T E C H N I C A L
R E P O R T

0315

A COPULA-GRAPHIC ESTIMATOR FOR THE CONDITIONAL SURVIVAL FUNCTION UNDER DEPENDENT CENSORING

R. BRAEKERS and N. VERAVERBEKE

I A P S T A T I S T I C S
N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

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

A copula-graphic estimator for the conditional survival function under dependent censoring.

Roel Braekers and Noël Veraverbeke
Limburgs Universitair Centrum, Universitaire Campus
B-3590 Diepenbeek, Belgium.
roel.braekers@luc.ac.be, noel.veraverbeke@luc.ac.be

Abstract

In survival analysis, it is very common to assume that the lifetime variable and the censoring variable are independent. In this case, the product limit estimator is the standard non-parametric estimator for the distribution function of the lifetime variable. When the assumption of independence is not satisfied, Zheng and Klein (1995) proposed a copula-graphic estimator where the dependence between lifetime and censoring variable is described by a copula. Rivest and Wells (2001) derived an explicit form for this estimator if the copula is Archimedean.

In this paper, we extend the estimator of Rivest and Wells (2001) to the fixed design regression case. For our copula-graphic estimator, we find an asymptotic representation and prove weak convergence to a Gaussian limit. We illustrate the estimation method with a classical dataset on bone marrow transplant patients.

1 Introduction

At fixed design points $0 \leq x_1 \leq \ldots \leq x_n \leq 1$, we have nonnegative responses $Y_1, \ldots, Y_n$ such as survival times or failure times. These responses are independent random variables and the distribution function of the response Y_i at x_i will be denoted by $F_{x_i}(t) = P(Y_i \leq t)$.

In many clinical or industrial trials, the responses $Y_1, \ldots, Y_n$ are subject to random right censoring. For each response, there is a censoring variable C_i with conditional distribution function $G_{x_i}(t) = P(C_i \leq t)$. The observed random variables at design point x_i are in fact Z_i and δ_i $(i = 1, \ldots, n)$, with

$$Z_i = \min(Y_i, C_i) \qquad \text{and} \qquad \delta_i = I(Y_i \leq C_i).$$

At a given fixed design value $x \in [0,1]$, we write F_x, G_x, H_x for the distribution function of respectively the response Y_x, the censoring variable C_x and the observed variable $Z_x = \min(Y_x, C_x)$ at x. Also we will write $\delta_x = I(Y_x \leq C_x)$. Note that for the design variables x_i, we write $Y_i, C_i, Z_i, F_i, \ldots$ instead of $Y_{x_i}, C_{x_i}, Z_{x_i}, F_{x_i}, \ldots$.

In order to estimate uniquely the distribution function F_x from the observed data, we have to make an assumption about the dependence between the Y_i and C_i for each i (Tsiatis 1975). It is very common in survival analysis to assume independence between these random variables (conditional on the covariate). However we see that in some practical situations, for example in oncology, this assumption does not hold. Therefore we assume that we can rewrite the joint survival function of the response Y_x and the censoring variable C_x at x as

$$S_x(t_1, t_2) = P(Y_x > t_1, C_x > t_2) = \mathcal{C}_x(\bar{F}_x(t_1), \bar{G}_x(t_2))$$

where $\mathcal{C}_x$ is a known copula function depending in a general way on x and $\bar{F}_x(t)$ (resp. $\bar{G}_x(t)$) is the survival function of Y_x (resp. C_x) at x. Without covariates x, this idea was introduced by Zheng and Klein (1995). However their copula-graphic estimator had no closed form expression. Rivest and Wells (2001) got around this problem by focusing on the class of Archimedian copulas. In this work, we will extend their ideas to the fixed design regression case.

We assume that at a fixed design value $x \in [0,1]$, the joint survival function is given by

$$S_x(t_1, t_2) = \varphi_x^{[-1]}(\varphi_x(\bar{F}_x(t_1)) + \varphi_x(\bar{G}_x(t_2))) \tag{1}$$

where, for each x, $\varphi_x : [0,1] \to [0, +\infty]$ is a known continuous, convex, strictly decreasing function with $\varphi_x(1) = 0$. $\varphi_x^{[-1]}$ is the pseudo-inverse of φ_x, as defined in Nelsen (1999).

We note from (1) that,

$$1 - H_x(t) = \bar{H}_x(t) = S_x(t,t) = \varphi_x^{[-1]}(\varphi_x(\bar{F}_x(t)) + \varphi_x(\bar{G}_x(t))).$$

This paper concerns a non-parametric estimation of $F_x(t)$ and is organized as follows. In Section 2, we define the distribution function estimator F_{xh} for F_x. It is an extension of the Beran estimator, as it was studied by Van Keilegom and Veraverbeke (1996, 1997a and 1997b). After specifying some assumptions in Section 3, we derive for this estimator an asymptotic representation in Section 4 and prove weak convergence in Section 5. In Section 6, we apply our estimator to a practical situation where we take different choices for the generator function φ_x.

2 Copula-graphic estimator

For a fixed design value $x \in [0,1]$, we derive an estimator for the distribution function $F_x(t)$. Since we only have observations at the design points $x_1, \ldots, x_n$, we use smoothing weights to give observations at a design point close to x a larger contribution in our estimator than observations at design points far away from x. In a fixed design regression it is natural to work with Gasser-Müller weights,

$$w_{ni}(x,h_n) = \frac{1}{c_n(x,h_n)} \int_{x_{i-1}}^{x_i} \frac{1}{h_n} K\left(\frac{x-z}{h_n}\right) dz \qquad (i=1,\ldots,n), \tag{2}$$

$$c_n(x,h_n) = \int_0^{x_n} \frac{1}{h_n} K\left(\frac{x-z}{h_n}\right) dz \tag{3}$$

where $x_0 = 0$, K is a known probability density function, called the kernel and $\{h_n\}$ is a sequence of positive constants, tending to zero as $n \to +\infty$, called the bandwidth sequence.

Let us assume that there are no ties in the observations. To find an estimator for $\bar{F}_x(t)$ (resp. $\bar{G}_x(t)$) at the design point x, we work as Rivest and Wells (2001) and look for the right continuous step function $\bar{F}_{xh}(t)$ (resp. $\bar{G}_{xh}(t)$) with $\bar{F}_{xh}(0) = 1$ (resp. $\bar{G}_{xh}(0) = 1$), which has jumps at the points Z_i with $\delta_i = 1$ (resp. $\delta_i = 0$) satisfying

$$\varphi_x^{[-1]}(\varphi_x(\bar{F}_{xh}(Z_i)) + \varphi_x(\bar{G}_{xh}(Z_i))) = \bar{H}_{xh}(Z_i)$$

where $\bar{H}_{xh}(t) = \sum_{i=1}^{n} w_{ni}(x,h_n) I(Z_i > t)$.

To get a closed form expression for $\bar{F}_{xh}$, we take a point Z_i with $\delta_i = 1$. The function $\bar{G}_{xh}$ has not jump in this point i.e. $\bar{G}_{xh}(Z_i^-) = \bar{G}_{xh}(Z_i)$, and the jump of $\bar{F}_{xh}$ at Z_i satisfies

$$\begin{aligned} \varphi_x(\bar{F}_{xh}(Z_i^-)) - \varphi_x(\bar{F}_{xh}(Z_i)) &= \varphi_x(\bar{H}_{xh}(Z_i^-)) - \varphi_x(\bar{H}_{xh}(Z_i)) \\ &= \varphi_x(\bar{H}_{xh}(Z_i^-)) - \varphi_x(\bar{H}_{xh}(Z_i^-) - w_{ni}(x,h_n)). \end{aligned}$$

Hence

$$\begin{aligned} \varphi_x(\bar{F}_{xh}(t)) &= -\sum_{Z_i \le t, \delta_i = 1} \varphi_x(\bar{F}_{xh}(Z_i^-)) - \varphi_x(\bar{F}_{xh}(Z_i)) \\ &= -\sum_{Z_i \le t, \delta_i = 1} \varphi_x(\bar{H}_{xh}(Z_i^-)) - \varphi_x(\bar{H}_{xh}(Z_i^-) - w_{ni}(x,h_n)) \end{aligned}$$

and

$$\bar{F}_{xh}(t) = \varphi_x^{[-1]}\left(-\sum_{Z_i \le t, \delta_i = 1} \varphi_x(\bar{H}_{xh}(Z_i^-)) - \varphi_x(\bar{H}_{xh}(Z_i^-) - w_{ni}(x,h_n))\right). \tag{4}$$

In (4) we can replace without any complications the pseudo inverse $\varphi_x^{[-1]}$ by the inverse φ_x^{-1}. Furthermore we note that this estimator in general does not tend to 0 as $t \to +\infty$. In order to have a proper distribution estimator, we use the modification

$$\bar{F}_{xh}(t) = \varphi_x^{-1}\left(-\sum_{Z_i \le t, \delta_i = 1} \varphi_x(\bar{H}_{xh}(Z_i^-)) - \varphi_x(\bar{H}_{xh}(Z_i^-) - w_{ni}(x, h_n))\right) I(t < Z_{(n)})$$

where $Z_{(n)}$ is the largest order statistic in the sample $Z_1, \ldots, Z_n$.

3 Regularity conditions

For the design points $x_1, \ldots, x_n$ we write $\underline{\Delta}_n = \min\limits_{1 \le i \le n}(x_i - x_{i-1})$ and $\bar{\Delta}_n = \max\limits_{1 \le i \le n}(x_i - x_{i-1})$. The notations $||K||_\infty = \sup\limits_{u \in \mathbb{R}} K(u)$, $||K||_2^2 = \int\limits_{-\infty}^{+\infty} K^2(u)du$, $\mu_1^K = \int\limits_{-\infty}^{+\infty} uK(u)du$, $\mu_2^K = \int\limits_{-\infty}^{+\infty} u^2 K(u)du$ will be used for the kernel K.

We use the following assumptions on the design and on the kernel.

(C1) $x_n \to 1$, $\bar{\Delta}_n = O(n^{-1})$, $\bar{\Delta}_n - \underline{\Delta}_n = o(n^{-1})$.

(C2) K is a probability density function with finite support $[-M, M]$ for some $M > 0$, $\mu_1^K = 0$ and K Lipschitz of order 1.

Note that, for $c_n(x, h_n)$ defined in (3), $c_n(x, h_n) = 1$ for n sufficiently large. Therefore we take $c_n(x, h_n) = 1$ in all proofs of asymptotic results.

If L is any (sub)distribution, then T_L denotes the right endpoint of its support ($T_L = \inf\{t : L(t) = L(+\infty)\}$). Here we have that $T_{H_x} \le \min(T_{F_x}, T_{G_x})$ where we attain the equality in case $\varphi_x(0) = +\infty$. For $\varphi_x(0) < +\infty$, it depends on the function φ_x whether or not we have an equality. To obtain our results, we need some smoothness conditions on the functions $H_x(t) = P(Z_x \le t)$ and $H_x^u(t) = P(Z_x \le t, \delta_x = 1)$. For a fixed $T > 0$,

(C3) $\dot{L}_x(t) = \frac{\partial}{\partial x} L_x(t)$ exists and is continuous in $(x, t) \in [0, 1] \times [0, T]$

(C4) $L_x'(t) = \frac{\partial}{\partial t} L_x(t)$ exists and is continuous in $(x, t) \in [0, 1] \times [0, T]$

(C5) $\ddot{L}_x(t) = \frac{\partial^2}{\partial x^2} L_x(t)$ exists and is continuous in $(x, t) \in [0, 1] \times [0, T]$

(C6) $L_x''(t) = \frac{\partial^2}{\partial t^2} L_x(t)$ exists and is continuous in $(x, t) \in [0, 1] \times [0, T]$

(C7) $\dot{L}'_x(t) = \frac{\partial^2}{\partial x \partial t} L_x(t)$ exists and is continuous in $(x,t) \in [0,1] \times [0,T]$

The generator $\varphi_x(v)$ of the Archimedean copula needs to satisfy the following properties.

(C8) $\varphi'_x(v) = \frac{\partial}{\partial v}\varphi_x(v)$ and $\varphi''_x(v) = \frac{\partial^2}{\partial v^2}\varphi_x(v)$ are Lipschitz in the x-direction with a bounded Lipschitz constant, and $\varphi'''_x(v) = \frac{\partial^3}{\partial v^3}\varphi_x(v) \leq 0$ exists and is continuous in $(x,v) \in [0,1]\times]0,1]$.

These assumptions and the fact that φ_x is a generator for an Archimedean copula, give that $\varphi'_x(v)$ is monotone increasing with $\varphi'_x(v) < 0$ and $\varphi''_x(v)$ is monotone decreasing with $\varphi''_x(v) \geq 0$.

4 Almost sure asymptotic representation

Before we derive an asymptotic representation for $F_{xh}(t)$, we give a lemma about the survival function F_x.

Lemma 1. If $H_x(t)$ and $H^u_x(t)$ satisfy (C4) in $[0,1] \times [0,T]$ with $T < T_{H_x}$ and $\varphi'_x(v)$ exists on $[0,1]\times]0,1]$, then under (1),

$$\bar{F}_x(t) = \varphi_x^{-1}\left(-\int_0^t \varphi'_x(\bar{H}_x(s))dH^u_x(s)\right).$$

Proof. Under (1) and with Tsiatis (1975), we get that

$$H^{u\prime}_x(t) = -\left.\frac{\partial}{\partial t_1}S_x(t_1,t_2)\right|_{t_1=t_2=t} = \frac{\varphi'_x(\bar{F}_x(t))F'_x(t)}{\varphi'_x(\bar{H}_x(t))}.$$

This leads to

$$\varphi_x^{-1}\left(-\int_0^t \varphi'_x(\bar{H}_x(s))dH^u_x(s)\right) = \varphi_x^{-1}\left(-\int_1^{\bar{F}_x(t)} \varphi'_x(w)dw\right) = \bar{F}_x(t).$$

Theorem 1. Assume (C1), (C2), $H_x(t)$ and $H^u_x(t)$ satisfy (C5), (C6) and (C7) in $[0,T]$ with $T < T_{H_x}$, φ_x satisfies (C8), $h_n \to 0$, $\frac{\log n}{nh_n} \to 0$, $\frac{nh_n^5}{\log n} = O(1)$, then, under (1) as $n \to +\infty$,

$$F_{xh}(t) - F_x(t) = \sum_{i=1}^n w_{ni}(x,h_n)g_{tx}(Z_i,\delta_i) + R_n(t)$$

where

$$\begin{aligned} g_{tx}(Z_i,\delta_i) &= \frac{-1}{\varphi_x'(\bar{F}_x(t))}\left[\int_0^t \varphi_x''(\bar{H}_x(s))(I(Z_i\le s)-H_x(s))dH_x^u(s)\right. \\ &\quad -\varphi_x'(\bar{H}_x(t))(I(Z_i\le t,\delta_i=1)-H_x^u(t)) \\ &\quad \left.-\int_0^t \varphi_x''(\bar{H}_x(s))(I(Z_i\le s,\delta_i=1)-H_x^u(s))dH_x(s)\right] \end{aligned}$$

and $\sup_{0\le t\le T}|R_n(t)| = O((nh_n)^{-3/4}(\log n)^{3/4})$ a.s.

Proof. Based on Lemma 1, we can write for $t < T_{H_{xh}}$,

$$\begin{aligned} F_{xh}(t)-F_x(t) &= \left[-\varphi_x^{-1}\left(-\sum_{Z_i\le t,\delta_i=1}\varphi(\bar{H}_{xh}(Z_i^-))-\varphi(\bar{H}_{xh}(Z_i^-)-w_{ni}(x,h_n))\right)\right. \\ &\quad \left.+\varphi_x^{-1}\left(-\sum_{Z_i\le t,\delta_i=1}\varphi_x'(\bar{H}_{xh}(Z_i))w_{ni}(x,h_n)\right)\right] \\ &\quad -\left[\varphi_x^{-1}\left(-\int_0^t\varphi_x'(\bar{H}_{xh}(s))dH_{xh}^u(s)\right)-\varphi_x^{-1}\left(-\int_0^t\varphi_x'(\bar{H}_x(s))dH_x^u(s)\right)\right]. \end{aligned}$$

Applying a first order Taylor expansion on the first term and a second order Taylor expansion on the second term, we get

$$\begin{aligned} F_{xh}(t)-F_x(t) &= \frac{-1}{\varphi_x'(\bar{F}_x(t))}\left[-\int_0^t\varphi_x'(\bar{H}_{xh}(s))dH_{xh}^u(s)+\int_0^t\varphi_x'(\bar{H}_x(s))dH_x^u(s)\right] \\ &\quad +R_{n1}(t)+R_{n2}(t) \end{aligned}$$

where

$$\begin{aligned} R_{n1}(t) &= \frac{\varphi_x''(\varphi_x^{-1}(\varepsilon_1))}{2\varphi_x'(\varphi_x^{-1}(\varepsilon_1))^3}\left[-\int_0^t\varphi_x'(\bar{H}_{xh}(s))dH_{xh}^u(s)+\int_0^t\varphi_x'(\bar{H}_x(s))dH_x^u(s)\right]^2 \\ R_{n2}(t) &= \frac{-1}{\varphi_x'(\varphi_x^{-1}(\varepsilon_2))}\left[-\sum_{Z_i\le t,\delta_i=1}(\varphi_x(\bar{H}_{xh}(Z_i^-))-\varphi_x(\bar{H}_{xh}(Z_i^-)-w_{ni}(x,h_n)))\right. \\ &\quad \left.+\sum_{Z_i\le t,\delta_i=1}\varphi_x'(\bar{H}_{xh}(Z_i))w_{ni}(x,h_n)\right] \end{aligned}$$

with ε_1 between $-\int_0^t\varphi_x'(\bar{H}_{xh}(s))dH_{xh}^u(s)$ and $-\int_0^t\varphi_x'(\bar{H}_x(s))dH_x^u(s)$, and ε_2 between $-\sum_{Z_i\le t,\delta_i=1}(\varphi_x(\bar{H}_{xh}(Z_i^-))-\varphi_x(\bar{H}_{xh}(Z_i^-)-w_{ni}(x,h_n)))$ and $-\sum_{Z_i\le t,\delta_i=1}\varphi_x'(\bar{H}_{xh}(Z_i))w_{ni}(x,h_n)$.

Furthermore, for $t < T_{H_{xh}}$:

$$-\int_0^t \varphi_x'(\bar{H}_{xh}(s))dH_{xh}^u(s) + \int_0^t \varphi_x'(\bar{H}_x(s))dH_x^u(s) = -\int_0^t (\varphi_x'(\bar{H}_{xh}(s)) - \varphi_x'(\bar{H}_x(s)))dH_x^u(s)$$

$$-\int_0^t \varphi_x'(\bar{H}_x(s))d(H_{xh}^u(s) - H_x^u(s)) - \int_0^t (\varphi_x'(\bar{H}_{xh}(s)) - \varphi_x'(\bar{H}_x(s)))d(H_{xh}^u(s) - H_x^u(s)).$$

On the integrand of the first term, we use a second order Taylor expansion and the second term can be rewritten by partial integration. So we get

$$-\int_0^t \varphi_x'(\bar{H}_{xh}(s))dH_{xh}^u(s) + \int_0^t \varphi_x'(\bar{H}_x(s))dH_x^u(s) = \int_0^t \varphi_x''(\bar{H}_x(s))(H_{xh}(s) - H_x(s))dH_x^u(s)$$

$$-\varphi_x'(\bar{H}_x(t))(H_{xh}^u(t) - H_x^u(t)) - \int_0^t \varphi_x''(\bar{H}_x(s))(H_{xh}^u(s) - H_x^u(s))dH_x(s) + R_{n3}(t) + R_{n4}(t) \quad (5)$$

where

$$\begin{aligned} R_{n3}(t) &= -\int_0^t \frac{\varphi_x'''(\varepsilon_3)}{2}(H_{xh}(s) - H_x(s))^2 dH_x^u(s) \\ R_{n4}(t) &= -\int_0^t (\varphi_x'(\bar{H}_{xh}(s)) - \varphi_x'(\bar{H}_x(s)))d(H_{xh}^u(s) - H_x^u(s)) \end{aligned}$$

with ε_3 between $\bar{H}_{xh}(s)$ and $\bar{H}_x(s)$.

Since $H_x(T) < 1$ and $H_{xh}(T) \to H_x(T)$ a.s. (Lemma A.2. of Van Keilegom and Veraverbeke (1997b)), we may suppose that $T < T_{H_{xh}}$. For $R_{n3}(t)$ we have

$$\begin{aligned} \sup_{0 \le t < leT} |R_{n3}(t)| &\le \frac{1}{2} \sup_{0 \le t \le T} (H_{xh}(t) - H_x(t))^2 \max(\sup_{0 \le t \le T} |\varphi_x'''(\bar{H}_{xh}(t))|, \sup_{0 \le t \le T} |\varphi_x'''(\bar{H}_x(t))|) \\ &= O((nh_n)^{-1} \log n) \qquad \text{a.s.} \end{aligned}$$

by applying Lemma A.4. of Van Keilegom and Veraverbeke (1997b). By Lemma 2 below, $\sup_{0 \le t \le T} |R_{n4}(t)| = O((nh_n)^{-3/4} (\log n)^{3/4})$ a.s.

From (5), Lemma A.4. of Van Keilegom and Veraverbeke (1997b) and the bounds on $R_{n3}(t)$ and $R_{n4}(t)$, we get

$$\sup_{0 \le t \le T} \left| -\int_0^t \varphi_x'(\bar{H}_{xh}(s))dH_{xh}^u(s) + \int_0^t \varphi_x'(\bar{H}_x(s))dH_x^u(s) \right| = O((nh_n)^{-1/2} (\log n)^{1/2}) \text{ a.s.}$$

This leads to $\sup_{0\le t\le T}|R_{n1}(t)| = O((nh_n)^{-1}\log n)$ a.s. Furthermore in Lemma 3 below, we show $\sup_{0\le t\le T}|R_{n2}(t)| = O((nh_n)^{-1})$ a.s. which finishes the proof of this theorem.

We still have to prove the two lemmas used above.

Lemma 2. Under the conditions of Theorem 1, as $n \to +\infty$,

$$\sup_{0\le t\le T}\left|-\int_0^t(\varphi_x'(\bar H_{xh}(s)) - \varphi_x'(\bar H_x(s)))d(H_{xh}^u(s) - H_x^u(s))\right| = O((nh_n)^{-3/4}(\log n)^{3/4}) \text{ a.s.}$$

Proof. Divide $[0,T]$ into $k_n = O((nh_n)^{1/2}(\log n)^{-1/2})$ subintervals $[t_i,t_{i+1}]$ of length $O((nh_n)^{-1/2}(\log n)^{1/2})$. We have ,as in the proof of Lemma 2 of Lo and Singh (1985), that

$$\begin{aligned}
&\sup_{0\le t\le T}\left|-\int_0^t(\varphi_x'(\bar H_{xh}(s)) - \varphi_x'(\bar H_x(s)))d(H_{xh}^u(s) - H_x^u(s))\right| \\
&\le 2\max_{1\le i\le k_n}\sup_{t_i\le y\le t_{i+1}}|\varphi_x'(\bar H_{xh}(y)) - \varphi_x'(\bar H_x(y)) - \varphi_x'(\bar H_{xh}(t_i)) + \varphi_x'(\bar H_x(t_i))| \\
&+k_n\sup_{0\le t\le T}|\varphi_x'(\bar H_{xh}(t)) - \varphi_x'(\bar H_x(t))|\max_{1\le i\le k_n}|H_{xh}^u(t_{i+1}) - H_x^u(t_{i+1}) - H_{xh}^u(t_i) + H_x^u(t_i)| \\
&\le 2\max_{1\le i\le k_n}\sup_{t_i\le y\le t_{i+1}}\varphi_x''(\bar H_x(t_{i+1}))|H_{xh}^u(y) - H_x^u(y) - H_{xh}^u(t_i) + H_x^u(t_i)| \\
&+k_n\sup_{0\le t\le T}|\varphi_x'(\bar H_{xh}(t)) - \varphi_x'(\bar H_x(t))|\max_{1\le i\le k_n}|H_{xh}^u(t_{i+1}) - H_x^u(t_{i+1}) - H_{xh}^u(t_i) + H_x^u(t_i)| \\
&+O((nh_n)^{-1}\log n).
\end{aligned}$$

In the last inequality we used a second order Taylor expansion and Lemma A.4. of Van Keilegom and Veraverbeke (1997b). As was done in Lemma 2.1 of the same article, we can prove that each of the terms on the right hand side is $O((nh_n)^{-3/4}(\log n)^{3/4})$ a.s.

Lemma 3. Assume (C1), (C2), $H_x(t)$ satisfies (C3) in $[0,T]$ with $T < T_{H_x}$, $h_n \to 0$, $\frac{\log n}{nh_n} \to 0$, φ_x satisfies (C8), then as $n \to +\infty$,

$$\begin{aligned}
&\sup_{0\le t\le T}\left|-\sum_{Z_i\le t,\delta_i=1}(\varphi_x(\bar H_{xh}(Z_i^-)) - \varphi_x(\bar H_{xh}(Z_i^-) - w_{ni}(x,h_n)) - \varphi_x'(\bar H_{xh}(Z_i))w_{ni}(x,h_n))\right| \\
&= O((nh_n)^{-1}) \text{ a.s.}
\end{aligned}$$

Proof. Because $H_x(T) < 1$ and $H_{xh}(T) \to H_x(T)$ a.s. (Lemma A.2. Van Keilegom and Veraverbeke (1997b)), we may suppose that $T < T_{H_{xh}}$. If $t < T$, then after applying a second order Taylor expansion, we get

$$-\sum_{Z_i\le t,\delta_i=1}(\varphi_x(\bar H_{xh}(Z_i^-)) - \varphi_x(\bar H_{xh}(Z_i^-) - w_{ni}(x,h_n)) - \varphi_x'(\bar H_{xh}(Z_i))w_{ni}(x,h_n))$$

$$= -\frac{1}{2} \sum_{Z_i \le t, \delta_i = 1} \varphi_x''(\varepsilon_i) w_{ni}^2(x, h_n)$$

with ε_i between $\bar{H}_{xh}(Z_i)$ and $\bar{H}_{xh}(Z_i) + w_{ni}(x, h_n)$.
Hence

$$\sup_{0 \le t \le T} \left| - \sum_{Z_i \le t, \delta_i = 1} (\varphi_x(\bar{H}_{xh}(Z_i^-)) - \varphi_x(\bar{H}_{xh}(Z_i^-) - w_{ni}(x, h_n)) - \varphi_x'(\bar{H}_{xh}(Z_i)) w_{ni}(x, h_n)) \right|$$
$$\le \frac{1}{2} \varphi_x''(\bar{H}(T)) \sum_{i=1}^{n} w_{ni}^2(x, h_n) = O((nh_n)^{-1}) \qquad \text{a.s.}$$

5 Weak convergence

In this section, we show the weak convergence of the copula-graphic estimator $F_{xh}(t)$ in the space $D[0, T]$ of right continuous functions with left hand limits, endowed with the Skorokhod topology. Before we go to the main theorem, we give two lemmas about the bias and variance of this estimator.

Lemma 4. Assume (C1), (C2), $H_x(t)$ and $H_x^u(t)$ satisfy (C3) and (C5) in $[0, T]$, $h_n \to 0$. Then, as $n \to +\infty$

$$\sup_{0 \le t \le T} \left| \sum_{i=1}^{n} w_{ni}(x, h_n) E g_{tx}(Z_i, \delta_i) + \frac{\mu_2^K h_n^2}{2\varphi_x'(\bar{F}_x(t))} \left(\int_0^t \varphi_x''(\bar{H}_x(s)) \ddot{H}_x(s) dH_x^u(s) - \int_0^t \varphi_x'(\bar{H}_x(s)) d\ddot{H}_x^u(s) \right) \right| = o(h_n^2) + O(n^{-1}).$$

Proof. For fixed $t \le T$,

$$\sum_{i=1}^{n} w_{ni}(x, h_n) E g_{tx}(Z_i, \delta_i) = \frac{-1}{\varphi_x'(\bar{F}_x(t))} \times$$
$$\left(\int_0^t \varphi_x''(\bar{H}_x(s))(EH_{xh}(s) - H_x(s)) dH_x^u(s) - \int_0^t \varphi_x'(\bar{H}_x(s)) d(EH_{xh}^u(s) - H_x^u(s)) \right)$$

By Lemma A.1.b of Van Keilegom and Veraverbeke (1997b), we get our result.

Lemma 5. Assume (C1), (C2), $H_x(t)$ and $H_x^u(t)$ satisfy (C3) in $[0, T]$ with $T < T_{H_x}$ and φ_x satisfies (C8), $h_n \to 0$, $nh_n \to +\infty$. Then, as $n \to +\infty$

$$\sup_{0 \le t \le T} \left| \sum_{i=1}^{n} w_{ni}^2(x, h_n) \mathrm{Cov}(g_{tx}(Z_i, \delta_i), g_{ts}(Z_i, \delta_i)) - \frac{1}{nh_n} \Gamma_x(t, s) \right| = o((nh_n)^{-1})$$

where

$$\Gamma_x(t,s) = \frac{||K||_2^2}{\varphi_x'(\bar{F}_x(t))\varphi_x'(\bar{F}_x(s))}\left\{\int_0^{\min(t,s)} \varphi_x'(\bar{H}_x(z))^2 dH_x^u(z)\right.$$

$$+\int_0^{\min(t,s)} (\varphi_x''(\bar{H}_x(w))\bar{H}_x(w) + \varphi_x'(\bar{H}_x(w)))\int_0^w \varphi_x''(\bar{H}_x(y))dH_x^u(y)dH_x^u(w) \tag{6}$$

$$+\int_0^{\min(t,s)} \varphi_x''(\bar{H}_x(w)) \int_w^{\max(t,s)} (\varphi_x''(\bar{H}_x(y))\bar{H}_x(y) + \varphi_x'(\bar{H}_x(y)))dH_x^u(y)dH_x^u(w)$$

$$\left.-\int_0^t(\varphi_x''(\bar{H}_x(y))\bar{H}_x(y) + \varphi_x'(\bar{H}_x(y)))dH_x^u(y)\int_0^s(\varphi_x''(\bar{H}_x(w))\bar{H}_x(w) + \varphi_x'(\bar{H}_x(w)))dH_x^u(w)\right\}$$

Proof. Some straightforward calculations show that

$$\mathrm{Cov}(g_{tx}(Z_i,\delta_i), g_{ts}(Z_i,\delta_i)) = \frac{1}{\varphi_x'(\bar{F}_x(t))\varphi_x'(\bar{F}_x(s))}\left\{\int_0^{\min(t,s)} \varphi_x'(\bar{H}_x(z))^2 dH_{x_i}^u(z)\right.$$

$$+\int_0^{\min(t,s)}\int_0^w \varphi_x''(\bar{H}_x(y))dH_x^u(y)[\varphi_x''(\bar{H}_x(w))\bar{H}_{x_i}(w)dH_x^u(w) + \varphi_x'(\bar{H}_x(w))dH_{x_i}^u(w)]$$

$$+\int_0^{\min(t,s)} \varphi_x''(\bar{H}_x(w)) \int_w^{\max(t,s)} [\varphi_x''(\bar{H}_x(y))\bar{H}_{x_i}(y)dH_x^u(y) + \varphi_x'(\bar{H}_x(y))dH_{x_i}^u(y)]dH_x^u(w)$$

$$-\int_0^t[\varphi_x''(\bar{H}_x(y))\bar{H}_{x_i}(y)dH_x^u(y) + \varphi_x'(\bar{H}_x(y))dH_{x_i}^u(y)] \times$$

$$\left.\int_0^s[\varphi_x''(\bar{H}_x(w))\bar{H}_{x_i}(w)dH_x^u(w) + \varphi_x'(\bar{H}_x(w))dH_{x_i}^u(w)]\right\}$$

from which the result follows via standard calculations of asymptotic variances.

Theorem 2. Assume (C1), (C2), $H_x(t)$ and $H_x^u(t)$ satisfy (C5), (C6), (C7) in $[0,T]$ with $T < T_{H_x}$ and φ_x satisfies (C8).

(a) If $nh_n^5 \to 0$ and $\frac{(\log n)^3}{nh_n} \to 0$, then, under (1), as $n \to +\infty$,

$$(nh_n)^{-1/2}(F_{xh}(\cdot) - F_x(\cdot)) \to W(\cdot|x) \qquad \text{in } D[0,T]$$

(b) If $h_n = Cn^{-1/5}$ for some $C > 0$, then, under (1), as $n \to +\infty$,

$$(nh_n)^{-1/2}(F_{xh}(\cdot) - F_x(\cdot)) \to \widetilde{W}(\cdot|x) \qquad \text{in } D[0,T]$$

where $W(\cdot|x)$ and $\widetilde{W}(\cdot|x)$ are Gaussian processes with covariance function given by (6) and for $\widetilde{W}(\cdot|x)$, mean function given by

$$b_{tx} = \frac{-C^{5/2}\mu_2^K}{2\varphi_x'(\bar{F}_x(t))} \int_0^t [\varphi_x''(\bar{H}_x(s))\ddot{H}_x(s)dH_x^u(s) - \varphi_x'(\bar{H}_x(s))d\ddot{H}_x^u(s)].$$

Remark. Note that when lifetime and censoring time are independent ($\varphi_x(t) = -\log(t)$), we obtain the well-known formulas for the asymptotic mean and variance of the Beran estimator as in Van Keilegom and Veraverbeke (1997a).

Proof. From Theorem 1 and Lemma 4, we find

$$F_{xh}(t) - F_x(t) = \sum_{i=1}^{n} w_{ni}(x, h_n)\xi_{tx}(Z_i, \delta_i) + h_n^2\bar{b}_{tx} + \bar{R}_n(t)$$

where $\xi_{tx}(Z_i, \delta_i) = g_{tx}(Z_i, \delta_i) - Eg_{tx}(Z_i, \delta_i)$, $\sup_{0\le t\le T} |\bar{R}_n(t)| = O((nh_n)^{-3/4}(\log n)^{3/4}) + o(h_n^2)$ a.s. and $\bar{b}_{tx} = \frac{-\mu_2^K}{2\varphi_x'(\bar{F}_x(t))} \int_0^t [\varphi_x''(\bar{H}_x(s))\ddot{H}_x(s)dH_x^u(s) - \varphi_x'(\bar{H}_x(s))d\ddot{H}_x^u(s)]$. The bias $(nh_n)^{1/2}h_n^2\bar{b}_{tx}$ is $o(1)$ under conditions (a) and equals b_{tx} under conditions (b). Hence it suffices to prove the weak convergence of $W_{hx}(\cdot) = (nh_n)^{1/2} \sum_{i=1}^{n} w_{ni}(x, h_n)\xi_{\cdot x}(Z_i, \delta_i)$ to the Gaussian process $W(\cdot|x)$ with mean zero and covariance function $\Gamma_x(t, s)$.

This will be done in two steps. First we show the convergence of the finite dimensional distributions. Next we verify the asymptotic tightness by Theorem 2.11.9 (Bracketing central limit theorem) of van der Vaart and Wellner (1996).

Convergence of the finite dimensional distributions is that for any $q = 1, 2, \ldots$ and any $0 \le t_1 \le \ldots \le t_q \le T$: $(W_{hx}(t_1), W_{hx}(t_2), \ldots, W_{hx}(t_q)) \xrightarrow{D} N(0, \Gamma_x(t_i, t_j))$. Since $W_{hx}(t_i) = \sum_{i=1}^{n} W_{nki}$ where $W_{nki} = (nh_n)^{1/2}w_{nk}(x, h_n)\xi_{t_ix}(Z_k, \delta_k)$, it suffices to check that (see e.g. Araujo and Giné (1980)),

$$\lim_{n\to+\infty} \sum_{k=1}^{n} E(W_{nki}W_{nkj}) = \Gamma_x(t_i, t_j) \qquad (1 \le i, j \le q)$$

$$\lim_{n\to+\infty} \sum_{k=1}^{n} \int_{\{|W_{nk}|>\varepsilon\}} |W_{nk}|^2 dP = 0$$

for every $\varepsilon > 0$, where $|W_{nk}|^2 = \sum_{i=1}^{q} W_{nki}^2$. Now, applying Lemma 5,

$$\sum_{k=1}^{n} E(W_{nki}W_{nkj}) = (nh_n)\sum_{k=1}^{n} w_{nk}^2(x, h_n)\text{Cov}(g_{t_ix}(Z_k, \delta_k), g_{t_jx}(Z_k, \delta_k)) = \Gamma_x(t_i, t_j) + o(1).$$

Since the functions $\xi_{t_ix}(Z_k,\delta_k)$ are uniformly bounded, it follows that $\max\limits_{1\le k\le n}|W_{nk}| = O((nh_n)^{-1/2})$ a.s. and $\sum\limits_{k=1}^{n}|W_{nk}|^2 = O(1)$ a.s., and hence,

$$\sum_{k=1}^{n}\int_{\{|W_{nk}|>\varepsilon\}}|W_{nk}|^2dP \le O(1)P(\max_{1\le k\le n}|W_{nk}|>\varepsilon) = o(1).$$

To prove the asymptotic tightness, we denote the process $W_{hx}(t)$ as $W_{hx}(t) = \sum\limits_{i=1}^{n} Z_{ni}(t)$ where $Z_{ni}(t) = (nh_n)^{1/2}w_{ni}(x,h_n)\xi_{tx}(Z_i,\delta_i)$.

To verify the three conditions of Theorem 2.11.9 of van der Vaart and Wellner (1996), we put on $\mathcal{F}=[0,T]$, the semimetric

$$\begin{aligned}\rho(t,t') \;=\; \max\Bigg\{&\left|\frac{-1}{\varphi_x'(\bar F_x(t))}+\frac{1}{\varphi_x'(\bar F_x(t'))}\right|, |\varphi_x'(\bar H_x(t))-\varphi_x'(\bar H_x(t'))|,\\ &\sup_{x'\in[0,1]}|H_{x'}^u(t)-H_{x'}^u(t')|, |H_x(t)-H_x(t')|, \sup_{x'\in[0,1]}\sqrt{|H_{x'}^u(t)-H_{x'}^u(t')|}\Bigg\}.\end{aligned}$$

In the third condition, we need the bracketing number $N_{[\,]}(\varepsilon,\mathcal{F},L_2^n)$. This number is defined as the minimal number of sets in a partition of $\mathcal{F}=[0,T]=\bigcup_j\mathcal{F}_{\varepsilon j}$ such that for every set $\mathcal{F}_{\varepsilon j}$:

$$\sum_{i=1}^{n}E\left[\sup_{t,t'\in\mathcal{F}_{\varepsilon j}}|Z_{ni}(t)-Z_{ni}(t')|^2\right]\le\varepsilon^2.$$

Let us divide $\mathcal{F}=[0,T]$ into subintervals $0=t_0\le t_1\le\ldots\le t_q=T$ where $\rho(t,t')\le C\varepsilon$ for all $t,t'\in[t_{j-1},t_j], j=1,\ldots,q$ with C some constant which we will determine further on. For the partition $\mathcal{F}=[0,t_1]\cup\bigcup\limits_{j=2}^{q}]t_{j-1},t_j]$, we find after some tedious calculations that

$$\begin{aligned}&|Z_{ni}(t)-Z_{ni}(t')| \le (nh_n)^{1/2}w_{ni}(x,h_n)\Bigg(-\frac{\varphi_x''(\bar H_x(T))}{\varphi_x'(1)}|H_x^u(t)-H_x^u(t')|\\ &+(\varphi_x''(\bar H_x(T))-2\varphi_x'(\bar H_x(T)))\left|\frac{-1}{\varphi_x'(\bar F_x(t))}+\frac{1}{\varphi_x'(\bar F_x(t'))}\right|+|\varphi_x'(\bar H_x(t))-\varphi_x'(\bar H_x(t'))|\\ &-\varphi_x'(\bar H_x(T))(|I(Z_i\le t,\delta_i=1)-I(Z_i\le t',\delta=1)|+|H_{x_i}^u(t)-H_{x_i}^u(t')|)\\ &+\frac{\varphi_x'(\bar H_x(T))}{\varphi_x'(1)}|H_x(t)-H_x(t')|\Bigg)\end{aligned}\tag{7}$$

So

$$\begin{aligned}\sup_{t,t'\in\mathcal{F}_{\varepsilon j}}|Z_{ni}(t)-Z_{ni}(t')|^2 \;\le\;& (nh_n)w_{ni}^2(x,h_n)\{C_1(C\varepsilon)^2\\ &+\; C_2(C\varepsilon)|I(Z_i\le t_j,\delta_i=1)-I(Z_i\le t_{j-1},\delta=1)|\\ &+\; \varphi_x'(\bar H_x(T))^2|I(Z_i\le t_j,\delta_i=1)-I(Z_i\le t_{j-1},\delta=1)|^2\}\end{aligned}$$

where C_1, C_2 are constants, uniquely determined by the right hand side of (7). For the appropriate choice of C, this leads to

$$\sum_{i=1}^{n} E\left[\sup_{t,t' \in \mathcal{F}_{\varepsilon j}} |Z_{ni}(t) - Z_{ni}(t')|^2\right] \leq \varepsilon^2.$$

Hence the bracketing number $N_{[\,]}(\varepsilon, \mathcal{F}, L_2^n)$ is equal to $O(\varepsilon^{-1})$ and we get

$$\int_0^{\delta_n} \sqrt{\log N_{[\,]}(\varepsilon, \mathcal{F}, L_2^n)}d\varepsilon = \int_0^{\delta_n} \sqrt{\log O(\varepsilon^{-1})}d\varepsilon \to 0$$

when $\delta_n \to 0$.

We do not need to verify the second condition of Theorem 2.11.9 in van der Vaart and Wellner (1996), since our partition of $\mathcal{F} = [0, T]$ is independent of n. As last condition we have to check whether for all $\eta > 0$,

$$\sum_{i=1}^{n} E\left[\sup_{0 \leq t \leq T} |Z_{ni}(t)|\{\sup_{0 \leq t \leq T} |Z_{ni}(t)| > \eta\}\right] \to 0 \text{ as } n \to +\infty.$$

However since $\xi_{tx}(Z_i, \delta_i)$ is bounded uniformly and $\max_{1 \leq i \leq n} w_{ni}(x, h_n) = O((nh_n)^{-1})$ a.s., this condition is satisfied. By Theorem 2.11.9 of van der Vaart and Wellner (1996), we have that $W_{hx}(\cdot) \to W(\cdot|x)$ in $D[0, T]$.

6 Example

In this section ,we apply our copula-graphic estimator on a practical dataset. The bone marrow transplantation data, which are described in Klein and Moeschberger (1997), follow 137 patients in their recovery from acute leukemia after a bone marrow transplantation. In this example, we focus on the disease-free survival time where we see that a patient can leave the study in 3 ways: with a relapse of leukemia, a disease-free death or disease-free alive at the end of study. Since we believe that time till relapse Y_x and time until death C_x are dependent, we will only work with the 83 patients who relapsed or died within the study. Furthermore we think that the time till relapse depends on the age of the patient at transplantation. In Figure 1, we show a scatterplot of age versus disease-free survival time where we distinguish between relapsed and dead patients. We note that most deaths occur around the age of 30 and that the number of relapsed patients is about the same as the number of dead patients.

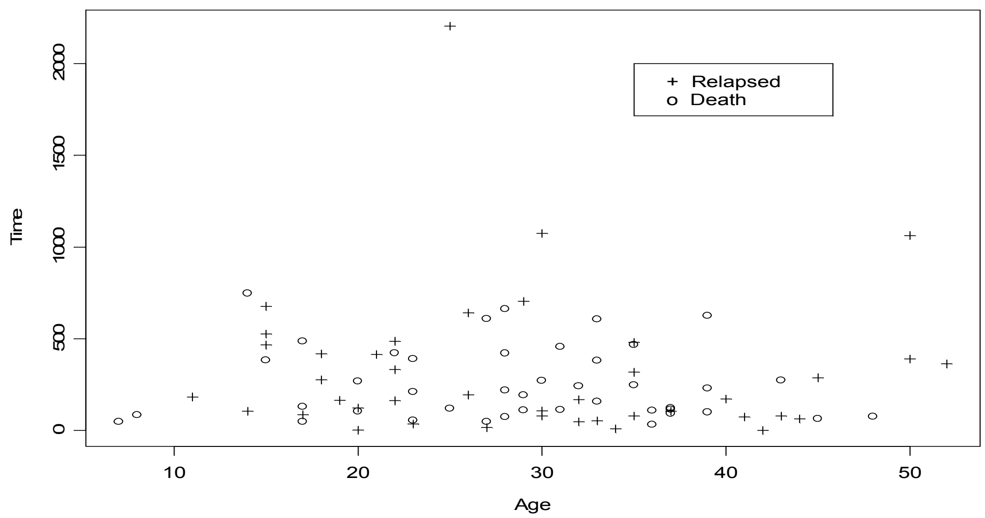

Figure 1. Scatterplot of age versus disease-free survival time

For these data, we construct the copula-graphic estimator for different choices of φ_x at ages 15 and 40. These ages are picked to represent two groups of patients, patients in puberty and still growing (age = 15) and patients who are fully grown and even start to age (age = 40). We believe that the survival function differs for each of these groups of patients.

In Figure 2, we show the survival functions at age 15 and 40, and for bandwidths 20 and 40. In each of the four plots, we have 3 choices of φ_x for which we construct the copula-graphic estimator. The solid line is the estimator when we assume independence between time till relapse and time until death ($\varphi_x(t) = -\log(t)$). The second choice (small dashed line) is the Fréchet-Hoeffding lower bound ($\varphi_x(t) = 1 - t$). With this copula, we assume that time till relapse and time until death are discordant. Informally this means that large values of time till relapse tend to be associated with small values of time until death. A more formal definition can be found in Nelsen (1999). As last choice for φ_x, we take a generator of the Frank family ($\varphi_x(t) = -\log(\frac{e^{-xt}-1}{e^{-x}-1})$) which, unlike the previous choices, depends on the covariate value x. For this family, time till relapse and time until death are discordant (resp. concordant) when the sign of the covariate x is negative (resp. positive). In this example, the covariate age is positive and so we assume with this family that time till relapse and time until death are concordant.

The association between time till relapse and time until death can be measured in several ways. Here we take Kendall's τ which is defined as $\tau(x) = 1 + 4\int_0^1 \frac{\varphi_x(t)}{\varphi'_x(t)}dt$ (Nelsen 1999) and has a range from -1 till 1. The association gets stronger when τ goes further away from zero. As we expect $\tau(x) = 0$ for the independence copula and we see that $\tau(x) = -1$ for the Fréchet-Hoeffding lower bound copula. Kendall's $\tau(x)$ is an increasing function of the covariate age for the Frank family copula. This means that we believe that the association between time till relapse and time until death is stronger at age $= 40$ then at age $= 15$.

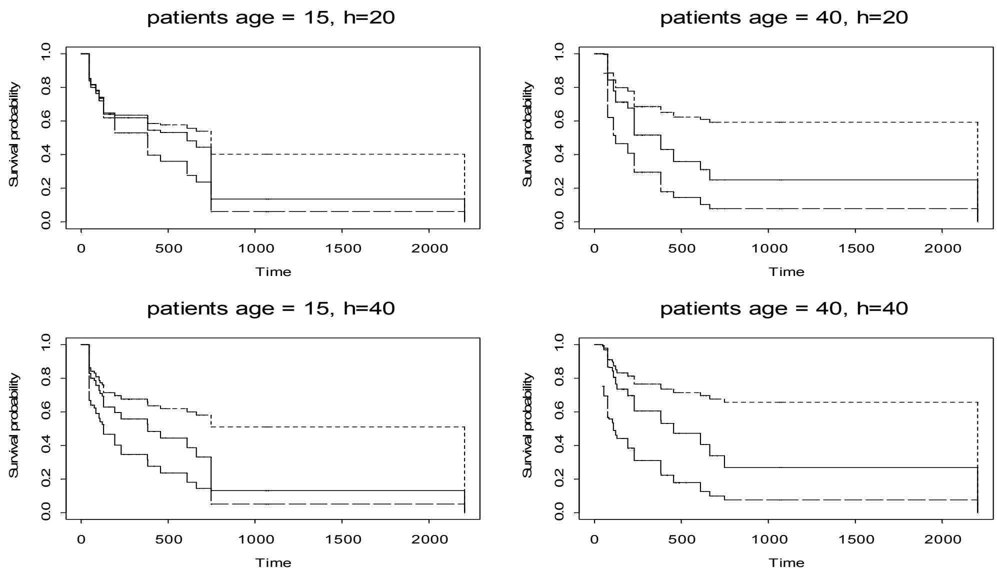

Figure 2. Copula-graphic estimator for choices of φ_x. Independence (solid line), Frank family (long dashed), Fréchet-Hoeffding lower bound (small dashed).

In Figure 2 we note that the survival function based on the Fréchet-Hoeffding lower bound copula always lies above the independence survival function, which in its turn, always lies above the Frank family survival function. When we compare the survival functions at age 15 with the functions at age 40, we get for the Frank family and the independence copula a higher survival function at age 15 than at age 40. This means a longer time till relapse for younger people than for older people. For the Fréchet-Hoeffding lower bound copula, we see the opposite result. The survival function at age 40 is higher than at age 15. Older people have in this case a longer time till relapse than younger people. From a medical point of view, the last conclusion is the correct one.

Acknowledgement

This work was supported by the Ministry of the Flemish Community (Project BIL00/28, International Scientific and Technological Cooperation) and by the IAP research network nr P5/24 of the Belgian State (Federal Office for Scientific, Technical and Cultural Affairs).

References

A. Araujo and E. Giné, (1980) *The central limit theorem for real and Banach valued random variables*, Wiley, New York.

J. P. Klein and M. L. Moeschberger, (1997) *Survival analysis: Techniques for censored and truncated data*, Springer-Verlag, New York.

S.-H. Lo and K. Singh, (1986) The product-limit estimator and the bootstrap: Some asymptotic representations, *Probab. Theory Related Fields.* **71**, 455-465.

R. B. Nelsen, (1999) *An introduction to copulas*, Springer-Verlag, New York.

A. Tsiatis, (1975) A nonidentifiability aspect of the problem of competing risks, *Proc. Nat. Acad. Sci. USA.* **72**, 20-22.

L. Rivest and M. T. Wells, (2001) A martingale approach to the copula-graphic estimator for the survival function under dependent censoring, *J. Multivariate Analysis.* **79**, 138-155.

A. W. van der Vaart and J. A. Wellner, (1996) *Weak convergence and empirical processes*, Springer-Verlag, New York.

I. Van Keilegom and N. Veraverbeke, (1996) Uniform strong convergence results for the conditional Kaplan-Meier estimator and its quantiles, *Comm. Statist.Theory Methods.* **25**, 2251-2265.

I. Van Keilegom and N. Veraverbeke, (1997a) Weak convergence of the bootstrapped conditional Kaplan-Meier process ant its quantile process, *Comm. Statist.Theory Methods.* **26**, 853-869.

I. Van Keilegom and N. Veraverbeke, (1997b) Estimation and bootstrap with censored data in fixed design nonparametric regression, *Ann. Inst. Statist. Math.* **49**, 467-491.

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