Universidade Vigo

Generalized copula-graphic estimator with left-truncated and right-censored data

Jacobo de Uña-Álvarez and Noël Veraverbeke

Report 14/01

Discussion Papers in Statistics and Operation Research

Departamento de Estatística e Investigación Operativa

Facultade de Ciencias Económicas e Empresariales
Lagoas-Marcosende, s/n · 36310 Vigo
Tfno.: +34 986 812440 - Fax: +34 986 812401
http://webs.uvigo.es/depc05/
E-mail: depc05@uvigo.es

Universidade Vigo

Generalized copula-graphic estimator with left-truncated and right-censored data

Jacobo de Uña-Álvarez and Noël Veraverbeke

Report 14/01

Discussion Papers in Statistics and Operation Research

Imprime: GAMESAL

Edita: Universidade Vigo

Facultade de CC. Económicas e Empresariales
Departamento de Estatística e Investigación Operativa
As Lagoas Marcosende, s/n 36310 Vigo
Tfno.: +34 986 812440

I.S.S.N: 1888-5756

Depósito Legal: VG 1402-2007

Generalized copula-graphic estimator with left-truncated and right-censored data

Jacobo de Uña-Álvarez (U. Vigo, Spain), Noël Veraverbeke (U. Hasselt, Belgium; Unit for BMI, North-West University, Potchefstroom, South Africa)

April 2014

Abstract

In this paper a copula-graphic estimator is proposed for left-truncated and right-censored survival data. It is assumed that there is some dependent censoring acting on the variable of interest, which may come from an existing competing risk. Furthermore, the full process is independently right-censored by some administrative censoring time, while there is an independent left-truncation variable which complicates the sampling procedure. The dependent censoring is modeled through an Archimedean copula function, which is supposed to be known. An asymptotic representation of the estimator as a sum of independent and identically distributed random variables is obtained and, consequently, a central limit theorem is established. These results extend to the truncated setting those in de Uña-Álvarez and Veraverbeke (2013). We investigate the finite sample performance of the estimator through simulations. A real data illustration is included.

Keywords: Almost sure representation; Archimedean copula; Cross-sectional data; Dependent censoring; Survival analysis

1 Introduction

Consider a situation in which two random variables Y and C censor each other. This occurs, for example, when Y and C represent the time up to event 1 and 2, respectively, in a competing risks model. In this situation, only one of the two events is observed, and the available information is given by $Z = \min(Y, C)$ and the event indicator $\delta = I(Y \leq C)$, which takes the value 1 when event 1 occurs ($\delta = 0$ otherwise). Tsiatis (1975) demonstrated that the marginal distribution functions F and G of Y and C cannot be identified, unless some information on the dependence structure between Y and C is available.

Assume that there exists a known Archimedean copula $\mathcal{C}(u_1, u_2)$ which relates the joint survival function of (Y, C) to the marginal survival functions $\overline{F}(t) = 1 - F(t)$ and $\overline{G}(t) = 1 - G(t)$:

$$P(Y > t_1, C > t_2) = \phi^{-1}(\phi(\overline{F}(t_1)) + \phi(\overline{G}(t_2))).$$

The function $\phi :]0, 1] \to [0, \infty[$ is called the generator of the copula $\mathcal{C}$. It is a known continuous, convex, strictly decreasing function with $\phi(1) = 0$. The variables Y and C are independent in the particular case $\phi(t) = -\ln t$; in that case, the Kaplan-Meier method provides consistent estimators of the marginal

distribution functions. In general, a broad family of generators have been used to model dependent random variables, see Nelsen (2006). Zheng and Klein (1995), see also Rivest and Wells (2001), introduced a nonparametric estimator for $\overline{F}(t)$, termed copula-graphic estimator, generalizing the product-limit Kaplan-Meier estimator to the dependent scenario. Their estimator, however, requires the direct observation of the pair (Z, δ), which is not always possible. This may be due to limitations in the follow-up period for the subjects, losses unrelated to the competing risks of interest, and so on. To overcome this issue, de Uña-Álvarez and Veraverbeke (2013) proposed a generalized copula-graphic estimator, by considering the presence of an independent censoring time.

To fix ideas, and to motivate the present work, introduce a potential censoring time D independent of (Z, δ). Rather than (Z, δ) we observe $(U, \rho, \rho\delta)$ where $U = \min(Z, D)$ and $\rho = I(Z \leq D)$; note that the value of δ (i.e. the event type) is observed only when Z is uncensored ($\rho = 1$). We put $\widetilde{G}$ for the distribution function of D. Denote $H(t) = P(Z \leq t)$, $\overline{H}(t) = 1 - H(t)$, and $H^1(t) = P(Z \leq t, \delta = 1)$. Then, if ϕ' exists and if H^1 is differentiable, we have from Tsiatis (1975)

$$\overline{F}(t) = \phi^{-1}\left(-\int_0^t \phi'(\overline{H}(s))dH^1(s)\right). \tag{1}$$

From this equation, de Uña-Álvarez and Veraverbeke (2013) introduced an estimator of $\overline{F}(t)$ by plugging in proper estimators for H and H^1, based on the observed values of $(U, \rho, \rho\delta)$. They also provided an asymptotic representation of the estimator as a sum of i.i.d. random variables. When $P(D = \infty) = 1$, the proposed estimator reduces to that in Rivest and Wells (2001).

A limitation of de Uña-Álvarez and Veraverbeke (2013)'s generalized copula-graphic estimator is that it does not take possible truncation effects into account. Random left-truncation often occurs in the field of survival analysis; delay entries or cross-sectional sampling schemes provide left-truncated data indeed. Therefore, we put T for a left-truncating variable (independent of Z) such that individual information is available only when $T \leq U$. The sample information is represented by $(T_i, U_i, \rho_i, \rho_i\delta_i)$, where $T_i \leq U_i$, $i = 1, ..., n$. That is, each $(T_i, U_i, \rho_i, \rho_i\delta_i)$ follows the conditional distribution of $(T, U, \rho, \rho\delta)$ given $T \leq U$. Ignoring left-truncation effects leads to a systematic underestimation of H, invalidating the methods proposed in the mentioned paper. In this work we adapt de Uña-Álvarez and Veraverbeke (2013)'s estimator to the presence of left-truncation and we perform the corresponding asymptotic analysis. The new estimator can be regarded as an extension of Tsai-Jewell-Wang (TJW) estimator (Tsai et al., 1987) for dependently censored data.

The rest of the paper is organized as follows. In Section 2 we introduce the estimator and we establish an almost sure asymptotic representation. In Section 3 we investigate the finite-sample performance of the estimator in a simulation

study. Section 4 gives an illustration of the method through the analysis of unemployment data. Main conclusions are reported in Section 5. Some needed lemmas and their proofs are given in the Appendix.

2 The estimator: main results

In this section we introduce the new estimator of F from equation (1). For that, we replace H in that equation by the TJW estimator for left-truncated and right-censored data, given by (assuming no ties)

$$1 - H_n(t) = \prod_{i=1}^{n} \left[1 - \frac{1}{nC_n(U_i)}\right]^{\rho_i I(U_i \leq t)}$$

where

$$C_n(t) = \frac{1}{n} \sum_{i=1}^{n} I(T_i \leq t \leq U_i)$$

is the 'proportion of individuals at risk' at time t. See Tsai et al. (1987). Note that this estimator behaves consistently since both T and D are assumed to be independent of Z. It can also be expressed as

$$H_n(t) = \sum_{i=1}^{n} W_{in} I(U_i \leq t)$$

where W_{in} is the TJW weight attached to U_i, which is given by

$$W_{in} = \frac{\rho_i}{nC_n(U_i)} \prod_{j=1}^{n} \left[1 - \frac{1}{nC_n(U_j)}\right]^{\rho_j I(U_j < U_i)}.$$

Asymptotic results for H_n were derived in a number of papers, see e.g. Zhou and Yip (1999) and references therein.

To estimate H^1 in (1) we proceed as in de Uña-Álvarez and Veraverbeke (2013); we consider δ_i as a 'covariate' for the possibly censored lifetime U_i. Following Sánchez-Sellero et al. (2005), we have that

$$H_n^1(t) = \sum_{i=1}^{n} W_{in} I(U_i \leq t, \delta_i = 1)$$

is an estimator for $H^1(t) = P(Z \leq t, \delta = 1)$. Since $W_{in} = 0$ whenever $\rho_i = 0$, we may write

$$H_n^1(t) = \sum_{i=1}^{n} W_{in} I(U_i \leq t, \rho_i \delta_i = 1)$$

which demonstrates that $H_n^1(t)$ can be constructed from the observed values $(T_i, U_i, \rho_i, \rho_i\delta_i)$, $i = 1, ..., n$. The estimator $H_n^1(t)$ can be seen as an adaptation to left-truncation of the empirical cumulative incidence function in a competing risks model, cfr. Kalbfleisch and Prentice (1980), p. 169, eq. (7.10).

For a general covariate vector X, Sánchez-Sellero et al. (2005) derived an almost sure representation of the weighted mean $\sum_{i=1}^n W_{in}\varphi(U_i, X_i)$ as a sum of i.i.d. random variables plus a remainder. Here, φ denotes an arbitrary (although known) real-valued function. From this result, asymptotic properties of $H_n^1(t)$ (such as strong consistency and distributional convergence to a normal) are easily obtained, when taking the special covariate $X = \delta$ and a particular indicator function for φ. The mentioned representation is crucial for our main result below. For identifiability reasons, Sánchez-Sellero et al. (2005) assumed (besides the needed support conditions) (i) the independence between (T, D) and (X, Z) and (ii) the independence between T and D. Condition (i) in our setting states the independence between (T, D) and (Z, δ), which holds provided that (T, D) is independent of the 'underlying process' (Y, C). In applications, this assumption will be realistic when the observation procedure is unrelated to the event times under investigation. Condition (ii) is not so clearly justified in practice; indeed, with cross-sectional sampling we often have $D = T + \tau$ for a certain constant τ which represents the maximum follow-up time from interception. Interestingly, this condition (ii) is not critical for the consistency of H_n^1, as it will be discussed later. See also our simulation results in Section 3.

The copula-graphic estimator for $\overline{F}(t)$ adapted to right-censoring and left-truncation is thus given by

$$\overline{F}_n(t) = \phi^{-1}\left(-\int_0^t \phi'(\overline{H}_n(s))dH_n^1(s)\right) \tag{2}$$

where $\overline{H}_n = 1 - H_n$. When the left-truncation is removed, the estimator (2) reduces to that in de Uña-Álvarez and Veraverbeke (2013). In the special case of no independent censoring ($D = \infty$), W_{in} is just the jump of the Lynden-Bell estimator for left-truncated data at time U_i (see e.g. Woodroofe, 1985), and therefore $\overline{F}_n(t)$ may be regarded as an adaptation of that estimator for dependently censored data. Indeed, if besides Y and C are independent ($\phi(t) = -\log t$), $\overline{F}_n(t)$ becomes the TJW estimator based on observations of (T, Z, δ). Equation (2) also leads to the TJW estimator in absence of dependent censoring ($Z = Y$, $\delta = 1$), based on observations of (T, U, ρ).

In order to formalize our assumptions and the main result, further notation is needed. For any distribution function K we put $a_K = \inf\{t : K(t) > 0\}$ and $b_K = \sup\{t : K(t) < 1\}$ for the lower and upper limits of the support of K. Also, the following functions will appear: $L(t) = P(T \leq t)$, $\widetilde{H}(t) = P(U \leq t)$,

$$C(t) = P(T \leq t \leq U | T \leq U) = \alpha^{-1}P(T \leq t \leq D)(1 - H(t))$$

with $\alpha = P(T \leq U) > 0$,

$$\widetilde{H}^1(t) = P(U \leq t, \rho = 1 | T \leq U) = \alpha^{-1} \int_0^t P(T \leq s \leq D) dH(s),$$

where we have used the independence between (T, D) and Z. The empirical version of $C(t)$ is the quantity $C_n(t)$ introduced above, while the empirical of $\widetilde{H}^1(t)$ is just

$$\widetilde{H}^1_n(t) = \frac{1}{n} \sum_{i=1}^{n} I(U_i \leq t, \rho_i = 1).$$

We prove an almost sure asymptotic representation for (2) with a uniform rate for the remainder. Put $a_{\widetilde{H}} = \min(a_F, a_G, a_{\widetilde{G}})$ and $b_{\widetilde{H}} = \min(b_F, b_G, b_{\widetilde{G}})$. Put $F_n = 1 - \overline{F}_n$. We will refer to the following conditions, where b is such that $a_{\widetilde{H}} \leq b < b_{\widetilde{H}}$:

(C1) F, G, L and $\widetilde{G}$ are continuous
(C2) (i) (T, D) is independent of (Y, C), and (ii) T and D are independent
(C3) H and H^1 have continuous first and second derivatives in $\left]a_{\widetilde{H}}, b\right]$
(C4) The copula generator ϕ has three continuous derivatives in $]0, 1]$ and $\phi'''(t) \leq 0$ for $t \in]0, 1]$
(C5) $a_L \leq a_{\widetilde{H}}$
(C6) $\int_{a_{\widetilde{H}}}^{b} C(t)^{-3} d\widetilde{H}^1(t) < \infty$

Conditions (C1)-(C4) here reduce to those considered in de Uña-Álvarez and Veraverbeke (2013) when there is no truncation. In the presence of truncation, condition (C5) ensures identifiability of the distribution of interest; note that, when $a_L > a_{\widetilde{H}}$, relevant information on F is missing due to left-truncation. Assumption (C6) was used by Zhou and Yip (1999) to obtain an almost sure uniform rate of convergence for the remainder in their Theorem 2.1. Basically, it controls the behavior of the truncation variable near the lower endpoint of $\widetilde{H}$. This condition (C6) is equivalent to

$$E\left[\Lambda(Z)^{-2} I(a_{\widetilde{H}} \leq Z \leq b)\right] < \infty$$

where $\Lambda(z) = P(T \leq z \leq D)$, which is enough for the purpose of applying Theorem 1 in Sánchez-Sellero et al. (2005) for the special family of functions $\varphi_t(d, u) = I(u \leq t)\phi'(\overline{H}(u))d$ with $a_{\widetilde{H}} \leq t \leq b$ (see the proof below). Finally, assumption (C2)(ii) was used in Sánchez-Sellero et al. (2005), Lemma 1, to get $\sup_{1 \leq i \leq n} C(U_i)/C_n(U_i) = O(\log n)$ almost surely. However, the independence between T and D may be removed as long as the function $C(t)$ remains bounded away from zero, since $\sup_{1 \leq i \leq n} C(U_i)/C_n(U_i) = O(1)$ almost surely holds in that case. In our real data illustration of Section 4, $D = T + \tau$ for a certain constant τ, and hence $\inf_{a_{\widetilde{H}} \leq t \leq b} C(t) > 0$ holds provided that $a_L < a_{\widetilde{H}}$, which is just slightly stronger than (C5) in this case. This demonstrates that our main

result could be extended at least for some dependence structures between T and D, namely, those for which $C(t)$ is bounded away from zero in the interval $\left[a_{\widetilde{H}}, b\right]$.

Theorem 1. Under (C1)-(C6) we have for $a_{\widetilde{H}} \leq t \leq b < b_{\widetilde{H}}$

$$F_n(t) - F(t) = -\frac{1}{\phi'(F(t))} \left\{ \sum_{i=1}^{n} \int_0^t \phi''(\overline{H}(s))\psi_i(s) dH^1(s) + \sum_{i=1}^{n} \widetilde{\psi}_i(t) \right\} + R_n(t)$$

where the ψ_i and $\widetilde{\psi}_i$ $(i = 1, ..., n)$ are i.i.d zero mean variables and

$$\sup_{a_{\widetilde{H}} \leq t \leq b} |R_n(t)| = O(n^{-3/4}(\log n)^{3/4}) \quad \text{a.s. as } n \to \infty.$$

Remark. (a) The ψ_i are defined as

$$\psi_i(t) = \overline{H}(t) \left\{ \frac{I(U_i \leq t, \rho_i = 1)}{C(U_i)} - \int_{a_{\widetilde{H}}}^t \frac{d\widetilde{H}^1(s)}{C(s)} - \int_{a_{\widetilde{H}}}^t \frac{I(T_i \leq s \leq U_i) - C(s)}{C(s)^2} d\widetilde{H}^1(s) \right\}.$$

(b) The $\widetilde{\psi}_i$ are defined as

$$\begin{aligned} \widetilde{\psi}_i(t) &= \varphi_t(\delta_i, U_i)\gamma_0(U_i)\rho_i - \gamma_1(U_i)\rho_i \\ &\quad + \gamma_2(T_i, U_i) - \gamma_3(T_i, U_i) \end{aligned}$$

where $\varphi_t(d, u) = \widetilde{\varphi}(u)d$ and $\widetilde{\varphi}(u) = I(u \leq t)\phi'(\overline{H}(u))$. The functions γ_0, γ_1, γ_2 and γ_3 are defined in Sánchez-Sellero et al. (2005). In our case they become:

$$\gamma_0(u) = \frac{1 - H(u)}{C(u)},$$

$$\gamma_1(u) = \frac{1}{C(u)} \int I(u < w)\widetilde{\varphi}(w)\gamma_0(w) d\widetilde{H}^{11}(w),$$

$$\gamma_2(s, u) = \int \int \frac{I(s < v < u, v < w)\widetilde{\varphi}(w)\gamma_0(w)}{C(v)^2} d\widetilde{H}^1(v) d\widetilde{H}^{11}(w),$$

$$\gamma_3(s, u) = \int_s^u \frac{\widetilde{\varphi}(w)\gamma_0(w)}{C(w)} d\widetilde{H}^{11}(w),$$

where $\widetilde{H}^{11}(w) = P(U \leq w, \rho = 1, \delta = 1 | T \leq U)$.

(c) In the special case of no truncation, the given representation reduces to that in de Uña-Álvarez and Veraverbeke (2013).

Proof to Theorem 1. Due to the regularity conditions in (C3)-(C4), from (1) and (2) we have

$$F_n(t) - F(t) = \frac{1}{\phi'(F(t))} \left\{ - \int_0^t \phi''(\overline{H}(s)) \left[\overline{H}_n(s) - \overline{H}(s) \right] dH^1(s) + \int_0^t \phi'(\overline{H}(s)) d \left[H_n^1(s) - H^1(s) \right] \right\} \quad (3)$$

$$+ R_{n1}(t) + R_{n2}(t) + R_{n3}(t)$$

where $R_{ni}(t), i = 1, 2, 3$, are remainder terms as in de Uña-Álvarez and Veraverbeke (2013). Lemmas 1 to 4 in the Appendix guarantee that these remainders satisfy the uniform rate given for $R_n(t)$. Now, in the first term of (3) we plug in the asymptotic representation for the TJW estimator due to Gijbels and Wang (1993) and sharpened by Zhou and Yip (1999), see their Theorem 2.2. Under (C1), (C2), (C5) and (C6) we have for $a_{\widetilde{H}} \leq t \leq b < b_{\widetilde{H}}$

$$H_n(t) - H(t) = \frac{1}{n} \sum_{i=1}^n \psi_i(t) + r_{n1}(t)$$

with $\sup_{a_{\widetilde{H}} \leq t \leq b} |r_{n1}(t)| = O(n^{-1} \log \log n)$ a.s. For the second term in (3), we use the result in Sánchez-Sellero et al. (2005), to obtain a suitable asymptotic representation; this is done by considering δ as a covariate. We obtain, under (C1), (C2), (C5) and (C6)

$$\int_0^t \phi'(\overline{H}(s)) d \left[H_n^1(s) - H^1(s) \right] = \frac{1}{n} \sum_{i=1}^n \widetilde{\psi}_i(t) + r_{n2}(t)$$

with $\sup_{a_{\widetilde{H}} \leq t \leq b} |r_{n2}(t)| = O(n^{-1} (\log n)^3)$ a.s. Under (C6) the integrability conditions in Sánchez-Sellero et al. (2005) are satisfied for the special family $\varphi_t(d, u) = I(u \leq t) \phi'(\overline{H}(u)) d$ with $a_{\widetilde{H}} \leq t \leq b < b_{\widetilde{H}}$, and the proof is complete.□

3 Simulation study

In this section we perform a simulation study to investigate the finite-sample performance of the proposed estimator. We consider a situation with two dependent, exponential survival times with rate 1, $Y \sim Exp(1)$ and $C \sim Exp(1)$. The variables Y and C follow a Clayton copula or a Frank copula. In the case of Clayton copula, the generator is given by $\phi_\theta(t) = t^{-\theta} - 1$, $\theta > 0$, i.e. the joint survival function is

$$P(Y > x_1, C > x_2) = \mathcal{C}(e^{-x_1}, e^{-x_2})$$

where

$$\mathcal{C}(u_1, u_2) = \left[u_1^{-\theta} + u_2^{-\theta} - 1\right]^{-1/\theta}.$$

This copula implies a Kendall's Tau $\tau_\theta = \theta/(\theta + 2)$; hence, only positive association is allowed. We consider the cases $\theta = 0.5$, 2, 10, corresponding to association levels of 0.2, 0.5 and 0.83 respectively. Specifically, the simulation algorithm is as follows (cfr. Exercise 4.17 in Nelsen (2006)):

Step 1. Generate independent random variables $V_1, V_2 \sim Exp(1)$
Step 2. Independently generate $Z_0 \sim \Gamma(1/\theta, 1)$, and compute $U_i = (1 + V_i/Z_0)^{-1/\theta}$, $i = 1, 2$
Step 3. Finally, compute $Y = -\ln(U_1)$, $C = -\ln(U_2)$

In the case of Frank copula, the generator is given by

$$\phi_\theta(t) = -\log\left[\frac{e^{-\theta t} - 1}{e^{-\theta} - 1}\right], \qquad \theta \neq 0.$$

Negative association is obtained when $\theta < 0$. The joint survival function is

$$P(Y > x_1, C > x_2) = \mathcal{C}(e^{-x_1}, e^{-x_2})$$

where

$$\mathcal{C}(u_1, u_2) = -\frac{1}{\theta}\log\left[1 + \frac{(e^{-\theta u_1} - 1)((e^{-\theta u_2} - 1))}{e^{-\theta} - 1}\right].$$

There is no explicit formula linking Kendall's Tau and θ for this model. In our simulations we consider $\theta = -12$, -5, and 2, with corresponding association levels of -0.71, -0.45, and 0.20. The data are generated by the inversion method, as follows:

Step 1. Generate independent random variables $V_1, V_2 \sim U(0, 1)$
Step 2. Compute $U_1 = V_1$ and $U_2 = \mathcal{C}^{-1}(V_2|U_1)$, where $\mathcal{C}(y|x) = \partial\mathcal{C}(x, y)/\partial x$
Step 3. Finally, compute $Y = -\ln(U_1)$, $C = -\ln(U_2)$

In Step 2, the following inverse function is needed:

$$\mathcal{C}^{-1}(y|x) = -\frac{1}{\theta}\log\left[1 + \frac{y(e^{-\theta} - 1)}{y + (1 - y)e^{-\theta x}}\right].$$

Once the variables Y and C are generated, we compute $Z = \min(Y, C)$ and $\delta = I(Y \leq C)$. The variable of interest is Y. We introduce independent censoring through a potential censoring time D independent of (Y, C), so the available information is $U = \min(Z, D)$, $\rho = I(Z \leq D)$, and $\rho\delta$. The distribution of D is $Exp(1)$ in Scenario 1 and $U(1, 1.5)$ in Scenario 2. Left-truncation is

introduced through an independent truncation time $T \sim U(0, t_m)$, where $t_m = 0.2$ or $t_m = 0.5$. Finally, the datum $(T, U, \rho, \rho\delta)$ is maintained only when $T \leq U$. A sample of size n is constructed following this scheme, where $n = 250$ or $n = 500$. Note that D and T are independent in these Scenarios 1 and 2. In order to simulate a situation in which D and T are dependent, we consider a third scenario (Scenario 3) in which $T \sim U(0, t_m)$ is drawn first and, afterwards, $D = T + \tau$ is computed. This represents the case in which Z is censored only when the residual time $Z - T$ exceeds the length of the follow-up period (τ) and, therefore, it mimics the situation of our real data application in Section 4 (see also the discussion of condition (C2)(ii) in Section 2). Here we take $\tau = 1$. Note that, in Scenario 3, D follows a $U(1, t_m + 1)$ distribution, which is also the distribution of D in Scenario 2 when $t_m = 0.5$. The truncation and censoring rates in these three scenarios are given in Table 1. The results on the performance of the proposed estimator are reported and discussed in Sections 3.1 (Clayton copula) and 3.2 (Frank copula). Results corresponding to the naive TJW estimator which ignores the dependent censoring are included to compare.

		Clayton			Frank	
	$\theta = 0.5$	$\theta = 2$	$\theta = 10$	$\theta = 2$	$\theta = -5$	$\theta = -12$
t_m		Scenario 1			Scenario 1	
0.2	24.4 (37.2)	23.4 (43.2)	21.0 (48.9)	23.9 (37.3)	25.7 (25.9)	25.7 (23.4)
0.5	47.0 (38.0)	44.5 (44.6)	40.1 (49.4)	46.1 (37.6)	51.5 (24.4)	52.3 (21.1)
		Scenario 2			Scenario 2	
0.2	17.2 (16.5)	16.0 (25.0)	13.2 (31.1)	16.6 (15.3)	18.7 (1.9)	18.7 (0.1)
0.5	35.0 (21.0)	31.6 (30.6)	25.6 (36.3)	33.9 (19.3)	41.5 (2.6)	42.5 (0.1)
		Scenario 3			Scenario 3	
0.2	17.1 (19.9)	16.0 (28.9)	13.1 (35.8)	16.6 (19.2)	18.6 (3.0)	18.7 (0.2)
0.5	35.0 (21.0)	31.6 (30.6)	25.6 (36.3)	33.9 (19.3)	41.5 (2.6)	42.5 (0.1)

Table 1. Truncation percentage and independent censoring rate (in brackets, $P(\rho = 0|T \leq U)$) for the simulated Scenarios. The percentage of dependent censoring ($P(\delta = 0|\rho = 1, T \leq U)$) is always 50%

3.1 Clayton copula

In Tables 2 to 7 we report the bias and the mean square error (MSE) of TJW and GCG survival estimators along 10,000 Monte Carlo trials for the Clayton copula and the three scenarios: Scenario 1 (Tables 2-3), Scenario 2 (Tables 4-5), and Scenario 3 (Tables 6-7). Both estimators are evaluated at the three quartiles of F: $t_1 = F^{-1}(0.25)$, $t_2 = F^{-1}(0.5)$, $t_3 = F^{-1}(0.75)$. In these tables we see that TJW is systematically biased while GCG is roughly unbiased. The bias of TJW grows with the association degree (as a result of the dependent censoring), being more severe at the right tail of F. The systematic bias of TJW

is responsible for its larger MSE (with the exception of $\theta = 0.5$, first quartile, where the MSE of TJW is smaller). On the other hand, the variances of TJW and GCG estimators are of the same order. In Scenario 3, the bias of the GCG estimator at the third quartile does not decrease when increasing the sample size if $t_m = 0.2$ (see Table 6). This is a result of the right-censoring variable D. Note that, when $t_m = 0.2$, the upper bound of the support of D is 1.2, smaller than $t_3 = 1.3863$. Therefore, no information on $F(t_3)$ is available in this case, and no estimator can be expected to be consistent.

In de Uña-Álvarez and Veraverbeke (2013), the performance of the GCG estimator without truncation was investigated in a Clayton copula model. We compare the results of our Scenario 1 to Tables 2 and 3, case $\lambda_C = \lambda_D = 1$ in that paper, which is the same scenario but without truncation. It is seen that the MSE of the GCG estimator is larger with truncation for the first and the second quartiles of Y, but the opposite is true for t_3 (particularly clear for $\theta = 0.5$ and $\theta = 2$). This is because left-truncation provokes some oversampling of relatively large lifetimes; as long as the final sample size n remains the same, this overinformation at large quantiles may result in a smaller variance in estimation. Regarding the bias, it is often the case for the simulated model that the absolute bias of the GCG estimator is smaller with truncated data, although in any case bias is negligible when compared to standard deviation.

The MSE grows with the truncation proportion (compare $t_m = 0.2$ to $t_m = 0.5$ in Tables), although some exceptions are found at the third quartile. For Scenarios 1 and 2, this may be again explained from the relative oversampling at large lifetimes which may be induced by a stronger left-truncation pattern. For Scenario 3, the explanation is different, and relates to the mentioned fact that, with $t_m = 0.2$, the variable D does not allow for the observation of Y around t_3.

We also may compare the results of Scenario 2 to those of Scenario 3 in the case $t_m = 0.5$, since both cases share the same percentages of truncation and censoring (Table 1). We see that the MSE of the GCG estimator is often (but not always) larger when T and D are dependent, indicating that the association degree between these two variables may influence the estimator's accuracy. Interestingly, the GCG estimator performs consistently even when $D = T + \tau$, suggesting that the independence between T and D is not substantial. See also our discussion of assumption (C2)(ii) in Section 2.

	$\theta =$	0.5		2		10	
		TJW	GCG	TJW	GCG	TJW	GCG
$n = 250$							
	t_1	0.0140	-0.0014	0.0419	-0.0012	0.0879	-0.0003
$t_m = 0.2$	t_2	0.0471	-0.0005	0.1158	0.0012	0.1832	0.0019
	t_3	0.0837	0.0007	0.1743	0.0025	0.2335	0.0035
	t_1	0.0137	-0.0027	0.0421	-0.0023	0.0888	-0.0004
$t_m = 0.5$	t_2	0.0467	-0.0008	0.1156	0.0011	0.1839	0.0030
	t_3	0.0838	0.0024	0.1742	0.0035	0.2339	0.0041
$n = 500$							
	t_1	0.0132	-0.0016	0.0415	-0.0009	0.0881	0.0003
$t_m = 0.2$	t_2	0.0472	-0.0001	0.1152	0.0003	0.1831	0.0007
	t_3	0.0840	0.0008	0.1743	0.0013	0.2327	0.0007
	t_1	0.0144	-0.0008	0.0413	-0.0016	0.0883	-0.0001
$t_m = 0.5$	t_2	0.0477	0.0004	0.1151	0.0006	0.1832	0.0015
	t_3	0.0836	0.0013	0.1742	0.0021	0.2329	0.0018

Table 2. Bias of the naive TJW estimator and of the generalized copula-graphic estimator (GCG) along 10,000 Monte Carlo trials. Scenario 1 with Clayton copula.

	$\theta =$	0.5		2		10	
		TJW	GCG	TJW	GCG	TJW	GCG
$n = 250$							
	t_1	0.0029	0.0031	0.0045	0.0035	0.0106	0.0042
$t_m = 0.2$	t_2	0.0048	0.0030	0.0161	0.0034	0.0364	0.0031
	t_3	0.0109	0.0041	0.0342	0.0035	0.0584	0.0026
	t_1	0.0049	0.0054	0.0069	0.0064	0.0122	0.0066
$t_m = 0.5$	t_2	0.0056	0.0038	0.0173	0.0045	0.0375	0.0041
	t_3	0.0103	0.0036	0.0340	0.0035	0.0583	0.0026
$n = 500$							
	t_1	0.0017	0.0018	0.0034	0.0020	0.0093	0.0023
$t_m = 0.2$	t_2	0.0036	0.0016	0.0148	0.0019	0.0350	0.0016
	t_3	0.0090	0.0022	0.0323	0.0019	0.0561	0.0012
	t_1	0.0028	0.0030	0.0045	0.0035	0.0103	0.0038
$t_m = 0.5$	t_2	0.0041	0.0021	0.0153	0.0023	0.0355	0.0023
	t_3	0.0087	0.0020	0.0322	0.0016	0.0562	0.0015

Table 3. MSE of the naive TJW estimator and of the generalized copula-graphic estimator (GCG) along 10,000 Monte Carlo trials. Scenario 1 with Clayton copula.

	$\theta =$	0.5		2		10	
		TJW	GCG	TJW	GCG	TJW	GCG
$n = 250$							
	t_1	0.0149	-0.0003	0.0420	-0.0010	0.0882	0.0011
$t_m = 0.2$	t_2	0.0472	0.0001	0.1156	0.0016	0.1834	0.0031
	t_3	0.0835	0.0024	0.1745	0.0039	0.2333	0.0042
	t_1	0.0134	-0.0033	0.0428	-0.0023	0.0901	0.0015
$t_m = 0.5$	t_2	0.0467	-0.0008	0.1161	0.0017	0.1848	0.0048
	t_3	0.0834	0.0031	0.1745	0.0045	0.2341	0.0050
$n = 500$							
	t_1	0.0148	0.0003	0.0409	-0.0013	0.0879	-0.0000
$t_m = 0.2$	t_2	0.0478	0.0009	0.1147	0.0003	0.1832	0.0012
	t_3	0.0839	0.0016	0.1733	0.0014	0.2334	0.0020
	t_1	0.0136	-0.0019	0.0416	-0.0017	0.0878	-0.0010
$t_m = 0.5$	t_2	0.0471	-0.0003	0.1153	0.0008	0.1830	0.0014
	t_3	0.0834	0.0013	0.1740	0.0023	0.2329	0.0016

Table 4. Bias of the naive TJW estimator and of the generalized copula-graphic estimator (GCG) along 10,000 Monte Carlo trials. Scenario 2 with Clayton copula.

	$\theta =$	0.5		2		10	
		TJW	GCG	TJW	GCG	TJW	GCG
$n = 250$							
	t_1	0.0029	0.0031	0.0044	0.0034	0.0107	0.0043
$t_m = 0.2$	t_2	0.0044	0.0025	0.0156	0.0028	0.0362	0.0028
	t_3	0.0096	0.0030	0.0330	0.0026	0.0572	0.0020
	t_1	0.0052	0.0058	0.0066	0.0062	0.0127	0.0069
$t_m = 0.5$	t_2	0.0055	0.0038	0.0169	0.0041	0.0378	0.0040
	t_3	0.0096	0.0032	0.0333	0.0028	0.0579	0.0024
$n = 500$							
	t_1	0.0019	0.0019	0.0033	0.0021	0.0093	0.0024
$t_m = 0.2$	t_2	0.0036	0.0014	0.0145	0.0017	0.0349	0.0015
	t_3	0.0084	0.0015	0.0314	0.0014	0.0559	0.0011
	t_1	0.0029	0.0031	0.0045	0.0036	0.0109	0.0046
$t_m = 0.5$	t_2	0.0040	0.0019	0.0153	0.0024	0.0359	0.0025
	t_3	0.0083	0.0014	0.0318	0.0016	0.0560	0.0013

Table 5. MSE of the naive TJW estimator and of the generalized copula-graphic estimator (GCG) along 10,000 Monte Carlo trials. Scenario 2 with Clayton copula.

	$\theta =$	0.5		2		10	
		TJW	GCG	TJW	GCG	TJW	GCG
$n = 250$							
	t_1	0.0145	-0.0008	0.0413	-0.0020	0.0891	-0.0002
$t_m = 0.2$	t_2	0.0474	0.0004	0.1155	0.0012	0.1840	0.0020
	t_3	0.1301	0.0552	0.2179	0.0556	0.2821	0.0573
	t_1	0.0145	-0.0020	0.0424	-0.0026	0.0874	-0.0014
$t_m = 0.5$	t_2	0.0478	0.0009	0.1157	0.0016	0.1829	0.0032
	t_3	0.0843	0.0046	0.1746	0.0050	0.2333	0.0045
$n = 500$							
	t_1	0.0139	-0.0008	0.0417	-0.0006	0.0875	-0.0006
$t_m = 0.2$	t_2	0.0470	-0.0001	0.1155	0.0009	0.1825	0.0006
	t_3	0.1289	0.0527	0.2184	0.0545	0.2812	0.0549
	t_1	0.0143	-0.0010	0.0419	-0.0019	0.0883	-0.0005
$t_m = 0.5$	t_2	0.0470	-0.0002	0.1158	0.0005	0.1833	0.0017
	t_3	0.0833	0.0015	0.1741	0.0017	0.2333	0.0023

Table 6. Bias of the naive TJW estimator and of the generalized copula-graphic estimator (GCG) along 10,000 Monte Carlo trials. Scenario 3 with Clayton copula.

	$\theta =$	0.5		2		10	
		TJW	GCG	TJW	GCG	TJW	GCG
$n = 250$							
	t_1	0.0029	0.0032	0.0046	0.0037	0.0104	0.0041
$t_m = 0.2$	t_2	0.0044	0.0026	0.0158	0.0030	0.0361	0.0027
	t_3	0.0203	0.0069	0.0509	0.0067	0.0828	0.0060
	t_1	0.0050	0.0057	0.0072	0.0070	0.0134	0.0082
$t_m = 0.5$	t_2	0.0055	0.0039	0.0172	0.0047	0.0379	0.0046
	t_3	0.0096	0.0036	0.0335	0.0034	0.0579	0.0026
$n = 500$							
	t_1	0.0016	0.0016	0.0032	0.0019	0.0095	0.0025
$t_m = 0.2$	t_2	0.0033	0.0013	0.0146	0.0015	0.0348	0.0015
	t_3	0.0185	0.0050	0.0495	0.0049	0.0810	0.0045
	t_1	0.0028	0.0031	0.0046	0.0037	0.0108	0.0044
$t_m = 0.5$	t_2	0.0039	0.0020	0.0155	0.0024	0.0359	0.0026
	t_3	0.0083	0.0017	0.0319	0.0015	0.0562	0.0017

Table 7. MSE of the naive TJW estimator and of the generalized copula-graphic estimator (GCG) along 10,000 Monte Carlo trials. Scenario 3 with Clayton copula.

3.2 Frank copula

Like for the Clayton copula, results on bias and MSE of TJW and GCG survival estimators for the Frank copula are reported in Tables 8-13. In this case, we see again that TJW is systematically biased, the bias being particularly visible as the association between Y and C gets stronger (as expected). Also, when $\tau_\theta < 0$, the bias of TJW is negative; this is because dependently censored values of Y are larger than what is expected under independence. Regarding the GCG estimator, it is virtually unbiased, although for the third quartile t_3 and large association degree the (positive) bias may be of the same order as the standard deviation for $n \leq 500$. This effect vanishes when considering larger sample sizes, although the bias of GCG at the third quartile remains always larger than at t_1 and t_2 (results not shown). Since large values of Y are attached to small values of its (dependent) censoring variable C when $\tau_\theta < 0$, these results are not entirely surprising; the 'effective sample size' for Y is smaller at the right tail, something that results in a larger standard deviation too. Our discussion also explains why the 'bias problem' is no longer present when considering the Frank copula with positive association (*e.g.* $\theta = 12$; results not shown). On the other hand, like for the Clayton copula, we see that $\overline{F}_n(t_3)$ has a systematic bias in Scenario 3 for $t_m = 0.2$; recall that, in this case, the maximum possible value of D is 1.2, which is smaller than t_3.

	$\theta =$	2		-5		-12	
		TJW	GCG	TJW	GCG	TJW	GCG
$n = 250$							
	t_1	0.0250	-0.0004	-0.0369	-0.0003	-0.0424	0.0002
$t_m = 0.2$	t_2	0.0576	0.0003	-0.1503	-0.0019	-0.2603	-0.0017
	t_3	0.0647	0.0004	-0.1514	-0.0121	-0.1961	0.0852
	t_1	0.0250	-0.0014	-0.0367	-0.0002	-0.0423	0.0006
$t_m = 0.5$	t_2	0.0580	0.0007	-0.1494	-0.0002	-0.2600	0.0003
	t_3	0.0644	0.0016	-0.1562	-0.0023	-0.2057	0.1010
$n = 500$							
	t_1	0.0245	-0.0004	-0.0364	0.0005	-0.0432	-0.0005
$t_m = 0.2$	t_2	0.0575	0.0001	-0.1497	-0.0006	-0.2602	-0.0011
	t_3	0.0649	0.0003	-0.1573	-0.0150	-0.2097	0.0570
	t_1	0.0234	-0.0020	-0.0363	0.0005	-0.0424	0.0002
$t_m = 0.5$	t_2	0.0566	-0.0007	-0.1491	0.0003	-0.2600	-0.0003
	t_3	0.0646	0.0009	-0.1581	-0.0041	-0.2174	0.0701

Table 8. Bias of the naive TJW estimator and of the generalized copula-graphic estimator (GCG) along 10,000 Monte Carlo trials. Scenario 1 with Frank copula.

	$\theta =$	2		-5		-12	
		TJW	GCG	TJW	GCG	TJW	GCG
$n = 250$							
	t_1	0.0032	0.0030	0.0042	0.0030	0.0045	0.0027
$t_m = 0.2$	t_2	0.0059	0.0027	0.0253	0.0033	0.0705	0.0031
	t_3	0.0084	0.0036	0.0278	0.0155	0.0421	0.0186
	t_1	0.0051	0.0051	0.0054	0.0042	0.0054	0.0037
$t_m = 0.5$	t_2	0.0066	0.0034	0.0247	0.0039	0.0697	0.0037
	t_3	0.0076	0.0033	0.0279	0.0114	0.0447	0.0181
$n = 500$							
	t_1	0.0020	0.0016	0.0028	0.0016	0.0033	0.0015
$t_m = 0.2$	t_2	0.0047	0.0013	0.0238	0.0017	0.0691	0.0017
	t_3	0.0063	0.0016	0.0274	0.0098	0.0459	0.0138
	t_1	0.0029	0.0027	0.0033	0.0022	0.0040	0.0023
$t_m = 0.5$	t_2	0.0049	0.0017	0.0234	0.0020	0.0687	0.0020
	t_3	0.0058	0.0015	0.0268	0.0064	0.0486	0.0116

Table 9. MSE of the naive TJW estimator and of the generalized copula-graphic estimator (GCG) along 10,000 Monte Carlo trials. Scenario 1 with Frank copula.

	$\theta =$	2		-5		-12	
		TJW	GCG	TJW	GCG	TJW	GCG
$n = 250$							
	t_1	0.0248	-0.0006	-0.0373	-0.0007	-0.0424	0.0003
$t_m = 0.2$	t_2	0.0574	0.0005	-0.1496	-0.0004	-0.2602	-0.0001
	t_3	0.0648	0.0021	-0.1531	0.0061	-0.2115	0.1101
	t_1	0.0256	-0.0015	-0.0364	0.0003	-0.0417	0.0009
$t_m = 0.5$	t_2	0.0581	0.0002	-0.1489	0.0011	-0.2588	0.0013
	t_3	0.0650	0.0025	-0.1564	0.0096	-0.2153	0.1165
$n = 500$							
	t_1	0.0246	-0.0001	-0.0368	0.0000	-0.0426	-0.0001
$t_m = 0.2$	t_2	0.0578	0.0008	-0.1494	-0.0000	-0.2602	-0.0004
	t_3	0.0644	0.0009	-0.1573	-0.0034	-0.2193	0.0845
	t_1	0.0247	-0.0008	-0.0366	-0.0000	-0.0423	0.0003
$t_m = 0.5$	t_2	0.0578	0.0005	-0.1493	0.0002	-0.2599	0.0000
	t_3	0.0647	0.0015	-0.1582	0.0015	-0.2222	0.0898

Table 10. Bias of the naive TJW estimator and of the generalized copula-graphic estimator (GCG) along 10,000 Monte Carlo trials. Scenario 2 with Frank copula.

	$\theta =$	2		-5		-12	
		TJW	GCG	TJW	GCG	TJW	GCG
$n = 250$							
	t_1	0.0036	0.0034	0.0041	0.0029	0.0044	0.0028
$t_m = 0.2$	t_2	0.0056	0.0024	0.0243	0.0029	0.0695	0.0028
	t_3	0.0071	0.0026	0.0264	0.0099	0.0464	0.0171
	t_1	0.0053	0.0054	0.0060	0.0049	0.0061	0.0045
$t_m = 0.5$	t_2	0.0065	0.0032	0.0243	0.0041	0.0686	0.0039
	t_3	0.0069	0.0026	0.0268	0.0092	0.0476	0.0189
$n = 500$							
	t_1	0.0022	0.0018	0.0027	0.0014	0.0032	0.0014
$t_m = 0.2$	t_2	0.0046	0.0012	0.0233	0.0014	0.0686	0.0015
	t_3	0.0056	0.0012	0.0265	0.0066	0.0490	0.0111
	t_1	0.0033	0.0031	0.0039	0.0028	0.0040	0.0024
$t_m = 0.5$	t_2	0.0051	0.0018	0.0234	0.0021	0.0683	0.0020
	t_3	0.0056	0.0014	0.0263	0.0051	0.0501	0.0117

Table 11. MSE of the naive TJW estimator and of the generalized copula-graphic estimator (GCG) along 10,000 Monte Carlo trials. Scenario 2 with Frank copula.

	$\theta =$	2		-5		-12	
		TJW	GCG	TJW	GCG	TJW	GCG
$n = 250$							
	t_1	0.0248	-0.0006	-0.0368	-0.0001	-0.0425	0.0003
$t_m = 0.2$	t_2	0.0577	0.0008	-0.1491	0.0004	-0.2602	-0.0003
	t_3	0.1192	0.0536	-0.1153	0.0556	-0.2045	0.1169
	t_1	0.0246	-0.0025	-0.0369	-0.0003	-0.0415	0.0008
$t_m = 0.5$	t_2	0.0578	0.0000	-0.1489	0.0010	-0.2596	0.0006
	t_3	0.0655	0.0030	-0.1546	0.0126	-0.2153	0.1160
$n = 500$							
	t_1	0.0241	-0.0006	-0.0367	0.0000	-0.0425	0.0002
$t_m = 0.2$	t_2	0.0571	-0.0001	-0.1492	0.0000	-0.2595	0.0004
	t_3	0.1184	0.0520	-0.1187	0.0505	-0.2108	0.0960
	t_1	0.0232	-0.0023	-0.0366	0.0001	-0.0423	0.0002
$t_m = 0.5$	t_2	0.0565	-0.0007	-0.1492	0.0003	-0.2596	0.0004
	t_3	0.0635	0.0005	-0.1577	0.0020	-0.2227	0.0893

Table 12. Bias of the naive TJW estimator and of the generalized copula-graphic estimator (GCG) along 10,000 Monte Carlo trials. Scenario 3 with Frank copula.

	$\theta =$	2		−5		−12	
		TJW	GCG	TJW	GCG	TJW	GCG
$n = 250$							
	t_1	0.0033	0.0031	0.0044	0.0033	0.0046	0.0029
$t_m = 0.2$	t_2	0.0055	0.0022	0.0242	0.0030	0.0695	0.0028
	t_3	0.0178	0.0063	0.0163	0.0120	0.0436	0.0186
	t_1	0.0056	0.0059	0.0057	0.0047	0.0059	0.0043
$t_m = 0.5$	t_2	0.0066	0.0034	0.0242	0.0039	0.0690	0.0038
	t_3	0.0070	0.0026	0.0262	0.0087	0.0476	0.0186
$n = 500$							
	t_1	0.0025	0.0021	0.0028	0.0016	0.0036	0.0019
$t_m = 0.2$	t_2	0.0047	0.0013	0.0232	0.0015	0.0683	0.0017
	t_3	0.0163	0.0046	0.0159	0.0084	0.0454	0.0135
	t_1	0.0037	0.0035	0.0036	0.0024	0.0040	0.0023
$t_m = 0.5$	t_2	0.0052	0.0020	0.0233	0.0021	0.0682	0.0021
	t_3	0.0055	0.0014	0.0262	0.0056	0.0503	0.0117

Table 13. MSE of the naive TJW estimator and of the generalized copula-graphic estimator (GCG) along 10,000 Monte Carlo trials. Scenario 3 with Frank copula.

4 Real data illustration

In this section we revisit the Galician unemployment data (see e.g. de Uña-Álvarez and Iglesias-Pérez, 2010). The data concern unemployment spells of 1,009 married women living in Galicia (NW of Spain), recruited by means of quarterly inquiries at the individuals' homes. The unemployment situation ends when the invididual finds a job or when she stops searching for a job. Here we denote by Y and C the latent variables "time to finding a job" and "time to stop the searching" respectively. These two variables are negatively correlated, since individuals with short values of Y are better prepared to find a new job and, therefore, they will find no reasons to stop their searching (large values of C). To model this negative correlation we use Frank's copula, which is Archimedean with generator given by

$$\phi_\theta(t) = -\log\left[\frac{e^{-\theta t}-1}{e^{-\theta}-1}\right], \qquad \theta \neq 0.$$

Negative values of θ result in negative association. The independent case is obtained in the limit ($\theta \to 0$). In particular, we take $\theta = -12$, -5 and -2 which lead to association levels (Kendall's Tau) of -0.71, -0.45 and -0.22 respectively.

The dataset reports 219 uncensored values of Y, 227 uncensored values of C, and 563 cases of administrative censoring (because of limitations in the

follow-up period). Note that realizations of Y (respectively C) imply some dependent censoring on C (resp. Y), due to the expected negative association (as discussed above). Besides, since the sampled information corresponds to women unemployed by the inquiry date, the data are left-truncated. Of course, ignoring left-truncation leads to a serious overestimation of the unemployment time (de Uña-Álvarez and Iglesias-Pérez, 2010). The truncation time T is just the time in unemployment by the inquiry date. The administrative censoring time D may be represented as $D = T + \tau$ where $\tau = 18$ (in months), leading to the violation of the independence assumption (C2)(ii). As discussed in Section 2 (see also simulations in Section 3), this is not crucial for the consistency of the new estimator nor for the validity of representation in Theorem 1.

Figure 1 below depicts the copula-graphic estimator of the cumulative distribution function of Y when using Frank's copula with the several choices of θ. The TJW estimator which incorrectly assumes independence between Y and C is included for comparison purposes. From Figure 1 it is seen that the copula-graphic estimator separates from the NPMLE as the correlation grows. Indeed, it becomes clear from this Figure 1 that TJW is underestimating the time to finding a job; this is because TJW ignores that individuals stop searching for a job at a given time are not representative of those who continue their searching, having less chances to return to the job market.

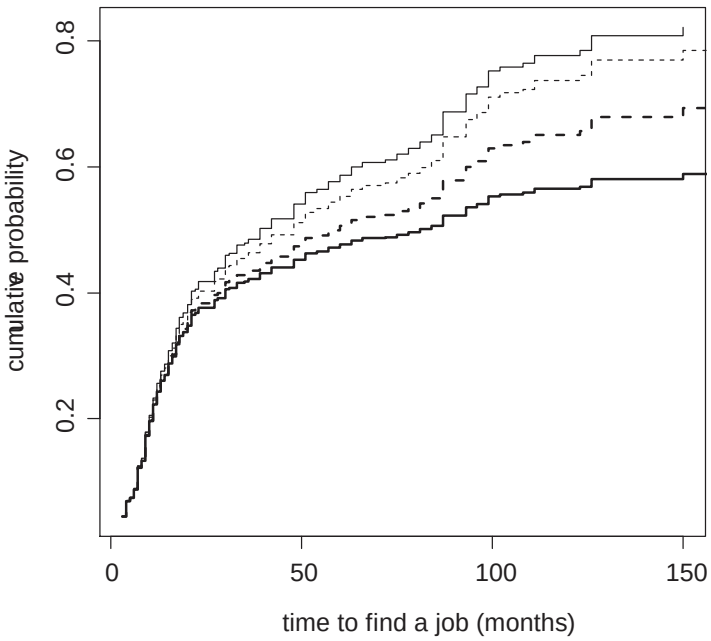

Figure 1. Cumulative distribution function of time to find a job based on Frank's copula: independent setting (thin solid line), and association levels of -0.22 (thin dashed), -0.45 (thick dashed) and -0.71 (thick solid line). Galician unemployment data.

5 Main conclusions

In this paper a new survival function estimator for left-truncated and right-censored data has been introduced, and an asymptotic iid representation for the estimator has been established. The new estimator is suitable for situations in which some dependent censoring is present. This is the case when, e.g., there are several competing risks acting on each individual. This dependent censoring is incorporated in the construction of the estimator through and Archimedean copula function, which is supposed to be known; this is because of the non-identifiability problem in the random right-censoring model (Tsiatis, 1975). Therefore, some information on the dependence structure between the lifetime of ultimate interest and the dependent right-censoring time is needed or pre-assumed. In our real data example on unemployment, this information comes from the fact that the individuals with long time to the next job are the worst prepared or qualified and, consequently, the ones leaving their searching sooner. A copula inducing negative association between time to finding a job

and time to stop the searching is suitable in this case. We have shown through simulated and real data that the naive estimator which ignores the dependent censoring may lead to a severe bias.

Besides the dependent censoring, the introduced model allows for the presence of some administrative censoring, which is regarded as independent of the lifetime. In practice, this administrative censoring time D may depend however on the left-truncation time T; this is indeed the case for the unemployment dataset considered in Section 4 and, in general, one will have a situation like this whenever the data come from a cross-section. Although some of the auxiliary results for left-truncated, right-censored data rely on the independence between D and T (e.g. Sánchez-Sellero et al., 2005), we have seen that this assumption may be skipped provided that the expected proportion of individuals at risk remains bounded away from zero along time. Our simulations have indeed confirmed that the proposed estimator behaves consistently when the pair (D, T) falls on a line.

Acknowledgements. Work supported by the Grant MTM2011-23204 of the Spanish Ministerio de Innovación y Ciencia (FEDER support included). The second author acknowledges the IAP Research Network P7/13 of the Belgian State (Begian Science Policy).

6 References

Arcones MA, Giné E (1995) On the law of the iterated logarithm for canonical U-statistics and processes. Stochastic Processes and their Applications 58, 217-245.

de Uña-Álvarez J, Iglesias-Pérez MC (2010) Nonparametric estimation of a conditional distribution from length-biased data. Annals of the Institute of Statistical Mathematics 62, 323-341.

de Uña-Álvarez J, Veraverbeke N (2013) Generalized copula-graphic estimator. Test 22, 343-360.

Gijbels I, Wang TJ (1993) Strong representations of the survival function estimator for truncated and censored data with applications. Journal of Multivariate Analysis 47, 210-229.

Kalbfleisch JD, Prentice RL (1980) *The Statistical Analysis of Failure Time Data*. Wiley, New York.

Nelsen RB (2006) *An Introduction to Copulas*. Springer, New York.

Rivest LP, Wells MT (2001) A martingale approach to the copula-graphic estimator for the survival function under dependent censoring. Journal of Multivariate Analysis, 79, 138-155.

Sánchez-Sellero C, González-Manteiga W, Van Keilegom I (2005) Uniform representation of product-limit integrals with applications. Scandinavian Journal of Statistics 32, 563-581.

Tsai WY, Jewell NP, Wang MC (1987) A note on the product-limit estimator under right censoring and left truncation. Biometrika 74, 883-886.

Tsiatis A (1975). A nonidentifiability aspect of the problem of competing risks. Proceedings of the National Academy of Sciences, 72, 20-22.

Woodroofe M (1985) Estimating a distribution function with truncated data. Annals of Statistics 13, 163-177.

Zheng M, Klein JP (1995). Estimates of marginal survival for dependent competing risks based on an assumed copula. Biometrika, 82, 127-138.

Zhou X, Sun L, Ren H (2000) Quantile estimation for left truncated and right censored data. Statistica Sinica 10, 1217-1229.

Zhou Y, Yip PSF (1999) A strong representation of the product-limit estimator for left truncated and right censored data. Journal of Multivariate Analysis 69, 261-280.

7 Appendix: technical lemmas

In this section we provide the technical lemmas used to prove our main result.

Lemma 1. Under the conditions of Theorem 1 we have

$$\sup_{a_{\widetilde{H}}\leq t\leq b}|R_{n1}(t)|=O(n^{-1}\log\log n)\quad \text{a.s. as } n\to\infty.$$

Proof. As in Lemma 1 in de Uña-Álvarez and Veraverbeke (2013) we have a.s.

$$\sup_{a_{\widetilde{H}}\leq t\leq b}|R_{n1}(t)|=O(\sup_{a_{\widetilde{H}}\leq t\leq b}|H_n(t)-H(t)|^2+\sup_{a_{\widetilde{H}}\leq t\leq b}\left|H_n^1(t)-H^1(t)\right|^2).$$

The first term is $O(n^{-1}\log\log n)$ a.s. by Zhou and Yip (1999), Corollary 2.2, which follows from their Theorem 2.2 and the standard functional LIL for a two-parameter Wiener process, as indicated in that paper. This rate also follows from Theorem 2.2 in Zhou and Yip (1999) and the LIL for empirical processes on VC-classes of functions, provided that the VC-class has a square integrable envelope, which is true under (C6) and $b<b_{\widetilde{H}}$; see Arcones and Giné (1995). For the second term we use the a.s. representation of Sánchez-Sellero et al. (2005) for the special family $\varphi_t(d,u)=I(u\leq t)d$ with $a_{\widetilde{H}}\leq t\leq b$, and (again) and the LIL for empirical processes on VC-classes of functions (Arcones and Giné, 1995), to get the given rate. The details are omitted.□

Lemma 2. Under the conditions of Theorem 1 we have

$$\sup_{a_{\widetilde{H}}\leq t\leq b}|R_{n2}(t)|=O(n^{-1}\log\log n)\quad \text{a.s. as } n\to\infty.$$

Proof. As in Lemma 2 in de Uña-Álvarez and Veraverbeke (2013) we have a.s.

$$\sup_{a_{\widetilde{H}}\leq t\leq b}|R_{n2}(t)|=O(\sup_{a_{\widetilde{H}}\leq t\leq b}|H_n(t)-H(t)|^2)$$

which is $O(n^{-1}\log\log n)$ a.s. by Zhou and Yip (1999), Corollary 2.2.□

Lemma 3. Under the conditions of Theorem 1 we have

$$\sup_{a_{\widetilde{H}}\le t\le b}|R_{n3}(t)| = O(n^{-3/4}(\log n)^{3/4}) \quad \text{a.s. as } n\to\infty.$$

Proof. Divide the interval $\left[a_{\widetilde{H}}, b\right]$ into $k_n = O(n^{1/2}(\log n)^{1/2})$ subintervals $[t_i, t_{i+1}]$ of length $O(n^{-1/2}(\log n)^{1/2})$. Then, as in de Uña-Álvarez and Veraverbeke (2013) we have

$$\sup_{c\le t\le d}|R_{n3}(t)| \le I + II$$

with I and II as in the above paper. For I we have by Taylor expansion and the result of Zhou and Yip (1999), Corollary 2.2,

$$\begin{aligned} I &\le 2\max_{1\le i\le k_n}\sup_{t_i\le y\le t_{i+1}}\left|\phi''(\overline{H}(t_{i+1})\right||H_n(y)-H(y)-H_n(t_i)+H(t_i)| \\ &\quad +O(n^{-1}\log\log n). \end{aligned}$$

Now, further subdivide each interval $[t_i, t_{i+1}]$ into $a_n = O(n^{1/4}(\log n)^{-1/4})$ subintervals of length $O(n^{-3/4}(\log n)^{3/4})$. By Bernstein's inequality we can show that this term is bounded a.s. by

$$c\max_{1\le i\le k_n}\max_{0\le j\le a_n-1}|H_n(t_{i,j+1})-H(t_{i,j+1})-H_n(t_i)+H(t_i)| + O(n^{-3/4}(\log n)^{3/4})$$

for some constant $c>0$. By the modulus of continuity result for the TJW estimator in Zhou et al. (2000), Lemma A.1 (which is a direct consequence of Theorem 2.2 in Zhou and Yip, 1999, and the LIL in Arcones and Giné, 1995), we obtain that $I = O(n^{-3/4}(\log n)^{3/4})$ a.s. Note that the interval $[c,d]$ with $a_{\widetilde{H}} < c$ in that Lemma A.1 may be expanded to $\left[a_{\widetilde{H}}, b\right]$ under (C6), which is enough to guarantee the existence of a square integrable envelope for the VC-class of functions considered in that paper. The II term is treated similarly and leads to the same order bound. It requires the a.s. rate

$$\sup_{a_{\widetilde{H}}\le t\le b}\left|H_n^1(t)-H^1(t)\right|^2 = O(n^{-1}\log\log n)$$

(see our Lemma 1) and also an almost sure order bound for the modulus of continuity of H_n^1. The latter follows from Lemma 4 below by taking $a_n = n^{-1/2}(\log n)^{1/2}$.□

Lemma 4. Let $\{a_n\}$ be a sequence of positive constants tending to zero with $a_n n(\log n)^{-5} > \Delta > 0$ for all n sufficiently large. Then, under the conditions in Theorem 1,

$$\sup_{a_{\widetilde{H}}\le t,s\le b,|t-s|\le a_n}\left|H_n^1(t)-H_n^1(s)-H^1(t)+H^1(s)\right| = O(a_n^{1/2}n^{-1/2}(\log n)^{1/2}) \quad \text{a.s.}$$

Proof. We proceed exactly in the same way as in Lemma 5 of de Uña-Álvarez and Veraverbeke (2013). We need the almost sure asymptotic representation for $H_n^1(t)$ as it can be derived from the result in Sánchez-Sellero et al. (2005), by taking δ as a covariate, with the choice $\varphi(d,u) \equiv \varphi_t(d,u) = I(u \leq t)d$:

$$H_n^1(t) = \frac{1}{n}\sum_{i=1}^{n} \xi_i(t) + J_n(t)$$

where $\xi_i(t)$ are the iid variables in that paper corresponding to such function φ_t, and where a.s.

$$\sup_{\widetilde{a}_H \leq t \leq b} |J_n(t)| = O(n^{-1}(\log n)^3).$$

Note that, similarly as in Section 2, under (C6) the integrability conditions in Sánchez-Sellero et al. (2005), Theorem 1, are satisfied for the special family φ_t, $a_{\widetilde{H}} \leq t \leq b$. Again as in Lemma 5 of de Uña-Álvarez and Veraverbeke (2013) it suffices to prove that $Var(\xi_i(t) - \xi_i(s))$ is bouded by a constant times $|t-s|$, for $a_{\widetilde{H}} \leq t, s \leq b$. This is shown by checking appropriate groups of terms in the variance. In the lengthy calculations, which are omitted here, the Lipschitz continuity of $\widetilde{H}^1$ and $\widetilde{H}^{11}$ is needed; but this follows by (C3) up to noting that $d\widetilde{H}^1(t) = \alpha^{-1}\Lambda(t)dH(t)$ and $d\widetilde{H}^{11}(t) = \alpha^{-1}\Lambda(t)dH^1(t)$.□