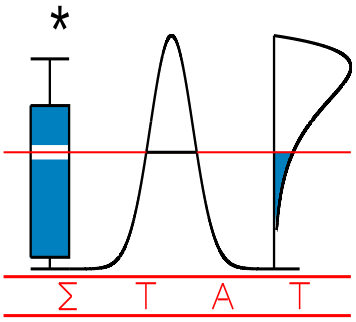
T E C H N I C A L

R E P O R T

11024

Likelihood based inference for semi-competing risks

HEUCHENNE, C., LAURENT, S., LEGRAND, C. and I. VAN KEILEGOM



I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

Likelihood based inference for semi-competing risks

Cédric Heuchenne * Stéphane Laurent † Catherine Legrand ‡
Ingrid Van Keilegom §

September 13, 2011

Abstract

Consider semi-competing risks data (two times to concurrent events are studied but only one of them is right-censored by the other one) where the link between the times Y and C to non-terminal and terminal events respectively, is modeled by a family of Archimedean copulas. Moreover, both Y and C are submitted to an independent right censoring variable D. A new methodology based on a maximum likelihood approach is developed to estimate the parameter of the copula and the resulting survival function of Y. The main advantage of this procedure is that it extends to multidimensional parameters copulas. We perform simulations to study the behavior of our proposed estimation procedure and its impact on other related estimators and we apply our method to real data coming from a study on the Hodgkin disease.

Key words: Bootstrap; Copulas; Dependent censoring; Kaplan-Meier estimator; Maximum Likelihood; Nonparametric estimation; Semi-competing risks; Survival analysis.

*QuantOM, HEC-Management School of University of Liège, Rue Louvrex 14, B-4000 Liège, Belgium and Institute of statistics, biostatistics and actuarial sciences, Université catholique de Louvain, E-mail: C.Heuchenne@ulg.ac.be

†QuantOM, HEC-Management School of University of Liège, Rue Louvrex 14, B-4000 Liège, Belgium, E-mail: slaurent@ulg.ac.be

‡Institute of statistics, biostatistics and actuarial sciences, Université catholique de Louvain, Voie du Roman Pays 20, B-1348 Louvain-la-Neuve, Belgium, E-mail: catherine.legrand@uclouvain.be

§Institute of statistics, biostatistics and actuarial sciences, Université catholique de Louvain, Voie du Roman Pays 20, B-1348 Louvain-la-Neuve, Belgium, E-mail: ingrid.vankeilegom@uclouvain.be

1 Introduction

A patient under medical follow-up can experience a terminal event (e.g. death) and/or a non-terminal event (e.g. occurence of some toxicity, recurrence of the disease). The time Y to the non-terminal event can be right-censored by the time C to the terminal event but not vice-versa. Furthermore, the random variable D denoting the time to the end of the study or the lost to follow-up also acts as a time to a terminal event since both Y and C are censored as soon as they are larger than D. In a modeling perspective, it is realistic to assume that D is independent of both Y and C but dependency between Y and C cannot be ignored in many cases. A convenient way to take it into account is to specify a parametric family of copulas for the joint law of Y and C.

The survival functions of C and D can be consistently estimated by Kaplan-Meier estimators, but estimating both the copula parameter and the survival function of Y seems more problematic. Fine *& al.* (2001) proposed a copula model fitting procedure assuming the parametric Clayton family of copulas. Their method consists in plugging an estimate $\hat{\theta}^{\text{Fine}}$ of the unknown parameter θ of the copula into a θ-dependent estimate of the survival function $S(\cdot)$ of Y. This has been criticized by Jiang *& al.* (2005) who proposed another Clayton copula based estimator of $S(\cdot)$ having better theoretical properties and seemingly reaching better performance in practice. Lakhal (2006) and Lakhal *& al.* (2008) generalized $\hat{\theta}^{\text{Fine}}$ to one-dimensional parametric families of Archimedean copulas and proposed a more attractive θ-dependent estimator $\hat{S}_\theta(\cdot)$ of $S(\cdot)$, defined similarly to the Rivest and Wells (2001) estimator in the simpler context of censored data. This estimator $\hat{S}_\theta(\cdot)$ has been studied in detail by Laurent (2011). Recently, Xu *& al.* (2010) developed another modeling aproach for this kind of data but not allowing to estimate $S(\cdot)$.

Our purpose is to propose a new estimating procedure for both $S(\cdot)$ and the copula parameter, allowing for multidimensional parameter families of copulas. In the one-parameter case, it can be compared to results obtained by the Lakhal *& al.* (2008) method. The need of multidimensional parameters copulas is especially important in this context since possibly complex dependency structures (already difficult to capture with one-dimensional parameter copulas and complete data) between Y and C can still be less easily detected due the "hiding" effects of the censoring mechanisms.

This paper is organized as follows. In the next section, we define our model and we prove its identifiability in the one-dimensional case. The estimation method is explained in Section 3. We study the performance of the resulting estimators by means of simulations in Section 4, and we apply our method to the Hodgkin disease data in Section 5. Finally, Section 6 gives some final recommendations for the application of the proposed methodology in practice and the Appendix contains some technical details.

2 Model

The model is defined as follows. For a given individual, let Y be the time to the event of interest (e.g., time until the first relapse of a certain disease), C the censoring time of the first type (e.g., the time until death), and D the censoring time of the second type (e.g., the time until lost to follow-up). We assume that D is independent of (Y, C), the marginal distributions of Y, C and D are unknown, and the survival copula for (Y, C) belongs to a parametric family $\{\mathcal{C}_\theta : \theta \in \Theta\}$ of copulas, where $\Theta \subset \mathbb{R}^d$, $d \geq 1$. We denote the unknown true parameter by θ_0 and the marginal survival functions of Y, C and D by $S(\cdot) = 1 - F(\cdot)$, $G_1(\cdot)$ and $G_2(\cdot)$ respectively.

The observable random variables are n independent replications $(T_i, Z_i, \Delta_{1i}, \Delta_{2i})$, $i = 1, \ldots, n$, of $(T, Z, \Delta_1, \Delta_2)$, where $T = \min(Y, C, D)$, $Z = \min(C, D)$ and the indicator variables Δ_1 and Δ_2 are defined by

$$\begin{cases} \Delta_1 = \mathbb{1}_{Y \leq C, Y \leq D} & \text{(no censoring)} \\ \Delta_2 = \mathbb{1}_{C \leq D} & \text{(first type censoring before second type censoring).} \end{cases}$$

Note that $\{\Delta_1 = 1, \Delta_2 = 1\}$ corresponds to the event for which we observe both Y and C. Moreover, we observe both Y and $\min(C, D)$ when $\Delta_1 = 1$, and only $\min(C, D)$ when $\Delta_1 = 0$. Finally, note that $T = \min(Y, C)$ corresponds to $\Delta_3 = 1$, where

$$\Delta_3 = \mathbb{1}_{\min(Y,C) \leq D} = \min(1, \Delta_1 + \Delta_2).$$

In the data example coming from a study on the Hodgkin disease (see Section 5 for more details), Y is the time to first relapse, C is the time to death and D is the time to lost to follow-up. Hence, it would be unrealistic to assume that Y and C are independent, whereas it is reasonable to assume no relation between (Y, C) and D. Hence, the above model can be reasonable for these data, but the main difficulty lies in the choice of the family of copulas $\{\mathcal{C}_\theta\}$. Of course the random variable Y is hypothetical since it has no physical interpretation for an individual who died before experiencing relapse. We refer to Prentice & *al* (1978) for a long discussion about latent failure times in survival modeling.

Our estimation method presented in the next section only concerns the case when $\{\mathcal{C}_\theta\}$ is a family of (strict) Archimedean copulas. A copula $\mathcal{C}$ is said to be (strictly) Archimedean when

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

for all $0 \leq u, v \leq 1$, where the (strict) generator $\varphi : (0, 1] \to I\!R^+$ is a decreasing convex function satisfying $\varphi(1) = 0$ and $\varphi(0^+) = +\infty$. In particular $\varphi(\cdot) = -\log(\cdot)$ corresponds to the independence case. Choosing a parametric family $\{\varphi_\theta : \theta \in \Theta\}$ of such generators yields a parametric family $\{\mathcal{C}_\theta : \theta \in \Theta\}$ of Archimedean copulas by defining $\mathcal{C}_\theta$ with φ_θ.

We will use the following common families of Archimedean copulas :

- the one-parameter *Clayton family* given by $\varphi_\theta(x) = \frac{x^{-\theta}-1}{\theta}$ with parameter $\theta \geq 0$ and the particular case $\varphi_0(x) = -\log(x)$,
- the one-parameter *Frank family* given by $\varphi_\theta(x) = -\log[\frac{\exp(-\theta x)-1}{\exp(-\theta)-1}]$ with parameter $\theta \in \mathbb{R}$ and the particular case $\varphi_0(x) = -\log(x)$,
- the one-parameter *Gumbel family* given by $\varphi_\theta(x) = (-\log x)^\theta$ with parameter $\theta \geq 1$,
- the *interior power Frank family* (hereafter named the two-parameter Frank family) with two-dimensional parameter $\theta = (\alpha, \beta)$, given by $\varphi_\theta(x) = \varphi_\alpha(x^\beta)$ with $\alpha \in \mathbb{R}$ and $\beta > 1$, where φ_α is the generator of the Frank family.

The Kendall rank correlation coefficient τ of the Archimedean copula with generator φ is given by $\tau = 1 + 4\int_0^1 \frac{\varphi(x)}{\varphi'(x)}dx$, where $\varphi'(x)$ denotes the derivative of $\varphi(x)$ with respect to x. Its range is $[0, 1[$ for the Clayton and Gumbel families and $]-1, 1[$ for the Frank and the two-parameter Frank families. For the one-parameter families, there is a one-to-one correspondence between the parameter θ and the Kendall's tau on the ranges of θ and τ specified above, whereas there is an infinity of values of α and β corresponding to a unique value of τ for the two-parameter Frank family (see Figure 1).

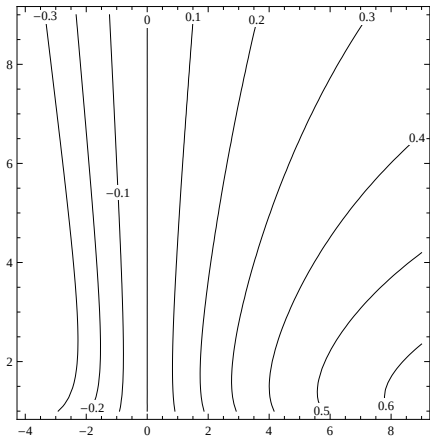


Figure 1: Kendall's tau of the two-parameter Frank family. Each curve corresponds to a value of the Kendall's tau, axes correspond to α and β.

The proof of the following theorem about the identifiability of the model can be found in the Appendix. Assumptions on the copula family are fulfilled for the Clayton, Gumbel and Frank families.

Theorem 2.1. *Assume that S, G_1 and G_2 are absolutely continuous and that $\{\mathcal{C}_\theta : \theta \in \Theta\}$ is a one-parameter family of Archimedean copulas such that $\varphi'_{\theta_1}/\varphi'_{\theta_2}$ is strictly monotone whenever $\theta_1 \neq \theta_2$. Then the model is identifiable; more precisely, if $(T, Z, \Delta_1, \Delta_2)$ has*

the same distribution as $(T', Z', \Delta_1', \Delta_2')$*, then the corresponding underlying parameters* (θ, S, G_1, G_2) *and* $(\theta', S', G_1', G_2')$ *defining the distributions of* (Y, C, D) *and* (Y', C', D') *are equal.*

3 Estimation

In this section, we propose appropriate estimators for the unknown parameters of the model θ_0, $S(\cdot)$, $G_1(\cdot)$ and $G_2(\cdot)$. Note that since we observe $Z = \min(C, D)$ and $\Delta_2 = I(C \leq D)$, and we assume that C and D are independent, we can use Kaplan-Meier estimators $\widehat{G}_1(\cdot)$ and $\widehat{G}_2(\cdot)$ of $G_1(\cdot)$ and $G_2(\cdot)$. The other part of the procedure (estimation of θ_0 and $S(\cdot)$) is then presented as follows. In a first step, the estimator $\widehat{S}_\theta(\cdot)$ of $S(\cdot)$ as a function of θ is displayed and discussed. Next, an existing estimator $\hat{\theta}_1$ of θ_0 is given and finally, a new estimator $\hat{\theta}_2$ of θ_0 is developed. The practical behavior of $\hat{\theta}_1$ and $\hat{\theta}_2$ as well as $\widehat{S}_1(\cdot) = \widehat{S}_{\hat{\theta}_1}(\cdot)$ and $\widehat{S}_2(\cdot) = \widehat{S}_{\hat{\theta}_2}(\cdot)$ are studied in Section 4.

Remark 3.1 Asymptotic properties of our estimators can be established by applying theorems for semi-parametric estimation of Chen, Linton and Van Keilegom (2003). Since the proofs are rather long and technical and in order to primarily focus on the practical behavior of our estimators, these asymptotic properties will be developed in a technical report.

3.1 Estimating S when θ is know

We introduce the θ-dependent estimator $\widehat{S}_\theta(\cdot)$ of $S(\cdot)$ as in Lakhal *& al.* (2008) and Laurent (2011) :

$$\hat{S}_\theta(t) = \varphi_\theta^{-1}\left[-\sum_{T_i \leq t, \Delta_{1i}=1} \left\{\varphi_\theta\left(\hat{\Gamma}(T_i-)\right) - \varphi_\theta\left(\hat{\Gamma}(T_i)\right)\right\}\right], \tag{1}$$

where $\hat{\Gamma}$ is the Kaplan-Meier estimator of the survival function $\Gamma(t) = \mathbb{P}(\min(Y, C) > t)$ of $\min(Y, C)$, which is available from the observations (T_i, Δ_{3i}), $i = 1, \ldots, n$.

The estimator $\hat{S}_\theta(\cdot)$ is a direct extension of the Rivest & Wells (2001) estimator to our context of right-censored semi-competing risks (the only difference is that $\hat{\Gamma}(\cdot)$ is the Kaplan-Meier estimator of $\Gamma(\cdot)$ and not the empirical survival function of $Y \wedge C$). Laurent (2011) studied its asymptotic behaviour when assuming an arbitrary continuous distribution for (Y, C). The assumptions on $\varphi_\theta(\cdot)$ given by Laurent (2011) are fulfilled for the Clayton, the Frank, and the two-parameter Frank families, but not for the Gumbel fam-

ily. However simulations show that inference seems to also be valid for the Gumbel family.

Remark 3.2 Laurent (2011) indicated a possible way to include covariates in the model and the inference procedure when θ is known: similarly to Braekers & Veraverbeke (2005) in the context of censored data, he proposes to include covariates X_i, $i = 1, \ldots, n$, $X_i \in \mathbb{R}^p$, $p \geq 1$, in the estimation of the survival function Γ of $\min(Y, C)$. Then, $\hat{S}_\theta(\cdot|X = x)$ could be defined similarly to (1) by replacing $\hat{\Gamma}(\cdot)$ by $\hat{\Gamma}(\cdot|X = x)$, an estimator of $\Gamma(\cdot|X = x) = P(\min(Y, C) > \cdot|X = x)$. For example, when $p = 1$, the estimate $\hat{\Gamma}(\cdot|X = x)$ can be obtained nonparametrically by Beran (1981) who defines (in the case of no ties) :

$$\hat{\Gamma}(y|x) \quad = \quad \prod_{T_i \leq y, \Delta_{3i}=1} \Big\{1 - \frac{W_i(x, a_n)}{\sum_{j=1}^n I(T_j \geq T_i) W_j(x, a_n)}\Big\}, \tag{2}$$

where

$$W_i(x, a_n) = \frac{K\left(\frac{x - X_i}{a_n}\right)}{\sum_{j=1}^n K\left(\frac{x - X_j}{a_n}\right)},$$

K is a kernel function and $\{a_n\}$ a bandwidth sequence. Obviously, they are different possible ways to take covariates into account. For example, Peng and Fine (2007) and Ding *& al.* (2009) postulated separate marginal regression models for both Y and C. Another direction for further research in this context is to consider estimation of the copula parameter for each value of the covariate (local copula).

3.2 Estimating θ_0 : existing methodology

When considering a one-parameter family, Lakhal *& al.* (2008) proposed an estimator $\hat{\theta}_1$ of the copula parameter θ_0 defined as the solution of $\hat{g}_n(\theta) = 0$ where the function $\hat{g}_n(\cdot)$ is derived as follows.

Consider an independent copy (Y', C', D') of (Y, C, D) and the corresponding four-tuple $(T', Z', \Delta_1', \Delta_2')$. The two four-tuples $(T, Z, \Delta_1, \Delta_2)$ and $(T', Z', \Delta_1', \Delta_2')$ represent the observable data for two individuals. Let $\tilde{Y} = \min(Y, Y')$, $\tilde{C} = \min(C, C')$, $\tilde{D} = \min(D, D')$, $\tilde{T} = \min(T, T')$ and $\tilde{Z} = \min(Z, Z')$. Then the event $A := \{\tilde{Y} \leq \tilde{C} \leq \tilde{D}\}$ is observable, and the event $B := \big\{(Y - Y')(C - C') > 0\big\}$ is observable on A, i.e., $B \cap A$ is observable.

Lakhal *& al.* (2008) introduced the conditional Kendall's tau $\tau_a := 2\mathbb{P}[B \mid A] - 1$. As noted by Oakes (1989),

$$\mathbb{P}[B \mid \tilde{Y} = y, \tilde{C} = c] = \chi_\theta\big(H(y, c)\big),$$

where $H(y,c) := \mathbb{P}(Y > y, C > c)$ is the joint survival function of (Y,C),

$$\chi_\theta(v) = \frac{-v\varphi''_\theta(v)}{-v\varphi''_\theta(v) + \varphi'_\theta(v)}$$

and $\varphi''_\theta(v)$ denotes the second derivative of $\varphi_\theta(v)$ with respect to v. Hence,

$$\mathbb{P}[B \mid A] = \mathbb{E}\left[\chi_\theta\big(H(\tilde{Y},\tilde{C})\big) \mid A\right].$$

Since $\tilde{Y} = \tilde{T}$ and $\tilde{C} = \tilde{Z}$ on the event A, we get that

$$g_n(\theta) = \binom{n}{2}^{-1} \sum_{i<j} \mathbb{1}_{A_{ij}} \left(\mathbb{1}_{B_{ij}} - \chi_\theta\big(H(\tilde{T}_{ij},\tilde{Z}_{ij})\big)\right),$$

with n independent observable four-tuples $(T_i, Z_i, \Delta_{1i}, \Delta_{2i})$, is a U-statistic having a null expectation, where A_{ij}, B_{ij}, $\tilde{T}_{ij}$ and $\tilde{Z}_{ij}$ are constructed similarly to A, B, $\tilde{T}$ and $\tilde{Z}$ with the two observable four-tuples $(T_i, Z_i, \Delta_{1i}, \Delta_{2i})$ and $(T_j, Z_j, \Delta_{1j}, \Delta_{2j})$. The joint survival function H of (Y,C) can be estimated in the upper wedge $\{0 \le s \le t\}$ by the Lin & Ying (1993) estimator

$$\hat{H}(s,t) = \frac{\hat{J}(s,t)}{\hat{G}_2(t)}, \tag{3}$$

where $\hat{J}(s,t)$ is the empirical joint survival function of (T,Z). Therefore the above U-statistic $g_n(\theta)$ can be approximated by

$$\hat{g}_n(\theta) = \binom{n}{2}^{-1} \sum_{i<j} \left(\mathbb{1}_{B_{ij}} - \chi_\theta\big(\hat{H}(\tilde{Y}_{ij},\tilde{C}_{ij})\big)\right) \mathbb{1}_{A_{ij}}. \tag{4}$$

The Lakhal *& al.* (2008) estimator $\hat{\theta}_1$ is then defined by $\hat{g}_n(\hat{\theta}_1) = 0$.

Remark 3.3 Lakhal *& al.* (2008) propose to use the following variant of (4):

$$\hat{g}_n(\theta) = \binom{n}{2}^{-1} \sum_{i<j} w(\tilde{Y}_{ij},\tilde{C}_{ij}) \left(\mathbb{1}_{B_{ij}} - \chi_\theta\big(\hat{H}(\tilde{Y}_{ij},\tilde{C}_{ij})\big)\right) \mathbb{1}_{A_{ij}}$$

where $w(\cdot,\cdot)$ is some random weight function converging to a deterministic function. In our numerical studies (see Section 4), we always use the function $\hat{g}_n$ given by (4), i.e., with weight function $w(\cdot,\cdot) \equiv 1$. We also used another weight function $w(\cdot,\cdot)$ suggested by Lakhal *& al.* (2008) but it did not yield any impact on our conclusions. Results corresponding to this weight function will be therefore omitted in Section 4.

3.3 Estimating θ_0 : likelihood based methodology

Now we describe our likelihood method to estimate θ_0. In a first step, we calculate the likelihood function of the unknown copula parameter θ which will be estimated in a second step.

Denote by t, z, δ_1 and δ_2 the possible realizations of T, Z, Δ_1 and Δ_2 respectively. We below derive the likelihood of θ given $(t, z, \delta_1, \delta_2)$ by considering the four possible situations for the values of δ_1 and δ_2. For each case we write a Radon-Nikodým derivative of the distribution of (T, Z) on the event $\{\Delta_1 = \delta_1, \Delta_2 = \delta_2\}$. The symbol "$\propto$" means that the two members on the left and right hand sides of this symbol are proportional functions (of θ), and we denote by $f^{(ij)}$ the partial (i, j)-th derivative of any bivariate function f.

- *(type A)* If $\delta_1 = \delta_2 = 0$ (hence $t = z$): on the event $\{\Delta_1 = 0, \Delta_2 = 0\}$, $T = Z = D$ and

$$\begin{aligned}\mathbb{P}(D \in \mathrm{d}z, \Delta_1 = 0, \Delta_2 = 0) &= \mathbb{P}(D \in \mathrm{d}z, Y > D, C > D) \\ &= \mathbb{P}(D \in \mathrm{d}z)\mathbb{P}(Y > z, C > z) \propto \mathcal{C}_\theta\left(S(z), G_1(z)\right)\mathrm{d}z.\end{aligned}$$

- *(type B)* If $\delta_1 = 0$ and $\delta_2 = 1$ (hence $t = z$): on the event $\{\Delta_1 = 0, \Delta_2 = 1\}$, $T = Z = C$ and

$$\begin{aligned}\mathbb{P}(C \in \mathrm{d}z, \Delta_1 = 0, \Delta_2 = 1) &= \mathbb{P}(C \in \mathrm{d}z, Y > C, C \leq D) \\ &= \mathbb{P}(D \geq z)\mathbb{P}(C \in \mathrm{d}z, Y > z) \propto \mathcal{C}_\theta^{(01)}\left(S(z), G_1(z)\right)\mathrm{d}z.\end{aligned}$$

- *(type C)* If $\delta_1 = 1$ and $\delta_2 = 0$: on the event $\{\Delta_1 = 1, \Delta_2 = 0\}$, $T = Y$, $Z = D$ and

$$\begin{aligned}\mathbb{P}(Y \in \mathrm{d}t, D \in \mathrm{d}z, \Delta_1 = 1, \Delta_2 = 0) &= \mathbb{P}(Y \in \mathrm{d}t, D \in \mathrm{d}z, C > D, Y \leq D) \\ &= \mathbb{P}(D \in \mathrm{d}z)\mathbb{P}(Y \in \mathrm{d}t, C > z) \\ &\propto \mathcal{C}_\theta^{(10)}\left(S(t), G_1(z)\right)\mathrm{d}t\,\mathrm{d}z.\end{aligned}$$

- *(type D)* If $\delta_1 = \delta_2 = 1$: on the event $\{\Delta_1 = 1, \Delta_2 = 1\}$, $T = Y$, $Z = C$ and

$$\begin{aligned}\mathbb{P}(Y \in \mathrm{d}t, C \in \mathrm{d}z, \Delta_1 = 1, \Delta_2 = 1) &= \mathbb{P}(Y \in \mathrm{d}t, C \in \mathrm{d}z, Y \leq C, C \leq D) \\ &= \mathbb{P}(D \geq z)\mathbb{P}(Y \in \mathrm{d}t, C \in \mathrm{d}z) \\ &\propto \mathcal{C}_\theta^{(11)}\left(S(t), G_1(z)\right)\mathrm{d}t\,\mathrm{d}z.\end{aligned}$$

Finally the likelihood L of θ given a single observation $(t, z, \delta_1, \delta_2)$ is given by

$$L(\theta; S, G_1 \mid t, z, \delta_1, \delta_2) \propto \mathcal{C}_\theta^{(\delta_1\delta_2)}\left(S(t), G_1(z)\right)$$
$$= [\varphi'_\theta\big(S(t)\big)]^{\delta_1}[\varphi'_\theta\big(G_1(z)\big)]^{\delta_2}(\varphi_\theta^{-1})^{(\delta_1+\delta_2)}\Big(\varphi_\theta\big(S(t)\big) + \varphi_\theta\big(G_1(z)\big)\Big), \tag{5}$$

where $(\varphi_\theta^{-1})^{(i)}(x)$ denotes the $i-$th derivative of $\varphi_\theta^{-1}(x)$ with respect to x. Note that this likelihood function depends on (θ and) $S(t)$ and $G_1(z)$ but not on $G_2(z)$.

Next, we define $\widehat{\theta}_2$ as the root of the profile derivative log-likelihood, i.e., the solution θ of the equation

$$\sum_{i=1}^{n} \frac{\partial \log L}{\partial \theta}\Big(\theta; \widehat{S}_\theta, \widehat{G}_1 \Big| T_i, Z_i, \Delta_{1i}, \Delta_{2i}\Big) = \mathbf{0}, \tag{6}$$

where $\mathbf{0}$ is the d-dimensional vector of zeros. Actually we cannot exactly use equation (6), since the member on the left hand side in (6) is possibly infinite or non-definite when $\widehat{S}_\theta(T_i), \widehat{G}_1(Z_i) \in \{0, 1\}$ for some i. To alleviate this problem, we define $\widehat{\theta}_2$ as the root of

$$\sum_{i=1}^{n} \mathbb{1}_{\widehat{S}_\theta(T_i)\in]0,1[,\widehat{G}_1(Z_i)\in]0,1[} \frac{\partial \log L}{\partial \theta}\Big(\theta; \widehat{S}_\theta, \widehat{G}_1 \Big| T_i, Z_i, \Delta_{1i}, \Delta_{2i}\Big) = \mathbf{0}. \tag{7}$$

3.4 Goodness-of-fit measure

Since there is no natural reason to a priori select a copula family, it is important to use a goodness-of-fit statistic. Following Lakhal (2006), we will use the statistic $\hat{Q}_{\hat{\theta}}$ (for $\hat{\theta} = \hat{\theta}_1$ or $\hat{\theta}_2$), where

$$\hat{Q}_\theta := \frac{1}{\sum_{i=1}^{n} \mathbb{1}_{\{T_i=Y_i, Z_i=C_i\}}} \sum_{i=1}^{n} \mathbb{1}_{\{T_i=Y_i, Z_i=C_i\}} \Big(\hat{H}(T_i, Z_i) - \mathcal{C}_\theta\big(\hat{S}_\theta(T_i), \hat{G}_1(Z_i)\big)\Big)^2.$$

The quantity $\hat{Q}_{\hat{\theta}}$ is a measure of the difference between the nonparametric estimate $\hat{H}(s, t)$ of the joint survival function of (Y, C) given by (3) and the semi-parametric estimate $\mathcal{C}_{\hat{\theta}}\big(\hat{S}_{\hat{\theta}}(\cdot), \hat{G}_1(\cdot)\big)$. Then, the smaller the value of $\hat{Q}_{\hat{\theta}}$, the best the fit to the data.

4 Simulations

Now, we study the practical behavior of our procedure via simulations. Our purpose is to compare both estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ but also the performances of related estimators based on them and useful in many applications. In this respect, beyond the survival function, other quantities like estimators of $\mathbb{P}(C > y \mid Y \wedge C > x)$ or $\mathbb{P}(C > y \mid Y = x, C > x)$ are investigated in some cases. As mentioned before, the parameter θ of the considered families $\{\mathcal{C}_\theta : \theta \in \Theta\}$ of Archimedean copulas can be multidimensional. However, (4) only enables to obtain a one-dimensional parameter estimator. Subsection 4.1 is therefore devoted to the comparison of $\hat{\theta}_1$ and $\hat{\theta}_2$ in the one-dimensional case while Subsection 4.2 is dealing with the analysis of our methodology in the two-dimensional case. All the simulations below are based on 1000 samples of size $n = 100$ or 200.

4.1 One parameter

First, we simulate a Frank copula for $H(\cdot, \cdot)$ with parameter corresponding to a Kendall's tau equal to either 0, 1/3 or 1/2. We use the R package *copula* to generate couples (U_i, V_i), $i = 1, \ldots, n$, from the copula. We then invert the survival functions of Y and C to obtain $Y_i = S^{-1}(U_i)$ and $C_i = G_1^{-1}(V_i)$. The random variable Y always has an exponential distribution with parameter 1, and C has an exponential distribution with a parameter such that $p := P(Y > C)$ is equal to either 0.25, 0.5 or 0.75. The censoring random variable D has a uniform distribution on $[0, a]$ with a chosen in order that $P(C > D) = 20\%$. Each time a Kaplan-Meier estimator has to be used at a point on the right of the largest data point, it is defined by its value at this largest data point.

Table 1 that corresponds to Table 1 in Lakhal *& al.* (2008) summarizes the obtained results. Globally, results are similar for both methods with a slight advantage for Lakhal *& al.* (2008). However, we generated data with the same characteristics as above but with a Clayton copula. In this case, the results seem to be inverted as it is illustrated in Table 2. On one side, using a likelihood procedure enables to fully introduce parametric information provided by $\mathcal{C}_\theta^{\delta_1\delta_2}(\cdot, \cdot)$ and $\hat{S}_\theta(\cdot)$ while on the other side adjusting a parametric function to a nonparametric conditional Kendall's tau leads to solve a simpler equation. Note that in the Clayton copula case, the function $\chi_\theta(v)$ is simply given by $(\theta+1)/(\theta+2)$ making equation (4) equivalent to the computation of a simple nonparametric conditional Kendall's tau, an operation using therefore a "minimal amount of information".

Tables 3 and 4 summarize the effects of the estimating procedures for θ on the estimation of $S(\cdot)$. Other simulations (with other copulas) led to similar results. Globally, the root of the average over the 1000 runs of the squared errors of both $\hat{S}_1(x)$ and $\hat{S}_2(x)$ at different values of x (denoted $\widehat{RMSE}(\hat{S}_k(x))$, $k = 1, 2$, in Tables 3 and 4) seems to

reflect a slightly better behavior for the method based on the likelihood approach. More precisely, in many situations, this effect seems to be obtained via a better behavior of $\hat{S}_2(\cdot)$ the rights tails. Note that the root of the average over the 1000 runs of the integrated squared errors of $\hat{S}_k(\cdot)$, $k = 1, 2$, describes the behavior of the estimators in a more global way since it only consists of a number. However, it is not reported here since its value depends on the way which the possibly inconsistent parts of $\hat{S}_k(\cdot)$, $k = 1, 2$, are dealt with. If these parts are simply not considered in the corresponding integrals (for each sample, integration is truncated at the largest data point of the sample on which Kaplan-Meier estimator $\hat{\Gamma}(\cdot)$ is constructed), the slightly better behavior of $\hat{S}_2(\cdot)$ is also observed.

Remark 4.1 Both Lakhal *& al.* (2008) and our estimation procedures for $S(\cdot)$ are only valid for strict Archimedean copulas. That is the reason why the case $\tau = 0$ is not reported in Table 2 (contrary to Table 1). Indeed, the Frank copula is strict for any $\tau \in]-1, 1[$ while the Clayton copula is strict only for $\tau \in [0, 1[$, a range corresponding to values of θ larger or equal to 0. As a consequence, since in the Clayton case, each sample will deliver an estimator for θ larger or equal to 0, the resulting non negative values of $\hat{\tau}$ will not be representative of the quality of the estimation procedure (positive bias always induced).

Finally, note that the Clayton copula can also be defined for $\theta \in]-1, +\infty[$ but it is not strict anymore in the sense that for negative values of θ, $\varphi_\theta(0) = -1/\theta$. In this case, some terms of the log-likelihood function can be maximized by negative values of θ that make them tend to $+\infty$. Avoiding that problem should be reached by defining a new restricted (profile derivative) log-likelihood function that prevents that behavior.

(a) Likelihood

n	τ	p	$\hat{E}(\hat{\tau})$	$\widehat{RMSE}(\hat{\tau})$
100	0	0.25	−0.006	0.080
		0.5	0.003	0.102
		0.75	0.027	0.161
	1/3	0.25	0.331	0.075
		0.5	0.344	0.095
		0.75	0.367	0.140
	1/2	0.25	0.499	0.060
		0.5	0.512	0.076
		0.75	0.541	0.117
200	0	0.25	−0.003	0.057
		0.5	0.003	0.070
		0.75	0.013	0.107
	1/3	0.25	0.332	0.049
		0.5	0.341	0.062
		0.75	0.358	0.092
	1/2	0.25	0.500	0.040
		0.5	0.508	0.050
		0.75	0.527	0.075

(b) Lakhal

n	τ	p	$\hat{E}(\hat{\tau})$	$\widehat{RMSE}(\hat{\tau})$
100	0	0.25	−0.001	0.077
		0.5	−0.003	0.096
		0.75	−0.006	0.143
	1/3	0.25	0.327	0.073
		0.5	0.327	0.090
		0.75	0.322	0.124
	1/2	0.25	0.493	0.061
		0.5	0.491	0.075
		0.75	0.491	0.103
200	0	0.25	0.001	0.056
		0.5	0.000	0.068
		0.75	−0.002	0.101
	1/3	0.25	0.330	0.049
		0.5	0.332	0.060
		0.75	0.333	0.083
	1/2	0.25	0.497	0.040
		0.5	0.498	0.049
		0.75	0.499	0.068

Table 1: Results obtained for the Frank copula corresponding to the above setting. $\hat{E}(\hat{\tau})$ corresponds to the average of the Kendall's tau over the 1000 runs and $\widehat{RMSE}(\hat{\tau})$ is the root of the average over these runs of the squared errors of the Kendall's tau. Results for our (Lakhal *& al.* (2008)) procedure are displayed on the left (right) part of the table.

(a) Likelihood

n	τ	p	$\hat{E}(\hat{\tau})$	$\widehat{RMSE}(\hat{\tau})$
100	1/3	0.5	0.336	0.083
200	1/3	0.5	0.338	0.058
100	1/2	0.5	0.501	0.071
200	1/2	0.5	0.501	0.049

(b) Lakhal

n	τ	p	$\hat{E}(\hat{\tau})$	$\widehat{RMSE}(\hat{\tau})$
100	1/3	0.5	0.329	0.094
200	1/3	0.5	0.336	0.064
100	1/2	0.5	0.496	0.078
200	1/2	0.5	0.498	0.054

Table 2: Results obtained for the Clayton copula corresponding to the above setting.

(a) Likelihood

$S(x)$	90	70	50	30	10
$\hat{E}\big(\hat{S}_2(x)\big)$	90.11	69.92	49.84	30.31	11.13
$\widehat{bias}\big(\hat{S}_2(x)\big)$	0.11	-0.08	-0.16	0.31	1.13
2.5%	83.38	58.75	37.26	18.44	2.77
97.5%	95.71	80.11	63.68	44.19	24.15
$\widehat{RMSE}\big(\hat{S}_2(x)\big)$	3.23	5.64	6.87	6.66	5.58

(b) Lakhal

$S(x)$	90	70	50	30	10
$\hat{E}\big(\hat{S}_1(x)\big)$	90.12	70.05	50.06	30.51	11.26
$\widehat{bias}\big(\hat{S}_1(x)\big)$	0.12	0.05	0.06	0.51	1.26
2.5%	83.33	59.04	37.72	18.54	2.80
97.5%	95.72	80.30	63.27	44.14	24.23
$\widehat{RMSE}\big(\hat{S}_1(x)\big)$	3.23	5.57	6.78	6.78	5.72

Table 3: Results obtained for the estimation of $S(\cdot)$ at different quantiles of $S(\cdot)$ (in %). The second, third and sixth lines of each table above correspond to the estimated mean, bias and root mean squared error of the estimated survival functions of Y at the different considered quantiles while the percentiles 2.5% and 97.5% of the obtained values of $\hat{S}_1(\cdot)$ and $\hat{S}_2(\cdot)$ are displayed on the fourth and fifth lines. The used copula is the Clayton's one with $n = 100$.

(a) Likelihood

$S(x)$	90	70	50	30	10
$\hat{E}\big(\hat{S}_2(x)\big)$	90.00	69.94	49.96	29.99	10.44
$\widehat{bias}\big(\hat{S}_2(x)\big)$	-0.00	-0.06	-0.04	-0.01	0.44
2.5%	85.44	62.56	40.90	21.90	4.26
97.5%	93.92	77.33	58.87	39.19	18.00
$\widehat{RMSE}\big(\hat{S}_2(x)\big)$	2.23	3.80	4.65	4.48	3.55

(b) Lakhal

$S(x)$	90	70	50	30	10
$\hat{E}\big(\hat{S}_1(x)\big)$	90.01	70.01	50.07	30.10	10.51
$\widehat{bias}\big(\hat{S}_1(x)\big)$	0.01	0.01	0.07	0.10	0.51
2.5%	85.42	62.82	41.35	22.05	4.34
97.5%	93.92	77.16	58.71	39.51	17.71
$\widehat{RMSE}\big(\hat{S}_1(x)\big)$	2.23	3.76	4.55	4.54	3.65

Table 4: Same results as in Table 3 but with $n = 200$.

4.2 Two parameters

As already said, a great interest of our methodology is that it can be applied to copulas with multidimensional parameters. To our knowledge, no other existing procedure makes it possible in this context. Since it is not obvious to a priori choose a parametric form of copula that correctly describes a data set, a practical way to capture characteristics of the dependency between two variables is to make the corresponding copula model more flexible by increasing the number of its parameters.

We simulate for $n = 200$ a two-parameter Frank copula for $H(\cdot, \cdot)$ with parameters $\alpha = 8$ and $\beta \approx 3.695$ in order to have $\tau = 0.5$. We used Algorithm 1 in Nelsen (2005) to generate couples (U_i, V_i), $i = 1, \ldots, n$, from the copula by numerically inverting the function $K(x) = x - \varphi_\theta(x)/\varphi'_\theta(x)$. The survival functions $S(\cdot)$ and $G_1(\cdot)$ are exponential with parameters 1 and such that $p = 0.5$ respectively. The censoring random variable D has a uniform distribution on $[0, a]$ with a chosen in order that $P(C > D) = 20\%$. Each time a Kaplan-Meier estimator has to be used at a point on the right of the largest data point, it is defined by its value at this largest data point. In order to illustrate the interest of modeling with multidimensional parameters copulas and to compare our procedure with Lakhal *& al.* (2008) on a robustness point of view, we fit different copulas to the data (two-parameter Frank, Frank, Clayton and Gumbel copulas).

Table 5 shows results for the Kendall's tau while Table 6 shows the impact of this estimation on the survival functions $\hat{S}_2(\cdot)$ and $\hat{S}_1(\cdot)$. As expected, the results obtained by the likelihood method with the two-parameter Frank copula are the best ones everywhere. Estimators of θ for the Frank, Clayton and Gumbel copulas does not necessarily correspond to the Kendall's tau assumed here anymore. Indeed, the link between θ and τ is valid if the assumed copula is true. Therefore, the obtained $\hat{E}[\hat{\tau}]$ in Table 5 does not necessarily need to correspond to 0.5. However, a distinction can be made between both procedures. Even though the assumed copula is false, Lakhal *& al.* (2008) consists in fitting a function of θ to a nonparametric conditional Kendall's tau; that suggests that this methodology will be more likely to obtain an estimator of θ that is linked to τ under the assumed false copula. This is not the case in the likelihood approach, the goal of which is to find the estimator of θ that correctly adjusts the assumed copula to the data. Therefore, although $\hat{E}[\hat{\tau}]$ seems to be further from 0.5 for the likelihood method, that has no negative impact on $\hat{S}_2(\cdot)$; on the contrary, it seems to lead in many cases to slightly better global behavior (also observed on the estimated integrated mean squared error, see previous subsection) than Lakhal *& al.* (2008). As in the previous subsection, this nice behavior is also observed in the right tails of the distributions, especially for the Clayton copula in Table 6. Finally, results about the Gumbel copula are also interesting. These are the best ones after the two-parameter Frank copula and results for both methods are

very close. Indeed, it is easy to check that in our case, the Gumbel copula is the one that can be made the closest to the considered two-parameter Frank copula.

Copula	Likelihood	Lakhal
two-parameter Frank	0.508	
	(0.057)	
Frank	0.514	0.531
	(0.058)	(0.065)
Clayton	0.324	0.574
	(0.186)	(0.094)
Gumbel	0.530	0.508
	(0.056)	(0.059)

Table 5: $\hat{E}[\hat{\tau}]$ and $\widehat{RMSE}(\hat{\tau})$ (between parentheses) for data generated with the two-parameter Frank copula and to which different copulas are fitted.

	two-parameter Frank	Frank		Clayton		Gumbel	
$S(x)$	Likelihood	Likelihood	Lakhal	Likelihood	Lakhal	Likelihood	Lakhal
90	90.0	92.2	92.2	93.1	92.6	90.5	90.6
	(2.4)	(3.1)	(3.0)	(3.6)	(3.3)	(2.2)	(2.3)
70	70.1	73.6	73.3	78.0	74.9	72.4	72.7
	(3.4)	(5.5)	(5.3)	(8.7)	(6.4)	(4.3)	(4.6)
50	50.0	49.6	49.2	56.1	48.9	50.5	51.1
	(4.1)	(4.6)	(4.6)	(7.9)	(5.3)	(4.3)	(4.6)
30	30.0	25.2	24.8	27.4	20.3	26.2	26.7
	(4.4)	(6.2)	(6.5)	(5.8)	(10.5)	(5.7)	(5.5)
10	10.1	6.7	6.5	4.5	3.0	6.0	6.3
	(5.2)	(4.9)	(5.0)	(6.3)	(7.3)	(5.4)	(5.3)

Table 6: $\hat{E}[\hat{S}_k(x)]$ (in %) and $\widehat{RMSE}(\hat{S}_k(x))$ (between parentheses), $k = 1, 2$, for the above two-parameter Frank copula model and different fitted copulas.

Based on the data generated by the above copula model, we finally study the impact of the estimation of θ on other quantities depending on the assumed copula at different levels. Lakhal *& al.* (2008) proposed to estimate the conditional probability $G_{1|Y\wedge C}(y|x) =$

$\mathbb{P}(C > y \mid Y \wedge C > x)$ in a nonparametric way by

$$\hat{G}^{NP}_{1|Y\wedge C}(y|x) = \hat{\mathbb{P}}(C > y \mid Y \wedge C > x) = \frac{\hat{H}(x,y)}{\hat{H}(x,x)},$$

where $\hat{H}(\cdot,\cdot)$ is the nonparametric estimator given by (3), and in a semi-parametric way by

$$\hat{G}^{1}_{1|Y\wedge C}(y|x) = \hat{\mathbb{P}}(C > y \mid Y \wedge C > x) = \frac{\mathcal{C}_{\hat{\theta}}(\hat{S}_1(x), \hat{G}_1(y))}{\mathcal{C}_{\hat{\theta}}(\hat{S}_1(x), \hat{G}_1(x))}. \tag{8}$$

Another quantity introducing derivatives of the copulas is $G_{1|Y,C}(y|x) = \mathbb{P}(C > y \mid Y = x, C > x)$ that can be estimated by

$$\hat{G}^{1}_{1|Y,C}(y|x) = \hat{\mathbb{P}}(C > y \mid Y = x, C > x) = \frac{\mathcal{C}^{(10)}_{\hat{\theta}}(\hat{S}_1(x), \hat{G}_1(y))}{\mathcal{C}^{(10)}_{\hat{\theta}}(\hat{S}_1(x), \hat{G}_1(x))}. \tag{9}$$

Obviously, $\hat{S}_1(x)$ can be replaced by $\hat{S}_2(x)$ in the above expressions to obtain the equivalent estimators (denoted $\hat{G}^{2}_{1|Y\wedge C}(y|x)$ and $\hat{G}^{2}_{1|Y,C}(y|x)$) using our methodology.

Table 7 shows results for the estimated conditional survival function $y \mapsto \hat{G}^{k}_{1|Y\wedge C}(y|x)$, $k = 1, 2$, where x is the median of Y whereas for the same value of x, Table 8 shows results for the estimated conditional survival function $y \mapsto \hat{G}^{k}_{1|Y,C}(y|x)$, $k = 1, 2$. These last two estimators of conditional survival functions are highly based on the two following copula characteristics: the parameric form of the copula and the value of its parameters. On one side, when the assumed parametric form is true, the results clearly stay excellent (see the first columns of Tables 7 and 8 for the two-parameter Frank copula) whereas when a Frank, a Clayton or a Gumbel copula is used, the results fastly deteriorate. A simple comparison with a nonparametric estimation (second column of Table 7) still exhibits this feature: an estimation obtained without any parametric assumption leads to very better results (especially in bias) than an estimation obtained with a false parametric form for the copula. On the other side, here again, quantities (8) and (9) highly depend on the semi-parametric bivariate survival function of (Y, C) (or its derivatives). As a consequence, a methodology that proposes to estimate θ by optimizing these quantities (through a maximum likelihood approach) with respect to the data seems to be better than a methodology that mainly obtains an estimator of θ via the computation of a conditional Kendall's tau. This effect is still more important in Table 8 in the case where copula derivatives are used. As already noticed, it is not observed in the case of the Gumbel copula which is the closest to the two-parameter Frank copula.

$G_{1\|Y\wedge C}(y\|x)$	two-parameter Frank Lik.	$\hat{G}^{NP}_{1\|Y\wedge C}(y\|x)$	Frank Lik.	Frank Lak.	Clayton Lik.	Clayton Lak.	Gumbel Lik.	Gumbel Lak.
90	90.1	89.9	92.0	92.1	91.3	93.2	91.4	91.2
	(3.1)	(3.9)	(3.3)	(3.3)	(3.0)	(4.0)	(3.0)	(3.0)
70	69.9	69.8	74.3	74.5	73.9	77.1	73.3	73.0
	(5.4)	(6.1)	(6.5)	(6.7)	(6.1)	(8.7)	(5.9)	(5.8)
50	50.3	50.1	55.1	55.2	55.9	58.2	54.4	54.1
	(6.1)	(6.9)	(8.0)	(8.1)	(8.3)	(10.4)	(7.4)	(7.3)
30	29.9	29.7	33.6	33.7	35.3	35.7	33.4	33.3
	(5.6)	(6.3)	(7.1)	(7.2)	(8.1)	(8.7)	(6.9)	(6.8)
10	10.1	9.87	11.6	11.6	12.7	12.2	11.7	11.7
	(4.5)	(5.2)	(5.4)	(5.4)	(6.2)	(5.9)	(5.4)	(5.4)

Table 7: $\hat{E}[\hat{G}^{NP}_{1|Y\wedge C}(y|x)]$, $\hat{E}[\hat{G}^{k}_{1|Y\wedge C}(y|x)]$ (in %), $\widehat{RMSE}(\hat{G}^{NP}_{1|Y\wedge C}(y|x))$ and $\widehat{RMSE}(\hat{G}^{k}_{1|Y\wedge C}(y|x))$ (between parentheses), $k = 1, 2$, for the above two-parameter Frank copula model and different fitted copulas. Lak. (respectively Lik.) stands for the estimators based on the Lakhal *& al.* (2008) (respectively our) procedure to estimate θ.

$G_{1\|Y,C}(y\|x)$	two-parameter Frank Likelihood	Frank Likelihood	Frank Lakhal	Clayton Likelihood	Clayton Lakhal	Gumbel Likelihood	Gumbel Lakhal
90	89.7	86.0	85.7	88.0	83.2	87.1	87.4
	(4.0)	(6.6)	(6.9)	(4.8)	(9.4)	(5.5)	(5.3)
70	68.9	58.0	56.9	62.9	48.5	61.0	61.9
	(6.5)	(14.5)	(15.5)	(10.1)	(23.9)	(11.5)	(10.8)
50	49.0	33.9	32.6	38.5	20.2	37.2	38.4
	(7.1)	(17.9)	(19.0)	(14.0)	(31.0)	(14.7)	(13.7)
30	29.0	15.5	14.6	16.6	4.3	17.2	18.3
	(5.9)	(15.3)	(16.2)	(14.6)	(26.0)	(13.8)	(12.9)
10	9.7	4.0	3.7	2.7	0.2	3.9	4.3
	(4.1)	(6.4)	(6.7)	(7.6)	(9.8)	(6.5)	(6.2)

Table 8: $\hat{E}[\hat{G}^{k}_{1|Y,C}(y|x)]$ (in %) and $\widehat{RMSE}(\hat{G}^{k}_{1|Y,C}(y|x))$ (between parentheses), $k = 1, 2$, for the above two-parameter Frank copula model and different fitted copulas.

5 Application to the Hodgkin disease

In this section, we apply our method to the Hodgkin disease data. We analyse data from 865 early stage Hodgkin lymphoma treated patients at the Princess Margaret Hospital (Toronto, Canada) between 1968 and 1986 (Pintilie, 2006). In this non-randomized cohort of patients, 249 were treated with chemotherapy (CMT) while 616 were treated with radiatio therapy (RT), and we do not know anything about the way the two groups have

been formed. There are:

- 146 ($\approx 59\%$) patients and 293 ($\approx 48\%$) patients in the CMT group and the RT group respectively who are censored due to lost to follow-up before occurence of any event ("type A");
- 42 ($\approx 17\%$) patients and 96 ($\approx 16\%$) patients in the CMT group and the RT group respectively who die without relapse ("type B");
- 12 ($\approx 5\%$) patients and 96 ($\approx 16\%$) patients in the CMT group and the RT group respectively who are censored due to lost to follow-up between occurence of relapse and death ("type C");
- 49 ($\approx 20\%$) patients and 131 ($\approx 21\%$) patients in the CMT group and the RT group respectively who experience both relapse and death ("type D").

Figure 2 displays the Kaplan-Meier estimates $\hat{G}_1(\cdot)$ and $\hat{G}_2(\cdot)$ of the respective survival functions of C (time to death) and D (time to lost to follow-up). They appear to be rather similar showing a similar risk of death in the two treatment groups.

Tables 9 and 10 display the estimated copula parameters with different choices of copula for (Y, C), the corresponding estimations of the Kendall's tau, and the value of the goodness-of-fit statistic introduced in Section 3.4. We see that the estimation of τ strongly differs according to the choice of the copula family, but each choice yields a larger $\hat{\tau}$ in the CMT group. We will not display all the estimated survival functions corresponding to each copula family but we checked that, contrary to $\hat{\tau}$, they appear very close to each other using either the Lakhal *& al.* (2008) method or our likelihood-based method. As expected, the two-parameter Frank family yields the smaller goodness-of-fit statistic, and throughout the sequel we pursue the analysis with this copula family only.

Figure 3 shows the contour plots of the Lakhal *& al.* (2008) function $\hat{g}_n(\theta)$ along with the estimates $\hat{\alpha}$ and $\hat{\beta}$ obtained with the likelihood method. It is interesting to note that $\hat{g}_n(\hat{\alpha}, \hat{\beta})$ is not far away from 0. This figure illustrates the drawback of the Lakhal *& al.* (2008) method: following this method we could only say that the estimate of (α, β) is on the zero-level curve. Note also that the values of (α, β) on this curve correspond to clearly different values of τ (compare with figure 1).

Table 11 shows the bootstrap percentile confidence intervals based on 500 bootstrap samples (each bootstrap sample consists of n four-tuples $(T_i^*, Z_i^*, \Delta_{1i}^*, \Delta_{2i}^*)$, $i = 1, \ldots, n$, randomly taken with replacement from an original sample of size n). Since the sample size in the RT group is 616 (in comparison with 249 in the CMT group), the confidence intervals in the RT group are expected to be very shorter than in the CMT group. This is not the case for the confidence interval of β. In fact, small values of $|\hat{\alpha}|$ are obtained in

the RT group ($[2.46; 4.65]$ in the RT group against $[6.46; 10.86]$ in the CMT group) and as it can be seen in Figure 1, τ is not very sensitive to fluctuations of β for small values of $|\alpha|$. This naturally suggests a larger variability of $\hat{\beta}$.

The estimate of the survival function $S(\cdot)$ of the time to relapse along with its point by point percentile bootstrap confidence interval is displayed on figure 4 for each treatment group. The risk of relapse appears to be lower in the CMT group. On the other hand, Figure 5 displays the estimate of the conditional survival function $\mathbb{P}(C > \cdot \mid C > x, Y = x)$ of the time to death given that death has not occured at time x yet and a relapse occured at time x ($x = 4$). Here again, pointwise percentile bootstrap confidence intervals are added.

The conditional risk of death appears to be higher in the CMT group. This estimate highly depends on the copula parametric shape and the value of its parameter. However, as seen in Section 4, since the flexible two-parameter Frank copula is assumed and its parameter is estimated with our likelihood procedure, a higher level of robustness of the results can be expected in both treatment groups.

From these results, it would be difficult to decide which treatment is the best one: the risk of death appears to be similar in both groups, the risk of relapse appears to be lower in the CMT group, but the risk of death appears to be higher in the CMT group for the individuals who experienced a relapse. These results should however be interpreted with care as we are comparing non-randomized groups of patients.

(a) CMT

Copula	Estimated parameter	$\hat{\tau}$	GoF statistic ($\times 10^4$)
Clayton	$\hat{\theta} \approx 8.15$	0.803	10.1
Frank	$\hat{\theta} \approx 12.42$	0.721	8.3
Gumbel	$\hat{\theta} \approx 2.44$	0.590	5.8
two-parameter Frank	$\hat{\alpha} \approx 8.03$ and $\hat{\beta} \approx 5.71$	0.412	5.1

(b) RT

Copula	Estimated parameter	$\hat{\tau}$	GoF statistic ($\times 10^4$)
Clayton	$\hat{\theta} \approx 1.79$	0.472	5.7
Frank	$\hat{\theta} \approx 3.87$	0.379	4.5
Gumbel	$\hat{\theta} \approx 1.52$	0.340	2.2
two-parameter Frank	$\hat{\alpha} \approx 3.33$ and $\hat{\beta} \approx 7.09$	0.218	1.6

Table 9: Estimated parameters and Kendall's tau obtained by the likelihood method in both the CMT and RT groups. The values of the goodness-of-fit statistic are multiplied by 10^4.

(a) CMT

Copula	Estimated parameter	$\hat{\tau}$	GoF statistic ($\times 10^4$)
Clayton	$\hat{\theta} \approx 9.62$	0.828	8.7
Frank	$\hat{\theta} \approx 12.47$	0.721	8.3
Gumbel	$\hat{\theta} \approx 2.38$	0.580	6.1

(b) RT

Copula	Estimated parameter	$\hat{\tau}$	GoF statistic ($\times 10^4$)
Clayton	$\hat{\theta} \approx 2.28$	0.533	5.6
Frank	$\hat{\theta} \approx 4.13$	0.398	4.6
Gumbel	$\hat{\theta} \approx 1.48$	0.325	2.1

Table 10: Estimated parameters and Kendall's tau obtained by the Lakhal *& al.* (2008) method in both the CMT and RT groups. The values of the goodness-of-fit statistic are multiplied by 10^4.

6 Conclusion

In this paper, a new method is proposed to estimate the parameters of a copula when considering semi-competing risks data. One of the concurrent events is censored by the

	(a) CMT			(b) RT	
	2.5%	97.5%		2.5%	97.5%
θ	6.46	10.86	θ	2.46	4.65
β	2.98	9.05	β	3.73	13.25
τ	0.30	0.59	τ	0.15	0.30

Table 11: Bootstrap confidence intervals for both components of θ and the Kendall's tau of the two-parameter Frank copula in both the CMT and RT groups.

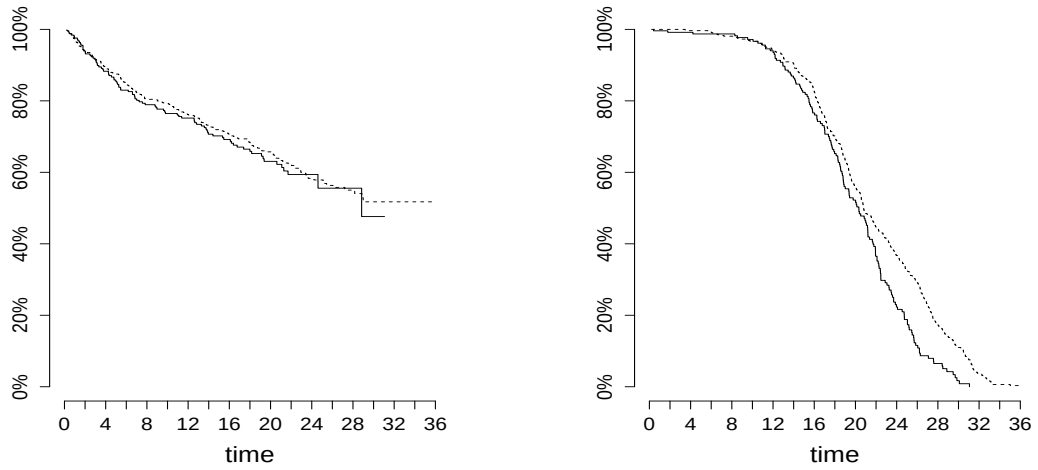
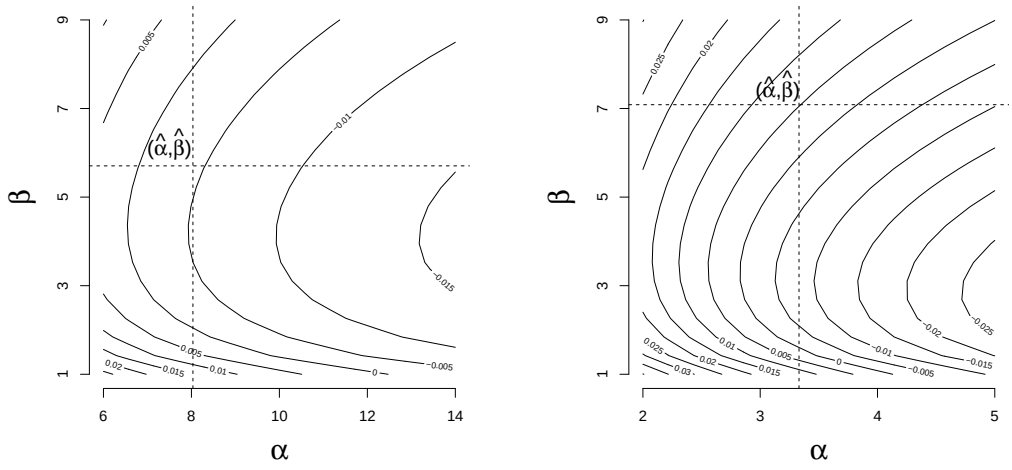


Figure 2: *Left:* Kaplan-Meier $\widehat{G}_1(\cdot)$ for treatment CMT (solid) and RT (dashed). *Right:* Kaplan-Meier $\widehat{G}_2(\cdot)$ for treatment CMT (solid) and RT (dashed).

other one but not vice versa. Their dependency is modeled by an Archimedean copula and they are both submitted to independent right censoring. The new methodology provides a maximum likelihood estimator that behaves well in many practical situations when it is inserted in related survival functions. From the achieved analysis, we can conclude that the methodology should be used in the following situations.

1. The most interesting case is when the dependency between Y and C is too complex (or not sufficiently observable -see also point 2. below-) to be modeled by a copula with a one-dimensional parameter. The only methodology that allows for multidimensional parameter copulas in this modeling context is the one proposed in this paper.

2. When the objective is to estimate survival functions of the type studied in this paper (highly depending on the assumed copula) and when the assumed copula is far from the model that generated the data (this case can occur when a copula is badly chosen due in particular to the censoring mechanisms that "hide" complex



(a) CMT (b) RT

Figure 3: Contour plots of the Lakhal *& al.* (2008) function $\hat{g}_n$ and the corresponding solution obtained by the likelihood method for both the CMT and RT groups.

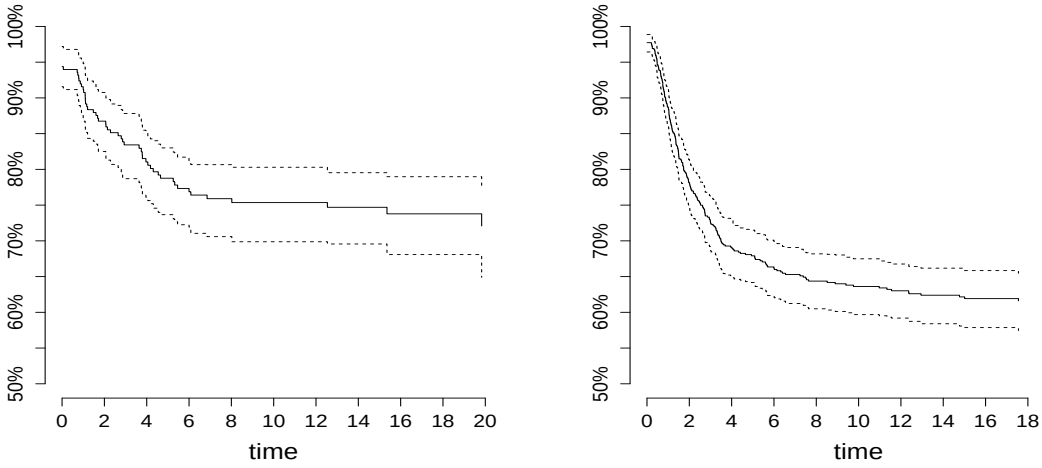


Figure 4: $\hat{S}_2(\cdot)$ and pointwise 95% bootstrap confidence bands. *Left:* CMT group. *Right:* RT group.

structures of dependency), our likelihood approach enables to make the copula as close as possible to the true model.

3. When the semi-parametric conditional Kendall's tau (in Lakhal *& al.* (2008)) is estimated without using any estimator of the bivariate survival function of (Y, C), the new methodology seems to provide better results (typically in the case of the strict Clayton copula).

4. At a lower level, the results obtained in the right tails of the survival function of Y

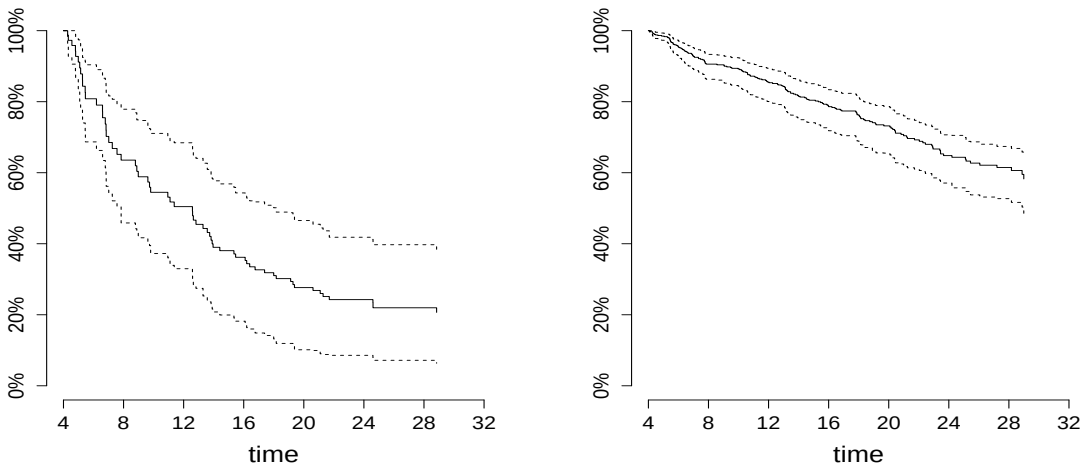


Figure 5: $y \mapsto \hat{\mathbb{P}}(C > \cdot \mid C > 4, Y = 4)$ and pointwise 95% bootstrap confidence bands.. *Left:* CMT group. *Right:* RT group.

seem to be better when the new methodology is used.

Appendix

Proof of Theorem 2.1.

Theorem 2.1 will be proved with the help of the following proposition. The first assertion was noted by Rivest & Wells (2001) and the second one is a copy of proposition 2 in Rivest & Wells (2001).

Proposition A.1. *Let U and V be two positive random variables each having an absolutely continuous distribution. Define the integrated hazard rate Λ of U subject to be censored by V by*

$$\Lambda(\mathrm{d}t) = \frac{\mathbb{P}(U \in \mathrm{d}t, U \leq V)}{\mathbb{P}(U \wedge V \geq t)}.$$

Given an Archimedean generator φ, define the function H_φ by

$$\varphi\big(H_\varphi(t)\big) = -\int_0^t \Gamma(s)\varphi'\big(\Gamma(s)\big)\,\mathrm{d}\Lambda(s) \qquad \text{for every } t \geq 0,$$

where $\Gamma(s) = \mathbb{P}(U \wedge V \geq s)$. Then,

(i) $H_\varphi(\cdot)$ is the survival function of U if the bivariate survival function of (U, V) is the Archimedean copula with generator φ;

(ii) $H_{\varphi_2}(t) < H_{\varphi_1}(t)$ for every $t > 0$ whenever $\varphi_1(\cdot)$ and $\varphi_2(\cdot)$ are two Archimedean generators such that $\varphi_1'(\cdot)/\varphi_2'(\cdot)$ is strictly increasing on $(0, 1)$.

Let Y, C, D be random variables such that D is independent of (Y, C) and the copula describing the dependency between Y and C is $\mathcal{C}_\theta$. We denote by $S(\cdot)$, $G_1(\cdot)$ and $G_2(\cdot)$ the survival functions of Y, C and D respectively. We construct the observable four-tuple $O = (T, Z, \Delta_1, \Delta_2)$ from Y, C and D. We have to prove that the distribution of O completely determines $S(\cdot)$, $G_1(\cdot)$, $G_2(\cdot)$ and θ.

Firstly, since $Z = C \wedge D$ and D is independent of C, we know by Berman (1963) that G_1 and G_2 are determined by the law of (Z, Δ_2). Consequently, the survival function Γ of $Y \wedge C$ is determined by the law of O beacuse of $\Gamma(t) = \mathbb{P}(Y > t, C > t) = \dfrac{\mathbb{P}(T > t, Z > t)}{G_2(t)}$. Now, introduce the integrated hazard rate Λ of C subject to be censored by $Y \wedge D$:

$$\Lambda(\mathrm{d}t) = \frac{\mathbb{P}(C \in \mathrm{d}t, C \leq Y \wedge D)}{\mathbb{P}(C \wedge Y \wedge D \geq t)} = \frac{\mathbb{P}(T \in \mathrm{d}t, \Delta_1 = 0, \Delta_2 = 1)}{\mathbb{P}(T \geq t)},$$

which is determined by the law of O. Moreover, Λ is also the integrated hazard rate of C subject to be censored by Y, because of the independence between D and (Y, C). Therefore θ is uniquely determined by the law of O using proposition A.1 and the fact that G_1, Γ and Λ are determined by the law of O. Finally, S is uniquely determined by the law of O since $\varphi_\theta\big(S(t)\big) + \varphi_\theta\big(G_1(t)\big) = \varphi_\theta\big(\Gamma(t)\big)$.

References

[1] Berman, S.-M. (1963). Note on extreme values, competing risks and semi-Markov processes. *Ann. Math. Statist.*, **34**, 1104–1106.

[2] Braekers, R. and Veraverbeke, N. (2005). A copula-graphic estimator for the conditional survival function under dependent censoring. *Canad. J. Statist.*, **33**, 429–447.

[3] Chen, X., Linton, O. and Van Keilegom, I. (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica*, **71**, 1591–1608.

[4] Ding, A.-A., Shi, G. ,Wang, W., Hsieh, J.-J. (2009). Marginal regression analysis for semicompeting risks data under dependent censoring. *Scand. J. Stat.*, **36**, 481–500.

[5] Fine, J.-P., Jiang, H. and Chappell, R. (2001). On semicompeting risks. *Biometrika,* **88**, 907–919.

[6] Jiang, H., Fine, J.-P., Kosorok, R. and Chappell, R. (2005). Pseudo self-consistent estimation of a copula model with informative censoring. *Scandinavian Journal of Statistics,* **33**, 1–20.

[7] Lakhal, L. (2006). Estimation de la dépendance et des lois marginales dans des modèles pour l'analyse des durées de vies multidimensionnelles. PhD dissertation, Université Laval, 2006.

[8] Lakhal, L., Rivest, L-P. and Abdous, B. (2008). Estimating Survival and Association in a Semicompeting Risks Model. *Biometrics,* **64**, 180–188.

[9] Laurent, S. (2011). Estimating the survival functions in a censored semi-competing risks model. Techninal report, Université de Liège, 2011.

[10] Nelsen, R.-B. (2005). Dependence modeling with Archimedean copulas. In Kolev, N. and Morettin, P., eds. *Proceedings of the Second Brazilian Conference on Statistical Modelling in Insurance and Finance, Institute of Mathematics and Statistics, University of São Paulo (2005)*, 45–54.

[11] Oakes, D. (1989). Bivariate survival models induced by frailties. *J. Amer. Statist.,* **84**, 487–493.

[12] Peng, L. and Fine, J.-P. (2007). Regression modeling of semi-competing risks data. *Biometrics,* **63**, 96–108.

[13] Prentice, R.-L., Kalbfleisch, J.-D., Peterson, A.-V., Flournoy, N., Farewell, V.-T., and Breslow, N.-E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics,* **34**, 541–554.

[14] Rivest, L. and Wells, M.-T. (2001). A martingale approach to the copula-graphic estimator for the survival function under dependent censoring. *J. Multiv. Anal.,* **79**, 138–155.

[15] Xu, J., Kalbfleisch, J.-D. and Tai, B., (2010). Statistical analysis of illness-death processes and semicompeting risks data. *Biometrics,* **66**, 716–25.