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

Jacobo de Uña Álvarez; Noël Veraverbeke

University of Vigo; U. Hasselt and North-West University, Potchefstroom

jacobo@uvigo.es

June 13, 2014

Overview

1. Introduction
2. The estimator
3. Simulation study
4. Real data
5. Main conclusions and Discussion
6. References

[Introduction] Some motivation

- Usual situation in survival analysis: the event time of interest Y is potentially censored by C
- Only $Z = min(Y, C)$ and $\delta = I(Y \leq C)$ are observed
- Typically Y and C are assumed to be independent
- Dependent censoring appears in practice
- Example: survival times with two causes of death
- Another example: $Y =$ time to finding a job, $C =$ time to stop searching for a job
- In general, event times are dependently censored under competing risks
- Kaplan-Meier estimator may be inconsistent with dependent censoring
- Existing solution: copula-graphic estimator: Zheng and Klein (1995), Rivest and Wells (2001)

[Introduction] Survival data with competing risks

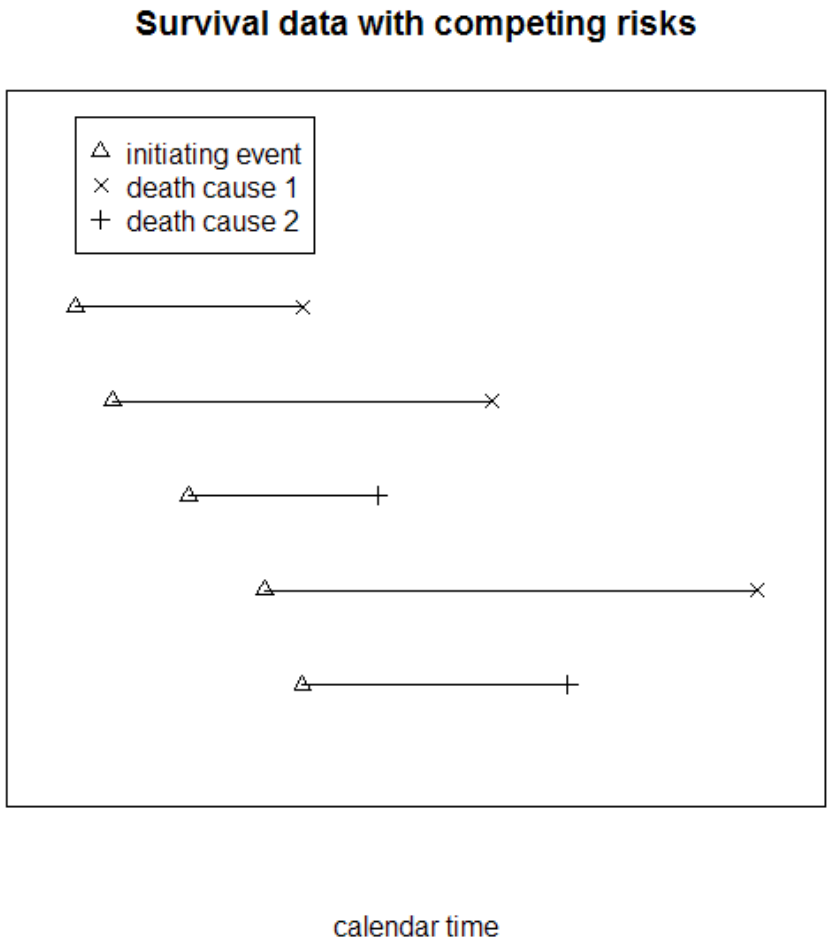

Figure: Y = time to death cause 1, C = time to death cause 2, δ = cause of death

[Introduction] Ordinary copula-graphic estimator

- 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 $\overline{G}(t) = 1 - G(t)$:

$$P(Y > t_1, C > t_2) = \mathcal{C}(\overline{F}(t_1), \overline{G}(t_2))$$

$$= \phi^{-1}(\phi(\overline{F}(t_1)) + \phi(\overline{G}(t_2))).$$

- 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 particular case $\phi(t) = -\log t$ leads to the product copula $\mathcal{C}(u_1, u_2) = u_1 u_2$ and corresponds to independence between Y and C
- More on copulas: Nelsen (2006).

[Introduction] Ordinary copula-graphic estimator

- Denote $H(t) = P(Z \leq t)$, $\overline{H}(t) = 1 - H(t)$, and $H^1(t) = P(Z \leq t, \delta = 1)$, where $Z = min(Y, C)$ and $\delta = I(Y \leq C)$
- Then, if ϕ' exists and if H^1 is differentiable, we have from Tsiatis (1975)

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

- An estimator of $\overline{F}(t)$ is obtained after plugging in ordinary empiricals for H and H^1; Zheng and Klein (1995), Rivest and Wells (2001)
- Problem 1: information on (Z, δ) may not be fully available, because of time limitations, losses, etc. (right censoring)
- Problem 2: Z may be subject to left-truncation (cross-sectional sampling, delayed entries)

[Introduction] Left-truncation and competing risks

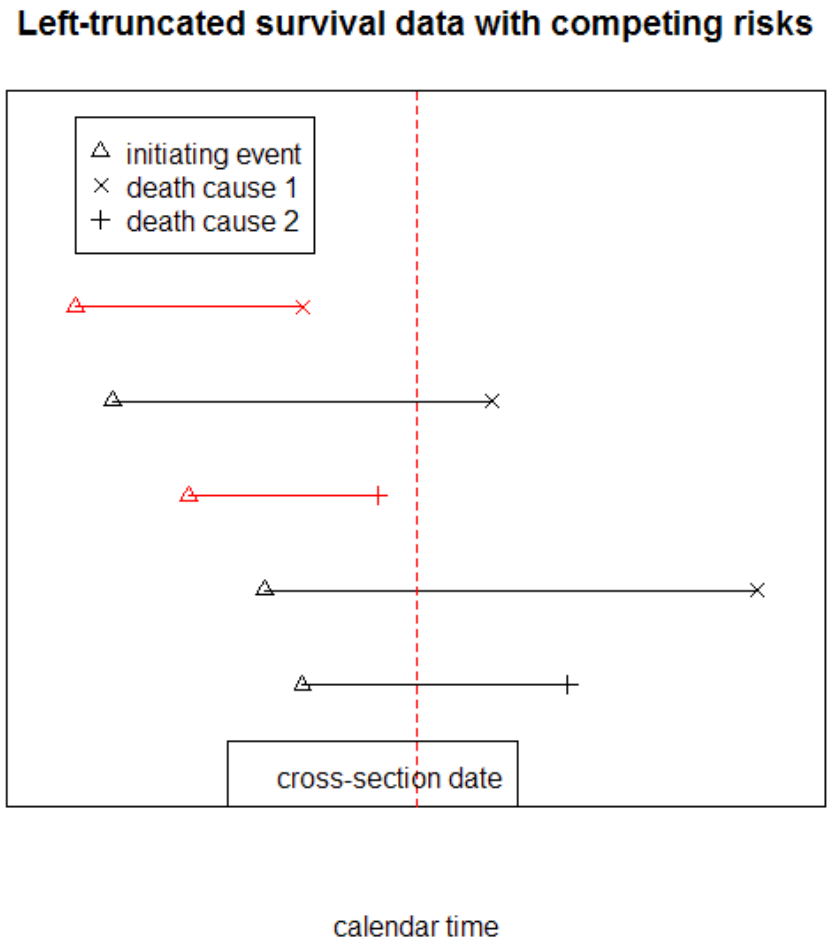

Figure: Y = time to death cause 1, C = time to death cause 2, δ = cause of death, T = time to cross-section (left-truncation time)

[Introduction] LTRC data and competing risks

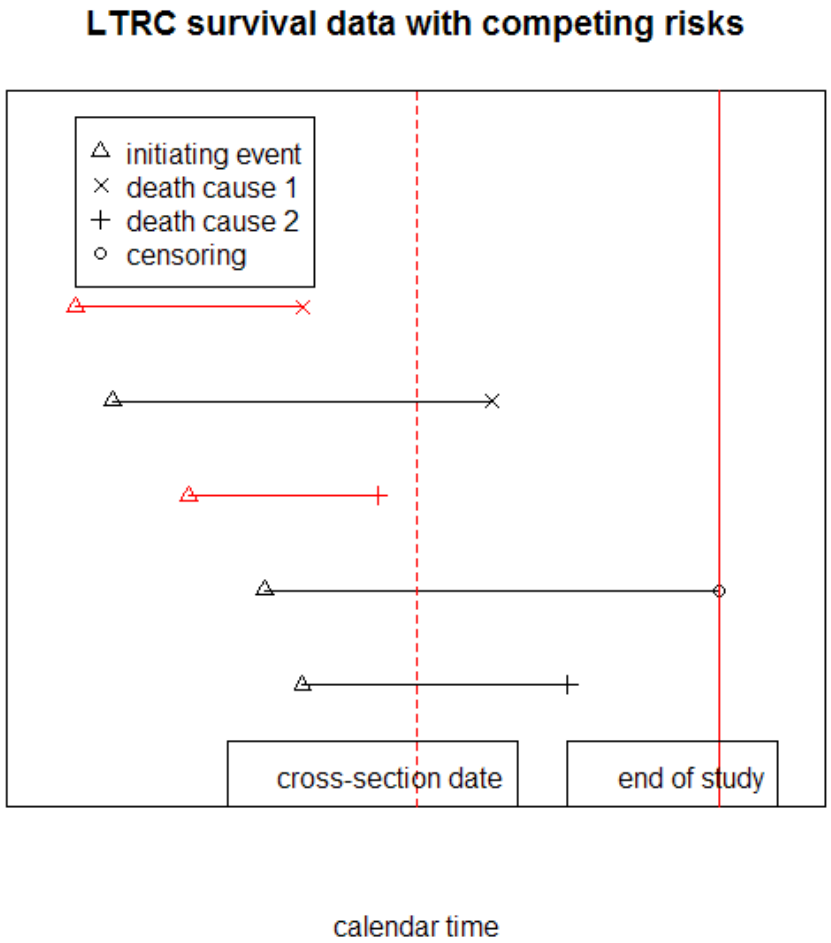

Figure: Y = time to death cause 1, C = time to death cause 2, δ = cause of death, D = time to administrative censoring, ρ = censoring status

[Introduction] Notations

- D is the (administrative) right-censoring time
- T is the left-truncation time
- We assume that (T, D) is independent of (Y, C) (and thus of (Z, δ))
- Rather than (Z, δ) we observe $(T, U, \rho, \rho\delta)$ conditionally on $T \leq U$ where $U = \min(Z, D)$ and $\rho = I(Z \leq D)$; note that the value of δ is observed only when Z is uncensored $(\rho = 1)$
- When $T > U$ nothing is observed
- We put $\widetilde{G}$ and L for the distribution functions of D and T respectively
- The sample is $(T_i, U_i, \rho_i, \rho_i\delta_i)$, $i = 1, ..., n$, iid observations of $(T, U, \rho, \rho\delta)$ conditionally on $T \leq U$

[The estimator] Estimating H and H^1

- Estimate H by the TJW estimator (Tsai et al., 1987):

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

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

- It can also be expressed as

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

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

[The estimator] Estimating H and H^1

- To estimate H^1 we consider δ_i as a covariable for the possibly censored lifetime U_i; following Sánchez-Sellero et al. (2005), we introduce

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

- $H_n^1(t)$ is just the usual estimator for a cumulative incidence function in a censored competing risks model, cfr. Kalbfleisch and Prentice (1980), adapted to left-truncation.
- An almost sure representation for $H_n^1(t)$ and for the corresponding product-limit integral $\sum_{i=1}^{n} W_{in} \varphi(U_i, \delta_i)$ for some general real-valued function φ can be found in Sánchez-Sellero et al. (2005), under the assumption of independence between T and D (but often $D = T + \tau$ in practice); their result can be extended to cope with dependencies (more on this later)

[The estimator] The GCG estimator

- We propose the following generalized copula-graphic estimator for $\overline{F}(t)$:

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

  where $\overline{H}_n = 1 - H_n$
- Without left-truncation, $\overline{F}_n(t)$ reduces to that in de Uña-Álvarez and Veraverbeke (2013)
- Without administrative censoring ($D = \infty$), W_{in} is just the jump of the Lynden-Bell estimator for left-truncated data (Woodroofe, 1985) and therefore $\overline{F}_n(t)$ extends that estimator to dependently censored data. If moreover 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 (1) also leads to the TJW estimator in absence of dependent censoring ($Z = Y$, $\delta = 1$), based on observations of (T, U, ρ).

[The estimator] Main result

- We prove an almost sure asymptotic representation for $\overline{F}_n(t)$ with a uniform rate for the remainder.
- For any distribution function K we put $a_K = \inf\{t : K(t) > 0\}$ and $b_K = \sup\{t : K(t) < 1\}$
- Introduce $\widetilde{H}(t) = P(U \leq t)$,

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

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

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

- Note $a_{\widetilde{H}} = \min(a_F, a_G, a_{\widetilde{G}})$ and $b_{\widetilde{H}} = \min(b_F, b_G, b_{\widetilde{G}})$. Put $F_n = 1 - \overline{F}_n$.

[The estimator] Main result

- 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) $\int_{a_{\widetilde{H}}}^{b} C(t)^{-3} d\widetilde{H}^1(t) < \infty$

[The estimator] Main result

- Conditions (C1)-(C4) reduce to those considered in de Uña-Álvarez and Veraverbeke (2013) when there is no truncation; while (C5) ensures identifiability.
- Assumption (C6) was used by Zhou and Yip (1999) to obtain an almost sure uniform representation for H_n; it is enough for the purpose of applying Theorem 1 in Sánchez-Sellero et al. (2005) too.
- 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 in that case.
- Often with cross-sectional data, $D = T + \tau$ 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 basically the identifiability assumption (C5)

[The estimator] Main result

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

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

$$+R_n(t)$$

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

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

Proof. Same steps as in de Uña-Álvarez and Veraverbeke (2013). We need: the almost sure representation for the TJW estimator in Zhou and Yip (1999); and in Sánchez-Sellero et al. (2005) for product-limit integrals.

[Simulation study] Simulated scenario

- $Y \sim Exp(1)$ and $C \sim Exp(1)$
- The variables Y and C follow a Clayton copula with generator $\phi_\theta(t) = t^{-\theta} - 1$, $\theta > 0$, i.e.

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

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

- This copula implies a Kendall's Tau $\tau_\theta = \theta/(\theta + 2)$. We consider the cases $\theta = 0.5$, 2, 10, corresponding to association levels of 0.2, 0.5 and 0.83 respectively
- $D \sim Exp(1)$ (Scenario 1) or $D \sim U(1, 1.5)$ (Scenario 2)
- $T \sim U(0, t_m)$ where $t_m = 0.2$ or $t_m = 0.5$; datum rejected if $T > min(Y, C, D)$
- $n = 250, 500$; results for GCG and TJW estimators at the three quartiles t_1, t_2, t_3 of $Exp(1)$ along $M = 10,000$ trials

[Simulation study] Simulated scenario

- Note that D and T are independent in Scenarios 1 and 2.
- 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 (here $\tau = 1$).
- 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 in many cross-sectional studies.
- In this Scenario 3, D follows a $U(1, t_m + 1)$ distribution, which is also the distribution of D in Scenario 2 when $t_m = 0.5$.

[Simulation study] Simulated scenario

- We also consider a Frank copula with generator

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

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.

[Simulation study] Clayton copula

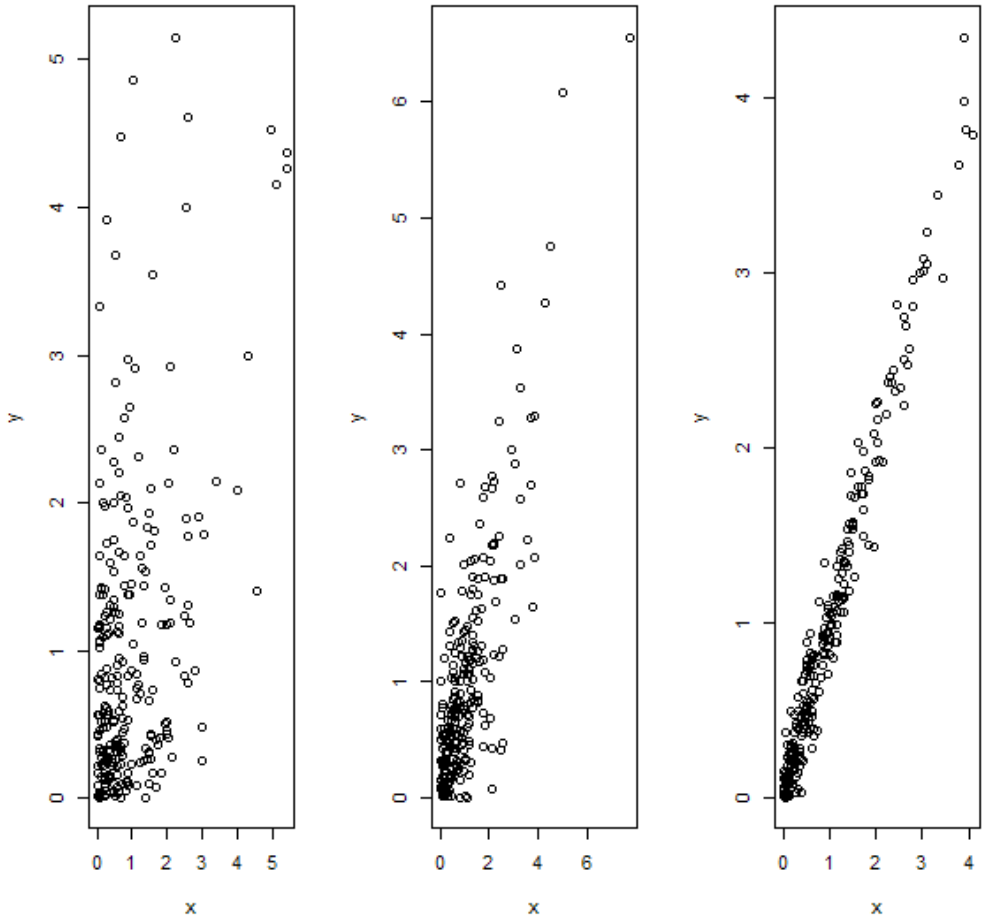

Figure: $n = 250$ simulated $Exp(1)$ marginals from Clayton copula, $\theta = 0.5, 2, 10$

[Simulation study] Frank copula

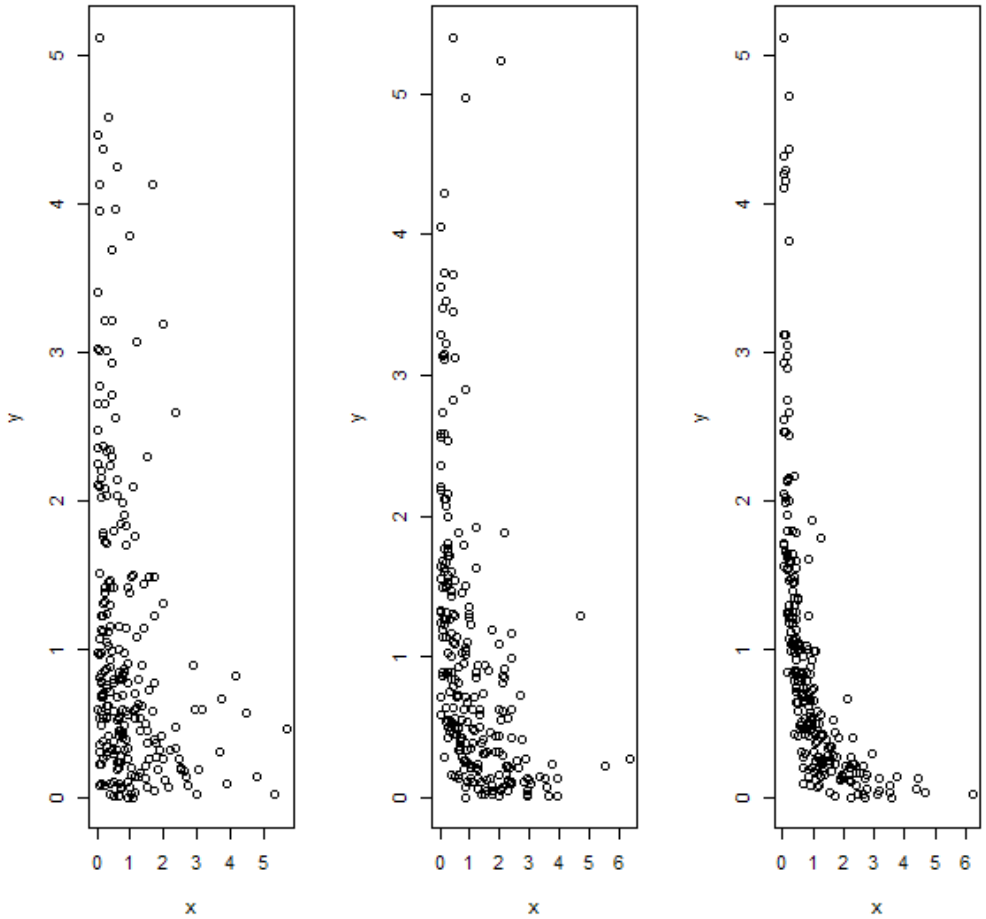

Figure: $n = 250$ simulated $Exp(1)$ marginals from Frank copula, $\theta = -2, -5, -12$

[Simulation study] Truncation and censoring rates

		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%

[Simulation study] Results (Clayton copula)

- TJW is systematically bias, the bias growing with the association degree (more visible at the right tail), while GCG is roughly unbiased
- In Scenario 3 with $t_m = 0.2$, the upper bound of the support of D is smaller than t_3 (systematic bias of GCG in this case)
- The variances of TJW and GCG are of the same order
- Compared to de Uña-Álvarez and Veraverbeke (2013) for the untruncated case, the MSE of GCG is larger with truncation for t_1, t_2, but the opposite is true for t_3 (according to the oversampling at the right tail)
- The MSE of GCG grows with the truncation proportion, but there are some exceptions at t_3 (again, due to the overinformation at this point which is provoked by left-truncation)
- Comparison of Scenarios 2 and 3 ($t_m = 0.5$) suggests that the MSE of GCG is often (not always) larger when T and D are dependent
- The GCG estimator performs consistently regardless the dependence between T and D

[Simulation study] Results (Frank copula)

- Roughly the same results for Frank and Clayton copulas when Y and C are positively associated
- With negative association, the increase of the dependent censoring on Y at the right tail (where the administrative censoring effects are stronger) results in more bias and variance of the GCG estimator

[Simulation study] Results

	$\theta =$	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.

[Simulation study] Results

	$\theta =$	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.

[Simulation study] Results

	$\theta =$	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.

[Simulation study] Results

	$\theta =$	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.

[Simulation study] Results

	$\theta =$	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.

[Simulation study] Results

	$\theta =$	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.

[Simulation study] Results

	$\theta =$	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.

[Simulation study] Results

	$\theta =$	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.

[Real data] Galician unemployment data

- 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 (e.g. de Uña-Álvarez and Iglesias-Pérez, 2010). The unemployment situation ends when the invididual finds a job or when she stops searching for a job.
- We denote by Y and C the latent variables "time to finding a job" and "time to stop the searching" respectively.
- Y and C 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, with association levels (Kendall's Tau) of -0.71, -0.45 and -0.22 ($\theta = -12$, -5 and -2 resp.)

[Real data] Galician unemployment data

- 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).
- Besides, since the sampled information corresponds to women unemployed by the inquiry date, the data are left-truncated. 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, this is not crucial for the consistency of the new estimator nor for the validity of representation in Theorem 1.

[Introduction] Galician unemployment data

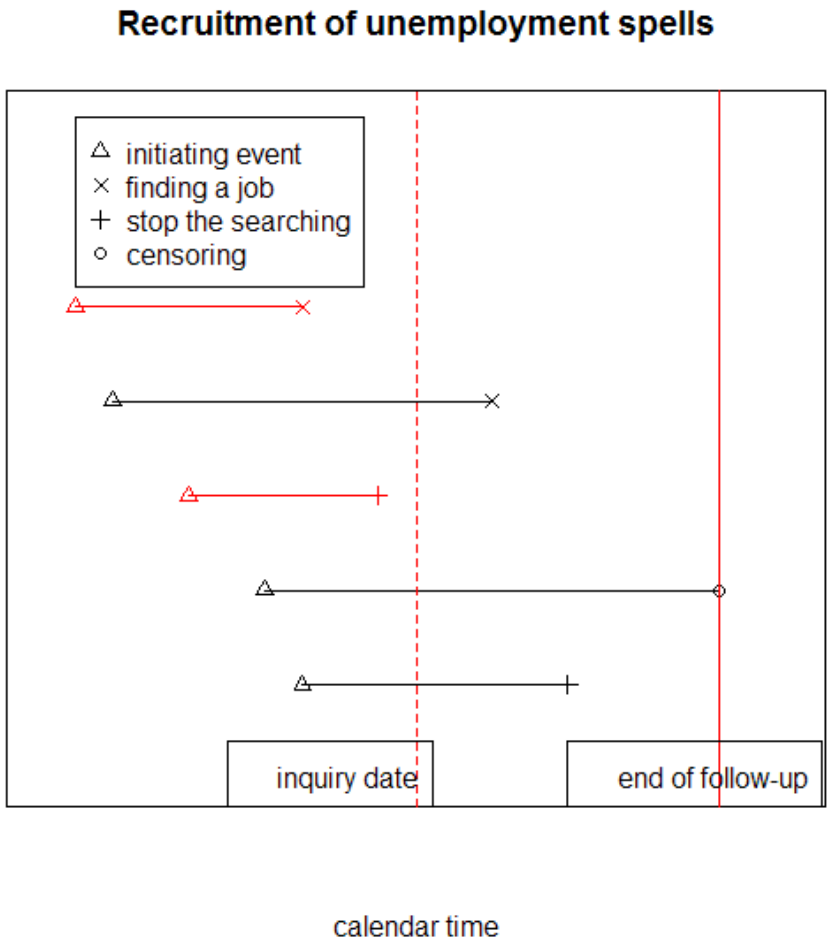

Figure: Y = time to finding a job, C = time to stop the searching, δ = employment/out-of-labour force indicator, D = time to administrative censoring

[Real data] Time to find a job under Frank copula

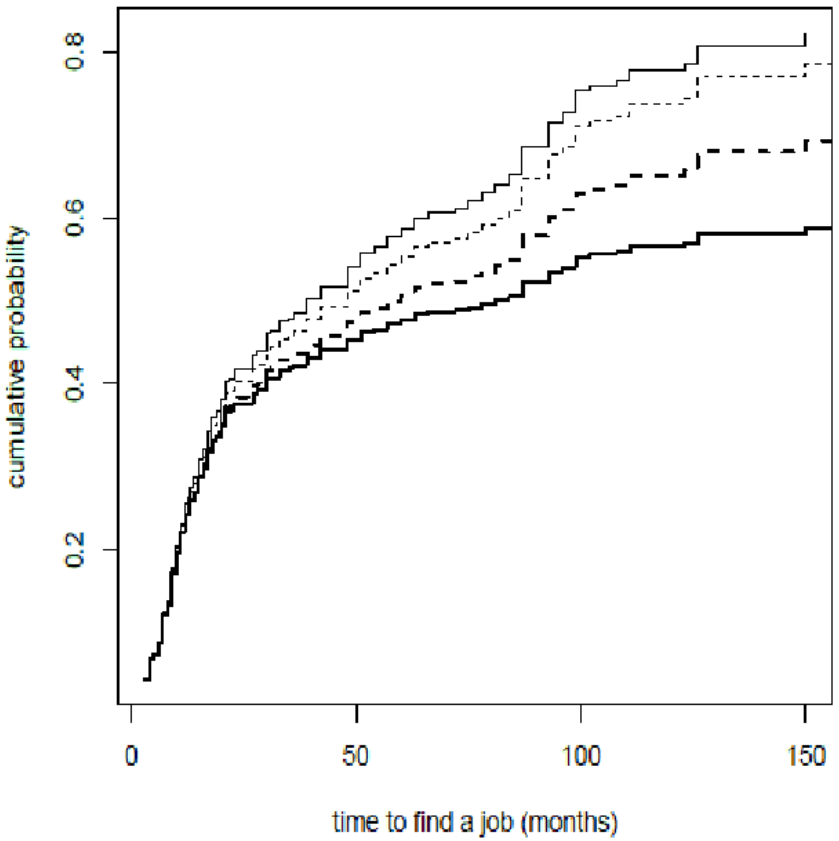

Figure: Independent setting (thin solid line), and association levels of -0.22 (thin dashed), -0.45 (thick dashed), and -0.71 (thick solid line).

[Main conclusions and Discussion] Main conclusions

- New estimator for left-truncated and right-censored data, extension of TJW for dependent censoring (e.g. competing risks)
- Asymptotic almost sure iid representation for the estimator
- Because of the non-identifiability problem (Tsiatis, 1975), some external information on the dependence structure is needed to choose a suitable copula
- 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
- The naive TJW estimator may lead to a severe bias
- 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

[Main conclusions and Discussion] Discussion

- Related but different model: semi-competing risks, C (time to a terminal event) censors Y (time to a non-terminal event) but not vice-versa (e.g. Lakhal et al. 2008; Heuchenne et al. 2014)
- Information on the truncation distribution could be relevant to reduce the variance (Asgharian et al. 2002)
- Covariate information may be included in the construction of the estimator (Braekers and Veraverbeke, 2005)

References

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.

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 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.

Thank you! Gracias!

Jacobo de Uña-Álvarez
Department of Statistics and OR -University of Vigo (Spain)
and
SiDOR Research Group
http://sidor.uvigo.es

E-mail: jacobo@uvigo.es
Web site: http://webs.uvigo.es/jacobo