INSTITUT NATIONAL DE LA STATISTIQUE ET DES ETUDES ECONOMIQUES
Série des Documents de Travail du CREST
(Centre de Recherche en Economie et Statistique)

n° 2004-02

Estimation in a Competing Risks Proportional Hazards Model under Length-biased Sampling with Censoring

J.-Y. DAUXOIS[1]
A. GUILLOUX[2]
S. N.U.A KIRMANI[3]

[1] CREST-ENSAI, Campus de Ker Lann, BP 37203, 35172 BRUZ Cédex, France. Mail : jean-yves.dauxois@ensai.fr
[2] CREST-ENSAI, Campus de Ker Lann, BP 37203, 35172 BRUZ Cédex, France. Mail : agathe.guilloux@ensai.fr
[3] Department of Mathematics, University of Northern Iowa, Cedar Falls, Iowa, 50614-0506, USA. Mail : kirmani@math.uni.edu

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

Jean-Yves Dauxois[1]
CREST-ENSAI, Rennes, France
Agathe Guilloux[1]
CREST-ENSAI, Rennes, France
Syed N.U.A. Kirmani[2]
University of Northern Iowa, Cedar Falls, Iowa

[1]CREST-ENSAI,
Campus de Ker Lann, BP 37203,
35 172 Bruz Cedex, FRANCE.
jean-yves.dauxois@ensai.fr, agathe.guilloux@ensai.fr

[2]Department of Mathematics, University of Northern Iowa,
Cedar Falls, Iowa 50614-0506
kirmani@math.uni.edu

Résumé :

Nous nous intéressons à des durées de vie soumises à deux risques concurrents, notés A et B. La durée de vie d'un individu qui serait uniquement soumis à la cause A (resp. B) est notée X (resp. Y). Dans notre situation, on observe donc uniquement la durée de vie $T = \inf(X, Y)$, qui est supposée n'être due qu'à une seule cause, et également l'indicateur δ qui spécifie la cause ayant entraîné le décès ou la défaillance. On suppose de plus que les observations que l'on fait de ces durées de vie sous risques concurrents sont soumises à biais de longueur. C'est à dire que l'on constitue l'échantillon des durées de vies observées en ne suivant que les individus en vie à une date t_0 donnée. Enfin, on autorise la présence d'une censure aléatoire droite.

L'objet de cet article est de proposer un estimateur de la fonction de survie de la cause A, i.e. la fonction définie, pour tout $t > 0$, par $\bar{G}_X(t) = P(X > t)$, dans le cas où les durées de vie X et Y sont à fonctions de risque proportionnelles. Nous établissons ensuite des résultats de convergence faible pour le processus associé.

Mots clés : Risques concurrents, risques proportionnels, échantillonnage biaisé en longueur, censure à droite, convergence faible des processus, topologie de Skorohod.

Abstract :

Consider a population of individuals who experience two causes of death. We observe the ones alive at time t_0 and follow them until death or possible censoring time. Given this length biased sample, we introduce an estimator of the survival function of "initial survival times" (i.e. for the entire population) under the assumption of proportional hazards for the two causes of death. The large sample behavior of our estimator is also studied.

Key words : Competing risks, Length-bias, Right-censored data, Proportional hazards, weak convergence of stochastic processes, Skorohod space.

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 experienced the initiating event, but not the terminating event, prior to t_0. In the context of epidemiology and survival analysis, cross-sectional sample studies prevalent rather than incident cases. As 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. Feinlieb and Zelen (1969) recognized LBS in screening for chronic diseases and 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; thus leading 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, 1985, 1989), Gill et. al. (1988) and Vardi and 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 random 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 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 and competing risks (CR). 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. Then the time gap between initiating and terminating events is of length $Z = X \wedge Y$ where X (Y) is the latent (potential) survival time associated with risk A (B). However, under LBS, Z 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 a certain time τ and each has the following four possibilities : (i) dies of cause A, (ii) dies of cause B (a natural cause or all causes other than A), (iii) withdraws from the study, and (iv) is alive and still under study at time τ. The possibilities (iii) and (iv) will be referred to as censoring and C_i will denote the potential censoring time for the ith member of the length-biased sample. Similarly, X_i^b (Y_i^b) will denote the potential survival time of the ith subject when facing risk A (B). The sample data thus consists of the n pairs (T_i, δ_i^b) where $T_i = Z_i^b \wedge C^i$ and δ_i^b indicates the mode of termination (death due to A, death due to B, censoring).

The main objective of the present paper is estimation of the survivor function $\bar{G}_X(t) = \mathbb{P}(X > t)$ in the LBS-CR set up described above. This problem has not been considered in the literature so far. Huang and 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 $\bar{G}_X$. We study the estimation of $\bar{G}_X$ when X and Y are independent but have proportional hazards.

The outline of this paper is as follows. Assuming that the initiating events follow a mixed Poisson process, the distribution of the time to the terminating event in a LBS-CR setup is derived in Section 2. An estimator $\hat{\bar{G}}_X$ for the survivor function $\bar{G}_X(t) = \mathbb{P}(X > t)$, the survivor function of primary interest, is developed in Section 3 on the basis of the observed length-biased sample under right censoring when the two independent risks A and B have proportional hazards. The main result of this paper is the weak convergence of the process $\sqrt{n}(\hat{\bar{G}}_X - \bar{G}_X)$ in the Skorohod space $\mathbb{D}[0, \infty]$. This result (Theorem 2) is based on another weak convergence result given in Lemma 1. The proof of Lemma 1, which may be of independent interest, is given in the appendix. In Section 4, we apply our technics to the dataset introduced by Bienen & van de Walle (1991). Eventually, to facilitate a visual comparison of $\hat{\bar{G}}_X$ with $\bar{G}_X$, we present confidence intervals in Section 5 for exponential and Weibull cases.

2. FRAMEWORK AND NOTATIONS

The objective of this section is to develop a framework for study of length-biased sampling (LBS) in the setup of competing risks (CR). For convenience, the initiating and terminating events of interest will be called birth and death, respectively. We shall consider a population ($i \in I$) of individuals 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 (Y) is a positive random variable representing the age at death in the hypothetical situation in which A (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. In any case, $Z_i = X_i \wedge Y_i$,

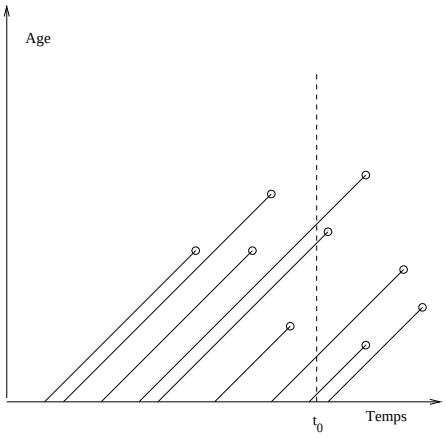

FIG. 1 – A Lexis diagram representation of lifespans

$i \in I$, will denote the lifetime of individual i. The calendar time of birth of the ith individual will be denoted by σ_i. A convenient graphical representation of the lifespan of an individual born at calendar time σ_i and experiencing death at age z_i is given by the well-known Lexis diagram (see Figure1). This diagram consists of line segments in a rectangular coordinate system with calendar time as abscissa and the age as ordinate such that the ith life is represented by the line segment joining the points $(\sigma_i, 0)$ and $(\sigma_i, \sigma_i + z_i)$. The Lexis diagram and associated point processes described in Brillinger (1986), Keiding (1990), and Lund (2000) provide useful settings for analyzing lifetimes. The Lexis diagram 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.

For correct analysis of survival times, it is important to note that a random sample cross-sectional selected at calendar time t_0 is not really a random sample from I but, in fact, from the set $J = \{i \in I : (\sigma_i, x_i, y_i) \in E\}$ where $E = \{(\sigma, x, y) : \sigma \leq t_0, \sigma + x \geq t_0, \sigma + y \geq t_0\}$ and X_i, Y_i are the latent survival times (corresponding to risks A, B, respectively) for individual i. Individuals with lifetime $Z_i = X_i \wedge Y_i$ shorter than $t_0 - \sigma_i$ are excluded from the population J. That is, the survival time Z_i is left truncated by the time $t_0 - \sigma_i$. Thus, the observable random variable is not Z_i but Z_i^b as a random variable whose probability distribution is the same as the conditional distribution of $Z_i = X_i \wedge Y_i$ given $(\sigma_i, X_i, Y_i) \in E$. We shall refer to Z_i^b as the length-biased version of Z_i.

The goal of this paper is to estimate the survivor function $\bar{G}_X(t) = \mathbb{P}(X > t)$ on the basis of a random sample of n from the aforementioned set J when A and B are two independent competing risks with β proportional hazards : $\bar{G}_Y(t) = \mathbb{P}(Y > t) = (\bar{G}_X(t))^\beta$ for all $t \geq 0$ and for some $\beta > 0$. We will allow the possibility of random right censoring in a manner to be defined later. The available data will include the cause of death for uncensored lifetimes. The

following proposition provides a key fact : it gives the probability distribution of Z^b defined as a random variable whose distribution is the same as the conditional distribution of $Z = X \wedge Y$ given $(\sigma, X, Y) \in E$. Here, σ is the calendar time of birth of a life having latent survival times X, Y. It will be seen that, under the assumptions made, the distribution of Z^b will be independent of σ. Thus, the survival times Z_1^b, ...,Z_n^b of the n individuals in the sample selected at t_0 will be independent copies of Z^b.

Theorem 1 *Suppose that :*
(i) the latent survival time X and Y are independent with respective probability density functions $g_X(t) = -d\bar{G}_X(t)/dt$ and $g_Y(t) = -d\bar{G}_Y(t)/dt$,
(ii) $\mathbb{E}|X| < \infty$ and $\mathbb{E}|Y| < \infty$, and
(iii) 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.
Then, Z^b has probability density function

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

for $z > 0$.

Proof of Theorem 1.
Although the above result is merely the competing risks statements of the well-known length-biased density (see, e.g., Lund (2000) and 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 B in $\mathbb{R}$, $\eta(B)$ is the random variable giving the number of births encountered in B. We assume that $a.s.$ $\eta(B) < \infty$. For each individual i, the birth time σ_i is "marked" by the pair of latent survival times (X_i, Y_i). We now define the Lexis point process

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

on $(\mathbb{R} \times \mathbb{R}_+^2, \mathcal{B}_{\mathbb{R}} \otimes \mathcal{B}_{\mathbb{R}_+^2})$, where $\mathcal{B}_{\mathbb{R}}$ denotes the Borel σ-algebra on $\mathbb{R}$. 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 $\lambda_{|\varphi}(\sigma, x, y) = \varphi g_X(x) g_Y(y)$ and with mean-measure $\Lambda_{|\varphi}(B) = \int_B \lambda_{|\varphi}(\sigma, x, y) d\sigma dx dy$ for each Borel set B on $\mathbb{R} \times \mathbb{R}_+^2$. We refer to Kingman (1993) for the marking theorem exploited here. Further, let $\mu_{|\varphi}(. \cap E)$ be the restriction of the Poisson process $\mu_{|\varphi}$ to the measurable set $E = \{(\sigma, x, y) : \sigma \leq t_0, \sigma + x \geq t_0, \sigma + y \geq t_0\}$. Then, by the well-known restriction theorem for Poisson processes (see, e.g., Kingman - 1993), $\mu_{|\varphi}(. \cap E)$ is a Poisson process on $\mathbb{R} \times \mathbb{R}_+^2$ with mean-measure $\Lambda_{E|\varphi}(B) = \Lambda_{|\varphi}(B \cap E) = \int_{B \cap E} \lambda_{|\varphi}(\sigma, x, y) d\sigma dx dy$ for each Borel set B in $\mathbb{R} \times \mathbb{R}_+^2$.

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. Kingman (1993), given $\mu_{|\varphi}(E) = N$, the points of the Poisson process $\mu_{|\varphi}(\cdot \cap E)$ look exactly like $\mu_{|\varphi}(E)$ independent random

variables, with common probability measure

$$\mathbb{P}_{E|\varphi}(\cdot) = \frac{\Lambda_{|\varphi}(\cdot \cap E)}{\Lambda_{|\varphi}(E)}$$

on Borel subsets of $\mathbb{R} \times \mathbb{R}_+^2$. Hayakawa (2000) showed that the order statistics property is a characterisation of mixed Poisson processes within the general class of point processes, this indicates that assumption (iii) can not be weakened.

Let X^b, Y^b denote the latent survival times (corresponding to risks A and B respectively) for an individual in J where, as defined earlier, $J = \{i \in I : (\sigma_i, X_i, Y_i) \in E\}$. Taking $B = \{(\sigma, x, y) : x \leq x_0, y \leq y_0\}$; $x_0, y_0 > 0$; it follows from the above discussion that, conditional on $\mu_{|\varphi}(E)$:

$$\begin{aligned}
& \mathbb{P}_{|\varphi}(X^b \leq x_0, Y^b \leq y_0) \\
= & \frac{\Lambda_{|\varphi}(B \cap E)}{\Lambda_{|\varphi}(E)} \\
= & \frac{\int_{B \cap E} \varphi g_{X,Y}(x,y) d\sigma dx dy}{\int_E \varphi g_{X,Y}(x,y) d\sigma dx dy} \\
= & \frac{\int_0^{y_0} \int_0^{x_0} \int_{t_0 - \inf(x,y)}^{t_0} g_{X,Y}(x,y) d\sigma dx dy}{\int_{-\infty}^{t_0} \bar{G}_X(t_0 - \sigma) \bar{G}_Y(t_0 - \sigma) d\sigma} \\
= & \frac{\int_0^{x_0} \int_0^{y_0} \inf(x,y) g_{X,Y}(x,y) dx dy}{\mathbb{E}(Z)}
\end{aligned}$$

where $Z = X \wedge Y$ and $\bar{G}_X$ and $\bar{G}_Y$ are the survival functions of X and Y, respectively. Since the last expression does not involve φ, integrating w.r. its distribution, we get

$$\mathbb{P}(X^b \leq x_0, Y^b \leq y_0) = \frac{\int_0^{x_0} \int_0^{y_0} \inf(x,y) g_{X,Y}(x,y) dx dy}{\mathbb{E}(Z)}.$$

The p.d.f. of the vector (X^b, Y^b) is then

$$f_{X^b, Y^b}(x,y) = \frac{1}{\mathbb{E}Z}(x \wedge y) g_X(x) g_Y(y). \tag{1}$$

Hence $Z^b = X^b \wedge Y^b$ has survival function

$$\begin{aligned}
\bar{F}_{Z^b}(z) &= \frac{1}{\mathbb{E}(Z)} \int_z^\infty \int_z^\infty (x \wedge y) g_X(x) g_Y(y) dx dy \\
&= \frac{1}{\mathbb{E}(Z)} [\int_z^\infty \int_z^\infty x I_{(x \leq y)} g_X(x) g_Y(y) dx dy \\
&+ \int_z^\infty \int_z^\infty y I_{(y \leq x)} g_X(x) g_Y(y) dx dy] \\
&= \frac{1}{\mathbb{E}Z} [\int_z^\infty x g_X(x) \bar{G}_Y(x) dx + \int_z^\infty y g_Y(y) \bar{G}_X(y) dy].
\end{aligned}$$

The proposition follows on taking the derivative.□

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 that these exits $\beta > 0$ such that for all $t > 0$:

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

where $\Lambda_X = -\ln(\bar{G}_X)$ and $\Lambda_Y = -\ln(\bar{G}_Y)$ are the cumulative hazard functions of X and Y, respectively. Equivalently,

$$\bar{G}_Y(x) = (\bar{G}_X(x))^\beta$$

for all $t > 0$. Under this assumption, the p.d.f. of Z^b reduces to

$$f_{Z^b}(z) = \frac{1}{\mathbb{E}Z}(1+\beta)zg_X(z)(\bar{G}_X(z))^\beta,\ z > 0. \tag{2}$$

The constant β gives the odds on death due to cause B. That is,

$$\frac{\mathbb{P}(Y^b \leq X^b)}{\mathbb{P}(X^b \leq Y^b)} = \beta.$$

Furthermore, it can be shown the the random variables $I(\{X^b \leq Y^b\})$ and $Z^b = X^b \wedge Y^b$ are independent. It may be noted here that, as it is evident from (1), the random variables X^b and Y^b are not independent. The initial independence of X and Y has been lost under the selection process.

3. ESTIMATION OF SURVIVAL FUNCTIONS

The survivor function of primary interest is $\bar{G}_X(t) = \mathbb{P}(X > t)$. We now construct an estimator of $\bar{G}_X$ on the basis of a length-biased sample described earlier. Adhering to the notations of the previous section and ignoring, for the moment, the possibility of right censoring, the observable random variable isZ^b rather than $Z = X \wedge Y$. The probability distribution of Z^b is the conditional distribution of Z given that $(\sigma, x, y) \in E$. Under the assumption that X and Y are independent with proportional hazards, the unconditional distribution of$Z$ had p.d.f.

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

Therefore, by Theorem 1,

$$f_{Z^b}(z) = \frac{1}{\mathbb{E}(Z)}zg_Z(z),\ z > 0.$$

That is, the density of Z^b is the length-biased version of the density of$Z$. Consequently, by the well-known inversion formula of Cox (1969), the distribution function $G_Z(t) = \mathbb{P}(Z \leq t)$ is expressible as

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

where $\bar{F}_{Z^b}(\cdot)$ is the survivor function of Z^b. On the other hand,

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

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

$$\hat{\bar{G}}_X(t) = \left(1 - \hat{G}_Z(t)\right)^{\hat{\alpha}}, \; t > 0 \tag{3}$$

where

$$\hat{G}_Z(t) = \frac{\int_0^t \frac{1}{z} d\hat{\bar{F}}_{Z^b}(z)}{\int_0^\infty \frac{1}{z}, d\hat{\bar{F}}_{Z^b}(z)}, \; t > 0 \tag{4}$$

$\hat{\bar{F}}_{Z^b}(\cdot)$ and $\hat{\alpha}$ are developed below.

To obtain an estimator $\hat{\bar{F}}_{Z^b}$, we now admit the possibility of right-censoring. However, the censoring to be considered will be of the type "end of study" or "loss of follow-up". Each individual in the sample, selected in the manner of section 2, is followed until death or censoring. The observed data then consists of n independent pairs (T_i, δ_i^b) where $T_i = Z_i^b \wedge C_i$ and

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

Here, $C_1, ...C_n$ are independent copies of a random variable C which is independent of Z^b and has survivor function $\bar{H}_C$. For later use, let $S(t) = \bar{F}_{Z^b}(t)\bar{H}_C(t)$ denote the survivor function of $T = Z^b \wedge C$. Further, let

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

be the counting process associated with the jth cause of death and let

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

be the at-risk process. Let the process J be defined by $J(t) = I(\{Y(t) > 0\})$. Defining

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

the survivor function $\bar{F}_{Z^b}$ will be estimated by the Kaplan-Meier estimator (Andersen et al. (1992))

$$\hat{\bar{F}}_{Z^b}(t) = \prod_{s \le t} \left(1 - \frac{J(s)\Delta N(s)}{Y(s)}\right).$$

As for the survivor functions of X^b and Y^b, they can not be estimated here because of the lack of independence between X^b and Y^b. However, the corresponding sub-survivor functions

$$\begin{aligned} F_1(t) &= \mathbb{P}(X^b \leq t, X^b \leq Y^b) \\ &= \mathbb{P}(Z^b \leq t, X^b \leq Y^b) \\ &\text{and} \\ F_2(t) &= \mathbb{P}(Z^b \leq t, Y^b \leq X^b) \end{aligned} \tag{5}$$

can be estimated from the available censored sample. Indeed, they can be estimated by the Aalen-Johansen estimators

$$\begin{aligned} \hat{F}_1(t) &= \int_0^t \hat{\bar{F}}(x-)\frac{dN_1(x)}{Y(x)} \\ &\text{and} \\ \hat{F}_2(t) &= \int_0^t \hat{\bar{F}}(x-)\frac{dN_2(x)}{Y(x)}. \end{aligned} \tag{6}$$

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

$$\alpha = \mathbb{E}\left(I(\{X^b \leq Y^b\})\right)$$

so that α may be estimated by

$$\hat{\alpha} = \hat{F}_1(\infty) \tag{7}$$

The proposed estimator of $\bar{G}_X$ is then given by (3) with plug-ins coming from (4), (5), (7) and (5).

Our main result consists of the weak convergence of the process $\sqrt{n}(\hat{G}_X - G_X)$ on the whole line. Let us define Assumption $\mathcal{A}$ as

$$\int_0^\infty \frac{dF_{Z^b}(x)}{\bar{H}_C(x)} < \infty.$$

Theorem 2 *If assumption $\mathcal{A}$ is fulfilled, the following weak convergence holds in the Skorohod space $\mathbb{D}[0,\infty]$:*

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

as n goes to ∞, where L is defined by

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

Z is a gaussian process defined on $[0,\infty]$ with covariance function given by

$$< Z(s), Z(t) >= \bar{F}_{Z^b}(s)\bar{F}_{Z^b}(t)\int_0^{s\wedge t} \frac{dF_{Z^b}(x)}{\bar{F}_{Z^b}(x)S(x)}$$

and U is a mean-zero normally distributed r.v. with variance given by

$$\begin{aligned}\mathbb{V}(U) &= \int_0^\infty (F_1(\infty)-F_1(x))^2 \frac{d\bar{F}_{Z^b}(x)}{\bar{F}_{Z^b}(x)S(x)} + \int_0^\infty \bar{F}_{Z^b}(x)^2 \frac{d\bar{F}_{Z^b}(x)}{\bar{F}_1(x)S(x)} \\ &- 2\int_0^\infty (F_1(\infty)-F_1(x))\bar{F}_{Z^b}(x)\frac{d\bar{F}_{Z^b}(x)}{\bar{F}_1(x)S(x)}.\end{aligned}$$

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

Lemma 1 *Under Assumption $\mathcal{A}$, as n goes to ∞, the following weak convergence holds in $\mathbb{D}[0,\infty]\times\mathbb{R}$*

$$\begin{pmatrix} \sqrt{n}\left(\hat{G}_Z - G_Z)\right) \\ \sqrt{n}(\hat{\alpha}-\alpha) \end{pmatrix} \rightsquigarrow \begin{pmatrix} L \\ U \end{pmatrix}. \tag{8}$$

Proof of Theorem 2.
In view of equation (3), we have

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

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

$$D\Psi_t(f,r).(h,u) = rf(t)^{r-1}h(t) + f(t)^r lnf(t)u.\square$$

4. DATASET ANALYSIS

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 and van de Walle (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

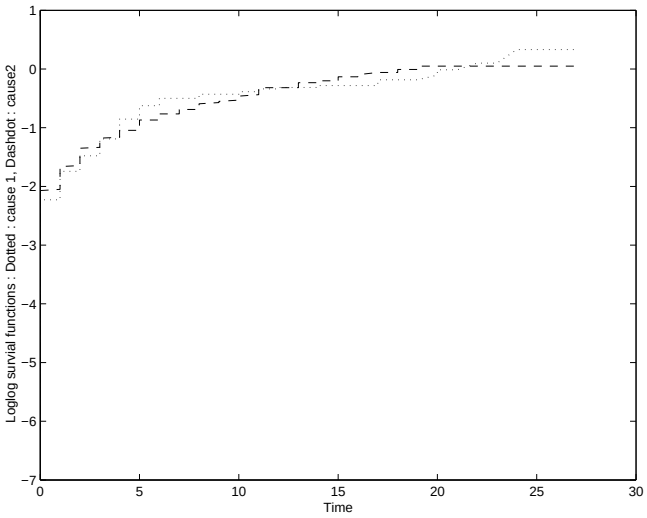

FIG. 2 – Loglog estimated survival functions with the initial sample

such leaders facing two competing risks : exit by constitutional means (risk A) and non constitutional means (risk B). We treat exit by death due to natural causes as a mode of independent random censoring. Continued stay in power until the end of study (the end of 1987) the other mode of this 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 Figure 2 which provides the plots of log-log survivor functions for the two risks against time. The log-log survivor functions of Figure 2 have been estimated from the initial sample. Figure 3 shows the survivor function associated to risk B and estimated from respectively the initial and length-biased samples

5. SIMULATIONS

In order to evaluate the performance of our estimator, we did the following two simulation studies.

5.1. Exponential distribution

The birth process is generated from a uniform distribution due to the property of a mixed Poisson process. The r.v. X has an exponential distribution

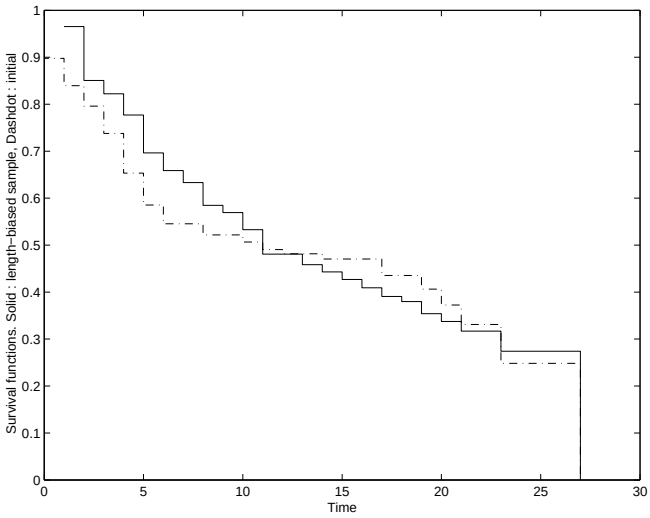

FIG. 3 – Estimated survival function for the second cause of death : initial and length-biased samples

with mean 1 and Y is such that :

$$\bar{G}_Y(x) = (\bar{G}_X(x))^{\beta}. \tag{9}$$

Hence Y has also an exponential distribution with mean $1/\beta$. We also generated a censoring r.v. C from an exponential distribution with mean $1/\lambda$. Our results are presented in Figure 4.

In our example, the complete population is of size $n = 5000$. For $\beta = 0.60$, $\lambda = 0.5$ and an uniform distribution on $[-10, 10]$, the length-biased sample is of size 172. In this biased sample, 44.19 percent of the individuals are censored. Since the variance function of our estimator is hard to compute, we estimated it by bootstrap.

5.2. Weibull distribution

In this case, X is assumed to have a Weibull distribution with parameters 1 and 1.5. Hence Y is also Weibull with parameters β and 1.5. The censoring r.v. has an exponential distribution with mean $\frac{1}{\lambda}$. In the example presented below, $\beta = 0.60$ and $\lambda = 0.55$. The complete observation sample's size is $n = 5000$ and the length-biased one's is 155. 31.6 percent of the sample's observations are censored. The variance of our estimator is again estimated by bootstrap. The results are shown in Figure 5.

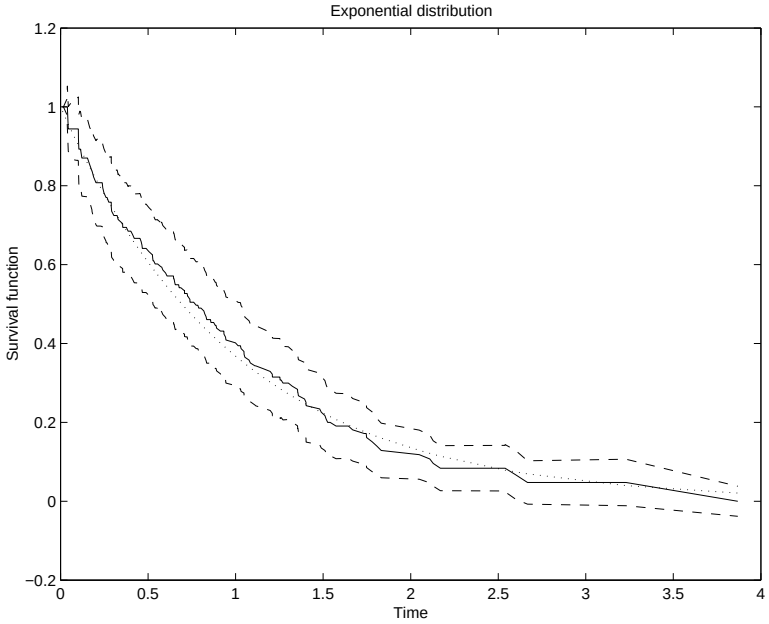

FIG. 4 – Estimator (solid), boostraped IC (Dashed) and true function(Dotted)

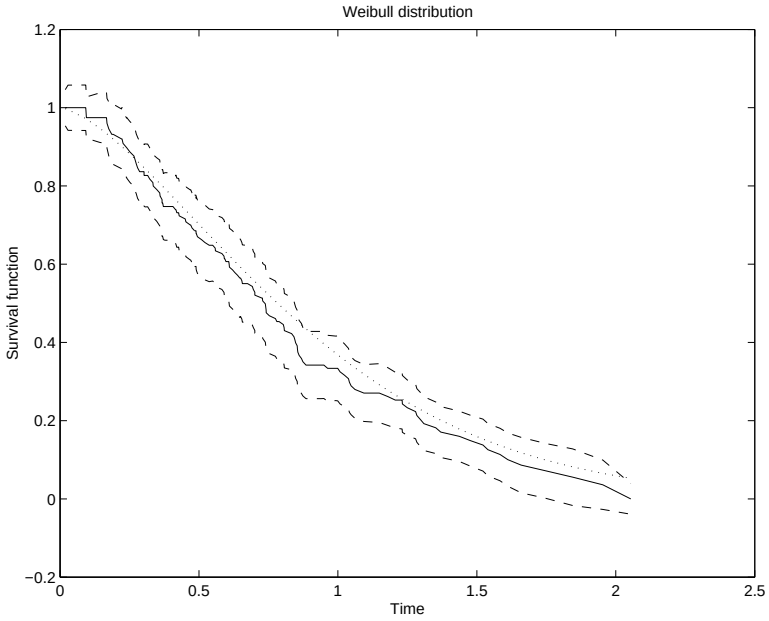

FIG. 5 – Estimator (solid), boostraped IC (Dashed) and true function(Dotted)

7. PROOF OF THE LEMMA

Lemma 1 *Under assumption $\mathcal{A}$, as n goes to ∞, the following weak convergence holds in $\mathbb{D}[0,\infty] \times \mathbb{R}$*

$$\begin{pmatrix} \sqrt{n}\left(\hat{G}_Z - G_Z\right) \\ \sqrt{n}(\hat{\alpha} - \alpha) \end{pmatrix} \rightsquigarrow \begin{pmatrix} L \\ U \end{pmatrix}.$$

Proof of Lemma 1. From Theorem 3 of Dauxois and Guilloux (2004), we have, under Assumption $\mathcal{A}$,

$$\sqrt{n}\begin{pmatrix} \hat{\bar{F}}_{Z^b} - \bar{F}_{Z^b} \\ \hat{F}_1 - F_1 \end{pmatrix} \rightsquigarrow Z = \begin{pmatrix} Z \\ Z_1 \end{pmatrix}$$

in $\mathbb{D}^2[0,\infty]$, where Z_1 is a gaussian process defined on $[0,\infty]$ with covariance function given by

$$\begin{aligned} < Z_1(s), Z_1(t) > &= \int_0^{s\wedge t} (F_j(t) - F_j(u))^2 \frac{dF_{Z^b}(u)}{\bar{F}_{Z^b}(u)S(u)} + \int_0^{s\wedge t} \bar{F}^2_{Z^b}(u)\frac{dF_{Z^b}(u)}{\bar{F}_1(u)S(u)} \\ &- \int_0^{s\wedge t} (F_j(t) - F_j(u))\bar{F}_{Z^b}(u)\frac{dF_{Z^b}(u)}{\bar{F}_1(u)S(u)}. \end{aligned}$$

It is easily seen that

$$\sqrt{n}\begin{pmatrix} \hat{G}_Z - G_Z \\ \hat{\alpha} - \alpha \end{pmatrix} = \sqrt{n}\begin{pmatrix} \Phi_t(\hat{\bar{F}}_{Z^b}) - \Phi^1_t(\bar{F}_{Z^b}) \\ \hat{F}_1(\infty) - F_1(\infty) \end{pmatrix}.$$

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

$$\Phi_t(f) = \frac{\int_0^t \frac{1}{z} df(z)}{\int_0^\infty \frac{1}{z} df(z)}$$

The map Φ_t is Hadamard-differentiable with derivative $D\Phi_t(\bar{F}_{Z^b}) \cdot h$ at $\bar{F}_{Z^b}$ applied to h in $\mathbb{D}[0,\infty]$ defined by

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

The functional delta method in its version of Theorem 3.9.4. of Van der Vaart & Wellner (1996) applies.□

8. REFERENCES

Références

[1] Allison P.D. (1995), *Survival analysis using the SAS system : a practical guide*. SAS Institute.

[2] Andersen P.K., Borgan O., Gill R.D. and Keiding N. (1992), *Statistical models based on counting processes*. Springer-Verlag.

[3] Armitage P. (1959), Comparison of survival curses, *J. R. Stat. Soc. Ser. A.* **122**, 279-300.

[4] Asgharian M., M'Lan C.E. and Wolfson D.B. (2002), Length-biased sampling with right censoring : an unconditional approach, *J. Amer. Statist. Assoc.* **97**, 201-209.

[5] Bienen H.S. and van de Walle N. (1991), *Time of power*. Stanford University Press.

[6] Blumenthal S. (1967). Limit theorems for functions of the shortest two-sample spacings and related test. *Ann. Math. Statist.* **38**, 108-116.

[7] Billingsley P. (1968). *Convergence of probability measures*. Wiley.

[8] Brillinger D.R. (1986). The natural variability of vital rates and associated statistics (with discussion), *Biometrics* **42**, 693-734

[9] Cheng P.E. and Lin G.D. (1987). Maximum likelihood estimation of a survival function under the Koziol-Green proportional hazards model. *Statist. Probab. Lett.* **5**, 75-80.

[10] Cox D.R. (1969), in *New development in survey sampling* eds Johnson and Smith, Wiley.

[11] Crowder M. (2001). *Classical competing risks*, Chapman and Hall.

[12] Csörgö S. (1988). Estimation in the proportional hazards model of random-censorship. *Statistics.* **19**, 437-463.

[13] Dauxois J.Y. (2000). A new method for proving weak convergence results applied to nonparametric estimators in survival analysis. *Stochastic Process. Appl.* **90**, 327-334.

[14] Dauxois J.Y. and Guilloux A. (2004). Estimating the cumulative incidence functions under length-biased sampling. Submitted and *Rapports techniques du CREST n° 2004-01.*

[15] Feinlieb M. and Zelen M. (1969). On the theory of screening for chronic diseases *Biometrika*. **56**, 601-614.

[16] Gather U. and Pawlitschko J. (1998) Estimating the survival function under a generalized Koziol-Green model with partially informative censoring *Metrika*. **48**, 189-207.

[17] Gill R.D., Vardi Y. and Wellner J.A. (1988). Large sample theory of empirical distributions in biased sampling model *Ann. Statis.*. **16**, 1069-1172.

[18] Hayakawa Y. (2000). A new characterisation property of mixed Poisson processes via Berman's theorem. *Jour. Appl. Prob.* **37**, 261-268.

[19] Huang Y. and Wang M-C (1995). Estimating the occurence rate for prevalent survival data in competing risks models *J. Amer. Statist. Assoc.* **90**, 1406-1415.

[20] Jakubowski A., Mémin J. and Pages G. (1989). Convergence en loi des suites d'intégrales stochastiques sur l'espace $\mathbb{D}1$ de Skohorod. *Probab. Theory Related Fields*, **81**, 111-137.

[21] Keiding N. (1990). Statistical inference for the Lexis Diagram. *J. R. Stat. Soc. Ser. A* **332**, 487-509

[22] Kingman J.F.C. (1993). *Poisson processes*. Oxford Science Publications.

[23] Kurtz T.G. and Protter P. (1991). Weak limit theorems for stochastic integrals and stochatic differential equations. *Annals of Probability*. **19**, 1035-1070.

[24] Lexis W. (1875). Einleitung in die Theorie der Bevölkerung-Statistik. Dans *Mathematical demography* (édition D. Smith and N. Keyfitz), *Biomathematics* **6**, 39-41 (1977), Springer-Verlag, Berlin

[25] Lund J. (2000). Sampling bias in population studies - How to use the Lexis diagram. *Scandinavian Journal of Statistics* **27**, 589-604

[26] McFadden J.A. (1962), On the lengths on intervals in stationary point process, *Journal of the Royal Society - Series B*. **24**, 364-382.

[27] Simon R. (1980), Length biased sampling in etiologic studies, *American Journal of Epidemiology*. **111**, 444-452.

[28] Tsai W-Y., Jewell N.P. and Wang M-C. (1987), A note on the product-limit estimator under right censoring and left truncation, *Biometrika*. **4**, 883-886.

[29] Turnbull B.W. (1976), The empirical distribution function with arbitrarily grouped, censored and truncated data , *J. R. Stat. Soc. Ser. B*. **38**, 290-295.

[30] 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.

[31] Vardi Y. (1982), Nonparametric estimation in presence of length bias, *Ann. Statist.* **10**, 616-620.

[32] Vardi Y. (1985), Empirical distributions in selection bias models, *Ann. Statist.* **13**, 178-205.

[33] Vardi Y. (1989), Multiplicative censoring, renewal processes, deconvolution and decreasing density, *Biometrika*. **76**, 751-761.

[34] Vardi Y. and Zhang C-H (1992), Large sample study of empirical distributions on a random-multiplicative censoring model, *Ann. Statist.* **20**, 1022-1039.

[35] Wang M-C., Jewell N.P. and Tsai W-Y (1986), Asymptotic properties of the product limit estimate under random truncation, *Ann. Statist.* **14**, 1597-1605.

[36] Wang M-C. (1991), Nonparametric estimation from cross-sectional survival data, *J. Amer. Statist. Assoc.* **86**, 130-043.

[37] Wang M-C., Brookmeyer R. and Jewell N.P. (1993),Statistical models for prevalent cohort, *Biometrics*. **49**, 1-11.

[38] Woodroofe M. (1985), Estimating a distribution function with truncated data, *Ann. Statist.* **13**, 163-177.