An extension of the Koziol-Green model under dependent censoring
Peer-reviewed author version

An Extension of the Koziol-Green Model
under Dependent Censoring

Auguste Gaddah and Roel Braekers [1]

Interuniversity Institute for Biostatistics and statistical Bioinformatics, Universiteit Hasselt,
Agoralaan 1, 3590 Diepenbeek, Belgium

July 9, 2010

Abstract

In survival analysis, the classical Koziol-Green model under random censorship is commonly used for informative censoring. We propose in this paper an extension of this model in which we derive a non-parametric estimator for the distribution function of a survival time under two types of informative censoring. As first type of informative censoring, we assume that the censoring time depends on the survival time through the expression of their joint distribution by an Archimedean copula. For the second type of informative censoring, we assume that the marginal distribution of the censoring time is a function of the marginal distribution of the survival time where this function is found through a section of a known copula function on the observed lifetime and the censoring indicator. We prove in this paper the uniform consistency of the new estimator and show the weak convergence of the associated process. Afterwards we give some finite sample simulation results and illustrate this estimator on a real life data set.

Keywords: Almost sure representation; copula function; exponential bound; informative censoring; right censoring; simulation; strong consistency; weak convergence.

MSC: 62N02; 62G05; 62G20.

1   Introduction

Suppose $Y_1, Y_2, ..., Y_n$ are independent and identically distributed copies of a non-negative random variable Y with continuous distribution function F. As often occurs in clinical trials or industrial life tests, these random variables are subject to right censoring. That is, for each Y_i $(i = 1, 2, 3, ..., n)$, there exists an independent and identically distributed random variable C_i of a censoring variable C with distribution function G such that we only observe the pair (Z_i, δ_i), where

$$Z_i = \min(Y_i, C_i) \quad \text{and} \quad \delta_i = \mathbf{1}\{Y_i \leq C_i\}$$

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

For each i, Z_i and δ_i are respectively copies of the observed lifetime $Z = \min(Y, C) \sim H$ and the censoring indicator $\delta = \mathbf{1}\{Y \leq C\}$.

To estimate the distribution function F from the observed data, we need to make a non-verifiable assumption about the dependency between Y and C (Tsiatsis 1975). In survival analysis, it is common to assume independence between these random variables. Under this assumption, Kaplan and Meier (1958) developed the product limit estimator as standard estimator for F. However we note that in some situations, this independence assumption is not satisfied. For example, in a cancer study where the time of interest is the time until recurrence of a cancer tumor and the censoring time is time until death, or in industrial testing, it may occur that a piece of equipment is taken away (i.e. censored) because it shows signs of future failure. Zheng and Klein (1995) proposed an extension of the product limit estimator under dependent censoring by using a copula function to describe the dependence structure between Y and C. This idea was later investigated by Rivest and Wells (2001) in the class of Archimedean copulas. They derived a closest form expression for this copula-graphic estimator.

Another type of informative censoring which is often used in survival analysis is the Koziol-Green sub-model (Koziol and Green (1976)). In this sub-model of the general random censorship model, the time until an event Y is assumed independent of the censoring time C but also the marginal survival function of the censoring time is assumed to be a power of the survival function of the survival time. An important consequence of these assumptions is that the assumption on the marginal survival functions is equivalent with the result that the observable variables Z and δ are independent. This sub-model was further studied by many authors. For example, Abdushukurov (1987) and Cheng and Lin (1987) independently pointed out that the survival distribution function estimator in this sub-model outperforms the general Kaplan-meier product limit estimator in terms of asymptotic efficiency. Csörgő (1988) developed a test to check the validity of this model based on the independence of the observable variables.

In the current paper we extend the Koziol-Green sub-model to allow on the one hand for dependency between Y and C, and on the other hand for a more general relationship between the marginal distribution functions of the time until an event and the censoring time. We generalize the classical Koziol-Green sub-model in two different ways at the same time. Hereby we wish to provide a solution for a major critic of this classical model that it would be too restrictive to be often true in practice. Furthermore, since this extended model will lead to a sub-model for the Rivest and Wells copula-graphic estimator in which also the marginal distribution of the censoring time contributes information for the estimation of the distribution F of the time until an event, the estimator for F in this extended model will also outperform the Rivest and Wells copula-graphic estimator in terms of asymptotic efficiency. To define the extended Koziol-Green model under dependent censoring we first assume, as in Rivest and Wells (2001), that the joint survival function of the time until an event and the censoring time is given by

$$S(t_1, t_2) = P\left(Y > t_1, C > t_2\right) = \varphi^{[-1]}\left(\varphi(\bar{F}(t_1)) + \varphi\left(\bar{G}(t_2)\right)\right) \tag{1}$$

where $\bar{F}(t) = 1 - F(t)$ and $\bar{G}(t) = 1 - G(t)$ are survival distribution functions of Y and C respectively. The function $\varphi : [0, 1] \to [0, \infty]$ is a generator of a known Archimedean copula and is a continuous, convex and strictly decreasing function with $\varphi(1) = 0$. We denote by $\varphi^{[-1]}$ the pseudo-inverse of φ which is defined as in Nelsen (2006),

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

Since a copula function only determines the association structure between the continuous variables Y and

C, we can model the marginal distribution functions separately. In this paper, we assume that the marginal distributions of Y and C are linked through

$$\bar{G}(t) = \mu(\bar{F}(t)), \quad t \geq 0 \tag{2}$$

where $\mu(w)$ is a non-decreasing function on $[0,1]$ with $\mu(0) = 0$ and $\mu(1) = 1$. We define this function $\mu(w)$ indirectly by assuming some known copula function $\mathcal{C}$ which satisfies

$$H^u(t) = P\left(Z \leq t, \delta = 1\right) = \mathcal{C}\left(\gamma, H(t)\right) \tag{3}$$

where $\gamma = P(\delta = 1)$ is the proportion of uncensored observations and $H(t) = P(Z \leq t)$ is the distribution function of the observed lifetime. We note in (3) that we do not need the full copula function to model the sub-distribution $H^u(t)$ of the uncensored observations, but only a section of it at γ. Since the censoring indicator δ is a discrete variable, we know from Sklar's theorem (Nelsen (2006)) that the copula function $\mathcal{C}$ is not unique. In Klement *et al.* (2007), the set of all copulas with the same vertical γ - section is studied. Furthermore from Genest and Nešlehová (2007), it is clear that this copula function alone is not sufficient to describe the association structure between Z and δ. The marginal distributions are also needed for this. However, the non-uniqueness of the copula function $\mathcal{C}$ does not have any consequence on the estimator of the distribution function F. In the following section, we will see that this estimator does not change when a copula function with the same vertical γ-section is chosen.

The remaining part of this paper is organized as follows. In the next section, we derive an estimator for the distribution function F in the extended Koziol-Green model under dependent censoring. Hereby we will show in some examples that the Archimedean copula in (1) and the copula function $\mathcal{C}$ in (3) can be different copulas function in practical settings. After specifying some definitions and assumptions in Section 3, we give an exponential bound and an almost sure representation for this estimator. Furthermore, we show in the same section the strong consistency and weak convergence of the process associated with this estimator. In Section 4, we investigate the finite sample performance of the survival distribution function estimator in the extended Koziol-Green model by means of a simulation study. Finally, we apply this estimator to a real data set in Section 5. In the Appendix, we give the proof of the theorems.

2   The Extended Koziol-Green Model

In this section, we develop an estimator for the distribution function F in the extended Koziol-Green model under dependent censoring. We distinguish in this model two different informative censoring structures. On the one hand, we assume that the censoring time C depends on the survival time Y through their joint survival function which is given by (1). On the other hand, for the second informative censoring structure, it is assumed that the marginal distributions of the censoring time and the survival time are linked as in (2) where the function $\mu(w)$ is indirectly defined by the relationship (3) between the observable random variables Z and δ.

To find an estimator for F in this model, we obtain from (1), as in Tsiatis (1975), that

$$\frac{dH^u(t)}{dt} = -\frac{\partial}{\partial t_1} S(t_1, t_2)\bigg|_{t=t_1=t_2} = \frac{\varphi'(\bar{F}(t))}{\varphi'\left(S(t,t)\right)} \frac{dF(t)}{dt} = \frac{\varphi'(\bar{F}(t))}{\varphi'\left(\bar{H}(t)\right)} \frac{dF(t)}{dt}$$

with $\varphi'(u) = \frac{d}{du}\varphi(u)$ and $S(t,t) = \varphi^{-1}\left(\varphi(\bar{F}(t)) + \varphi(\bar{G}(t))\right) = 1 - H(t) = \bar{H}(t)$.

3

Reorganizing this equation, gives

$$\varphi'\left(\bar{F}(t)\right)\frac{dF(t)}{dt} = \varphi'\left(\bar{H}(t)\right)\frac{dH^u(t)}{dt}$$

By integrating on both sides and with $\varphi(\bar{F}(0)) = \varphi(1) = 0$, we obtain that

$$\bar{F}(t) = \varphi^{-1}\left(-\int_0^t \varphi'(\bar{H}(s))dH^u(s)\right). \tag{4}$$

This result is also given in Rivest and Wells (2001) and is valid in any model where the underlying association structure between Y and C is assumed to follow an Archimedean copula as in (1).

For the second informative censoring structure in this extended Koziol-Green model, we find from (2) that

$$dH^u(s) = \mathcal{C}_2\left(\gamma, H(s)\right)dH(s)$$

where $\mathcal{C}_2(u,v) = \frac{\partial}{\partial v}\mathcal{C}(u,v)$ is the partial derivative of the general copula function $\mathcal{C}(u,v)$ with respect to the second coordinate.

Introducing this result into (4), we obtain after a variable transformation, that

$$\bar{F}(t) = \varphi^{-1}\left(-\int_0^{H(t)} \varphi'(1-w)\mathcal{C}_2\left(\gamma, w\right)dw\right) \tag{5}$$

We now find an estimator for the distribution function $F(t)$ in the extended Koziol-Green model under dependent censoring by replacing γ and $H(t)$ in (5) by their empirical counterparts which are defined as

$$H_n(t) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{Z_i \leq t\} \quad \text{and} \quad \gamma_n = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{\delta_i = 1\}.$$

Hence, we find an estimator for the distribution function $F(t)$ of the survival time by

$$\bar{F}_n(t) = \varphi^{-1}\left(-\int_0^{H_n(t)} \varphi'(1-w)\mathcal{C}_2\left(\gamma_n, w\right)dw\right). \tag{6}$$

For brevity, we shall henceforth refer to this estimator as the EKG estimator.

Some remarks:

- If we take $\mathcal{C}$ such that Z and δ are independent $(H^u(t) = \gamma H(t))$, we see that the estimator in (6) simplifies to

$$\bar{F}_n(t) = \varphi^{-1}\left(\gamma_n\varphi(\bar{H}_n(t))\right). \tag{7}$$

This is an unconditional version of an estimator which was studied by Braekers and Veraverbeke (2008) in a fixed design regression setting.

Moreover, if we also assume that the censoring time and the survival time are independent, this estimator reduces to the classical Koziol-Green estimator,

$$\bar{F}_n(t) = (1 - H_n(t))^{\gamma_n}.$$

This estimator was also studied by Abdushuskurov (1987) and Cheng and Lin (1987). It is known as the ACL estimator.

- We note that we do not need a closed form expression for the function $\mu(w)$ in (2) to find an expression for the estimator of the distribution function F. It suffices to know the γ-section of the copula $\mathcal{C}$ in (3). Numerically, it is possible to find the function $\mu(w)$ as the solution of the set of equations, $\forall w \in [0, 1]$,

$$
\begin{aligned}
w &= \varphi^{-1}\left(-\int_{m(w)}^{1} \varphi'(s)\mathcal{C}_2(\gamma, 1 - s)ds \right) \\
\mu(w) &= \varphi^{-1}(\varphi(m(w)) - \varphi(w))
\end{aligned}
$$

- Due to the dependent censoring, we can consider the EKG model as a competing risks problem in which an assumption is made on the association structure between some underlying latent variables. Hereby we assume a known Archimedean copula function to describe this association. On the other hand, we model a link function between the marginal distribution functions of the lifetime and the censoring time indirectly through condition (3) on the observed sub-distribution function $H^u(t)$ of the uncensored observations. We can rewrite this condition in terms of the sub-hazard function of these observations,

$$
\lambda^*(t) = \frac{dH^u(t)}{\bar{H}(t)} = \mathcal{C}_2(\gamma, H(t))\frac{dH(t)}{\bar{H}(t)} = \mathcal{C}_2(\gamma, H(t))\lambda(t)
$$

with $\lambda(t)$ the overall hazard function.

3 Asymptotic Results

In this section, we give three theorems which summarize some asymptotic results of the EKG estimator given in (6). We give an exponential bound and a strong consistency result in Theorem 1 while in Theorem 2 and 3, we present an almost sure representation of this estimator and show the weak convergence of the associated process. The proof of the different theorems is relegated to the Appendix. Before we state the theorems, we first give the following basic definitions and regularity conditions that are important in the development of the asymptotic results. For any distribution function K, we denote the right endpoint of its support by $T_K = \inf\{t : K(t) = 1\}$.

A1. For a copula function $\mathcal{C}$, we define $\mathcal{C}_1(u, v) = \frac{\partial}{\partial u}\mathcal{C}(u, v)$, $\mathcal{C}_2(u, v) = \frac{\partial}{\partial v}\mathcal{C}(u, v)$, $\mathcal{C}_{12}(u, v) = \mathcal{C}_{21}(u, v) = \frac{\partial^2}{\partial u \partial v}\mathcal{C}(u, v)$, $\mathcal{C}_{11}(u, v) = \frac{\partial^2}{\partial u^2}\mathcal{C}(u, v)$, $\mathcal{C}_{22}(u, v) = \frac{\partial^2}{\partial v^2}\mathcal{C}(u, v)$. Furthermore we assume that $\mathcal{C}_{11}(u, v)$, $\mathcal{C}_{22}(u, v)$ and $\mathcal{C}_{12}(u, v)$ exist and are continuous for all $(u, v) \in [0, 1]^2$.

A2. For the generator of an Archimedean copula φ, we define $\varphi'(u) = \frac{d}{du}\varphi(u)$, $\varphi''(u) = \frac{d^2}{du^2}\varphi(u)$, $\varphi'''(u) = \frac{d^3}{du^3}\varphi(u)$ and assume that $\varphi'''(u)$ exists and is continuous for all $u \in (0, 1]$.

Theorem 1. *Assume A1, A2, $\varphi'(1) < 0$ and $T < T_H$, then*
(a) For all $\varepsilon > 0$, we have

$$
P\left(\sup_{0 \leq t \leq T} |F_n(t) - F(t)| > \varepsilon \right) \leq 2\exp\left(-\frac{n\alpha^2}{6(3\gamma + \beta)} \right) + D\exp\left(-n\alpha^2 \right)
$$

where D is a finite positive constant,

$$
\alpha = \frac{\varphi'(1)\varepsilon}{2\varphi'\left(\varphi^{[-1]}\left(\varphi(1 - H(T)) - \frac{\varphi'(1)\varepsilon}{2} \right) \right)} \quad and \quad \beta = 1 - H(T) - \varphi^{[-1]}\left(\varphi(1 - H(T)) - \frac{\varphi'(1)\varepsilon}{2} \right)
$$

(b) As $n \to \infty$, then
$$\sup_{0 \le t \le T} |F_n(t) - F(t)| = O\left(n^{-1/2}(\log n)^{1/2}\right) \quad a.s.$$

Theorem 2. *Assume A1, A2, $\varphi'(1) < 0$ and $T < T_H$. Then, as $n \to \infty$,*
$$F_n(t) - F(t) = \frac{1}{n}\sum_{i=1}^{n} k(t; Z_i, \delta_i) + r_n(t)$$

where

$$k(t; Z_i, \delta_i) = \frac{1}{\varphi'(\bar{F}(t))}\left\{(\mathbf{1}\{\delta_i = 1\} - \gamma)\int_0^{H(t)}\varphi'(1 - w)\mathcal{C}_{12}(\gamma, w)\,dw \right. \tag{8}$$
$$\left. + (\mathbf{1}\{Z_i \le t\} - H(t))\,\varphi'(\bar{H}(t))\mathcal{C}_2(\gamma, H(t))\right\}$$

and

$$\sup_{0 \le t \le T}|r_n(t)| = O\left(n^{-1}\log n\right) \quad a.s.$$

Theorem 3. *Assume the conditions of Theorem 2. As $n \to \infty$, then*
$$\sqrt{n}\,(F_n(\cdot) - F(\cdot)) \to W(\cdot) \quad in \quad \ell^\infty[0, T]$$

where $W(\cdot)$ is a zero mean Gaussian process with covariance function

$$\Gamma(s, t) = \frac{1}{\varphi'(\bar{F}(s))\varphi'(\bar{F}(t))}\left\{\gamma(1 - \gamma)\int_0^{H(s)}\varphi'(1 - w)\mathcal{C}_{12}(\gamma, w)\,dw\int_0^{H(t)}\varphi'(1 - w)\mathcal{C}_{12}(\gamma, w)\,dw \right.$$
$$+ (H^u(s) - \gamma H(s))\,\varphi'(\bar{H}(s))\mathcal{C}_2(\gamma, H(s))\int_0^{H(t)}\varphi'(1 - w)\mathcal{C}_{12}(\gamma, w)\,dw$$
$$+ (H^u(t) - \gamma H(t))\,\varphi'(\bar{H}(t))\mathcal{C}_2(\gamma, H(t))\int_0^{H(s)}\varphi'(1 - w)\mathcal{C}_{12}(\gamma, w)\,dw$$
$$\left. + (H(s \wedge t) - H(s)H(t))\,\varphi'(\bar{H}(s))\varphi'(\bar{H}(t))\mathcal{C}_2(\gamma, H(s))\,\mathcal{C}_2(\gamma, H(t))\right\}$$

4 A Simulation Study

In this section, we investigate the finite sample performance of the proposed nonparametric EKG estimator via a simulation study. Hereto we assume that the survival times $Y_i \sim \text{Exp}(\lambda)$ $(i = 1, 2, 3, \ldots, n)$ are independent random variables with $\lambda = 0.5$. Since there are two different types of informative censoring in this model, we take on the one hand that the joint survival function of the survival time and the censoring time is defined by the Archimedean Clayton copula with generator function $\varphi(t) = \frac{1}{2}\left(t^{-2} - 1\right)$. For the second type of informative censoring, we assume that the marginal distribution of the censoring time is a function of the distribution function of the survival time. We consider a section of the Plackett copula in (3), given by
$$\mathcal{C}(u, v) = \frac{1 + (\theta - 1)(u + v) - \sqrt{[1 + (\theta - 1)(u + v)]^2 - 4uv\theta(\theta - 1)}}{2(\theta - 1)} \tag{9}$$

	EKG			RW			BV		
n	$\gamma = 0.25$	$\gamma = 0.50$	$\gamma = 0.75$	$\gamma = 0.25$	$\gamma = 0.50$	$\gamma = 0.75$	$\gamma = 0.25$	$\gamma = 0.50$	$\gamma = 0.75$
20	0.2811	0.0021	0.0034	0.1758	0.0001	0.0010	0.3917	0.0051	0.0584
50	0.1540	0.0008	0.0006	0.1425	0.0007	0.0004	0.3911	0.0075	0.0580
100	0.0727	0.0011	0.0003	0.1166	0.0006	0.0003	0.3909	0.0089	0.0405
150	0.0463	0.0003	0.0007	0.0955	0.0002	0.0003	0.3911	0.0087	0.0595
200	0.0292	0.0003	0.0003	0.0897	0.0007	0.0005	0.3904	0.0078	0.0586
250	0.0226	0.0002	0.0004	0.0770	0.0001	0.0004	0.3906	0.0087	0.0413

Table 1: Absolute biases for the Extended Koziol-Green estimator (EKG), Rivest and Wells (2001) estimator (RW) and Braekers and Veraverbeke (2008) estimator (BV)

	EKG			RW		
n	$\gamma = 0.25$	$\gamma = 0.50$	$\gamma = 0.75$	$\gamma = 0.25$	$\gamma = 0.50$	$\gamma = 0.75$
20	0.4221	0.2524	0.2499	0.3146	0.2524	0.2495
50	0.3430	0.2505	0.2495	0.2988	0.2504	0.2494
100	0.2896	0.2499	0.2494	0.2913	0.2498	0.2494
150	0.2713	0.2497	0.2493	0.2862	0.2496	0.2493
200	0.2615	0.2495	0.2493	0.2833	0.2495	0.2492
250	0.2577	0.2495	0.2493	0.2809	0.2495	0.2493

Table 2: Variances under the Extended Koziol-Green estimator (EKG), Rivest and Wells (2001) estimator (RW).

with $\theta = 20$, to express the link-function between both marginal distributions.

The data generation process in the simulation study consists of the following steps:

1. We generate two independent uniform (0,1) random variables u and t.

2. We set $v = c_u^{-1}(t)$, where $c_u(v) = \frac{\partial}{\partial u} \left\{ \varphi^{[-1]} \left(\varphi(u) + \varphi(v) \right) \right\}$ and c_u^{-1} is the inverse of c_u.

3. We obtain h as a solution to the equation

$$1 - h = \varphi^{[-1]} \left(-\int_0^h \varphi'(1-w)\mathcal{C}_2\left(\gamma, w\right) dw + \varphi(u) \right)$$

   with γ being the desired proportion of uncensored observations.

4. We set $Y = -\frac{1}{\lambda}\log(v)$ and $C = -\frac{1}{\lambda}\log\left(\varphi^{-1}\left(\varphi(1-h) - \varphi(u)\right)\right)$

5. We set $Z = \min(Y, C)$ and $\delta = \mathbf{1}\{Y \leq C\}$

In the algorithm above, Step 2 generates a pair of uniform variables whose joint distribution is an Archimedean copula with generator function φ. By this step, the future couple (Y, C) will satisfy (1). Afterwards in Step 3 and 4, we alter the marginal distributions of the pair (Y, C) such that also (3) holds. We investigate in this simulation study the effect of the censoring intensity on the EKG estimator by considering different values of γ (i.e. $\gamma = 0.25, 0.50, 0.75$). For each value of γ and for different sample sizes n (i.e. $n = 20, 50, 100, 150, 200, 250$), we generate 10000 samples. For each sample size, we determine the bias in estimating $F(t)$ at a pre-specified point $t = 1.5$ as the average difference between $F_n(1.5)$ and $F(1.5)$.

In Table 1, we give for the different sample sizes and censoring intensities, the absolute biases for the EKG estimator. For the purpose of comparison, this table also includes the biases based on the estimators of Rivest and Wells (2001), and Braekers and Veraverbeke (2008) as given in (7). We observe that the results based on the EKG estimator and the estimator of Rivest and Wells (2001) are close. This was expected since the EKG model is a sub-model of the Rivest and Wells model. Their estimator only considers the dependent censoring between the lifetime Y and the censoring time C, but leaves the marginal distributions of both times unspecified. Therefore we have that data which is generated under the EKG model also satisfies the Rivest and Wells model and the biases for both estimators are almost the same. On the other hand, if we take a different sub-model of the Rivest and Wells model by considering a different copula function in (3) as in the estimator linked to Braekers and Veraverbeke (2008), we note in Table 1 that the bias of this estimator is for each sample size extremely larger than the bias on the EKG estimator. For an increasing sample size, these biases will probably diminish but a substantial bias will always remain due to the wrong copula function for $\mathcal{C}$ on Z and δ.

In Table 2 we compare the variances of the simulated estimates of the EKG estimator and the Rivest and Wells estimator. We note that the variances for the Rivest and Wells estimator and the EKG estimator are almost the same when there is a large probability of uncensored observations. For a small amount of uncensored observations and a large sample size, we have that the variance of the EKG estimator is smaller than for the Rivest and Wells estimator. We expect such a behavior since the EKG model is more efficient than the Rivest and Wells model for the generated data. Taking the extra information contained in the distribution of the censored observations into account improves the estimate of the survival time distribution when the amount of censoring is high in the data set. For a small sample size and a small probability of uncensored observations, we note in Table 2 that the variance of the Rivest and Wells estimator is smaller than for the EKG estimator. This result contradicts the asymptotic efficiency of the EKG model, however we are still investigating whether this difference is substantial or merely the result of our simulation.

To get further insight, we repeated the simulation process for other choices of the underlying Archimedean copula between the survival time Y and censoring time C in (1). However, we do not report these additional results because the conclusions are the same as above.

5   Real Data Application: Worcester heart attack study

In this section, we apply the estimator developed in Section 2 to a real data set. The data comes from the Worcester heart attack study (WHAS) which has information on more than 8000 admissions. The main objective of this study was to describe trends over time in the incident and survival rates following hospital admission of acute miocardia infarction (AMI) patients. However, we will only consider the 10% random sample of the original data set presented by Hosmer and Lemeshow (1999,p.24,25). Only a small subset of variables as well as patients with no missing values are included in this sub-sample. As a result, the data we utilize in the section has information on only 481 patients. Of these patients, 82 (17%) died while in admission (uncensored), whereas 399 (83%) were discharged (censored). Also, we will mainly be concerned about the time until death in hospital of such patients. Note that the results of this section are only for illustrative purpose, and not to be compared with the analysis of the complete data set. See Hosmer and Lemeshow (1999) for more details and pointers towards the findings from the complete WHAS data.

In this study, we observe that a patient with severe health conditions is likely to die within the first few
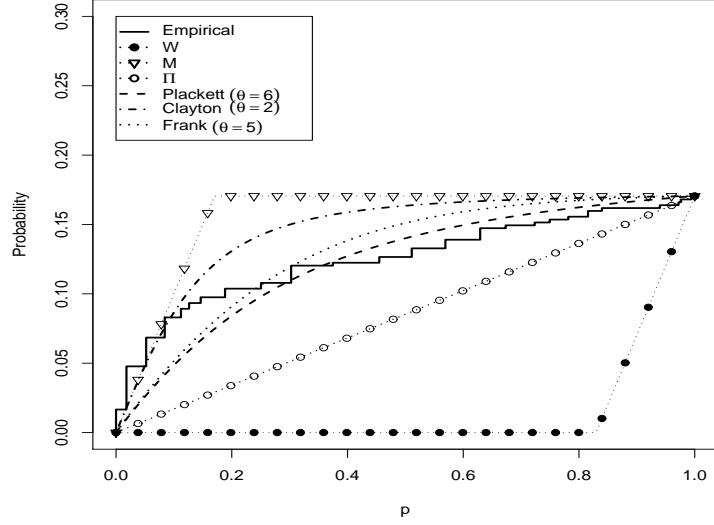
Figure 1: Graphical test of the copula function for the relationship between observed variables

days of admission. However, if such patient does not die, then he/she is most likely to spend many days in a hospital bed. Therefore, we believe that the time until death in the hospital Y_i increases with the time spent in the hospital C_i (i.e. the time spent in the hospital has a positive influence on the time until death in hospital). Moreover, we believe that the occurrence of these censoring is a manifestation of the patients medical condition. Therefore, we suspect that censoring time is additionally informative to the time until death in the hospital through its distribution function. As such, the EKG estimator could be the outstanding candidate to estimate the survival distribution of the time spent in hospital admission of such patients.

In this practical data illustration, we first assume that the underlying association structure of the time until death and the time until discharge is given by the Clayton copula with generator function $\varphi(v) = \frac{1}{0.9} \left(v^{-0.9} - 1 \right)$. Since this is an unverifiable assumption, we will afterwards perform a small sensitivity analysis in which we assume that the time until death and the time until discharge are independent. For the second type of informative censoring based the marginal distributions, we have to find a copula function $\mathcal{C}$ such that the condition (3) holds. Unlike the underlying association structure, we can verify this condition from the observed data. To find a suitable copula function we look at the empirical counterpart of condition (3),

$$H_n^u(t) = \mathcal{C}\left(\gamma_n, H_n(t)\right)$$

where

$$H_n^u(t) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{Z_i \leq t, \delta_i = 1\} \quad , \quad H_n(t) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{Z_i \leq t\} \quad , \quad \gamma_n = \sum_{i=1}^n \mathbf{1}\{\delta_i = 1\}$$

with Z_i and δ_i denoting respectively, the observed time and censoring indicator for each $i(i = 1, 2, 3, ..., n)$. This readily suggests an informal visual procedure to find a copula function $\mathcal{C}$. In Figure 1, we compare the empirical quantity $H_n^u\left(H_n^{-1}(p)\right)$ where $H_n^{-1}(p) = \inf\{t : H_n(t) > p\}$ is the quantile function of $H_n(t)$, with the γ_n-section of different known copula functions as Fréchet-Hoeffding lower bound (W), Fréchet-Hoeffding upper bound (M), Product (Π), Plackett, Clayton and Frank copulas. We note in this figure that the Plackett
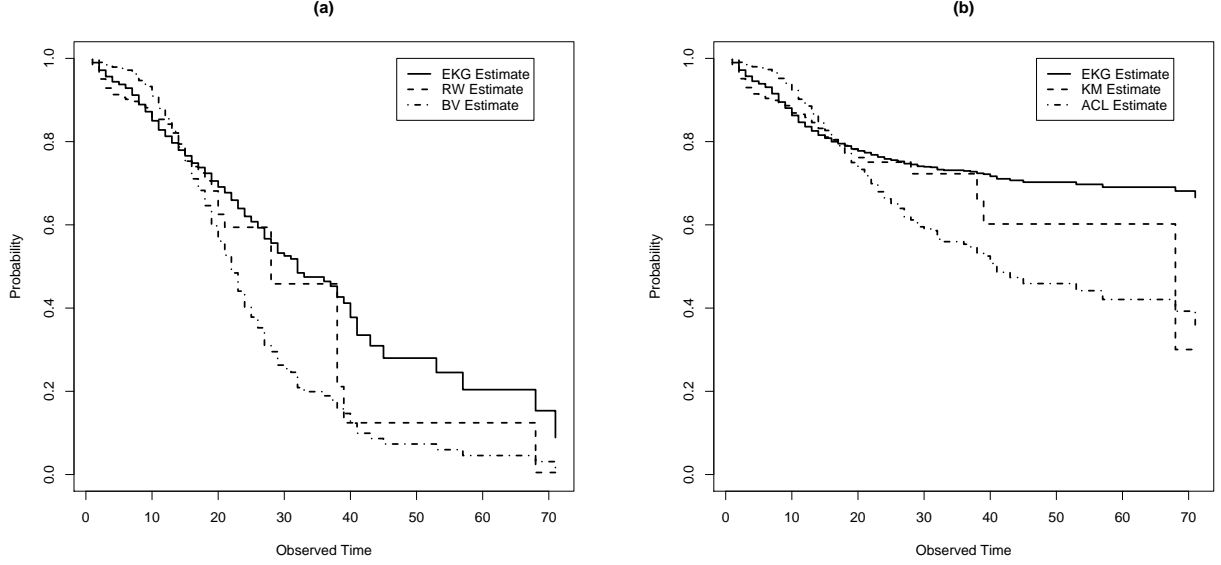
9

Figure 2: Survival distribution estimates for the acute miocardia infarction patients. In (a), a Clayton copula is assumed for the underlying association structure between death time and censoring time, while in (b) death time and censoring time are assumed independent.

copula with parameter $\theta = 6$ gives the best approximation to the empirical quantity and as such it is the most plausible candidate for this data set.

In Figure 2(a), we present the EKG estimate of the distribution function for the time until death. We assume in the figure that the underlying association between the time until death and the time until discharge is given by a Clayton copula as mentioned above. For the second type of informative censoring, we assume a Plackett copula with $\theta = 6$ based on the result in Figure 1. Furthermore we show in this figure also the Rivest and Wells (2001) copula graphic estimate (RW estimate), and Braekers and Veraverbeke (2008) estimate (BV estimate) of the survival distribution. This allows us to compare between the different models. Unlike for the BV estimate, we do not observe much difference between RW and EKG estimates, except for longer death times. This is expected since the RW estimate only uses the uncensored observations and these have mainly smaller values. Therefore, we know that the choice for the Plackett copula function in condition (3) is satisfied in this data set and is clearly different from the independence copula as in the BV estimate.

In Figure 2(b), we assume that the death time and censoring time are independent. Although we know that this assumption may be wrong for this data set, we show it as a form of sensitivity analysis. Under this assumption, the RW and BV estimates reduce to the Kaplan-Meier (KM) and the ACL estimates respectively. From the figure, we observe that the EKG and KM estimates are close to each other. Again, this suggest that the true underlying relationship between the observed time and censoring indicator is well approximated in this data set. Comparing both association structures in Figure 2, we observe that the probability of dying in the hospital is overestimated (bias) under the independence assumption. This last observation is a well known phenomenon in survival analysis. It has also been pointed out by Zheng and Klein (1995), and Rivest and Wells (2001), among others.

Acknowledgement

References

Abdushukurov, A. A. (1987), Nonparametric estimation in the proportional hazards model of random censorship, *Akad. Nauk Uz Tashkent. VINITI*, No. 3448-V (in Russian).

Braekers, R., and Veraverbeke, N. (2005) A copula-graphic estimator for the conditional survival function under dependent censoring, *Canadian Journal of Statistics*, 33, 429-447.

Braekers, R., and Veraverbeke, N. (2008), A conditional Koziol-Green model under dependent censoring, *Statistics and Probability Letters*, 78, 927-937.

Cheng, P.E., and Lin, G.D. (1987), Maximum likelihood estimation of a survival function under the Koziol-Green proportional hazards model, *Statistics and Probability Letters*, 5, 75-80.

Csörgó, S. (1988), Testing for the proportional hazards model of random censorship, *Proceedings of the 4th Prague Symposium on Asymptotic Statistics, Prague*.

Genest,C., and Nešlehová, J. (2007), A primer on copulas for count data, *Austin Bulletin*, 37, 475-515.

Hosmer, W.D., and Lemeshow, S. (1999), *Applied Survival Analysis: Regression Modeling of Time to Event Data*, Wiley, New York.

Kaplan, E.L., and Meier, P. (1958), Non-parametric estimation from incomplete observations, *Journal of American Statistical Association*, 53, 457-481.

Klement, P.E., Kolesárová, A., Mesiar, R., and Sempi, C. (2007), Copulas constructed from horizontal sections, *Communications in Statistics-Theory and Methods*, 36, 2901-2911.

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

Nelsen, R.B. (2006), *An Introduction to Copulas*, Springer-Verlag, New York.

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

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

Serfling, R.J. (1980), *Approximation Theorems of Mathematical Statistics*, Wiley, New York.

van der Vaart, A.W. (1998), *Asymptotic Statistics*, Cambridge University Press.

van der Vaart, A.W., and Welner, J.A. (2000), *Weak Convergence and Empirical Processes*, Springer, New. York.

Zheng, M., and Klein, J.P.(1995), Estimates of marginal survival for dependent competing risks based on an

assumed copula, *Biometrika*, 82, 127-138.

Appendix

Here, we give the proof of the Theorems presented in Section 3. First, we give the following two lemmas which are vital in establishing the proofs.

Lemma 1. *Suppose $0 \leq a < 1 - H(T)$, $b \geq 0$ and $a = \frac{b\varphi'(1)}{2\varphi'(1-H(T)-a)}$. Then for all $T < T_H$,*

$$\frac{b\varphi'(1)}{2\varphi'\left(\varphi^{[-1]}\left(\varphi(1-H(T)) - \frac{b}{2}\varphi'(1)\right)\right)} \leq a \leq 1 - H(T) - \varphi^{[-1]}\left(\varphi(1-H(T)) - \frac{b}{2}\varphi'(1)\right)$$

Proof. By the mean value theorem, we have

$$\varphi(1 - H(T) - a) - \varphi(1 - H(T)) = -\varphi'(1 - H(T) - a^*)a \tag{10}$$

where a^* is a point between zero and a.

Next, we note from the conditions of the lemma that

$$-\frac{b}{2}\varphi'(1) = -\varphi'(1 - H(T) - a)a \geq -\varphi'(1 - H(T) - a^*)a \tag{11}$$

Substituting (11) into (10), gives after some straight forward calculations that

$$a \leq 1 - H(T) - \varphi^{[-1]}\left(\varphi(1-H(T)) - \frac{b}{2}\varphi'(1)\right) \tag{12}$$

Using (12), we also get that

$$a = \frac{b\varphi'(1)}{2\varphi'(1-H(T)-a)} \geq \frac{b\varphi'(1)}{2\varphi'\left(\varphi^{[-1]}\left(\varphi(1-H(T)) - \frac{b}{2}\varphi'(1)\right)\right)}$$

which concludes the proof. $\square$

Proof of Theorem 1: We have that

$$
\begin{aligned}
F_n(t) - F(t) &= \left(1 - \bar{F}_n(t)\right) - \left(1 - \bar{F}(t)\right) = \bar{F}(t) - \bar{F}_n(t) \\
&= -\left\{\varphi^{-1}\left(-\int_0^{H_n(t)} \varphi'(1-w)\mathcal{C}_2(\gamma_n, w)dw\right) - \varphi^{-1}\left(-\int_0^{H(t)} \varphi'(1-w)\mathcal{C}_2(\gamma, w)dw\right)\right\}
\end{aligned}
$$

Using the mean value theorem, we get

$$F_n(t) - F(t) = (\gamma_n - \gamma)A(\gamma^*, H^*(t)) + (H_n(t) - H(t))B(\gamma^*, H^*(t))$$

where

$$A(\gamma^*, H^*(t)) = \frac{\int_0^{H^*(t)} \varphi'(1-w)\mathcal{C}_{12}(\gamma^*, w)dw}{\varphi'\left(\varphi^{-1}\left(-\int_0^{H^*(t)} \varphi'(1-w)\mathcal{C}_2(\gamma^*, w)dw\right)\right)}$$

and

$$B(\gamma^*, H^*(t)) = \frac{\varphi'(1 - H^*(t))\mathcal{C}_2(\gamma^*, H^*(t))}{\varphi'\left(\varphi^{-1}\left(-\int_0^{H^*(t)} \varphi'(1-w)\mathcal{C}_2(\gamma^*, w)dw\right)\right)}$$

with γ^* between γ_n and γ, and $H^*(t)$ between $H_n(t)$ and $H(t)$. Using integration by parts, and noting that $\varphi'(1)\mathcal{C}_1(\gamma^*, 0) = 0$, we further obtain

$$A(\gamma^*, H^*(t)) = \frac{\varphi'(1 - H^*(t))\mathcal{C}_1(\gamma^*, H^*(t)) + \int_0^{H^*(t)} \varphi''(1-w)\mathcal{C}_1(\gamma^*, w)dw}{\varphi'\left(\varphi^{-1}\left(-\int_0^{H^*(t)} \varphi'(1-w)\mathcal{C}_2(\gamma^*, w)dw\right)\right)}$$

Using Condition A1, this gives after some calculations that

$$\sup_{0 \leq t \leq T} |A(\gamma^*, H^*(t))| \leq \frac{3}{|\varphi'(1)|} \sup_{0 \leq t \leq T} |\varphi'(1 - H^*(t))| \tag{13}$$

Similarly, we also find that

$$\sup_{0 \leq t \leq T} |B(\gamma^*, H^*(t))| \leq \frac{1}{|\varphi'(1)|} \sup_{0 \leq t \leq T} |\varphi'(1 - H^*(t))| \tag{14}$$

Using (13) and (14), we find for all $\varepsilon > 0$ and $\eta > 0$ that

$$P\left(\sup_{0 \leq t \leq T} |F_n(t) - F(t)| > \varepsilon\right) \leq P\left(\frac{3}{|\varphi'(1)|} \sup_{0 \leq t \leq T} |\varphi'(1 - H^*(t))| \, |\gamma_n - \gamma| > \frac{\varepsilon}{2}\right)$$
$$+ P\left(\frac{1}{|\varphi'(1)|} \sup_{0 \leq t \leq T} |\varphi'(1 - H^*(t))| \sup_{0 \leq t \leq T} |H_n(t) - H(t)| > \frac{\varepsilon}{2}\right)$$

for which the right hand side of the inequality can be written as

$$P\left(\frac{3}{|\varphi'(1)|} \sup_{0 \leq t \leq T} |\varphi'(1 - H^*(t))| \, |\gamma_n - \gamma| > \frac{\varepsilon}{2}, \sup_{0 \leq t \leq T} |H_n(t) - H(t)| \leq \eta\right)$$
$$+ P\left(\frac{3}{|\varphi'(1)|} \sup_{0 \leq t \leq T} |\varphi'(1 - H^*(t))| \, |\gamma_n - \gamma| > \frac{\varepsilon}{2}, \sup_{0 \leq t \leq T} |H_n(t) - H(t)| > \eta\right)$$
$$+ P\left(\frac{1}{|\varphi'(1)|} \sup_{0 \leq t \leq T} |\varphi'(1 - H^*(t))| \sup_{0 \leq t \leq T} |H_n(t) - H(t)| > \frac{\varepsilon}{2}, \sup_{0 \leq t \leq T} |H_n(t) - H(t)| \leq \eta\right)$$
$$+ P\left(\frac{1}{|\varphi'(1)|} \sup_{0 \leq t \leq T} |\varphi'(1 - H^*(t))| \sup_{0 \leq t \leq T} |H_n(t) - H(t)| > \frac{\varepsilon}{2}, \sup_{0 \leq t \leq T} |H_n(t) - H(t)| > \eta\right)$$

With $0 < \eta < 1 - H(T)$ such that $\sup_{0 \leq t \leq T} |\varphi'(1 - H^*(t))| < |\varphi'(1 - H(T) - \eta)|$, the preceding quantity is further bounded above by

$$P\left(|\gamma_n - \gamma| > \frac{\varphi'(1)\varepsilon}{6\varphi'(1 - H(T) - \eta)}\right) + P\left(\sup_{0 \leq t \leq T} |H_n(t) - H(t)| > \frac{\varphi'(1)\varepsilon}{2\varphi'(1 - H(T) - \eta)}\right)$$
$$+ 2P\left(\sup_{0 \leq t \leq T} |H_n(t) - H(t)| > \eta\right)$$

Choosing $\eta = \frac{\varphi'(1)\varepsilon}{2\varphi'(1 - H(T) - \eta)}$, we easily find that

$$P\left(\sup_{0 \leq t \leq T} |F_n(t) - F(t)| > \varepsilon\right) \leq P\left(|\gamma_n - \gamma| > \frac{\eta}{3}\right) + 3P\left(\sup_{0 \leq t \leq T} |H_n(t) - H(t)| > \eta\right)$$

Using Bernstein's inequality on the first term at the right hand side of the preceding inequality followed by an application of Dvoretzky, Kiefer and Wolfowitz theorem on the second term of the same inequality, yields

$$P\left(\sup_{0 \leq t \leq T} |F_n(t) - F(t)| > \varepsilon\right) \leq 2\exp\left(-\frac{n\eta^2}{6(3\gamma + \eta)}\right) + D\exp\left(-2n\eta^2\right)$$

13

where D is a finite positive constant. Using the Lemma 1, we see that this is further bounded above by

$$2 \exp\left(-\frac{n\alpha^2}{6(3\gamma + \beta)}\right) + D \exp\left(-2n\alpha^2\right)$$

with $\alpha = \frac{\varphi'(1)\varepsilon}{2\varphi'\left(\varphi^{[-1]}\left(\varphi(1-H(T))-\frac{\varphi'(1)\varepsilon}{2}\right)\right)}$ and $\beta = 1 - H(T) - \varphi^{[-1]}\left(\varphi(1-H(T)) - \frac{\varphi'(1)\varepsilon}{2}\right)$

If we take $\varepsilon_n = \varepsilon = Kn^{-1/2}(\log n)^{1/2}$ for some positive constant K, then we note that ε_n is small for large n. Thus, by the Borel-Cantelli lemma we find the strong consistency of F_n. $\qquad\square$

Proof of Theorem 2: To establish the asymptotic representation of $F_n(t)$, we start with a second order Taylor expansion to get

$$F_n(t) - F(t) = \frac{1}{\varphi'(\bar{F}(t))}\left\{\int_0^{H_n(t)} \varphi'(1-w)\mathcal{C}_2(\gamma_n, w)\, dw - \int_0^{H(t)} \varphi'(1-w)\mathcal{C}_2(\gamma, w)\, dw\right\} + r_{n1}(t) \quad (15)$$

where

$$r_{n1}(t) = \frac{\varphi''(\varphi^{-1}(\eta(t)))}{2\left[\varphi'(\varphi^{-1}(\eta(t)))\right]^3}\left\{\int_0^{H_n(t)} \varphi'(1-w)\mathcal{C}_2(\gamma_n, w)\, dw - \int_0^{H(t)} \varphi'(1-w)\mathcal{C}_2(\gamma, w)\, dw\right\}^2$$

with $\eta(t)$ between $-\int_0^{H_n(t)} \varphi'(1-w)\mathcal{C}_2(\gamma_n, w)\, dw$ and $-\int_0^{H(t)} \varphi'(1-w)\mathcal{C}_2(\gamma, w)\, dw$. We denote by

$$I(t) = \int_0^{H_n(t)} \varphi'(1-w)\mathcal{C}_2(\gamma_n, w)\, dw - \int_0^{H(t)} \varphi'(1-w)\mathcal{C}_2(\gamma, w)\, dw.$$

and find that

$$\sup_{0 \le t \le T} |r_{n1}(t)| \le \frac{1}{|\varphi'(1)|^3} \sup_{0 \le t \le T} \varphi''(\varphi^{-1}(\eta(t))) \sup_{0 \le t \le T} |I(t)|^2$$

Using Condition $A1$, we note that $-\int_0^{H_n(t)} \varphi'(1-w)\mathcal{C}_2(\gamma_n, w)\, dw$ and $-\int_0^{H(t)} \varphi'(1-w)\mathcal{C}_2(\gamma, w)\, dw$ are respectively bounded above by $\varphi(1 - H_n(T))$ and $\varphi(1 - H(T))$. On noting that $\varphi''(\cdot)$ is a decreasing function, we further get

$$\sup_{0 \le t \le T} |r_{n1}(t)| \le \frac{1}{|\varphi'(1)|^3} \sup_{0 \le t \le T} \varphi''(\varphi^{-1}(\eta(t))) \sup_{0 \le t \le T} |I(t)|^2 \le \frac{1}{|\varphi'(1)|^3}\varphi''(1 - H_M(T)) \sup_{0 \le t \le T} |I(t)|^2$$

where $H_M(T) = \max(H_n(T), H(T))$. Furthermore, we use the mean value theorem to get

$$I(t) = [\gamma_n - \gamma]\int_0^{H^*(t)} \varphi'(1-w)\mathcal{C}_{12}(\gamma^*, w)\, dw + [H_n(t) - H(t)]\varphi'(1 - H^*(t))\mathcal{C}_2(\gamma^*, H^*(t))$$

with γ^* between γ_n and γ; and $H_n^*(t)$ between $H_n(t)$ and $H(t)$. This gives

$$\sup_{0 \le t \le T} |I(t)| \le |\gamma_n - \gamma| \sup_{0 \le t \le T}\left|\int_0^{H^*(t)} \varphi'(1-w)\mathcal{C}_{12}(\gamma^*, w)\, dw\right|$$
$$+ \sup_{0 \le t \le T} |H_n(t) - H(t)| \sup_{0 \le t \le T} |\varphi'(1 - H^*(t))\mathcal{C}_2(\gamma^*, H^*(t))|$$

Using integration by parts and recalling that $\mathcal{C}_1(\gamma^*, 0) = 0$ for all $\gamma^* \in (0, 1]$, we obtain

$$\sup_{0 \le t \le T}\left|\int_0^{H^*(t)} \varphi'(1-w)\mathcal{C}_{12}(\gamma^*, w)\, dw\right| = \sup_{0 \le t \le T} |\varphi'(1 - H^*(t))\mathcal{C}_1(\gamma^*, H^*(t))|$$
$$+ \sup_{0 \le t \le T}\left|\int_0^{H^*(t)} \varphi''(1-w)\mathcal{C}_1(\gamma^*, w)\, dw\right|$$

14

By similar deductions as in the proof of Theorem 1, we obtain

$$\sup_{0 \le t \le T} \left| \int_0^{H^*(t)} \varphi'(1 - w)\mathcal{C}_{12}(\gamma^*, w)dw \right| \le 3|\varphi'(1 - H_M(T))|$$

Using the preceding inequality, we get after some calculations that

$$\sup_{0 \le t \le T} |I(t)| \le 3|\varphi'(1 - H_M(T))||\gamma_n - \gamma| + |\varphi'(1 - H_M(T))| \sup_{0 \le t \le T} |H_n(t) - H(t)|$$

As a consequence of the Law of Iterated Logarithm, we also have that

$$|\gamma_n - \gamma| = O\left(n^{-1/2} (\log n)^{1/2}\right) \quad \text{a.s.} \quad \text{and} \quad \sup_{0 \le t \le T} |H_n(t) - H(t)| = O\left(n^{-1/2} (\log n)^{1/2}\right) \quad \text{a.s.} \tag{16}$$

Since $H(T) < 1$ and $H_n(T) \to H(T)$ a.s. (see Strong Law of large numbers), we may suppose that $T < T_{H_n}$. Hence, we obtain that

$$\sup_{0 \le t \le T} |I(t)| = O\left(n^{-1/2} (\log n)^{1/2}\right) \quad a.s.$$

and

$$\sup_{0 \le t \le T} |r_{n1}(t)| = O\left(n^{-1} \log n\right) \quad \text{a.s.}$$

We can further decompose the main term in (15) by using a second order Taylor expansion to get

$$\int_0^{H_n(t)} \varphi'(1 - w)\mathcal{C}_2(\gamma_n, w) \, dw - \int_0^{H(t)} \varphi'(1 - w)\mathcal{C}_2(\gamma, w) \, dw$$

$$= [\gamma_n - \gamma] \int_0^{H(t)} \varphi'(1 - w)\mathcal{C}_{12}(\gamma, w) \, dw + [H_n(t) - H(t)] \varphi'(1 - H(t))\mathcal{C}_2(\gamma, H(t)) \tag{17}$$

$$+ r_{n2}(t) + r_{n3}(t) + r_{n4}(t)$$

where

$$r_{n2}(t) = \frac{1}{2} [\gamma_n - \gamma]^2 \int_0^{H^*(t)} \varphi'(1 - w)\mathcal{C}_{112}(\gamma^*, w) \, dw$$

$$r_{n3}(t) = \frac{1}{2} [H_n(t) - H(t)]^2 \{\varphi'(1 - H^*(t))\mathcal{C}_{22}(\gamma^*, H^*(t)) - \varphi''(1 - H^*(t))\mathcal{C}_2(\gamma^*, H^*(t))\}$$

$$r_{n4}(t) = [\gamma_n - \gamma] [H_n(t) - H(t)] \varphi'(1 - H^*(t))\mathcal{C}_{12}(\gamma^*, H^*(t))$$

with $\mathcal{C}_{112}(u, v) = \frac{\partial^3 \mathcal{C}(u,v)}{\partial u^2 \partial v}$ and γ^* between γ_n and γ; and $H^*(t)$ between $H_n(t)$ and $H(t)$. We now determine the rate of convergence of $r_{n2}(t), r_{n3}(t)$ and $r_{n4}(t)$. Integrating by parts, we have

$$r_{n2}(t) = \frac{1}{2}[\gamma_n - \gamma]^2 \left\{ \varphi'(1 - H^*(t))\mathcal{C}_{11}(\gamma^*, H_n^*(t)) - \varphi'(1)\mathcal{C}_{11}(\gamma^*, 0) + \int_0^{H^*(t)} \varphi''(1 - w)\mathcal{C}_{11}(\gamma^*, w)dw \right\}$$

Working as before, we get

$$\sup_{0 \le t \le T} |r_{n2}(t)| \le 4|\varphi'(1 - H_M(T))| \sup_{0 \le u,v \le 1} |\mathcal{C}_{11}(u, v)| [\gamma_n - \gamma]^2.$$

For $r_{n3}(t)$, we have

$$\sup_{0 \le t \le T} |r_{n3}(t)| \le \left\{ \varphi'(1 - H_M(T)) \sup_{0 \le u,v \le 1} |\mathcal{C}_{22}(u, v)| + \varphi''(1 - H_M(T)) \right\} \sup_{0 \le t \le T} |H_n(t) - H(t)|^2$$

15

Since $H(T) < 1$ and $H_n(T) \to H(T)$ a.s. (see Strong Law of large numbers), we suppose as before that $T < T_{H_n}$. Consequently, we obtain

$$\sup_{0 \le t \le T} |r_{n2}(t)| = O\left(n^{-1} \log n\right) \quad \text{a.s.} \quad \text{and} \quad \sup_{0 \le t \le T} |r_{n3}(t)| = O\left(n^{-1} \log n\right) \quad \text{a.s.}$$

Similarly, we also get that

$$\sup_{0 \le t \le T} |r_{n4}(t)| = O\left(n^{-1} \log n\right) \quad \text{a.s.}$$

Let $r_n(t) = r_{n1}(t) + \frac{1}{\varphi'(\bar{F}(t))}(r_{n2}(t) + r_{n3}(t) + r_{n4}(t))$. It is straight forwards to see that

$$\sup_{0 \le t \le T} |r_n(t)| = O\left(n^{-1} \log n\right) \quad \text{a.s.}$$

since $\varphi'(\bar{F}(t)) \le \varphi'(1) < 0$ for all $t \in [0, T]$. Using this together with (17) and (15), concludes the proof. $\quad\square$

Proof of Theorem 3: To show the weak convergence of the process $\sqrt{n}\left(F_n(\cdot) - F(\cdot)\right)$, we note that by Theorem 2 we only have to prove the weak convergence of the main term in the asymptotic representation which is given by

$$W_n(t) = n^{-1/2} \sum_{i=1}^{n} k\left(t; Z_i, \delta_i\right).$$

with $k\left(t; Z_i, \delta_i\right)$ i.i.d. and defined by (8). After some calculations we get that, for all $t \in [0, T]$,

$$E\left[k\left(t; Z, \delta\right)\right] = 0$$

and, for all $0 \le s, t \le T$,

$$\Gamma(s, t) = \text{Cov}\left[k\left(s; Z, \delta\right), k\left(t; Z, \delta\right)\right] = E\left[k\left(s; Z, \delta\right), k\left(t; Z, \delta\right)\right]$$

$$= \frac{1}{\varphi'(\bar{F}(s))\varphi'(\bar{F}(t))} \times$$

$$\left\{ \int_0^{H(s)} \varphi'(1-w)\mathcal{C}_{12}\left(\gamma, w\right) dw \int_0^{H(t)} \varphi'(1-w)\mathcal{C}_{12}\left(\gamma, w\right) dw E\left[\left(\mathbf{1}\{\delta = 1\} - \gamma\right)^2\right] \right.$$

$$+ \varphi'(\bar{H}(s))\mathcal{C}_2\left(\gamma, H(s)\right) \int_0^{H(t)} \varphi'(1-w)\mathcal{C}_{12}\left(\gamma, w\right) dw E\left[\left(\mathbf{1}\{\delta = 1\} - \gamma\right)\left(\mathbf{1}\{Z \le s\} - H(s)\right)\right]$$

$$+ \varphi'(\bar{H}(t))\mathcal{C}_2\left(\gamma, H(t)\right) \int_0^{H(s)} \varphi'(1-w)\mathcal{C}_{12}\left(\gamma, w\right) dw E\left[\left(\mathbf{1}\{\delta = 1\} - \gamma\right)\left(\mathbf{1}\{Z \le t\} - H(t)\right)\right]$$

$$\left. + \varphi'(\bar{H}(s))\varphi'(\bar{H}(t))\mathcal{C}_2\left(\gamma, H(s)\right)\mathcal{C}_2\left(\gamma, H(t)\right) E\left[\left(\mathbf{1}\{Z \le s\} - H(s)\right)\left(\mathbf{1}\{Z \le t\} - H(t)\right)\right] \right\}$$

$$= \frac{1}{\varphi'(\bar{F}(s))\varphi'(\bar{F}(t))} \times$$

$$\left\{ \gamma\left(1-\gamma\right) \int_0^{H(s)} \varphi'(1-w)\mathcal{C}_{12}\left(\gamma, w\right) dw \int_0^{H(t)} \varphi'(1-w)\mathcal{C}_{12}\left(\gamma, w\right) dw \right.$$

$$+ \left(H^u(s) - \gamma H(s)\right) \varphi'(\bar{H}(s))\mathcal{C}_2\left(\gamma, H(s)\right) \int_0^{H(t)} \varphi'(1-w)\mathcal{C}_{12}\left(\gamma, w\right) dw$$

$$+ \left(H^u(t) - \gamma H(t)\right) \varphi'(\bar{H}(t))\mathcal{C}_2\left(\gamma, H(t)\right) \int_0^{H(s)} \varphi'(1-w)\mathcal{C}_{12}\left(\gamma, w\right) dw$$

$$\left. + \left(H(s \wedge t) - H(s)H(t)\right) \varphi'(\bar{H}(s))\varphi'(\bar{H}(t))\mathcal{C}_2\left(\gamma, H(s)\right)\mathcal{C}_2\left(\gamma, H(t)\right) \right\}$$

16

To prove the weak convergence of $W_n(t)$, we verify the conditions of Theorem 2.5.6 of van der Vaart and Wellner (2000). Hereto we show that the class of functions $\mathcal{F}$ given by

$$\mathcal{F} = \{k(t; z, d) | t \in [0, T]\}$$

is Donsker.

For the first term in (8), we note that $\frac{(\mathbf{1}\{d=1\}-\gamma)}{\varphi'(\bar{F}(t))} \int_0^{H(t)} \varphi'(1-w)\mathcal{C}_{12}(\gamma, w) \, dw$ is uniformly bounded over t. Furthermore we see that the second function $z \to \frac{(\mathbf{1}\{z \leq t\}-H(t))}{\varphi'(\bar{F}(t))}\varphi'(\bar{H}(t))\mathcal{C}_2(\gamma, H(t))$ is uniformly bounded over t and is a monotone function of z. Hence, we have that $k(t; z, d)$ is a monotone function of z and

$$
\begin{aligned}
\sup_{0 \leq t \leq T} |k(t; z, d)| &= \sup_{0 \leq t \leq T} \left| \frac{(\mathbf{1}\{d=1\}-\gamma)}{\varphi'(\bar{F}(t))} \left(\varphi'(\bar{H}(t))\mathcal{C}_1(\gamma, H(t)) + \int_0^{H(t)} \varphi'(1-w)\mathcal{C}_1(\gamma, w)dw \right) \right. \\
&\qquad \left. + \frac{(\mathbf{1}\{z \leq t\}-H(t))}{\varphi'(\bar{F}(t))}\varphi'(\bar{H}(t))\mathcal{C}_2(\gamma, H(t)) \right| \\
&\leq \frac{\varphi'(\bar{H}(T))}{\varphi'(1)} \left(3 + \sup_{0 \leq t \leq T} \mathcal{C}_2(\gamma, H(t)) \right) \leq M
\end{aligned}
$$

where M is a finite positive constant. Using Theorem 2.7.5 of van der Vaart and Wellner (2000), we get that the bracketing number $N_{[\,]}(\varepsilon, \mathcal{F}, L_2(P)) = O\left(\exp\left(\frac{K}{\varepsilon}\right)\right)$ with K a positive constant. Hence, we get that

$$\int_0^\infty \sqrt{\log N_{[\,]}(\varepsilon, \mathcal{F}, L_2(P))} d(\varepsilon) < \infty$$

which shows the class $\mathcal{F}$ is Donsker and completes the proof. $\qquad \square$