Estimation in a Competing Risks Proportional Hazards Model Under Length-biased Sampling With Censoring

Jean-Yves Dauxois, Agathe Guilloux, Syed N.U.A. Kirmani

ESTIMATION IN A COMPETING RISKS PROPORTIONAL HAZARDS MODEL UNDER LENGTH-BIASED SAMPLING WITH CENSORING

JEAN-YVES DAUXOIS, AGATHE GUILLOUX, AND SYED N.U.A. KIRMANI

ABSTRACT. What population does the sample represent? The answer to this question is of crucial importance when estimating a survivor function in duration studies. As is well-known, in a stationary population, survival data obtained from a cross-sectional sample taken from the population at time t_0 represents not the target density $f(t)$ but its length-biased version proportional to $tf(t)$, for $t > 0$. The problem of estimating survivor function from such length-biased samples becomes more complex, and interesting, in presence of competing risks and censoring. This paper lays out a sampling scheme related to a mixed Poisson process and develops nonparametric estimators of the survivor function of the target population assuming that the two independent competing risks have proportional hazards. Two cases are considered: with and without independent consoring before length biased sampling. In each case, the weak convergence of the process generated by the proposed estimator is proved. A well-known study of the duration in power for political leaders is used to illustrate our results. Finally, a simulation study is carried out in order to assess the finite sample behaviour of our estimators.

1. INTRODUCTION

The central problem in the analysis of duration data is the efficient estimation of the distribution of the time Z between two specified events under different sampling scenarios. The two events whose gap time is of interest will be called the initiating and terminating events. The two events may be HIV infection and death, successive hospitalizations due to a disease or entry and exit from the workforce. Frequently, the distribution of Z must be estimated from a cross-sectional sample at time t_0 consisting of subjects who have experienced the initiating event, but not the terminating event, prior to t_0. In the context of epidemiology and survival analysis, cross-sectional studies are concerned with prevalent rather than incident cases. As

Date: July 3, 2011.
Key words and phrases. Cross-sectional sample, Cumulative incidence function, Functional delta-method, Gaussian process, Lexis diagram, Mixed Poisson process, Nonparametric estimation, Weak convergence.

it is well-known, such data suffer from length-bias in the sense that Z^b, the time gap between initiating and terminating events for a cross-sectionally selected subject, is stochastically larger than Z with $dP(Z^b < t)$ proportional to $tdP(Z < t)$. This phenomenon, to be referred to as length-biased sampling (LBS), was noted by McFadden (1962) for lengths of intervals in a stationary point process, by Blumenthal (1967) in industrial life testing and by Cox (1969) for estimating the distribution of fiber lengths in a fabric. Zelen & Feinleib (1969) recognized LBS in screening for chronic diseases while Simon (1980) noted its relevance in etiologic studies. The source of LBS is the simple fact that, when drawing observations from a set of subjects in a particular state, the probability of being included in the sample is proportional to the sojourn time in that state. This, therefore, leads to disproportionate representation of longer durations. Vardi (1982) was the first to consider nonparametric estimation in the presence of LBS. He derived and studied the unconditional nonparametric maximum likelihood estimate (NPMLE) of the distribution function of Z on the basis of two independent samples, one a sample from Z and the other a sample from its length-biased version Z^b. We refer to Vardi (1982), Vardi (1985), Vardi (1989), Gill et al. (1988) and Vardi & Zhang (1992) for further theoretical developments. More recently, Asgharian et al. (2002) obtained the unconditional NPMLE of the survivor function of Z and its asymptotic properties when the data are purely length-biased with a special case of right censoring.

Length-biased data can be considered as a special case of left-truncation if the occurrence time of the initiating event is uniformly distributed. Here, truncation refers to the fact that a subject can not be observed at t_0 if it has experienced the terminating event before t_0. There is an extensive literature on nonparametric estimation under left truncation. We refer to Turnbull (1976), Woodroofe (1985), Wang et al. (1986), Tsai et al. (1987), Wang (1991) and Wang et al. (1993).

The motivation for the present paper comes from the conjunction of LBS, competing risks (CR) and Proportional Hazards (PH). Suppose that the terminating event can occur in either of two competing ways A and B, e.g. A may be death due to a specific disease, say cancer, and B death from a natural cause. Let X (resp. Y) be the latent, or potential survival time associated with risk A (resp. B) and let us assume that X and Y have proportional hazards. Suppose that the terminal event can also be due to independent right censoring. Then the time gap between initiating and terminating events is of length $T = (X \wedge Y) \wedge C$ where C is the censoring time and $x \wedge y$ denotes the minimum of x and y. However, under LBS, the random variable (r.v.) T is not observable. To be precise, we shall consider the following situation. The observed sample consists of n independent individuals, cross-sectionally selected at t_0, who were exposed to risk A at known time points $\sigma_i \leq t_0$, $i = 1, ..., n$. These individuals are followed up to death, from cause A or B, or censoring time. For the ith member of the length-biased sample, the r.v. X_i^b (resp. Y_i^b and C_i^b) will denote the potential survival time of the ith subject when facing risk A (resp. B and censoring). The sample data thus consists of the n pairs (T_i^b, δ_i^b) where $T_i^b = X_i^b \wedge Y_i^b \wedge C_i^b$, and δ_i^b indicates the mode of termination (death due to A, death due to B or censoring). Our main objective is to estimate the survivor function $\bar{G}_X(t) = \mathrm{pr}(X > t)$ based on such a sample.

The setup described above, namely, the LBS-CR-PH data with independent censoring preceding LBS is the first framework to be considered. We will refer to it as "Case 1: Independent censoring before LBS".

In a number of practical situations, two risks A and B, which have proportional hazards, compete unhindered by the risk of censoring. An estimator of $\bar{G}_X(\cdot)$, based on a LBS, can be introduced and its large sample properties studied. However, without any additional serious mathematical complications, it is possible to introduce a more general estimator of $\bar{G}_X(\cdot)$ and study its large sample behavior even if we allow the possibility of independent random censoring after the cross-sectional sample has been selected. Such post-LBS censoring may or may not be justified in specific practical situations. An excellent discussion, with examples, of various censoring issues in biased-sampling situations is given by Tsai (2009). We will allow the post-LBS censoring scenario in our "Case 2: No censoring before LBS". In this case, the observable random variables are $T_i = Z_i^b \wedge C_i$ where Z_i^b is the LBS observation of $Z_i = X_i \wedge Y_i$ and δ_i which gives the type of the observed terminal event. Case 2 encompasses the "no possibility of censoring" scenario and we easily obtain the estimator and its large sample properties for the without censoring case from our results for Case 2.

As far as we know, these problems have not been considered in the literature so far. Huang & Wang (1995) did consider the LBS-CR set up but they were concerned with estimation of crude hazard functions and occurrence probabilities rather than estimation of $\bar{G}_X(\cdot)$. Dauxois & Guilloux (2008) have considered the problem of the nonparametric inference of the Cumulative Incidence Functions under competing risks and selection-biased sampling. But no proportional hazards assumption was made in their work.

The outline of this paper is as follows. The two cases described above are considered respectively in Section 2 and 3. Estimators of $\hat{\bar{G}}_X(\cdot)$ are obtained in each case and their large sample behaviors are studied. In Section 4, we apply our methodology to the data set introduced by Bienen &

M'Lan (1991), whereas in Section 5 we study the behaviour of our estimator through Monte Carlo simulation of its mean integrated squared error (MISE). An appendix details the proofs of technical lemmas used in Sections 2 and 3.

2. CASE 1: INDEPENDENT CENSORING BEFORE LENGTH BIASED SAMPLING

The objective of this section is to develop a framework for study of length-biased sampling (LBS) in the setup of competing risks (CR). From now on and for convenience, the initiating and terminating events of interest will be called birth and death, respectively.

2.1. **Initial population.** We shall consider a population of individuals (to be called *initial population*) who are subject to two competing causes, A and B, of death. The CR model will be described in terms of latent survival times X and Y where X (resp. Y) is a positive random variable (r.v.) representing the age at death in the hypothetical situation in which A (resp. B) is the only possible cause of death. Frequently, there is a primary cause of interest. For example, the target interest of study may be death due to breast cancer. In such cases, we shall take A as the primary risk of interest and all other causes will be lumped together as B. The individual lifetime will be denoted by $Z = X \wedge Y$.

In the present paper, we are concerned with the important special case in which the risks A and B have proportional hazards. Thus, it will be assumed in this paper that there exists $\beta > 0$ such that for all $t > 0$:

$$\Lambda_Y(t) = \beta \Lambda_X(t)$$

where $\Lambda_X(\cdot)$ and $\Lambda_Y(\cdot)$ are the cumulative hazard functions of X and Y, respectively. Equivalently, denoting by $\bar{G}_X(\cdot)$ and $\bar{G}_Y(\cdot)$ the survival functions of X and Y, we will assume in the following that

$$\bar{G}_Y(t) = (\bar{G}_X(t))^\beta \tag{1}$$

for all $t > 0$. This model, often called "Koziol-Green" model, has been widely studied in classical survival analysis literature, see e.g. Chen & Lin (1987), Csörgő (1988), Gather & Pawlitschko (1998) and Kirmani & Dauxois (2004).

The constant β gives the odds on death due to cause B, i.e.:

$$\frac{\text{pr}(Y \leq X)}{\text{pr}(X \leq Y)} = \beta,$$

while the theoretical proportion α of deaths from cause A among all deaths is given by:

$$\alpha = \mathrm{pr}\,(X \leq Y) = \frac{1}{\beta+1}. \tag{2}$$

We will assume that the lifetime Z may suffer from independent random-right censoring. Let C denote the censoring time and $\bar{H}(\cdot)$ its survival function. Then, $T = Z \wedge C$ denotes the age at terminating event (death from cause A, from cause B or censoring) and δ indicates the mode of termination:

$$\delta = \begin{cases} 0 & \text{if} \quad C < Z \\ 1 & \text{if} \quad T = Z = X \\ 2 & \text{if} \quad T = Z = Y \end{cases}.$$

It has to be noted that under this independent censoring mechanism, the proportion of deaths from cause A among all termination causes (A, B or censoring) is still equal to α, that is:

$$\alpha = \frac{\mathrm{pr}\,(X \leq Y, X \leq C)}{\mathrm{pr}\,(X \wedge Y \leq C)}. \tag{3}$$

2.2. **Length-biased population.** Let $\{i \in I\}$ denotes the initial population described in the previous subsection. Let X_i and Y_i be the latent survival times (corresponding to risks A and B, respectively) and C_i be the latent censoring time for individual i. The age of individual i at terminal event is $T_i = Z_i \wedge C_i = X_i \wedge Y_i \wedge C_i$ and δ_i indicates the mode of termination.

Now, let σ_i be the calendar time of birth of the individual. A convenient graphical representation of the lifespan of an individual born at calendar time σ_i and experiencing a terminal event at age t_i is given by the well-known Lexis diagram (see Fig. 1). This diagram consists of line segments in a rectangular coordinate system with calendar time as abscissa and the age as ordinate such that the life (or time from birth to censoring) is represented by the line segment joining the points $(\sigma_i, 0)$ and $(\sigma_i + t_i, t_i)$. The Lexis diagram and associated point processes described in Brillinger (1986), Keiding (1990), and Lund (2000) provide useful settings for analyzing lifetimes. It is particularly important in describing sampling patterns for selection of individuals in a study. It also helps in visualizing follow-up patterns and truncation of lifetimes.

A random sample cross-sectionally selected at calendar time t_0 is not really a random sample from the initial population I but, in fact, from the population $J = \{i \in I : (\sigma_i, x_i, y_i, c_i) \in E\}$ where $E = \{(\sigma, x, y, c) : \sigma \leq t_0, \sigma + x \geq t_0, \sigma + y \geq t_0, \sigma + c \geq t_0\}$. Individuals with age at terminal event $T_i = Z_i \wedge C_i = X_i \wedge Y_i \wedge C_i$ shorter than $t_0 - \sigma_i$ are excluded from

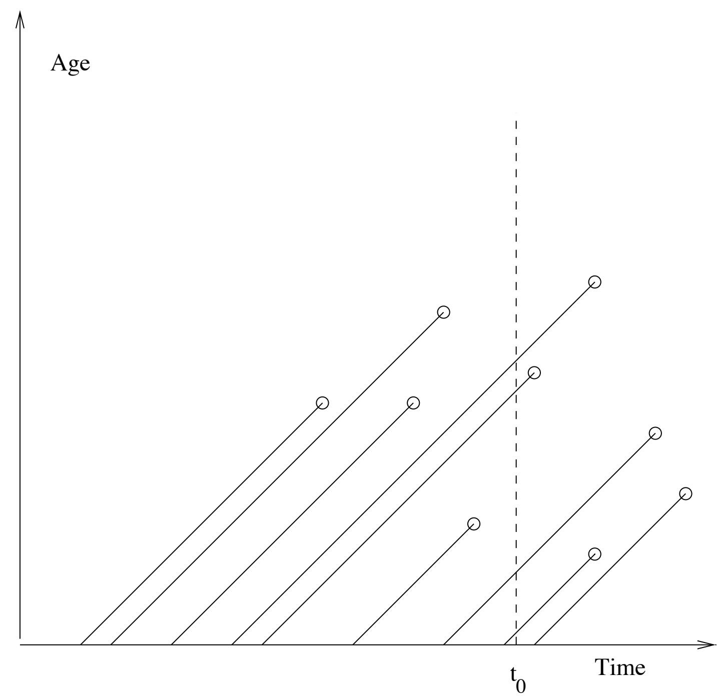

FIGURE 1. A Lexis diagram representation of lifespans

the population J. That is, the time T_i is left truncated by the time $t_0 - \sigma_i$. Individuals with birthtimes $\sigma_i \geq t_0$ are also excluded from the sample.

Thus, the observable r.v. is not T_i but T_i^b, a r.v. whose probability distribution is the same as the conditional distribution of $T_i = Z_i \wedge C_i$ given $\{(\sigma_i, X_i, Y_i, C_i) \in E\}$. The mode of termination associated with T_i^b will be denoted δ_i^b. We shall refer to T_i^b as the length-biased version of T_i and $\{j \in J\}$ as the *length-biased population*.

The following proposition provides a key fact: it gives the probability distribution of (T^b, δ^b) defined above. It will be seen that, under the assumptions made, the distribution of (T^b, δ^b) will be independent of σ. Thus, the pairs $(T_1^b, \delta_1^b), \ldots, (T_n^b, \delta_n^b)$ of the n individuals in the sample selected at t_0 will be independent copies of (T^b, δ^b).

Theorem 1. *Suppose that:*
(i) the birth process $\eta = \sum_{i \in I} \varepsilon_{\sigma_i}$*, where* ε_{σ_i} *denotes the random measure concentrated on* σ_i*, is a mixed Poisson process with random intensity* φ*;*
(ii) conditionally on the process η*, the vectors* (X_i, Y_i, C_i)*, for* $i \in I$*, are independent and identically distributed with common probability density function (p.d.f.)* $g_X(\cdot)g_Y(\cdot)h(\cdot)$ *with respect to the Lebesgue measure on* $\mathbb{R}_+^3$ *(where* $g_X(\cdot)$*,* $g_Y(\cdot)$ *and* $h(\cdot)$ *are respectively the p.d.f of* X*,* Y *and* C*);*
(iii) $E(X) < \infty$ *and* $E(Y) < \infty$ *and* $E(C) < \infty$*.*

Then, the distribution of the pair (T^b, δ^b) is specified by the following expression of its three sub-distribution functions:

$$\begin{array}{rclcl} F_0^b(t) & = & \mathrm{pr}\left(T^b \leq t, \delta^b = 0\right) & = & E(T)^{-1} \int_0^t ch(c)\bar{G}_X(c)\bar{G}_Y(c)dc \\ F_1^b(t) & = & \mathrm{pr}\left(T^b \leq t, \delta^b = 1\right) & = & E(T)^{-1} \int_0^t xg_X(x)\bar{G}_Y(x)\bar{H}(x)dx \\ F_2^b(t) & = & \mathrm{pr}\left(T^b \leq t, \delta^b = 2\right) & = & E(T)^{-1} \int_0^t yg_Y(y)\bar{G}_X(y)\bar{H}(y)dy \end{array}$$

for $t > 0$.

Proof of Theorem 1.
Although the above result is merely the competing risks statements of the well-known length-biased density Lund (2000); van Es et al. (2000), we offer the following derivation. First note that η is a point process on $\mathbb{R}$ such that, for each Borel set S in $\mathbb{R}$, the r.v. $\eta(S)$ gives the number of births encountered in S. We assume that $\eta(S) < \infty$ almost surely. For each individual i, the birth time σ_i is marked by the pair of latent survival times (X_i, Y_i, C_i). We now define the Lexis point process

$$\mu = \sum_{i \in I} \varepsilon_{(\sigma_i, X_i, Y_i, C_i)}$$

on $(\mathbb{R} \times \mathbb{R}_+^3, \mathcal{B}_{\mathbb{R}} \otimes \mathcal{B}_{\mathbb{R}_+^3})$, where $\mathcal{B}_{\mathbb{R}}$ (resp. $\mathcal{B}_{\mathbb{R}_+^3}$) denotes the Borel σ-algebra on $\mathbb{R}$ (resp. $\mathbb{R}_+^3$). This has the advantage of showing that $\mu_{|\varphi}$, the process μ conditional on the intensity φ of the mixed Poisson process, is Poisson with intensity

$$(\sigma, x, y, c) \mapsto \lambda_{|\varphi}(\sigma, x, y, c) = \varphi g_X(x) g_Y(y) h(c)$$

and with mean-measure $\Lambda_{|\varphi}(\cdot)$ defined, for each Borel set S on $\mathbb{R} \times \mathbb{R}_+^3$, by

$$\Lambda_{|\varphi}(S) = \int_S \lambda_{|\varphi}(\sigma, x, y, c) d\sigma dx dy dc.$$

We refer to Kingman (1993) for the marking theorem exploited here. Further, let $\mu_{E|\varphi}(\cdot) = \mu_{|\varphi}(\cdot \cap E)$ be the restriction of the Poisson process $\mu_{|\varphi}$ to the measurable set $E = \{(\sigma, x, y, c) : \sigma \leq t_0, \sigma + x \geq t_0, \sigma + y \geq t_0, \sigma + c \geq t_0\}$. Then, by the well-known restriction theorem for Poisson processes Kingman (1993), the process $\mu_{E|\varphi}$ is Poisson on $\mathbb{R} \times \mathbb{R}_+^3$ with mean-measure $\Lambda_{E|\varphi}(\cdot)$ defined, for all Borel set S in $\mathbb{R} \times \mathbb{R}_+^3$, by

$$\Lambda_{E|\varphi}(S) = \Lambda_{|\varphi}(S \cap E) = \int_{S \cap E} \lambda_{|\varphi}(\sigma, x, y, c) d\sigma dx dy dc.$$

Our mode of sampling is equivalent to selecting a random subset $E^* \subset E$ such that $\mu_{|\varphi}(E^* \cap E) = n$ is the sample size. By the order statistics property of Poisson processes, see e.g. Crump (1975), given $\mu_{|\varphi}(E) = N$, the points of the Poisson process $\mu_{|\varphi}(\cdot \cap E)$ look exactly like $\mu_{|\varphi}(E)$ independent r.v.'s, with common probability measure

$$\text{pr}_{E|\varphi}(\cdot) = \frac{\Lambda_{|\varphi}(\cdot \cap E)}{\Lambda_{|\varphi}(E)}$$

on Borel subsets of $\mathbb{R} \times \mathbb{R}_+^3$. Hayakawa (2000) showed that the order statistics property characterizes a mixed Poisson process within the general class of point processes. This indicates that assumption (i) can not be weakened.

Let X^b and Y^b denote the latent survival times (corresponding to risks A and B respectively) and C^b the censoring time for an individual in J where, as defined earlier, $J = \{i \in I : (\sigma_i, X_i, Y_i, C_i) \in E\}$. Let x_0, y_0 and c_0 be positive reals and $S = \{(\sigma, x, y, c) : x \leq x_0, y \leq y_0, c \leq c_0\}$. It follows from the above discussion that, conditionally on $\mu_{|\varphi}(E) = N$:

$$\begin{aligned} \text{pr}_{|\varphi}(X^b \leq x_0, Y^b \leq y_0, C^b \leq c_0) &= \frac{\Lambda_{|\varphi}(S \cap E)}{\Lambda_{|\varphi}(E)} \\ = \quad & \frac{\int_{S\cap E} \varphi g_X(x) g_Y(y) h(c) d\sigma dx dy dc}{\int_E \varphi g_X(x) g_Y(y) h(c) d\sigma dx dy dc} \\ = \quad & \frac{\int_0^{x_0}\int_0^{y_0}\int_0^{c_0}\int_{t_0-(x\wedge y\wedge c)}^{t_0} g_X(x) g_Y(y) h(c) d\sigma dx dy dc}{\int_{-\infty}^{t_0} \bar{G}_X(t_0-\sigma)\bar{G}_Y(t_0-\sigma)\bar{H}(t_0-\sigma) d\sigma} \\ = \quad & \frac{\int_0^{x_0}\int_0^{y_0}\int_0^{c_0} (x\wedge y\wedge c) g_X(x) g_Y(y) h(c) dx dy dc}{E(T)}. \end{aligned}$$

Since the last expression does not involve φ, integrating w.r. its distribution, we get

$$\begin{aligned} & \text{pr}(X^b \leq x_0, Y^b \leq y_0, C^b \leq c_0) \\ = \quad & \frac{\int_0^{x_0}\int_0^{y_0}\int_0^{c_0} (x\wedge y\wedge c) g_X(x) g_Y(y) h(c) dx dy dc}{E(T)}. \end{aligned}$$

The proposition then follows by differentiation and integration on the proper sets.□

It has to be noted that the distribution function of the r.v. T^b is the length-biased version of the distribution function of T. Indeed, from Theorem 1, we get for all $t \geq 0$:

$$\text{pr}\left(T^b \leq t\right) = F^b(t) = F_0^b(t) + F_1^b(t) + F_2^b(t) = \frac{1}{E(T)} \int_0^t u dF(u),$$

where $F(\cdot)$ is the distribution function of the r.v. T. Consequently, by the knwon inversion formula of Cox (1969), the distribution function $F(\cdot)$ is expressible as:

$$F(t) = E(T)\int_0^t \frac{1}{u}dF^b(u) = \frac{\int_0^t \frac{1}{u}dF^b(u)}{\int_0^\infty \frac{1}{u}dF^b(u)}. \tag{4}$$

In the present paper we are concerned with the special case where the risks A and B have proportional hazards (see (1)). Under this assumption, the sub-distribution functions of (T^b, δ^b) given in Theorem 1 have simplify as follows:

$$\begin{array}{rcl} F_0^b(t) &=& E(T)^{-1}\int_0^t ch(c)\bar{G}_X^{\beta+1}(c)dc \\ F_1^b(t) &=& E(T)^{-1}\int_0^t xg_X(x)\bar{G}_X^{\beta}(x)\bar{H}(x)dx \\ F_2^b(t) &=& E(T)^{-1}\int_0^t y\beta g_X(y)\bar{G}_X^{\beta}(y)\bar{H}(y)dy \end{array}, \tag{5}$$

for all $t > 0$.

2.3. **Statistical Inference.** Our aim is to estimate the survivor function $\bar{G}_X(\cdot) = \mathrm{pr}(X > \cdot)$ of the cause of primary interest on the basis of a length-biased sample obtained from the initial population described earlier (competing risks with proportional hazards and independent censoring). Adhering to the notations of the previous section, the observable r.v.'s are T^b and δ^b rather than T and δ. Recall that the probability distribution of (T^b, δ^b) is the conditional distribution of (T, δ) given $\{(\sigma, X, Y, C) \in E\}$.

Under the assumption that X and Y are independent with proportional hazards, the unconditional distribution of $Z = X \wedge Y$ has p.d.f.

$$g_Z(z) = (1+\beta)g_X(z)(\bar{G}_X(z))^\beta,\ z > 0,$$

and distribution function $\bar{G}_Z(\cdot) = \bar{G}_X^{\beta+1}(\cdot)$. Moreover, from Theorem 1 one can see that the sub-distribution function $F_{12}^b(\cdot)$ defined, for all $t \geq 0$, by:

$$F_{12}^b(t) = \mathrm{pr}\left(T^b \leq t, \delta^b \neq 0\right) = F_1^b(t) + F_2^b(t),$$

may be rewritten as:

$$F_{12}^b(t) = \frac{\int_0^t zg_Z(z)\bar{H}(z)dz}{E(T)},\ \text{for all } t \geq 0. \tag{6}$$

That is, the sub-distribution function $F_{12}^b(\cdot)$ is a weighted version of the distribution of Z, with weight function $t \mapsto t\bar{H}(t)$. It has to be noted that it is not a length-biased version since it is not proportional to $\int_0^t zg_Z(z)dz$, for all $t \geq 0$. Consequently, the well-known inversion formula of Cox (1969) does not apply here. We will instead follow the approach of de Uña-Álvarez (2004).

By taking the derivative in Equation (6), we get for all $t > 0$:

$$E(T)\frac{1}{t}dF_{12}^b(t) = g_Z(t)\bar{H}(t)dt.$$

Equivalently,

$$E(T)\frac{1}{t\bar{G}_Z(t)\bar{H}(t)}dF_{12}^b(t) = \frac{g_Z(t)}{\bar{G}_Z(t)}dt.$$

Note that the independence between Z and C gives us $\bar{F}(\cdot) = 1 - F(\cdot) = \bar{G}_Z(\cdot)\bar{H}(\cdot)$. Hence, integrating the last equality and using Equation (4) we obtain, for all $t > 0$:

$$\int_0^t \frac{1}{z\int_z^\infty \frac{1}{u}dF^b(u)}dF_{12}^b(z) = \int_0^t \frac{g_Z(z)}{\bar{G}_Z(z)}dz = \Lambda_Z(t), \tag{7}$$

where $\Lambda_Z(\cdot) = -\log(\bar{G}_Z(\cdot))$ is the cumulative hazard function of Z.

Moreover, as X and Y have proportional hazards, we have: $\Lambda_Z(\cdot) = \Lambda_X(\cdot) + \Lambda_Y(\cdot) = (1+\beta)\Lambda_X(\cdot)$. Finally, using the product integral notion Andersen et al. (1993) we get:

$$\bar{G}_X(t) = \prod_{s\in[0,t]}(1 - d\Lambda_X(s)) = \prod_{s\in[0,t]}(1 - d(\alpha\Lambda_Z(s))),$$

for all $t \geq 0$.

Hence, a natural estimator of $\bar{G}_X(\cdot)$ is the plug-in estimator

$$\begin{aligned} \widehat{\bar{G}}_X(t) &= \prod_{s\in[0,t]}\left(1 - d\left(\hat{\alpha}\hat{\Lambda}_Z(s)\right)\right) \\ &= \prod_{s\in[0,t]}\left(1 - \frac{\hat{\alpha}}{s\int_s^\infty \frac{1}{u}d\hat{F}^b(u)}d\hat{F}_{12}^b(s)\right), \text{ for all } t > 0, \end{aligned} \tag{8}$$

where $\hat{F}_{12}^b(\cdot)$, $\hat{F}^b(\cdot)$ and $\hat{\alpha}$ are estimators of respectively $F_{12}^b(\cdot)$, $F^b(\cdot)$ and α that we will introduce now. Note that the estimator $\hat{\Lambda}_Z(\cdot)$ is obtained by plug-in in equation (7).

Recall that each individual in the sample, selected in the manner of Section 2.2, is followed until death or censoring. Then the observed data consists of n independent pairs (T_i^b, δ_i^b) where $T_i^b = Z_i^b \wedge C_i^b$ and

$$\delta_i^b = \begin{cases} 0 & \text{if} \quad C_i^b < Z_i^b \\ 1 & \text{if} \quad T_i^b = Z_i^b = X_i^b \\ 2 & \text{if} \quad T_i^b = Z_i^b = Y_i^b \end{cases}.$$

The sub-distribution functions $F_0^b(\cdot)$, $F_1^b(\cdot)$ and $F_2^b(\cdot)$ associated with (T^b, δ^b) and defined in Theorem 1 can be estimated from the available sample by

$$\hat{F}_k^b(t) = \frac{1}{n}\sum_{i=1}^n I(\{T_i^b \leq t, \delta_i^b = k\}), \tag{9}$$

for all $t \geq 0$ and $k = 0, 1, 2$.

As the distribution function $F^b(\cdot)$ of the r.v. T^b is equal to $F_1^b(\cdot) + F_2^b(\cdot) + F_3^b(\cdot)$, one can estimate it by

$$\hat{F}^b(\cdot) = \hat{F}_0^b(\cdot) + \hat{F}_1^b(\cdot) + \hat{F}_2^b(\cdot). \tag{10}$$

Also,

$$\hat{F}^b_{12}(\cdot) = \hat{F}^b_1(\cdot) + \hat{F}^b_2(\cdot)$$

gives us an estimator of $F^b_{12}(\cdot)$.

To estimate $\alpha = 1/(1+\beta)$, we first note that

$$\alpha = \frac{\text{pr}\,(X \leq Y, X \leq C)}{\text{pr}\,(X \wedge Y \leq C)} = \frac{\text{pr}\left(X^b \leq Y^b, X^b \leq C^b\right)}{\text{pr}\,(X^b \wedge Y^b \leq C^b)},$$

where the first equality is given by Equation (3) and the second is a straightforward consequence of Thereom 1. By the definition of the sub-distribution functions $F^b_k(\cdot)$, for $k = 0, 1, 2$, the second equality is equivalent to $\alpha = F^b_1(+\infty)/F^b_{12}(+\infty)$. As a consequence α may be estimated by

$$\hat{\alpha} = \frac{\hat{F}^b_1(+\infty)}{\hat{F}^b_{12}(+\infty)}. \tag{11}$$

Estimators given in (9), (10) and (11) complete the definition of the estimator $\hat{\bar{G}}_X(\cdot)$ of $\bar{G}_X(\cdot)$ given in (8).

2.4. **Large sample behaviour.** Our aim in this section is to obtain the weak convergence of the process $\sqrt{n}(\hat{\bar{G}}_X(\cdot) - \bar{G}_X(\cdot))$, as n tends to $+\infty$. One can see from the above section that our estimator $\hat{\bar{G}}_X(\cdot)$ is a function of the estimators $\hat{F}^b_k(\cdot)$, for $k = 0, 1, 2$. Since asymptotic results are available for the estimators of sub-distribution functions, the expected weak convergence result can be obtained by using the functional delta-method van der Vaart & Wellner (1996). But, it has to be noted that one needs a weak convergence on the whole line $[0, +\infty]$ of the empirical processes associated with $\hat{F}^b_k(\cdot)$, for $k = 0, 1, 2$. Such a result is available in Dauxois & Guilloux (2008). Their Theorem 1 is written with two competing risks (for ease of notation) and in presence of independent random right censoring. After an easy adaptation of Dauxois & Guilloux (2008), considering the case with 3 competing risks and without censoring, we get the following weak convergence in $\mathbb{D}^3[0, +\infty]$, where $\mathbb{D}[0, +\infty]$ is the space of càdlàg (right-continuous with left-hand limits) functions. As $n \to +\infty$, one has

$$\sqrt{n}\begin{pmatrix}\hat{F}^b_0(\cdot) - F^b_0(\cdot)\\ \hat{F}^b_1(\cdot) - F^b_1(\cdot)\\ \hat{F}^b_2(\cdot) - F^b_2(\cdot)\end{pmatrix} \xrightarrow{\mathcal{D}} \begin{pmatrix}Z_0(\cdot)\\ Z_1(\cdot)\\ Z_2(\cdot)\end{pmatrix}, \tag{12}$$

where $(Z_0(\cdot), Z_1(\cdot), Z_2(\cdot))'$ is a trivariate mean-zero gaussian process with covariance function

$$\begin{aligned}\langle Z_k(s), Z_k(t)\rangle &= \sum_{l\neq k}\int_0^{s\wedge t}(F_k^b(s)-F_k^b(u))(F_k^b(t)-F_k^b(u))\frac{dF_l^b(u)}{(\bar F^b(u))^2}\\ &+ \int_0^{s\wedge t}(F_k^b(s)-F_k^b(u)-\bar F^b(u))(F_k^b(t)-F_k^b(u)-\bar F^b(u))\frac{dF_k^b(u)}{(\bar F^b(u))^2}\end{aligned}$$

and, for $k\neq l$

$$\begin{aligned}\langle Z_k(s), Z_l(t)\rangle &= \int_0^{s\wedge t}(F_k^b(s)-F_k^b(u))(F_l^b(t)-F_l^b(u))\frac{dF_j^b(u)}{(\bar F^b(u))^2}\\ &+ \int_0^{s\wedge t}(F_l^b(t)-F_l^b(u))(F_k^b(s)-F_k^b(u)-\bar F^b(u))\frac{dF_k^b(u)}{(\bar F^b(u))^2}\\ &+ \int_0^{s\wedge t}(F_k^b(s)-F_k^b(u))(F_l^b(t)-F_l^b(u)-\bar F^b(u))\frac{dF_l^b(u)}{(\bar F^b(u))^2},\end{aligned}$$

where j is different from k and l.

As the first step in the derivation of the large sample behaviour of the process $\sqrt{n}(\hat{\bar G}_X(\cdot)-\bar G_X(\cdot))$, we introduce the following preliminary result, whose proof is given in the Appendix.

Lemma 1. *As n goes to $+\infty$, we have the following weak convergence in $\mathbb{D}[0,\infty]\times\mathbb{R}$:*

$$\sqrt{n}\begin{pmatrix}\hat\Lambda_Z(\cdot)-\Lambda_Z(\cdot)\\ \hat\alpha-\alpha\end{pmatrix}\xrightarrow{\mathcal{D}}\begin{pmatrix}L(\cdot)\\ U\end{pmatrix}, \tag{13}$$

where

$$\begin{aligned}L(\cdot) &= \int_0^{\cdot}\frac{1}{z\int_z^{+\infty}\frac{1}{u}dF^b(u)}dZ_{12}(z)-\int_0^{\cdot}\frac{z\int_z^{+\infty}\frac{1}{u}dZ(u)}{\left(z\int_z^{+\infty}\frac{1}{u}dF^b(u)\right)^2}dF_{12}^b(z),\\ Z(\cdot) &= Z_0(\cdot)+Z_1(\cdot)+Z_2(\cdot),\\ Z_{12}(\cdot) &= Z_1(\cdot)+Z_2(\cdot)\end{aligned}$$

and

$$U=\frac{1}{\mathrm{pr}(\delta^b\neq 0)}[(1-\alpha)Z_1(+\infty)-\alpha Z_2(+\infty)].$$

We are now in a position to give the main result of this section.

Theorem 2. *The following weak convergence holds in the Skorohod space $\mathbb{D}[0,\infty]$:*

$$\sqrt{n}(\hat{\bar G}_X(\cdot)-\bar G_X(\cdot))\xrightarrow{\mathcal{D}}\xi(\cdot)=-\bar G_X(\cdot)\left(U\Lambda_Z(\cdot)+\alpha L(\cdot)\right),$$

as n tends to $+\infty$.

Proof of Theorem 2.
In view of equation (8), we have for all t

$$\sqrt{n}\left(\hat{\bar{G}}_X(t) - \bar{G}_X(t)\right) = \sqrt{n}\left(\Psi_t(\hat{\Lambda}_Z(\cdot), \hat{\alpha}) - \Psi_t(\Lambda_Z(\cdot), \alpha)\right),$$

where $\Psi_\cdot$ is a map from $\mathbb{D}_{BV}[0,\infty] \times \mathbb{R}$ to $\mathbb{D}[0,\infty)$ defined by

$$\Psi_\cdot(f(\cdot), r) = \prod_{t\in[0,\cdot]}(1 - d\,(rf(t))),$$

where $\mathbb{D}_{BV}[0,\infty]$ is the space of càdlàg functions of bounded variations on $[0,\infty]$. Using for example the Chain rule Lemma 3.9.3 of van der Vaart & Wellner (1996) and the Hadamard differentiability of the product integral Andersen et al. (1993), one can see that the map $\Psi_\cdot$ is Hadamard-differentiable with differential $D\Psi_{(f(\cdot),r)}$ at $(f(\cdot), r)$ in $\mathbb{D}_{BV}[0,\infty]\times\mathbb{R}$ given, for all $(h(\cdot), u) \in \mathbb{D}[0,\infty] \times \mathbb{R}$, by

$$D\Psi_{(f(\cdot),r)}(h(\cdot), u) = -\prod_{t\in[0,\cdot]}(1 - d\,(rf(t)))\,(uf(\cdot) + rh(\cdot))\,.$$

An application of the functional delta method (see Theorem 3.9.4. of van der Vaart & Wellner (1996)) on the weak convergence of Lemma 3 gives us

$$\sqrt{n}(\hat{\bar{G}}_X(\cdot) - \bar{G}_X(\cdot)) \xrightarrow{\mathcal{D}} D\Psi_{(\Lambda_Z(\cdot),\alpha)}(L(\cdot), U),$$

as n tends to $+\infty$. From the above expression of the differential, one obtains the limiting process $\xi(\cdot)$ of Theorem 2.$\square$

3. Case 2: No censoring before Length-Biased Sampling

3.1. **New Framework.** We now consider the case of no censoring before length-biased sampling but assume, as before, that the risks A and B have proportional hazards. It can be easily shown that $Z = X \wedge Y$ has pdf

$$g_Z(z) = (1+\beta)g_X(z)(\bar{G}_X(z))^\beta,\ z > 0.$$

Our goal is again to estimate $\bar{G}_X(t) = \text{pr}(X > t)$ on the basis of observations on Z^b, the length length-biased version of Z, and the associated cause of death. An easy adaptation of Theorem 1 (the set E is now $E = \{(\sigma, x, y) : \sigma \leq t_0, \sigma + x \geq t_0, \sigma + y \geq t_0\}$ since there is no censoring at this stage) shows that the length biased observation Z^b of Z has probability density function

$$g_{Z^b}(z) = \frac{z}{E(Z)}(g_X(z)\bar{G}_Y(z) + g_Y(z)\bar{G}_X(z)),$$

for $z > 0$. Now, thanks to the proportional hazard property assumed on the r.v. X and Y, this p.d.f. reduces to

$$g_{Z^b}(z) = \frac{1}{EZ}(1+\beta)zg_X(z)(\bar{G}_X(z))^\beta,\ z > 0$$

and finally one can write

$$g_{Z^b}(z) = \frac{1}{E(Z)} z g_Z(z),\, z > 0. \tag{14}$$

Thus, the p.d.f. of Z^b appears as the length-biased version of the pdf of Z. The corresponding survivor function of Z^b will be denoted by $\bar{G}_{Z^b}(\cdot)$.

It has to be noted that the constant β gives also the odds of death due to cause B after length-biased sampling, i.e.

$$\frac{\text{pr}(Y^b \leq X^b)}{\text{pr}(X^b \leq Y^b)} = \beta.$$

It can also be shown that the random variables $I(\{X^b \leq Y^b\})$ and $Z^b = X^b \wedge Y^b$ are independent. However, as in Section 2, the initial independence between X and Y has been lost under the selection process, i.e. the r.v. X^b and Y^b are not independent.

Now, as we will see, assuming independent right censoring on the length-biased observation Z^b doesn't substantially complicate the following derivation of the estimator of $\bar{G}_X(\cdot)$ and its large sample behaviour study. The situation without any censoring, the one of preliminary interest, will be obtained as a simple particular case of our results under right censoring. This will be detailed at this end of this section.

Each individual in the sample, selected according to the above procedure, is followed until death or censoring. The observed data then consists of n independent pairs (T_i, δ_i) where $T_i = Z_i^b \wedge C_i$ and

$$\delta_i = \begin{cases} 0 & \text{if} \quad C_i < Z_i^b \\ 1 & \text{if} \quad T_i = Z_i^b = X_i^b \\ 2 & \text{if} \quad T_i = Z_i^b = Y_i^b \end{cases}.$$

Here, the r.v. $C_1, \ldots, C_n$ are independent copies of a random variable C which is assumed to be independent of Z^b and with survivor function $\bar{H}_C(\cdot)$. For later use, let $S(\cdot) = \bar{G}_{Z^b}(\cdot)\bar{H}_C(\cdot)$ denotes the survivor function of $T = Z^b \wedge C$.

3.2. **Statistical inference.** From Equation (14) we know that the p.d.f. of Z^b is the length-biased version of the one of Z. Consequently, by the well-known inversion formula of Cox (1969), the distribution function $G_Z(\cdot) = \text{pr}(Z \leq \cdot)$ is expressible, for $t \geq 0$, as

$$G_Z(t) = \frac{\int_0^t \frac{1}{z} d\bar{G}_{Z^b}(z)}{\int_0^{+\infty} \frac{1}{z} d\bar{G}_{Z^b}(z)}.$$

On the other hand we can write

$$\bar{G}_X(\cdot) = \left(\bar{G}_Z(\cdot)\right)^{\alpha},$$

where $\alpha = 1/(1+\beta)$. Hence, a natural estimator of $\bar{G}_X(\cdot)$ is the plug-in estimator

$$\tilde{\bar{G}}_X(t) = \left(1 - \hat{G}_Z(t)\right)^{\tilde{\alpha}}, \text{ for all } t > 0, \tag{15}$$

where

$$\hat{G}_Z(t) = \frac{\int_0^t \frac{1}{z} d\hat{\bar{G}}_{Z^b}(z)}{\int_0^{+\infty} \frac{1}{z} d\hat{\bar{G}}_{Z^b}(z)}, t > 0, \tag{16}$$

and $\hat{\bar{G}}_{Z^b}(\cdot)$ and $\hat{\alpha}$ are estimators to be introduced below.

Let

$$N_j(t) = \sum_{i=1}^n I(\{T_i \le t, \delta_i = j\}), \text{ for } j = 1, 2,$$

for $t \ge 0$, be the counting process associated with the jth cause of death and let

$$Y(t) = \sum_{i=1}^n I(\{T_i \ge t\}),$$

for $t \ge 0$, be the at-risk process. Moreover let $J(\cdot)$ and $N(\cdot)$ be two processes defined respectively by $J(t) = I(\{Y(t) > 0\})$ and

$$N(t) = \sum_{i=1}^n I(\{T_i \le t, \delta_i \ne 0\}) = N_1(t) + N_2(t),$$

for all $t \ge 0$.

The survivor function $\bar{G}_{Z^b}(\cdot)$ can be estimated by the Kaplan-Meier estimator (cf e.g. Andersen et al. (1993))

$$\widehat{\bar{G}}_{Z^b}(t) = \prod_{i:T_{(i)} \le t} \left(1 - \frac{\Delta N(T_{(i)})}{Y(T_{(i)})}\right), \tag{17}$$

where $T_{(1)} \le \cdots \le T_{(n)}$ are the ordered statistics and $\Delta N(u) = N(u) - N(u^-)$, for all $u \ge 0$. The estimator $\hat{G}_Z(\cdot)$ given in (16) is now completly defined.

In order to introduce our estimator of α, let

$$G_1(t) \;\; = \;\; \text{pr}(X^b \le t, X^b \le Y^b) = \text{pr}(Z^b \le t, X^b \le Y^b).$$

be the cumulative incidence function associated to cause A. Then,

$$\alpha = \text{pr}\left(X^b \le Y^b\right) = G_1(+\infty).$$

Estimating $G_1(\cdot)$ by the Aalen-Johansen estimator Andersen et al. (1993)

$$\hat{G}_1(t) \;\; = \;\; \int_0^t \hat{\bar{G}}_{Z^b}(x^-) \frac{dN_1(x)}{Y(x)},$$

an estimator of α is given by

$$\tilde{\alpha} = \hat{G}_1(+\infty). \tag{18}$$

Thus, the estimator $\tilde{\bar{G}}_X(\cdot)$ given in (15) is completly defined thanks to (16), (17) and (18).

3.3. **Large sample behaviour.** In order to get the weak convergence of the process $\sqrt{n}(\tilde{\bar{G}}_X(\cdot) - \bar{G}_X(\cdot))$ stated in the next theorem, the following assumption is needed:

$$\textbf{Assumption } \mathcal{A} : \int_0^{+\infty} \frac{dG_{Z^b}(x)}{\bar{H}_C(x-)} < +\infty.$$

Theorem 3. *If assumption $\mathcal{A}$ is fulfilled, the following weak convergence holds in the Skorohod space $\mathbb{D}[0,\infty]$:*

$$\sqrt{n}(\tilde{\bar{G}}_X(\cdot) - \bar{G}_X(\cdot)) \xrightarrow{\mathcal{D}} \xi(\cdot) = \alpha\bar{G}_Z(\cdot)\tilde{L}(\cdot) + \tilde{U}\bar{G}_Z(\cdot)\ln(\bar{G}_Z(\cdot)),$$

as n goes to ∞, where $\tilde{L}(\cdot)$ is a mean zero gaussian process defined by

$$\tilde{L}(\cdot) = G_{Z^b}(\cdot)\frac{\int_0^{+\infty}\frac{1}{x}d\tilde{Z}(x)}{\int_0^{+\infty}\frac{1}{x}d\bar{G}_{Z^b}(x)} - \frac{\int_0^{\cdot}\frac{1}{x}d\tilde{Z}(x)}{\int_0^{+\infty}\frac{1}{x}d\bar{G}_{Z^b}(x)},$$

$\tilde{Z}(\cdot)$ is a mean zero gaussian process defined on $[0,+\infty]$ with covariance function given by

$$< \tilde{Z}(s), \tilde{Z}(t) >= \bar{G}_{Z^b}(s)\bar{G}_{Z^b}(t)\int_0^{s\wedge t}\frac{dG_{Z^b}(x)}{\bar{G}_{Z^b}(x)S(x^-)}$$

and $\tilde{U}$ is a mean-zero normally distributed r.v. with variance given by

$$\begin{aligned}
\mathrm{v}(\tilde{U}) &= \int_0^{+\infty}(G_1(+\infty) - G_1(x))^2\frac{d\bar{G}_{Z^b}(x)}{\bar{G}_{Z^b}(x)S(x)} + \int_0^{+\infty}\bar{G}^2_{Z^b}(x)\frac{d\bar{G}_{Z^b}(x)}{\bar{G}_1(x)S(x)}\\
&- 2\int_0^{+\infty}(G_1(+\infty) - G_1(x))\bar{G}_{Z^b}(x)\frac{d\bar{G}_{Z^b}(x)}{\bar{G}_1(x)S(x)}.
\end{aligned}$$

The proof of the above theorem requires the following key result proved in appendix.

Lemma 2. *Under Assumption $\mathcal{A}$, the following weak convergence holds in $\mathbb{D}[0,\infty]\times\mathbb{R}$*

$$\sqrt{n}\begin{pmatrix}\widehat{G}_Z(\cdot) - G_Z(\cdot)\\ \tilde{\alpha} - \alpha\end{pmatrix} \xrightarrow{\mathcal{D}} \begin{pmatrix}\tilde{L}(\cdot)\\ \tilde{U}\end{pmatrix}, \tag{19}$$

as n goes to ∞.

Proof of Theorem 2.
In view of equation (15), we have

$$\sqrt{n}\left(\tilde{\bar{G}}_X(t) - \bar{G}_X(t)\right) \;=\; \sqrt{n}\left(\Phi_t(\widehat{G}_Z(\cdot),\tilde{\alpha}) - \Phi_t(G_Z(\cdot),\alpha)\right).$$

where $\Phi(\cdot,\cdot)$ is a map from $\mathbb{D}[0,\infty]\times\mathbb{R}$ to $[0,\infty)$ defined by $\Phi_t(f(\cdot),r) = (1-f(t))^r$. A two-terms Taylor expansion of the map $(x,y)\mapsto h(x,y) = (1-x)^y$ assures that $\Phi(\cdot,\cdot)$ is Hadamard-differentiable with differential $D\Phi_{(f(\cdot),r)}$ at $(f(\cdot),r)$ defined, for all $(h(\cdot),u)$ in $\mathbb{D}[0,\infty]\times\mathbb{R}$, by

$$D\Phi_{(f(\cdot),r)}(h(\cdot),u) = u(1-f(t))^r\ln(1-f(t)) - r(1-f(t))^{r-1}h(t).$$

The functional delta method in its version of Theorem 3.9.4. of van der Vaart & Wellner (1996) applies and gives the result of Theorem 2.□

We now come back to the "without censoring" case which was of original interest. In this case, the observations are given by Z^b and δ^b where the latter is now defined by

$$\delta = \begin{cases} 1 & \text{if} \quad Z^b = X^b \\ 2 & \text{if} \quad Z^b = Y^b \end{cases}.$$

As an estimator of $\bar{G}_X(\cdot)$ one can still use

$$\tilde{\bar{G}}_X(t) = \left(1 - \hat{G}_Z(t)\right)^{\tilde{\alpha}},\ \text{for all } t>0,$$

where

$$\hat{G}_Z(t) = \frac{\int_0^t \frac{1}{z} d\hat{\bar{G}}_{Z^b}(z)}{\int_0^{+\infty} \frac{1}{z} d\hat{\bar{G}}_{Z^b}(z)}, t>0.$$

But, in the absence of censoring, the Kaplan-Meier estimator $\hat{\bar{G}}_{Z^b}(\cdot)$ is nothing but the empirical survivor function defined by :

$$\hat{\bar{G}}_{Z^b}(t) = \frac{1}{n}\sum_{i=1}^{n} I(Z_i^b > t).$$

The statistic $\hat{G}_1(+\infty)$ still gives us an estimate of α and, now, we have the simplified expression:

$$\tilde{\alpha} = \frac{N_1(+\infty)}{n},$$

which is the observed proportion of death due to cause A.

The following corollary derives the asymptotic behaviour of our estimator in the "without censoring" case. It is easily obtained from Theorem 3 on noting that $S(\cdot)$ is now equal to $\bar{G}_{Z^b}(\cdot)$.

Corollary 1. *The following weak convergence holds in the Skorohod space* $\mathbb{D}[0,\infty]$*:*

$$\sqrt{n}(\tilde{\bar{G}}_X(\cdot) - \bar{G}_X(\cdot)) \xrightarrow{\mathcal{D}} \xi(\cdot) = \alpha \bar{G}_Z(\cdot)\tilde{L}(\cdot) + \tilde{U}\bar{G}_Z(\cdot)\ln(\bar{G}_Z(\cdot)),$$

as n *goes to* ∞*, where* $\tilde{L}(\cdot)$ *is a mean zero gaussian process defined by*

$$\tilde{L}(\cdot) = G_{Z^b}(\cdot)\frac{\int_0^{+\infty}\frac{1}{x}d\tilde{Z}(x)}{\int_0^{+\infty}\frac{1}{x}d\bar{G}_{Z^b}(x)} - \frac{\int_0^{\cdot}\frac{1}{x}d\tilde{Z}(x)}{\int_0^{+\infty}\frac{1}{x}d\bar{G}_{Z^b}(x)},$$

$\tilde{Z}(\cdot)$ *is a mean zero gaussian process defined on* $[0,+\infty]$ *with covariance function given by*

$$< \tilde{Z}(s), \tilde{Z}(t) >= \bar{G}_{Z^b}(\max(s,t)) - \bar{G}_{Z^b}(s)\bar{G}_{Z^b}(t)$$

and $\tilde{U}$ *is a real r.v. with distribution* $N(0,\alpha(1-\alpha))$.

4. Illustrative Example

The statistical analysis of the proportional hazards competing risks model developed here under the length-biased sampling scheme is of wide ranging interest. Its applicability extends well beyond the epidemiologic studies involving follow up of prevalent cases identified through a cross-sectional study. Here, we present an application to a well-known problem in political science. In those parts of the world where democratic institutions and constitutional practices are firmly entrenched, change of government frequently occurs through non-constitutional means (such as coups). In such situations, it is of interest to be able to estimate and predict the duration for which political and executive leaders hold power. The question is of more than academic interest as the length of a leader's stay in power may affect economic and human right issues. Bienen & M'Lan (1991) is a pioneering study of the time of power for primary leaders of countries world-wide. They provide, analyze, and interpret data on duration (in years) in power for 2,256 leaders from 167 countries for a 100 years period terminating in 1987. However, we are interested only in a subset of the original data, confined to countries outside of Europe, North America, and Australia; and restricted to leaders who were in power in 1972. There were 99 such leaders facing two competing risks: exit by constitutional means (risk A) and non-constitutional means (risk B). We treat other termination modes as censoring. Bienen and van de Walle's data is rich in covariates. Allison (1995) gives an analysis of covariates effects via Cox models for a subset consisting of 472 spells of time in power beginning in 1960 or later. Although our analysis is not concerned with covariates and, unlike Allison (1995), we are estimating in the length-biased set up; we note from Allison (1995), that the two risks - constitutional and non-constitutional exits - have proportional hazards. This proportionality is indicated by Fig. 2 which provides the plots of log-log survivor functions for the two risks against time. Notice that the log-log survivor functions of Fig. 2 have been estimated from the initial sample (the sample with 472 spells used by Allison (1995)). Figure 3 shows the survivor function corresponding to risk B when estimated from the initial and length-biased samples. Our estimator, although based on a length-biased sample selected from the initial sample, performs quite well as compared to the Kaplan-Meier estimator computed from the whole initial sample. The length-biased sample is only about 20% of the initial sample of 472 spells.

5. Simulation Study

We carried out Monte Carlo simulations to compare our estimator in Case 1 (independent censoring preceding length-biased sampling) with the true survivor function when the two independent competing risks are Weibull distributed. More precisely, we consider two scenarios:

- Scenario 1: $\bar{G}_X(t) = \exp(-t^{1.5})$ and $\bar{G}_Y(t) = \exp(-0.6t^{1.5})$, for all $t > 0$
- Scenario 2: $\bar{G}_X(t) = \exp(-0.6t^{1.5})$ and $\bar{G}_Y(t) = \exp(-t^{1.5})$, for all $t > 0$

In the first scenario, the lifetime of interest X is stochastically smaller than Y, while in the second scenario the reverse is true.

We generated a population of n_I individuals whose birth times follow a homogeneous Poisson process with intensity $\lambda = 1$ on the interval $(-10, 1)$. For each individual in this initial population, a censoring time was simulated according to an exponential distribution with parameter μ. The length-biased sample consists of the individuals alive, but not censored, at time $t_0 = 0.5$. We chose the values of the parameters n_I and μ in order to get censoring levels of approximatively 5%, 10% or 30% and sample sizes of $n = 100$ or $n = 1000$ for the length-biased data.

The simulation design described above was replicated 1000 times. Tables 1 and 2 give the resulting Monte Carlo estimates of the classical mean integrated squared error (MISE) and a scaled mean integrated squared error (SMISE). The MISE is defined as

$$\text{MISE}(\hat{\bar{G}}_X) = \int_0^\infty E\left[\left(\hat{\bar{G}}_X(t) - \bar{G}_X(t)\right)^2 I(t \leq T^b_{(n-1)})\right] dt,$$

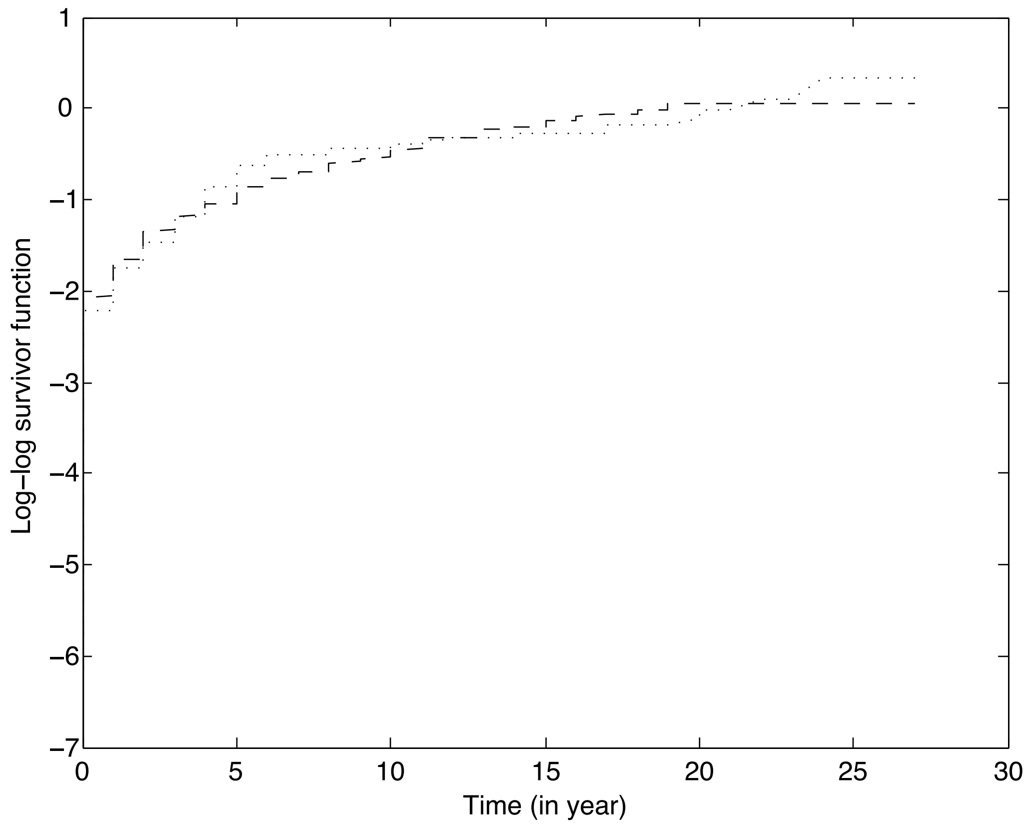

FIGURE 2. Estimates of the Log-log survivor functions for cause 1 (dotted line) and cause 2 (dot-dashed line), using the initial sample.

where $T^b_{(n-1)}$ is the $(n-1)$-th ordered statistic in our sample, while SMISE is defined as

$$\text{SMISE}(\hat{\bar{G}}_X) = \int_0^\infty E\left[\frac{1}{T^b_{(n-1)}}\left(\hat{\bar{G}}_X(t) - \bar{G}_X(t)\right)^2 I(t \leq T^b_{(n-1)})\right] dt.$$

The number of grid points to approximate the integrals is set as 1000.

One can see in these tables that both the MISEs and SMISEs always decrease with the number of observations. This illustrates the consistency of our estimators.

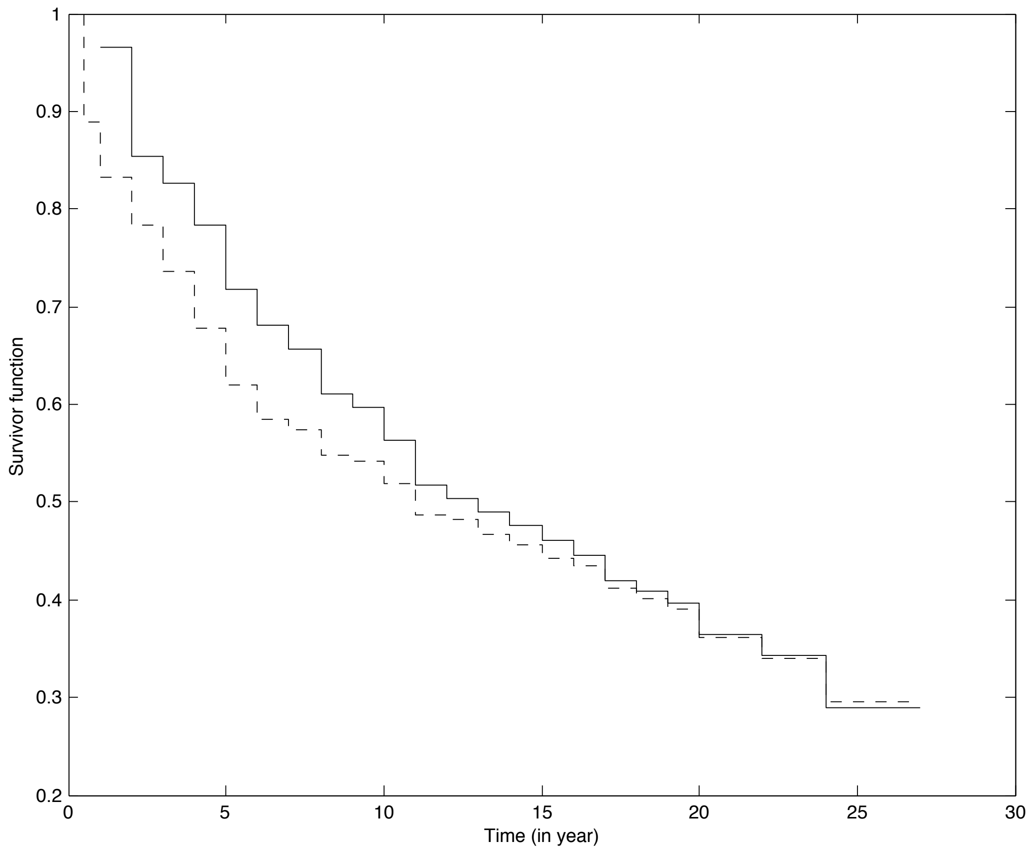

FIGURE 3. Estimates of the survivor function for cause 2: Kaplan-Meier estimator (hyphenated line) based on the whole initial population and Case 1 estimator (solid line) based on the 1972 cross-sectional sample from the initial sample.

In Table 1 and 2, one can see that the MISE does not necessarily increase when the proportion of censoring increases. This can be explained by the fact that the interval $[0, T^b_{(n-1)}]$ on which the MISE is calcuted decreases when the censoring increases, see e.g. Geffray & Guilloux (2011) for details. On the other hand, the SMISE, which is normalized with respect to the length of the interval $[0, T^b_{(n-1)}]$, has the expected behavior: it increases as the censoring increases.

		Censoring					
		5%		15%		30%	
n	100	4·6	(3·2)	5·7	(3·84)	5·3	(4·4)
	1000	1·0	(0·5)	1·0	(0·6)	1·2	(0·7)

TABLE 1. Simulation results under Scenario 1. Monte Carlo estimates of *MISE**10^3 and *SMISE**10^3 (in parentheses) for Case 1 estimator

		Censoring					
		5%		15%		30%	
n	100	3·0	(2·1)	3·3	(2·6)	3·2	(2·7)
	1000	0·9	(0·5)	0·9	(0·6)	0·8	(0·6)

TABLE 2. Simulation results under Scenario 2. Monte Carlo estimates of *MISE**10^3 and *SMISE**10^3 (in parentheses) for Case 1 estimator

REFERENCES

ALLISON, P. (1995). *Survival analysis using the SAS system: a pratical guide*. SAS Institute.

ANDERSEN, P. K., BORGAN, Ø., GILL, R. D. & KEIDING, N. (1993). *Statistical models based on counting processes*. Springer Series in Statistics. New York: Springer-Verlag.

ASGHARIAN, M., M'LAN, C. E. & WOLFSON, D. B. (2002). Length-biased sampling with right censoring: an unconditional approach. *J. Amer. Statist. Assoc.* **97**, 201–209.

BIENEN, H. & M'LAN, VAN DE WALLE, N. (1991). *Time of power*. Stanford University Press.

BLUMENTHAL, S. (1967). Limit theorems for functions of shortest two-sample spacings and a related test. *Ann. Math. Statist.* **38**, 108–116.

BRILLINGER, D. R. (1986). The natural variability of vital rates and associated statistics. *Biometrics* **42**, 693–734. With discussion and a reply by the author.

CHEN, P. & LIN, G. (1987). Maximum likelihood estimation of a survival function under the koziol-green proportional hazards model. *Statistics and Probability Letters* **5**, 75–80.

COX, D. (1969). *New development in survey sampling*. Johnson and Smith, Wiley.

CRUMP, K. S. (1975). On point processes having an order statistic structure. *Sankhyā Ser. A* **37**, 396–404.

CSÖRGŐ, S. (1988). Estimation in the proportional hazards model of random censorship. *Statistics* **19**, 437–463.

DAUXOIS, J.-Y. & GUILLOUX, A. (2008). Nonparametric inference under competing risks and selection-biased sampling. *J. Multivariate Anal.* **99**, 589–605.

DE UÑA-ÁLVAREZ, J. (2004). Nelson-Aalen and product-limit estimation in selection bias models for censored populations. *J. Nonparametr. Stat.* **16**, 761–777.

GATHER, U. & PAWLITSCHKO, J. (1998). Estimating the survival function under a generalized Koziol-Green model with partially informative censoring. *Metrika* **48**, 189–207 (1999).

GEFFRAY, S. & GUILLOUX, A. (2011). Maximum likelihood estimator for cumulative incidence functions under proportionality constraint. *To appear in Sankhya Series A* .

GILL, R. D., VARDI, Y. & WELLNER, J. A. (1988). Large sample theory of empirical distributions in biased sampling models. *Ann. Statist.* **16**, 1069–1112.

HAYAKAWA, Y. (2000). A new characterisation property of mixed Poisson processes via Berman's theorem. *J. Appl. Probab.* **37**, 261–268.

HUANG, Y. & WANG, M.-C. (1995). Estimating the occurrence rate for prevalent survival data in competing risks models. *J. Amer. Statist. Assoc.* **90**, 1406–1415.

KEIDING, N. (1990). Statistical inference in the Lexis diagram. *Philos. Trans. Roy. Soc. London Ser. A* **332**, 487–509.

KINGMAN, J. F. C. (1993). *Poisson processes*, vol. 3 of *Oxford Studies in Probability*. New York: The Clarendon Press Oxford University Press. Oxford Science Publications.

KIRMANI, S. N. U. A. & DAUXOIS, J.-Y. (2004). Testing the Koziol-Green model against monotone conditional odds for censoring. *Statist. Probab. Lett.* **66**, 327–334.

LUND, J. (2000). Sampling bias in population studies—how to use the Lexis diagram. *Scand. J. Statist.* **27**, 589–604.

MCFADDEN, J. A. (1962). On the lengths of intervals in a stationary point process. *J. Roy. Statist. Soc. Ser. B* **24**, 364–382.

SIMON, R. (1980). Length biased sampling in etiologic studies. *American Journal of Epidemiology* **111**, 444–452.

TSAI, W. Y. (2009). Pseudo-partial likelihood for proportional hazards models with biased-sampling data. *Biometrika* **96**, 601–615.

TSAI, W.-Y., JEWELL, N. P. & WANG, M.-C. (1987). A note on the product-limit estimator under right censoring and left truncation. *Biometrika* **74**, pp. 883–886.

TURNBULL, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *J. Roy. Statist. Soc. Ser. B* **38**, 290–295.

VAN DER VAART, A. W. & WELLNER, J. A. (1996). *Weak convergence and empirical processes*. Springer Series in Statistics. New York: Springer-Verlag. With applications to statistics.

VAN ES, B., KLAASSEN, C. A. J. & OUDSHOORN, K. (2000). Survival analysis under cross-sectional sampling: length bias and multiplicative censoring. *J. Statist. Plann. Inference* **91**, 295–312. Prague Workshop on Perspectives in Modern Statistical Inference: Parametrics, Semi-parametrics, Non-parametrics (1998).

VARDI, Y. (1982). Nonparametric estimation in the presence of length bias. *Ann. Statist.* **10**, 616–620.

VARDI, Y. (1985). Empirical distributions in selection bias models. *Ann. Statist.* **13**, 178–205. With discussion by C. L. Mallows.

VARDI, Y. (1989). Multiplicative censoring, renewal processes, deconvolution and decreasing density: nonparametric estimation. *Biometrika* **76**, 751–761.

VARDI, Y. & ZHANG, C.-H. (1992). Large sample study of empirical distributions in a random-multiplicative censoring model. *Ann. Statist.* **20**, 1022–1039.

WANG, M.-C. (1991). Nonparametric estimation from cross-sectional survival data. *J. Amer. Statist. Assoc.* **86**, 130–143.

WANG, M.-C., BROOKMEYER, R. & JEWELL, N. P. (1993). Statistical models for prevalent cohort data. *Biometrics* **49**, 1–11.

WANG, M.-C., JEWELL, N. P. & TSAI, W.-Y. (1986). Asymptotic properties of the product limit estimate under random truncation. *Ann. Statist.* **14**, 1597–1605.

WOODROOFE, M. (1985). Estimating a distribution function with truncated data. *Ann. Statist.* **13**, 163–177.

ZELEN, M. & FEINLEIB, M. (1969). On the theory of screening for chronic diseases. *Biometrika* **56**, 601–614.

APPENDIX 1: PROOF OF THE LEMMAS

Lemma 3. *As n goes to $+\infty$, we have the following weak convergence in $\mathbb{D}[0,\infty] \times \mathbb{R}$:*

$$\sqrt{n}\begin{pmatrix} \hat{\Lambda}_Z(\cdot) - \Lambda_Z(\cdot) \\ \hat{\alpha} - \alpha \end{pmatrix} \xrightarrow{\mathcal{D}} \begin{pmatrix} L(\cdot) \\ U \end{pmatrix},$$

where

$$
\begin{aligned}
L(\cdot) &= \int_0^{\cdot} \frac{1}{z\int_z^{+\infty}\frac{1}{u}dF^b(u)}dZ_{12}(z) - \int_0^{\cdot} \frac{z\int_z^{+\infty}\frac{1}{u}dZ(u)}{\left(z\int_z^{+\infty}\frac{1}{u}dF^b(u)\right)^2}dF_{12}^b(z),\\
Z(\cdot) &= Z_0(\cdot)+Z_1(\cdot)+Z_2(\cdot),\\
Z_{12}(\cdot) &= Z_1(\cdot)+Z_2(\cdot)
\end{aligned}
$$

and

$$U = \frac{1}{\mathrm{pr}(\delta^b \neq 0)}[(1-\alpha)Z_1(+\infty) - \alpha Z_2(+\infty)].$$

Proof of Lemma 1. From the expression of $\hat{\Lambda}_Z(\cdot)$, $\Lambda_Z(\cdot)$, $\hat{\alpha}$ and α given in Section 2.3, one can write:

$$\sqrt{n}\begin{pmatrix}\hat{\Lambda}_Z(\cdot)-\Lambda_Z(\cdot)\\ \hat{\alpha}-\alpha\end{pmatrix} = \sqrt{n}\left(\psi\begin{pmatrix}\hat{F}_0^b(\cdot)\\ \hat{F}_1^b(\cdot)\\ \hat{F}_2^b(\cdot)\end{pmatrix} - \psi\begin{pmatrix}F_0^b(\cdot)\\ F_1^b(\cdot)\\ F_2^b(\cdot)\end{pmatrix}\right),$$

where ψ is the function defined on $\mathbb{D}_{BV}^3[0,\infty]$ to $\mathbb{D}[0,\infty]\times\mathbb{R}$ by

$$\psi\begin{pmatrix}f_0(\cdot)\\ f_1(\cdot)\\ f_2(\cdot)\end{pmatrix} = \begin{pmatrix}\int_0^{\cdot}\frac{1}{z\int_z^{+\infty}\frac{1}{u}d(f_0(u)+f_1(u)+f_2(u))}d(f_1(z)+f_2(z))\\ \frac{f_1(+\infty)}{f_1(+\infty)+f_2(+\infty)}\end{pmatrix}.$$

Let us denote by ψ^1 and ψ^2 respectively the first and second coordinate of the function ψ. Rather straightforward arguments of differential calculus give us that the differential $D\psi^1_{(f_0(\cdot),f_1(\cdot),f_2(\cdot))}$ of ψ^1 at $(f_0(\cdot),f_1(\cdot),f_2(\cdot))\in\mathbb{D}_{BV}^3[0,\infty]$ is, for all $(g_0(\cdot),g_1(\cdot),g_2(\cdot))$:

$$
\begin{aligned}
&D\psi^1_{(f_0(\cdot),f_1(\cdot),f_2(\cdot))}(g_0(\cdot),g_1(\cdot),g_2(\cdot))\\
&= \int_0^{\cdot}\frac{1}{z\int_z^{+\infty}\frac{1}{u}d(f_0(u)+f_1(u)+f_2(u))}d(g_1(z)+g_2(z))\\
&- \int_0^{\cdot}\frac{z\int_z^{+\infty}\frac{1}{u}d(g_0(u)+g_1(u)+g_2(u))}{\left(z\int_z^{+\infty}\frac{1}{u}d(f_0(u)+f_1(u)+f_2(u))\right)^2}d(f_1(z)+f_2(z)).
\end{aligned}
$$

On the other hand, the differential of ψ^2 is

$$
\begin{aligned}
D\psi^2_{(f_0(\cdot),f_1(\cdot),f_2(\cdot))}(g_0(\cdot),g_1(\cdot),g_2(\cdot)) &= \frac{f_2(+\infty)}{(f_1(+\infty)+f_2(+\infty))^2}g_1(+\infty)\\
&- \frac{f_1(+\infty)}{(f_1(+\infty)+f_2(+\infty))^2}g_2(+\infty).
\end{aligned}
$$

We are thus in a position to apply the functional delta-method on the weak convergence (12) of the sub-distribution empirical processes. This gives us the expected weak convergence of Lemma 1. The expression of the limiting process is easily obtained if one note that we have:

$$F^b(\cdot) = F_0^b(\cdot) + F_1^b(\cdot) + F_2^b(\cdot),\ F_{12}^b(\cdot) = F_1^b(\cdot) + F_2^b(\cdot),$$
$$F_{12}^b(+\infty) = \mathrm{pr}(\delta^b \neq 0)$$

and $\alpha = F_1^b(+\infty)/F_{12}^b(+\infty)$. □

Lemma 4. *Under assumption $\mathcal{A}$, as n goes to ∞, the following weak convergence holds in $\mathbb{D}[0,\infty] \times \quad R$*

$$\sqrt{n}\begin{pmatrix} \hat{G}_Z(\cdot) - G_Z(\cdot) \\ \tilde{\alpha} - \alpha \end{pmatrix} \xrightarrow{\mathcal{D}} \begin{pmatrix} \tilde{L}(\cdot) \\ \tilde{U} \end{pmatrix}.$$

Proof of Lemma 2. From Theorem 3 of Dauxois & Guilloux (2008), we have, under Assumption $\mathcal{A}$,

$$\sqrt{n}\begin{pmatrix} \hat{\bar{G}}_{Z^b}(\cdot) - \bar{G}_{Z^b}(\cdot) \\ \hat{G}_1(\cdot) - G_1(\cdot) \end{pmatrix} \xrightarrow{\mathcal{D}} \begin{pmatrix} \tilde{Z}(\cdot) \\ \tilde{Z}_1(\cdot) \end{pmatrix}$$

in $\mathbb{D}^2[0,\infty]$, where $\tilde{Z}(\cdot)$ is defined in Theorem 3 and $\tilde{Z}_1$ is a mean-zero gaussian process defined on $[0,\infty]$ with covariance function given by

$$\begin{aligned} & < \tilde{Z}_1(s), \tilde{Z}_1(t) > \\ = \; & \int_0^{s\wedge t} (G_1(t) - G_1(u))^2 \frac{dG_{Z^b}(u)}{\bar{G}_{Z^b}(u)S(u-)} + \int_0^{s\wedge t} \bar{G}_{Z^b}^2(u)\frac{dG_{Z^b}(u)}{\bar{G}_1(u)S(u-)} \\ - \; & \int_0^{s\wedge t} (G_j(t) - G_j(u))\bar{G}_{Z^b}(u)\frac{dG_{Z^b}(u)}{\bar{G}_1(u)S(u-)}. \end{aligned}$$

It is easily seen that

$$\sqrt{n}\begin{pmatrix} \widehat{G}_Z(\cdot) - G_Z(\cdot) \\ \tilde{\alpha} - \alpha \end{pmatrix} = \sqrt{n}\begin{pmatrix} \Phi(\hat{\bar{G}}_{Z^b}(\cdot)) - \Phi(\bar{G}_{Z^b}(\cdot)) \\ \hat{G}_1(\infty) - G_1(\infty) \end{pmatrix},$$

where Φ is a map from $\mathbb{D}[0,\infty]$ to $\mathbb{D}[0,\infty]$ defined by

$$\Phi(f(\cdot)) = \frac{\int_0^{\cdot} \frac{1}{z} df(z)}{\int_0^{\infty} \frac{1}{z} df(z)}.$$

The map Φ is Hadamard-differentiable with differential $D\Phi_{\bar{G}_{Z^b}(\cdot)}$ at $\bar{G}_{Z^b}(\cdot)$ defined, for all $h(\cdot)$ in $\mathbb{D}[0,\infty]$, by

$$D\Phi_{\bar{G}_{Z^b}(\cdot)} h(\cdot) = \frac{\int_0^{\cdot} \frac{1}{z} dh(z)}{\int_0^{+\infty} \frac{1}{z} d\bar{G}_{Z^b}(z)} - G_{Z^b}(\cdot)\frac{\int_0^{+\infty} \frac{1}{z} dh(z)}{\int_0^{+\infty} \frac{1}{z} d\bar{G}_{Z^b}(z)}.$$

The functional delta method in its version of Theorem 3.9.4. of van der Vaart & Wellner (1996) ends the proof of this lemma.□