Nonparametric density estimation in presence of bias and censoring.

E. Brunel[1]

IUT de Paris V, MAP5, UMR CNRS 8145. email: elodie.brunel@univ-paris5.fr

F. Comte[2]

Université Paris V, MAP5, UMR CNRS 8145. email: fabienne.comte@univ-paris5.fr.

A. Guilloux[3]

L.S.T.A., Université Paris VI. email: aguillou@ccr.jussieu.fr

Summary. We consider projection estimator methods for the nonparametric estimation of the density of i.i.d. biased observations with a general known bias function w and under right censoring. Adaptive procedures to catch the optimal estimator among a collection by contrast penalization are investigated and proved to give efficient estimators with optimal nonparametric rates of convergence.

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

Keywords. Adaptive estimation. Minimax rate. Biased data. Right-Censoring. Nonparametric penalized contrast estimator.

1. Introduction

We consider here the problem of estimating the probability density function (p.d.f.) f, with associated distribution function (c.d.f) F of a random variable (r.v.) X on a compact set $A = [0, 1]$ for simplicity. If independent and identically distributed (i.i.d.) observations $X_1, \ldots, X_n$ were available, the p.d.f. f may be estimated by usual penalized projection methods developed by Barron and Cover (1991), **?**). Here we assume that it is impossible to draw a direct sample from the distribution of X. Instead the observable r.v. is X_w with p.d.f. given, for all $x > 0$, by :

$$f_w(x) = \frac{w(x)f(x)}{\mu} \text{ where } \mu = \int w(u)f(u)du. \tag{1}$$

In this case, the r.v. X_w 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 $w(x)$.

The relevance of selection bias in statistical inference has been firstly pointed out by Fisher (1934), who suggested that it is an important source of error in estimation. Since then, many authors noticed its presence in data from a wide range of fields. We refer to Cox (1969) for industrial applications, Chakraborty and Rao (2000) for biomedical applications, and Heckman (1985) in Economics, among many others. The review by Patil and 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 $X_{w,1}, \ldots, X_{w,n}$ has received a lot of attention. We refer to Gill *et al.* (1988) and Efromovich (2004, a) 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).

As claimed above, we aim to estimate the p.d.f f from a selection-biased sample. As far as we know, only Efromovich (2004, a) deals with that problem. He considers a minimax estimation procedure with a known weight function w and states asymptotic results for several estimators. More precisely, in Efromovich (2004, b) the c.d.f. is estimated via projection on trigonometric polynomial spaces and therefore the density estimator is obtained by derivation. The results are obtained for a class of analytic c.d.f. (which are infinitely differentiable functions) and in this context, the standard minimax rate of $O(\ln(n)/n)$ is achieved. The framework of Efromovich (2004, a) is more general and closer to the results presented here. He considers a blockwise shrinkage estimator, based on a projection estimator on trigonometric bases again and with a threshold. Two types of results are obtained. First, global minimax standard results with an asymptotic risk of order $O(n^{-2m/2m+1})$ for a Sobolev class of m-differentiable density functions are proved; secondly, he considers a local approach by introducing a fixed density called the anchor density, not necessary a Sobolev one, and he defines a class of functions uniformly close to it. The local estimator is based on trigonometric expansions of both the anchor density and a Sobolev function t whose uniform norm is arbitrarily small. Then, this local approach achieves superefficiency rates of convergence for a class of densities in the neighborhood of the anchor density i.e. a logarithmic factor is gained on.

In what follows, also assuming that the bias function w is known, we construct and study adaptive estimators of f in the exact case described above but also in presence of a random right censorship. In the present work, the general techniques of model selection referenced above are applied in a particular framework related to "loss" of information. The estimation of the density function for censored data has been widely studied and an important literature deals with nonparametric methods including Tanner and Wong (1983), Mielniczuk (1985), Marron and 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 introduces a suitable correction of the Kaplan-Meier estimator for taking the length bias problem into account. The kernel density (and hazard) estimator is directly obtained by convolution of a kernel function with this corrected estimator of the survival function see Marron and 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.

The paper is organized as follows; in Section 2, the projection spaces with their key properties are described and also, the class of regularity of the density functions. Then, the penalized projection density estimator is defined in Section 3 by correcting the usual density contrast in order to take into account the biasing problem and its quadratic risk is evaluated. It follows from Section 2 that it is adaptive to the unknown smoothness of the function. A Monte-Carlo study shows the performance of the penalized estimator for various form of the underlying density and the rate of convergence for complete sample and biased sample are compared. But the main contribution of this paper stands in Section 4, where the previous results are extended to both biased and censored data. The proofs are postponed in the final section.

2. Framework

2.1. Projection bases and their fundamental properties

The projection spaces $(S_m)_{m\in\mathcal{M}_n}$ described hereafter are finite-dimensional linear subspaces of $\mathbb{L}^2([0,1])$, with dimension $\dim(S_m) = D_m$ such that $D_m \leq n$, $\forall m \in \mathcal{M}_n$ and $\mathcal{M}_n$ is the set of all possible m. An orthonormal basis of S_m is denoted by $(\varphi_\lambda)_{\lambda\in\Lambda_m}$ where $|\Lambda_m| = D_m$. Without loss of generality, we set $A = [0,1]$ for the simplicity of the description of the bases which are of the same form up to a scaling change.

[T] *Trigonometric spaces*: $\varphi_0 = 1$, $\varphi_j(x) = \sqrt{2}\cos(2\pi jx)$ and $\varphi_{j+m+1}(x) = \sqrt{2}\sin(2\pi jx)$ for $j = 1,\dots,m$ }, $D_m = 2m+1$ and $\mathcal{M}_n = \{1,\dots,[n/2]-1\}$.

[P] *Regular piecewise polynomial spaces*: $(r+1)$ polynomials of degree $k = 0,1,\dots,r$ on each

subinterval $[(j-1)/m, j/m]$, for $j = 1, \dots m$, $D_m = (r+1)m$, $m \in \mathcal{M}_n = \{1, 2, \dots, [n/(r+1)]\}$. Usual examples are the orthogonal collection in $\mathbb{L}^2([-1,1])$ of Legendre polynomials or histogram basis (with $r = 0$).
[DP] *Dyadic piecewise polynomial spaces* : dyadic instead of uniform subdivisions are needed for adaptive results with $m = 2^q$ and $D_m = (r+1)\, 2^q$.
[W] *Dyadic wavelet generated spaces* with regularity r and compact support, as described e.g. in Donoho and Johnstone (1998). The generating basis is of cardinality $D_m = 2^{m+1}$ and $m \in \mathcal{M}_n = \{1, 2, \dots, [\ln(n)/2] - 1\}$.

What we often refer to the *norm connection* property of the projection spaces is the basic tool to obtain the adequate order of the risk bound and in many adaptation results. All the spaces $(S_m)_{m\in\mathcal{M}_n}$ with dimension D_m $(\forall n, D_m \le n)$ will satisfy this property formulated as follows,

$(\mathcal{H}_1)$

$$\exists \Phi_0 > 0, \forall m \in \mathcal{M}_n, \forall t \in S_m, \|t\|_\infty \le \Phi_0 \sqrt{D_m}\|t\|. \tag{2}$$

where $\|t\|^2 = \int_0^1 t^2(x)dx$, for t in $\mathbb{L}^2([0,1])$.

It follows from Lemma 1 in Birgé and Massart (1998), that Property (2) in the context of $(\mathcal{H}_1)$ is equivalent to

$$\exists \Phi_0 > 0, \|\sum_{\lambda\in\Lambda_m} \varphi_\lambda^2\|_\infty \le \Phi_0^2 D_m. \tag{3}$$

Moreover, for the results concerning the adaptive estimators, we need the following additional assumption:

$(\mathcal{H}_2)$ $(S_m)_{m\in\mathcal{M}_n}$ is a collection of nested models, we denote by $\mathcal{S}_n$ the space belonging to the collection, such that $\forall m \in \mathcal{M}_n, S_m \subset \mathcal{S}_n$. We denote by N_n the dimension of $\mathcal{S}_n$: $\dim(\mathcal{S}_n) = N_n$ $(\forall m \in \mathcal{M}_n, D_m \le N_n)$.

All the described spaces satisfy $(\mathcal{H}_1)$, with for instance $\Phi_0 = \sqrt{2}$ for collection [T] and $\Phi_0 = \sqrt{2r+1}$ for collection [P]. Moreover, [T], [DP] and [W] satisfy $(\mathcal{H}_2)$.
Moreover, the following balls are of interest in the proofs $B_m(0,1) = \{t \in S_m, \|t\| = 1\}$ for all $m \in \mathcal{M}_n$, $B_n(0,1) = \{t \in \mathcal{S}_n, \|t\| \le 1\}$ and $B_{m\vee m'}(0,1) = \{t \in S_m + S_{m'}, \|t\| \le 1\}$ for all $m, m' \in \mathcal{M}_n$ with $S_m + S_{m'} = S_{m\vee m'}$ for nested collection of models.

2.2. Class of density functions

Given a density function f belonging to some class of smooth functions $\mathcal{F}$, the $\mathbb{L}^2$-norm on $\mathcal{F}$ and a collection $(S_m)_{m\in\mathcal{M}_n}$ of linear subspaces of $\mathbb{L}^2([0,1])$ described in Section 2.1 of dimension bounded by D_m, the goal is to provide strategies to find an optimal value $m^* \in \mathcal{M}_n$ and the corresponding element f_{m^*} in the collection for the quadratic risk. It is well-known that the element of best approximation is the orthogonal projection f_m of f on S_m in $\mathbb{L}^2([0,1])$ and standard calculations give the following upper bound for the quadratic risk

$$\mathbb{E}\|f - f_{m^*}\|^2 \le C\left(\|f - f_m\|_2^2 + \frac{D_m}{n}\right), \forall m \in \mathcal{M}_n,$$

where C is a constant. Both terms in the above upper bound depend on D_m and as a consequence, we should make the usual squared bias/variance compromise in order to minimize the risk. The way of solving this minimization problem is to put a smoothness assumption on f. For all three collections [T], [P] or [W], the $\mathbb{L}^2$ projection f_m on the linear subspace S_m achieves the best rate of approximation in $\mathbb{L}^2([0,1])$ for the Besov class of functions $\mathcal{F} = \mathcal{B}^{\alpha}_{2,\infty}(A)$ (see Lemma 12, ?)) i.e.

$$\|f - f_m\| \le C(\varphi, \alpha)\,|f|_{\alpha,2}\, D_m^{-\alpha}, \quad \text{for } f \in \mathcal{B}^{\alpha}_{2,\infty}(A)\,,$$

where $|f|_{\alpha,2}$ denotes the semi-norm of f in $\mathcal{B}^{\alpha}_{2,\infty}(A)$ and $C(\varphi,\alpha)$ is a positive constant depending on the basis $(\varphi_\lambda)_{\lambda\in\Lambda_m}$ and α. Therefore, balancing the approximation and variance terms leads to an optimal rate of order $n^{-2\alpha/(2\alpha+1)}$ for a dimension of order $1/n^{2\alpha+1}$. Details on the approximation properties of the bases considered in Section 2.1 can be found in DeVore and Lorentz (1993). Recall briefly that the semi-norm in $\mathcal{B}^{\alpha}_{2,\infty}(A)$ is defined by $|f|_{\alpha,2} = \sup_{y>0} y^{-\alpha}\omega_r(f,y) < +\infty$ with $r = [\alpha]+1$ and the modulus of smoothness is

$$\omega_r(f,y) = \sup_{0<h\le y} \|\sum_{k=0}^{r} \left(\begin{array}{c} r \\ k \end{array}\right) (-1)^{r-k} f(x+kh)\|\,.$$

We want to point out here that Besov spaces provide a generalization of Sobolev spaces of differentiable functions in the sense that, $W_2^{\alpha}(A) \subset \mathcal{B}^{\alpha}_{2,\infty}(A)$. In the (heuristically) simplest case of Sobolev functions, ($f \in W_2^{\alpha}(A)$ means that $f, f', \ldots, f^{(\alpha)} \in \mathbb{L}^2(A)$ with α an integer) the expression of the semi-norm reduces to $|f|_{\alpha,2} = \sum_{j=1}^{\alpha} \|f^{(j)}\|$. Moreover the semi-norm $|f|_{\alpha,2}$ involving the derivatives of f up to the order α is generalized thanks to the r-th modulus of smoothness ω_r on Besov spaces with the index of regularity α not necessarily being an integer. The norm on the general Besov space is then defined by $\|f\|_{\mathcal{B}^{\alpha}_{2,\infty}(A)} = |f|_{\alpha,2} + \|f\|$, for $f \in \mathcal{B}^{\alpha}_{2,\infty}(A)$.

3. Biased data

3.1. Model and associated contrast

Consider first in this section the estimation of the density f under bias selection but without censoring. We have at our disposal an i.i.d. sample $X_{w,1},\ldots,X_{w,n}$ with density f_w defined by (1) subject to a known selection bias w. To this end, we consider the following natural contrast function

$$\gamma_n(t) = \|t\|^2 - \frac{2\hat\mu}{n}\sum_{i=1}^{n} w^{-1}(X_{w,i})t(X_{w,i}), \;\; \hat\mu = \left(\frac{1}{n}\sum_{i=1}^{n} w^{-1}(X_{w,i})\right)^{-1}, \tag{4}$$

where t is an arbitrary function belonging to S_m. The estimate of μ is standard, see for instance, Cox (1969), Patil and Rao (1977), Efromovich (2004, a). Moreover the following assumption is set on w:

$$\exists w_1, w_2 > 0 \;\text{ such that }\; 0 < w_1 \le w(x) \le w_2 < +\infty, \;\; \forall x \in A. \tag{5}$$

Then we consider the following estimator of $f_{|A}$ with A a compact set,

$$\hat f_m = \arg\min_{t\in S_m} \gamma_n(t). \tag{6}$$

In the sequel we will use the abusive notation $f = f_{|A}$ and the corresponding projection estimator $\hat f_m$ is an estimator of the restricted density $f_{|A}$. Note that such a projection estimator is very simple to compute since

$$\hat f_m = \sum_{\lambda\in\Lambda_m} \hat a_\lambda \varphi_\lambda \quad \text{with } \hat a_\lambda = \frac{\hat\mu}{n}\sum_{i=1}^{n}\frac{\varphi_\lambda(X_{w,i})}{w(X_{w,i})}\,. \tag{7}$$

3.2. Risk bounds for the estimators

It is easy to see that the following decomposition of the contrast holds for $t, s \in \mathbb{L}^2(A)$,

$$\begin{aligned}\gamma_n(t) - \gamma_n(s) &= \|t-f\|^2 - \|s-f\|^2 + 2\langle t-s, f\rangle - \frac{2}{n}\hat{\mu}\sum_{i=1}^{n} w^{-1}(X_{w,i})(t-s)(X_{w,i}) \\ &= \|t-f\|^2 - \|s-f\|^2 - 2\,\frac{\hat{\mu}-\mu}{\mu}\left[\langle t-s, f\rangle + \nu_n(t-s)\right] - 2\nu_n(t-s), \qquad (8)\end{aligned}$$

where, since $\mathbb{E}[\mu\, w^{-1}(X_{w,i})\, t(X_{w,i})] = \langle t, f\rangle$, we denote by $\nu_n(t)$ the centered empirical process of interest

$$\nu_n(t) = \frac{1}{n}\sum_{i=1}^{n}\left[\mu w^{-1}(X_{w,i}) t(X_{w,i}) - \langle t, f\rangle\right].$$

Let f_m be the orthogonal projection of f on S_m. It follows from (6) that $\gamma_n(\hat{f}_m) \le \gamma_n(f_m)$, and using decomposition (8), we find that

$$\|\hat{f}_m - f\|^2 \le \|f_m - f\|^2 + 2\frac{\hat{\mu}-\mu}{\mu}\langle \hat{f}_m - f_m, f\rangle + 2\nu_n(\hat{f}_m - f_m) + 2\frac{\hat{\mu}-\mu}{\mu}\nu_n(\hat{f}_m - f_m).$$

Let us denote by $B_m(0,1) = \{t \in S_m, \|t\| = 1\}$. The linearity of the process ν_n ensures that $2\,\mathbb{E}|\nu_n(\hat{f}_m - f_m)| \le (1/8)\mathbb{E}\|\hat{f}_m - f_m\|^2 + 8\mathbb{E}\left[\sup_{t\in B_m(0,1)} \nu_n^2(t)\right]$ and also that $\mathbb{E}\left[\sup_{t\in B_m(0,1)} \nu_n^2(t)\right] \le \sum_{\lambda\in\Lambda_m} \mathbb{E}[\nu_n^2(\varphi_\lambda)] \le n^{-1}\sum_{\lambda\in\Lambda_m}\mathbb{E}\left[\mu^2\varphi_\lambda^2(X_{w,1})/w^2(X_{w,1})\right]$. It follows immediately that

$$2\,\mathbb{E}|\nu_n(\hat{f}_m - f_m)| \;\le\; \frac{1}{4}\mathbb{E}(\|\hat{f}_m - f\|^2) + \frac{1}{4}\|f_m - f\|^2 + \frac{8\mu^2\Phi_0^2}{w_1^2}\frac{D_m}{n}. \qquad (9)$$

Moreover, it is easy to check that $\mathbb{E}(\hat{\mu}-\mu)^2 \le w_2^2\mu^2/(w_1^2 n)$, and so

$$\begin{aligned}2\,\mathbb{E}|\frac{\hat{\mu}-\mu}{\mu}\langle \hat{f}_m - f_m, f\rangle| &\le \frac{1}{8\|f\|^2}\mathbb{E}(\langle \hat{f}_m - f_m, f\rangle^2) + 8\frac{\|f\|^2}{\mu^2}\mathbb{E}(\hat{\mu}-\mu)^2 \\ &\le \frac{1}{4}\mathbb{E}(\|\hat{f}_m - f\|^2) + \frac{1}{4}\|f_m - f\|^2 + \frac{8\|f\|^2 w_2^2}{w_1^2}\frac{1}{n}.\end{aligned}$$

Lastly,

$$\begin{aligned}2\,\mathbb{E}\left|\frac{\hat{\mu}-\mu}{\mu}\nu_n(\hat{f}_m - f_m)\right| &\le \frac{1}{8}\,\mathbb{E}\|\hat{f}_m - f_m\|^2 + \frac{8}{\mu^2}\,\mathbb{E}[(\hat{\mu}-\mu)^2\sup_{t\in B_m(0,1)}\nu_n^2(t)] \\ \le\; \frac{1}{4}\mathbb{E}\|\hat{f}_m - f\|^2 &+ \frac{1}{4}\|f_m - f\|^2 + \frac{8}{\mu^2}\mathbb{E}^{1/2}[(\hat{\mu}-\mu)^4]\mathbb{E}^{1/2}\left[\sup_{t\in B_m(0,1)}\nu_n^4(t)\right]. \qquad (10)\end{aligned}$$

It is easy to see that $\mathbb{E}[(\hat{\mu}-\mu)^4] \le (\mu^4 w_2^4 m_4^4/n^2)$ where $m_4^4 = \mathbb{E}[(w^{-1}(X_{w,1}) - \mu^{-1})^4] \le 8(w_1^{-4} + \mu^{-4})$. On the other hand

$$\begin{aligned}\mathbb{E}\left[\sup_{t\in B_m(0,1)}\nu_n^4(t)\right] &\le \mathbb{E}\left[\sum_{\lambda\in\Lambda_m}\nu_n^2(\varphi_\lambda)\right]^2 \le D_m\sum_{\lambda\in\Lambda_m}\mathbb{E}[\nu_n^4(\varphi_\lambda)] \\ &\le 4D_m\sum_{\lambda\in\Lambda_m}\mathbb{E}\left(\frac{\mu^4\varphi_\lambda^4(X_{w,1})}{w^4(X_{w,1})}\right)\frac{1}{n^2} \le \frac{4\Phi_0^4\,\mu^4}{w_1^4}\frac{D_m^3}{n^2}.\end{aligned}$$

Gathering all terms, if $D_m \leq n$, roughly bounding D_m^3/n^2 by D_m/n, we find

$$2\,\mathbb{E}\left|\frac{\hat{\mu}-\mu}{\mu}\nu_n(\hat{f}_m - f_m)\right| \leq \frac{1}{4}\mathbb{E}\|\hat{f}_m - f\|^2 + \frac{1}{4}\|f_m - f\|^2 + \frac{16\,w_2^2 m_4^2 \Phi_0^2 \mu^2}{w_1^2}\frac{D_m}{n}. \tag{11}$$

Moreover if $D_m \leq n^{2/3}$ the last term would be of order $1/n$. Finally, with inequalities (9), (10) and (11) we obtain for one estimator $\hat{f}_m$ in the collection

$$\mathbb{E}(\|\hat{f}_m - f\|^2) \leq \frac{7}{4}\|f_m - f\|^2 + \frac{3}{4}\mathbb{E}(\|\hat{f}_m - f\|^2) + \frac{8\Phi_0^2\mu^2}{w_1^2}\left(1 + 2w_2^2 m_4^2\right)\frac{D_m}{n} + \frac{8\|f\|^2 w_2^2}{w_1^2}\frac{1}{n},$$

and therefore, the following result:

Proposition 3.1. *Let $\hat{f}_m$ be an estimator of f on $A = [0,1]$ given by (6) on a linear space S_m satisfying ($\mathcal{H}_1$), then*

$$\mathbb{E}(\|f - \hat{f}_m\|^2) \leq 7\|f - f_m\|^2 + K_1\frac{D_m}{n} + \frac{K_2}{n},$$

where $K_1 = 32\Phi_0^2\mu^2\left(1 + 2w_2^2 m_4^2\right)/w_1^2$, $K_2 = 32\|f\|^2 w_2^2/w_1^2$.

Remark 3.1. Note that w is assumed as in all the section to satisfy (5), but the assumption $w(x) > w_1$ in Proposition 3.1 can be replaced by $\mathbb{E}[w^{-4}(X_{w,1})] < +\infty$, and this is fulfilled for a length bias $w(x) = x$ and some Gamma distributions $\Gamma(p,\theta)$ with $p > 3$. As we shall see in the simulation study, it does not seem to be a problem when this type of condition is not fulfilled.

As claimed in Section 2.2, this non asymptotic bound provides a way for balancing the approximation term $\|f - f_m\|^2$ and the variance one and therefore this leads to standard rates on Besov spaces $B_{2,\infty}^{\alpha}(A)$ provided that the dimension D_m of the projection space is well chosen in function of the index of regularity α of the function. If $f \in \mathcal{B}_{2,\infty}^{\alpha}(A)$, the resulting estimator $\hat{f}_m$ of the density f in the collection defined by (7) achieves the usual optimal nonparametric rate $n^{-2\alpha/(2\alpha+1)}$. Of course, the regularity α is unknown and an automatic data-driven choice for D_m must be performed by penalization of the contrast. It is clear that most elements of the above proof, together with a straightforward application of Talagrand's inequality will give an adaptive result as shown in the following subsection.

3.3. Automatic selection of the optimal dimension

The adaptive counterpart of the estimator leads to an automatic selection of the space S_m in the collection via a penalization term in the contrast function:

$$\hat{m} = \arg\min_{m \in \mathcal{M}_n} \gamma_n(\hat{f}_m) + \mathrm{pen}(m) \tag{12}$$

with

$$\mathrm{pen}(m) = \kappa\mu^2\Phi_0^2\,\mathbb{E}[w^{-2}(X_{w,1})]\frac{D_m}{n} \tag{13}$$

standing for a penalty function with κ a numerical constant. Note that the penalty depends on constants which do not have the same status. Indeed, the constant Φ_0 related to the basis is known, but the constant μ and $\mathbb{E}[w^{-2}(X_{w,1})]$ are unknown and must be replaced by estimators. The adequate data-driven penalty therefore equals

$$\widehat{\mathrm{pen}}(m) = \kappa\hat{\mu}^2\Phi_0^2\left(\frac{1}{n}\sum_{i=1}^{n} w^{-2}(X_{w,i})\right)\frac{D_m}{n} \tag{14}$$

where the constant κ is a universal constant usually determined by simulations experiments (see Section 3.4).

THEOREM 3.1. *Consider a collection of models satisfying* $(\mathcal{H}_1)$ *and* $(\mathcal{H}_2)$ *with* $N_n \leq n^{2/3}$. *Let* $\hat{f}_{\hat{m}}$ *be the adaptive estimator defined by*

$$\hat{m} = \arg\min_{m\in\mathcal{M}_n} \gamma_n(\hat{f}_m) + \widehat{\mathrm{pen}}(m) \tag{15}$$

with the random penalty $\widehat{\mathrm{pen}}(m)$ *given by (13), then*

$$\mathbb{E}(\|\hat{f}_{\hat{m}} - f\|^2) \leq C \inf_{m\in\mathcal{M}_n} \left[\|f_m - f\|^2 + \mathrm{pen}(m)\right] + \frac{C'}{n}, \tag{16}$$

where C *and* C' *are constants depending on* Φ_0, $\|f\|, w_1, w_2$ *and* μ.

The condition $N_n \leq n^{2/3}$ leads to some mild restriction on the regularity α of the considered density functions in Besov spaces $\mathcal{B}^{\alpha}_{2,\infty}(A)$. Since we have already seen that the optimal rate is achieved for $D_m = O(n^{1/(2\alpha+1)})$, this requires $\alpha > 1/4$.

This adaptive result is proved in two steps. First, the nonasymptotic bound in (16) is proved for the theoretical penalty given by (13). Then, it is easy to extend to the case of the random penalty given by (14) (see e.g. Brunel and Comte (2005)). Moreover, as a consequence of Section 2.2, we obtain straightforwardly from Theorem 3.1, a rate for functions belonging to Besov spaces.

COROLLARY 3.1. *Assume that an estimator* $\tilde{f}$ *of* f *defined by (15) satisfies (16) under assumptions of Theorem 3.1, then*

$$\mathbb{E}\|f - \tilde{f}\|^2 \leq C(\alpha, f, \Phi_0)\, n^{-\frac{2\alpha}{2\alpha+1}} \tag{17}$$

for all $f \in \mathcal{B}^{\alpha}_{2,\infty}(A)$ *and with* $C(\alpha, f, \Phi_0)$ *a constant depending on the basis and the smoothness of the density* f.

REMARK 3.2. This adaptive result has to be compared to the results obtained for the Efromovich-Pinsker adaptive estimator in Efromovich (2004, b). The global minimax approach reaches very similar rates for a class of Sobolev densities but for trigonometric bases only. Here, we consider a larger variety of models. On the other hand the local approach for densities in a neighborhood of some fixed density f_0 beats all the others in term of nonparametric superefficiency rates since there is a logarithmic gain in the asymptotics more precisely the risk is $O(q_n^{1/2m} \times n^{-2m/(2m+1)})$ with $q_n = O(\ln(n))^{-C_2}$, for some positive constant C_2. But, the fixed density f_0 corresponds to a low frequency part of the unknown density and requires an *a priori* knowledge on it.

3.4. Simulation study

We present here simulation results for the estimator $\hat{f}_{\hat{m}}$ obtained when using the trigonometric basis. Consider an i.i.d. sample $X_{w,1}, \ldots, X_{w,n}$ with common p.d.f. f_w, we aim to estimate the unbiased p.d.f. f. We assume that the observations are in $[a,b]$. For $m \geq 0$, the dimension of the projection space S_m is given by $D_m = 2m+1$. For a fixed $m \geq 0$ and all $x \in [a,b]$, the estimator $\hat{f}_m$ is defined by :

$$\hat{f}_m(x) \quad = \quad \frac{\hat{a}_1}{\sqrt{b-a}} + \sqrt{\frac{2}{b-a}}\left\{\sum_{k=0}^{m}\left[\hat{a}_{2k}\cos\left(\frac{2\pi k(x-a)}{b-a}\right) + \hat{a}_{2k+1}\sin\left(\frac{2\pi k(x-a)}{b-a}\right)\right]\right\}$$

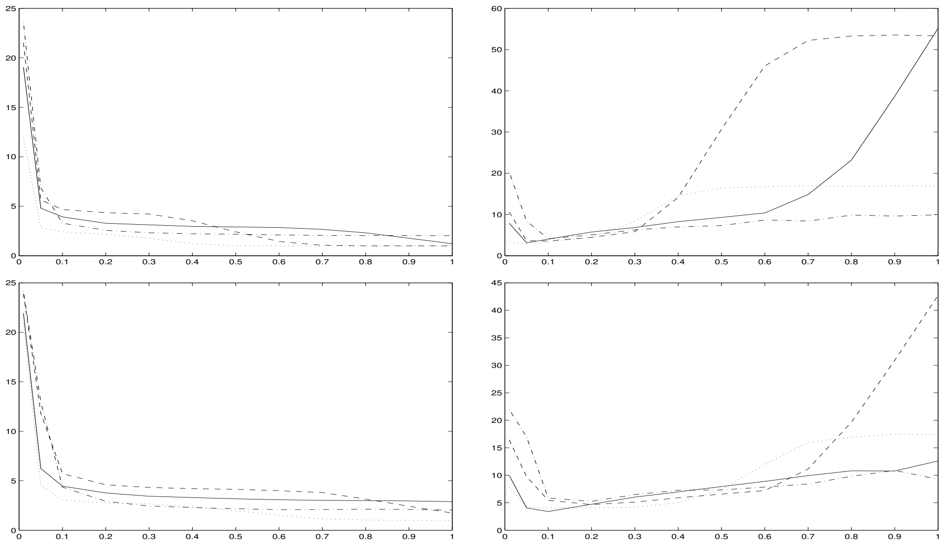

Fig. 1. Left: Selected dimension in function of κ. Right: MISE in function of the constant κ. Unbiased case at the top, length-biased case at bottom. Plain : a), dotted : b), dashed-dotted : c), dashed : d).

where $\hat{a}_1 = (\hat{\mu}/n)\sum_{i=1}^n w^{-1}(X_{w,i}) = 1$ and for $k \in \{1, \dots, n\}$,

$$\hat{a}_{2k} \quad = \quad \frac{\hat{\mu}\sqrt{2}}{n\sqrt{b-a}} \sum_{i=1}^n \frac{\cos\left(\frac{2\pi k (X_{w,i}-a)}{b-a}\right)}{w(X_{w,i})} \text{ and } \hat{a}_{2k+1} = \frac{\hat{\mu}\sqrt{2}}{n\sqrt{b-a}} \sum_{i=1}^n \frac{\sin\left(\frac{2\pi k (X_{w,i}-a)}{b-a}\right)}{w(X_{w,i})}.$$

Now, to construct the penalized contrast estimator $\hat{f}_{\hat{m}}$, the dimension $D_{\hat{m}}$ of the selected projection space has to minimize the penalized contrast

$$\gamma_n(\hat{f}_m) + \widehat{\text{pen}}(m) = -\left(\hat{a}_1^2 + \sum_{k=0}^m \left(\hat{a}_{2k}^2 + \hat{a}_{2k+1}^2\right)\right) + \kappa\hat{\mu}^2\Phi_0^2\left(\frac{1}{n}\sum_{i=1}^n w^{-2}(X_{w,i})\right)\frac{D_m}{n},$$

where $\Phi_0 = \sqrt{2}$ (for the trigonometric basis) and the constant κ has to be suitably chosen.

In what follows, the constant κ is equal to 0.15 (see Figure 1) and, following Birgé and Rozenholc (2005), we replace D_m by $D_m + \ln(D_m)^{2.5}$ in the penalty term of the above contrast. In addition, we remove observations under the 5th percentile when computing $(1/n)\sum_{i=1}^n w^{-2}(X_{w,i})$.

We consider different situations : $w(t) = 1$, in which data do not suffer from selection bias (unbiased case, UB), $w(t) = t$, which corresponds to a length-bias (LB) and $w(t) = t^{1/3}$ called power-bias (PB). Four distributions for the X_i's are studied:

a) the Gamma distribution with shape parameter 5 and scale 1.

b) the mixed Gamma distribution. In this case, the p.d.f f of the X_i's is given, for $x > 0$, by :

$$f(x) = \frac{0.3}{4!}x^4\exp(-x) + \frac{0.7}{12!}x^{12}\exp(-x).$$

c) the Normal distribution with mean 3 and variance 1.

d) the mixed Normal distribution. Here the p.d.f. f of the X_i's is given for, $x \in \mathbb{R}$, by :

$$f(x) = 0.3\Phi(x+3) + 0.7\Phi(x+8)$$

where Φ is the standard normal p.d.f.

Concerning the interval $[a,b]$ of estimation and for each of these distributions, the lower bound a can be taken equal to 0 and the upper bound b to $\max_{i=1,\dots,n} X_{w,i}$. Notice moreover that in cases a) and b), the moment condition mentioned in Remark 3.1 is fulfilled whereas this is not true in cases c) and d). It does not seem to be a problem in the simulation results.

For a p.d.f. f and its penalized contrast estimator $\hat{f}_{\hat{m}}$, the integrated squared error (ISE) is $\text{ISE} = \int_{[a,b]}(\hat{f}_{\hat{m}}(x) - f(x))^2 dx$ and is computed as follows:

$$\text{ISE} = \frac{b-a}{K}\sum_{k=0}^{K-1}\left[\hat{f}_{\hat{m}}\left(a + \frac{k(b-a)}{K}\right) - f\left(a + \frac{k(b-a)}{K}\right)\right]^2,$$

where K is an integer. The Mean ISE (MISE) is the mean of those values over 500 simulated samples. Figure 1 illustrates the behavior of the MISE (Figure 1-left) and of the selected dimension of the projection space (Figure 1-right) in both unbiased case (Figure 1-top) and length-biased sampling case (Figure 1-bottom) when the constant κ is varying. This explains our choice $\kappa = 0.15$, for which a global minimum for the different MISE seems to be reached. Figure 1-right shows clearly that the increase of the error for too small penalties is much faster than the increase for too great ones. From Figure 1-left, we can see that too small penalties are associated with high selected dimensions and therefore high variance terms.

In Figure 2, different examples of estimated p.d.f (dotted) versus true p.d.f. (plain) are given. For each distribution, we compute an estimator from two samples of size 50 (on the left) and size 200 (on the right). As expected, estimators behave better for larger samples. Notice that to avoid the usual problem of scarcity of the observations at the end of the interval, all the plots given in Figure 2 are truncated at the 90th observed percentile. Clearly estimation for mixed distributions is less accurate than in cases a) and c) for small samples, which is not surprising.

		Distribution			
n	Bias	Gamma	Mixed Gamma	Normal	Mixed Normal
50	LB	11.41 (7.46)	10.72 (4.86)	15.72 (8.54)	23.84 (13.95)
	PB	13.39 (10.57)	12.97 (3.97)	13.28 (11.74)	38.38 (18.36)
	UB	14.93 (13.31)	12.82 (3.64)	13.05 (12.53)	44.95 (13.84)
100	LB	6.79 (2.93)	5.12 (2.94)	10.15 (6.87)	8.12 (5.51)
	PB	7.70 (3.94)	7.57 (5.23)	7.44 (5.71)	7.53 (7.33)
	UB	7.54 (3.71)	9.31 (5.52)	7.15 (6.24)	9.13 (10.92)
200	LB	4.06 (2.03)	4.06 (1.24)	5.40 (3.49)	4.86 (1.18)
	PB	4.52 (2.07)	3.63 (1.44)	4.52 (2.51)	3.91 (2.10)
	UB	5.03 (2.17)	3.92 (1.71)	4.59 (3.25)	4.20 (2.40)
500	LB	2.34 (0.91)	3.62 (0.74)	2.55 (1.33)	2.65 (1.18)
	PB	2.40 (0.90)	2.47 (0.49)	1.81 (1.60)	2.45 (1.00)
	UB	2.41 (1.08)	2.59 (0.64)	1.82 (1.55)	2.54 (1.07)

Table 1: MISE$\times$1000 for 500 samples with size 50, 100, 200, 500 in the length-biased case (LB), the power-biased case (PB) and the unbiased case (UB) with standard deviations in the parenthesis.

In Table 1 below, we report the MISE (multiplicated by 1000) in each considered case of our penalized contrast estimator computed over 500 replications of samples of size 50, 100, 200 and

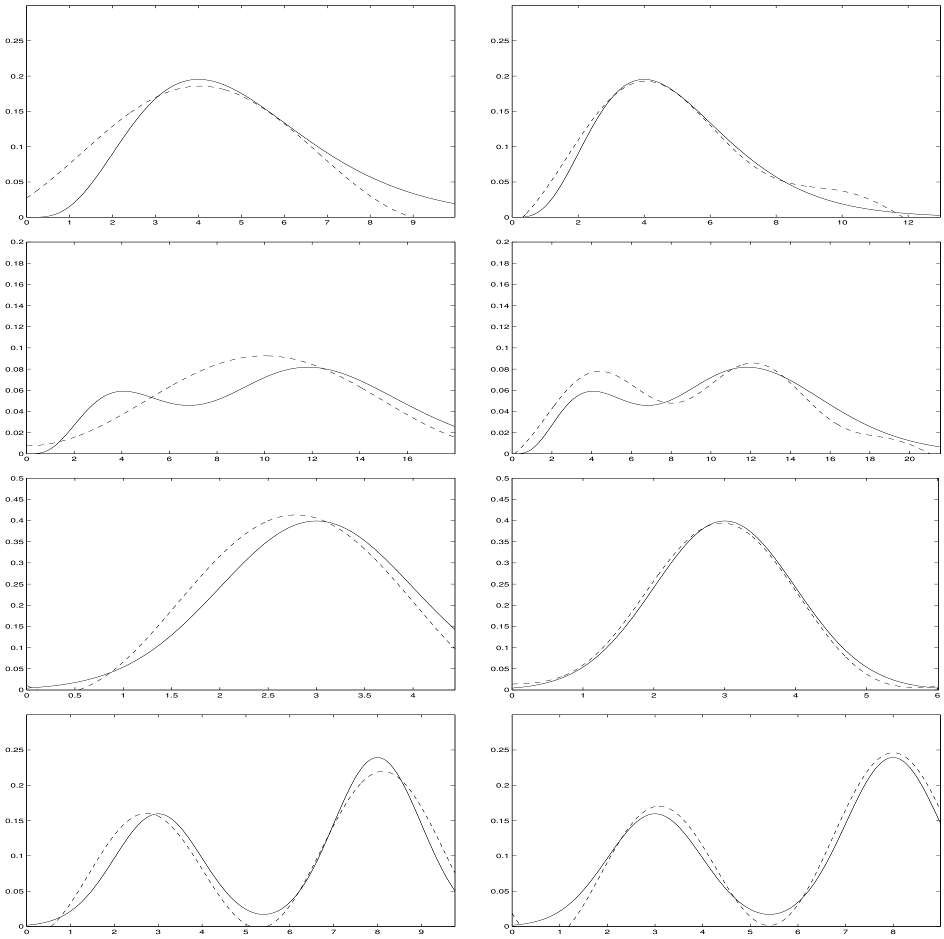

Fig. 2. Examples of estimated f in presence of length bias, for $n = 50$ (left side) and $n = 200$ (right side), and from top to bottom, distributions a), b), c), d). Full line: true density. Dotted line : estimated density

500. Constant K in the computation of the MISE is equal to 500. For each sample size and each distribution, we compare the MISE in the length-biased case, the power-biased case and the unbiased case. As expected again, the value of the MISE reduces when the sample size increases. Moreover, the values of the MISE are of the same order in all cases but contrary to our intuition, the MISE in the unbiased case is not always smaller that the MISE in the biased cases.

4. The case of biased and censored data

4.1. Model with both bias and censoring

We consider now that the biased observations are lifetime data observed with censoring. In other words, we observe

$$Y_i = X_{w,i} \wedge C_i \text{ and } \delta_i = \mathbf{I}_{\{X_{w,i} \le C_i\}}$$

where $a \wedge b$ is the minimum between a and b and the C_i's are i.i.d. censoring times, independent of the $X_{w,i}$'s. We denote by g the probability density function of the C_i's, by G the associated cumulative distribution function and by $\bar{G} = 1 - G$ the survival function. Recall that the density of the unobserved sample is denoted by f, the biased sample has density f_w with cumulative distribution function F_w and survival function $\bar{F}_w = 1 - F_w$. Since the inverse of the survival functions $\bar{G}^{-1}$ and $\bar{F}_w^{-1}$ are not square integrable in general, estimates will only be computed over a bounded interval $A \subsetneq [0, \tau]$ with $\tau = \inf\{x \in \mathbb{R}_+ : (1-F_w)(1-G)(x) = 0\}$. This does not imply any practical restriction since, for estimation purpose, we can choose the interval A such that the largest uncensored $X_{w,i}$'s belongs to it. Up to a scaling change, we set $A = [0, 1]$ without loss of generality, and we have

$$\exists c_G > 0, \;\; \forall x \in A, \;\; c_G \leq 1 - G(x) < 1, \tag{18}$$

by taking $c_G = \inf_{x \in A} [1 - G(x)]$. The same property holds for F_w with c_G replaced by c_{F_w}. Note that we keep assuming that w is known and satisfies Inequality (5).

4.2. Contrast functions and preliminary results

We still aim at estimating f on a compact set A with the sample $(Y_1, \delta_1), \ldots, (Y_n, \delta_n)$. We consider the contrast function

$$\breve{\gamma}_n(t) = \|t\|^2 - \frac{2}{n}\tilde{\mu} \sum_{i=1}^{n} \frac{\delta_i t(Y_i)}{w(Y_i)\hat{\bar{G}}_n(Y_i)}$$

where $\hat{\bar{G}}_n$ is an estimator of the survival function $\bar{G} = 1 - G$ of the censoring times and $\tilde{\mu}$ is an estimator of μ defined by :

$$\tilde{\mu} = \left(\frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i}{w(Y_i)\hat{\bar{G}}_n(Y_i)} \right)^{-1}.$$

Then the projection estimator of the density f is defined by

$$\breve{f}_m = \arg\min_{t \in S_m} \breve{\gamma}_n(t) \tag{19}$$

An explicit formula for this estimator is simply obtained as in Section 3,

$$\breve{f}_m = \sum_{\lambda \in \Lambda_m} \breve{a}_\lambda \varphi_\lambda \quad \text{with } \breve{a}_\lambda = \frac{\tilde{\mu}}{n} \sum_{i=1}^{n} \frac{\delta_i \varphi_\lambda(Y_i)}{w(Y_i)\,\hat{\bar{G}}_n(Y_i)} \tag{20}$$

and model selection is performed by using the criterion:

$$\breve{m} = \arg\min_{m \in \mathcal{M}_n} [\breve{\gamma}_n(\breve{f}_m) + \breve{\text{pen}}(m)] \tag{21}$$

where $\breve{\text{pen}}$ is a penalty function to be found. Now, we have to estimate the survival function $\bar{G}$. A natural choice is to take the Kaplan-Meier product-limit estimator $\hat{\bar{G}}_n$, modified in the way suggested by Lo *et al.* (1989), and defined by

$$\hat{\bar{G}}_n(x) = \prod_{X_{w,(i)} \leq x} \left(\frac{n-i+1}{n-i+2} \right)^{1-\delta_{(i)}} \tag{22}$$

with $X_{w,(i)}$ the i-th order statistic and $\delta_{(i)}$ the corresponding indicator of non-censoring. Indeed, we have the following properties: $\hat{\bar{G}}_n(x) \geq 1/(n+1)$, $\forall x$ and compared to the standard Kaplan-Meier estimate

$$\hat{\bar{G}}_{n,0} = \prod_{X_{w,(i)} \leq x} \left(\frac{n-i}{n-i+1} \right)^{1-\delta_{(i)}}$$

we have, according to Lo *et al.* (1989),

$$\sup_{0\le x\le T} |\hat{\bar{G}}_{n,0}(x) - \hat{\bar{G}}_n(x)| = O(n^{-1}),\ a.s. \tag{23}$$

for all $0 < T < \inf_{x\ge 0}\{x \ge 0 \ / \ G(x) = 1\}$ and in particular this is true for $T = \tau$. Moreover, the following lemma will be useful to control the probability of the uniform deviation of the estimator of the survival function $\hat{\bar{G}}_n$.

LEMMA 4.1. *For all $k \in \mathbb{N}^*$, there exists a constant $C(k)$ depending on k and c_{F_w} such that*

$$\mathbb{E}\left(\sup_{x\in\mathbb{R}} |\hat{\bar{G}}_n(x) - \bar{G}(x)|^{2k}\right) \le C(k)n^{-k}.$$

This implies in particular the following useful result for the moments of the estimator $\tilde{\mu}$.

LEMMA 4.2. *There exist $c_1, c_2 > 0$ satisfying $\mathbb{E}[(\tilde{\mu}-\mu)^2] \le c_1/n$ and $\mathbb{E}[(\tilde{\mu}-\mu)^4] \le c_2/n^2$.*

4.3. Risk bounds of the adaptive estimator

Now, analogous decomposition of the contrast as in Section 3 gives the desired risk bound. Now, we have to take into account a term involving the censoring variable with the inverse of the survival function $\bar{G}^{-1}$ and its empirical version. More precisely, for $t, s \in \mathbb{L}^2([0,1])$

$$\begin{aligned}
\breve{\gamma}_n(t) - \breve{\gamma}_n(s) &= \|t-f\|^2 - \|s-f\|^2 + 2\langle t-s, f\rangle - \frac{2}{n}\tilde{\mu}\sum_{i=1}^n \frac{\delta_i\,(t-s)(Y_i)}{w(Y_i)\hat{\bar{G}}_n(Y_i)} \\
&= \|t-f\|^2 - \|s-f\|^2 - 2R_n(t-s) \\
&\quad -2\,\frac{(\tilde{\mu}-\mu)}{\mu}\,\breve{\nu}_n(t-s) - 2\,\breve{\nu}_n(t-s) - 2\,\frac{\tilde{\mu}-\mu}{\mu}\,\langle t-s, f\rangle
\end{aligned}$$

where $\breve{\nu}_n(t)$ is the centered empirical process defined by

$$\breve{\nu}_n(t) = \frac{1}{n}\sum_{i=1}^n \left(\frac{\mu\delta_i\, t(Y_i)}{w(Y_i)\,\bar{G}(Y_i)} - \langle t, f\rangle\right)$$

and $R_n(t)$ is a residual term defined by

$$R_n(t) = \frac{\tilde{\mu}}{n}\sum_{i=1}^n \frac{\delta_i\, t(Y_i)}{w(Y_i)}\left(\frac{1}{\hat{\bar{G}}_n(Y_i)} - \frac{1}{\bar{G}(Y_i)}\right).$$

Following the approach leading to Proposition 3.1, we prove:

PROPOSITION 4.1. *Let $\hat{f}_m$ be an estimator of f on the compact set $A = [0,1]$ given by (19) on a linear space S_m satisfying ($\mathcal{H}_1$), then*

$$\mathbb{E}(\|\breve{f}_m - f\|^2) \le 9\|f_m - f\|^2 + K_1'\frac{D_m}{n} + \frac{K_2'}{n},$$

where K_1' and K_2' are constants depending on w_1, w_2, c_G, Φ_0 and μ.

As suggested by the proof of Proposition 4.1, we propose the following penalty function

$$\breve{\mathrm{pen}}(m) = \kappa \Phi_0^2 \mu^2 \mathbb{E}\left(\frac{\delta_1^2}{w^2(Y_1)\bar{G}^2(Y_1)}\right)\frac{D_m}{n}, \tag{24}$$

where κ is a numerical constant. As in the simple biasing case, the penalty depends on several constants and a relevant random penalty will be used to draw up the adaptive version of the estimator. The latter is easy to define by replacing the unknown quantities by their empirical counterparts:

$$\widehat{\breve{\mathrm{pen}}}(m) = \kappa \Phi_0^2 \tilde{\mu}^2 \left(\frac{1}{n}\sum_{i=1}^n \frac{\delta_i}{w^2(Y_i)\hat{\bar{G}}_n^2(Y_i)}\right)\frac{D_m}{n}. \tag{25}$$

THEOREM 4.1. *Assume that f is bounded on the compact set $A = [0,1]$ with $\sup_{x\in A}|f(x)| \leq f_1$. Consider the collections [DP] or [W] with $N_n \leq O(n^{2/3})$ or the collection [T] with $N_n \leq \sqrt{n}(4\sqrt{f_1})$ and let $\breve{f}_{\breve{m}}$ be the adaptive estimator defined by (21) with the penalty function $\widehat{\breve{\mathrm{pen}}}(m)$ given by (25), then*

$$\mathbb{E}(\|\breve{f}_{\breve{m}} - f\|^2) \leq C \inf_{m\in\mathcal{M}_n}\left(\|f_m - f\|^2 + \breve{\mathrm{pen}}(m)\right) + \frac{C'}{n}, \tag{26}$$

where C and C' are constants depending on Φ_0, $\|f\|_\infty, w_1, w_2, c_G$ and μ and $\breve{\mathrm{pen}}(m)$ is given by (24).

Restricted conditions $D_m \leq O(n^{2/3})$ for collections [DP] and [W] and $D_m \leq \sqrt{n}/(4\sqrt{f_1})$ for collection [T] imply a condition on the regularity α of the considered functions respectively $\alpha \geq 1/4$ and $\alpha \geq 1/2$.

5. Proofs

5.1. Proof of Theorem 3.1

It follows from the study of the quadratic risk of $\hat{f}_m$, from $(\mathcal{H}_2)$ and from the assumption $D_m^3 \leq n^2$ that, by writing now that, $\forall m \in \mathcal{M}_n$,

$$\gamma_n(\hat{f}_{\hat{m}}) + \mathrm{pen}(\hat{m}) \leq \gamma_n(f_m) + \mathrm{pen}(m)$$

we obtain, following the proof of Proposition 3.1 and for $N_n \leq n^{2/3}$, that

$$\begin{aligned}\mathbb{E}(\|\hat{f}_{\hat{m}} - f\|^2) \quad \leq \quad & 7\|f_m - f\|^2 + \frac{32w_2^2}{w_1^2}(2\,\Phi_0^2\mu^2 m_4^2 + \|f\|^2)\frac{1}{n} + 8\,\mathbb{E}\left(\sup_{t\in B_{\hat{m}\vee m}(0,1)} \nu_n^2(t)\right) \\ & + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}).\end{aligned} \tag{27}$$

where $B_{\hat{m}\vee m}(0,1)$ equals $\{t \in S_{\hat{m}} + S_m \,/\, \|t\| \leq 1\}$. For nested collections of models as described by $(\mathcal{H}_2)$, $S_m + S_{\hat{m}}$ is equal to S_m or $S_{\hat{m}}$ and with $\dim(S_m + S_{\hat{m}}) = D_m \vee D_{\hat{m}} = \max(D_m, D_{\hat{m}})$.

Let $p(m,m')$ be such that

$$8\,p(m,m') \leq \mathrm{pen}(m) + \mathrm{pen}(m') \tag{28}$$

for all $m, m' \in \mathcal{M}_n$.

$$\mathbb{E}(\|\hat{f}_{\hat{m}} - f\|^2) \leq 7\|f_m - f\|^2 + 8\,\mathbb{E}\left(\sup_{t\in B_{\hat{m}\vee m}(0,1)} \nu_n^2(t) - p(m,\hat{m})\right) + 2\,\mathrm{pen}(m) + \frac{C(f,w,\Phi_0)}{n} \tag{29}$$

with $C(f,w,\Phi_0)=32w_2^2(2\,\Phi_0^2 m_4^2\mu^2+\|f\|^2)/w_1^2$. If for all $m\in\mathcal{M}_n$

$$\mathbb{E}(\|\hat{f}_{\hat{m}}-f\|^2) \quad\leq\quad 7\,\|f_m-f\|^2+2\,\mathrm{pen}(m)+\frac{K}{n} \tag{30}$$

then the proof will be complete. So, with (29) it is enough to prove that

$$\mathbb{E}\left(\sup_{t\in B_{m\vee\hat{m}}(0,1)}\nu_n^2(t)-p(m,\hat{m})\right)_+ \leq \sum_{m'\in\mathcal{M}_n}\mathbb{E}\left(\sup_{t\in B_{m\vee m'}(0,1)}\nu_n^2(t)-p(m,m')\right)_+ \leq \frac{K'}{n}$$

The proof is based on the application of Talagrand's Inequality which can be formulated as follows (see Talagrand (1996)). Firstly, from $(\mathcal{H}_1)$, we have for all $m,m'\in\mathcal{M}_n$,

$$\sup_{t\in B_{m\vee m'}(0,1)}\|\mu\,w^{-1}\,t\|_\infty\leq\mu\,w_1^{-1}\,\Phi_0\sqrt{D_{m'}\vee D_m}:=M_1$$

$$\begin{aligned}\sup_{t\in B_{m\vee m'}(0,1)}\mathrm{Var}(\mu\,w^{-1}(X_{w,1})\,t(X_{w,1})) &\leq \sup_{t\in B_{m\vee m'}(0,1)}\int\mu\,w^{-1}(x)\,t^2(x)\,f(x)\,dx\\ &\leq \mu\,w_1^{-1}\Phi_0\,\sqrt{D_{m'}\vee D_m}\|f\|:=v\,,\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left(\sup_{t\in B_{m\vee m'}(0,1)}\nu_n^2(t)\right) &\leq \sum_{\lambda\in\Lambda_{m\vee m'}}\frac{1}{n}\,\mathrm{Var}(\mu\frac{\varphi_\lambda(X_{w,1}))}{w(X_{w,1})}\leq\frac{1}{n}\sum_{\lambda\in\Lambda_{m\vee m'}}\mu^2\,\mathbb{E}\left(\frac{\varphi_\lambda^2(X_{w,1})}{w^2(X_{w,1})}\right)\\ &\leq \frac{\mu^2}{n}\sum_{\lambda\in\Lambda_{m\vee m'}}\int\frac{\varphi_\lambda^2(x)}{w^2(x)}f_w(x)\,dx\leq\Phi_0^2\,\mu^2\,\mathbb{E}[w^{-2}(X_{w,1})]\frac{D_{m'}\vee D_m}{n}:=H^2\end{aligned}$$

and finally, with $C(\epsilon)=(\sqrt{1+\epsilon}-1)\wedge 1$ and κ_1 standing for a universal constant, the Talagrand's Inequality applies,

$$\mathbb{E}\left(\sup_{t\in B_{m\vee m'}(0,1)}\nu_n^2(t)-p(m,m')\right)_+ \quad\leq\quad \frac{6}{\kappa_1}\left(\frac{v}{n}\,e^{-\kappa_1\epsilon\frac{n\,H^2}{v}}+\frac{8M_1^2}{\kappa_1\,n^2\,C^2(\epsilon)}\,e^{-\frac{\kappa_1 C(\epsilon)\sqrt{\epsilon}}{\sqrt{2}}\frac{n\,H}{M_1}}\right),$$

with $p(m,m')=2(1+2\epsilon)H^2$.
Now replacing M_1, v and H^2 by the values derived above, we obtain for $\epsilon=1/2$, $p(m,m')=4H^2=(4\mu^2\Phi_0^2\mathbb{E}[w^{-2}(X_{w,1})])(D_m\vee D_{m'})/n$ which gives the announced penalty (13) by using (28), and the following upper bound,

$$\mathbb{E}\left(\sup_{t\in B_{m\vee m'}(0,1)}\nu_n^2(t)-p(m,m')\right)_+ \leq\frac{\alpha_0}{n}\left(\sqrt{D_{m'}\vee D_m}\,e^{-\alpha_1\sqrt{D_{m'}\vee D_m}}+\alpha_2\,e^{-\alpha_3\sqrt{n}}\right)$$

with $\alpha_0=6\mu\Phi_0\|f\|/(w_1\kappa_1)$, $\alpha_1=\kappa_1\Phi_0\mu w_1\mathbb{E}(w^{-2}(X_{w,1}))/(2\|f\|)$, $\alpha_2=8\mu\Phi_0/[C^2(1/2)w_1\kappa_1\|f\|]$ and $\alpha_3=\kappa_1 C(1/2)w_1\mathbb{E}^{1/2}[w^{-2}(X_{w,1})]/2$. For $|\mathcal{M}_n|\leq n$ and if $(\mathcal{H}_2)$ holds, which ensures that the subspaces are nested,

$$\sum_{m'\in\mathcal{M}_n}\sqrt{D_m\vee D_{m'}}e^{-\alpha_1\sqrt{D_m\vee D_{m'}}}\leq D_m^{3/2}e^{-\alpha_1\sqrt{D_m}}+\sum_{k=1}^{n}\sqrt{k}e^{-\alpha_1 k},$$

with $\sum_{k=1}^n \sqrt{k}e^{-\alpha_1 k} \le S(\alpha_1) < +\infty$ for all $\alpha_1 > 0$ and since $|\mathcal{M}_n| \le n$, we obtain

$$\begin{aligned}
&\sum_{m'\in\mathcal{M}_n} \mathbb{E}\left(\sup_{t\in B_{m\vee m'}(0,1)} \nu_n^2(t) - p(m,m')\right)_+ \\
\le\ & \frac{\alpha_0}{n}\left(\sum_{m'\in\mathcal{M}_n} \sqrt{D_m \vee D_{m'}}\, e^{-\alpha_1\sqrt{D_m\vee D_{m'}}} + \alpha_2|\mathcal{M}_n|\, e^{-\alpha_3\sqrt{n}}\right) \le \frac{\alpha_0}{n}\left(S(\alpha_1)+c\right), \quad (31)
\end{aligned}$$

where $c > 0$ is an arbitrary constant. This ends the proof of the theorem for a penalized estimator with a theoretical penalty. Now, we have to give the proof when $\mathrm{pen}(m)$ is replaced by $\widehat{\mathrm{pen}}(m)$. We set $\Omega_{1,\varepsilon} = \{\omega, |\hat\mu/\mu - 1| < \varepsilon\}$ and $\Omega_{2,\varepsilon} = \left\{\omega, \left|n^{-1}\sum_{i=1}^n w^{-2}(X_{w,i})/\mathbb{E}[w^{-2}(X_{w,1})] - 1\right| < \varepsilon\right\}$. On $\Omega_{1,\varepsilon}\cap\Omega_{2,\varepsilon}$, we mimick the above proof with $\mathrm{pen}(m)$ replaced by $\widehat{\mathrm{pen}}(m)$. Now let us define

$$p(m,m') = \Phi_0^2\mu^2\mathbb{E}[w^{-2}(X_{w,1})]\frac{D_m + D_{m'}}{n}$$

On $\Omega_{1,\varepsilon}\cap\Omega_{2,\varepsilon}$, we have $\hat\mu < (1+\varepsilon)\mu$ and $n^{-1}\sum_{i=1}^n w^{-2}(X_{w,i}) < (1+\varepsilon)\mathbb{E}[w^{-2}(X_{w,1})]$ and also $\mu < \hat\mu/(1-\varepsilon)$ and $\mathbb{E}[w^{-2}(X_{w,1})] < n^{-1}\sum_{i=1}^n w^{-2}(X_{w,i})1/(1-\varepsilon)$. Consequently, with $\kappa = 8/(1-\varepsilon)^2$ we find

$$8\,p(m,\hat m) - \widehat{\mathrm{pen}}(\hat m) + \widehat{\mathrm{pen}}(m) \le \frac{16\Phi_0^2}{(1-\varepsilon)^2}(1+\varepsilon)^2\,\mu\,\mathbb{E}[w^{-2}(X_{w,1})]\frac{D_m}{n}.$$

In the same way as in equation (29), it follows that for all $m\in\mathcal{M}_n$,

$$\begin{aligned}
\mathbb{E}(\|\hat f_{\hat m} - f\|^2\mathbb{I}_{\Omega_{1,\varepsilon}\cap\Omega_{2,\varepsilon}}) \ \le\ & 7\,\|f_m - f\|^2 + 8\sum_{m'\in\mathcal{M}_n}\mathbb{E}\left(\sup_{t\in B_{m,m'}(0,1)}[\nu_n^2(t) - p(m,m')]\right)_+ \\
& + \frac{16\Phi_0^2}{(1-\varepsilon)^2}(1+\varepsilon)^2\,\mu\,\mathbb{E}[w^{-2}(X_{w,1})]\frac{D_m}{n}
\end{aligned}$$

and with equation (31), it is easy to conclude that

$$\mathbb{E}(\|\hat f_{\hat m} - f\|^2\mathbb{I}_{\Omega_{1,\varepsilon}\cap\Omega_{2,\varepsilon}}) \ \le\ 7\,\|f_m - f\|^2 + +\frac{16\Phi_0^2}{(1-\varepsilon)^2}(1+\varepsilon)^2\,\mu\,\mathbb{E}[w^{-2}(X_{w,1})]\frac{D_m}{n} + \frac{B}{n}.$$

Thus, we only have to prove that $\mathbb{E}(\|\hat f_{\hat m} - f\|^2\mathbb{I}_{\Omega_{1,\varepsilon}\cup\Omega_{2,\varepsilon}}) \le B'/n$. Since by definition $\gamma_n(\hat f_{\hat m}) - \gamma_n(f_{\hat m}) = \|\hat f_{\hat m} - f\|^2 - \|f_{\hat m} - f\|^2 - 2\nu_n(\hat f_{\hat m} - f_{\hat m}) \le 0$, we easily write that

$$\|\hat f_{\hat m} - f\|^2 \le 3\|f\|^2 + 8\sup_{t\in S_{\hat m},\|t\|=1}\nu_n^2(t).$$

Then $\sup_{t\in S_{\hat m},\|t\|=1}\nu_n^2(t) \le \sum_{m\in\mathcal{M}_n}\left(\sup_{t\in S_m,\|t\|=1}\nu_n^2(t) - \mathrm{pen}(m)\right)_+ + \mathrm{pen}(\hat m)$. We know in the same way of equation (31) that, for some well chosen κ in $\mathrm{pen}(m)$,

$$\mathbb{E}\left(\sum_{m\in\mathcal{M}_n}\left(\sup_{t\in S_m,\|t\|=1}(\nu_n^2(t) - \mathrm{pen}(m))\right)_+\mathbf{I}_{\Omega_{1,\varepsilon}^c\cup\Omega_{2,\varepsilon}^c}\right) \le \frac{A}{n}.$$

On the other hand, $\forall m\in\mathcal{M}_n$, $\mathrm{pen}(m) \le \kappa\Phi_0^2\mu\mathbb{E}[w^{-2}(X_{w,1})] := A'$, so that $\mathbb{E}\left(\mathrm{pen}(\hat m)\mathbf{I}_{\Omega_{1,\varepsilon}^c\cup\Omega_{2,\varepsilon}^c}\right) \le A'\,\mathbb{P}(\Omega_{1,\varepsilon}^c\cup\Omega_{2,\varepsilon}^c)$. Therefore

$$\mathbb{E}\left(\sup_{t\in S_{\hat m},\|t\|=1}\nu_n(t)^2\mathbf{I}_{\Omega_{1,\varepsilon}^c\cup\Omega_{2,\varepsilon}^c}\right) \le \frac{A}{n} + A'\mathbb{P}(\Omega_{1,\varepsilon}^c\cup\Omega_{2,\varepsilon}^c). \quad (32)$$

Lastly, the result is proved provided that $\mathbb{P}(\Omega^c_{1,\varepsilon}\cup\Omega^c_{2,\varepsilon})\leq \mathbb{P}(\Omega^c_{1,\varepsilon})+\mathbb{P}(\Omega^c_{2,\varepsilon})\leq A"/n$ and this holds by simple use of Tchebychev Inequality, with $\eta_1=\varepsilon\,\mu$

$$\mathbb{P}(\Omega^c_{1,\varepsilon})=\mathbb{P}\left(|\hat{\mu}-\mu|>\eta_1\right) \quad\leq\quad \frac{\mathrm{Var}\,(\hat{\mu})}{\eta_1^2}\leq\frac{\mu^2w_2^2}{nw_1^2\eta_1^2}=\frac{w_2^2}{nw_1^2\varepsilon^2}$$

and on the other hand with $\eta_2=\varepsilon\,\mathbb{E}[X_{w,1}^{-2}]$

$$\mathbb{P}(\Omega^c_{2,\varepsilon})=\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n w^{-2}(X_{w,i})-\mathbb{E}[X_{w,1}^{-2}]\right|>\eta_2\right)\quad\leq\quad\frac{1}{\eta_2^2}\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n w^{-2}(X_{w,i})\right)=\frac{1}{\eta_2^2}\frac{1}{w_1^4}\frac{1}{n}.$$

□

5.2. Proof of Lemma 4.1

First note that with the remark (23), it is enough for $\hat{\hat{G}}_n$ to prove the result for $\hat{\hat{G}}_{n,0}$. We use a nonasymptotic exponential bound for the Kaplan-Meier estimator shown by Földes and Rejtö (1981) which can be formulated as follows (see Bitouzé *et al.*, 1999), there exists a constant $c>0$ such that for any positive λ

$$\mathbb{P}\left(\sqrt{n}\|(1-F_w)\,(\hat{\hat{G}}_{n,0}-\bar{G})\|_\infty>\lambda\right)\leq 2.5\,e^{-2\lambda^2+c\lambda}$$

and so

$$\begin{aligned}\mathbb{E}\left[\left(\sup_{x\in[0,1]}|\hat{\hat{G}}_{n,0}(x)-\bar{G}(x)|\right)^{2k}\right]&=2k\int_0^{+\infty}u^{2k-1}\,\mathbb{P}(\sup_{x\in[0,1]}|\hat{\hat{G}}_{n,0}(x)-\bar{G}(x)|>u)\,du\\&\leq 2k\int_0^{+\infty}u^{2k-1}\,\mathbb{P}(\sqrt{n}\|(1-F_w)\,(\hat{\hat{G}}_{n,0}-\bar{G})\|_\infty>c_{F_w}\sqrt{n}\,u)\,du\\&\leq 5ke^{c^2/8}\int_0^{\infty}u^{2k-1}\exp\left(-2c_{F_w}^2n\left[u-\frac{c}{4\sqrt{n}c_{F_w}}\right]^2\right)du\\&\leq\frac{5e^{c^2/8}k}{2^kc_{F_w}^{2k}}\left(\int_{-c/(2\sqrt{2})}^{+\infty}\left(z+\frac{1}{2\sqrt{2}}\right)^{2k-1}e^{-z^2}dz\right)n^{-k}=C(k)n^{-k}.\square\end{aligned}$$

5.3. Proof of Lemma 4.2

To prove the Lemma, let us introduce the estimator of μ if G is known

$$\tilde{\mu}_0=\left(\frac{1}{n}\sum_{i=1}^n\frac{\delta_i}{w(Y_i)\bar{G}(Y_i)}\right)^{-1}$$

and recall the definition of the set $\Omega_G=\{\hat{\hat{G}}_n(x)\geq c_G/2,\forall x\in[0,1]\}$. On the subset Ω_G^c, we have that $\|\bar{G}-\hat{\hat{G}}_n\|_\infty\geq c_G/2$. Then by writing that

$$\mathbb{E}[(\tilde{\mu}-\mu)^2]\leq 2\mathbb{E}[(\tilde{\mu}-\tilde{\mu}_0)^2]+2\mathbb{E}[(\tilde{\mu}_0-\mu)^2],$$

we find easily that

$$\mathbb{E}[(\tilde{\mu}_0-\mu)^2]\leq\mu^2\,w_2^2\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n\frac{\delta_i}{w(Y_i)\bar{G}(Y_i)}-\frac{1}{\mu}\right)^2\right]\leq\frac{\mu^2w_2^2}{n}\mathrm{Var}\left(\frac{\delta_1}{w(Y_1)\bar{G}(Y_1)}\right)\leq\frac{\mu^2w_2^2}{nw_1^2c_G^2}.$$

On the other hand

$$\begin{aligned}\mathbb{E}[(\tilde\mu-\tilde\mu_0)^2\mathbb{I}_{\Omega_G}] &\le w_2^4\mathbb{E}\left\{\left[\frac{1}{n}\sum_{i=1}^n\frac{\delta_i}{w(Y_i)}\left(\frac{1}{\bar G(Y_i)}-\frac{1}{\hat{\bar G}_n(Y_i)}\right)\right]^2\mathbb{I}_{\Omega_G}\right\}\\ &\le \frac{4w_2^4}{w_1^2c_G^4}\mathbb{E}(\|\bar G-\hat{\bar G}_n\|_\infty^2)\le\frac{4w_2^4C(1)}{w_1^2\,c_G^4}\frac{1}{n}:=\frac{K}{n}\end{aligned}$$

by using Lemma 4.1 with $k=1$.

$$\begin{aligned}\mathbb{E}[(\tilde\mu-\tilde\mu_0)^2\mathbb{I}_{\Omega_G^c}] &\le \frac{w_2^4(n+1)^2}{w_1^2c_G^2}\mathbb{E}(\|\bar G-\hat{\bar G}_n\|_\infty^2\mathbb{I}_{\Omega_G^c})\\ &\le \frac{2^4w_2^4(n+1)^2}{w_1^2c_G^6}\mathbb{E}(\|\bar G-\hat{\bar G}_n\|_\infty^6)\le\frac{K'}{n}\end{aligned}$$

by using Lemma 4.1 with $k=3$. Finally, we get $\mathbb{E}[(\tilde\mu-\mu)^2]\le c_1/n$ with c_1 a positive constant depending on w, c_{F_w}, c_G and μ. The bound for the order 4 is proved analogously. □

5.4. *Proof of Proposition 4.1*

By definition of $\breve f_m$, it is clear that $\breve\gamma_n(\breve f_m)\le\breve\gamma_n(f_m)$. Then, the decomposition of the contrast gives

$$\begin{aligned}\|\breve f_m-f\|^2 &\le \|f_m-f\|^2+2R_n(\breve f_m-f_m)+2\,\frac{\tilde\mu-\mu}{\mu}\,\breve\nu_n(\breve f_m-f_m)\\ &\quad+2\,\breve\nu_n(\breve f_m-f_m)+2\,\frac{\tilde\mu-\mu}{\mu}\,\langle\breve f_m-f_m,f\rangle\end{aligned}$$

The last term is decomposed as in (10) with $\hat\mu$ replaced by $\tilde\mu$ and by using Lemma 4.2, it follows that

$$2\mathbb{E}\left|(\tilde\mu-\mu)\frac{\langle\breve f_m-f_m,f\rangle}{\mu}\right|\le\frac15\mathbb{E}(\|\breve f_m-f\|^2)+\frac15\|f_m-f\|^2+\frac{10\,\|f\|^2}{\mu^2}\frac{c_1}{n}. \tag{33}$$

with c_1 defined in Lemma 4.2. The standard method gives that

$$\mathbb{E}\left(\sup_{t\in B_m(0,1)}\breve\nu_n^2(t)\right)\le\sum_{\lambda\in\Lambda_m}\frac1n\mathrm{Var}\left(\frac{\mu\,\delta_1\,\varphi_\lambda(Y_1)}{w(Y_1)\bar G(Y_1)}\right)\le\frac{\Phi_0^2\mu^2}{c_G^2w_1^2}\frac{D_m}{n}$$

so that

$$2\mathbb{E}|\breve\nu_n(\breve f_m-f_m)|\le\frac15\mathbb{E}(\|\breve f_m-f\|^2)+\frac15\|f_m-f\|^2+\frac{10\Phi_0^2\mu^2}{c_G^2w_1^2}\frac{D_m}{n}\,. \tag{34}$$

Moreover Lemma 4.2 and an upper bound for $\mathbb{E}\left[\sup_{t\in B_m(0,1)}\breve\nu_n^4(t)\right]$ of order D_m^3/n^2 as previously for the term (10), allows to bound $2\,\mathbb{E}|(\tilde\mu-\mu)\breve\nu_n(\breve f_m-f_m)/\mu|$ by

$$\frac15\mathbb{E}\|\breve f_m-f\|^2+\frac15\|f_m-f\|^2+\frac{20\,\Phi_0^2}{c_G^2\,w_1^2}c_2^{1/2}\frac{D_m^{3/2}}{n^2}$$

with c_2 defined in Lemma 4.2 and the last term being of order $1/n$ if $D_m\le n^{2/3}$.

Now, we have to examine the residual term involving the censoring distribution.

$$2\mathbb{E}(|R_n(\breve{f}_m - f_m)|) \leq \frac{1}{10}\mathbb{E}\|\breve{f}_m - f_m\|^2 + 10\,\mathbb{E}\left(\sup_{t\in B_m(0,1)} R_n^2(t)\right).$$

Let $\Omega_G = \{\omega, \hat{\tilde{G}}_n(x) \geq c_G/2, \forall x \in [0,1]\}$. Then

$$\begin{aligned}\mathbb{E}\left(\sup_{t\in B_m(0,1)} R_n^2(t)\mathbb{I}_{\Omega_G}\right) &\leq \sum_{\lambda\in\Lambda_m} \mathbb{E}\left(R_n^2(\varphi_\lambda)\mathbb{I}_{\Omega_G}\right) \leq 4\frac{D_m\Phi_0^2}{w_1^2c_G^4}\mathbb{E}(\|\bar{G} - \hat{\tilde{G}}_n\|_\infty^2\tilde{\mu}^2)\\ &\leq 4\frac{D_m\Phi_0^2}{w_1^2c_G^4}\mathbb{E}^{1/2}(\|\bar{G} - \hat{\tilde{G}}_n\|_\infty^4)\mathbb{E}^{1/2}(\tilde{\mu}^4) \leq 4\frac{\Phi_0^2\sqrt{C(2)}\,c_\mu}{w_1^2c_G^4}\frac{D_m}{n},\end{aligned}$$

because Lemma 4.2 implies also that there exists a positive constant c_μ such that $\mathbb{E}(\tilde{\mu}^4) \leq c_\mu^2$. On the other hand, on Ω_G^c, $\|\bar{G} - \hat{\tilde{G}}_n\|_\infty \geq c_G/2$ so that

$$\mathbb{E}(\|\bar{G} - \hat{\tilde{G}}_n\|_\infty^{2p}\mathbb{I}_{\Omega_G^c}) \leq (2/c_G)^{2q}\,\mathbb{E}(\|\bar{G} - \hat{\tilde{G}}_n\|_\infty^{2p+2q}) \leq (2/c_G)^{2q}(C(p+q)/n)^{p+q}$$

for any integers p and q, by using Lemma 4.1. Therefore, with $p = 2$ and $q = 6$, we immediately obtain

$$\begin{aligned}\mathbb{E}\left(\sup_{t\in B_m(0,1)} R_n^2(t)\mathbb{I}_{\Omega_G^c}\right) &\leq \sum_{\lambda\in\Lambda_m} \mathbb{E}\left(R_n^2(\varphi_\lambda)\mathbb{I}_{\Omega_G^c}\right)\\ &\leq 2^6\frac{D_m(n+1)^2\Phi_0^2c_\mu}{w_1^2c_G^8}\mathbb{E}^{1/2}(\|\bar{G} - \hat{\tilde{G}}_n\|_\infty^{16}\mathbb{I}_{\Omega_G^c}) \leq K\frac{c_\mu\Phi_0^2}{w_1^2c_G^8}\frac{1}{n}\end{aligned}$$

where $K = 2^7\sqrt{C(8)}$ suits. By gathering all terms, we obtain the result with $K_1' = 50\Phi_0^2(\mu^2 + 2\sqrt{c_2} + 4\sqrt{C(2)}c_\mu)/(w_1^2c_G^2)$ and $K_2' = 50(\|f\|^2c_1/\mu^2 + 2^7\sqrt{C(8)}c_\mu\Phi_0^2/(w_1^2c_G^2))$. □

5.5. *Proof of Theorem 4.1*

It follows from the study of $\breve{f}_m$, from $(\mathcal{H}_2)$ and from the assumption $D_m^3 \leq n^2$ that, by writing now that, $\forall m \in \mathcal{M}_n$,

$$\breve{\gamma}_n(\breve{f}_{\breve{m}}) + \breve{\text{pen}}(\breve{m}) \leq \breve{\gamma}_n(f_m) + \breve{\text{pen}}(m)$$

we obtain

$$\begin{aligned}\mathbb{E}(\|\breve{f}_{\breve{m}} - f\|^2) &\leq 9\|f_m - f\|^2 + c(w_1, w_2, c_G, \Phi_0, \mu)\frac{1}{n} + 50\,\mathbb{E}\left(\sup_{t\in B_{\breve{m}\vee m}(0,1)} \breve{\nu}_n^2(t)\right)\\ &\quad +5[\breve{\text{pen}}(m) - \breve{\text{pen}}(\breve{m})] + 50\,\mathbb{E}\left(\sup_{t\in B_{\breve{m}\vee m}(0,1)} R_n^2(t)\mathbb{I}_{\Omega_G}\right),\end{aligned}$$

where $B_{\breve{m}\vee m}(0,1)$ equals $\{t \in S_{\breve{m}} + S_m \,/\, \|t\| \leq 1\}$. For nested collections of models as described by

$(\mathcal{H}_2)$, $S_m + S_{\breve{m}}$ is equal to S_m or $S_{\breve{m}}$ and with $\dim(S_{\breve{m}} + S_m) = D_m \vee D_{\breve{m}} = \max(D_{\breve{m}}, D_m)$.
Let $p(m, m')$ be such that $5\,p(m, m') \leq \breve{\text{pen}}(m) + \breve{\text{pen}}(m')$ for all $m, m' \in \mathcal{M}_n$.

$$\begin{aligned}\mathbb{E}(\|\breve{f}_{\breve{m}} - f\|^2) &\leq 9\|f_m - f\|^2 + 50\,\mathbb{E}\left(\sup_{t\in B_{\breve{m}\vee m}(0,1)} \breve{\nu}_n^2(t) - p(m, \breve{m})\right) \qquad (35)\\ &\quad +6\,\text{pen}(m) + \frac{c(w_1, w_2, c_G, \Phi_0, \mu)}{n} + 50\,\mathbb{E}\left(\sup_{t\in B_n(0,1)} R_n^2(t)\mathbb{I}_{\Omega_G}\right)\end{aligned}$$

where $B_n(0,1)=\{t\in\mathcal{S}_n, \|t\|\le 1\}$. If for all $m\in\mathcal{M}_n$

$$\mathbb{E}(\|\breve{f}_{\breve{m}}-f\|^2) \quad \le \quad 9\,\|f_m-f\|^2+6\,\mathrm{pen}(m)+\frac{K}{n}$$

then the proof will be complete. So, with (35) it is enough to prove that

$$\mathbb{E}\left(\sup_{t\in B_{m\vee\breve{m}}(0,1)}\breve{\nu}_n^2(t)-p(m,\hat{m})\right)_+ \le \sum_{m'\in\mathcal{M}_n}\mathbb{E}\left(\sup_{t\in B_{m\vee m'}(0,1)}\breve{\nu}_n^2(t)-p(m,m')\right)_+ \le \frac{K'}{n}$$

and that

$$\mathbb{E}\left(\sup_{t\in B_n(0,1)}R_n^2(t)\mathbb{I}_{\Omega_G}\right)\le\frac{K"}{n}$$

with Ω_G defined in Proposition 4.1. The proof is based again on the application of Talagrand's Inequality. For all $m,m'\in\mathcal{M}_n$,

$$\sup_{t\in B_{m\vee m'}(0,1)}\|\,\mu\,w^{-1}\bar{G}^{-1}\,t\|_\infty\le\,\mu\,w_1^{-1}c_G^{-1}\,\Phi_0\sqrt{D_{m'}\vee D_m}:=M_1$$

$$\begin{aligned}\sup_{t\in B_{m\vee m'}(0,1)}\mathrm{Var}\left(\frac{\mu\delta_1 t(Y_1)}{w(Y_1)\bar{G}(Y_1)}\right) &\le \sup_{t\in B_{m\vee m'}(0,1)}\mathbb{E}\left[\left(\frac{\mu\delta_1 t(Y_1)}{w(Y_1)\bar{G}(Y_1)}\right)^2\right]\\ &\le \frac{\mu\,\Phi_0}{w_1c_G}\,\|f\|\sqrt{D_{m'}\vee D_m}:=v\,,\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left(\sup_{t\in B_{m\vee m'}(0,1)}\breve{\nu}_n^2(t)\right) &\le \sum_{\lambda\in\Lambda_{m\vee m'}}\frac{1}{n}\mathrm{Var}\left(\frac{\mu\delta_1\varphi_\lambda(Y_1)}{w(Y_1)\bar{G}(Y_1)}\right)\\ &\le \frac{\Phi_0^2(D_{m'}\vee D_m)}{n}\mathbb{E}\left(\frac{\mu^2\delta_1^2}{w^2(Y_1)\bar{G}^2(Y_1)}\right):=H^2\end{aligned}$$

and the orders being the same as previously, the end of the study of this term is as previously.

Now, it remains to study R_n on the set Ω_G. The study is conducted in the same way as in Brunel and Comte (2005) by using the results of Baraud (2002).

$$\begin{aligned}\sup_{t\in B_n(0,1)}[R_n^2(t)\mathbb{I}_{\Omega_G}] &\le \frac{4}{w_1^2c_G^4}\tilde{\mu}^2\|\bar{G}-\hat{\bar{G}}_n\|_\infty^2\sup_{t\in B_n(0,1)}\left(\frac{1}{n}\sum_{i=1}^n|t(X_{w,i})|\right)^2\\ &\le \frac{4}{w_1^2c_G^4}\tilde{\mu}^2\|\bar{G}-\hat{\bar{G}}_n\|_\infty^2\left\{\sup_{t\in B_n(0,1)}[|\nu_n"(t^2)|+\mathbb{E}(t^2(X_{w,1}))]\right\}\\ &\le \frac{4}{w_1^2c_G^4}\tilde{\mu}^2\|\bar{G}-\hat{\bar{G}}_n\|_\infty^2\left[\sup_{t\in B_n(0,1)}|\nu_n"(t^2)|+\frac{f_1w_2}{\mu}\right],\end{aligned}$$

where

$$\nu_n"(t)=\frac{1}{n}\sum_{i=1}^n[t(X_{w,i})-\mathbb{E}(t(X_{w,i}))].$$

Therefore since $\mathbb{E}(\tilde{\mu}^2)$ and $\mathbb{E}(\tilde{\mu}^4)$ are bounded and with the bound on $\mathbb{E}(\|\bar{G}-\hat{\bar{G}}_n\|_\infty^{2k})$ for $k=1,2$ given by Lemma 4.1, we obtain a bound that can be written

$$\mathbb{E}\left(\sup_{t\in B_n(0,1)}R_n^2(t)\mathbb{I}_{\Omega_G}\right)\le\frac{K}{n}\left[\mathbb{E}^{1/2}\sup_{t\in B_n(0,1)}[\nu_n"(t^2)]^2+1\right].$$

Let $\tilde{f}_1 = f_1 w_2/\mu$. It follows from Baraud (2002) that for all $\rho > 0$,

$$\mathbb{P}\left(\sup_{t\in B_n(0,1)} |\nu_n"(t^2)| \geq \rho\right) \leq |\Lambda_n|^2 \exp\left(-\frac{n\rho^2}{4\tilde{f}_1 L_n(\varphi)}\right)$$

where $L_n(\varphi)$ is a quantity associated to the orthonormal basis of the largest space $\mathcal{S}_n = \bigcup_{m\in\mathcal{M}_n} S_n$ of the (nested) collection and $(\varphi_\lambda)_{\lambda\in\Lambda_n}$ denote the orthonormal basis of $\mathcal{S}_n$, dim $\mathcal{S}_n = N_n = |\Lambda_n|$. We know from Baraud (2002) that $L_n(\varphi) \leq K_\varphi N_n$ for the basis [DP] and [W], and $L_n(\varphi) \leq N_n^2$ for [T]. Moreover, it is checked in Brunel and Comte (2005) that, by integration of the previous inequality

$$\mathbb{E}\left(\sup_{t\in B_n(0,1)} [\nu_n"(t^2)]^2\right) \leq \ln(n) + 4\tilde{f}_1 N_n L_n(\varphi)\exp\left(-\frac{n\ln(n)}{4\tilde{f}_1 L_n(\varphi)}\right).$$

It follows that if $L_n(\varphi) \leq n/(16\tilde{f}_1)$,

$$\mathbb{E}\left(\sup_{t\in B_n(0,1)} [\nu_n"(t^2)]^2\right) \leq \ln(n) + \frac{1}{4n^2} \leq 2\ln(n) \ \text{ if } n \geq 2.$$

Therefore, if $N_n \leq n/(16\tilde{f}_1 K\varphi)$ for [DP] and [W] and if $N_n \leq \sqrt{n}/(4\sqrt{\tilde{f}_1})$ for [T], then

$$\mathbb{E}\left(\sup_{t\in B_n(0,1)} R_n^2(t)\mathbb{I}_{\Omega_G}\right) \leq \frac{w_1}{n} + \frac{w_2\sqrt{\ln(n)}}{n}.$$

This ends the proof of the theorem for a penalized estimator with a theoretical penalty. Now, the proof when $\breve{\text{pen}}(m)$ is replaced by $\widehat{\breve{\text{pen}}}(m)$ being similar to the end of the proof of Theorem 3.1 is omitted. □

References

Antoniadis, A., Grégoire, G. and Nason, G. (1999). Density and hazard rate estimation for right-censored data by using wavelet methods. *J. R. Stat. Soc.*, Ser. B Stat. Methodol. **61**, 63-84.

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.

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

Baraud, Y., Comte, F. and Viennet, G. (2001). Adaptive estimation in autoregression or β-mixing regression via model selection. *Ann. Statist* **29**, 839-875.

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

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

Birgé, L. and Massart, P. (1998). Minimum contrast estimators on sieves : exponential bounds and rates of convergence. *Bernoulli* **4** (3), 329-375.

Birgé, L. and Rozenholc, Y. (2005). How many bins must be put in a regular histogram. Forthcoming in *ESAIM Probab. Statist.* Preprint LPMA, University of Paris 6, no.721.

Brunel, E. and Comte, F. (2005). Penalized contrast estimation of hazard rate with censored data. Forthcoming in *Sankhya*. Preprint MAP5 2004-18, www.math-info.univ-paris5.fr/map5/.

Chakraborty, R., and Rao, C.R., (2000). Selection biases of samples and their resolutions. In Handbook of Statistics, Vol. 18 (P.K. Sen and C.R. Rao, eds), 675-712. Elsevier Science.

Cox, D.R. (1969). In *New development in survey sampling.* Eds Johnson and Smith, Wiley.

Dauxois, J.-Y. and Guilloux, A. (2005). Estimating the cumulative incidence functions under selection-biased sampling. Comptes-Rendus du CREST 2004-01.

DeVore, R.A. and Lorentz, G.G. (1993). *Constructive approximation.* Springer-Verlag.

Donoho, D.L. and Johnstone, I.M. (1998). Minimax estimation with wavelet shrinkage. *Ann. Statist.* **26**, 879-921.

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

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

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

Földes, A. and Rejtö, L. (1981). LIL type result for the product limit estimator. *Z. Wahrscheinlichkeitstheor. Verw. Geb.* **56**, 75-86.

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

Heckman, J. (1985). Selection bias and self-selection. In The new Palgrave : A Dictionary of Economics, 287-296. MacMillan Press, Stockton, N.Y.

Lo, S.H., Mack, Y.P. and 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.

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

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

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

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

Patil, G.P. and 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.

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

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

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

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