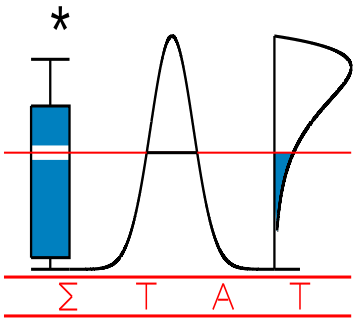
T E C H N I C A L

R E P O R T

0624

A CONDITIONAL KOZIOL-GREEN MODEL 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 conditional Koziol-Green model under dependent censoring

Roel Braekers [1] and Noël Veraverbeke
Hasselt University, Center for Statistics
Campus Diepenbeek, Agoralaan building D
3590 Diepenbeek, Belgium

Abstract

In survival analysis, we are interested in the distribution function of the lifetime of some event. Due to different practical reasons we only observe a lower bound of the true lifetime. Under independence, Kaplan and Meier (1958) developed a non-parametric estimator for the distribution function of the lifetime. However in several studies it was seen that the distribution of the censoring time also contains information about the distribution of the lifetime. Therefore Koziol and Green (1976) considered a sub-model of the previous model. They assumed that the survival function of the censoring time is a power of the survival function of the lifetime. In their paper, Veraverbeke and Cadarso Suárez (2000) extended the Koziol-Green model to a fixed design regression model and called it the conditional Koziol-Green model.

We extend, in this paper, the conditional Koziol-Green model to also accommodate for dependent censoring such that we have a model in which two types of informative censoring are introduced. On the one hand we have the relationship between the conditional distributions of the censoring time and the lifetime, and on the other hand we see a possible dependence of the censoring time on the lifetime via a dependence structure, given on the joint distribution function of both variables. This generalized conditional Koziol-Green model is a sub-model of the copula-graphic model given in Braekers and Veraverbeke (2005) and also has the property that the observed lifetime and the censoring indicator are conditionally independent, given the covariate value.

We derive in this model a copula-graphic estimator for the conditional distribution of the lifetime. For this estimator we establish an exponential bound which serves as the starting point for an almost sure consistency result. Furthermore we give an almost sure asymptotic representation and an asymptotic normality result. Afterwards we apply this estimator on a real data set about the survival of Atlantic halibut.

Keywords: Archimedean copula, dependent censoring, fixed design regression, Koziol-Green model
MSC2000 Classification: Primary 62N01, Secondary 62N02, 62G08

Running head: Koziol-Green under dependent censoring

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

[1] Corresponding Author
E-mail: roel.braekers@uhasselt.be

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 this assumption clearly does not hold. For example in medicine when the event of interest is death due to a given disease and the censoring event is death due to other diseases. In industrial testing, it may occur that some piece of equipment is taken away (is censored) because it shows some sign of future failure. Therefore a dependence model is used in which the dependence structure is given by specifying a copula for the joint distribution of Y_x and C_x. Assume that the joint survival function of the response Y_x and the censoring variable C_x at x can be written 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 Archimedean copulas. As in Braekers and Veraverbeke (2005), 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) and given by

$$\varphi_x^{[-1]}(s) = \begin{cases} \varphi_x^{-1}(s) & 0 \leq s \leq \varphi_x(0) \\ 0 & \varphi_x(0) \leq s \leq +\infty \end{cases}.$$

We note from (1) that,

$$1 - H_x(t) = \bar{H}_x(t) = S_x(t, t) = \varphi_x^{[-1]}(\varphi_x(\bar{F}_x(t)) + \varphi_x(\bar{G}_x(t))). \tag{2}$$

However in the design of some clinical trials, we see another type of informative censoring in which the distribution function of the failure time and the censoring time are related. Koziol and Green (1976) considered a sub-model for the Kaplan-Meier estimator in which they assumed that the survival function of the censoring variable is a power of the survival function of the failure time. This sub-model has the advantage that the estimator for the distribution function of the failure time has a simpler form.

Other results, like a test-procedure to check the validity of this assumption (Csörgő(1988)) were derived. Veraverbeke and Cadarso Suárez (2000) extended this model for the fixed design regression situation.

In this paper we will further extend this sub-model to the case where the failure time Y_x depends on the censoring variable C_x. We therefore use the fact that the classical Koziol-Green model is characterized by the conditional independence of Z_x and δ_x. Translating the latter property into our model (1) leads to the following assumption: for each covariate value $x \in [0,1]$,

$$\bar{G}_x(t) = \varphi_x^{[-1]}(\beta_x \varphi_x(\bar{F}_x(t))), \qquad \forall t > 0 \tag{3}$$

where $\beta_x > 0$ is a constant depending only on x.

When we consider both types of informative censoring, we rewrite (2) as

$$\begin{aligned} \bar{H}_x(t) &= \varphi_x^{[-1]}\left(\varphi_x(\bar{F}_x(t)) + \beta_x \varphi_x(\bar{F}_x(t))\right) \\ &= \varphi_x^{[-1]}\left((\beta_x + 1)\varphi_x(\bar{F}_x(t))\right). \end{aligned} \tag{4}$$

This relation will be used to find a conditional distribution estimator F_{xh} for F_x where $x \in]0,1[$ is a fixed design value. We organize this paper as follows. In Section 2, we define the distribution function estimator F_{xh} and show that it is an extension of the conditional Koziol-Green estimator, as it was studied by Veraverbeke and Cadarso Suárez (2000). After specifying some assumptions in Section 3, we give for this estimator an exponential bound with consistency result in Section 4. In Section 5, we derive an almost sure asymptotic representation which we use in Section 6 to find an asymptotic normality result for this estimator. In Section 7 we apply the estimator to a practical situation in which we explore different choices for the generator function φ_x. In the Appendix we give the proofs of the results in Section 4, 5 and 6.

2 The conditional Koziol-Green model

We develop in this section a non-parametric estimator for the conditional distribution function F_x of the failure time Y_x under two different types of informative censoring. As first type, we assume that the failure time Y_x depends on the censoring variable C_x as given in (1). While for the second type, we assume that the distribution of the censoring variable C_x is related to the distribution function of Y_x as given by (3).

From (4), we find an estimator for F_x since we can rewrite this equation as

$$\bar{F}_x(t) = \varphi_x^{[-1]}\left(\gamma_x \varphi_x(\bar{H}_x(t))\right) \tag{5}$$

with

$$\gamma_x = \frac{1}{\beta_x + 1} = P(\delta_x = 1).$$

The last equation follows from

$$\begin{aligned} p_{x1} = P(\delta_x = 1) &= \int_0^{+\infty} -\frac{\partial}{\partial t_1} S_x(t_1, t_2)\Big|_{t=t_1=t_2} dt = \int_0^{+\infty} \frac{\varphi_x'(\bar{F}_x(t))}{\varphi_x'(\bar{H}_x(t))} dF_x(t) \\ p_{x0} = P(\delta_x = 0) &= \int_0^{+\infty} -\frac{\partial}{\partial t_2} S_x(t_1, t_2)\Big|_{t=t_1=t_2} dt = \beta_x \int_0^{+\infty} \frac{\varphi_x'(\bar{F}_x(t))}{\varphi_x'(\bar{H}_x(t))} dF_x(t) \end{aligned}$$

so that $\beta_x = \frac{p_{x0}}{p_{x1}}$.

To find an estimator for F_x in this model, we replace in (5), the different quantities $H_x(t)$ and γ_x by estimators. As in other work with a non-parametric regression problem (Veraverbeke and Cadarso Suárez (2000), Braekers and Veraverbeke (2005)), we consider estimators which involve a sequence of smoothing weights $\{w_{ni}(x, h_n)\}$, depending on a positive bandwidth sequence $\{h_n\}$, tending to zero, as $n \to +\infty$. In our present situation of fixed design points, it is customary to take the Gasser-Müller type weights, given by

$$\begin{aligned} w_{ni}(x, h_n) &= \frac{1}{c_n(x, h_n)} \int_{x_{i-1}}^{x_i} \frac{1}{h_n} K\left(\frac{x - z}{h_n}\right) dz, \qquad (i = 1, \ldots, n) \\ c_n(x, h_n) &= \int_0^{x_n} \frac{1}{h_n} K\left(\frac{x - z}{h_n}\right) dz. \end{aligned} \tag{6}$$

Here $x_0 = 0$ and K is a known probability density function (kernel).

For the conditional distribution function $H_x(t)$, we take a Stone type estimator (Stone (1977)) given by

$$H_{xh}(t) = \sum_{i=1}^{n} w_{ni}(x, h_n) \mathrm{I}(Z_i \leq t).$$

A similar estimator is taken for the exponent γ_x and is given by

$$\gamma_{xh} = \sum_{i=1}^{n} w_{ni}(x, h_n) I(\delta_i = 1).$$

Hence we find an estimator for the conditional distribution function $F_x(t)$ by

$$\bar{F}_{xh}(t) = \varphi_x^{[-1]} \left(\gamma_{xh} \varphi_x(\bar{H}_{xh}(t))\right).$$

Note that the estimator $\bar{F}_{xh}(t)$ has a simpler structure than the copula-graphic estimator of Braekers and Veraverbeke (2005) for the more general model under dependent censoring. Furthermore we see that in our estimator the estimator for γ_x only depends on the δ_i while the estimator for $H_x(t)$ only depends on the Z_i. This result follows from assumption (3), which is equivalent to the assumption that Z_x and δ_x are conditionally independent.

If we take $\varphi_x(t) = -\log(t)$, we see that this estimator equals the estimator of Veraverbeke and Cadarso Suárez (2000) as we expected.

3 Regularity conditions

For the design points $x_1, \ldots, x_n$ we write $\underline{\Delta}_n = \min_{1 \leq i \leq n} (x_i - x_{i-1})$ and $\bar{\Delta}_n = \max_{1 \leq i \leq n} (x_i - x_{i-1})$. The notations $\|K\|_\infty = \sup_{u \in \mathbb{R}} K(u)$, $\|K\|_2^2 = \int_{-\infty}^{+\infty} K^2(u) du$, $\mu_1^K = \int_{-\infty}^{+\infty} uK(u) du$, $\mu_2^K = \int_{-\infty}^{+\infty} u^2 K(u) du$ 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.

The assumption (C1) expresses that the chosen design points are asymptotically equidistant points, selected uniformly over the whole interval $[0, 1]$. This implies that, for $c_n(x, h_n)$ defined in (6), $c_n(x, h_n) = 1$ for n sufficiently large. Therefore we may take $c_n(x, h_n) = 1$ in all proofs of the asymptotic results.

If L is any distribution, then T_L denotes the right endpoint of its support ($T_L = \inf\{t : L(t) = L(+\infty)\}$). We note that $T_{H_x} = T_{F_x} = T_{G_x}$. To obtain our results, we need some smoothness conditions. For a fixed $0 < T < T_{F_x}$,

(C3) $\dot{F}_x(t) = \frac{\partial}{\partial x} F_x(t)$, $\ddot{F}_x(t) = \frac{\partial^2}{\partial x^2} F_x(t)$ exist and are continuous in $(x, t) \in [0, 1] \times [0, T]$

(C4) $\dot{\beta}_x = \frac{\partial}{\partial x}\beta_x$, $\ddot{\beta}_x = \frac{\partial^2}{\partial x^2}\beta_x$ exist and are continuous in $x \in [0, 1]$

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

(C5) $\varphi'_x(v) = \frac{\partial}{\partial v}\varphi_x(v)$ 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 Exponential Bound and Strong Consistency

In the first theorem we establish conditions under which $F_{xh}(t)$ is a strongly consistent estimator for $F_x(t)$ and we also obtain the rate of this convergence. These results follow from an exponential inequality result. We postpone the proofs of these results to the Appendix.

Theorem 1. Assume (C1) - (C5), $h_n \to 0$, $T < T_{F_x}$,

(a) For $\varepsilon > 0$, n sufficiently large and

$$\begin{aligned} \eta_1 &\geq 2(\|\dot{\gamma}_x\|_\infty \bar{\Delta}_n + \|\ddot{\gamma}_x\|_\infty \mu_2^K h_n^2), \\ \eta_2 &\geq \max\left(\sqrt{6}\|K\|_2 (nh_n)^{-1/2}, 2(\|\dot{H}_x\|_\infty \bar{\Delta}_n + \|\ddot{H}_x\|_\infty \mu_2^K h_n^2)\right) \end{aligned}$$

we have

$$P\left(\sup_{0\leq t\leq T} |F_{xh}(t) - F_x(t)| > \varepsilon\right) \leq 4e^{-C_1\eta_1^2 nh_n} + d_0\eta_2 nh_n e^{-d_1 nh_n\eta_2^2/4}$$

where C_1, d_0, d_1 are constants and η_1, η_2 satisfy $\eta_1 = \frac{\varepsilon}{2M_1(\gamma_x - \eta_1)}$ and $\eta_2 = \frac{\varepsilon}{2M_2(H_x(T) - \eta_2)}$ with $M_1(z) = \max\limits_{0\leq y\leq 1} \frac{-\varphi_x(y)}{\varphi'_x(\varphi_x^{-1}(z\varphi_x(y)))}$ and $M_2(z) = \max\limits_{0\leq y\leq 1} \frac{y\varphi'_x(z)}{\varphi'_x(\varphi_x^{-1}(z\varphi_x(y)))}$.

(b) If $\frac{nh_n^5}{\log n} = O(1)$, then, as $n \to +\infty$,

$$\sup_{0\leq t\leq T} |F_{xh}(t) - F_x(t)| = O\left((nh_n)^{-1/2}(\log n)^{1/2}\right) \text{ a.s.}$$

5 Almost sure representation

As in the work of Veraverbeke and Cadarso Suárez (2000) for the conditional Koziol-Green estimator and in Braekers and Veraverbeke (2005) for the copula-graphic estimator, we derive in this section an almost sure representation for the estimator $F_{xh}(t)$. In this result we rewrite this estimator as a weighted sum plus a remainder term. This representation forms a basic tool in finding further asymptotic results. For the proof of Theorem 2 we again refer to the Appendix.

Theorem 2. Assume (C1) - (C5), $h_n \to 0$, $\frac{\log n}{nh_n} \to 0$, $\frac{nh_n^5}{\log n} = O(1)$, $T < T_{F_x}$. Then, for $t < T_{F_x}$,

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

where

$$g_{tx}(Z_i, \delta_i) = -\frac{\varphi_x(\bar{H}_x(t))}{\varphi'_x(\bar{F}_x(t))} (I(\delta_i = 1) - \gamma_x) + \frac{\gamma_x \varphi'_x(\bar{H}_x(t))}{\varphi'_x(\bar{F}_x(t))} (I(Z_i \leq t) - H_x(t))$$

and as $n \to +\infty$,

$$\sup_{0\leq t\leq T} |R_n(x, t)| = O((nh_n)^{-1} \log n) \text{ a.s.}$$

6 Asymptotic normality

In Theorem 3, we show the asymptotic normality of $(nh_n)^{1/2}(F_{xh}(t) - F_x(t))$. Due to the asymptotic representation in Theorem 2, we only have to use the main term to find the bias and variance expressions of this result. The proof of Theorem 3 is given in the Appendix.

Theorem 3. Assume (C1) - (C5), $T < T_{H_x}$,

(a) if $nh_n^5 \to 0$ and $(nh_n)^{-1/2} \log n \to 0$, then for $t \leq T$, as $n \to \infty$,

$$(nh_n)^{1/2}(F_{xh}(t) - F_x(t)) \to N(0, s_x^2(t))$$

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

$$(nh_n)^{1/2}(F_{xh}(t) - F_x(t)) \to N(b_x(t), s_x^2(t))$$

where

$$\begin{aligned} b_{tx} &= \frac{1}{2}\mu_2^K C^{5/2} \left\{ \frac{-\varphi_x(\bar{H}_x(t))}{\varphi'_x(\bar{F}_x(t))} \ddot{\gamma}_x + \frac{\gamma_x \varphi'_x(\bar{H}_x(t))}{\varphi'_x(\bar{F}_x(t))} \ddot{H}_x(t) \right\} \\ s_x^2(t) &= ||K||_2^2 \left\{ \left(\frac{\varphi_x(\bar{H}_x(t))}{\varphi'_x(\bar{F}_x(t))} \right)^2 \gamma_x(1 - \gamma_x) + \left(\frac{\gamma_x \varphi'_x(\bar{H}_x(t))}{\varphi'_x(\bar{F}_x(t))} \right)^2 H_x(t)(1 - H_x(t)) \right\}. \end{aligned}$$

7 Example: Survival of Atlantic Halibut

We apply in this section the conditional Koziol-Green estimator on a practical data set about survival of Atlantic halibut, studied by Neilson, Waiwood and Smith (1989). An important issue was the survival time of the fish after it was caught and handled as in the commercial fishery. For this purpose they had installed special holding tanks on the research vessel in which they placed the fish. Each fish was followed until it died. However some fish, mainly large fish, were removed after 48 hours to make space for other experimental animals. So the time until death was censored by the time that the animal had spent in the holding tank. Also the fish that were alive at the end of the experiment, were treated as censored observations. The researchers recorded several covariates among which we focus on the fork length of the fish. In previous analyses of the data set, a significant effect of fork length on survival time had been found. In Figure 1 we show a scatter plot of the survival time versus the fork length of each animal, where we use + for uncensored observations and O for censored observations. The main causes of death for the fish were the stress of the new environment and infections caused by sick fishes in the tank. Therefore we believe that the survival time Y_x of a fish depends on the time that this fish has spent in the holding tank C_x, where the time in the holding tank has a negative influence on the survival time.

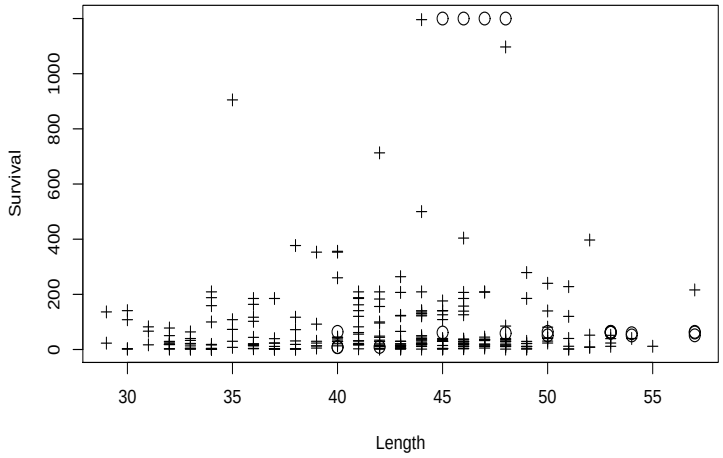


Figure 1: Atlantic halibut data set: Survival times (in hours) versus fork length (in cm). Fish died in the holding tanks: +, fish removed from the holding tanks or alive at the end of the study: O.

For these data, we construct a conditional Koziol-Green estimate for different choices of φ_x at fork lengths 32 cm and 53 cm, representing typical small fishes and typical large fishes. The four choices of the function φ_x that we will consider here, will lead each time to a different association for the dependence structure between the survival time and the time spent in the holding tank. The first choice is the independent copula ($\varphi_x(t) = -\log(t)$). This is the (possibly wrong) choice used in previous analyses of the data. In the other choices of φ_x we express that the time spent in the holding tank has a negative influence on the survival time. Nelsen (1999) formally defined this as discordance. For the second choice of φ_x we take the Fréchet-Hoeffding lower bound ($\varphi_x(t) = 1 - t$), which is the most

extreme discordance that can be considered. In the next choices we allow the generator function φ_x to depends on the fork length x. Our third choice is the Frank family 1 copula given by

$$\varphi_x(t) = -\log\left(\frac{e^{(x-20)t}-1}{e^{x-20}-1}\right)$$

and the fourth choice is the Frank family 2 copula given by

$$\varphi_x(t) = -\log\left(\frac{e^{(60-x)t}-1}{e^{60-x}-1}\right).$$

The Frank family 1 copula gives a stronger discordant association for larger fishes than for small fishes, while for the Frank family 2 copula, there is a stronger discordant association for small fishes.

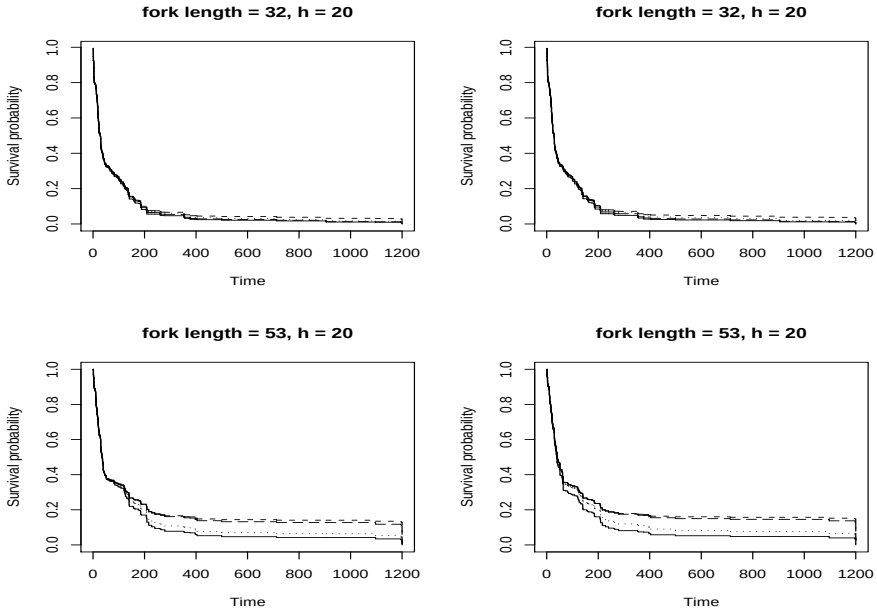


Figure 2: Copula-graphic (left column) and Koziol-Green (right column) estimates for the conditional survival function at lengths 32 cm and 53 cm with bandwidth 20. Independence (solid line), Fréchet - Hoeffding lower bound (dashed line), Frank family 1 (longdashed line) and Frank family 2 (dotted line).

In Figure 2, we show in the right column the Koziol-Green estimates for the conditional survival function $\bar{F}_x(t)$ at fork lengths of 32 cm and 53 cm for bandwidth 20. In the left column we give the copula-graphic estimates of Braekers and Veraverbeke (2005) for this conditional survival function under a general model. By comparing both estimates we can verify whether the Koziol-Green sub-model is satisfied. In each of the four plots, we construct the copula-graphic estimator for the four choices of φ_x. We use in this data set the Gasser-Müller weights with a biquadratic kernel given by $K(z) = (15/16)(1-z^2)^2 I(|z| \leq 1)$. As we saw in Figure 1, the covariate fork length of a fish is measured crudely on a scale of whole centimeters so that the observations form vertical lines on this plot. It is therefore possible to treat this covariate as fixed. It is also easy to see that our results for the copula-graphic estimator remain valid in the interval $[25, 60]$ for the covariate x, instead of the standard interval $[0, 1]$. The choice of the bandwidth is selected here for illustration purpose only. We also calculated estimates under other bandwidths but the results did not change. It is possible to set up a bandwidth selection criterium using, for example,

the asymptotic mean squared error expression, but this would lead us into a field of research that we do not enter at this moment.

We note in Figure 2 that the different estimates for the conditional survival function lie close together in each of the four plots and lie even almost on top of each other in the plots of the first row. This means that the choice of the generator function φ_x does not have a great influence on the survival time of small fishes. In the plots at the fork length of 53 cm we see that both the copula-graphic and the Koziol-Green estimates can be divided in two groups. The Fréchet-Hoeffding lower bound copula and the Frank family 1 copula give estimates that lie almost on top of each other but that are clearly different from the estimates of the independent copula and the Frank family 2 copula which form the second group. By this division in two groups, we see that this data set reacts differently to two different situations. The choices of φ_x in the first group have in common that they assume a large discordant association between survival time and time spent in the holding tank for larger fishes. In the second group, the choices of φ_x assume practically no discordant association for larger fishes. This influences the estimates for the survival function. With a φ_x from the first group, the estimated survival function for larger fish is higher than with a φ_x from the second group (in particular the φ_x that describes independence). This means that when we ignore stress in the fish caused by the catch, the handling and the living conditions in the holding tank, we underestimate the true survival time for larger fishes. To finish this section, we compare the estimates for the Koziol-Green sub-model with the estimates for the general copula-graphic model of Braekers and Veraverbeke (2005). We note that there is not much difference between the plots in the two columns. Therefore we know that the Koziol-Green assumption is satisfied in this data set and the Koziol-Green sub-model gives a better insight into the data.

Appendix: Proofs

Proof of Theorem 1: (a) To establish strong consistency of the estimator $F_{xh}(t)$, we first use the bivariate mean value theorem.

$$\begin{aligned} F_{xh}(t) - F_x(t) &= \bar{F}_x(t) - \bar{F}_{xh}(t) \\ &= -\frac{\varphi_x(\varepsilon_2(t))}{\varphi_x'(\varphi_x^{-1}(\varepsilon_1\varphi_x(\varepsilon_2(t))))}(\gamma_{xh} - \gamma_x) + \frac{\varepsilon_1\varphi_x'(\varepsilon_2(t))}{\varphi_x'(\varphi_x^{-1}(\varepsilon_1\varphi_x(\varepsilon_2(t))))}(H_{xh}(t) - H_x(t)) \end{aligned}$$

with ε_1 between γ_{xh} and γ_x, and $\varepsilon_2(t)$ between $\bar{H}_{xh}(t)$ and $\bar{H}_x(t)$.

Hence for all $\varepsilon > 0$, $\eta_1 > 0$, $\eta_2 > 0$,

$$\begin{aligned} & P\left(\sup_{0\le t\le T}|F_{xh}(t) - F_x(t)| > \varepsilon\right) \\ \le\ & P\left(M_1(\varepsilon_1)|\gamma_{xh} - \gamma_x| > \frac{\varepsilon}{2}\right) + P\left(\sup_{0\le t\le T} M_2(\varepsilon_2(t)) \sup_{0\le t\le T}|H_{xh}(t) - H_x(t)| > \frac{\varepsilon}{2}\right) \\ \le\ & P\left(M_1(\varepsilon_1)|\gamma_{xh} - \gamma_x| > \frac{\varepsilon}{2}, |\gamma_{xh} - \gamma_x| \le \eta_1\right) + P\left(|\gamma_{xh} - \gamma_x| > \eta_1\right) + P\left(\sup_{0\le t\le T}|H_{xh}(t) - H_x(t)| > \eta_2\right) \\ +\ & P\left(\sup_{0\le t\le T} M_2(\varepsilon_2(t)) \sup_{0\le t\le T}|H_{xh}(t) - H_x(t)| > \frac{\varepsilon}{2}, \sup_{0\le t\le T}|H_{xh}(t) - H_x(t)| \le \eta_2\right) \end{aligned}$$

If we assume that $0 < \eta_1 < \gamma_x$ and $0 < \eta_2 < \bar{H}_x(T)$, by Lemma 1 below, this expression is bounded by

$$P\left(|\gamma_{xh} - \gamma_x| > \frac{\varepsilon}{2M_1(\gamma_x - \eta_1)}\right) + P\left(|\gamma_{xh} - \gamma_x| > \eta_1\right) + P\left(\sup_{0\le t\le T} |H_{xh}(t) - H_x(t)| > \eta_2\right)$$
$$+ P\left(\sup_{0\le t\le T} |H_{xh}(t) - H_x(t)| > \frac{\varepsilon}{2M_2(\bar{H}_x(T) - \eta_2)}\right).$$

Choosing η_1 and η_2 such that $\eta_1 = \frac{\varepsilon}{2M_1(\gamma_x - \eta_1)}$ and $\eta_2 = \frac{\varepsilon}{2M_2(\bar{H}_x(T) - \eta_2)}$, we have that

$$P\left(\sup_{0\le t\le T} |F_{xh}(t) - F_x(t)| > \varepsilon\right) \le 2P(|\gamma_{xh} - \gamma_x| > \eta_1) + 2P\left(\sup_{0\le t\le T} |H_{xh}(t) - H_x(t)| > \eta_2\right). \qquad (7)$$

For the second probability on the right hand side of (7), we have the bound

$$P\left(\sup_{0\le t\le T} |H_{xh}(t) - H_x(t)| > \eta_2\right) \le \frac{1}{2} d_0 \eta_2 n h_n e^{-d_1 n h_n \eta_2^2/4}$$

where d_0 and d_1 are constants. This result is given by Van Keilegom and Veraverbeke (1997) under the condition that
$\eta_2 \ge \max\left(\sqrt{6}\|K\|_2 (nh_n)^{-1/2}, 2\|\dot{H}_x\|_\infty \bar{\Delta}_n + 2\mu_2^K \|\ddot{H}\|_\infty h_n^2\right)$.

From Veraverbeke and Cadarso Suárez (2000), we find a similar result for the first probability of (7),

$$P(|\gamma_{xh} - \gamma_x| > \eta_1) \le 2e^{-C_1 \eta_1^2 n h_n}$$

with C_1 an absolute constant and under the condition that $\eta_1 \ge 2(\|\dot{\gamma}_x\|_\infty \bar{\Delta}_n + \|\ddot{\gamma}_x\|_\infty \mu_2^K h_n^2)$.

Combining these results gives the exponential bound for $F_{xh}(t)$.

(b) From Lemma 2 below, we find, for small $\varepsilon > 0$, a new upper bound

$$P\left(\sup_{0\le t\le T} |F_{xh}(t) - F_x(t)| > \varepsilon\right) \le 4e^{-\frac{C_1 \varepsilon^2 n h_n}{4(M_1(\gamma_x)+1)^2}} + d_0 \frac{\varepsilon}{2} n h_n e^{-\frac{d_1 n h_n}{4} \frac{\varepsilon^2}{4(M_2(\bar{H}_x(T))+1)^2}}$$

If we take $\varepsilon_n = C_2 (nh_n)^{-1/2} (\log n)^{1/2}$, we note that ε_n is small for large n and by the Borel-Cantelli Lemma we find the result.

Lemma 1. We have, for $z \in]0,1[$, that

$$M_1(z) = \max_{0\le y\le 1} \frac{-\varphi_x(y)}{\varphi_x'(\varphi_x^{-1}(z\varphi_x(y)))} \quad \text{and} \quad M_2(z) = \max_{0\le y\le 1} \frac{y\varphi_x'(z)}{\varphi_x'(\varphi_x^{-1}(y\varphi_x(z)))}$$

are non-increasing functions of z.

Proof: Due to the assumptions on $\varphi_x(t)$ and $\varphi_x'(t)$, we see that, for every $y, z \in]0,1[$

$$\frac{-\varphi_x(y)}{\varphi_x'(\varphi_x^{-1}(z\varphi_x(y)))} \quad \text{and} \quad \frac{y\varphi_x'(z)}{\varphi_x'(\varphi_x^{-1}(y\varphi_x(z)))}$$

are well-defined. Furthermore we can show that, for every $z \in]0,1[$, $0 \le \lim_{y\to 0} \frac{-\varphi_x(y)}{\varphi_x'(\varphi_x^{-1}(z\varphi_x(y)))} \le 1$, $\lim_{y\to 1} \frac{-\varphi_x(y)}{\varphi_x'(\varphi_x^{-1}(z\varphi_x(y)))} = 0$, $0 \le \lim_{y\to 0} \frac{y\varphi_x'(z)}{\varphi_x'(\varphi_x^{-1}(y\varphi_x(z)))} < +\infty$, $\lim_{y\to 1} \frac{y\varphi_x'(z)}{\varphi_x'(\varphi_x^{-1}(y\varphi_x(z)))} = 1$ such that $M_1(z)$ and

$M_2(z)$ are well-defined. To prove that these functions are non-increasing, we fix for the moment a value of $y \in [0,1]$. For $z_1, z_2 \in]0,1[$ with $z_1 < z_2$, we find that

$$\frac{-\varphi_x(y)}{\varphi_x'(\varphi_x^{-1}(z_1\varphi_x(y)))} \geq \frac{-\varphi_x(y)}{\varphi_x'(\varphi_x^{-1}(z_2\varphi_x(y)))} \quad \text{and} \quad \frac{y\varphi_x'(z_1)}{\varphi_x'(\varphi_x^{-1}(y\varphi_x(z_1)))} \geq \frac{y\varphi_x'(z_2)}{\varphi_x'(\varphi_x^{-1}(y\varphi_x(z_2)))}$$

if we rewrite the second inequality as a constraint optimization problem. Hence $M_1(z_2) \leq M_1(z_1)$ and $M_2(z_2) \leq M_2(z_1)$.

Lemma 2. If ε is sufficiently small, we find that

$$\begin{aligned} \frac{\varepsilon}{2(M_1(\gamma_x)+1)} \leq &\quad \min(\eta_1, \gamma_x) \quad &\leq \min(\eta_1^*, \gamma_x) \\ \frac{\varepsilon}{2(M_2(\bar{H}_x(T))+1)} \leq &\quad \min(\eta_2, \bar{H}_x(T)) \quad &\leq \min(\eta_2^*, \bar{H}_x(T)) \end{aligned}$$

where $\eta_1,\ \eta_1^*,\ \eta_2,\ \eta_2^* > 0$ is, respectively, a root of $\eta_1 = \frac{\varepsilon}{2M_1(\gamma_x-\eta_1)}$, $M_1(\gamma_x - \eta_1^*) = M_1(\gamma_x) + 1$, $\eta_2 = \frac{\varepsilon}{2M_2(\bar{H}_x(T)-\eta_2)}$ and $\quad M_2(\bar{H}_x(T) - \eta_2^*) = M_2(\bar{H}_x(T)) + 1$.

Proof: We only prove the results for η_1. For η_2 we work similarly.

From Lemma 1, we know that $M_1(\gamma_x - \eta_1)$ is a non-decreasing function of η_1 on the interval $[0, \gamma_x]$. Therefore there either exists a value η_1^* in this interval such that $M_1(\gamma_x - \eta_1^*) = M_1(\gamma_x) + 1$ or we have that $M_1(\gamma_x - \eta_1) < M_1(\gamma_x) + 1,\ \forall\ \eta_1 \in [0, \gamma_x]$. Hence $M_1(\gamma_x - \eta_1) \leq M_1(\gamma_x) + 1,\ \forall\ \eta_1 \leq \min(\eta_1^*, \gamma_x)$.

Define $\varepsilon^* := 2(M_1(\gamma_x) + 1)\min(\eta_1^*, \gamma_x)$. We note that for $\varepsilon < \varepsilon^*$,

$$\frac{\varepsilon}{2(M_1(\gamma_x)+1)} < \frac{\varepsilon^*}{2(M_1(\gamma_x)+1)} = \min(\eta_1^*, \gamma_x).$$

Furthermore we see that if η_1^* exists, $\eta_1^* M_1(\gamma_x - \eta_1^*) = (M_1(\gamma_x) + 1)\eta_1^* = \frac{\varepsilon^*}{2}$ else $\gamma_x M_1(0) < (M_1(\gamma_x) + 1)\gamma_x = \frac{\varepsilon^*}{2}$. Since the function $\eta_1 M_1(\gamma_x - \eta_1)$ is an increasing function of η_1 on the interval $[0, \gamma_x]$, this means that for $\varepsilon < \varepsilon^*$, the root η_1 of

$$\eta_1 = \frac{\varepsilon}{2M_1(\gamma_x - \eta_1)} \Leftrightarrow \eta_1 M_1(\gamma_x - \eta_1) = \frac{\varepsilon}{2}$$

is smaller than η_1^*. If $\eta_1 M_1(\gamma_x - \eta_1) < \frac{\varepsilon}{2},\ \forall\ \eta_1 \in [0, \gamma_x]$, we take $\eta_1 = \gamma_x$. So $\min(\eta_1, \gamma_x) \leq \min(\eta_1^*, \gamma_x)$.

If η_1 exists, we have that

$$\frac{\varepsilon}{2(M_1(\gamma_x)+1)} M_1\left(\gamma_x - \frac{\varepsilon}{2(M_1(\gamma_x)+1)}\right) \leq \frac{\varepsilon(M_1(\gamma_x)+1)}{2(M_1(\gamma_x)+1)} = \frac{\varepsilon}{2} = \eta_1 M_1(\gamma_x - \eta_1),$$

such that $\frac{\varepsilon}{2(M_1(\gamma_x)+1)} \leq \eta_1$. If η_1 does not exists, we see that

$$\frac{\varepsilon}{2(M_1(\gamma_x)+1)} \leq \frac{\varepsilon^*}{2(M_1(\gamma_x)+1)} = \gamma_x.$$

Hence

$$\frac{\varepsilon}{2(M_1(\gamma_x)+1)} \leq \min(\eta_1, \gamma_x).$$

Proof of Theorem 2: To find an asymptotic representation of $F_{xh}(t)$, we use a second order Taylor expansion.

$$\begin{aligned} F_{xh}(t) - F_x(t) &= \varphi_x^{-1}\left(\gamma_x\varphi_x(\bar{H}_x(t))\right) - \varphi_x^{-1}\left(\gamma_{xh}\varphi_x(\bar{H}_{xh}(t))\right) \\ &= -\frac{1}{\varphi_x'(\bar{F}_x(t))}\left\{\gamma_{xh}\varphi_x(\bar{H}_{xh}(t)) - \gamma_x\varphi_x(\bar{H}_x(t))\right\} + \frac{\varphi_x''(\varphi_x^{-1}(\varepsilon(t)))}{2\varphi_x'(\varphi_x^{-1}(\varepsilon(t)))^3}\left\{\gamma_{xh}\varphi_x(\bar{H}_{xh}(t)) - \gamma_x\varphi_x(\bar{H}_x(t))\right\}^2 \end{aligned}$$

with $\varepsilon(t)$ between $\gamma_{xh}\varphi_x(\bar{H}_{xh}(t))$ and $\gamma_x\varphi_x(\bar{H}_x(t))$.

After adding and subtracting some terms, and applying a second order Taylor expansion to the second term, we get

$$\begin{aligned}
&= -\frac{1}{\varphi'_x(\bar{F}_x(t))}\left\{\varphi_x(\bar{H}_x(t))(\gamma_{xh}-\gamma_x)+\gamma_x(\varphi_x(\bar{H}_{xh}(t))-\varphi_x(\bar{H}_x(t)))\right\}\\
&- \frac{(\varphi_x(\bar{H}_{xh}(t))-\varphi_x(\bar{H}_x(t)))(\gamma_{xh}-\gamma_x)}{\varphi'_x(\bar{F}_x(t))}+\frac{\varphi''_x(\varphi_x^{-1}(\varepsilon(t)))}{2\varphi'_x(\varphi_x^{-1}(\varepsilon(t)))^3}\left\{\gamma_{xh}\varphi_x(\bar{H}_{xh}(t))-\gamma_x\varphi_x(\bar{H}_x(t))\right\}^2\\
&= -\frac{\varphi_x(\bar{H}_x(t))}{\varphi'_x(\bar{F}_x(t))}(\gamma_{xh}-\gamma_x)+\frac{\gamma_x\varphi'_x(\bar{H}_x(t))}{\varphi'_x(\bar{F}_x(t))}(H_{xh}(t)-H_x(t))+R_{n1}(t)+R_{n2}(t)+R_{n3}(t)\\
&= \sum_{i=1}^{n} w_{ni}(x,h_n)g_{tx}(Z_i,\delta_i)+R_{n1}(t)+R_{n2}(t)+R_{n3}(t)
\end{aligned}$$

with $R_{n1}(t)=-\frac{\gamma_x\varphi''_x(\eta(t))}{2\varphi'_x(\bar{F}_x(t))}(H_{xh}(t)-H_x(t))^2$, $R_{n2}(t)=-\frac{(\varphi_x(\bar{H}_{xh}(t))-\varphi_x(\bar{H}_x(t)))(\gamma_{xh}-\gamma_x)}{\varphi'_x(\bar{F}_x(t))}$, $R_{n3}(t)=\frac{\varphi''_x(\varphi_x^{-1}(\varepsilon(t)))}{2\varphi'_x(\varphi_x^{-1}(\varepsilon(t)))^3}\left\{\gamma_{xh}\varphi_x(\bar{H}_{xh}(t))-\gamma_x\varphi_x(\bar{H}_x(t))\right\}^2$ where $\eta(t)$ lies between $\bar{H}_{xh}(t)$ and $\bar{H}_x(t)$. In the remaining part of this proof we show the rate of convergence for each of the remainder terms. For $R_{n1}(t)$, we find

$$\sup_{0<t\leq T}|R_{n1}(t)|\leq \varphi''_x(\eta(T))\sup_{0<t\leq T}(H_{xh}(t)-H_x(t))^2.$$

While for $R_{n2}(t)$, we use a first order Taylor expansion and have

$$\sup_{0<t\leq T}|R_{n2}(t)|\leq \varphi'_x(\nu(T))\sup_{0<t\leq T}|H_{xh}(t)-H_x(t)|.|\gamma_{xh}-\gamma_x|$$

where $\nu(t)$ lies between $\bar{H}_{xh}(t)$ and $\bar{H}_x(t)$. After adding and subtracting some terms, and also using a first order Taylor expansion, we find for $R_{n3}(t)$,

$$\begin{aligned}
R_{n3}(t) &= \frac{\varphi''_x(\varphi_x^{-1}(\varepsilon(t)))}{2\varphi'_x(\varphi_x^{-1}(\varepsilon(t)))^3}\left\{\varphi_x(\bar{H}_{xh}(t))^2(\gamma_{xh}-\gamma_x)^2+\gamma_x^2\varphi'_x(\xi(t))^2(H_{xh}(t)-H_x(t))^2\right.\\
&\left.-2\gamma_x\varphi_x(\bar{H}_{xh}(t))\varphi'_x(\xi(t))(\gamma_{xh}-\gamma_x)(H_{xh}(t)-H_x(t))\right\}
\end{aligned}$$

with $\xi(t)$ between $\bar{H}_{xh}(t)$ and $\bar{H}_x(t)$. Since $H_x(T)<1$ and $H_{xh}(T)\to H_x(T)$ a.s. (Lemma A2, Van Keilegom and Veraverbeke (1997)), we may suppose that $T<T_{H_{xh}}$. Furthermore we have that $\sup_{0<t\leq T}|H_{xh}(t)-H_x(t)|\to 0$ a.s. and $\gamma_{xh}\to\gamma_x$ a.s. (Lemma A4, Van Keilegom and Veraverbeke (1997), Lemma A1, Braekers and Veraverbeke (2005)). Hence, we have that $\sup_{0<t\leq T}|R_{n1}(t)|$, $\sup_{0<t\leq T}|R_{n2}(t)|$ and $\sup_{0<t\leq T}|R_{n3}(t)|$ are all $O((nh_n)^{-1}\log n)$ a.s.

Proof of Theorem 3: To prove asymptotic normality, we first calculate the bias and variance expressions. Due to Theorem 2 we only need to consider the main term in the asymptotic representation. After a straightforward calculation (see e.g. Aerts, Janssen and Veraverbeke (1994)), we find

$$\begin{aligned}
&\sum_{i=1}^{n} w_{ni}(x,h_n)E[g_{tx}(Z_i,\delta_i)]=\sum_{i=1}^{n} w_{ni}(x,h_n)\left\{-\frac{\varphi_x(\bar{H}_x(t))}{\varphi'_x(\bar{F}_x(t))}(\gamma_{x_i}-\gamma_x)+\frac{\gamma_x\varphi'_x(\bar{H}_x(t))}{\varphi'_x(\bar{F}_x(t))}(H_{x_i}(t)-H_x(t))\right\}\\
&= \frac{1}{2}\mu_2^K h_n^2\left\{-\frac{\varphi_x(\bar{H}_x(t))}{\varphi'_x(\bar{F}_x(t))}\ddot{\gamma}_x+\frac{\gamma_x\varphi'_x(\bar{H}_x(t))}{\varphi'_x(\bar{F}_x(t))}\ddot{H}_x(t)\right\}+o(h_n^2)+O(n^{-1})
\end{aligned}$$

$$\sum_{i=1}^{n} w_{ni}^2(x,h_n)\mathrm{Var}(g_{tx}(Z_i,\delta_i))$$

$$= \sum_{i=1}^{n} w_{ni}^2(x,h_n)\left\{\frac{\varphi_x(\bar{H}_x(t))^2}{\varphi_x'(\bar{F}_x(t))^2}\gamma_{x_i}(1-\gamma_{x_i}) + \left(\frac{\gamma_x\varphi_x'(\bar{H}_x(t))}{\varphi_x'(\bar{F}_x(t))}\right)^2 H_{x_i}(t)(1-H_{x_i}(t))\right\}$$

$$= \frac{\|K\|_2^2}{nh_n}\left\{\frac{\varphi_x(\bar{H}_x(t))^2}{\varphi_x'(\bar{F}_x(t))^2}\gamma_x(1-\gamma_x) + \left(\frac{\gamma_x\varphi_x'(\bar{H}_x(t))}{\varphi_x'(\bar{F}_x(t))}\right)^2 H_x(t)(1-H_x(t))\right\} + o((nh_n)^{-1})$$

The asymptotic normality result for the estimator can now be obtained by checking Liapunov's condition for $(nh_n)^{1/2}\sum_{i=1}^{n} w_{ni}(x,h_n)\left(g_{tx}(Z_i,\delta_i) - E[g_{tx}(Z_i,\delta_i)]\right)$. The (b)-part of Theorem 3 deals with the optimal bandwidth $h_n = Cn^{-1/5}$ for some $C > 0$, which minimizes the mean squared error.

Acknowledgement

Financial support from the IAP research network nr P5/24 of the Belgian Government (Belgian Science Policy) is gratefully acknowledged.

References

[1] M. Aerts, P. Janssen and N. Veraverbeke, (1994) Bootstrapping regression quantiles, *J. Nonpar. Statist.*, **4**, (1994), 1-20.

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

[3] S. Csörgő, (1988). Testing for the proportional hazards model of random censorship,*Proceedings of the Fourth Prague Symposium on Asymptotic Statistics* (P. Mandl and M. Hušková, eds.), 41-53, Charles University Press, Prague.

[4] E. L. Kaplan and P. Meier, (1958) Nonparametric estimation from incomplete observations, *J. Amer. Statist. Assoc.*, **53**, 457-481.

[5] J. A. Koziol and S. B. Green, (1976). A Cramér-von Mises statistic for randomly censored data, *Biometrika*, **63**, 465-474.

[6] J. D. Neilson, K. G. Waiwood and S. J. Smith, (1989) Survival of Atlantic halibut (Hippoglossus hippoglossus) caught by longline and otter trawl gear, *Canadian Journal of Fisheries and Aquatic Sciences*, **46**, 887-897.

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

[8] C. J. Stone, (1977) Consistent nonparametric regression, *Ann. Statist.*, **5**, 595-645.

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

[10] 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.

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

[12] N. Veraverbeke and C. Cadarso Suárez, (2000) Estimation of the conditional distribution in a conditional Koziol-Green model, *Test*, **9**, 97-122.

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