the copula parameter, leading to a less restrictive model.

More concretely, we will use the likelihood (3.1) from Czado and Van Keilegom (2023) and the CG - estimator (3.2) from Rivest and Wells (2001). Within this likelihood, we will replace the parametric distribution of T with the non-parametric distribution provided by the copula graphic estimator in (Rivest and Wells, 2001). By employing this method, we aim to estimate the copula's parameter. One advantage of this method is that T becomes non-parametric and that we do not need to know the parameter of the copula; we will estimate it. Finally, we will validate our method through simulation.

To achieve our objective, we need to consider several factors. Czado and Van Keilegom (2023) model the bivariate distribution of (T, C), denoted as $F_{T,C}$, using a copula. Meanwhile, in the second paper, Rivest and Wells (2001) utilize a copula to model the survival function of (T, C), denoted as $S_{T,C}$. We need to establish a connection between the two formulas. Another point to note is that in the second paper, the CG-estimator given by (3.2) is a step function, which will lead to challenges when estimating the density of T.

From one copula to another

As mentioned above, Czado and Van Keilegom (2023) model $F_{T,C}$ using a copula, while Rivest and Wells (2001) model $S_{T,C}$ using a copula. We need to find a way to transition between these two modeling approaches. We take our inspiration from Delhelle and Van Keilegom (2024), where, in the supplementary material, a link between a copula and a survival copula

is described. The following equations demonstrate this relationship:

$$\begin{aligned}
P(T > t, C > c) &= 1 - P(T \leq t) - P(C \leq c) + P(T \leq t, C \leq c) \\
&= 1 - P(T \leq t, C \leq \infty) - P(T \leq \infty, C \leq c) + P(T \leq t, C \leq c) \\
&= 1 - \mathscr{C}(F_T(t), 1) - \mathscr{C}(1, F_C(c)) + \mathscr{C}(F_T(t), F_C(c)) \\
&= \tilde{\mathscr{C}}(1 - F_T(t), 1 - F_C(c)) \\
&= \tilde{\mathscr{C}}(S_T(t), S_C(c))
\end{aligned}$$

We find that $\tilde{\mathscr{C}}(S_T(t), S_C(c))$, the copula of the second paper, is given by $1-\mathscr{C}(F_T(t), 1)-\mathscr{C}(1, F_C(c))+\mathscr{C}(F_T(t), F_C(c))$, where $\mathscr{C}$ is the copula of the first paper. To confirm that this relationship defines a copula, we verify conditions (2.4), (2.5) and (2.6). as follows,

- When $S_T(t) = 0$, meaning $F_T(t) = 1$ (and similarly for $S_C(c)$),

$$\begin{aligned}
\tilde{\mathscr{C}}(0, S_C(c)) &= 1 - \mathscr{C}(1, 1) - \mathscr{C}(1, F_C(c)) + \mathscr{C}(1, F_C(c)) \\
&= 1 - 1 \\
&= 0 \\
\tilde{\mathscr{C}}(S_T(t), 0) &= 1 - \mathscr{C}(F_T(t), 1) - \mathscr{C}(1, 1) + \mathscr{C}(F_T(t), 1) \\
&= 1 - 1 \\
&= 0
\end{aligned}$$

- When $S_T(t) = 1$, thus $F_T(t) = 0$ (and similarly for $S_C(c)$),

$$\begin{aligned}
\tilde{\mathscr{C}}(1, S_C(c)) &= 1 - \mathscr{C}(0, 1) - \mathscr{C}(1, F_C(c)) + \mathscr{C}(0, F_C(c)) \\
&= 1 - F_C(c) \\
&= S_C(c) \\
\tilde{\mathscr{C}}(S_T(t), 1) &= 1 - \mathscr{C}(F_T(t), 1) - \mathscr{C}(1, 0) + \mathscr{C}(F_T(t), 0) \\
&= 1 - F_T(t) \\
&= S_T(t)
\end{aligned}$$

- Since $\mathscr{C}$ is a copula, it verify the condition (2.6). This translates directly to $\tilde{\mathscr{C}}$.

In this way, we establish that $\tilde{\mathscr{C}}(S_T(t), S_C(c)) = 1-\mathscr{C}(F_T(t), 1)-\mathscr{C}(1, F_C(c))+\mathscr{C}(F_T(t), F_C(c))$ is indeed a copula. This approach links the copula methodologies of the two papers.

Moving forward with our model, it is more practical to choose between modeling survival time by a copula or modeling the bivariate distribution of (T, C) by a copula. We decided to adapt the model of Czado and Van Keilegom (2023) for the case where the survival function is modeled by a copula,

$$P(T > t, C > c) = \mathscr{C}_\theta(S_T(t), S_C(c))$$

To implement this, we need to adapt the likelihood (3.1) as was done in Deresa and Van Keilegom (2024), leading to the following formulation:

$$\begin{aligned}\ell(\alpha; \mathcal{D}) &= \sum_{i=1}^{n} \log\{f_{Y_i,\Delta,\alpha}(Y_i, \delta_i)\} \\ &= \sum_{\delta_i=1} \log\{f_{T,\theta_T}(Y_i)[h_{C|T,\theta}(1 - F_{C,\theta_C}(Y_i)|1 - F_{T,\theta_T}(Y_i)]\} \\ &\quad + \sum_{\delta_i=0} \log\{f_{C,\theta_C}(Y_i)[h_{T|C,\theta}(1 - F_{T,\theta_T}(Y_i)|1 - F_{C,\theta_C}(Y_i)]\}\end{aligned} \tag{3.3}$$

Here, $h_{C|T,\theta}$ and $h_{T|C,\theta}$ are the derivatives of the survival copula function expressed in terms of their associated copulas.

Density function of T

In their paper, Rivest and Wells (2001) provide the estimate of the survival function, $\hat{S}_T$. We know that the cumulative distribution function, F_T, is simply given by $1 - S_T$. Therefore, based on the estimation of the survival function, we can obtain the estimation of the cumulative distribution function of T, $\hat{F}_T = 1 - \hat{S}_T$. Given that $\hat{S}_T$ is a step function, $\hat{F}_T$ is also a step function. As mentioned above, the density of T does not exist, we replace it by the jump size :

$$\Delta\hat{S}_T(Y_i) = \hat{F}_T(Y_i) - \hat{F}_T(Y_{i^-}) = \hat{S}_T(Y_{i^-}) - \hat{S}_T(Y_i)$$

Implementation of our model

With all components duly specified, we can more clearly explain our model. Our model follows an iterative process, starting by setting a value for the copula parameter θ. We use this

parameter to calculate the copula graphic estimator, $\hat{S}_T$, from which we derive the cumulative distribution function of the survival time, T. We then use the jump size in the adapted likelihood (3.3), enabling us to estimate the copula parameter θ and the parameter of the distribution of C. We'll repeat this procedure until convergence.

Two of the three parameters in our model are constrained: the variance of C must be positive, and depending on the Archimedean copula used, the parameter θ is also constrained. To remove these constraints, we will transform these two parameters so that they take values on the entire real axis. The variance of C will simply be transformed by taking its logarithm. We know that there exists a link between Kendall's tau and the parameter of copulas, as we have seen in Section 2.2.2. As done in Czado and Van Keilegom (2023), we will estimate the hyperbolic arctangent of Kendall's tau, which is an unconstrained parameter.

Our algorithm is the following, for simplicity of notation we have written the model in terms of constrained parameters:

1. Set the copula parameter, θ, to a fixed value, such as $\theta^{(0)} = 0$.

2. Using the approach of Czado and Van Keilegom (2023) with $\theta^{(0)}$ to estimate the copula graphic estimator,

$$\hat{S}_T^{(1)}(t) = \phi^{-1}\left[-\sum_{Y_i \leq t, \delta_i = 1} \phi(\hat{\pi}(Y_i)) - \phi(\hat{\pi}(Y_i) - 1/n)\right],$$

   where ϕ is the generator function of an Archimedean copula, depending on the parameter $\theta^{(0)}$. Based on $\hat{S}_T$, we estimate the cdf of T and the jump's size,

$$\hat{F}_T^{(1)}(t) = 1 - \hat{S}_T^{(1)}(t) \qquad \Delta\hat{S}_T^{(1)}(Y_i) = \hat{S}_T^{(1)}(Y_{i^-}) - \hat{S}_T^{(1)}(Y_i)$$

3. Maximizing the adapted likelihood (3.3) of Czado and Van Keilegom (2023) with $\Delta\hat{S}_T^{(1)}(Y_i)$ $\hat{F}_T{}^{(1)}$ to estimate $\hat{\theta}^{(1)}$,

$$\begin{aligned}\ell(\alpha; \mathcal{D}) = &\sum_{\delta_i = 1} \log\{\Delta\hat{S}_T^{(1)}(Y_i)[h_{C|T,\hat{\theta}^{(0)}}(1 - F_{C,\theta_C}(Y_i)|1 - \hat{F}_T^{(1)}(Y_i)]\} \\ &+ \sum_{\delta_i = 0} \log\{f_{C,\theta_C}(Y_i)[h_{T|C,\hat{\theta}^{(0)}}(1 - \hat{F}_T^{(1)}(Y_i)|1 - F_{C,\theta_C}(Y_i)]\}\end{aligned}$$

By maximizing this likelihood we obtain $\hat{\theta}^{(1)}$, the copula parameter and $\hat{\theta_C}$, the estimation of the parameter of the parametric distribution of C.

4. Repeat steps 2 and 3, until convergence is achieved.

We implement our model in R. The code can be found in the Appendix. In the next section, we will test our model using several simulation scenarios before applying it to real data.

4 Simulation

In this section, we will evaluate the performance of our model across different marginal distributions and Frank and Gumbel copulas. Additionally, we will assess the quality of key model parameters, specifically the distribution parameter of the censoring time, C, and Kendall's tau, τ, related to the copula. Kendall's tau is a measure of dependence, which is equivalent to the theta parameter of the copula and easier to interpret. As previously demonstrated, it is defined by the following formula:

$$\tau = 4 \int_0^1 \int_0^1 \mathscr{C}(u, v) c(u, v) du dv - 1,$$

where $c(u, v)$ is the copula density.

We need to choose appropriate starting values to maximize the likelihood. To find these starting values, we took our inspiration from the paper by Delhelle and Van Keilegom (2024). For the parameter of the distribution of C, we select the maximum likelihood estimators under the distribution followed by C using only the uncensored observations. To this, we add 4 perturbations by multiplying these estimators by 4 random values, U_1, U_2, U_3 and U_4 from a uniform distribution, $U(0.5, 1.5)$. This leaves us, for now, with 5 different starting values: the undisturbed one and four disturbed ones. We use 10 starting values for Kendall's tau: 0.05, 0.15, 0.25, ..., 0.95. Combining the different starting values for the distribution of C with the grid of values of Kendall's tau, leads to 50 starting parameter vectors.

In order to maximize the likelihood, we have to simulate data for the survival time, T, and the censoring time, C. We will investigate two main scenarios and three secondary scenarios. Our main cases will be where both distributions are log normal. The parameters of a lognormal variable, Y, are μ and σ corresponding, respectively, to the mean and the standard deviation of $\log(Y)$. We investigate two scenarios for the marginal distribution of T and C. The parameters of the first scenario are $(1.5, 1)$ for the distribution of T and $(1.4, 0.6)$ for the distribution of C. In the second scenario, the parameters of T and the parameters of C are both $(0, 1)$. For these two scenarios, we study three values of Kendall's tau, $\tau = 0.2$, $\tau = 0.5$ and $\tau = 0.7$ and four sample sizes: $n = 250$, $n = 500$, $n = 1000$ and $n = 2000$. We repeated each simulation 200 times. We report the bias between our estimator and the true

value, e.g. for the mean parameter of C, the bias is $\hat{\mu}_C - \mu_C$. If the bias is negative, it means that our estimator underestimates the true value of the parameter, and if the bias is positive, our estimator overestimates the true parameter. For all the scenarios, the tables with the estimated values of the parameters instead of the biases can be found in Appendix 2. We also report the empirical standard error of the estimate and the mean absolute deviation between the true distribution function and our estimator and between the true distribution function and the Kaplan-Meier estimator. The mean absolute deviation (MAD) is defined for our estimator as,

$$\text{MAD}(\hat{F}_T) = |\hat{F}_T - F_T|$$

where $\hat{F}_T$ is given by our method and F_T is the true cdf of T. The mean absolute deviation for the Kaplan-Meier estimator is defined as,

$$\text{MAD}(\hat{F}_{T,\text{KM}}) = |\hat{F}_{T,\text{KM}} - F_T|$$

where $\hat{F}_{T,\text{KM}}$ is the cdf of T calculated by using the Kaplan-Meier estimator.

Note that in the results table presented, the columns labeled "$\hat{F}_T$" and "$\hat{F}_{T,\text{KM}}$" do not display bias values but rather mean absolute deviations (MAD) between the true distribution function and our estimator, and between the true distribution function and the Kaplan-Meier estimator, respectively. For clarity and ease of reading, these MAD values are reported alongside the bias, although they represent a different type of statistical measure.

Results are shown for Frank and Gumbel copulas, in Table 4 and Table 5 respectively for the first scenario, and in Table 6 and Table 7 for the second scenario. We also investigate three secondary scenarios for only the Frank copula. As for the primary scenario, we study three values of Kendall's tau, 0.2, 0.5, and 0.7. We repeated the simulation 200 times and reported the estimate, the empirical standard error of the estimate and the mean absolute deviation between the true distribution function and our estimator and between the true distribution function and the Kaplan-Meier estimator. In the first secondary scenario (scenario 3), the distributions of T and C are normal, we only investigate the case where $T \sim N(0,1)$ and $C \sim N(0,1)$. The second secondary scenario (scenario 4) studies the case where T follows a lognormal distribution with parameters $(0, 1.7)$ and C follows a Weibull distribution

with parameters for shape $k = 1$ and scale $\lambda = 1.5$. A larger λ indicates a more spread-out distribution, while k controls the distribution's shape. In our final secondary scenario, we explore what happens when C is assumed to follow a lognormal distribution for parameter estimation, despite actually following a Weibull distribution. For this last scenario, T is simulated with a lognormal distribution of parameter $\mu_T = 0$ and $\sigma_T = 1$ and C with both parameters, scale and shape, of 1. Results are shown in Table 8 for the first secondary scenario, in Table 9 for scenario 4, and in Table 10 for the last scenario. A summary of all scenarios is available in Table 3

The uncensoring probabilities, $P(\delta = 1)$, are presented in Table 2. In scenarios 1 and 4, we observed a decrease in uncensoring probabilities as the strength of dependence increased. For scenarios 2 and 3, the uncensoring probabilities remain constant regardless of the studied dependence strength. Scenario 4 displays the lowest uncensoring probabilities, which, similar to the previous two scenarios, remain constant across different levels of dependence.

Scenario	Copula	$\tau = 0.2$	$\tau = 0.5$	$\tau = 0.7$
1	Frank	0.46	0.44	0.41
	Gumbel	0.47	0.47	0.45
2	Frank	0.50	0.50	0.50
	Gumbel	0.50	0.50	0.50
3	Frank	0.50	0.50	0.50
4	Frank	0.47	0.45	0.44
5	Frank	0.39	0.39	0.39

Table 2: Uncensoring probabilities for all scenario

We observe numerical errors in less than 1.5% of cases for all scenarios. These errors occurred more frequently with larger sample sizes.

	Copule	dist.T	dist.C	par.T	par.C	Result
Scénario 1	Frank	lnorm	lnorm	(1.5,1)	(1.4,0.6)	Table 4
	Gumbel	lnorm	lnorm	(1.5,1)	(1.4,0.6)	TAble 5
Scenario 2	Frank	lnorm	lnorm	(0,1)	(0,1)	Table 6
	Gumbel	lnorm	lnorm	(0,1)	(0,1)	Table 7
Scenario 3	Frank	norm	norm	(0,1)	(0,1)	Table 8
Scenario 4	Frank	lnorm	Weibull	(0,1.7)	(1,1.5)	Table 9
Scenario 5	Frank	lnorm	Weibull	(0,1)	(1,1)	Table 10

Table 3: Summary of all scenarios

4.1 Main scenario

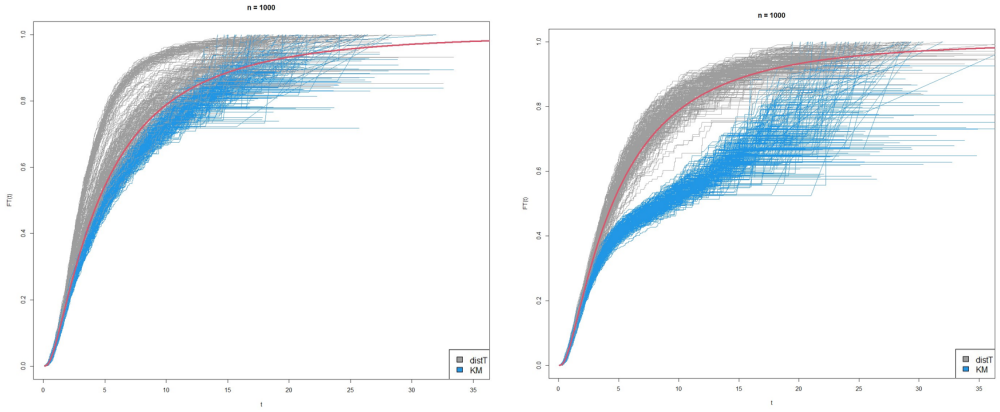
Scenario 1

Table 4 presents results for the first scenario involving a Frank copula. The parameter of primary interest, and the most challenging to estimate, is the τ parameter. Before focusing on this parameter, we analyze the results for the parameters of the C distribution: $\hat{\mu}_C$ and $\hat{\sigma}_C$. Across all sample sizes, the bias and the standard deviation for $\hat{\mu}_C$ decrease as Kendall's tau increases, indicating that the estimated value progressively aligns more closely with the true value and becomes more accurate. Larger sample sizes yield more accurate estimations for μ_C. The bias of the estimated standard deviation, $\hat{\sigma}_C$, also decreases with increasing sample size and τ values.

Next, we consider the estimator of Kendall's tau, $\hat{\tau}$. At a true value of 0.2 for Kendall's tau, our estimator performs poorly for smaller sample sizes (n = 250 and n = 500). The bias for $\hat{\tau}$ is positive and high, meaning that our estimator overestimates the true value of the dependency and gives a value that misrepresents the weak dependence between censoring time and survival time. As the sample size increases, the bias of τ decreases, meaning that the estimator gets closer to the true value. For a large sample size, $n = 2000$, the bias of $\hat{\tau}$ is 0.093, meaning that our estimator correctly estimates the value of τ even if the dependency is low. The standard deviation remains fairly constant as the sample size increases.

For $\tau = 0.5$, the bias of our estimator begins to decrease substantially from a sample size of 1000. When $\tau = 0.7$, the reduction in bias starts from a sample size of 500. Additionally, the estimated standard deviation tends to decrease as the value of τ increases, indicating an increase in the precision of the estimator as the dependency strengthens.

The last two columns of Table 4 represent the MAD of our estimator and the Kaplan-Meier estimator. For small dependencies, when $\tau = 0.2$, the Kaplan-Meier estimator seems better than ours, especially when the sample size is small. When the sample size increases, the bias of the two estimators is the same. This can be explained by the fact that Kaplan Meier is an estimator for cases where the survival time is independent of the censoring time. A 0.2 value of τ represents a situation where the dependency between the censoring time, C, and the survival time, T, is low. When the dependence between T and C increases, i.e. for higher τ values, our estimator is less biased than the Kaplan-Meier estimator, especially for large sample sizes. Note that the bias of our estimator tends to decrease as the sample size increases, for all τ values. To illustrate the biases between the estimators of the cdf of T and the true values of the cdf of T, we show two graphs, in Figure 6.



(a) $\tau = 0.2$ (b) $\tau = 0.7$

Figure 6: Graph of Kaplan-Meier estimator (blue) and or estimator (grey) for scenario 1 with a Frank copula. Result for $n = 1000$ and for $\tau = 0.2$ and $\tau = 0.7$

		$n = 250$					$n = 500$				
τ		$\hat{\mu}_C$	$\hat{\sigma}_C$	$\hat{\tau}$	$\hat{F}_T$	$\hat{F}_{T,\text{KM}}$	$\hat{\mu}_C$	$\hat{\sigma}_C$	$\hat{\tau}$	$\hat{F}_T$	$\hat{F}_{T,\text{KM}}$
0.2	Bias	-0.230	0.040	0.603	0.101	0.035	-0.187	0.033	0.489	0.0854	0.031
	SD	0.103	0.04	0.232	0.034	0017	0.114	0.036	0.282	0.041	0.014
0.5	Bias	-0.099	0.052	0.266	0.059	0.073	-0.074	0.034	0.493	0.047	0.070
	SD	0.086	0.078	0.193	0.025	0.024	0.079	0.044	0.181	0.024	0.017
0.7	Bias	-0.035	0.037	0.116	0.041	0.104	-0.036	0.027	0.073	0.031	0.102
	SD	0.067	0.051	0.144	0.015	0.026	0.051	0.039	0.096	0.014	0.020
		$n = 1000$					$n = 2000$				
0.2	Bias	-0.276	0.015	0.278	0.052	0.029	-0.038	0.003	0.093	0.025	0.028
	SD	0.120	0.031	0.306	0.044	0.010	0.087	0.021	0.222	0.031	0.007
0.5	Bias	-0.032	0.014	0.084	0.028	0.070	-0.027	0.004	0.041	0.018	0.069
	SD	0.066	0.028	0.154	0.019	0.012	0.044	0.015	0.105	0.011	0.009
0.7	Mean	-0.026	0.013	0.041	0.021	0.102	0.006	0.004	0.018	0.015	0.102
	SD	0.038	0.27	0.076	0.011	0.013	0.028	0.017	0.060	0.007	0.010

Table 4: Result of scenario 1 with a Frank copula. C follows a Lognormal distribution with true parameters (1.4, 0.6), and T is simulated with a Lognormal distribution of parameter $(1.5, 1)$.

For the Gumbel copula, the results for the parameters of the distribution of C are similar to those for the Frank copula. For the estimate of τ, the results are less good than for the Frank copula, when the dependency between T and C is 0.2 and 0.5, the bias of $\hat{\tau}$ is never beyond 0.100, even for large sample sizes ($n = 100$ and $n = 2000$). When Kendall's tau takes a value of 0.7, the bias of $\hat{\tau}$ is smaller. However, we can say that overall, the larger the sample size and the greater the dependence between survival time and censoring time, the more accurate the estimator appears to be. We also illustrate the MAD between estimators and true distribution with the graphs in Figure 7.

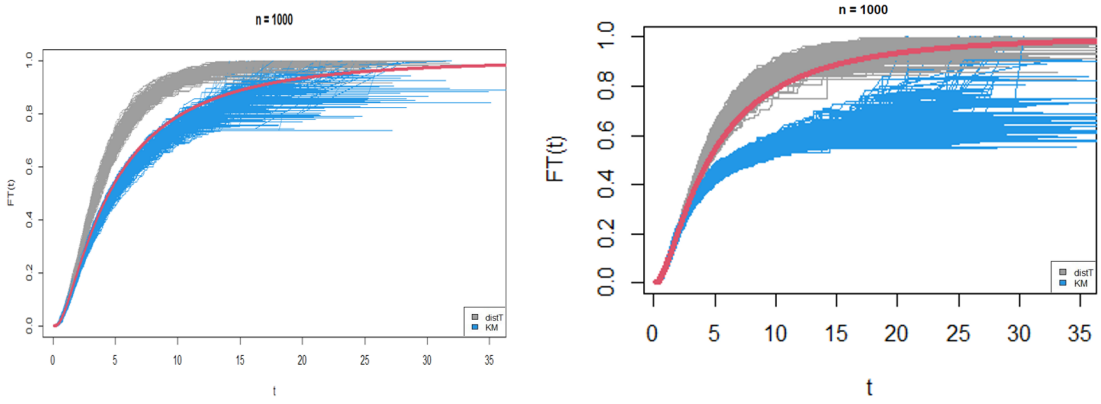
		$n = 250$					$n = 500$				
τ		$\hat{\mu}_C$	$\hat{\sigma}_C$	$\hat{\tau}$	$\hat{F}_T$	$\hat{F}_{T,\text{KM}}$	$\hat{\mu}_C$	$\hat{\sigma}_C$	$\hat{\tau}$	$\hat{F}_T$	$\hat{F}_{T,\text{KM}}$
0.2	Bias	-0.230	0.030	0.520	0.095	0.032	-0.199	0.034	0.459	0.087	0.026
	SD	0.065	0.034	0.138	0.023	0.015	0.057	0.023	0.134	0.019	0.012
0.5	Bias	-0.117	0.046	0.275	0.064	0.058	-0.098	0.039	0.227	0.056	0.055
	SD	0.068	0.042	0.129	0.021	0.022	0.060	0.033	0.130	0.019	0.016
0.7	Mean	-0.054	0.041	0.134	0.047	0.083	0.041	0.031	0.099	0.039	0.080
	SD	0.063	0.051	0.105	0.017	0.024	0.054	0.040	0.103	0.016	0.017
		$n = 1000$					$n = 2000$				
0.2	Bias	-0.164	0.025	0.369	0.076	0.024	-0.148	0.020	0.325	0.071	0.023
	SD	0.041	0.018	0.093	0.013	0.009	0.024	0.011	0.050	0.008	0.007
0.5	Bias	-0.075	0.029	0.173	0.047	0054	-0.053	0.016	0.115	0.040	0.054
	SD	0.049	0.026	0.110	0.015	0.011	0.040	0.018	0.087	0.011	0.008
0.7	Bias	-0.032	0.023	0.077	0.031	0.080	-0.029	0.012	0.044	0.024	0.079
	SD	0.038	0.029	00076	0.012	0.012	0.033	0.022	0.068	0.011	0.009

Table 5: Result of scenario 1 with a Gumbel copula. C follows a Lognormal distribution with true parameters (1.4, 0.6), and T is simulated with a Lognormal distribution of parameter $(1.5, 1)$.

Based on this first simulation, we can conclude that when the sample size is large, our estimator performs well with the Frank copula, even for small dependency values. This performance appears to increase as the dependency between \$T\$ and \$C\$ increases. The estimator with the Gumbel copula performs better with a larger dependency between survival and censoring times. For small sample sizes, our estimator does not perform as well, but it still outperforms the Kaplan-Meier estimator for $\tau = 0.5$ and $\tau = 0.7$ for both the Frank and Gumbel copulas.

Scenario 2

The results of scenario 2 are shown in Table 6 for the Frank copula and in Table 7 for the Gumbel copula. We start by looking at the estimation of the parameters of the distribution



(a) $\tau = 0.2$ (b) $\tau = 0.7$

Figure 7: Graph of Kaplan-Meier estimator (blue) and or estimator (grey) for scenario 1 with a Gumbel copula. Result for $n = 1000$ and for $\tau = 0.2$ and $\tau = 0.7$

of C. The conclusions are the same as for the previous scenario: both parameters, $\hat{\mu}_c$ and $\hat{\sigma}_C$, are more accurately estimated when either the sample size increases or the dependence between survival time and censoring time increases.

We begin by examining the results for the Frank copula. We examine the results by τ value, starting with the smallest. When $\tau = 0.2$, as the sample size increases, the bias of the estimated value of τ decreases, but the estimates are still not good. The bias is always high and positive, meaning that the estimated value of tau is always greater than the true value. As in the first scenario, when the dependence between T and C is small, the MAD for the Kaplan-Meier estimator is smaller than for our estimator. We note that the MAD of our estimator tends to decrease as the sample size increases and as the dependency increases.

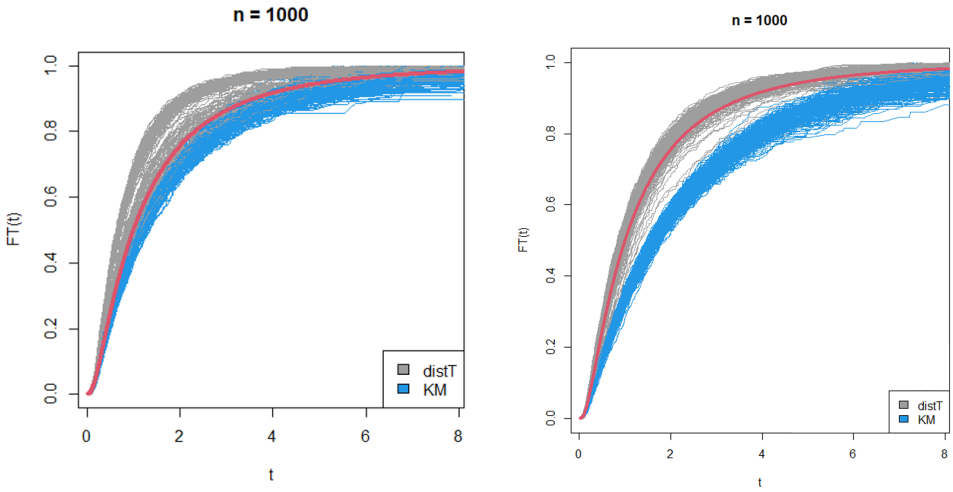
The results for $\tau = 0.5$ are also not good. The bias of the estimated value of tau is high and positive. For the last value of Kendall's tau, i.e. $\tau = 0.7$ the bias and the standard deviation of $\hat{\tau}$ are smaller than for the two previous values of τ. However, for $\tau = 0.5$ and $\tau = 0.7$, the MAD of our estimator becomes smaller than the MAD of the Kaplan-Meier estimator. This means that as the dependency between survival time and censoring time increases, our estimator becomes less biased than its competitor, Kaplan Meier. Figure 8

shows the Kaplan-Meier estimator and our estimator for a sample size of 1000 and Kendall's tau values of 0.2 and 0.7.

If we look globally at all the estimated τ values for the different cases (different Kendall's tau values and different sample sizes) obtained by summing the true value of Kendall's tau and the bias, we see that overall the average estimated τ values are very high: for sample sizes of 250 and 500 the tau values are between 0.7 and 0.8. The estimate may look better when Kendall's tau is 0.7, but that's because the estimated value returned by our estimator is always high.

		$n = 250$					$n = 500$				
τ		$\hat{\mu}_C$	$\hat{\sigma}_C$	$\hat{\tau}$	$\hat{F}_T$	$\hat{F}_{T,\text{KM}}$	$\hat{\mu}_C$	$\hat{\sigma}_C$	$\hat{\tau}$	$\hat{F}_T$	$\hat{F}_{T,\text{KM}}$
0.2	Bias	-0.339	-0.097	0.624	0.104	0.037	-0.278	-0.078	0.515	0.090	0.034
	SD	0.128	0.053	0.216	0.029	0.017	0.172	0.060	0.298	0.040	0.013
0.5	Bias	-0.168	-0.028	0.319	0.058	0.081	-0.157	-0.025	0.300	0.054	0.081
	SD	0.115	0.053	0.180	0.022	0.022	-0.105	-0.045	0.181	0.021	0.016
0.7	Bias	-0.061	-0.005	0.147	0.034	0.115	0.060	0.006	0.138	0.029	0.114
	SD	0.106	0.055	0.143	0.015	0.023	0.071	0.038	0.113	0.013	0.017
		$n = 1000$					$n = 2000$				
0.2	Bias	-0.261	-0.075	0.488	0.085	0.033	-0.173	-0.051	0.322	0.061	0.032
	SD	0.179	0.055	0.322	0.044	0.010	0.192	0.061	0.351	0.048	0.007
0.5	Bias	-0.150	-0.024	0.291	0.052	0.081	-0.126	-0.022	0.238	0.045	0.082
	SD	0.101	0.031	0.181	0.021	0.011	0.106	0.030	0.201	0.022	0.008
0.7	Bias	-0.047	-0.004	0.113	0.028	0.115	-0.044	-0.007	0.098	0.020	0.116
	SD	0.086	0.031	0.147	0.016	0.011	0.057	0.021	0.108	0.012	0.008

Table 6: Result of scenario 2 with Frank copula. C follows a Lognormal distribution with true parameters (0,1) and T is simulated with a Lognormal distribution of parameter $(0, 1)$.



(a) $\tau = 0.2$ (b) $\tau = 0.7$

Figure 8: Graph of Kaplan-Meier estimator (blue) and or estimator (grey) for scenario 2 with a Frank copula. Result for $n = 1000$ and for $\tau = 0.2$ and $\tau = 0.7$

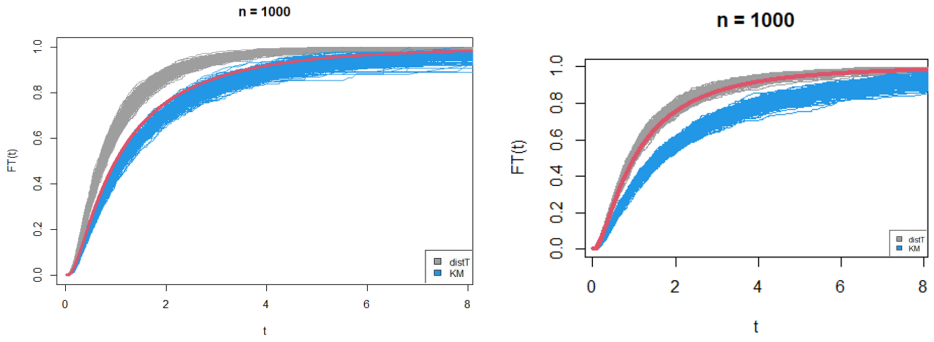
For the Gumbel copula, the average estimated τ is about 0.7 for all the true values of Kendall's tau and the sample sizes. As with the Frank copula, the results may seem better for Kendall's tau of 0.7, but this is because our estimator always returns a value of around 0.7. We also show the graph of the Kaplan Meier and our estimator with a Gumbel copula for a sample size of 1000 and Kendall's tau values of 0.2 and 0.7 in Figure 9.

We can therefore conclude that for the lognormal distribution with both distribution parameters $(0, 1)$, our model does not correctly estimate Kendall's tau value of the copula. However, for large sample sizes, our estimator is still better than its competitor, the Kaplan-Meier estimator.

With these first two simulation cases, we can see that in some situations our estimator works well, but in others, it doesn't perform very well. It is still better than its competitor, the Kaplan-Meier estimator when the sample size is large and there is sufficient dependence between survival time and censoring time.

		$n = 250$					$n = 500$				
τ		$\hat{\mu}_C$	$\hat{\sigma}_C$	$\hat{\tau}$	$\hat{F}_T$	$\hat{F}_{T,\text{KM}}$	$\hat{\mu}_C$	$\hat{\sigma}_C$	$\hat{\tau}$	$\hat{F}_T$	$\hat{F}_{T,\text{KM}}$
0.2	Bias	-0.315	-0.097	0.555	0.103	0.036	-0.304	-0.081	0.534	0.100	0.033
	SD	0.089	0.048	0.120	0.021	0.018	0.069	0.038	0.120	0.016	0.014
0.5	Bias	-0.152	-0.047	0.270	0.057	0.080	-0.139	-0.043	0.242	0.052	0.080
	SD	0.088	0.049	0.110	0.020	0.022	0.071	0.041	0.116	0.016	0.0116
0.7	Bias	0.032	0.010	0.072	0.030	0.115	0.02-	-0.010	0.051	0.024	0.114
	SD	0.092	0.050	0.116	0.013	0.023	0.074	0.042	0.114	0.011	0.016
		$n = 1000$					$n = 2000$				
0.2	Bias	-0.286	0.084	0.497	0.096	0.033	-0.273	-0.080	0.464	0.093	0.032
	SD	0.071	0.033	0.121	0.015	0.010	0.068	0.028	0.116	0.013	0.007
0.5	Bais	-0.131	-0.040	0.229	0.050	0.081	-0.122	-0.037	0.209	0.048	0.082
	SD	0.072	0.034	0.122	0.014	0.012	0.067	0.027	0.110	0.013	0.008
0.7	Bias	-0.016	-0.006	0.038	0.020	0.115	0.012	0.004	0.025	0.018	0.116
	SD	-0.070	-0.033	0.114	0.010	0.012	-0.070	-0.026	0.112	0.010	0.008

Table 7: Result of scenario 2 with Gumbel copula. TC follows a Lognormal distribution with true parameters (0,1) and T is simulated with a Lognormal distribution of parameter $(0, 1)$.



(a) $\tau = 0.2$ (b) $\tau = 0.7$

Figure 9: Graph of Kaplan-Meier estimator (blue) and or estimator (grey) for scenario 2 with a Gumbel copula. Result for $n = 1000$ and for $\tau = 0.2$ and $\tau = 0.7$

4.2 Secondary scenario

Scenario 3

The results, given in Table 8 for this scenario, illustrate a consistent trend observed in the bias of the estimated parameters of the C distribution. As the sample size increases, the bias of $\hat{\mu}_C$ decreases, indicating an improvement in the estimation of the mean of the C distribution. This improvement is more pronounced as the dependency, represented by Kendall's tau, increases. Regarding the estimation of $\hat{\sigma}_C$, similar to the outcomes of the previous simulations, it remains robust across all conditions tested.

In this scenario, the estimation of Kendall's tau, $\hat{\tau}$, shows considerable improvement. We observe a smaller bias compared to the two previous scenarios for cases where the true value of Kendall's tau is 0.2 and the sample size is 250. The bias here is approximately half that observed in other simulations. Furthermore, this estimate consistently improves as the sample size increases, approaching a value near zero for $n = 2000$ and $\tau = 0.2$. As the dependence between censoring time and survival time strengthens, our estimator also becomes more accurate. This trend of improving accuracy is more pronounced with larger sample sizes.

Consistent with earlier simulations, the MAD between our estimator and the true distribution of T decreases as both the sample size and the dependence between T and C increase. The Kaplan-Meier estimator outperforms ours when the dependence is low. However, as the dependence strengthens, with $\tau = 0.5$ and $\tau = 0.7$, our estimator regains superiority, demonstrating better performance in scenarios of moderate to high dependence.

		$n = 250$					$n = 500$				
τ		$\hat{\mu}_C$	$\hat{\sigma}_C$	$\hat{\tau}$	$\hat{F}_T$	$\hat{F}_{T,\text{KM}}$	$\hat{\mu}_C$	$\hat{\sigma}_C$	$\hat{\tau}$	$\hat{F}_T$	$\hat{F}_{T,\text{KM}}$
0.2	Bias	-0.201	-0.041	0.282	0.077	0.037	-0.135	-0.037	0.194	0.062	0.034
	SD	0.198	0.048	0.453	0.042	0.017	0.141	0.043	0.398	0.035	0.013
0.5	Bias	-0.160	-0.020	0.255	0.061	0.081	-0.092	-0.016	0.159	0.043	0.081
	SD	0.112	0.048	0.296	0.022	0.022	0.081	0.038	0.230	0.021	0.016
0.7	Bias	-0.062	-0.002	0.116	0.037	0.115	-0.005	-0.002	0.136	0.025	0.114
	SD	0.098	0.051	0.214	0.022	0.023	0.062	0.033	0.173	0.020	0.017
		$n = 1000$					$n = 2000$				
0.2	Bias	-0.155	-0.038	0.188	0.064	0.033	-0.093	-0.026	0.037	0.049	0.032
	SD	.0179	0.047	0.418	0.040	.010	0.136	0.038	0.340	0.027	0.007
0.5	Bias	-0.149	-0.023	0.271	0.053	0.081	-0.096	-0.020	0.161	0.034	0.082
	SD	0.099	0.029	0.219	0.021	0.011	0.079	0.028	0.153	0.017	0.008
0.7	Bias	-0.046	-0.004	0.114	0. 028	0.116	-0.012	-0.004	0.031	0.013	0.116
	SD	0.086	0.031	0.151	0.016	0.011	0.051	0.023	0.082	0.009	0.008

Table 8: Results of scenario 3 with a Frank copula, C follows a Normal distribution with true parameters $(0, 1)$ and T is simulated with a Normal distribution of parameter $(0, 1)$

Scenario 4

For all values of n and τ, the shape parameter $\hat{k}_C$ is well estimated. The estimation of the scale parameter improves with increasing sample size and/or with an increase in dependence between T and C. As in the previous scenario, the estimation of Kendall's tau shows considerable improvement, reaching its minimal bias reaching its minimal bias for the case of $n = 250$ and $\tau = 0.2$, compared to the same case in other scenarios.

If we look at the value of Kendall's tau of 0.2, the bias of its estimation decreases as the sample size increases reaching its smallest value at a sample size of 2000. This means that, when the censoring time follows a Weibull distribution, our estimator works very well for small dependencies and large sample sizes. The MAD of our estimator becomes smaller than the MAD of the Kaplan Meier estimator from $n = 1000$.

For the other values of Kendall's tau, $\tau = 0.5$ and $\tau = 0.7$, the MAD of our estimator is always smaller than the MAD of Kaplan Meier, meaning our estimator performs better for small and large sample sizes. For a sample size greater than 250, the bias of $\hat{\tau}$ is almost always smaller than 0.1.

		$n = 250$					$n = 500$				
τ		$\hat{k}_C$	$\hat{\lambda}_C$	$\hat{\tau}$	$\hat{F}_T$	$\hat{F}_{T,\text{KM}}$	$\hat{\mu}_C$	$\hat{\sigma}_C$	$\hat{\tau}$	$\hat{F}_T$	$\hat{F}_{T,\text{KM}}$
0.2	Bias	-0.027	-0.212	0.263	0.055	0.036	-0.010	-0.075	0.099	0.033	0.032
	SD	0.074	0.286	0.320	0.036	0.018	0.047	0.234	0.221	0.024	0.014
0.5	Bias	-0.054	-0.155	0.202	0.048	0.076	-0.029	-0.082	0.106	0.034	0.075
	SD	0.081	0.227	0.243	0.022	0.024	0.060	0.196	0.189	0.021	0.017
0.7	Bias	-0.056	-0.067	0.117	0.036	0.106	-0.033	-0.050	0.071	0.026	0.105
	SD	0.078	0.157	0.159	0.013	0.027	0.057	0.114	0.111	0.011	0.019
		$n = 1000$					$n = 2000$				
0.2	Bias	-0.001	-0.040	0.050	0.020	0.031	-0.001	-0.011	0.009	0.013	0.030
	SD	0.032	0.141	0.124	0.010	0.010	0.022	0.102	0.087	0.007	0.007
0.5	Bias	-0.013	-0.044	0.054	0.021	0.076	-0.005	-0.027	0.028	0.014	0.076
	SD	0.042	0.129	0.124	0.014	0.014	0.029	0.091	0.086	0.006	0.009
0.7	Bias	-0.021	-0.038	0.047	0.020	0.106	-0.010	-0.024	0.027	0.014	0.106
	SD	0.047	0.100	0.088	0.10	0.013	0.033	0.075	0.069	0.007	0.008

Table 9: Results of scenario 4 with a Frank copula, C follows a Weibull distribution with true parameters $(1, 1.5)$ and T is simulated with a lognormal distribution of parameter $(0, 1.7)$

Scenario 5

In this last scenario, we examine the implications of a mismatch between the true distribution and the assumed distribution model for censoring time, C. The outcomes of this approach are detailed in Table 10. Unlike the other scenarios, we report the estimated values and not the bias, since we are estimating the parameters of a lognormal distribution, whereas the

true distribution is a Weibull distribution. It makes no sense to compare the estimated parameters with the true parameters. In this scenario, we're mainly interested in the estimated value of Kendall's tau.

We can see that in this scenario none of the $\hat{\tau}$ values are well estimated (except for the very first value, in the case where $n = 250$). For sample sizes ranging from 500 to 2500, the estimated value of $\hat{\tau}$ is always very close to 0, which means that our model does not detect the presence of a dependency between survival time and censoring time.

		$n = 250$					$n = 500$				
τ		$\hat{\mu}_C$	$\hat{\sigma}_C$	$\hat{\tau}$	$\hat{F}_T$	$\hat{F}_{T,\text{KM}}$	$\hat{\mu}_C$	$\hat{\sigma}_C$	$\hat{\tau}$	$\hat{F}_T$	$\hat{F}_{T,\text{KM}}$
0.2	Mean	-0.472	1.389	0.211	0.082	0.048	-0.394	1.470	-0.002	0.51	0.047
	SD	0.179	0.209	0.378	0.052	0.017	0.089	0.099	0.110	0.020	0.015
0.5	Mean	-0.440	1.387	0.234	0.116	0.135	-0.386	1.446	0.016	0.134	0.135
	SD	0.133	0.164	0.347	0.040	0.021	0.072	0.083	0.086	0.018	0.016
0.7	Mean	-0.496	1.323	0.383	0.133	0.232	-0.454	1.376	0.059	0.216	0.231
	SD	0.101	0.133	0.356	0.096	0.022	0.062	0.078	0.139	0.045	0.017
		$n = 1000$					$n = 2500$				
0.2	Mean	-0.387	1.479	-0.021	0.051	0.046	-0.395	1.476	-0.021	0.051	0.046
	SD	0.055	0.070	0.022	0.011	0.010	0.036	0.044	0.013	0.008	0.007
0.5	Mean	-0.382	1.452	0.001	0.136	0.136	-0.384	1.453	-0.002	0.136	0.136
	SD	0.046	0.059	0.013	0.011	0.011	0.033	0.042	0.010	0.008	0.008
0.7	Mean	-0.448	1.385	0.021	0.227	0.236	-0.448	1.385	0.019	0.228	0.231
	SD	0.041	0.051	0.008	0.011	0.011	0.300	0.037	0.005	0.008	0.008

Table 10: Results for scenario 4 with a Frank copula, C follows a Weibull distribution with true parameters $(1, 1)$, but parameters are estimated assuming C follows a lognormal distribution. T is simulated with a Lognormal distribution with parameters $(0, 1)$

This scenario shows us the importance of correctly specifying the margins for our estimators. If these are not well specified, the results are not good at all. Now that we've tested our estimator with simulations, we can test it on real data in the next section.

5 Data illustration

In this section, we apply our model to pancreatic cancer data. We have chosen the same database as the paper Czado and Van Keilegom (2023) and we therefore select the same observations as they do. The database initially contains 143,157 observations and 10 variables, including survival time in months, status (alive or dead), cause of death (COD), gender, age of patients at study entry, race (white, black, other), and other tumor information (such as size, stage,...). We focus on survival time in months for black patients with localized pancreatic tumors and exclude patients with zero survival time. Leaving us with 1549 observations, 777 died of pancreatic cancer and 772 died of other causes or were still alive at the end of the study. When patients are censored due to death from another disease, their censoring time may be associated with their unobserved survival time, as various diseases often have overlapping risk factors like stress, dietary habits, and physical condition, (Czado and Van Keilegom, 2023).

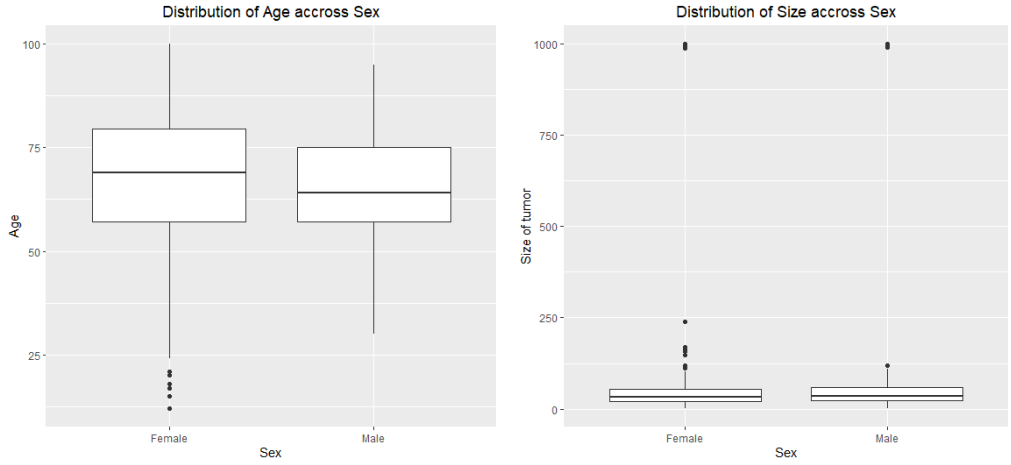
Table 11 presents a small summary of the median and average times for all observations, as well as for censored and uncensored observations. The median time for all observations is 11 months. The median time for uncensored observations, i.e. patients who died of pancreatic cancer, is much shorter, at 6 months. The median time for censored observations, i.e. patients who did not experience the event of interest before the end of the study, was 23 months, which means that censoring tends to be longer than survival time.

δ	Min	$q_{0.25}$	Median	Mean	$q_{0.75}$	Max
$\delta = 0, 1$	1.0	4.0	11.0	23.3	29.0	151.0
$\delta = 1$	1.0	3.0	6.0	10.3	13.0	88.0
$\delta = 0$	1.0	8.0	23.0	36.4	54.0	151.0

Table 11: Summary of the survival and censoring time

Although our model does not account for covariates, assessing the balance of our database is interesting. The database includes 919 black females and 630 black males. Figure 10a and 10b illustrate the distributions of age and tumor size between genders. Both variables are similarly distributed across genders, with the average ages being 65 for men and 67 for

women. Regarding tumor size, we have noted that there are a total of 153 missing values (86 for women and 67 for men), and additionally, 90 values are recorded as 999. Given that there are also some sizes like 995, 994, etc., it is difficult to ascertain whether these values are also missing. As we can see in Figure 10b, the majority of tumor sizes are smaller than 100. Since we are not analyzing the impact of covariates with our models, the presence of missing values for the size variable is not of concern to us, as there are no missing values for survival time and censoring.



(a) Distribution of Age across Sex (b) Distribution of Size across Sex

Figure 10: Descriptive analysis

In their paper, Czado and Van Keilegom (2023) found that the Gumbel copula with lognormal margins best fit the data. We will apply our estimator to this setting using both Gumbel and Frank copulas with a lognormal margin for the censoring time. The results are presented in the following table (Table 12),

Copula	μ_C	σ_C	τ	θ	loglikelihood
Gumbel	2.95	1.30	0.344	1.52	-9859.5
Frank	2.75	1.20	0.68	10.51	-9799.0

Table 12: Result of the data application for Frank and Gumbel copula with a log-normal margin.

If we compare our estimators with those obtained by Czado and Van Keilegom (2023), we can see that we obtain parameters for the distribution of C quite similar to theirs. However, for Kendall's tau, we obtain a value of 0.34 with the Gumbel copula, while they report 0.78. For the Frank copula, we found a tau value of 0.68, compared to their 0.80. Notably, our best model uses the Frank copula, which yields results closer to those reported in the paper, but with a smaller log-likelihood, indicating a lesser fit to the data.

6 Conclusion

6.1 Summary of the results

We evaluated the performance of our estimators through 4 simulation scenarios and two types of Archimedean copulas, Frank's copula and Gumbel's copula. Two of these four scenarios, scenarios 1 and 3, showed that the parameters of the C distributions and the copula association parameters were well estimated, with similar performances for the Frank and Gumbel copulas. These scenarios showed us that the greater the dependence between survival time and censoring time, and the larger the sample size, the better our estimator performed, i.e. the more precise the estimates of the three parameters.

Scenario 2, which was similar to scenario 1 with different T and C distribution parameters, showed us that in some cases our estimators were less good at estimating the copula dependence parameter, always estimating a high dependence parameter. However, the distribution parameters of C were well estimated in this scenario. It's also worth noting that, for fairly high dependence (from $\tau = 0.5$), our estimate of the distribution of T is better than its Kaplan Meier competitor.

The last scenario showed us the importance of specifying the distribution of C. In this scenario, C followed a Weibull distribution, but its distribution was estimated as if it followed a lognormal distribution. Estimates of the copula's dependence parameter are not good at all. Except in the case of a sample size of 250, the estimates of τ are always very close to 0. This means that the model does not detect the dependence between survival time and censoring time. In this scenario, the MAD of our estimator and that of Kaplan Meier are similar, which means that our estimator is no better than Kaplan Meier in this scenario.

We can therefore conclude that, when we specify the margins correctly and there is a certain dependence between T and C or a sufficiently large sample size, our estimator is better than its Kaplan Meier competitor.

When applied to real data, we obtain parameter estimates for C quite similar to those obtained in Czado and Van Keilegom (2023). For Kendall's tau parameter, we don't obtain the same value as that obtained by Czado and Van Keilegom (2023) in their paper. Moreover, the likelihood of their paper is higher than ours, indicating that their model fits the data better.

6.2 Limits and future research

The instability observed in Kendall's tau estimator for certain distribution specifications raises some questions. It remains unclear whether this instability is due to numerical problems or to inherent limitations of the estimator itself. Further study of this issue would be beneficial in determining the underlying causes and possibly refining the robustness of the estimator.

Regarding our proposed estimators, here are several possibilities for future improvement:

1. **Extension to other Archimedean copulas**: Future work could explore the adaptation of our estimator to other types of Archimedean copulas, such as Clayton or Gaussian copulas. This would broaden the applicability of the estimator across different statistical models and real-world scenarios.

2. **Non-parametric censoring times**: Currently, our estimator assumes a parametric distribution for censoring times. Adapting our approach to non-parametric censoring distributions would increase its flexibility and could address the problem of margin misspecification. Fitting a non-parametric distribution for censoring time could be done by adapting the formula 3.2 present in **?**.

3. **Inclusion of covariates**: The inclusion of covariates in the model represents an important improvement, as it would enable the model to improve its prediction and relevance by allowing an analysis of the factors influencing the variables under study.

List of Figures

1 Theoretical survival function, (Kleinbaum and Klein, 2012) 5
2 Survival function in practice, (Kleinbaum and Klein, 2012) 5
3 Hazard functions of different shapes (Van Keilegom, 2023) 6
4 Wrong assumption of independence between T and C 10
5 Impact of the association parameter 17
6 Graph of Kaplan-Meier estimator (blue) and or estimator (grey) for scenario 1 with a Frank copula. Result for $n = 1000$ and for $\tau = 0.2$ and $\tau = 0.7$ . . . 32
7 Graph of Kaplan-Meier estimator (blue) and or estimator (grey) for scenario 1 with a Gumbel copula. Result for $n = 1000$ and for $\tau = 0.2$ and $\tau = 0.7$ . . 35
8 Graph of Kaplan-Meier estimator (blue) and or estimator (grey) for scenario 2 with a Frank copula. Result for $n = 1000$ and for $\tau = 0.2$ and $\tau = 0.7$ . . . 37
9 Graph of Kaplan-Meier estimator (blue) and or estimator (grey) for scenario 2 with a Gumbel copula. Result for $n = 1000$ and for $\tau = 0.2$ and $\tau = 0.7$ . . 38
10 Descriptive analysis 44

List of Tables

1 Example of Kaplan-Meier estimator 8
2 Uncensoring probabilities for all scenario 30
3 Summary of all scenarios 31
4 Result of scenario 1 with a Frank copula. C follows a Lognormal distribution with true parameters (1.4, 0.6), and T is simulated with a Lognormal distribution of parameter $(1.5, 1)$. 33
5 Result of scenario 1 with a Gumbel copula. C follows a Lognormal distribution with true parameters (1.4, 0.6), and T is simulated with a Lognormal distribution of parameter $(1.5, 1)$. 34
6 Result of scenario 2 with Frank copula. C follows a Lognormal distribution with true parameters (0,1) and T is simulated with a Lognormal distribution of parameter $(0, 1)$. 36

7 Result of scenario 2 with Gumbel copula. TC follows a Lognormal distribution with true parameters (0,1) and T is simulated with a Lognormal distribution of parameter $(0, 1)$. 38
8 Results of scenario 3 with a Frank copula, C follows a Normal distribution with true parameters $(0, 1)$ and T is simulated with a Normal distribution of parameter $(0, 1)$ 40
9 Results of scenario 4 with a Frank copula, C follows a Weibull distribution with true parameters $(1, 1.5)$ and T is simulated with a lognormal distribution of parameter $(0, 1.7)$ 41
10 Results for scenario 4 with a Frank copula, C follows a Weibull distribution with true parameters $(1, 1)$, but parameters are estimated assuming C follows a lognormal distribution. T is simulated with a Lognormal distribution with parameters $(0, 1)$ 42
11 Summary of the survival and censoring time 43
12 Result of the data application for Frank and Gumbel copula with a log-normal margin. 45
13 Result of scenario 1 with Frank copula. C follows a Lognormal distribution with true parameters (1.4, 0.6), and T is simulated with a Lognormal distribution of parameter $(1.5, 1)$. 53
14 Result of scenario 1 with Gumbel copula. C follows a Lognormal distribution with true parameters (1.4, 0.6), and T is simulated with a Lognormal distribution of parameter $(1.5, 1)$. 54
15 Result of scenario 2 with Frank copula. C follows a Lognormal distribution with true parameters (0, 1), and T is simulated with a Lognormal distribution of parameter $(0, 1)$. 55
16 Result of scenario 2 with Gumbel copula. C follows a Lognormal distribution with true parameters (0, 1), and T is simulated with a Lognormal distribution of parameter $(0, 1)$. 56
17 Result of scenario 3 with Frank copula. C follows a Normal distribution with true parameters (0, 1), and T is simulated with a Normal distribution of parameter $(0, 1)$. 57

18 Result of scenario 4 with Frank copula. C follows a Weibull distribution with true parameters (1, 1.5), and T is simulated with a lognormal distribution of parameter $(0, 1.7)$. 58

Appendices

A Derivation of the Kaplan-Meier formula

This demonstration is based on Van Keilegom (2023), we use the same notations as in section 2.1.2, i.e. :

- n observations, $Y_1, ..., Y_n$ with censoring indicators $\delta_1, ..., \delta_n$
- $Y_{(1)}, ..., Y_{(n)}$ the ordered events
- k distinct event times ($k \leq n$), $Y^*_{(1)}, ..., Y^*_{(n)}$ and $d_{(1)}, ..., d_{(k)}$ the corresponding counts of events
- $r_{(j)}$ the size of event risk at event time $Y^*_{(j)}$

The likelihood for right-censored data is,

$$L = \prod_{i=1}^{n} f(Y_i)^{\delta_i} S(Y_i)^{1-\delta_i}$$

And so, the log-likelihood is given by,

$$\ell = \sum_{i=1}^{n} (\delta_i \log f(Y_i) + (1 - \delta_i) \log S_T(Y_i))$$

We replace the density function $f(Y_i)$ by $S_T(Y_{i^-}) - S_T(Y_i)$ and so obtained the non parametric loglikelihood,

$$\ell = \sum_{i=1}^{n} (\delta_i \log(S_T(Y_{i^-}) - S_T(Y_i)) + (1 - \delta_i \log(S_T(Y_i)))$$

Unlike a classical log-likelihood maximization over parameters, the aim here is to maximize over a class of function. The aim is to find an estimator of the survival function, $\hat{S}_T(.)$ which maximizes ℓ. It can be shown, Kaplan and Meier (1958), that the form of the survival function that maximizes the nonparametric loglikelihood is,

$$\hat{S}_T(t) = \prod_{j:Y^*_{(j)} \leq t} (1 - h_{(j)})$$

for some parameters $h_{(1)}, ..., h_{(k)}$. This is now a parametric problem which depends on k parameters. We have to find these k parameters in such a way that the log-likelihood is maximum. Let's plugging-in $\hat{S}_T(.)$ in the log-likelihood, and after some calculation we find,

$$\ell = \sum_{j=1}^{k} \left(d_{(j)} \log h_{(j)} + (r_{(j)} - d_{(j)}) \log(1 - h_{(j)}) \right)$$

To maximize this expression, we take its derivative and set it equal to 0

$$\frac{d}{dh_{(j)}} \ell = 0$$

$$\sum_{j=1}^{k} \left(d_{(j)} \frac{1}{h_{(j)}} - (r_{(j)} - d_{(j)}) \frac{1}{1 - h_{(j)}} \right) = 0$$

$$\hat{h}_{(j)} = \frac{d_{(j)}}{r_{(j)}}$$

And so, the Kaplan-Meier estimator is given by,

$$\hat{S}_T(t) = \prod_{j: Y^*_{(j)} \leq t} \frac{r_{(j)} - d_{(j)}}{r_{(j)}}$$

B Simulation tables of estimated parameters

B.1 Scenario 1

		$n = 250$					$n = 500$				
τ		$\hat{\mu}_C$	$\hat{\sigma}_C$	$\hat{\tau}$	$\hat{F}_T$	$\hat{F}_{T,\mathrm{KM}}$	$\hat{\mu}_C$	$\hat{\sigma}_C$	$\hat{\tau}$	$\hat{F}_T$	$\hat{F}_{T,\mathrm{KM}}$
0.2	Mean	1.170	0.640	0.803	0.101	0.035	1.213	0.633	0.689	0.085	0.031
	SD	0.103	0.042	0.232	0.034	0.017	0.114	0.036	0.282	0.041	0.014
0.5	Mean	1.301	0.652	0.766	0.059	0.073	1.327	0.638	0.693	0.047	0.070
	SD	0.086	0.048	0.193	0.025	0.024	0.079	0.044	0.182	0.025	0.017
0.7	Mean	1.365	0.637	0.815	0.041	0.104	1.374	0.627	0.773	0.031	0.102
	SD	0.067	0.051	0.144	0.015	0.026	0.051	0.039	0.096	0.014	0.020
		$n = 1000$					$n = 2000$				
0.2	Mean	1.294	0.615	0.478	0.052	0.029	1.362	0.603	0.294	0.025	0.028
	SD	0.120	0.031	0.306	0.044	0.010	0.087	0.021	0.222	0.031	0.007
0.5	Mean	1.368	0.614	0.584	0.028	0.070	1.383	0.604	0.541	0.018	0.069
	SD	0.066	0.028	0.154	0.019	0.012	0.044	0.015	0.105	0.011	0.009
0.7	Mean	1.384	0.613	0.741	0.021	0.102	1.394	0.604	0.718	0.015	0.102
	SD	0.038	0.027	0.076	0.011	0.013	0.028	0.017	0.060	0.007	0.010

Table 13: Result of scenario 1 with Frank copula. C follows a Lognormal distribution with true parameters (1.4, 0.6), and T is simulated with a Lognormal distribution of parameter $(1.5, 1)$.

		$n = 250$					$n = 500$				
τ		$\hat{\mu}_C$	$\hat{\sigma}_C$	$\hat{\tau}$	$\hat{F}_T$	$\hat{F}_{T,\text{KM}}$	$\hat{\mu}_C$	$\hat{\sigma}_C$	$\hat{\tau}$	$\hat{F}_T$	$\hat{F}_{T,\text{KM}}$
0.2	Mean	1.177	0.640	0.722	0.095	0.032	1.201	0.634	0.659	0.087	0.026
	SD	0.065	0.034	0.138	0.023	0.015	0.057	0.023	0.134	0.019	0.012
0.5	Mean	1.283	0.646	0.775	0.064	0.058	1.302	0.639	0.727	0.056	0.055
	SD	0.068	0.042	0.129	0.021	0.022	0.060	0.033	0.130	0.019	0.016
0.7	Mean	1.346	0.641	0.834	0.047	0.083	1.359	0.631	0.799	0.039	0.080
	SD	0.063	0.051	0.105	0.017	0.024	0.054	0.040	0.103	0.016	0.017
		$n = 1000$					$n = 2000$				
0.2	Mean	1.236	0.625	0.569	0.076	0.024	1.252	0.620	0.525	0.071	0.023
	SD	0.041	0.018	0.093	0.013	0.009	0.024	0.011	0.050	0.008	0.007
0.5	Mean	1.325	0.628	0.673	0.047	0.054	1.347	0.616	0.615	0.040	0.054
	SD	0.049	0.026	0.110	0.015	0.011	0.040	0.018	0.087	0.011	0.008
0.7	Mean	1.368	0.623	0.777	0.031	0.080	1.381	0.612	0.744	0.024	0.079
	SD	0.038	0.029	0.076	0.012	0.012	0.033	0.022	0.068	0.011	0.009

Table 14: Result of scenario 1 with Gumbel copula. C follows a Lognormal distribution with true parameters (1.4, 0.6), and T is simulated with a Lognormal distribution of parameter $(1.5, 1)$.

B.2 Scenario 2

		$n = 250$					$n = 500$				
τ		$\hat{\mu}_C$	$\hat{\sigma}_C$	$\hat{\tau}$	$\hat{F}_T$	$\hat{F}_{T,\text{KM}}$	$\hat{\mu}_C$	$\hat{\sigma}_C$	$\hat{\tau}$	$\hat{F}_T$	$\hat{F}_{T,\text{KM}}$
0.2	Mean	-0.339	0.903	0.824	0.104	0.037	-0.278	0.922	0.715	0.090	0.034
	SD	0.128	0.053	0.216	0.029	0.017	0.172	0.060	0.298	0.040	0.013
0.5	Mean	-0.168	0.972	0.819	0.058	0.081	-0.157	0.975	0.800	0.054	0.081
	SD	0.115	0.053	0.180	0.022	0.022	0.105	0.045	0.181	0.021	0.016
0.7	Mean	-0.061	0.995	0.847	0.034	0.115	-0.060	0.994	0.838	0.029	0.114
	SD	0.106	0.055	0.143	0.015	0.023	0.071	0.038	0.113	0.013	0.017
		$n = 1000$					$n = 2000$				
0.2	Mean	-0.261	0.925	0.688	0.085	0.033	-0.173	0.949	0.522	0.061	0.032
	SD	0.179	0.055	0.322	0.044	0.010	0.192	0.061	0.351	0.048	0.007
0.5	Mean	-0.150	0.976	0.791	0.052	0.081	-0.126	0.978	0.738	0.045	0.082
	SD	0.101	0.031	0.181	0.021	0.011	0.106	0.030	0.201	0.022	0.008
0.7	Mean	-0.047	0.996	0.813	0.028	0.115	-0.044	0.993	0.798	0.020	0.116
	SD	0.086	0.031	0.147	0.016	0.011	0.057	0.021	0.108	0.012	0.008

Table 15: Result of scenario 2 with Frank copula. C follows a Lognormal distribution with true parameters (0, 1), and T is simulated with a Lognormal distribution of parameter $(0, 1)$.

		$n = 250$					$n = 500$				
τ		$\hat{\mu}_C$	$\hat{\sigma}_C$	$\hat{\tau}$	$\hat{F}_T$	$\hat{F}_{T,\text{KM}}$	$\hat{\mu}_C$	$\hat{\sigma}_C$	$\hat{\tau}$	$\hat{F}_T$	$\hat{F}_{T,\text{KM}}$
0.2	Mean	-0.315	0.903	0.755	0.103	0.035	-0.305	0.909	0.734	0.100	0.033
	SD	0.088	0.048	0.120	0.021	0.018	0.069	0.038	0.120	0.016	0.014
0.5	Mean	-0.152	0.953	0.770	0.057	0.080	-0.139	0.957	0.742	0.052	0.080
	SD	0.088	0.049	0.110	0.020	0.022	0.071	0.041	0.116	0.016	0.016
0.7	Mean	-0.032	0.990	0.772	0.030	0.115	-0.025	0.990	0.751	0.024	0.114
	SD	0.092	0.050	0.116	0.013	0.023	0.074	0.042	0.114	0.011	0.016
		$n = 1000$					$n = 2000$				
0.2	Mean	-0.286	0.916	0.697	0.096	0.033	-0.273	0.920	0.664	0.093	0.032
	SD	0.071	0.033	0.121	0.015	0.010	0.068	0.028	0.116	0.013	0.007
0.5	Mean	-0.131	0.960	0.729	0.050	0.081	-0.123	0.963	0.709	0.048	0.082
	SD	0.072	0.034	0.122	0.014	0.012	0.067	0.027	0.110	0.013	0.008
0.7	Mean	-0.016	0.994	0.738	0.020	0.115	-0.012	0.996	0.725	0.018	0.116
	SD	0.070	0.033	0.114	0.010	0.012	0.070	0.026	0.112	0.010	0.008

Table 16: Result of scenario 2 with Gumbel copula. C follows a Lognormal distribution with true parameters (0, 1), and T is simulated with a Lognormal distribution of parameter $(0, 1)$.

B.3 Scenario 3

		$n = 250$					$n = 500$				
τ		$\hat{\mu}_C$	$\hat{\sigma}_C$	$\hat{\tau}$	$\hat{F}_T$	$\hat{F}_{T,\text{KM}}$	$\hat{\mu}_C$	$\hat{\sigma}_C$	$\hat{\tau}$	$\hat{F}_T$	$\hat{F}_{T,\text{KM}}$
0.2	Mean	-0.201	0.959	0.483	0.077	0.037	-0.135	0.963	0.394	0.062	0.034
	SD	0.198	0.050	0.453	0.042	0.017	0.141	0.043	0.398	0.035	0.013
0.5	Mean	-0.160	0.980	0.755	0.061	0.081	-0.092	0.984	0.659	0.043	0.081
	SD	0.112	0.048	0.296	0.022	0.022	0.081	0.038	0.230	0.021	0.016
0.7	Mean	-0.062	0.998	0.815	0.037	0.115	-0.005	0.998	0.736	0.025	0.114
	SD	0.098	0.051	0.214	0.022	0.023	0.062	0.033	0.173	0.020	0.017
		$n = 1000$					$n = 2000$				
0.2	Mean	-0.155	0.962	0.388	0.064	0.033	-0.093	0.974	0.237	0.049	0.032
	SD	0.179	0.047	0.418	0.040	0.010	0.136	0.038	0.340	0.027	0.007
0.5	Mean	-0.149	0.977	0.771	0.053	0.081	-0.096	0.980	0.661	0.034	0.082
	SD	0.099	0.029	0.219	0.021	0.011	0.079	0.028	0.153	0.017	0.008
0.7	Mean	-0.046	0.996	0.814	0.028	0.115	-0.012	0.996	0.731	0.013	0.116
	SD	0.086	0.031	0.151	0.016	0.011	0.051	0.023	0.082	0.009	0.008

Table 17: Result of scenario 3 with Frank copula. C follows a Normal distribution with true parameters (0, 1), and T is simulated with a Normal distribution of parameter $(0, 1)$.

B.4 Scenario 4

		$n = 250$					$n = 500$				
τ		$\hat{\mu}_C$	$\hat{\sigma}_C$	$\hat{\tau}$	$\hat{F}_T$	$\hat{F}_{T,\mathrm{KM}}$	$\hat{\mu}_C$	$\hat{\sigma}_C$	$\hat{\tau}$	$\hat{F}_T$	$\hat{F}_{T,\mathrm{KM}}$
0.2	Mean	0.973	1.288	0.463	0.055	0.036	0.990	1.425	0.299	0.033	0.032
	SD	0.074	0.286	0.320	0.036	0.018	0.047	0.234	0.221	0.024	0.014
0.5	Mean	0.946	1.345	0.702	0.048	0.076	0.971	1.418	0.606	0.034	0.075
	SD	0.081	0.227	0.243	0.022	0.024	0.060	0.196	0.189	0.021	0.017
0.7	Mean	0.954	1.433	0.817	0.036	0.106	0.967	1.450	0.771	0.026	0.105
	SD	0.078	0.157	0.159	0.013	0.027	0.057	0.114	0.111	0.011	0.019
		$n = 1000$					$n = 2000$				
0.2	Mean	0.998	1.460	0.250	0.020	0.031	1.000	1.489	0.209	0.013	0.030
	SD	0.032	0.141	0.124	0.010	0.010	0.022	0.102	0.087	0.007	0.007
0.5	Mean	0.987	1.456	0.554	0.021	0.076	0.995	1.473	0.528	0.014	0.076
	SD	0.042	0.129	0.124	0.014	0.012	0.029	0.091	0.086	0.009	0.009
0.7	Mean	0.979	1.462	0.747	0.020	0.106	0.990	1.476	0.727	0.014	0.106
	SD	0.047	0.100	0.088	0.010	0.013	0.033	0.075	0.069	0.008	0.009

Table 18: Result of scenario 4 with Frank copula. C follows a Weibull distribution with true parameters (1, 1.5), and T is simulated with a lognormal distribution of parameter $(0, 1.7)$.

C Code

Functions

```
CG_Frank <- function (Y, delta, theta, S.plot = TRUE, S.col = "black")
{
  if (abs(theta) < 1e-06) {theta = 1e-06}
  n = length(Y)
  R = n:1
  t.sort = sort(Y)
  d.sort = delta[order(Y)]
  A = log((exp(-theta * (R - 1)/n) - 1)/(exp(-theta * R/n) - 1))
  S = -1/theta * log(1 + (exp(-theta) - 1) * exp(cumsum(A * d.sort)))
  if (d.sort[n] == 0){S[n] = S[n-1]}
  if (d.sort[n] == 1){S[n] = 0}
  if (S.plot == TRUE) {
    plot(c(0, t.sort), c(1, S), type = "s", lwd = 3, ylim = c(0,1), ylab =
    "Survival probability", xlab = "Time", col = S.col)
    points(t.sort[d.sort == 0], S[d.sort == 0], pch = 3, cex = 2, col = S.
   col)
  }
  func1 = function(x) {
    x/(exp(x) - 1)
  }
  tau = 1 - 4/theta * (1 - integrate(func1, 0, theta)$value/theta)

  list(tau = tau, Y.sort = t.sort, delta.sort = d.sort, n.risk = R, surv =
    S)
}

dist_T <- function(Y, delta, theta, fam){
  if(fam=="frank"){
    S_hat = CG_Frank(Y, delta, theta, S.plot = FALSE)$surv
  }
  if(fam=="gumbel"){
    S_hat = CG_Gumbel(Y, delta, theta, S.plot = FALSE)$surv
  }
  FT = pmax(0, 1 - S_hat)
```

```
  fT = c(1-S_hat[1], -diff(S_hat))
  list(fT = fT, FT = FT, S_hat = S_hat)
}

CG_Gumbel <- function (Y, delta, theta, S.plot = TRUE, S.col = "black")
{
  alpha = max(alpha, 0)
  n = length(t.vec)
  R = n:1
  t.sort = sort(t.vec)
  d.sort = d.vec[order(t.vec)]
  A = (-log((R - 1)/n))^(alpha + 1) - (-log(R/n))^(alpha +1)
  S = exp(-(cumsum(A * d.sort))^(1/(1 + alpha)))
  if (d.sort[n] == 0){S[n] = S[n-1]}
  if (d.sort[n] == 1){S[n] = 0}
  if (S.plot == TRUE) {
    plot(c(0, t.sort), c(1, S), type = "s", lwd = 3, ylim = c(0,
                                                              1), ylab = "
   Survival probability", xlab = "Time",
         col = S.col)
    points(t.sort[d.sort == 0], S[d.sort == 0], pch = 3,
           cex = 2, col = S.col)
  }
  list(tau = (alpha/(alpha + 1)), time = t.sort, n.risk = R,
       surv = S)
}

likelihood <- function(param = c(parC1, log.parC2, z.tau), Y, delta, distT
   ){
  par.tau = tanh(param[3])
  theta = sign(par.tau) * as.vector(ktau_to_par("frank", max(1e-20, abs(
   par.tau))))
  par.C = c(param[1], exp(param[2]))

  frank_cop = bicop_dist("frank",rot, theta)
  FT = distT$FT
  fT = distT$fT
  if(dist.C=="lnorm"|dist.C=="weibull_false") {
```

```
    FC = plnorm(Y,par.C[1],par.C[2])
    fC = dlnorm(Y,par.C[1],par.C[2])
  }
  if(dist.C=="norm") {
    FC = pnorm(Y,par.C[1],par.C[2])
    fC = dnorm(Y,par.C[1],par.C[2])
  }
  if(dist.C=="weibull"){
    FC = pweibull(Y,par.C[1],par.C[2])
    fC = dweibull(Y,par.C[1],par.C[2])
  }

  hC.condT = hbicop(cbind(1-FT,1-FC), 1,frank_cop)
  hT.condC = hbicop(cbind(1-FT,1-FC), 2,frank_cop)

  l1 <- sum(log(fT[delta==1])+log(hC.condT[delta==1]))
  l0 <- sum(log(fC[delta==0])+log(hT.condC[delta==0]))
  l1 + l0
}

sim <- function(i,n,par.tau,fam,rot,dist.T,dist.C,par.T,par.C,k1,k2,theta)
    {
  if (i/5==ceiling(i/5)) print(i)
  set.seed(seed+i)

  u.TC= rbicop(n,fam,rot,theta)
  if(dist.T=="lnorm") xT = qlnorm(1-u.TC[,1],par.T[1],par.T[2])
  if(dist.C=="lnorm") xC = qlnorm(1-u.TC[,2],par.C[1],par.C[2])
  if(dist.T=="norm") xT = qnorm(1-u.TC[,1],par.T[1],par.T[2])
  if(dist.C=="norm") xC = qnorm(1-u.TC[,2],par.C[1],par.C[2])
  if(dist.T=="lnorm") xT = qlnorm(1-u.TC[,1],par.T[1],par.T[2])
  if(dist.C=="weibull"|dist.C=="weibull_false") xC = qweibull(1-u.TC[,2],
    shape = par.C[1], scale = par.C[2])

  delta = as.numeric(xT<=xC)
  Y = pmin(xT,xC)
  delta = delta[order(Y)]
  Y = sort(Y)
```

```
103
104 if(dist.C=="lnorm"|dist.C=="weibull_false") {
105   mle_paraC = survreg(Surv(Y,1-delta)~1,dist="lognormal")
106   p = c(mle_paraC$icoef[1], exp(mle_paraC$icoef[2]))
107 }
108 if(dist.C=="norm") {
109   mle_paraC = survreg(Surv(Y,1-delta)~1,dist="gaussian")
110   p = c(mle_paraC$icoef[1], mle_paraC$icoef[2])
111 }
112 if(dist.C=="weibull"){
113   mle_paraC = survreg(Surv(Y,1-delta)~1,dist="weibull")
114   p = c(mle_paraC$icoef[1], exp(mle_paraC$icoef[2]))
115 }
116
117 pm = matrix(p, k1, 2, byrow=TRUE)
118 un = c(1,runif(k1-1,0.5,1.5))
119 X0 = pm*un
120 X0 = matrix(t(X0),k1*k2,2,byrow=TRUE)
121 t = seq(0.5/k2,1-0.5/k2,1/k2) #+ runif(k2,-0.5/k2,0.5/k2)
122 th = NULL
123 for (k in 1:k2) th = c(th,ktau_to_par("frank", t[k]))
124 X0 = cbind(X0,rep(t,each=k1), rep(th, each=k1))
125 X0[,2] = log(X0[,2])
126 X0[,3] = atanh(X0[,3])
127
128 KK = k1*k2
129 m = 1
130 result = matrix(0,m*KK,4)
131
132 for (j in 1:KK)
133 {
134   dist.T_temp = dist_T(Y,delta,X0[j,4], fam)
135   test = likelihood(X0[j,1:3],Y,delta,dist.T_temp)
136   if ((is.na(test)==TRUE) | (test==Inf) | (test==-Inf)) result[(m*(j-1)
  +1):(m*j),1] = -Inf
137   else
138   {
139     result1 = try(optim(X0[j,1:3],likelihood,Y=Y,delta=delta,distT=dist.
```

```
    T_temp, control=list(trace=0,fnscale=-1,maxit=700),method="BFGS"))
140       if ((class(result1)[1]=="try-error") )  result[m*(j-1)+1,1] = -Inf
    else result[m*(j-1)+1,] = c(result1$value,result1$par)
141     }
142   }
143
144   Lik = result[,1]
145   hatParams = result[,2:4]
146
147   Tri = order(Lik)[length(Lik)]
148   ppar = hatParams[Tri,]
149   ttheta = sign(tanh(ppar[3])) * as.numeric(ktau_to_par("frank", max(1e
     -20, abs(tanh(ppar[3])))))
150   distT = dist_T(Y,delta,ttheta,fam)$FT
151   KM = 1-survfit(Surv(Y,delta)~1)$surv
152   if(dist.T=="norm") {
153     bias = mean(abs(distT-pnorm(Y,par.T[1],par.T[2])))
154     biasKM = mean(abs(KM-pnorm(Y,par.T[1],par.T[2])))
155   }
156   if(dist.T=="lnorm") {
157     bias = mean(abs(distT-plnorm(Y,par.T[1],par.T[2])))
158     biasKM = mean(abs(KM-plnorm(Y,par.T[1],par.T[2])))
159   }
160   lines(Y,distT,col=8)
161   try(lines(Y,KM,col=4))
162   legend("bottomright",
163          legend=c("distT", "KM"),
164          fill=c(8, 4))
165   return(c(ppar[1],exp(ppar[2]),tanh(ppar[3]),bias,biasKM))
166 }
```

simulation

```
167 require(rvinecopulib)
168 require(rafalib)
169 library(compound.Cox)
170 library(plyr)
```

```r
library(doParallel)
library(EnvStats)
library(foreach)
library(doParallel)
library(stats)

no_cores=detectCores()-2

setwd("C:/Users/Laura/Documents/Ucl/M2/memoire/code/CODE OK")
source("functions.R")
seed = 123

par.tau = 0.7
fam = "frank"
rot = 0

##case1##
#dist.T = "lnorm"
#dist.C = "lnorm"
#par.T = c(1.5,1)
#par.C = c(1.4,0.6)

##case2##
#dist.T = "lnorm"
#dist.C = "lnorm"
#par.T = c(0,1)
#par.C = c(0,1)

##case3##
#dist.T = "norm"
#dist.C = "norm"
#par.T = c(0,1)
#par.C = c(0,1)

##case4##
#dist.T = "lnorm"
#dist.C = "weibull"
#par.T = c(0,1.7)
```

```
#par.C = c(1,1.5)

##case5##
#dist.T = "lnorm"
#dist.C = "weibull_false"
#par.T = c(1.5,1)
#par.C = c(1,1)

nn = c(250,500,1000,2000)
nIter = 200
k1 = 5
k2 = 10
parallel = FALSE

theta = sign(par.tau) * as.vector(ktau_to_par("frank", max(1e-20, abs(par.
    tau))))

if(dist.T=="norm") {
  yy = sort(rnorm(10000,par.T[1],par.T[2]))
  pyy = pnorm(yy,par.T[1],par.T[2])
  ymin = qnorm(0.02,par.T[1],par.T[2])
  ymax = qnorm(0.98,par.T[1],par.T[2])
}
if(dist.T=="lnorm") {
  yy = sort(rlnorm(10000,par.T[1],par.T[2]))
  pyy = plnorm(yy,par.T[1],par.T[2])
  ymin = qlnorm(0.02,par.T[1],par.T[2])
  ymax = qlnorm(0.98,par.T[1],par.T[2])
}

for (l in 1:length(nn)) {

start.time = Sys.time()
n = nn[l]

plot(yy,pyy,xlim=c(ymin,ymax),cex=0.5,col=2,xlab="t",ylab="FT(t)",main=
    paste("n =",n))

```

```
if (parallel==TRUE){
  registerDoParallel(cores=no_cores)
  res = foreach(i=1:nIter,.combine='rbind') %dopar% {sim(i,n,par.tau,fam,
   rot,dist.T,dist.C,par.T,par.C,k1,k2,theta)}
}
else {
  res = matrix(0,nIter,5)
  for (i in 1:nIter) res[i,] = sim(i,n,par.tau,fam,rot,dist.T,dist.C,par.T
   ,par.C,k1,k2,theta)
}

lines(yy,pyy,xlim=c(ymin,ymax),lwd=4,col=2)

mean_estim = apply(res,2,mean)
sd_estim = apply(res,2,sd)

print(c("n =",n),quote=FALSE)
par.mat = rbind(c(par.C,par.tau,0,0),mean_estim,sd_estim)
rownames(par.mat) = c("true","mean(estimate)","sd(estimate)")
colnames(par.mat) = c("E(C)","Var(C)","tau","Bias distT","Bias KM")
print(round(par.mat,digits=7))

end.time = Sys.time()
time.taken = end.time - start.time
print(time.taken)
start.time = end.time
}
```

References

Basu, A. and Ghosh, J. (1978). Identifiability of the multinormal and other distributions under competing risks model. *Journal of Multivariate Analysis*, 8(3):413–429.

Braekers, R. and Veraverbeke, N. (2005). A copula-graphic estimator for the conditional survival function under dependent censoring. *Canadian Journal of Statistics*, 33(3):429–447.

Camilleri, L. (2019). History of survival analysis. *The Sunday Times of Malta*, page 53.

Chen, H.-Y., Yu, S.-L., Chen, C.-H., Chang, G.-C., Chen, C.-Y., Yuan, A., Cheng, C.-L., Wang, C.-H., Terng, H.-J., Kao, S.-F., et al. (2007). A five-gene signature and clinical outcome in non–small-cell lung cancer. *New England Journal of Medicine*, 356(1):11–20.

Czado, C. (2019). *Analyzing dependent data with vine copulas*, volume 222. Springer.

Czado, C. and Van Keilegom, I. (2023). Dependent censoring based on parametric copulas. *Biometrika*, 110(3):721–738.

Deheuvels, P. (1979). La fonction de dépendance empirique et ses propriétés. un test non paramétrique d'indépendance. *Bulletins de l'Académie Royale de Belgique*, 65(1):274–292.

Delhelle, M. and Van Keilegom, I. (2024). Copula based dependent censoring in cure models. *(under revision)*.

Deresa, N. W. and Van Keilegom, I. (2020). Flexible parametric model for survival data subject to dependent censoring. *Biometrical Journal*, 62(1):136–156.

Deresa, N. W. and Van Keilegom, I. (2021). On semiparametric modelling, estimation and inference for survival data subject to dependent censoring. *Biometrika*, 108(4):965–979.

Deresa, N. W. and Van Keilegom, I. (2023). Copula based Cox proportional hazards models for dependent censoring. *Journal of the American Statistical Association*, pages 1–11.

Deresa, N. W. and Van Keilegom, I. (2024). Semiparametric transformation models for survival data with dependent censoring. *(Submitted)*.

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

Emoto, S. E. and Matthews, P. C. (1990). A Weibull model for dependent censoring. *The Annals of Statistics*, 18(4):1556–1577.

Emura, T. and Chen, Y.-H. (2016). Gene selection for survival data under dependent censoring: A copula-based approach. *Statistical Methods in Medical Research*, 25(6):2840–2857.

Emura, T. and Chen, Y.-H. (2018). *Analysis of survival data with dependent censoring: copula-based approaches*, volume 450. Springer.

Escarela, G. and Carriere, J. F. (2003). Fitting competing risks with an assumed copula. *Statistical Methods in Medical Research*, 12(4):333–349.

Fan, T.-H., Wang, Y.-F., and Ju, S.-K. (2019). A competing risks model with multiply censored reliability data under multivariate weibull distributions. *IEEE Transactions on Reliability*, 68(2):462–475.

Genest, C. and Favre, A.-C. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering*, 12(4):347–368.

Härdle, W. K. and Simar, L. (2019). *Applied Multivariate Statistical analysis*. Springer Nature.

Huang, X. and Zhang, N. (2008). Regression survival analysis with an assumed copula for dependent censoring: a sensitivity analysis approach. *Biometrics*, 64(4):1090–1099.

Jackson, D., White, I. R., Seaman, S., Evans, H., Baisley, K., and Carpenter, J. (2014). Relaxing the independent censoring assumption in the Cox proportional hazards model using multiple imputation. *Statistics in Medicine*, 33(27):4681–4694.

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

Klein, J. P. and Moeschberger, M. (1988). Bounds on net survival probabilities for dependent competing risks. *Biometrics*, pages 529–538.

Kleinbaum, D. G. and Klein, M. (2012). *Survival Analysis: A Self-Learning Text*. Statistics for Biology and Health. Springer New York, New York, NY.

Nelsen, R. B. (2006). Archimedean copulas. *An introduction to copulas*, pages 109–155.

Rivest, L.-P. and Wells, M. T. (2001). A Martingale Approach to the Copula-Graphic Estimator for the Survival Function under Dependent Censoring. *Journal of Multivariate Analysis*, 79(1):138–155.

Roth, M. (2012). *On the multivariate t distribution*. Linköping University Electronic Press.

Scharfstein, D. O. and Robins, J. M. (2002). Estimation of the failure time distribution in the presence of informative censoring. *Biometrika*, 89(3):617–634.

Schmidt, T. (2006). Coping with copulas in "copulas: From theory to applications in finance". *Bloomberg Financial*.

Shih, J.-H., Lee, W., Sun, L.-H., and Emura, T. (2019). Fitting competing risks data to bivariate pareto models. *Communications in Statistics-Theory and Methods*, 48(5):1193–1220.

Sklar, M. (1959). Fonctions de répartition à n dimensions et leurs marges. In *Annales de l'ISUP*, volume 8, pages 229–231.

Slud, E. V. and Rubinstein, L. V. (1983). Dependent competing risks and summary survival curves. *Biometrika*, 70(3):643–649.

Una-Álvarez, J. and Veraverbeke, N. (2013). Generalized copula-graphic estimator. *Test*, 22:343–360.

Van Keilegom, I. (2022-2023). Analyse des données de survie et de durée. *LSTAT2220*.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the econometric society*, pages 1–25.

Zheng, M. and Klein, J. P. (1995). Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika*, 82(1):127–138.

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
Faculté des sciences
Place des sciences, 2 bte L6.06.01, 1348 Louvain-la-Neuve, Belgique | **www.uclouvain.be/sc**