ADAPTIVE NONPARAMETRIC STRATEGIES FOR CENSORED LIFETIMES WITH UNKNOWN SAMPLING BIAS.

E. BRUNEL[1], F. COMTE[2], AND A. GUILLOUX[3]

Abstract. In addition to the right-censoring of the lifetimes, sampling bias possibly occurs in many situations while collecting the data. In practical situations, the form of the bias function is always assumed to be known, but if no information is available on the selection bias this can turn out to be unbearable. We provide a model selection strategy for estimating the hazard and the density functions in presence of both censoring and unknown sampling bias. We use a new sampling scheme description based on the Lexis diagram including time-window or cohort studies. Adaptive projection estimators on trigonometric bases are developed by contrast penalization and optimal nonparametric rates of convergence are given. Both estimators are practically studied through simulation experiments in the case of the time-window study.

June 13, 2006

AMS Classification (2001): 62G07; 62N01.

Keywords. Adaptive estimation. Biased data. Right-Censoring. Hazard rate. Nonparametric penalized contrast estimator. Nonhomogeneous Poisson process.

1. Introduction

We consider here the problem of estimating the probability density function (p.d.f.) f and the hazard rate $\lambda = f/(1-F)$, where F is the distribution function (c.d.f) F, associated with a random variable (r.v.) X. Our nonparametric estimates are built on a compact set A, and we often take $A = [0,1]$ for simplicity. If independent and identically distributed (i.i.d.) observations $X_1, \dots, X_n$ were available, the p.d.f. f may be estimated by usual penalized projection methods developed by Barron & Cover (1991), Barron *et al.* (1999). Here we assume that it is impossible to draw a direct sample from the distribution of X. Instead the observable r.v. has p.d.f. given by $w(x)f(x)/\int w(u)f(u)du$, for all $x > 0$. In this case, the r.v. is said to suffer from a "selection bias" and one may heuristically say that selection bias arises when the event $\{X = x\}$ is observed with a probability proportional to $w(x)$.

Bias may occur in the following kind of model. Consider, in a population of individuals I, the r.v.'s of their birth dates $(\sigma_i)_{i\in I}$, and the non-negative r.v.'s of

$^{(1)}$ IUT de Paris V, MAP5, UMR CNRS 8145. email: elodie.brunel@univ-paris5.fr.
$^{(2)}$ Université Paris V, MAP5, UMR CNRS 8145. email: fabienne.comte@univ-paris5.fr.
$^{(3)}$ L.S.T.A., Université Paris VI. email: aguillou@ccr.jussieu.fr.

their lifetimes $(X_i)_{i\in I}$. In the Lexis (1875) diagram, an individual can be represented by his life-line, $\mathcal{L}(\sigma, X) = \{(\sigma+y, y),\ 0 \le y \le X\}$, which is a unit-slope line whose points have as coordinates the calendar time $(\sigma+y)$ and the age (y). Lexis diagrams have been recently considered for such modelization purpose by Keiding (1990), Lund (2000) and Guilloux (2006). Let $\mathcal{S}$ be a deterministic Borel set in the Lexis diagram. Individuals with life-lines intersecting $\mathcal{S}$ are included in the study, i.e. only pairs (σ, X) such that $\mathcal{L}(\sigma, X)\cap\mathcal{S} \neq \emptyset$. Time-window or cohort studies can be described by such a sampling pattern with a particular choice of $\mathcal{S}$.

Let $\sigma_{\mathcal{S}}$ denote the birth time and $X_{\mathcal{S}}$ the lifetime for the included individuals. From now on, the pair $(\sigma_{\mathcal{S}}, X_{\mathcal{S}})$ will be referred to as the observable r.v.'s as opposed to the unobservable pair (σ, X). Straightforwardly, we have for all $s \in \mathbb{R}$ and $x \ge 0$:

$$\begin{aligned} \mathbb{P}(\sigma_{\mathcal{S}} \le s, X_{\mathcal{S}} \le x) &= \mathbb{P}\left(\sigma \le s, X \le x | \mathcal{L}(\sigma, X)\cap\mathcal{S} \neq \emptyset\right) \\ &\neq \mathbb{P}(\sigma \le s, X \le x). \end{aligned}$$

More precisely, we know from Guilloux (2006) that, under some condition on the birth-dates $(\sigma_i)_{i\in I}$, we have, for all $x \ge 0$:

$$F_{\mathcal{S}}(x) = \mathbb{P}\left(X_{\mathcal{S}} \le x\right) = \frac{\int_0^x w(v)dF(v)}{\mu_{\mathcal{S}}}, \tag{1}$$

where F is the c.d.f. of the r.v. X and w is a non-negative weight function, which depends only on the distribution of the r.v. σ and $\mu_{\mathcal{S}} = \int_0^\infty w(v)dF(v)$.

The lifetimes can also be subject to right-censoring. In this model, we can thus address the question of estimating the density f or the hazard rate λ of the underlying r.v. X, *without knowing the bias function.*

The relevance of selection bias in statistical inference has been first pointed out by Fisher (1934). Since then, many authors noticed its presence in data from a wide range of fields. We refer to Chakraborty & Rao (2000) for biomedical applications, and Heckman (1985) in Economics, among many others. The review by Patil & Rao (1977) gives numerous practical examples of weighted distributions.

The problem of estimating the cumulative distribution function (c.d.f.) F of X given an i.i.d. biased sample $X_{w,1}, \dots, X_{w,n}$ has been widely studied. We refer to Gill *et al.* (1988) and Efromovich (2004a) for theoretical results in the general case. The special case where $w(x) = x$ for all $x > 0$, called "length-biased sampling", has received a particular attention, see Vardi (1982), de Uña-Àlvarez (2002) and Asgharian *et al.* (2002). Efromovich (2004b) dealt with the problem of nonparametric estimation of f. He considered a minimax estimation procedure with a known weight function w and states asymptotic results for several estimators. Brunel *et al.* (2005) have also considered this problem with additional right-censoring of the observations.

Nonparametric methods for the estimation of the density function and the hazard rate under censorship have also been widely studied, see Tanner & Wong (1983), Mielniczuk (1985), Marron & Padgett (1987) and Lo *et al.* (1989) for kernel estimators and more recently Antoniadis *et al.* (1999) for wavelet methods. In the setting of both biased and censored data, we have to mention de Uña-Àlvarez (2002) who introduced a suitable correction of the Kaplan-Meier estimator for taking the length bias problem into account. The kernel density (or hazard) estimator is directly obtained by convolution of a kernel function with this corrected estimator of the survival function see Marron & de Uña-Àlvarez (2004). But, it is worth noticing that more general selection bias models with censoring have rarely been simultaneously investigated up to now.

In what follows, we construct and study adaptive estimators of the hazard rate λ and the density f in the general model described above. This means that we propose and study an estimation procedure under bias and censoring without assuming that the bias is known. Nonparametric projection methods are developed as in Brunel *et al.* (2005), by taking advantage of the Dvorestky-Kiefer-Wolfowitz type inequalities proved by Guilloux (2006).

The paper is organized as follows. In Section 2, the bias and censoring model is described, some of its properties are recalled (see Guilloux (2006)). Section 3 defines the penalized estimators of the functions λ and f, and provides nonasymptotic bounds for their respective mean integrated squared errors (MISE). In Section 4, simulation experiments illustrate the penalized estimators in the case of the time-window study. Most proofs are gathered in Section 5.

2. Bias and censoring model

2.1. **The sampling model.**

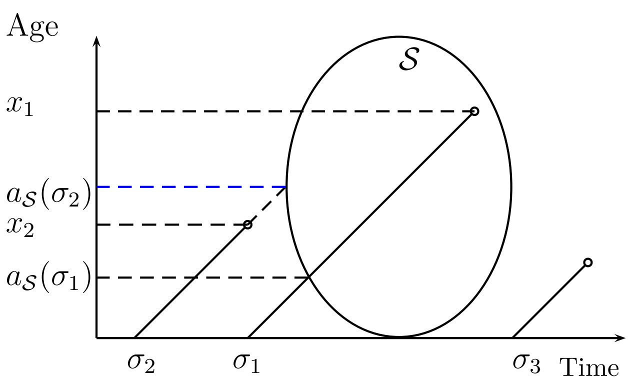


Figure 2.1. Sampling scheme

Let $\mathcal{S}$ be a Borel set and let us denote by $a_{\mathcal{S}}(\sigma)$ the age at inclusion for a birth-time σ:

$$\begin{cases} a_{\mathcal{S}}(\sigma) = \inf\{y \geq 0,\ (\sigma + y, y) \in \mathcal{S}\} \\ a_{\mathcal{S}}(s) = \infty \text{ if the infimum does not exist.} \end{cases}$$

The individual i with birth date σ_i and lifetime X_i is included in the sample if:

$$\mathcal{L}(\sigma_i, X_i) \cap \mathcal{S} \neq \emptyset \Leftrightarrow a_{\mathcal{S}}(\sigma_i) < \infty \text{ and } X_i \geq a_{\mathcal{S}}(\sigma_i). \tag{2}$$

Figure 2.1 gives a representation of the Lexis diagram and the sampling process described above : individual 1 is included in the sample, individual 2 could have been included but died before his inclusion since $x_2 < a_{\mathcal{S}}(\sigma_2)$, whereas individual 3 is not included because $a_{\mathcal{S}}(\sigma_3) = \infty$.

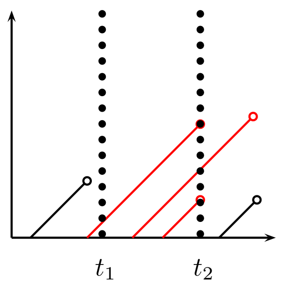
2.2. **Two examples for $\mathcal{S}$.** Before going on in the presentation, let us give two examples of sampling patterns, which can be described through a Borel set in $\mathcal{B}_{\mathbb{R}\times\mathbb{R}_+}$: the time-window and cohort studies.

Time-window study. The individuals alive at time t_1 or born between t_1 and t_2 are included in the study, t_1 and t_2 are fixed. The Borel set to consider is :

$$\mathcal{S}_{tw} = \{(s, y), t_1 \leq s \leq t_2, y \in \mathbb{R}_+\}.$$

The age $a_{\mathcal{S}_{tw}}$ at inclusion is given by:

$$\begin{cases} a_{\mathcal{S}_{tw}}(\sigma) & = & t_1 - \sigma & \text{if} & -\infty < \sigma \leq t_1 \\ a_{\mathcal{S}_{tw}}(\sigma) & = & 0 & \text{if} & t_1 < \sigma \leq t_2 \\ a_{\mathcal{S}_{tw}}(\sigma) & = & +\infty & \text{if} & \sigma > t_2. \end{cases}$$



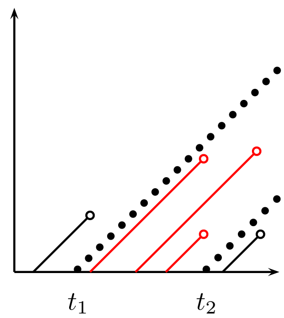
Example 1: Time-window study

Cohort study. The individuals born between times t_1 and t_2 are included in the study, t_1 and t_2 are fixed. In case, the Borel set $\mathcal{S}_c$ is given by:

$$\mathcal{S}_c = \{(s + u, u), t_1 \leq s \leq t_2, u \in \mathbb{R}_+\}.$$

The age $a_{\mathcal{S}_c}$ at inclusion is then given by:

$$\begin{cases} a_{\mathcal{S}_c}(\sigma) & = & +\infty & \text{if} & -\infty < \sigma \leq t_1 \\ a_{\mathcal{S}_c}(\sigma) & = & 0 & \text{if} & t_1 < \sigma \leq t_2 \\ a_{\mathcal{S}_c}(\sigma) & = & +\infty & \text{if} & \sigma > t_2. \end{cases}$$



Example 2: Cohort study

We refer to Lund (2000) for an extensive review of studies described through the Lexis diagram.

2.3. **Theoretical representation of the model and results on the involved counting processes.** In this subsection, we recall some results obtained by Guilloux (2006) useful for the present purpose.

2.3.1. *Bias.* Following Lund (2000), we consider the following modelization associated with the Lexis diagram. Let the point process $\eta = \sum_{i\in I} \varepsilon_{\sigma_i}$, with the birth times as occurrence times, be a non-homogeneous Poisson process on $\mathbb{R}$ with intensity φ (where ε_a is the Dirac measure at point a). Assume furthermore, that the lifetimes X_i, for $i \in I$, are i.i.d. with common probability density function f. The marking theorem (see Kingman (1993)) ensures that the point process $\mu = \sum_{i\in I} \varepsilon_{(\sigma_i, X_i)}$ is a nonhomogeneous Poisson process with intensity φf.

Here, we are interested in the individuals whose life-lines intersect the Borel set $\mathcal{S}$. In other words, we are interested in the restriction $\mu_{|\mathcal{S}}$ of the process μ to the Borel set $\mathcal{S}$. The restriction theorem (see Kingman (1993)) ensures that the restriction $\mu_{|\mathcal{S}}$ is a Poisson process with mean measure $\int_{B\cap\mathcal{S}} \varphi f / \int_{\mathcal{S}} \varphi f$, for any Borel set B in $\mathcal{B}_{\mathbb{R}\times\mathbb{R}_+}$. Finally, by the order statistics property for Poisson processes, given the number $\mu(\mathcal{S})$ of points in the Borel set $\mathcal{S}$, the points of Poisson process $\mu_{|\mathcal{S}}$ look exactly like independent random variables, with common probability measure $\mathbb{P}(\cdot) = \int_{\cdot\cap\mathcal{S}} \varphi f / \int_{\mathcal{S}} \varphi f$ on Borel subsets of $\mathbb{R}\times\mathbb{R}_+$.

As a consequence, using (2), we have, for all $s \in \mathbb{R}$ and $x \in \mathbb{R}_+$:

$$\begin{aligned} \mathbb{P}\left(\sigma_{\mathcal{S}} \le s, X_{\mathcal{S}} \le x\right) &= \frac{\iint_{]-\infty,s]\times[0,x]} \mathbb{1}_{\{(u,v)\in\mathcal{S}\}}\varphi(u)f(v)dudv}{\iint_{\mathbb{R}\times\mathbb{R}_+} \mathbb{1}_{\{(u,v)\in\mathcal{S}\}}\varphi(u)f(v)dudv} \\ &= \frac{1}{\mu_{\mathcal{S}}}\iint_{]-\infty,s]\times[0,x]} \varphi(u)f(v)\mathbb{1}_{\{a_{\mathcal{S}}(u)<\infty\}}\mathbb{1}_{\{a_{\mathcal{S}}(u)\le v\}}dudv, \end{aligned} \tag{3}$$

where $\mathbb{1}_{\{.\}}$ stands for the indicator function and

$$\mu_{\mathcal{S}} = \iint_{\mathbb{R}\times\mathbb{R}_+} \mathbb{1}_{\{a_{\mathcal{S}}(u)<\infty\}}\mathbb{1}_{\{a_{\mathcal{S}}(u)\le v\}}\varphi(u)f(v)dudv.$$

Hence the marginal distribution of the r.v. $X_{\mathcal{S}}$ is given, for all $x \in \mathbb{R}_+$, by :

$$F_{\mathcal{S}}(x) = \mathbb{P}\left(X_{\mathcal{S}} \le x\right) = \frac{1}{\mu_{\mathcal{S}}}\int_0^x w(v)f(s)ds, \tag{4}$$

with

$$w(x) = \int_{-\infty}^{\infty} \mathbb{1}_{\{a_{\mathcal{S}}(u)\le x\}}\varphi(u)du. \tag{5}$$

For example, in the time-window study (see Section 2.2), the weight function w is given, for $x \ge 0$, by:

$$w(x) = \int_{t_1-x}^{t_2} \varphi(u)du.$$

In the particular case where $t_1 = t_2$ and φ is a constant, such a sample is called a "length-biased sample", see Asgharian *et al.* (2002) and de Uña-Àlvarez (2002).

In the example of the cohort study (see Section 2.2), the weight function w is constant and given, for $x \ge 0$, by $w(x) = \int_{t_1}^{t_2} \varphi(u)du$.

2.3.2. *Censoring.* Now only the individuals whose life-lines intersect the Borel set $\mathcal{S}$ are included in the study. For an included individual i, with birth date $\sigma_{\mathcal{S},i}$ and lifetime $X_{\mathcal{S},i}$, we assume that his age at inclusion $a_{\mathcal{S}}(\sigma_{\mathcal{S},i})$ is observable. The lifetime $X_{\mathcal{S},i}$ can straightforwardly be written as follows:

$$X_{\mathcal{S},i} \quad = \quad \underbrace{a_{\mathcal{S}}(\sigma_{\mathcal{S},i})}_{\text{age at inclusion}} \quad + \quad \underbrace{(X_{\mathcal{S},i} - a_{\mathcal{S}}(\sigma_{\mathcal{S},i}))}_{\text{time spent in the study}}.$$

As the time spent in the study is given by $X_{\mathcal{S},i} - a_{\mathcal{S}}(\sigma_{\mathcal{S},i})$, we shall assume that this time can be censored. It is the case, for example, if an individual i leaves the study before his death, see Asgharian (2003) and Winter & Földes (1988).

For that matter, we introduce a non-negative r.v. C with p.d.f. h and c.d.f. H, independent of $X_{\mathcal{S}}$ and $a_{\mathcal{S}}(\sigma_{\mathcal{S}})$, such that the observable time for individual i is

$$Z_i = a_{\mathcal{S}}(\sigma_{\mathcal{S},i}) \quad + \quad (X_{\mathcal{S},i} - a_{\mathcal{S}}(\sigma_{\mathcal{S},i})) \wedge C_i.$$

As usual, we assume furthermore that the r.v. $\mathbf{1}_{\{X_{\mathcal{S},i} - a_{\mathcal{S}}(\sigma_{\mathcal{S},i}) \le C\}}$ is observable. As a consequence, the available data are i.i.d. replications of:

$$\text{(6)} \qquad \begin{cases} \sigma_{\mathcal{S},i} \\ Z_i = a_{\mathcal{S}}(\sigma_{\mathcal{S},i}) + (X_{\mathcal{S},i} - a_{\mathcal{S}}(\sigma_{\mathcal{S},i})) \wedge C_i, & \text{for i=1,...,n.} \\ \mathbf{1}_{\{X_{\mathcal{S},i} - a_{\mathcal{S}}(\sigma_{\mathcal{S},i}) \le C_i\}} \end{cases}$$

We seek to estimate the density function f and the hazard rate $\lambda = f/(1-F)$ of the unbiased r.v. X with the available data described by (6).

2.3.3. *Counting processes for the estimation.* In this context, Guilloux (2006) introduces the following counting processes. For all $x \ge 0$, let

$$\text{(7)} \qquad D(x) = \sum_{i=1}^{n} \mathbf{1}_{\{Z_i \le x, X_{\mathcal{S},i} - a_{\mathcal{S}}(\sigma_{\mathcal{S},i}) \le C_i\}}.$$

For $x \ge 0$, the r.v. $D(x)$ is the "number of observed deaths before age x" in the sample. Let furthermore the process O be defined, for all $x \ge 0$, by:

$$\text{(8)} \qquad O(x) \quad = \quad \sum_{i=1}^{n} \mathbf{1}_{\{a_{\mathcal{S}}(\sigma_{\mathcal{S},i}) \le x \le Z_i\}} = \sum_{i=1}^{n} \mathbf{1}_{\{a_{\mathcal{S}}(\sigma_{\mathcal{S},i}) \le x \le X_{\mathcal{S},i}, x \le a_{\mathcal{S}}(\sigma_{\mathcal{S},i}) + C_i\}}$$

The r.v. $O(x)$ represents the "number of individuals at risk at age x". In the sampling situation considered here, to be at risk at age x for an individual means that he was included in the study at age less than x and is neither dead nor censored before age x.

Let Λ denote the cumulative hazard function of the r.v. X and be defined as:

$$\Lambda(x) = \int_0^x \frac{f(s)ds}{1 - F(s)},$$

for all $x \geq 0$. As in classical survival analysis, it seems natural to define the estimator $\widehat{\Lambda}_n^\epsilon$ by:

$$\widehat{\Lambda}_n^\epsilon(x) = \int_0^x \frac{dD(s)}{O(s) + n\epsilon}, \tag{9}$$

for all $x \geq 0$. Guilloux (2006) studies $\widehat{\Lambda}_n^{\epsilon_n}(.)$ where $(\epsilon_n)_{n\geq 1}$ is a sequence of positive numbers such that $\epsilon_n \to 0$ and $\sqrt{n}\varepsilon_n \to 0$ as $n \to \infty$.

We can also define the estimator $\widehat{F}_n$ for the c.d.f. F of the r.v. X. Mimicking the construction of the Kaplan-Meier estimator in classical survival analysis, we define, for all $x \geq 0$:

$$\hat{F}_n(x) = 1 - \prod_{i:Z_i \leq x} \left(1 - \frac{\mathbb{1}_{\{X_{\mathcal{S},i} - a_{\mathcal{S}}(\sigma_{\mathcal{S},i}) \leq C_i\}}}{O(Z_i) + n\epsilon_n}\right). \tag{10}$$

Let us define

$$\tau = \inf(\tau_F, \tau_H) \text{ with } \tau_L = \{x \in \mathbb{R}^+ : (1 - L)(x) = 0\}, \text{ for any c.d.f. } L. \tag{11}$$

It is useful to mention the following result proved in Guilloux (2006).

Theorem 2.1. *Assume that there exist w_1 and w_2 such that $w_1 \leq w(x) \leq w_2$, for all $x \geq 0$. Then, for all $u > 0$:*

$$\mathbb{P}\left(\sqrt{n} \sup_{x\in[0,\tau]} \left| \left(\hat{F}_n(x) - F(x)\right)(1 - H)^2(x) w_1 \right| > u \right) \leq 2.5 \exp(-2u^2 + Cu),$$

where C is an universal constant (which does not depend on F, H nor w).

3. Estimation procedure

3.1. **Assumptions on the model.** Let $A \subsetneq [0, \tau]$ with τ defined by (11). Up to a scaling change, we set $A = [0, 1]$ without loss of generality. First, we assume that

$$\exists w_1, w_2 > 0, \text{ such that } 0 < w_1 \leq w(x) \leq w_2 < +\infty, \ \forall x \in A. \tag{12}$$

This is a very mild condition since A is a compact set.
It follows from (1) that $F_{\mathcal{S}}(x) = \int_0^{\tau_F \wedge x} w(s) f(s) ds / \mu_{\mathcal{S}}$ with $\mu_{\mathcal{S}} = \int_0^{\tau_F} w(s) f(s) ds$, so that $\tau_{F_{\mathcal{S}}} \leq \tau_F$. Moreover, the function w is nondecreasing and by (12) it is lower bounded, thus $\tau_F = \tau_{F_{\mathcal{S}}}$.
Secondly, the following property holds for H:

$$\exists c_H > 0, \forall x \in A, c_H \leq 1 - H(x) \leq 1, \tag{13}$$

by taking $c_H = \inf_{x \in A}[1 - H(x)]$.
Lastly, the same property as (13) holds also for F with c_H replaced by c_F.
In the following, estimates are computed over the compact set $A \subsetneq [0, \tau]$. Note that this does not imply any practical restriction since, for estimation purpose, we can choose the interval A such that the largest uncensored $X_{\mathcal{S},i}$ belongs to it.

3.2. **Description of the projection bases.** Let $\psi_0(x) = 1$, $\psi_{2j}(x) = \sqrt{2}\cos(2\pi jx)$, $\psi_{2j-1}(x) = \sqrt{2}\sin(2\pi jx)$ and consider

$$S_m = \mathrm{Span}(\psi_j(x), j = 0, \dots, m-1).$$

Then S_m is a finite dimensional subspace with dimension m of $\mathbb{L}^2([0,1])$ endowed with the scalar product $\langle u, v\rangle = \int_0^1 u(x)v(x)dx$. The functions $(\psi_j)_{0\le j\le m-1}$ constitute an orthonormal basis of S_m so that $t = \sum_{j=0}^{m-1} a_j\psi_j$, for any function t in S_m and with $a_j = \langle t, \psi_j\rangle$. A key property of the trigonometric basis relies on the link between the supremum norm and the $\mathbb{L}^2$ norm:

$$\forall m, \forall t \in S_m, \|t\|_\infty \le \sqrt{2m}\|t\| \quad \text{and} \quad \|\sum_{j=0}^{m-1}\psi_j^2\|_\infty \le 2m. \tag{14}$$

For $m, m' \in \mathcal{M}_n$, $m \le m'$ implies that $S_m \subset S_{m'}$ which means that the subspaces are nested. Then $\mathcal{S}_n = \cup_{m\in\mathcal{M}_n} S_m$ stands for the largest subspace of the collection and we denote by N_n the dimension of $\mathcal{S}_n$.

3.3. **Definition of the estimators.** Using the observations described by (6), we can define contrast functions to estimate both the density f and the hazard rate λ.

3.3.1. *Estimation of the hazard rate.* Let us define

$$\Phi(x) = \mathbb{P}(a_{\mathcal{S}}(\sigma_{\mathcal{S},1}) \le x \le Z_1),$$

which stands for the probability for an individual to be at risk at age x. This means clearly that the individual was included in the study at age less than x and is neither dead nor censored before x, so that we can write

$$\Phi(x) = \mathbb{P}(a_{\mathcal{S}}(\sigma_{\mathcal{S},1}) \le x \le X_{\mathcal{S},1}, x \le a_{\mathcal{S}}(\sigma_{\mathcal{S},1}) + C_1). \tag{15}$$

From definition (8) of the process $O(x)$, it is obvious that $\mathbb{E}(O(x)/n) = \Phi(x)$. This is why we take

$$\Phi_n(x) = \frac{O(x)}{n}, \tag{16}$$

as an estimator of $\Phi(x)$. Note that

$$\forall n \in \mathbb{N}^*,\ \Phi_n(Z_i) \ge 1/n\,, \text{ for } i = 1, \dots, n. \tag{17}$$

The following lemma states useful properties of Φ and Φ_n with first a useful consequence of Theorem 2.1.

Lemma 3.1. *1) Assume that there exists w_1 such that, for all $x \ge 0$, $w_1 \le w(x)$. For all $k \in \mathbb{N}^*$, there exists a constant $C_F(k)$ depending on k, w and c_H such that*

$$\mathbb{E}\left(\sup_{x\in A}|\hat{F}_n(x) - F(x)|^{2k}\right) \le C_F(k)n^{-k}.$$

2) There exists a constant c_Φ such that, for all $x \in A$, $\Phi(x) \geq c_\Phi$.
3) For all $k \in \mathbb{N}^*$, there exists a constant $C_\Phi(k)$ such that

$$\mathbb{E}\left(\sup_{x\in\mathbb{R}^+} |\Phi_n(x) - \Phi(x)|^{2k}\right) \leq C_\Phi(k) n^{-k}.$$

Remark 3.1. *The condition $w_1 \leq w(x)$ holds for the time-window and the cohort studies as soon as the interior of $\mathcal{S} \cap \{(x,0), x \in \mathbb{R}\}$ is non empty. It means that a death can occur immediately at the inclusion of an individual.*

Moreover, the definition of the contrast for estimating λ is deduced from the following property:

Proposition 3.1. *For any function t in S_m,*

$$\mathbb{E}\left(\frac{\delta_1 t(Z_1)}{\Phi(Z_1)}\right) = \langle t, \lambda\rangle \text{ where } \langle u, v\rangle = \int_0^1 u(x)v(x)dx.$$

Therefore, by replacing $\Phi(x)$ by its natural estimator, we obtain the following projection contrast function, for $t \in S_m$,

$$\gamma_n(t) = \|t\|^2 - \frac{2}{n}\sum_{i=1}^n \frac{\delta_i t(Z_i)}{\Phi_n(Z_i)}. \tag{18}$$

Clearly, $\gamma_n(t)$ is the empirical version of $\|t\|^2 - 2\langle t,\lambda\rangle = \|t-\lambda\|^2 - \|\lambda\|^2$, where $\|t\|^2 = \int_0^1 t^2(x)dx$. Moreover, the contrast function can also be seen as

$$\gamma_n(t) = \|t\|^2 - 2\int_0^1 t(x) d\widehat{\Lambda}_n^0(x)$$

where $\widehat{\Lambda}_n^0$ is defined by (9).
Then we can define the projection estimator

$$\hat{\lambda}_m = \arg\min_{t\in S_m} \gamma_n(t)$$

which gives an estimator of λ_m, the orthogonal projection of λ on S_m. Lastly, we select

$$\hat{m} = \arg\min_{m\in\mathcal{M}_n} \left\{\gamma_n(\hat{\lambda}_m) + \text{pen}(m)\right\}$$

which gives an estimator $\hat{\lambda}_{\hat{m}}$ of λ on a random space $S_{\hat{m}}$. Here, pen(.) is a penalty function that ensures the squared-bias variance compromise, it is given by

$$\text{pen}(m) = \kappa\left(\frac{1}{n}\sum_{i=1}^n \frac{\delta_i}{\Phi_n^2(Z_i)}\right)\frac{m}{n}. \tag{19}$$

Note that

$$\lambda_m = \sum_{j=0}^{m-1} a_j\psi_j \text{ with } a_j = \langle\lambda, \psi_j\rangle,$$

and

$$\hat{\lambda}_m = \sum_{j=0}^{m-1} \hat{a}_j \psi_j, \text{ with } \hat{a}_j = \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i \psi_j(Z_i)}{\Phi_n(Z_i)}.$$

Moreover $\gamma_n(\hat{\lambda}_m) = -\sum_{j=1}^{m-1} \hat{a}_j^2$.

3.3.2. *Estimation of the density.* We can deduce from Proposition 3.1 and from $f = \lambda(1-F) = \lambda \bar{F}$, the following projection contrast for the estimation of f:

$$\tilde{\gamma}_n(t) = \|t\|^2 - \frac{2}{n} \sum_{i=1}^{n} \delta_i t(Z_i) \frac{1 - \hat{F}_n(Z_i)}{\Phi_n(Z_i)}, \tag{20}$$

where $\hat{F}_n$ and Φ_n are the estimators of F and Φ defined by (10) and (16). Again $\tilde{\gamma}_n(t)$ is an empirical version of $\|t\|^2 - 2\int_0^1 t(x)f(x)dx = \|t-f\|^2 - \|f\|^2$. Then we can define

$$\hat{f}_m = \arg\min_{t \in S_m} \tilde{\gamma}_n(t)$$

and $\tilde{m} = \arg\min \tilde{\gamma}_n(\hat{f}_m) + \widetilde{\text{pen}}(m)$. As previously, $\widetilde{\text{pen}}(.)$ is a penalty function that ensures the squared-bias variance compromise, it is given by

$$\widetilde{\text{pen}}(m) = \kappa \left(\frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i (1 - \hat{F}_n(Z_i))^2}{\Phi_n^2(Z_i)} \right) \frac{m}{n}. \tag{21}$$

We also have, $f_m = \sum_{j=0}^{m-1} b_j \psi_j$ with $b_j = \langle f, \psi_j \rangle$, and $\hat{f}_m = \sum_{j=0}^{m-1} \hat{b}_j \psi_j$, with $\hat{b}_j = (1/n)\sum_{i=1}^{n} \delta_i (1 - \hat{F}_n(Z_i))\psi_j(Z_i)/\Phi_n(Z_i)$. Moreover $\tilde{\gamma}_n(\hat{f}_m) = -\sum_{j=1}^{m-1} \hat{b}_j^2$.

3.4. **MISE bounds for the adaptive estimators of the hazard rate and the density.** The study of the estimator relies on the following decompositions. For λ, we have

$$\gamma_n(t) - \gamma_n(s) = \|t - \lambda\|^2 - \|s - \lambda\|^2 - 2\nu_n(t-s) - 2R_n(t-s)$$

where

$$\nu_n(t) = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\delta_i t(Z_i)}{\Phi(Z_i)} - \langle t, \lambda \rangle \right], \quad R_n(t) = \frac{1}{n} \sum_{i=1}^{n} \delta_i t(Z_i) \frac{\Phi(Z_i) - \Phi_n(Z_i)}{\Phi(Z_i)\Phi_n(Z_i)}.$$

For f, we have

$$\tilde{\gamma}_n(t) - \tilde{\gamma}_n(s) = \|t - f\|^2 - \|s - f\|^2 - 2\tilde{\nu}_n(t-s) - 2R_{n,1}(t-s) - 2R_{n,2}(t-s) \tag{22}$$

where $\tilde{\nu}_n$ is a centered empirical process

$$\tilde{\nu}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\delta_i t(Z_i) \bar{F}(Z_i)}{\Phi(Z_i)} - \langle t, f \rangle \right]$$

and $R_{n,1}$ and $R_{n,2}$ are residual terms:
(23)

$$R_{n,1}(t) = \frac{1}{n}\sum_{i=1}^{n}\delta_i t(Z_i)\frac{F(Z_i)-\hat{F}_n(Z_i)}{\Phi_n(Z_i)},\ \ R_{n,2}(t) = \frac{1}{n}\sum_{i=1}^{n}\delta_i t(Z_i)\bar{F}(Z_i)\frac{\Phi(Z_i)-\Phi_n(Z_i)}{\Phi_n(Z_i)\Phi(Z_i)}.$$

By proving that the residual terms are negligible and applying Talagrand (1996) inequality to ν_n or $\tilde{\nu}_n$, we obtain the following results.

Theorem 3.1. *Consider the estimators $\hat{\lambda}_{\hat{m}}$ and $\hat{f}_{\tilde{m}}$ defined in Section 3.3.1 and 3.3.2 and the collection of models $(S_m)_{m\in\mathcal{M}_n}$ with $N_n \le c/\sqrt{n}$. Assume that $f(x) \le f_1$ and that $w_1 \le w(x) \le w_2$ for all $x \in A$, then*

$$\mathbb{E}(\|\hat{g}_{\hat{m}} - g\|^2) \le C_g \inf_{m\in\{1,\dots,n\}}\left(\|g-g_m\|^2 + \frac{m}{n}\right) + \frac{C_g'\sqrt{\ln(n)}}{n}, \tag{24}$$

where C_g and C_g' are constant depending on f, w, h and g, g_m and $\hat{g}_{\hat{m}}$ stands successively for λ, λ_m and $\hat{\lambda}_{\hat{m}}$ and for f, f_m and $\hat{f}_{\tilde{m}}$.

Theorem 3.1 states that the adaptive estimator automatically realizes a compromises between the squared-bias $\|f - f_m\|^2$ (or $\|\lambda - \lambda_m\|^2$) and the variance order m/n, given by the penalties.

Moreover, when the function g (i.e. λ or f) has regularity α (for instance belongs to a Besov space on the interval with indexes α, ∞) then it is well known that $\|g - g_m\|$ is of order $m^{-\alpha}$. Therefore, the automatic optimization performed in $\inf_{m\in\{1,\dots,n\}}(\|g-g_m\|^2 + m/n)$ yields an order $n^{-2\alpha/(2\alpha+1)}$. This is the standard optimal rate in the minimax sense for hazard or density estimation.

4. Simulations

4.1. **Simulated sampling pattern.** We consider the time-window study with fixed values t_1 and t_2, for the implementation of the algorithm. We have to build a n-sample of $(\sigma_{\mathcal{S}}, Z, \delta)$ and for this purpose, the steps are the following:

1) Given t_1, $t_2 > 0$ with $t_1 < t_2$, draw N birth-times $(\sigma_i)_{i=1,\dots,N}$ on $\mathbb{R}^+$ as realizations of a Poisson process with given intensity $\varphi(s) \propto 1$ (homogeneous Poisson process) which corresponds to unbiased case (denoted by UB hereafter) or linear bias (denoted by LB) or $\varphi(s) \propto s^{-2/3}$ (inhomogeneous Poisson process) resulting in a power bias (denoted by PB) $w(t) \propto t^{1/3}$. For the simulations of inhomogeneous Poisson process, we refer to Devroye (1986). Moreover, the intensity φ is "calibrated" in function of the values t_1 and t_2 such that the generated birth-times satisfy both conditions: with probability near of 1, at least n lifetimes X_i's fall in the time-window $[t_1, t_2]$ and the largest simulated birth-time σ_N would exceed t_2. Note that $N >> n$, but for example, $N = 1000$ ensure that $n = 200$ biased lifetimes will be included in the study after step 4).

2) For each birth-time σ_i, draw the associated lifetime X_i with given c.d.f F.

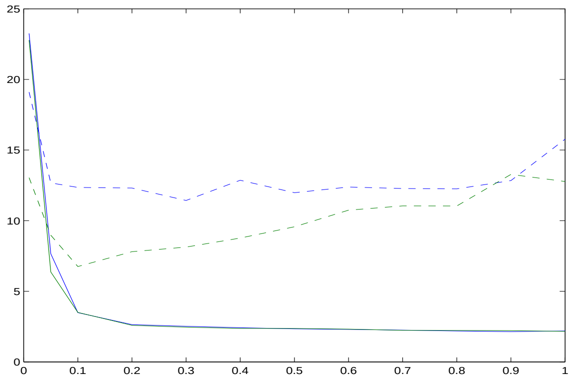


FIGURE 4.1. Choice of κ: the dimension (solid line) and the MSE2$*10^4$ (dashed line) in function of κ, the two curves correspond to unbiased and power-biased cases.

3) Inclusion in the study: the age at inclusion ξ_i in the study is given by $\xi_i = t_1 - \sigma_i$ if $\sigma_i \le t_1$ and $\xi_i = 0$ if $t_1 \le \sigma_i \le t_2$. So, choose the $X_i = X_{\mathcal{S},i}$ such that $\xi_i \le t_1$ and $X_i \ge \xi_i$. This mechanism of inclusion in the study involves a bias selection and we keep only selected couples $(\xi_i, X_{\mathcal{S},i})$.

4) Draw a n-sample $(\sigma_{\mathcal{S},1}, X_{\mathcal{S},1}), \dots, (X_{\mathcal{S},n}, \sigma_{\mathcal{S},n})$ by uniform random sampling among all the individuals falling in the time-window and selected in step 3).

5) Draw the censoring variables $C_1, \dots, C_n$ with exponential c.d.f. $\mathcal{E}xp(c)$ for several values of $c > 0$ involving different censoring proportions.

6) Put $Z_i = (t_1 - \sigma_{\mathcal{S},i}) + (X_{\mathcal{S},i} - (t_1 - \sigma_{\mathcal{S},i})) \wedge C_i$ if $\sigma_{\mathcal{S},i} \le t_1$ and $Z_i = X_{\mathcal{S},i} \wedge C_i$ if $t_1 \le \sigma_{\mathcal{S},i} \le t_2$.

Note that in the UB case, the step 3) is skipped.

4.2. **Practical implementation of the estimators.** Two models are investigated as in Antoniadis *et al.* (1999), Reynaud-Bouret (2006) and Brunel & Comte (2005):

(a) The first example is the *Gamma case*: the X_i's are generated following a Gamma distribution with shape parameter 5 and scale 1. The hazard rate is a monotone curve.

(b) The second data set is called the *Bimodal case*: The X_i's are distributed from a bimodal density defined by $f = 0.8u + 0.2v$ where u is the p.d.f of $\exp(Z/2)$ with $Z \sim \mathcal{N}(0, 1)$ and $v = 0.17Z + 2$.

The expression of the hazard estimator on $A = [a, b]$ is the following:

$$\hat{\lambda}_m(x) = \frac{\hat{a}_0}{\sqrt{b-a}} + \sqrt{\frac{2}{b-a}} \left(\sum_{j=1}^{[(m-1)/2]} \hat{a}_{2j-1} \sin(\frac{2\pi j(x-a)}{b-a}) + \sum_{j=1}^{[m/2]} \hat{a}_{2j} \cos(\frac{2\pi j(x-a)}{b-a}) \right)$$

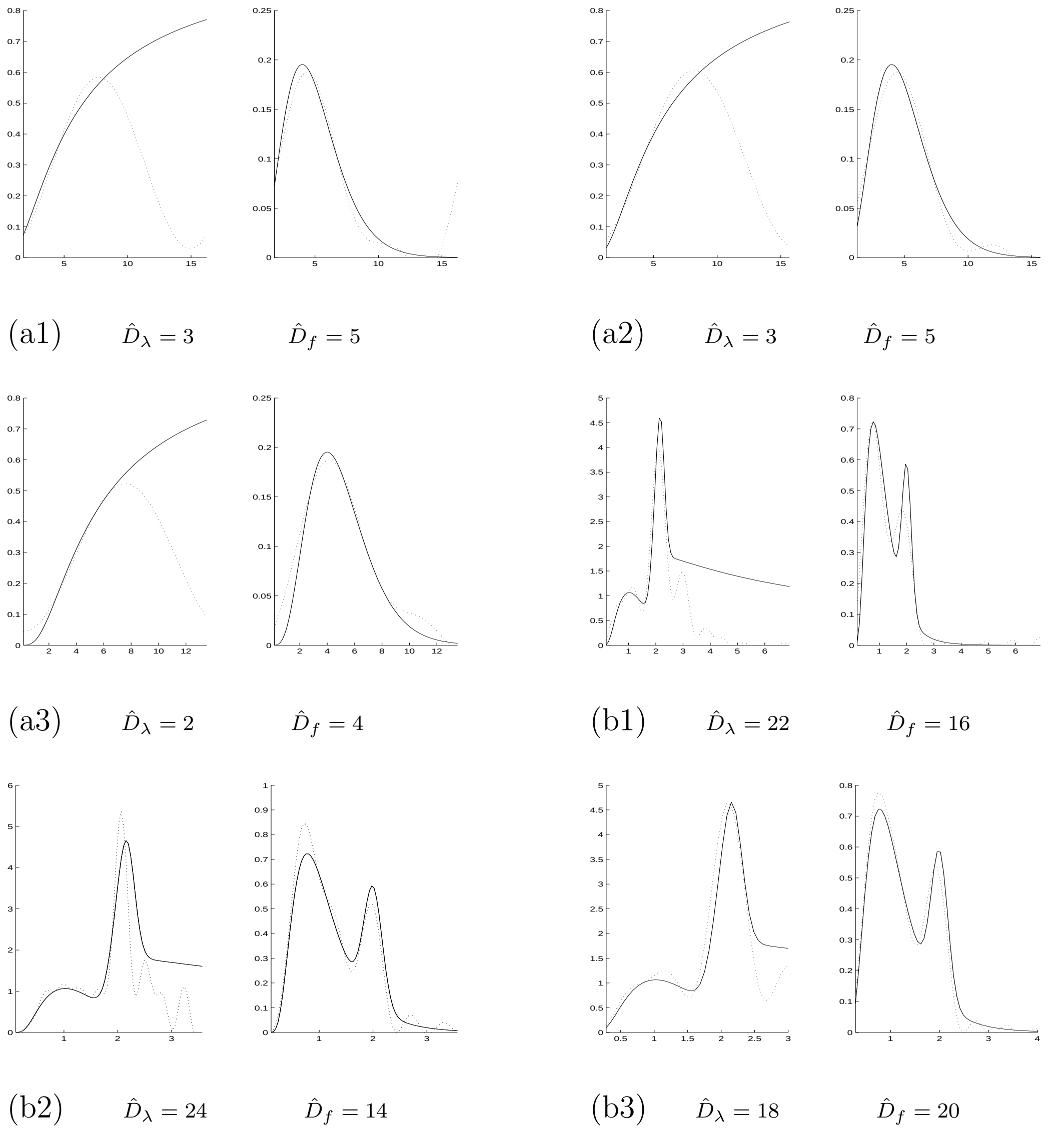


FIGURE 4.2. Penalized estimators of λ and f with power bias. Model (a): (a1) $n = 200$, no censure; (a2) $n = 500$, no censure; (a3) $n = 500$, censure 38%. Model (b): (b1) $n = 500$, no censure; (b2) $n = 500$, censure 33%; (b3) $n = 1000$, no censure. For each picture, on the left, the estimator of λ and on the right the estimator of f, $\hat{D}_\lambda$ and $\hat{D}_f$ denote the selected dimension for λ and f respectively. True curve (full) and estimate (dotted).

n		Gamma				Bimodal			
		200		500		200		500	
Censoring		0%	40-50%	0%	40-50%	0%	40-50%	0%	40-50%
UB	MSE	0.11	0.11	0.12	0.13	1.65	1.80	1.18	1.50
	MSE2	0.0018	0.004	0.011	0.029	0.19	0.17	0.11	0.08
LB	MSE	0.12	0.11	0.13	0.12	1.45	1.50	1.11	1.27
	MSE2	0.0014	0.0017	0.0008	0.0013	0.15	0.09	0.08	0.04
PB	MSE	0.12	0.101	0.13	0.10	1.37	1.32	1.02	1.07
	MSE2	0.0012	0.0014	0.0007	0.0009	0.17	0.05	0.08	0.02

TABLE 4.1. Monte-Carlo results for the estimator of λ, for $J = 200$ replications.

n		Gamma				Bimodal			
		200		500		200		500	
Censoring		0%	40-50%	0%	40-50%	0%	40-50%	0%	40-50%
UB	MSE	0.0008	0.0011	0.0004	0.0008	0.02	0.02	0.01	0.01
	MSE2	0.0005	0.0003	0.0003	0.0002	0.01	0.01	0.008	0.004
LB	MSE	0.0005	0.0005	0.0002	0.0003	0.01	0.01	0.005	0.003
	MSE2	0.0002	0.0003	0.0001	0.0002	0.007	0.008	0.003	0.002
PB	MSE	0.0006	0.0005	0.0003	0.0004	0.01	0.009	0.005	0.003
	MSE2	0.0002	0.0003	0.0001	0.0002	0.006	0.006	0.003	0.002

TABLE 4.2. Monte-Carlo results for the estimator of f, for $J = 200$ replications

with $\hat{a}_0 = \sum_{i=1}^{n} \delta_i/n$ and for $\tilde{Z}_i = (Z_i - a)/(b - a)$,

$$\hat{a}_{2j} = \frac{\sqrt{2}}{n\sqrt{b-a}} \sum_{i=1}^{n} \frac{\delta_i \cos(2\pi j \tilde{Z}_i)}{\Phi_n(\tilde{Z}_i)}, \hat{a}_{2j+1} = \frac{\sqrt{2}}{n\sqrt{b-a}} \sum_{i=1}^{n} \frac{\delta_i \sin(2\pi j \tilde{Z}_i)}{\Phi_n(\tilde{Z}_i)}.$$

The optimal dimension $\hat{m}$ is chosen to minimize the penalized contrast $-\sum_{j=0}^{m-1} \hat{a}_j^2 + \kappa\,(n^{-1}\sum_{i=1}^{n} \delta_i/\Phi_n^2(Z_i))\, m/n$. Here, the universal constant has been approximated by simulation experiments as shown in Figure 4.1 and taken as $\kappa = 0.15$. We also took a modified version of the estimator $1/\Phi_n(x) = \max(1.5; \min(1/(1 + (O(x) - n)/(n + 5)); \sqrt{n}))$ for numerical reasons.

Plots of both estimators are given in Figure 4.2. The estimators are only illustrated in the PB case since Tables 4.1 and 4.2 show that the estimation errors are quite the same whatever the form of the bias is. As expected, the behaviour of the hazard estimator at the end of the interval is erratic due to the sparsity of the observations there. This problem vanishes when considering the density estimator. Note that the dimensions selected by the algorithm for both λ and f are much smaller in the Gamma case (a) than in the Bimodal case (b). This reflects that the algorithm makes automatically the squared Bias/Variance compromise and is able to adapt to various forms of curves. Now, to study the quality of the estimation procedure, we compute over J replications of samples of size $n = 200$ and $n = 500$ the mean squared errors (MSE) over a grid of $K = 64$ regularly spaced points $t_1, \dots, t_K$ of $[a, b]$:

$$\mathrm{MSE}_j = \frac{1}{n}\sum_{k=1}^{K}[g(t_k) - \hat{g}_{\hat{m}_j}(t_k)]^2$$

where $\hat{g}_{\hat{m}_j}(t_k)$ is the penalized estimator of λ or f computed for the jth replication for $j = 1, \dots, J$. Then, the MSE given in Table 4.1 is the arithmetic mean of the J MSE_j. In order to take into account the sparsity of the observations at the end of the interval, ($\mathbb{P}(X > 6) = 0.25$ in the Gamma case and $\mathbb{P}(X > 2) = 0.16$ in the Bimodal case), we also compute an error MSE2 defined by the same kind of mean squared error but with a truncated mean over the t_k's less than 6 in the Gamma case and 2 in the Bimodal case. In the unbiased case with censoring, we get very conforming results to those in Brunel & Comte (2005) so that we can consider the estimator as a generalized version of the estimator therein. We can also compare our results to those in Brunel *et al.* (2005) for density estimation in presence of bias but without censoring. Indeed, the Gamma case is also studied therein and the MSE's up to the interval length factor ($b-a$ is about 15 here) are of the same order. The MSE2's for the hazard estimators are really better than the full MSE's (from 10 to 100 times smaller) which is not very surprising since the estimation breaks up at the end of the interval. This improvement still occurs for the density but in a less striking way. Besides, the censoring effect for both density and hazard does not seem to degrade the estimation error. We can even notice that for hazard estimation in the Bimodal case, the censoring (with significant rate about $40 - 50\%$) systematically seems to improve the MSE2's.

5. Proofs

5.1. **Proof of Lemma 3.1.** *Proof of 1).* The result follows by simply integrating the deviation given in Theorem 2.1.

$$\mathbb{E}\left[\left(\sup_{x\in[0,1]} |\hat{F}_n(x) - F(x)|\right)^{2k}\right] = 2k\int_0^{+\infty} u^{2k-1}\,\mathbb{P}(\sup_{x\in[0,1]} |\hat{F}_n(x) - F(x)| > u)\,du$$

$$\leq 2k\int_0^{+\infty} u^{2k-1}\,\mathbb{P}(\sqrt{n}\|(1-H)^2\,(\hat{F}_n - F)w_1\|_\infty > c_H^2 w_1\sqrt{n}\,u)\,du$$

$$\leq 5ke^{C^2/8}\int_0^{\infty} u^{2k-1}\exp\left(-2(c_H^2w_1)^2n\left[u-\frac{C}{4\sqrt{n}c_H^2w_1}\right]^2\right)du$$

$$\leq \frac{5e^{C^2/8}k}{2^k(c_H^2w_1)^{2k}}\left(\int_{-C/(2\sqrt{2})}^{+\infty}\left(z+\frac{1}{2\sqrt{2}}\right)^{2k-1}e^{-z^2}dz\right)n^{-k}=C_F(k)n^{-k}.$$

Proof of 2). First, using the definition of Φ given by (15) and the expression of the density probability $f_{\sigma_{\mathcal{S}},X_{\mathcal{S}}}$ of $(\sigma_{\mathcal{S}},X_{\mathcal{S}})$ deduced from (3), we write that

$$\begin{aligned}\Phi(x) &= \iiint \mathbf{1}_{\{a_{\mathcal{S}}(\sigma)\leq x\}}\mathbf{1}_{\{y\geq x\}}\mathbf{1}_{\{x\leq a_{\mathcal{S}}(\sigma)+c\}}f_{\sigma,X_{\mathcal{S}}}(\sigma,y)h(c)d\sigma dy dc\\ &= \frac{1}{\mu_{\mathcal{S}}}\int \mathbf{1}_{\{y\geq x\}}f(y)\left[\int h(c)\left(\int \mathbf{1}_{\{x-c\leq a_{\mathcal{S}}(\sigma)\leq x\}}\varphi(\sigma)d\sigma\right)dc\right]dy.\end{aligned}$$

By reminding that $w(x)=\int \mathbf{1}_{\{a_{\mathcal{S}}(u)\leq x\}}\varphi(u)du$, we get

$$\int \mathbf{1}_{\{x-c\leq a_{\mathcal{S}}(\sigma)\leq x\}}\varphi(\sigma)d\sigma=\begin{cases} w(x)-w(x-c) & \text{if } x-c>0\\ w(x) & \text{otherwise,}\end{cases}$$

since $w(0^-)=0$. It follows that

$$\begin{aligned}\Phi(x) &= \frac{\bar{F}(x)}{\mu_{\mathcal{S}}}\left[\int_0^x h(c)(w(x)-w(x-c))dc+\int_x^{+\infty}h(c)w(x)dc\right]\\ &= \frac{\bar{F}(x)}{\mu_{\mathcal{S}}}\left(w(x)-\int_0^x h(c)w(x-c)dc\right). \qquad (25)\end{aligned}$$

For $x\geq s$, we have $w(x)-w(s)=\int \mathbf{1}_{\{s<a_{\mathcal{S}}(\sigma)\leq t\}}\varphi(\sigma)d\sigma\geq 0$. Thus, for c in $[0,x]$, we have $w(x-c)\leq w(x)$ and

$$\Phi(x)\geq \frac{\bar{F}(x)}{\mu_{\mathcal{S}}}w(x)\left(1-\int_0^x h(c)dc\right)=\frac{\bar{F}(x)}{\mu_{\mathcal{S}}}w(x)\bar{H}(x).$$

This implies that $\forall x\in A,\ \Phi(x)\geq c_Fc_Hw_1/\mu_{\mathcal{S}}:=c_\Phi$.
Proof of 3). We just have to write that $\Phi(x)=\mathbb{P}(a_{\mathcal{S}}(\sigma_{\mathcal{S},1})\leq x\leq Z_1)$ is a difference of two distribution functions $\Phi(x)=\mathbb{P}(a_{\mathcal{S}}(\sigma_{\mathcal{S},1})\leq x)-\mathbb{P}(Z_1\leq x)$. Then we can use the exponential deviation inequality between theoretical and empirical distribution functions for i.i.d. random variables, as given in Massart (1990) (which gives that $\forall\lambda>0,\mathbb{P}(\sqrt{n}\|F_{n,Z}-F_Z\|_\infty\geq\lambda)\leq 2e^{-2\lambda^2}$, where $F_{n,Z}(x)=(1/n)\sum_{i=1}^n \mathbf{1}_{\{Z_i\leq x\}}$ and $F_Z(x)=\mathbb{P}(Z_1\leq x)$). Then we can integrate the resulting exponential probabilities as above (see the proof of 1)).

5.2. **Proof of Proposition 3.1.** Using equations (6) and (3), we obtain:

$$\begin{aligned}\mathbb{E}\left(\frac{\delta_1 t(Z_1)}{\Phi(Z_1)}\right) &= \mathbb{E}\left(\mathbf{1}_{\{X_{\mathcal{S},1}-a_{\mathcal{S}}(\sigma_{\mathcal{S},1})\le C_1\}}\frac{t(X_{\mathcal{S},1})}{\Phi(X_{\mathcal{S},1})}\right)\\ &= \iiint \mathbf{1}_{\{x-a_{\mathcal{S}}(\sigma)\le c\}}\mathbf{1}_{\{a_{\mathcal{S}}(\sigma)\le x\}}\frac{t(x)}{\Phi(x)}h(c)\varphi(\sigma)\frac{f(x)}{\mu_{\mathcal{S}}}dc\,d\sigma\,dx\\ &= \int \frac{t(x)f(x)}{\mu_{\mathcal{S}}\Phi(x)}\left[\int h(c)\left(\int \mathbf{1}_{\{x-c\le a_{\mathcal{S}}(\sigma)\le x\}}\varphi(\sigma)d\sigma\right)dc\right]dx.\end{aligned}$$

Then it follows from (25) that

$$\int h(c)\left(\int \mathbf{1}_{\{x-c\le a_{\mathcal{S}}(\sigma)\le x\}}\varphi(\sigma)d\sigma\right)dc = \frac{\Phi(x)}{\bar F(x)}\mu_{\mathcal{S}}.$$

Thus

$$\mathbb{E}\left(\frac{\delta_1 t(Z_1)}{\Phi(Z_1)}\right) = \int \frac{t(x)f(x)}{\Phi(x)\mu_{\mathcal{S}}}\frac{\Phi(x)\mu_{\mathcal{S}}}{\bar F(x)}dx = \int t(x)\frac{f(x)}{1-F(x)}dx = \langle t,\lambda\rangle.$$

5.3. **Proof of Theorem 3.1.** For the sake of brevity, we prove the result only in the more complicated case, that is for the estimation of f. The proof for the estimation of λ would follow the same line.

For the sake of simplicity, we also prove the result with a deterministic penalty function,

$$\widetilde{\text{pen}}(m) = \kappa\mathbb{E}\left(\frac{\delta_1\bar F^2(Z_1)}{\Phi^2(Z_1)}\right)\frac{m}{n}.$$

Going from the above penalty to the random one given in the definitions of the estimators is now standard (see Comte and Brunel (2005, p.465)).

Step 1: Decomposition of the $\mathbb{L}^2$ *risk.* Let $B_{m,m'}(0,1)=\{t\in S_m+S_{m'},\ \|t\|=1\}$. We deduce from (22) and the definition of $\hat f_{\tilde m}$ that

$$\begin{aligned}\|\hat f_{\tilde m}-f\|^2 &\le \|f-f_m\|^2+2\tilde\nu_n(\hat f_{\tilde m}-f_m)+2R_{n,1}(\hat f_{\tilde m}-f_m)+2R_{n,2}(\hat f_{\tilde m}-f_m)\\ &\quad +\widetilde{\text{pen}}(m)-\widetilde{\text{pen}}(\tilde m)\\ &\le \|f-f_m\|^2+2\|\hat f_{\tilde m}-f_m\|\left(\sup_{t\in B_{m,\tilde m}(0,1)}(|\tilde\nu_n(t)|+|R_{n,1}(t)|+|R_{n,2}(t)|)\right)\\ &\quad +\widetilde{\text{pen}}(m)-\widetilde{\text{pen}}(\tilde m)\end{aligned}$$

We use that for $x, y \geq 0$, $2xy \leq (1/4)x^2 + 4y^2$ and $(x+y+z)^2 \leq 2x^2 + 4y^2 + 4z^2$ and we obtain

$$\begin{aligned}
\|\hat{f}_{\tilde{m}} - f\|^2 &\leq \|f - f_m\|^2 + \frac{1}{4}\|\hat{f}_{\tilde{m}} - f_m\|^2 + 8 \sup_{t \in B_{m,\tilde{m}}(0,1)} \tilde{\nu}_n^2(t) + 16 \sup_{t \in B_{m,\tilde{m}}(0,1)} R_{n,1}^2(t) \\
&\quad + 16 \sup_{t \in B_{m,\tilde{m}}(0,1)} R_{n,2}^2(t) + \widetilde{\text{pen}}(m) - \widetilde{\text{pen}}(\tilde{m}) \\
&\leq \|f - f_m\|^2 + \frac{1}{4}\|\hat{f}_{\tilde{m}} - f_m\|^2 + 8 \left(\sup_{t \in B_{m,\tilde{m}}(0,1)} (\tilde{\nu}_n)^2(t) - p(m, \tilde{m}) \right)_+ \\
&\quad + 16 \sup_{t \in B_{m,\tilde{m}}(0,1)} R_{n,1}^2(t) + 16 \sup_{t \in B_{m,\tilde{m}}(0,1)} R_{n,2}^2(t) \\
&\quad + 8p(m, \tilde{m}) + \widetilde{\text{pen}}(m) - \widetilde{\text{pen}}(\tilde{m})
\end{aligned}$$

Next $p(m, m')$ is chosen such that

$$\mathbb{E}\left(\sup_{t \in B_{m,\tilde{m}}(0,1)} \tilde{\nu}_n^2(t) - p(m, \tilde{m}) \right)_+ \leq \frac{C}{n}, \tag{26}$$

and the penalty function $\widetilde{\text{pen}}(m)$ is deduced from $p(.,.)$ by setting the constraint

$$\forall m, m' \in \mathcal{M}_n, \ 8p(m, m') \leq \widetilde{\text{pen}}(m) + \widetilde{\text{pen}}(m'). \tag{27}$$

Step 2: Use of Talagrand (1996)'s Inequality. From the linear centered empirical process $\tilde{\nu}_n(t)$, we compute the following constants in view of applying Talagrand's Inequality. The connection between the norms stated in (14) is very useful in all the following.

$$\begin{aligned}
\mathbb{E}\left(\sup_{t \in B_{m,m'}(0,1)} \tilde{\nu}_n^2(t) \right) &\leq \frac{1}{n} \sum_{j=0}^{m \vee m' - 1} \text{Var}\left(\frac{\psi_j(X_{\mathcal{S},1})\bar{F}(X_{\mathcal{S},1})}{\Phi(X_{\mathcal{S},1})} \right) \\
&\leq \frac{1}{n}\mathbb{E}\left(\frac{\sum_{j=0}^{m \vee m' - 1} \psi_j^2(X_{\mathcal{S},1})\bar{F}^2(X_{\mathcal{S},1})}{\Phi^2(X_{\mathcal{S},1})} \right) \leq \frac{2(m \vee m')}{n} \sup_{x \in A} \frac{\bar{F}^2(x)}{\Phi^2(x)} \qquad (28) \\
&\leq \frac{2\mu_{\mathcal{S}}^2}{w_1^2 c_H^2} \frac{m \vee m'}{n} := H^2. \qquad (29)
\end{aligned}$$

Indeed, it follows from (25) that $\Phi(t)/\bar{F}(t) \geq w(t)\bar{H}(t)/\mu_{\mathcal{S}}$.

$$\begin{aligned}
\sup_{t \in B_{m,m'}(0,1)} \text{Var}\left(\frac{t(X_{\mathcal{S},1})\bar{F}(X_{\mathcal{S},1})}{\Phi(X_{\mathcal{S},1})} \right) &\leq \sup_{t \in B_{m,m'}} \int \frac{t^2(x)\bar{F}^2(x)}{\Phi^2(x)} \frac{w(x)f(x)}{\mu_{\mathcal{S}}} dx \\
&\leq \frac{\mu_{\mathcal{S}} w_2 f_1}{c_H^2 w_1^2} \sup_{t \in B_{m,m'}} \int_A t^2(x)dx \leq \frac{\mu_{\mathcal{S}} w_2 f_1}{c_H^2 w_1^2} := v \qquad (30)
\end{aligned}$$

$$\sup_{t\in B_{m,m'}(0,1)} \sup_{x\in A} \left|\frac{t(x)\bar{F}(x)}{\Phi(x)}\right| \leq \frac{\mu_{\mathcal{S}}}{w_1 c_H}\sqrt{2(m\vee m')} := M_1. \tag{31}$$

Thus, we obtain, with $C(\epsilon) = (\sqrt{1+\epsilon}-1)\wedge 1$ and K_1 standing for numerical constant,

$$\mathbb{E}\left(\sup_{t\in B_{m,m'}(0,1)} \tilde{\nu}_n^2(t) - 2(1+2\epsilon)H^2\right)_+ \leq \frac{6}{K_1}\left(\frac{v}{n}e^{-K_1\epsilon nH^2/v} + \frac{8M_1^2}{K_1 n^2 C^2(\epsilon)} e^{-\frac{K_1 C(\epsilon)\sqrt{\epsilon}}{\sqrt{2}}\frac{nH}{M_1}}\right)$$

Now, by taking $\epsilon = 1/2$ and with the values of H^2, v and M_1 computed above, we can write the resulting upper bound in this context, for all m, m'

$$\mathbb{E}\left(\sup_{t\in B_{m,m'}(0,1)} \tilde{\nu}_n^2(t) - 4H^2\right)_+ \leq \frac{6/K_1}{n}\left(ae^{-b(m\vee m')} + ce^{-d\sqrt{n}}\right)$$

with $a = \mu_{\mathcal{S}} w_2 f_1/(c_H^2 w_1^2)$, $b = K_1\mu_{\mathcal{S}}/(w_2 f_1)$, $c = 16\mu_{\mathcal{S}}^2/(K_1 w_1^2 c_H^2 C^2(1/2))$ and $d = K_1 C(1/2)/2$. Finally, with the choice $p(m,m') = 4H^2$

$$\mathbb{E}\left(\sup_{t\in B_{m,\tilde{m}}(0,1)} \tilde{\nu}_n^2(t) - p(m,\tilde{m})\right)_+ \leq \sum_{m'\leq n} \mathbb{E}\left(\sup_{t\in B_{m,m'}(0,1)} \tilde{\nu}_n^2(t) - p(m,m')\right)_+ \leq \frac{C}{n}.$$

Then, if (26) and (27) hold, then we have, since $\|\hat{f}_{\tilde{m}} - f_m\|^2 \leq 2\|\hat{f}_{\tilde{m}} - f\|^2 + 2\|f_m - f\|^2$,

$$\begin{aligned}\frac{1}{2}\mathbb{E}(\|\hat{f}_{\tilde{m}} - f\|^2) \;&\leq\; \frac{3}{2}\|f - f_m\|^2 + 2\widetilde{\mathrm{pen}}(m) + \frac{8C}{n}\\ &+16\mathbb{E}\left(\sup_{t\in B_{m,\tilde{m}}(0,1)} R_{n,1}^2(t)\right) + 16\mathbb{E}\left(\sup_{t\in B_{m,\tilde{m}}(0,1)} R_{n,2}^2(t)\right)\end{aligned}$$

Then, for nested model collections described in Section 3.2, $S_m + S_{m'} \subset \mathcal{S}_n$ for all m, m' implies that the supremun taken over the unit ball $B_{m,\tilde{m}}(0,1)$ is lower than the supremun over $\mathcal{B}_n(0,1) = \{t \in \mathcal{S}_n, \|t\| \leq 1\}$. As a consequence, it remains to study the residual terms involving $R_{n,1}$ and $R_{n,2}$.

Step 3: Study of the two residual terms.

Lemma 5.1. *Let $\Omega_\Phi = \{\omega, \Phi_n(x) \geq c_\Phi/2, \forall x \in [0,1]\}$. Assume that $f(x) \leq f_1$ and $w_1 \leq w(x) \leq w_2$, for all $x \in [0,1]$,*

$$L_n(\varphi) \leq N_n^2 \quad \textit{with} \;\; N_n = dim\,\mathcal{S}_n, \tag{32}$$

then, for $i = 1, 2$, we have

$$\mathbb{E}\left(\sup_{t\in\mathcal{B}_n(0,1)} R_{n,i}^2(t)\mathbf{I}_{\Omega_\Phi}\right) \leq K_{i+1}\frac{\sqrt{\ln(n)}}{n} \tag{33}$$

with $\mathcal{B}_n(0,1) = \{t \in \mathcal{S}_n, \|t\| = 1\}$ and for numerical constants $K_2 > 0$ and $K_3 > 0$.

On the complementary of Ω_Φ, we use that $\Phi_n(x) \geq 1/n$ for all i, and that $\sup_{x\in[0,1]} |\Phi_n(x) - \Phi(x)| > c_\Phi/2$, to obtain with Lemma 3.1,

$$\mathbb{E}\left(\sup_{t\in\mathcal{B}_n(0,1)} R_{n,1}^2(t)\mathbf{1}_{\Omega_\Phi^c}\right) \leq c/n.$$

Indeed, on Ω_Φ^c, we can write

$$\begin{aligned}\mathbb{E}\left(\sup_{t\in\mathcal{B}_n(0,1)} R_{n,1}^2(t)\mathbf{1}_{\Omega_\Phi^c}\right) &\leq 2n^2N_n\mathbb{E}\left(\|F-\hat{F}_n\|^2\mathbf{1}_{\Omega_\Phi^c}\right)\\ &\leq 2n^2N_n\mathbb{E}^{1/2}\left(\|F-\hat{F}_n\|^4\right)\mathbb{P}^{1/2}(\Omega_\Phi^c).\end{aligned}$$

With Lemma 3.1 1), we have $\mathbb{E}(\|F-\hat{F}_n\|_\infty^4) \leq C_F(2)/n^2$ and as $N_n \leq n$,

$$\mathbb{P}(\Omega_\Phi^c) \leq \mathbb{P}(\|\Phi_n - \Phi\|_\infty > c_\Phi/2) \leq \mathbb{E}\left[\left(\frac{2}{c_\Phi}\right)^{2k}\|\Phi_n-\Phi\|_\infty^{2k}\right] \leq \left(\frac{2}{c_\Phi}\right)^{2k}\frac{C_\Phi(k)}{n^k}$$

by using Lemma 3.1 2). Then choosing $k \geq 12$ gives the result.

The same approach gives the same order for $\mathbb{E}\left(\sup_{t\in\mathcal{B}_n(0,1)} R_{n,2}^2(t)\mathbf{1}_{\Omega_\Phi^c}\right)$.

5.4. **Proof of Lemma 5.1.**

$$\mathbb{E}\left(\sup_{t\in\mathcal{B}_n(0,1)} R_{n,1}^2(t)\mathbb{I}_{\Omega_\Phi}\right) \leq \frac{1}{c_\Phi^2}\mathbb{E}\left[\sup_{x\in[0,1]}|\hat{F}_n(x)-F(x)|^2\sup_{t\in\mathcal{B}_n(0,1)}\left(\frac{1}{n}\sum_{i=1}^n|t(X_{\mathcal{S},i})|\right)^2\right]$$

$$\leq \frac{1}{c_\Phi^2}\mathbb{E}\left\{\sup_{x\in[0,1]}|\hat{F}_n(x)-F(x)|^2\left[\sup_{t\in\mathcal{B}_n(0,1)}[|\nu_n^*(t^2)|+\mathbb{E}(t^2(X_{\mathcal{S},1}))]\right]\right\}$$

$$\leq \frac{1}{c_\Phi^2}\mathbb{E}\left\{\sup_{x\in[0,1]}|\hat{F}_n(x)-F(x)|^2\left[\sup_{t\in\mathcal{B}_n(0,1)}|\nu_n^*(t^2)|+\frac{f_1w_2}{\mu_{\mathcal{S}}}\right]\right\},$$

with $\nu_n^*(t) = \frac{1}{n}\sum_{i=1}^n[t(X_{\mathcal{S},i}) - \mathbb{E}(t(X_{\mathcal{S},i}))]$. Now, by applying Schwartz Inequality and Lemma 3.1 1) with $k=2$, we obtain

$$\mathbb{E}\left(\sup_{t\in\mathcal{B}_n(0,1)} R_{n,1}^2(t)\mathbb{I}_{\Omega_\Phi}\right) \leq \frac{C_F^{1/2}(2)}{c_\Phi^2}\frac{1}{n}\left[\mathbb{E}^{1/2}\sup_{t\in\mathcal{B}_n(0,1)}[\nu_n^*(t^2)]^2+\frac{f_1w_2}{\mu_{\mathcal{S}}}\right]$$

From Baraud (2002), we have the following exponential inequality, for all $\rho > 0$,

$$\mathbb{P}\left(\sup_{t\in\mathcal{B}_n(0,1)}|\nu_n^*(t^2)| \geq \rho\right) \leq |\Lambda_n|^2\exp\left(-\frac{n\rho^2}{4\tilde{f}_1L_n(\varphi)}\right)$$

where we denote by $\tilde{f}_1 = f_1w_2/\mu_{\mathcal{S}}$ and $L_n(\varphi)$ is a quantity associated to the orthonormal basis $\psi = (\psi_j)_{j\in\{0,\dots,N_n-1\}}$ of the largest space $\mathcal{S}_n = \bigcup_{m\in\mathcal{M}_n} S_n$ of the (nested) collection satisfying (32). Moreover, in the proof of Theorem 4.2. by Brunel & Comte (2005), it is checked that

$$\mathbb{E}\left(\sup_{t\in B_n(0,1)}[\nu_n^*(t^2)]^2\right) \leq 2\ln(n) \ \text{ for } n \geq 2,$$

as soon as $L_n(\varphi) \leq n/(16\tilde{f}_1)$. Therefore, if $N_n \leq \sqrt{n}/(4\sqrt{\tilde{f}_1})$, we easily deduce Inequality (33).

References

[1] Antoniadis, A., Grégoire, G. & Nason, G. (1999). Density and hazard rate estimation for right-censored data by using wavelet methods. *J. Roy. Statist. Soc. Ser. B* **61**, 63-84.

[2] 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.

[3] Asgharian, M. (2003). Biased sampling with right censoring: a note on Sun, Cui & Tiwari (2002). *Canad. J. Statist.* **30**, 475-490.

[4] Baraud, Y. (2002). Model selection for regression on a random design. *ESAIM Probab. Statist.* **6**, 127-146.

[5] Barron, A.R., Birgé, L. & Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields* **113**, 301-413.

[6] Barron, A.R. & Cover, T. M. (1991). Minimum complexity density estimation. *IEEE Trans. Inform. Theory* **37**, 1034-1054.

[7] Brunel, E. & Comte, F. (2005). Penalized contrast estimation of hazard rate with censored data. *Sankhya*, **67**, 441-475.

[8] Brunel, E., Comte, F. & Guilloux, A. (2005). Nonparametric density estimation in presence of bias and censoring. Preprint MAP5, 2005-13, www.math-info.univ-paris5.fr/map5/.

[9] Chakraborty, R. & Rao, C.R. (2000). Selection biases of samples and their resolutions. In *Handbook of Statistics*, Vol. 18 (P.K. Sen & C.R. Rao, eds), 675-712. Elsevier Science, North-Holland, Amsterdam.

[10] Devroye, L. (1986). Non-Uniform Random Variate Generation (originally published with Springer-Verlag, New York).

[11] Efromovich, S. (2004, a). Distribution estimation for biased data. *J. Statist. Plann. Inference* **124**, 1-43.

[12] Efromovich, S. (2004, b). Density estimation for biased data. *Ann. Statist.* **32**, no. 3, 1137-1161.

[13] Fisher, R.A. (1934). The effect of methods of ascertainment upon the estimation of frequencies. *Ann. Eugen.* **6**, 13-25.

[14] Gill, R.D., Vardi, Y. & Wellner, J.A. (1988). Large sample theory of empirical distributions in biased sampling models. *Ann. Statist.* **16**, 1069-1172.

[15] Guilloux, A. (2006) Nonparametric estimation for censored lifetimes suffering from an unknown selection bias. Preprint LSTA 2006-4. www.lsta.upmc.fr/prepublications.html.

[16] Heckman, J. (1985). Selection bias and self-selection. In *The new Palgrave: A Dictionary of Economics*, 287-296. MacMillan Press, Stockton, New York.

[17] Keiding, N. (1990). Statistical inference for the Lexis Diagram. *R. Soc. Lond. Philos. Trans. Ser. A Math. Phys. Eng. Sci.* **332**, 487-509.

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

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

[20] Lo, S.H., Mack, Y.P. & Wang, J.L. (1989). Density and hazard rate estimation for censored data via strong representation of the Kaplan-Meier estimator. *Probab. Theory Related Fields* **80**, 461-473.

[21] Lund, J. (2000). Sampling bias in population studies - How to use the Lexis diagram. *Scand. J. Statist.* **27**, 589-604.

[22] Marron, J.S. & de Uña-Àlvarez, J. (2004). SiZer for length biased, censored density and hazard estimation. *J. Statist. Plann. Inference* **121**, 149-161.

[23] Marron, J.S. & Padgett, W.J. (1987). Asymptotically optimal bandwith selection for kernel density estimators from randomly right censored samples. *Ann. Statist.* **15**, 1520-1535.

[24] Massart, P. (1990). The tight constant in Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Probab.* **18**, 1269-1283.

[25] Mielniczuk, J. (1985). Properties of some kernel estimators and of the adapted Loftsgarden-Quesenberry estimator of a density function for censored data. *Period. Math. Hungar.* **16**, 69-81

[26] Patil, G.P. & Rao, C.R., (1977). The weighted distributions : A survey of their applications. In *Applications of Statistics*, (P.R. Krishnaiah, ed.), 383-405. North-Holland Publishing Co., Amsterdam.

[27] Reynaud-Bouret, P. (2006). Penalized projection estimators of the Aalen multiplicative intensity. Forthcoming in *Bernoulli.*

[28] Talagrand, M. (1996). New concentration inequalities in product spaces. *Invent. Math.* **126**, 505-563.

[29] Tanner, M. A. & Wong, W.H. (1983). The estimation of the hazard function from randomly censored data by the kernel method. *Ann. Statist.* **11**, 989-993.

[30] de Uña-Àlvarez, J. (2002). Product-limit estimation for length-biased censored data. *Test* **11**, 109-125.

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

[32] Winter, B.B. & Földes, A. (1988). A product-limit estimator for use with length-biased data. *Canad. J. Statist.* **16**, 337-355.