Biometrika (2023), **110**, 3, *pp.* 721–738 https://doi.org/10.1093/biomet/asac067
Advance Access publication 20 December 2022

Dependent censoring based on parametric copulas

By C. CZADO

Department of Mathematics, Technical University of Munich, Boltzmannstr. 3, 85748 Garching b. München, Germany
cczado@ma.tum.de

and I. VAN KEILEGOM

Operations Research and Statistics Research Group, KU Leuven, Naamsestraat 69, 3000 Leuven, Belgium
ingrid.vankeilegom@kuleuven.be

Summary

Consider a survival time T that is subject to random right censoring, and suppose that T is stochastically dependent on the censoring time C. We are interested in the marginal distribution of T. This situation is often encountered in practice. Consider, for example, the case where T is a patient's time to death from a certain disease. Then the censoring time C could be the time until the patient leaves the study or the time until death from another cause. If the reason for leaving the study is related to the health condition of the patient, or if the patient dies from a disease that has similar risk factors to the disease of interest, then T and C are likely to be dependent. In this paper we propose a new model that takes such dependence into account. The model is based on a parametric copula for the relationship between T and C, and on parametric marginal distributions for T and C. Unlike most other authors, we do not assume that the parameter defining the copula is known. We give sufficient conditions on these parametric copulas and marginals under which the bivariate distribution of (T, C) is identified. These sufficient conditions are then checked for a wide range of common copulas and marginals. We also study the estimation of the model, and carry out extensive simulations and analysis on a pancreatic cancer dataset to illustrate the proposed model and estimation procedure.

Some key words: Copula; Dependent censoring; Identifiability; Inference.

1. Introduction

A very common situation in survival analysis is that durations are right censored. This can occur for several reasons. In medical studies it often happens that patients who are followed over time until they die of a certain disease remain alive at the end of the study, leave the study before the end for various reasons, or die from another cause. In most work on this topic it is assumed that the survival time T is independent of the censoring time C, where independent censoring should here be interpreted as stochastic independence, but there are many instances in which this independence assumption is violated. Consider, for example,

© The Author(s) 2022. Published by Oxford University Press on behalf of the Biometrika Trust.
All rights reserved. For permissions, please email: journals.permissions@oup.com

Downloaded from https://academic.oup.com/biomet/article/110/3/721/6947986 by Central University of Punjab user on 11 November 2023

the case where the patient leaves the study for reasons relating to their health, or where they die of another related disease. In such cases it is important to take the dependence between T and C into account in the model. However, the seminal paper by Tsiatis (1975) shows that the bivariate distribution of T and C is not identifiable in a completely nonparametric setting. Some authors have therefore proposed parametric or semiparametric models for this bivariate distribution that are identifiable. A popular model is the copula, which allows one to model the marginal laws of T and C separately from the relation between T and C. The first paper in this context is that of Zheng & Klein (1995). They worked with a completely known copula and proposed a nonparametric estimator of the marginal distribution of T and C, called a copula-graphic estimator, that generalizes the Kaplan & Meier (1958) estimator to the case of dependent censoring. Later, Rivest & Wells (2001) focused on the case of Archimedean copulas and obtained a closed-form expression for the copula-graphic estimator. However, a major drawback of this estimator is that it relies on a completely known copula, such as a Frank copula with a known association parameter. In practice, the association parameter is often not known and can have a major influence on the resulting estimator of the marginal distributions. Other contributions in the same vein that involve a known copula include Escarela & Carriere (2003), Braekers & Veraverbeke (2005), Huang & Zhang (2008), Chen (2010), de Uña-Álvarez & Veraverbeke (2017), Emura & Chen (2018) and Sujica & Van Keilegom (2018), whereas other approaches not based on copulas have been studied by Nádas (1971), Basu & Ghosh (1978), Basu (1988), Emoto & Matthews (1990), Scharfstein & Robins (2002), Jackson et al. (2014), Collett (2015), Hsu et al. (2015) and Deresa & Van Keilegom (2020a,b, 2021), among others. The latter papers only consider rather specific (semi)parametric model assumptions, which cannot easily be extended to other contexts. There are also a number of works that have attempted to estimate the copula parameter in a parametric model, but in none of the existing studies is the identifiability of the model shown; see, e.g., the work of Shih et al. (2019) and Fan et al. (2019), who work with a restrictive parametric model, and of Emura & Chen (2016), who estimate the dependence parameter based on a cross-validation criterion. Finally, Schwarz et al. (2013) showed that the class of strict Archimedean copulas is identifiable, provided that the marginal distributions of T and C are fully known and equal.

In this paper we show that the assumption of a completely known copula is not necessary for identifying the joint distribution of T and C. We prove that if the marginal distributions of T and C and the copula function are all modelled parametrically, then under certain conditions the joint model is identifiable. In particular, the association parameter of the copula function is identifiable, which represents an important step forward in the use of copulas in survival analysis. This might seem surprising, since under right censoring one observes either T or C, but in general not both. We will show that the identification of the relation between T and C does not follow standard lines of reasoning commonly used in parametric models; instead, it is a delicate exercise that makes use of the available information in an optimal way. The price paid for the identifiability of the association parameter is that the marginals are no longer fully nonparametric. We believe, nonetheless, that this is an acceptable price, given that one often has no idea how to choose the association parameter in practice. We will derive sufficient conditions on the families of marginal distributions and on the family of copula functions under which the joint model is identifiable. These sufficient conditions are satisfied for a wide range of common parametric marginal distributions and copula functions, making the model very useful in practice.

Downloaded from https://academic.oup.com/biomet/article/110/3/721/6947986 by Central University of Punjab user on 11 November 2023

2. The model, notation and definitions

Let T be a survival time and C a censoring time. Because of random right censoring we observe $Y = \min(T, C)$ and $\Delta = I(T \leqslant C)$. We allow for stochastic dependence between T and C, and will model this dependence with a copula. Throughout the paper we assume that T and C are nonnegative, and that the marginal distributions F_T and F_C of T and C are continuous, and belong to parametric families:

$$F_T \in \{F_{T,\theta_T} : \theta_T \in \Theta_T\}, \quad F_C \in \{F_{C,\theta_C} : \theta_C \in \Theta_C\} \tag{1}$$

for certain parameter spaces Θ_T and Θ_C. We denote their densities by f_T and f_C, or by f_{T,θ_T} and f_{C,θ_C} in the parametric families. We model the bivariate distribution $F_{T,C}$ of (T, C) by a copula-based model. A copula is a bivariate distribution function $\mathcal{C} : [0,1] \times [0,1] \to [0,1]$ with uniform margins. Thanks to Sklar (1959) and the continuity of the marginal distributions F_T and F_C, we know that there is a unique copula $\mathcal{C}$ for which

$$F_{T,C}(t, c) = \mathcal{C}\{F_T(t), F_C(c)\} \tag{2}$$

holds for any $t, c \geqslant 0$. We model the copula parametrically,

$$\mathcal{C} \in \{\mathcal{C}_\theta : \theta \in \Theta\} \tag{3}$$

for some parameter space Θ. For our approach we need to express conditional distribution functions in terms of their associated copulas and so define

$$h_{C|T,\theta}(v \mid u) = \frac{\partial}{\partial u}\mathcal{C}_\theta(u, v), \quad h_{T|C,\theta}(u \mid v) = \frac{\partial}{\partial v}\mathcal{C}_\theta(u, v).$$

Further, we let $F_{Y,\Delta}(y, \delta) = \mathrm{pr}(Y \leqslant y, \Delta = \delta)$ and $f_{Y,\Delta}(y, \delta) = (\mathrm{d}/\mathrm{d}y)F_{Y,\Delta}(y, \delta)$ for $\delta = 0, 1$. Finally, let $F_Y(y) = \sum_{\delta=0}^{1} F_{Y,\Delta}(y, \delta)$ and $f_Y(y) = \sum_{\delta=0}^{1} f_{Y,\Delta}(y, \delta)$ be the distribution and density of the observable random variable Y, respectively. We can then write

$$\begin{aligned}
&F_{T|C}(t \mid c) = h_{T|C}\{F_T(t) \mid F_C(c)\}, \quad F_{C|T}(c \mid t) = h_{C|T}\{F_C(c) \mid F_T(t)\},\\
&F_Y(y) = F_T(y) + F_C(y) - \mathcal{C}\{F_T(y), F_C(y)\},\\
&f_{Y,\Delta}(y, 1) = f_T(y)[1 - h_{C|T}\{F_C(y) \mid F_T(y)\}],\\
&f_{Y,\Delta}(y, 0) = f_C(y)[1 - h_{T|C}\{F_T(y) \mid F_C(y)\}].
\end{aligned} \tag{4}$$

For a derivation of the results in (4) we refer to the Appendix. For all the functions introduced above, copula parameters θ and marginal parameters θ_T and θ_C can be added to indicate the parametric versions of these functions; for example, with $\alpha = (\theta, \theta_T, \theta_C)^{\mathrm{T}}$, we write $f_{Y,\Delta,\alpha}(y, 1) = f_{T,\theta_T}(y)[1 - h_{C|T,\theta}\{F_{C,\theta_C}(y) \mid F_{T,\theta_T}(y)\}]$.

3. Identifiability

The study of the identifiability of models (1)–(3) forms the backbone of this paper, and we present detailed results in this section. All proofs are given in the Appendix.

Downloaded from https://academic.oup.com/biomet/article/110/3/721/6947986 by Central University of Punjab user on 11 November 2023

We start with a general result that gives sufficient conditions under which the model is identifiable. By identifiability we mean that the parameters $(\theta, \theta_T, \theta_C) \in \Theta \times \Theta_T \times \Theta_C$ determine in a unique way the density of the observable random variables (Y, Δ), i.e., if $f_{Y,\Delta,\alpha_1} \equiv f_{Y,\Delta,\alpha_2}$, then $\alpha_1 = \alpha_2$, where $\alpha_j = (\theta_j, \theta_{Tj}, \theta_{Cj})^{\mathrm{T}}$ $(j = 1, 2)$.

THEOREM 1. *Suppose that the following two conditions hold.*

(i) *For $\theta_{T1}, \theta_{T2} \in \Theta_T$ and $\theta_{C1}, \theta_{C2} \in \Theta_C$ we have the four equivalences*

$$\lim_{t\to 0} \frac{f_{T,\theta_{T1}}(t)}{f_{T,\theta_{T2}}(t)} = 1 \iff \theta_{T1} = \theta_{T2}, \qquad \lim_{t\to\infty} \frac{f_{T,\theta_{T1}}(t)}{f_{T,\theta_{T2}}(t)} = 1 \iff \theta_{T1} = \theta_{T2},$$

$$\lim_{t\to 0} \frac{f_{C,\theta_{C1}}(t)}{f_{C,\theta_{C2}}(t)} = 1 \iff \theta_{C1} = \theta_{C2}, \qquad \lim_{t\to\infty} \frac{f_{C,\theta_{C1}}(t)}{f_{C,\theta_{C2}}(t)} = 1 \iff \theta_{C1} = \theta_{C2}.$$

(ii) *The parameter space $\Theta \times \Theta_T \times \Theta_C$ is such that*

$$\lim_{t\to 0} h_{T|C,\theta}(u_t \mid v_t) = 0 \quad \forall\, (\theta, \theta_T, \theta_C) \in \Theta \times \Theta_T \times \Theta_C$$

or

$$\lim_{t\to\infty} h_{T|C,\theta}(u_t \mid v_t) = 0 \quad \forall\, (\theta, \theta_T, \theta_C) \in \Theta \times \Theta_T \times \Theta_C,$$

and similarly for $h_{C|T,\theta}(v_t \mid u_t)$, where $u_t = F_{T,\theta_T}(t)$ and $v_t = F_{C,\theta_C}(t)$.

Then the model specified in (1)–(3) is identified.

Condition (i) of Theorem 1 refers to the margins only, while condition (ii) involves both the margins and the copula family. Condition (i) is valid for a wide range of parametric families for the marginal densities f_T and f_C, as shown in the next theorem. For other families not mentioned below, condition (i) can be easily checked as well, but we restrict our attention to the most important parametric families in survival analysis. The equivalence relation in condition (i) needs to be satisfied for t tending to both 0 and ∞, so whenever one of the two limits equals 1, the parameter vectors should be equal.

THEOREM 2. *Condition* (i) *of Theorem 1 is satisfied for the families of lognormal, log-Student-t, Weibull and log-logistic densities.*

We now study two important classes of copulas in more detail, namely Archimedean and Gaussian copulas. We will check in which cases condition (ii) is valid for these classes. Archimedean copulas can be written as

$$\mathcal{C}(u, v) = \psi^{[-1]}\{\psi(u) + \psi(v)\}, \tag{5}$$

where ψ is a generator, i.e., $\psi : [0, 1] \to [0, \infty)$ is a continuous, strictly decreasing and convex function such that $\psi(1) = 0$. Here $\psi^{[-1]}$ is the pseudo-inverse of ψ, i.e., $\psi^{[-1]}(t) = \psi^{-1}(t)$ if $0 \leqslant t \leqslant \psi(0)$ and $\psi^{[-1]}(t) = 0$ if $t \geqslant \psi(0)$. Important families of Archimedean copulas include the Frank family, corresponding to $\psi_\theta(u) = -\log[\{\exp(-\theta u) - 1\}/\{\exp(-\theta) - 1\}]$ with $\theta \in \mathbb{R}\backslash\{0\}$, the Clayton family, for which $\psi_\theta(u) = \theta^{-1}(u^{-\theta} - 1)$ with $\theta \in [-1, \infty)\backslash\{0\}$, and the Gumbel family, defined by $\psi_\theta(u) = \{-\log(u)\}^\theta$ with $\theta \in [1, \infty)$.

Downloaded from https://academic.oup.com/biomet/article/110/3/721/6947986 by Central University of Punjab user on 11 November 2023

Differentiation of (5) gives

$$h_{T|C}(u \mid v) = \frac{\psi'(v)}{\psi'\big[\psi^{-1}\{\psi(u)+\psi(v)\}\big]}, \quad h_{C|T}(v \mid u) = \frac{\psi'(u)}{\psi'\big[\psi^{-1}\{\psi(u)+\psi(v)\}\big]}$$

for $0 < u, v < 1$, provided the derivatives and inverses in the formula exist. The following lemma helps in assessing whether the first part of condition (ii) is satisfied or not. An analogous result for the second part holds.

LEMMA 1. *Suppose the generator ψ is differentiable on* $(0,1)$. *If* $\lim_{v\to 1}\psi'(v) \in (-\infty, 0)$, *then* $\lim_{t\to\infty} h_{T|C,\theta}\{F_{T,\theta_T}(t) \mid F_{C,\theta_C}(t)\} = 1$.

Lemma 1 can be used to assess whether condition (ii) holds, in the sense that if the statement in the lemma is true, then the limit as t tends to ∞ will not lead to the required result, and instead the limit as t tends to 0 should be calculated. Below we do this for some well-known parametric Archimedean copula families. We will also consider the Gaussian copula, which is not an Archimedean copula and is defined by

$$C_\theta(u, v) = \Phi_\theta\{\Phi^{-1}(u), \Phi^{-1}(v)\},$$

where Φ is the cumulative distribution function of a standard normal random variable, and Φ_θ is the cumulative distribution function of a bivariate standard normal random vector with correlation θ.

THEOREM 3 (FRANK, GUMBEL AND GAUSSIAN COPULAS). *Condition* (ii) *of Theorem 1 is satisfied by the following:*

(i) *the Frank copula, independently of the marginal distributions and the parameter space;*
(ii) *the Gumbel copula if* $\lim_{t\to 0}\log F_{T,\theta_T}(t)/\log F_{C,\theta_C}(t) \in (0,\infty)$ *for all* $(\theta_T, \theta_C) \in \Theta_T \times \Theta_C$;
(iii) *the Gaussian copula if*

$$\lim_{t\to 0} A_{\theta,F_{T,\theta_T},F_{C,\theta_C}}(t) = -\infty \quad \forall\,(\theta,\theta_T,\theta_C) \in \Theta\times\Theta_T\times\Theta_C$$

or

$$\lim_{t\to \infty} A_{\theta,F_{T,\theta_T},F_{C,\theta_C}}(t) = -\infty \quad \forall\,(\theta,\theta_T,\theta_C) \in \Theta\times\Theta_T\times\Theta_C,$$

and similarly for $A_{\theta,F_{C,\theta_C},F_{T,\theta_T}}$, *where* $A_{\theta,F_1,F_2}(t) = \Phi^{-1}\{F_1(t)\} - \theta\Phi^{-1}\{F_2(t)\}$.

It can easily be seen that $\lim_{t\to 0}\log F_{T,\theta_T}(t)/\log F_{C,\theta_C}(t) = \sigma_C^2/\sigma_T^2$ if $\log T \sim N(\mu_T, \sigma_T^2)$ and $\log C \sim N(\mu_C, \sigma_C^2)$, and that this limit equals ρ_T/ρ_C if $T \sim \text{Wei}(\lambda_T, \rho_T)$ and $C \sim \text{Wei}(\lambda_C, \rho_C)$; we refer to (A3) in the Appendix for the definition of the Weibull parameters. Hence, because of Theorem 3(ii), the Gumbel copula can be used for these two marginal specifications.

Regarding the Gaussian copula, for any distribution F with positive support, the function $\Phi^{-1}\{F(t)\}$ tends to $-\infty$ and $+\infty$ as t tends to zero and infinity, respectively. Hence, the limit of $A_{\theta,F_1,F_2}(t)$ will be determined by which of the functions F_1 and F_2 dominates in the limit. As an example, consider the case where $\log T \sim N(\mu_T, \sigma_T^2)$ and $\log C \sim N(\mu_C, \sigma_C^2)$. Then

Downloaded from https://academic.oup.com/biomet/article/110/3/721/6947986 by Central University of Punjab user on 11 November 2023

$\Phi^{-1}\{F_T(t)\} = \{\log(t) - \mu_T\}/\sigma_T$, and hence $A_{\theta,F_T,F_C}(t) = (1/\sigma_T - \theta/\sigma_C)\log(t) - (\mu_T/\sigma_T - \theta\mu_C/\sigma_C)$, and this tends to $-\infty$ as t tends to either zero or infinity depending on the sign of $\sigma_C - \theta\sigma_T$. This shows that for lognormal margins and for a Gaussian copula, condition (ii) of Theorem 1 is satisfied locally around the true parameter values. See also Nádas (1971), Basu & Ghosh (1978) and Deresa & Van Keilegom (2020a), who considered the bivariate normal case. For other margins the condition is harder to verify, since the normal quantile function Φ^{-1} is difficult to handle. Numerical calculation of the function $A_{\theta,F_T,F_C}(t)$ when $T \sim \text{Wei}(\lambda_T, \rho_T)$ and $C \sim \text{Wei}(\lambda_C, \rho_C)$ supports, however, that the limit of this function equals $-\infty$ if $\rho_T < \rho_C$ and t tends to ∞, or if $\rho_T > \rho_C$ and t tends to 0.

Condition (ii) is a sufficient condition for identification, but not, per se, a necessary condition. In fact, the proof of Theorem 1 does not, in a sense, make use of all the available information, since it is essentially based on the behaviour of the density of (Y, Δ) as time goes to either infinity or zero, but does not take into account the behaviour at intermediate time-points. We will illustrate this for the Clayton copula with $\theta > 0$; in this case the Clayton copula is strict, i.e., $\psi(0) = \infty$. One can easily show that for the Clayton copula we have that

$$\begin{aligned} h_{T|C,\theta}\{F_T(t) \mid F_C(t)\} &= \left[1 + \left\{\frac{F_C(t)}{F_T(t)}\right\}^{\theta} - F_C(t)^{\theta}\right]^{-(\theta+1)/\theta}, \\ h_{C|T,\theta}\{F_C(t) \mid F_T(t)\} &= \left[1 + \left\{\frac{F_T(t)}{F_C(t)}\right\}^{\theta} - F_T(t)^{\theta}\right]^{-(\theta+1)/\theta}; \end{aligned} \tag{6}$$

see also Aas et al. (2009). Hence, it is not possible that both functions tend to zero as t tends to zero. This is because $\lim_{t\to 0} h_{T|C,\theta}\{F_T(t) \mid F_C(t)\} = 0$ if and only if $F_C(t)/F_T(t) \to \infty$ and $\lim_{t\to 0} h_{C|T,\theta}\{F_C(t) \mid F_T(t)\} = 0$ if and only if $F_C(t)/F_T(t) \to 0$. Since both functions converge to 1 as t tends to infinity, we see that condition (ii) is not satisfied. However, from the fact that $f_{Y,\Delta}(y, 0) = f_C(y)[1 - h_{T|C}\{F_T(y) \mid F_C(y)\}]$ for all y and similarly for $f_{Y,\Delta}(y, 1)$, we can identify the parameters θ, θ_T and θ_C, as the following theorem shows.

THEOREM 4 (CLAYTON COPULA). *Suppose that condition* (i) *of Theorem 1 holds,* $\Theta_T \times \Theta_C$ *is such that* $\lim_{t\to 0} F_{T,\theta_T}(t)/F_{C,\theta_C}(t)$ *is either* 0 *or* $+\infty$ *for all* $\theta_T \in \Theta_T$ *and* $\theta_C \in \Theta_C$, *and the copula* $\mathcal{C}_\theta$ *is a Clayton copula with* $\theta > 0$. *Then the model specified in* (1)–(3) *is identified.*

The condition that $\lim_{t\to 0} F_{T,\theta_T}(t)/F_{C,\theta_C}(t) = 0$ or $+\infty$ is satisfied for many parametric families. For example, consider the lognormal family for T and C, $\log T \sim N(\mu_T, \sigma_T^2)$ and $\log C \sim N(\mu_C, \sigma_C^2)$. Then

$$\begin{aligned} \lim_{t\to 0} \frac{F_T(t)}{F_C(t)} &= \lim_{t\to 0} \frac{f_T(t)}{f_C(t)} \\ &= \frac{\sigma_C}{\sigma_T} \lim_{t\to 0} \exp\left\{-\frac{1}{2}\frac{(\log t - \mu_T)^2}{\sigma_T^2} + \frac{1}{2}\frac{(\log t - \mu_C)^2}{\sigma_C^2}\right\} \\ &= \frac{\sigma_C}{\sigma_T} \lim_{t\to 0} \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma_T^2} - \frac{1}{\sigma_C^2}\right)(\log t)^2 + \left(\frac{\mu_T}{\sigma_T^2} - \frac{\mu_C}{\sigma_C^2}\right)\log t - \frac{1}{2}\left(\frac{\mu_T^2}{\sigma_T^2} - \frac{\mu_C^2}{\sigma_C^2}\right)\right\}. \end{aligned}$$

Downloaded from https://academic.oup.com/biomet/article/110/3/721/6947986 by Central University of Punjab user on 11 November 2023

This converges to 0 if $\sigma_T < \sigma_C$ or if $\sigma_T = \sigma_C$ and $\mu_T > \mu_C$. Similarly, the expression converges to infinity if $\sigma_T > \sigma_C$ or if $\sigma_T = \sigma_C$ and $\mu_T < \mu_C$. Hence, the model is identified locally around the true (μ_T, σ_T) and (μ_C, σ_C). If $\mu_T = \mu_C$ and $\sigma_T = \sigma_C$, it is easily seen that $h_{T|C}\{F_T(t) \mid F_C(t)\} = \{2 - F_C(t)^\theta\}^{-(\theta+1)/\theta}$ for the Clayton copula. Therefore, $\lim_{t\to 0} h_{T|C}\{F_T(t) \mid F_C(t)\} = 2^{-(\theta+1)/\theta}$. It follows that $\lim_{t\to 0} f_{C,\mu_{C1},\sigma_{C1}}(t)/f_{C,\mu_{C2},\sigma_{C2}}(t) = \{1-2^{-(\theta_2+1)/\theta_2}\}/\{1-2^{-(\theta_1+1)/\theta_1}\}$ for two sets of parameters $(\theta_1, \mu_{C1}, \sigma_{C1})$ and $(\theta_2, \mu_{C2}, \sigma_{C2})$. Since the limit $\lim_{t\to 0} f_{C,\mu_{C1},\sigma_{C1}}(t)/f_{C,\mu_{C2},\sigma_{C2}}(t)$ can only equal 0, 1 or ∞ for the lognormal density, it follows that $\theta_1 = \theta_2$ and hence $(\mu_{C1}, \sigma_{C1}) = (\mu_{C2}, \sigma_{C2})$. This shows that the model is also identifiable when $\mu_T = \mu_C$ and $\sigma_T = \sigma_C$.

Using similar, but easier, calculations we can show that if $T \sim \text{Wei}(\lambda_T, \rho_T)$ and $C \sim \text{Wei}(\lambda_C, \rho_C)$, then $\lim_{t\to 0} F_T(t)/F_C(t)$ equals 0 if $\rho_T > \rho_C$, ∞ if $\rho_T < \rho_C$, and λ_T/λ_C if $\rho_T = \rho_C$; we refer to (A3) for the definition of the Weibull parameters.

Remark 1. As pointed out by a referee, we could also define our model in terms of the bivariate survival function instead of the bivariate distribution function. In that case, (2) should be replaced by $S_{T,C}(t,c) = \bar{\mathcal{C}}\{S_T(t), S_C(c)\}$, where $S_{T,C}(t,c) = \text{pr}(T > t, C > c)$, $S_T(t) = \text{pr}(T > t)$, $S_C(c) = \text{pr}(C > c)$ and $\bar{\mathcal{C}}(u,v) = u + v + \mathcal{C}(1-u, 1-v) - 1$. It is easy to see that this is equivalent to requiring that $F_{T,C}(t,c) = \mathcal{C}\{F_T(t), F_C(c)\}$. Hence, our identifiability result in Theorem 1 can also be used when working with bivariate survival functions.

4. Estimation

In this section we consider parameter estimation of the joint parametric model for the survival time T and the censoring time C specified in (1)–(3). For this we assume that we have an independent and identically distributed sample $\mathcal{D} = \{(y_i, \delta_i),\ i = 1, \ldots, n\}$ available. Then the joint loglikelihood for the parameter vector $\alpha = (\theta, \theta_T, \theta_C)^{\mathrm{T}}$ is

$$\begin{aligned}\ell(\alpha; \mathcal{D}) &= \sum_{i=1}^{n} \log\{f_{Y,\Delta,\alpha}(y_i, \delta_i)\} \\ &= \sum_{\delta_i=1} \log\big(f_{T,\theta_T}(y_i)[1 - h_{C|T,\theta}\{F_{C,\theta_C}(y_i) \mid F_{T,\theta_T}(y_i)\}]\big) \\ &\quad + \sum_{\delta_i=0} \log\big(f_{C,\theta_C}(y_i)[1 - h_{T|C,\theta}\{F_{T,\theta_T}(y_i) \mid F_{C,\theta_C}(y_i)\}]\big). \end{aligned} \tag{7}$$

We follow a maximum likelihood approach by maximizing the loglikelihood specified in (7), i.e., we define parameter estimators by

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

where $A = \Theta \times \Theta_T \times \Theta_C$. If we use, for instance, lognormal margins for T and C, and single-parameter copula families, we need to optimize over five parameters. In the simulation study and the data application we use unconstrained optimization.

To obtain the asymptotic normality of $(\hat{\theta}, \hat{\theta}_T, \hat{\theta}_C)$, we make use of the results of White (1982), in particular sufficient conditions under which a parameter estimator defined as the maximizer of a certain criterion function is asymptotically normal. The results allow

Downloaded from https://academic.oup.com/biomet/article/110/3/721/6947986 by Central University of Punjab user on 11 November 2023

Table 1. *Parameter specifications for the simulation scenarios with lognormal margins with mean parameters μ_T and μ_C and standard deviation parameters σ_T and σ_C for T and C, respectively, and dependence measured by Kendall's τ*

Scenario	μ_T	σ_T	μ_C	σ_C	τ	θ_{Frank}	θ_{Clayton}	θ_{Gumbel}	θ_{Gauss}
1	2.2	1.0	2.0	0.25	0.2	1.86	0.50	1.25	0.31
					0.5	5.74	2.00	2.00	0.71
					0.7	11.74	4.67	3.33	0.89
2	2.5	1.0	2.0	0.50	0.2	1.86	0.50	1.25	0.31
					0.5	5.74	2.00	2.00	0.71
					0.7	11.74	4.67	3.33	0.89

for misspecification of the parametric model. Let $\alpha^* = (\theta^*, \theta_T^*, \theta_C^*)^{\mathrm{T}}$ be the parameter vector that minimizes the Kullback–Leibler information criterion $E\{\log f_{Y,\Delta}(Y,\Delta) - \log f_{Y,\Delta,\alpha}(Y,\Delta)\}$, and let $d = \dim(\Theta) + \dim(\Theta_T) + \dim(\Theta_C)$. Then we have the following result, which is based on White (1982). The regularity conditions (A1)–(A6) in White (1982) are assumptions regarding the true density $f_{Y,\Delta}(y,\delta)$, the density $f_{Y,\Delta,\alpha}(y,\delta)$ under our assumed model, and the derivatives of this density with respect to both α and y.

THEOREM 5. (i) *Under the regularity conditions* (A1)–(A3) *in White (1982),*

$$(\hat{\theta}, \hat{\theta}_T, \hat{\theta}_C) \to (\theta^*, \theta_T^*, \theta_C^*)$$

in probability as $n \to \infty$.

(ii) *Under the regularity conditions* (A1)–(A6) *in White (1982),*

$$n^{1/2}\{(\hat{\theta}, \hat{\theta}_T, \hat{\theta}_C) - (\theta^*, \theta_T^*, \theta_C^*)\} \to N(0, V)$$

in distribution as $n \to \infty$, where $V = A(\alpha^*)^{-1}B(\alpha^*)A(\alpha^*)^{-1}$, with

$$A(\alpha) = \left[E\left\{ \frac{\partial^2}{\partial\alpha_j \partial\alpha_k} \log f_{Y,\Delta,\alpha}(Y,\Delta) \right\} \right]_{j,k=1}^{d},$$

$$B(\alpha) = \left[E\left\{ \frac{\partial}{\partial\alpha_j} \log f_{Y,\Delta,\alpha}(Y,\Delta) \frac{\partial}{\partial\alpha_k} \log f_{Y,\Delta,\alpha}(Y,\Delta) \right\} \right]_{j,k=1}^{d}.$$

If the model is correctly specified, V equals $A(\alpha)^{-1}$, the inverse Fisher matrix.

5. SIMULATION STUDY

We study the performance of the maximum likelihood estimators of θ, θ_T and θ_C for four parametric copula families, namely the Frank, Clayton, Gumbel and Gauss copulas, and lognormal margins for T and C. The parameters of a lognormal random variable X are μ and σ, corresponding to the mean and standard error of $\log(X)$, respectively. Two simulation scenarios are investigated with parameter specifications given in Table 1.

A visualization of the resulting theoretical density, survival and hazard functions of Y for both scenarios is given in the Supplementary Material. For Scenario 1, in particular, the resulting marginal density of the observable random variable Y is nonstandard. This

Downloaded from https://academic.oup.com/biomet/article/110/3/721/6947986 by Central University of Punjab user on 11 November 2023

Table 2. *Uncensoring probabilities* $\mathrm{pr}(\Delta = 1)$ *for all simulation scenarios*

Scenario	Copula	$\tau = 0.2$	$\tau = 0.5$	$\tau = 0.7$
1	Frank	0.42	0.40	0.39
	Clayton	0.43	0.43	0.42
	Gumbel	0.41	0.40	0.39
	Gauss	0.42	0.41	0.40
2	Frank	0.30	0.27	0.25
	Clayton	0.31	0.29	0.27
	Gumbel	0.30	0.27	0.26
	Gauss	0.30	0.28	0.26

is expected since it is the sum of two subdensities. We also see that the strength of the dependence between T and C influences the skewness of Y.

The resulting uncensoring probabilities $\mathrm{pr}(\Delta = 1)$ for all scenarios are reported in Table 2. We see that in Scenario 1 the uncensoring probabilities are fairly constant over the studied dependence strength, and are similar for all investigated copula families. For Scenario 2 we observe lower uncensoring probabilities, which decrease as the dependence strength increases.

We study two sample sizes, $n = 200$ and $n = 500$, and repeat each simulation setting 200 times. For the asymptotic standard errors we determine the needed Hessian matrix numerically. Since numerical evaluation of the Hessian becomes unstable for extremely large copula parameter values, we truncate the copula parameter such that the resulting Kendall's τ is bounded by 0.8. This is not a severe restriction since for the Gaussian copula it corresponds to a correlation of 0.95. We report for the simulation experiments the average estimate, the empirical standard error of the average estimate, the average of the asymptotic standard error estimates for the parameter estimators, and the empirical root mean squared error based on 200 replications.

The results for Scenario 1 are shown in Tables 3, 4, 5 and 6 for the Frank, Clayton, Gumbel and Gauss copulas, respectively. We observe satisfactory performance of the estimation procedure. As expected, the average root mean squared error goes down as the sample size increases. The marginal parameters are well estimated in all cases. The asymptotic standard error estimates for the parameter estimates agree well with the empirical standard errors. Further, since the marginal parameters are almost perfectly unbiased, even when $n = 200$, the root mean squared error agrees almost perfectly with the empirical standard errors. Although the number of Monte Carlo replications, 200, seems small, the results show that it is sufficient. The results for Scenario 2 are given in the Supplementary Material and also show satisfactory performance.

In at most 2.5% of the cases in Scenario 1 and 5% of the cases in Scenario 2, we observed numerical problems for the asymptotic standard error estimation. This occurred most often for the Clayton and Gauss scenarios. The computation time for the simulation is moderate, about 25 minutes per scenario, which was estimated by running 10 replications per parameter setting of each copula family.

6. Data illustration

We apply the dependent censoring model specified in (1)–(3) to a dataset on pancreatic cancer from the Surveillance, Epidemiology, and End Results, SEER, database, available at `https://seer.cancer.gov/data-software/`. In particular, we consider the

Downloaded from https://academic.oup.com/biomet/article/110/3/721/6947986 by Central University of Punjab user on 11 November 2023

Table 3. *Simulation results for the Frank copula in Scenario* 1

		$n = 200$						$n = 500$					
τ		μ_T	σ_T	μ_C	σ_C	θ	τ	μ_T	σ_T	μ_C	σ_C	θ	τ
0.2	aver.est	2.19	0.98	2.00	0.25	2.22	0.22	2.21	1.00	2.00	0.25	1.91	0.20
	sd.aver.est	0.09	0.09	0.04	0.02	1.94	0.16	0.07	0.07	0.03	0.01	1.18	0.11
	aver.asderr	0.11	0.10	0.05	0.02	2.02	0.18	0.07	0.06	0.03	0.01	1.20	0.12
	RMSE	0.09	0.09	0.04	0.02	1.97	0.16	0.07	0.07	0.03	0.01	1.18	0.11
0.5	aver.est	2.19	0.99	2.00	0.25	6.22	0.49	2.21	1.00	2.00	0.25	5.78	0.49
	sd.aver.est	0.11	0.10	0.04	0.02	2.80	0.14	0.07	0.06	0.02	0.01	1.42	0.08
	aver.asderr	0.11	0.10	0.04	0.02	2.36	0.12	0.07	0.06	0.02	0.01	1.40	0.08
	RMSE	0.11	0.10	0.04	0.02	2.84	0.14	0.07	0.06	0.02	0.01	1.42	0.08
0.7	aver.est	2.20	1.00	2.00	0.25	11.85	0.70	2.20	1.00	2.00	0.25	11.87	0.70
	sd.aver.est	0.10	0.10	0.03	0.02	2.79	0.07	0.06	0.05	0.01	0.01	1.77	0.04
	aver.asderr	0.10	0.09	0.02	0.02	2.99	0.06	0.06	0.06	0.02	0.01	1.84	0.04
	RMSE	0.10	0.10	0.03	0.02	2.82	0.07	0.06	0.05	0.01	0.01	1.82	0.04

aver.est, average estimate; sd.aver.est, empirical standard error of the average estimate; aver.asderr, average of the asymptotic standard error estimates for the parameter estimators; RMSE, empirical root mean squared error.

Table 4. *Simulation results for the Clayton copula in Scenario* 1

		$n = 200$						$n = 500$					
τ		μ_T	σ_T	μ_C	σ_C	θ	τ	μ_T	σ_T	μ_C	σ_C	θ	τ
0.2	aver.est	2.18	0.96	1.99	0.26	0.86	0.25	2.20	0.99	2.00	0.25	0.59	0.20
	sd.aver.est	0.11	0.10	0.05	0.03	0.84	0.19	0.07	0.06	0.04	0.02	0.47	0.14
	aver.asderr	0.10	0.09	0.06	0.03	0.93	0.24	0.07	0.06	0.04	0.02	0.54	0.16
	RMSE	0.11	0.10	0.06	0.03	0.91	0.20	0.07	0.06	0.04	0.02	0.47	0.14
0.5	aver.est	2.20	1.00	2.00	0.25	2.18	0.48	2.21	1.00	2.00	0.25	2.06	0.49
	sd.aver.est	0.12	0.10	0.04	0.03	1.20	0.14	0.07	0.06	0.03	0.02	0.74	0.09
	aver.asderr	0.10	0.09	0.04	0.03	1.17	0.15	0.06	0.06	0.03	0.02	0.70	0.09
	RMSE	0.12	0.10	0.04	0.03	1.22	0.14	0.07	0.06	0.03	0.02	0.74	0.10
0.7	aver.est	2.20	0.99	2.00	0.25	5.08	0.70	2.20	1.00	2.00	0.25	4.81	0.70
	sd.aver.est	0.10	0.08	0.03	0.02	2.09	0.08	0.07	0.05	0.02	0.01	0.98	0.04
	aver.asderr	0.10	0.09	0.03	0.02	1.73	0.07	0.06	0.06	0.02	0.01	1.03	0.04
	RMSE	0.10	0.08	0.03	0.02	2.13	0.08	0.07	0.05	0.02	0.01	0.98	0.04

aver.est, average estimate; sd.aver.est, empirical standard error of the average estimate; aver.asderr, average of the asymptotic standard error estimates for the parameter estimators; RMSE, empirical root mean squared error.

monthly survival times of black patients with localized pancreatic cancer between 2000 and 2015. We exclude all patients with zero survival times, leaving us with data on 1549 patients with localized pancreatic cancer, of which 777 died of pancreatic cancer and 772 were still alive at the end of the study period or died of other causes. We view the 772 patients as censored observations. When patients are censored because they died of another disease, their censoring time is likely related to their unobserved survival time, since many diseases share common risk factors, such as stress, eating habits, physical condition, etc.

We fit model (1)–(3) to this dataset for the independence, Frank, Clayton, Gumbel and Gauss copula models with lognormal margins. The resulting parameter maximum likelihood estimates are presented in Table 7. The unconstrained parameters μ and $\log(\sigma)$ for the lognormal margins are estimated and transformed to the constrained ones. Two types

Downloaded from https://academic.oup.com/biomet/article/110/3/721/6947986 by Central University of Punjab user on 11 November 2023

Table 5. *Simulation results for the Gumbel copula in Scenario* 1

		$n = 200$						$n = 500$					
τ		μ_T	σ_T	μ_C	σ_C	θ	τ	μ_T	σ_T	μ_C	σ_C	θ	τ
0.2	aver.est	2.20	0.99	1.99	0.25	1.36	0.23	2.20	0.99	2.00	0.25	1.27	0.20
	sd.aver.est	0.11	0.10	0.04	0.02	0.35	0.15	0.07	0.06	0.02	0.01	0.16	0.09
	aver.asderr	0.11	0.10	0.04	0.02	0.28	0.14	0.07	0.06	0.02	0.01	0.15	0.09
	RMSE	0.11	0.10	0.04	0.02	0.36	0.15	0.07	0.06	0.02	0.01	0.16	0.09
0.5	aver.est	2.20	0.99	2.00	0.25	2.13	0.50	2.19	1.00	2.00	0.25	2.13	0.52
	sd.aver.est	0.12	0.10	0.04	0.02	0.59	0.13	0.07	0.06	0.02	0.01	0.38	0.08
	aver.asderr	0.11	0.10	0.03	0.02	0.56	0.13	0.07	0.06	0.02	0.01	0.36	0.08
	RMSE	0.12	0.10	0.04	0.02	0.60	0.13	0.07	0.06	0.02	0.01	0.40	0.08
0.7	aver.est	2.21	0.98	2.00	0.25	3.50	0.69	2.19	0.99	2.00	0.25	3.42	0.70
	sd.aver.est	0.11	0.10	0.03	0.02	0.91	0.08	0.06	0.06	0.02	0.01	0.55	0.05
	aver.asderr	0.10	0.09	0.03	0.02	0.90	0.08	0.06	0.06	0.02	0.01	0.55	0.05
	RMSE	0.11	0.10	0.03	0.02	0.92	0.08	0.06	0.06	0.02	0.01	0.56	0.05

aver.est, average estimate; sd.aver.est, empirical standard error of the average estimate; aver.asderr, average of the asymptotic standard error estimates for the parameter estimators; RMSE, empirical root mean squared error.

Table 6. *Simulation results for the Gauss copula in Scenario* 1

		$n = 200$						$n = 500$					
τ		μ_T	σ_T	μ_C	σ_C	θ	τ	μ_T	σ_T	μ_C	σ_C	θ	τ
0.2	aver.est	2.18	0.97	1.99	0.25	0.33	0.22	2.20	1.00	2.00	0.25	0.29	0.19
	sd.aver.est	0.11	0.10	0.04	0.02	0.24	0.17	0.07	0.06	0.03	0.01	0.18	0.12
	aver.asderr	0.12	0.10	0.05	0.02	0.29	0.20	0.07	0.06	0.03	0.01	0.20	0.13
	RMSE	0.11	0.11	0.04	0.02	0.25	0.18	0.07	0.06	0.03	0.01	0.18	0.12
0.5	aver.est	2.20	1.00	2.00	0.25	0.69	0.50	2.20	1.00	2.00	0.25	0.69	0.49
	sd.aver.est	0.12	0.10	0.04	0.02	0.15	0.13	0.07	0.06	0.02	0.01	0.09	0.08
	aver.asderr	0.11	0.10	0.04	0.02	0.15	0.13	0.07	0.06	0.02	0.01	0.10	0.09
	RMSE	0.12	0.10	0.04	0.02	0.15	0.12	0.07	0.06	0.02	0.01	0.09	0.08
0.7	aver.est	2.20	1.00	2.00	0.24	0.88	0.69	2.21	1.00	2.00	0.25	0.90	0.71
	sd.aver.est	0.09	0.08	0.02	0.02	0.06	0.07	0.07	0.05	0.02	0.01	0.03	0.05
	aver.asderr	0.10	0.09	0.03	0.02	0.07	0.08	0.06	0.06	0.02	0.01	0.03	0.04
	RMSE	0.09	0.08	0.02	0.02	0.06	0.07	0.07	0.05	0.02	0.01	0.03	0.05

aver.est, average estimate; sd.aver.est, empirical standard error of the average estimate; aver.asderr, average of the asymptotic standard error estimates for the parameter estimators; RMSE, empirical root mean squared error.

of standard error estimates were determined: one based on the asymptotic theory and one based on using 1000 parametric bootstrap samples with replacement. The parametric bootstrap data are generated by drawing pairs of survival and censoring times from the bivariate distribution $\mathcal{C}_{\hat{\theta}}\{F_{T,\hat{\theta}_T}(\cdot), F_{C,\hat{\theta}_C}(\cdot)\}$, and then keeping only the smallest of the two times and the censoring indicator. In view of the asymptotic theory in Theorem 5, we expect the bootstrap procedure to give valid standard error estimates. There is some disagreement between the asymptotic and bootstrap standard errors for the copula parameter and Kendall's τ estimates in the Clayton case. For this choice of copula, the mean of the largest 5% of values of the empirical Kendall's τ is 0.77, while the 95% empirical quantile is 0.17, which results in a large bootstrap standard error value for θ and τ. A similar, though much smaller, effect can be observed for the Gumbel copula. However, these outlying estimates have little effect on the corresponding bootstrap loglikelihood values.

Downloaded from https://academic.oup.com/biomet/article/110/3/721/6947986 by Central University of Punjab user on 11 November 2023

Table 7. *Estimated marginal and copula parameters and the fitted and average bootstrapped loglikelihoods for the SEER data for several copulas and lognormal margins, together with asymptotic and bootstrap standard error estimates*

Copula	μ_T	σ_T	μ_C	σ_C	θ	τ	loglikelihood	boot loglik
indep	3.25	1.92	3.34	1.31			−7312.37	−7311.00
asd	0.06	0.05	0.04	0.03				
bsd	0.07	0.05	0.04	0.03				
Frank	2.66	1.50	2.68	1.19	17.81	0.80	−7231.42	−7226.68
asd	0.04	0.04	0.04	0.02	3.10	0.03		
bsd	0.04	0.03	0.04	0.02	2.51	0.02		
Clayton	3.16	1.88	3.23	1.32	0.27	0.12	−7309.74	−7307.96
asd	0.07	0.05	0.07	0.03	0.15	0.06		
bsd	0.15	0.09	0.15	0.04	3.57	0.15		
Gumbel	2.53	1.42	2.59	1.17	4.65	0.78	−7203.12	−7203.59
asd	0.05	0.04	0.04	0.03	0.74	0.03		
bsd	0.04	0.02	0.03	0.02	0.29	0.01		
Gauss	2.57	1.47	2.57	1.23	0.95	0.80	−7251.10	−7249.11
asd	0.04	0.03	0.03	0.02	0.01	0.03		
bsd	0.04	0.02	0.03	0.02	0.00	0.01		

asd, asymptotic standard error estimate; bsd, bootstrap standard error estimate; boot loglik, average bootstrap loglikelihood.

Table 8. *Asymptotic correlation estimates between parameter estimates for the SEER data based on the Gumbel copula*

	μ_T	σ_T	μ_C	σ_C	θ
μ_T	1.00	0.41	0.94	−0.31	−0.55
σ_T	0.41	1.00	0.49	0.14	−0.55
μ_C	0.94	0.49	1.00	−0.44	−0.69
σ_C	−0.31	0.14	−0.44	1.00	0.66
θ	−0.55	−0.55	−0.69	0.66	1.00

Next we compare the copula specifications. Since the model complexities of these single-parameter copula models are the same, considering the maximized loglikelihood is sufficient. With regard to the copula specification, the model with the Gumbel copula has the highest loglikelihood. Comparing the dependent censoring models with the independent censoring one, we see a drop of more than 100 for the Gumbel specification. Using the likelihood-based test of Vuong (1989) for model comparison gives test statistics of 13.15, 4.48, 12.93 and 12.71 for testing preference of the Gumbel specification over the independence, Frank, Clayton and Gauss specifications, respectively. This yields p-values of less than 10^{-6} for all cases; therefore the Gumbel copula is the preferred copula model. For this specification, Table 8 reports estimates of the correlation between the parameter estimates based on the asymptotic theory. We see high dependence between the mean parameters for T and C. This indicates that upper tail dependence is present between the survival times and the censoring times, which means that long survival or censoring times occur much more often together than do short times. This conclusion is also supported by the worse performance of the Clayton copula-based model, which allows for lower tail dependence, but no upper tail dependence. The symmetric but no-tail-dependence Gaussian and Frank copula

Downloaded from https://academic.oup.com/biomet/article/110/3/721/6947986 by Central University of Punjab user on 11 November 2023

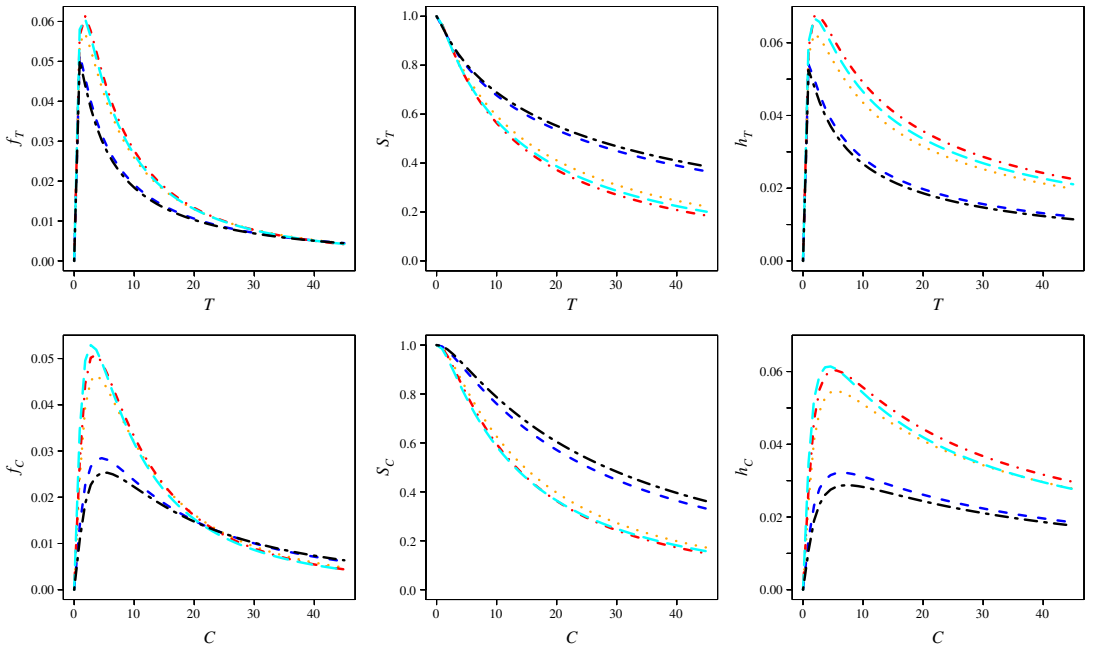

Fig. 1. Fitted density (left), survival function (middle) and hazard (right) of T (top) and C (bottom) for the SEER data based on several copulas and lognormal margins: Clayton copula (blue dashed), Frank copula (yellow dotted), Gumbel copula (red dot-dashed), Gauss copula (light blue long dashed) and independence copula (black dot-dashed).

models perform better than the Clayton copula, but worse than the Gumbel copula. The worse performance of the Frank copula relative to the Gauss copula can be attributed to the fact that the Frank copula has even lighter joint tails than does the Gaussian copula.

The fitted marginal probabilities $\text{pr}(\Delta = 1) = \text{pr}(T \leqslant C)$ are reported in the Supplementary Material. From these we see that the copula specifications give fitted marginal probabilities that are close to the empirical values. In the Supplementary Material it is shown that the fitted density of the censored observations has a heavier tail than that of the uncensored observations. Further, the Gumbel copula-based model is a little closer to the empirical density for both censored and uncensored observations. Finally, in Fig. 1 we display the fitted density of the survival times T and censoring times C under all copula specifications. The effect of the copula family is visible for both the survival and the censoring times. Moreover, the Gumbel and Gauss copulas yield very similar fits.

7. Discussion and future research

To the best of our knowledge, this work is the first to prove the identifiability of a copula model for dependent censoring, where the association parameter of the copula is not assumed to be known. This represents a major advantage of our model over existing copula models. To focus on this benefit, the setting of our study has been kept simple. Our model can, however, be extended in many ways. First, it can be made more flexible by considering semiparametric or nonparametric margins. The question is then whether the identifiability of the model can be guaranteed. Second, covariates can be added to the model. The simplest case is that of a fully parametric regression model for the survival and censoring times, but other models such as semiparametric Cox proportional hazards or accelerated failure

Downloaded from https://academic.oup.com/biomet/article/110/3/721/6947986 by Central University of Punjab user on 11 November 2023

time models would be worth studying as well. Third, more general survival models, such as competing risks models, cure models, administrative censoring and truncation, used in combination with dependent censoring, also look very promising. The important question of the effect of misspecification will also be pursued in future work. Another topic for further research is the development of goodness-of-fit tests for the proposed model. This could involve developing a test statistic that measures the L_2 distance between two estimators of the distribution of $\min(T, C)$, namely a nonparametric one, such as the empirical distribution, and the parametric estimator obtained under our dependent censoring model. If the model is correct, the two estimators should be similar.

ACKNOWLEDGEMENT

Czado acknowledges support from the German Science Foundation. Van Keilegom acknowledges support from the European Research Council. The authors thank Johanna Neslehova for very helpful discussions and the editor, associate editor and two referees for their valuable comments and suggestions.

SUPPLEMENTARY MATERIAL

The Supplementary Material contains further simulations and information on the data analysis.

APPENDIX

Proof of (4). Upon differentiating (2) we can express the joint density of (T, C) as

$$f_{T,C}(t, c) = c\{F_T(t), F_C(c)\}f_T(t)f_C(c), \tag{A1}$$

where c is the copula density. We are interested in determining the conditional distribution of T given C and vice versa. From (A1) it is straightforward to see that the conditional densities are

$$f_{T|C}(t \mid c) = c\{F_T(t), F_C(c)\}f_T(t),$$
$$f_{C|T}(c \mid T) = c\{F_T(t), F_C(c)\}f_C(c),$$

and the conditional distribution function of T given $C = c$ can be derived as

$$\begin{aligned} F_{T|C}(t \mid c) &= \int_0^t c\{F_T(t^*), F_C(c)\}f_T(t^*)\,\mathrm{d}t^* \\ &= \int_0^t \frac{\partial^2}{\partial u \partial v}\mathcal{C}(u, v)\Big|_{u=F_T(t^*), v=F_C(c)} \frac{\mathrm{d}F_T(t)}{\mathrm{d}t}\Big|_{t=t^*} \mathrm{d}t^* \\ &= \frac{\partial}{\partial v}\mathcal{C}(u, v)\Big|_{u=F_T(t), v=F_C(c)} = h_{T|C}\{F_T(t) \mid F_C(c)\}. \end{aligned}$$

Similarly we have $F_{C|T}(c \mid t) = h_{C|T}\{F_C(c) \mid F_T(t)\}$. We now derive the marginal distribution of $Y = \min(T, C)$:

$$\begin{aligned} F_Y(y) &= 1 - \mathrm{pr}(Y > y) = 1 - \mathrm{pr}(T > y, C > y) \\ &= 1 - \big[1 - F_C(y) - F_T(y) + \mathcal{C}\{F_T(y), F_C(y)\}\big] \\ &= F_C(y) + F_T(y) - \mathcal{C}\{F_T(y), F_C(y)\}. \end{aligned}$$

Downloaded from https://academic.oup.com/biomet/article/110/3/721/6947986 by Central University of Punjab user on 11 November 2023

Finally, we derive the expressions for the joint mixed density $f_{Y,\Delta}$ by observing that

$$F_{Y,\Delta}(y,1) = \mathrm{pr}(T \leqslant y, T \leqslant C) = \int_0^y \mathrm{pr}(C \geqslant t \mid T = t) f_T(t)\,\mathrm{d}t$$
$$= \int_0^y \big[1 - h_{C|T}\{F_C(t) \mid F_T(t)\}\big] f_T(t)\,\mathrm{d}t,$$

and hence $f_{Y,\Delta}(y,1) = [1 - h_{C|T}\{F_C(y) \mid F_T(y)\}]f_T(y)$. In a similar manner we obtain $f_{Y,\Delta}(y,0) = [1 - h_{T|C}\{F_T(y) \mid F_C(y)\}]f_C(y)$. □

Proof of Theorem 1. Recall from (4) that

$$\mathrm{pr}(Y \leqslant t, \Delta = 1) = \mathrm{pr}(T \leqslant t, T \leqslant C) = \int_0^t \mathrm{pr}(C \geqslant y \mid T = y) f_T(y)\,\mathrm{d}y.$$

Hence $f_{Y,\Delta}(t,1) = \{1 - F_{C|T}(t \mid t)\}f_T(t) = [1 - h_{C|T}\{F_C(t) \mid F_T(t)\}]f_T(t)$. From condition (ii) we know that $\lim_{t\to a} h_{C|T}\{F_C(t) \mid F_T(t)\} = 0$ for $a = 0$ or $a = \infty$. Hence $\lim_{t\to a} f_{Y,\Delta}(t,1) = \lim_{t\to a} f_T(t)$. Suppose now that $f_{Y,\Delta,\alpha_1}(t,1) = f_{Y,\Delta,\alpha_2}(t,1)$ for all t, where $\alpha_j = (\theta_j, \theta_{Tj}, \theta_{Cj})^{\mathrm{T}}$ $(j = 1, 2)$. Then

$$1 = \lim_{t\to a} \frac{f_{Y,\Delta,\alpha_1}(t,1)}{f_{Y,\Delta,\alpha_2}(t,1)} = \lim_{t\to a} \frac{f_{T,\theta_{T1}}(t)}{f_{T,\theta_{T2}}(t)}.$$

It follows from condition (i) that $\theta_{T1} = \theta_{T2}$. In a similar way we can show that $\theta_{C1} = \theta_{C2}$. Finally, to show that $\theta_1 = \theta_2$, notice that $F_{Y,\alpha_j}(t) = F_{T,\theta_{T1}}(t) + F_{C,\theta_{C1}}(t) - \mathcal{C}_{\theta_j}\{F_{T,\theta_{T1}}(t), F_{C,\theta_{C1}}(t)\}$ and $F_{Y,\alpha_1}(t) = F_{Y,\alpha_2}(t)$ for all t. Therefore, since the copula is unique, it follows that $\theta_1 = \theta_2$. □

Proof of Theorem 2. Consider first the lognormal density for T depending on $\theta_T = (\mu, \sigma)$:

$$\lim_{t\to 0} \frac{f_{T,\mu_1,\sigma_1}(t)}{f_{T,\mu_2,\sigma_2}(t)} = \lim_{t\to 0} \frac{\frac{1}{t\sigma_1}\phi\big(\frac{\log t - \mu_1}{\sigma_1}\big)}{\frac{1}{t\sigma_2}\phi\big(\frac{\log t - \mu_2}{\sigma_2}\big)} = \lim_{t'\to -\infty} \frac{\sigma_2}{\sigma_1} \frac{\phi\big(\frac{t' - \mu_1}{\sigma_1}\big)}{\phi\big(\frac{t' - \mu_2}{\sigma_2}\big)}.$$

It is easily seen that this can equal 1 only when $\mu_1 = \mu_2$ and $\sigma_1 = \sigma_2$. The same is true when taking the limit as $t \to \infty$.

Similarly, for the log-Student-t density, we have

$$\lim_{t\to 0,\infty} \frac{f_{T,\nu_1,\mu_1,\sigma_1}(t)}{f_{T,\nu_2,\mu_2,\sigma_2}(t)} = \frac{c_{\nu_1,\sigma_1}}{c_{\nu_2,\sigma_2}} \lim_{t\to 0,\infty} \frac{\big\{1 + \frac{1}{\nu_1}\big(\frac{\log t - \mu_1}{\sigma_1}\big)^2\big\}^{-(\nu_1+1)/2}}{\big\{1 + \frac{1}{\nu_2}\big(\frac{\log t - \mu_2}{\sigma_2}\big)^2\big\}^{-(\nu_2+1)/2}}, \tag{A2}$$

where $c_{\nu,\sigma} = \Gamma\{(\nu+1)/2\}/\{\sigma(\nu\pi)^{1/2}\Gamma(\nu/2)\}$, with Γ denoting the gamma function. It can easily be seen that the limit in (A2) is equal to 1 if and only if $(\nu_1, \mu_1, \sigma_1) = (\nu_2, \mu_2, \sigma_2)$.

This is also the case for the Weibull density, since

$$\lim_{t\to 0,\infty} \frac{f_{T,\lambda_1,\rho_1}(t)}{f_{T,\lambda_2,\rho_2}(t)} = \frac{\lambda_1\rho_1}{\lambda_2\rho_2} \lim_{t\to 0,\infty} \frac{t^{\rho_1 - 1}\exp(-\lambda_1 t^{\rho_1})}{t^{\rho_2 - 1}\exp(-\lambda_2 t^{\rho_2})}, \tag{A3}$$

and this equals 1 only if $\rho_1 = \rho_2$ and $\lambda_1 = \lambda_2$.

Finally, for the log-logistic density, the limit is

$$\lim_{t\to 0,\infty} \frac{f_{T,\lambda_1,\kappa_1}(t)}{f_{T,\lambda_2,\kappa_2}(t)} = \frac{\kappa_1\lambda_1^{\kappa_1}}{\kappa_2\lambda_2^{\kappa_2}} \lim_{t\to 0,\infty} \frac{t^{\kappa_1 - 1}\{1 + (\lambda_2 t)^{\kappa_2}\}^2}{t^{\kappa_2 - 1}\{1 + (\lambda_1 t)^{\kappa_1}\}^2},$$

Downloaded from https://academic.oup.com/biomet/article/110/3/721/6947986 by Central University of Punjab user on 11 November 2023

and again this can only equal 1 when $\kappa_1 = \kappa_2$ and $\lambda_1 = \lambda_2$. Hence, condition (i) is satisfied for each of these densities. □

Proof of Lemma 1. We use the abbreviated notation $u_t = F_{T,\theta_T}(t)$ and $v_t = F_{C,\theta_C}(t)$. Note that $\psi(1) = 0$, $\lim_{t\to 0} \psi^{-1}(t) = 1$ and $\lim_{u\to 1} \psi'(u) = c \in (-\infty, 0)$. Hence

$$\lim_{t\to\infty} h_{T|C,\theta}(u_t \mid v_t) = \lim_{t\to\infty} \frac{\psi'(v_t)}{\psi'\big[\psi^{-1}\{\psi(u_t) + \psi(v_t)\}\big]} = \frac{c}{c} = 1.$$

□

Proof of Theorem 3. We start with the Frank copula. Straightforward calculations show that $\lim_{u\to 1} \psi'(u) = \theta \exp(-\theta)/\{\exp(-\theta) - 1\} < 0$ for $\theta \neq 0$, and hence we know from Lemma 1 that $\lim_{t\to\infty} h_{T|C}\{F_T(t) \mid F_C(t)\} = 1$. Thus, the first part of condition (ii) can be satisfied only if $\lim_{t\to 0} h_{T|C}\{F_T(t) \mid F_C(t)\} = 0$. Some straightforward but tedious calculations show that

$$\begin{aligned}\lim_{t\to 0} h_{T|C}\{F_T(t) \mid F_C(t)\} &= \lim_{t\to 0} \frac{\psi'\{F_C(t)\}}{\psi'\big(\psi^{-1}\big[\psi\{F_T(t)\} + \psi\{F_C(t)\}\big]\big)} \\ &= \lim_{t\to 0} \frac{\exp\{-\theta F_C(t)\}\big[\exp\{-\theta F_T(t)\} - 1\big]}{\big[\exp\{-\theta F_C(t)\} - 1\big]\big[\exp\{-\theta F_T(t)\} - 1\big] + \exp(-\theta) - 1} = 0\end{aligned}$$

for $\theta \neq 0$. Hence, the first part of condition (ii) is satisfied. Similarly it can be shown that the second part is satisfied.

For the Gumbel family we have that $\lim_{u\to 1} \psi'(u) = 0$, and hence Lemma 1 is not applicable. Therefore, we calculate the limit as t tends to 0 and to infinity:

$$\begin{aligned}\lim_{t\to 0,\infty} h_{T|C}\{F_T(t) \mid F_C(t)\} = \lim_{t\to 0,\infty} &\Big[1 + \{-\log F_C(t)\}^{-\theta}\{-\log F_T(t)\}^{\theta}\Big]^{-1+1/\theta} \\ &\times \lim_{t\to 0,\infty} \exp\Big(-\big[\{-\log F_T(t)\}^{\theta} + \{-\log F_C(t)\}^{\theta}\big]^{1/\theta} - \log F_C(t)\Big);\end{aligned}$$

see Aas et al. (2009) for the formula for $h_{T|C}\{F_T(t) \mid F_C(t)\}$ for the Gumbel family. Under the assumption that $\log F_T(t)/\log F_C(t) \to c$ for some $0 < c < \infty$ as t tends to 0, the exponential factor above tends to 0 as t tends to 0 and tends to 1 as t tends to infinity, whereas the factor in front of this exponential factor converges to some constant in the interval [0, 1], depending on the limit of $\log F_T(t)/\log F_C(t)$ for t tending to 0 or infinity. This shows that the product of the two limits equals 0 as t tends to 0, whereas when t tends to infinity the limit can be zero, but may also be strictly positive, depending on the limit of $\log F_T(t)/\log F_C(t)$ for t growing to infinity.

Finally, we consider the Gaussian copula. Observe that

$$\mathrm{pr}\big[\Phi^{-1}\{F_T(T)\} \leqslant t,\ \Phi^{-1}\{F_C(C)\} \leqslant c\big] = \mathrm{pr}\big[T \leqslant F_T^{-1}\{\Phi(t)\},\ C \leqslant F_C^{-1}\{\Phi(c)\}\big] = \Phi_\theta(t, c),$$

and hence $\Phi^{-1}\{F_T(T)\} \mid \Phi^{-1}\{F_C(C)\} \sim N[\theta\Phi^{-1}\{F_C(C)\}, 1 - \theta^2]$. It follows that, omitting the parameters θ, θ_T and θ_C for simplicity,

$$\begin{aligned}h_{T|C}\{F_T(t) \mid F_C(t)\} = F_{T|C}(t \mid t) &= \mathrm{pr}\Big[\Phi^{-1}\{F_T(T)\} \leqslant \Phi^{-1}\{F_T(t)\} \,\Big|\, \Phi^{-1}\{F_C(C)\} = \Phi^{-1}\{F_C(t)\}\Big] \\ &= \Phi\left[\frac{\Phi^{-1}\{F_T(t)\} - \theta\Phi^{-1}\{F_C(t)\}}{(1-\theta^2)^{1/2}}\right] = \Phi\left\{\frac{A_{\theta,F_T,F_C}(t)}{(1-\theta^2)^{1/2}}\right\}.\end{aligned}$$

Downloaded from https://academic.oup.com/biomet/article/110/3/721/6947986 by Central University of Punjab user on 11 November 2023

Since $A_{\theta,F_T,F_C}(t)$ tends to $-\infty$ as t tends to either 0 or ∞, it follows that $h_{T|C}\{F_T(t) \mid F_C(t)\}$ tends to 0 as t tends to either 0 or ∞. Thus, the first part of condition (ii) is satisfied. In a similar way we can prove the second part. □

Proof of Theorem 4. Suppose that $\lim_{t\to 0} F_{T,\theta_T}(t)/F_{C,\theta_C}(t) = \infty$; the case where the limit equals 0 can be handled similarly. Then, it follows from (6) that $\lim_{t\to 0} h_{C|T,\theta}\{F_{C,\theta_C}(t) \mid F_{T,\theta_T}(t)\} = 0$. Hence, using similar arguments to those in the proof of Theorem 1, it follows from condition (i) that θ_T is identifiable. From the formula of $f_{Y,\Delta}(\cdot, 1)$ given in (4) it then follows that the function $t \to h_{C|T,\theta}\{F_{C,\theta_C}(t) \mid F_{T,\theta_T}(t)\}$ is identifiable.

Next, for t large enough, $F_T(t)^\theta/F_{C,\theta_C}(t)^\theta - F_T(t)^\theta$ is close to zero, where we have omitted θ_T since it is identifiable, and hence a Taylor expansion can be used to write

$$\begin{aligned}\log h_{C|T,\theta}\{F_{C,\theta_C}(t) \mid F_T(t)\} &= -\frac{\theta+1}{\theta}\log\left\{1 + \frac{F_T(t)^\theta}{F_{C,\theta_C}(t)^\theta} - F_T(t)^\theta\right\} \\ &= -\frac{\theta+1}{\theta}\sum_{k=1}^{\infty}\frac{(-1)^{k-1}}{k}\left[F_T(t)^\theta\{F_{C,\theta_C}(t)^{-\theta} - 1\}\right]^k\end{aligned}$$

for t large. This is a polynomial in $u_t = F_T(t)$. Therefore, for two sets of parameters (θ, θ_C) and (θ^*, θ_C^*) we have that

$$\frac{\theta+1}{\theta}\sum_{k=1}^{\infty}\frac{(-1)^{k-1}}{k}\left\{F_{C,\theta_C}(t)^{-\theta} - 1\right\}^k u_t^{\theta k} = \frac{\theta^*+1}{\theta^*}\sum_{k=1}^{\infty}\frac{(-1)^{k-1}}{k}\left\{F_{C,\theta_C^*}(t)^{-\theta^*} - 1\right\}^k u_t^{\theta^* k}$$

for t large, and this is only possible if $\theta = \theta^*$ and $\theta_C = \theta_C^*$. □

References

Aas, K., Czado, C., Frigessi, A. & Bakken, H. (2009). Pair-copula constructions of multiple dependence. *Insur. Math. Econ.* **44**, 182–98.

Basu, A. P. (1988). Multivariate exponential distributions and their applications in reliability. In *Quality Control and Reliability*, P. R. Krishnaiah & C. R. Rao, eds., vol. 7 of *Handbook of Statistics*. Amsterdam: Elsevier, pp. 467–77.

Basu, A. P. & Ghosh, J. K. (1978). Identifiability of the multinormal and other distributions under competing risks model. *J. Mult. Anal.* **8**, 413–29.

Braekers, R. & Veraverbeke, N. (2005). A copula-graphic estimator for the conditional survival function under dependent censoring. *Can. J. Statist.* **33**, 429–47.

Chen, Y.-H. (2010). Semiparametric marginal regression analysis for dependent competing risks under an assumed copula. *J. R. Statist. Soc.* B **72**, 235–51.

Collett, D. (2015). *Modelling Survival Data in Medical Research*. Boca Raton, Florida: Taylor & Francis, 3rd ed.

de Uña-Álvarez, J. & Veraverbeke, N. (2017). Copula-graphic estimation with left-truncated and right-censored data. *Statistics* **51**, 387–403.

Deresa, N. W. & Van Keilegom, I. (2020a). Flexible parametric model for survival data subject to dependent censoring. *Biomet. J.* **62**, 136–56.

Deresa, N. W. & Van Keilegom, I. (2020b). A multivariate normal regression model for survival data subject to different types of dependent censoring. *Comp. Statist. Data Anal.* **144**, 106879.

Deresa, N. W. & Van Keilegom, I. (2021). On semiparametric modelling, estimation and inference for survival data subject to dependent censoring. *Biometrika* **108**, 965–79.

Emoto, S. E. & Matthews, P. C. (1990). A Weibull model for dependent censoring. *Ann. Statist.* **18**, 1556–77.

Emura, T. & Chen, Y.-H. (2016). Gene selection for survival data under dependent censoring: A copula-based approach. *Statist. Meth. Med. Res.* **25**, 2840–57.

Emura, T. & Chen, Y.-H. (2018). *Analysis of Survival Data with Dependent Censoring: Copula-Based Approaches*. Singapore: Springer.

Downloaded from https://academic.oup.com/biomet/article/110/3/721/6947986 by Central University of Punjab user on 11 November 2023

Escarela, G. & Carriere, J. F. (2003). Fitting competing risks with an assumed copula. *Statist. Meth. Med. Res.* **12**, 333–49.
Fan, T.-H., Wang, Y.-F. & Ju, S.-K. (2019). A competing risks model with multiply censored reliability data under multivariate Weibull distributions. *IEEE Trans. Reliab.* **68**, 462–75.
Hsu, C., Taylor, J. & Hu, C. (2015). Analysis of accelerated failure time data with dependent censoring using auxiliary variables via nonparametric multiple imputation. *Statist. Med.* **34**, 2768–80.
Huang, X. & Zhang, N. (2008). Regression survival analysis with an assumed copula for dependent censoring: A sensitivity analysis approach. *Biometrics* **64**, 1090–9.
Jackson, D., White, I. R., Seaman, S., Evans, H., Baisley, K. & Carpenter, J. (2014). Relaxing the independent censoring assumption in the Cox proportional hazards model using multiple imputation. *Statist. Med.* **33**, 4681–94.
Kaplan, E. L. & Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Am. Statist. Assoc.* **53**, 457–81.
Nádas, A. (1971). The distribution of the identified minimum of normal pair determines the distribution of the pair. *Technometrics* **13**, 201–2.
Rivest, L. P. & Wells, M. T. (2001). A martingale approach to the copula-graphic estimator for the survival function under dependent censoring. *J. Mult. Anal.* **79**, 138–55.
Scharfstein, D. O. & Robins, J. M. (2002). Estimation of the failure time distribution in the presence of informative censoring. *Biometrika* **89**, 617–34.
Schwarz, M., Jongbloed, G. & Van Keilegom, I. (2013). On the identifiability of copulas in bivariate competing risks models. *Can. J. Statist.* **41**, 291–303.
Shih, J.-H., Lee, W., Sun, L.-H. & Emura, T. (2019). Fitting competing risks data to bivariate Pareto models. *Commun. Statist.* A **48**, 1193–220.
Sklar, M. (1959). Fonctions de répartition à *n* dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris* **8**, 229–31.
Sujica, A. & Van Keilegom, I. (2018). The copula-graphic estimator in censored nonparametric location-scale regression models. *Economet. Statist.* **7**, 89–114.
Tsiatis, A. (1975). A nonidentifiability aspect of the problem of competing risks. *Proc. Nat. Acad. Sci.* **72**, 20–2.
Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* **57**, 307–33.
White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–25.
Zheng, M. & Klein, J. P. (1995). Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika* **82**, 127–38.

[*Received on* 25 *May* 2021. *Editorial decision on* 29 *November* 2022]

Downloaded from https://academic.oup.com/biomet/article/110/3/721/6947986 by Central University of Punjab user on 11 November 2023