PENALIZED CONTRAST ESTIMATION OF DENSITY AND HAZARD RATE WITH CENSORED DATA.

E. BRUNEL[1] AND F. COMTE[2]

Abstract. We consider projection estimator methods for the estimation of density and hazard rate functions based on randomly right-censored data. Two types of adaptive hazard estimators are considered. The first one is a two-step estimator defined as the ratio of a penalized contrast estimator of the subdensity and of the empirical survival function of the data. The second estimator is built by using another penalized projection contrast. Both estimators are proved to achieve automatically the standard optimal rate associated with the unknown regularity of the hazard function, but with some restriction for the "ratio" estimator. In the examples studied here, the ratio estimator seems to be slightly better than the direct estimator.

Revised Version, January 2005

AMS Classification (2001): 62G05; 62G20.

Keywords and phrases. Adaptive estimation. Hazard rate. Minimax rate. Right-Censoring. Nonparametric penalized contrast estimator.

1. Introduction

This paper considers a model which is most commonly used in reliability or survival analysis: more precisely, we are interested in lifetimes (or failure times) of some individuals in presence of right-censoring. Right-censoring occurs for example when some of the individuals under study are not observed until the end (death, remission, recovery); in that case, only a lower bound of their lifetimes is observed, which is called the censoring time. In the end, the observation consists in the minimum of the lifetime and the censoring time, and in the knowledge whether the survival time is censored or not. Two functions are of interest in this context, the density of the

[1] IUT de Paris V et Laboratoire MAP5, UMR CNRS 8145. email: brunel@math-info.univ-paris5.fr.

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

data which are not censored, called the subdensity function, and the derivative of the log-survival probability called the hazard rate function. We consider hereafter some nonparametric and adaptive methods in order to estimate these functions.

Let us first describe the nonparametric methods used in the literature. Different strategies are possible to build a nonparametric estimator of the hazard rate. One method is to use the relationship of the hazard rate with the subdensity. Since Blum and Susarla (1980) kernel estimators of the subdensity and the hazard rate in presence of right censoring have been studied by many authors: Mielniczuk (1986), Diehl and Stute (1988), Lo *et al.* (1989), Uzunoḡullari and Wang (1992). The bandwidth selection remains an important problem in this context: indeed, the optimal bandwidth depends on the unknown function under estimation. Antoniadis et al. (1999) consider both subdensity and hazard estimators via some wavelet estimator, but the optimal wavelet resolution also depends on the unknown function. In the present work, an adaptive estimator of the subdensity is built. More precisely, a projection estimator on a finite dimensional space is defined and the dimension of the projection space is selected by using a data driven penalty function. This estimator can therefore automatically reach the optimal rate of convergence, in term of its integrated mean square risk. Then, we point out the possible sub-optimality of the resulting hazard rate estimator, due to its ratio feature.

Other kinds of non parametric estimators of the hazard rate are constructed by direct regularization (by convolving with a kernel for instance) of some cumulative hazard estimator as the Nelson-Aalen or the Kaplan-Meier estimators; early results for such kernel methods can be found in Tanner and Wong (1983), Ramlau-Hansen (1983) and Yandell (1983). In the same way, Wu and Wells (2003), proposed a wavelet-type estimator based on the transform of a Nelson-Aalen cumulative hazard estimator. Let us mention also Kooperberg et al. (1995) who study the L^2 convergence rate of a hazard rate estimator in a context of tensor product splines. Only few works deal with adaptive procedures of estimation and can therefore be compared with the present paper. We can cite Dölher and Rüschendorf (2002), who introduce an adaptive sieved maximum likelihood method but the rate of convergence of their estimator involves a logarithmic loss which makes their procedure slightly suboptimal. Reynaud-Bouret (2002) obtains adaptive results and minimax rates for penalized projection estimators of the Aalen multiplicative intensity process. She first considers histogram-type projection spaces, which are therefore more suitable for the estimation of highly non regular functions than for smooth ones. She also considers Fourier strategies and trigonometric projection spaces, to which we compare our method. Lastly, Brunel and Comte (2004) consider penalized contrast estimator using the Kaplan-Meier cumulative hazard estimator and a larger variety of models. In the present work, we propose, as for the subdensity, a direct projection estimator of the hazard rate, with automatic selection of the projection space by contrast penalization. Our estimator is simple to define and easy to compute (the simulation section illustrates it). Moreover, it is adaptive in the sense that its risk reaches the optimal rate of convergence without any prior information on the unknown function h. Note that the lower bounds for minimax rates in hazard estimation have been recently proved in Huber and Mac Gibbon (2004). Note also that our estimator may be related to the one studied in Patil (1997), who considers however uncensored data and does not prove any adaptive results. One original feature of our estimator comes from its definition without any resort to the Nelson-Aalen or Kaplan-Meier cumulative hazard rate as it is commonly done in the methods described above.

The method used in this work follows the mainstream of model selection methods introduced and developed in different frameworks by Barron and Cover (1991), Birgé and Massart (1997) and Barron *et al.* (1999). Most proofs rely on the powerful Talagrand's (1996) inequality for empirical centered processes. Some more technical properties, proved in a regression framework by Baraud et al. (2001) and Baraud (2002), are also used.

The outline of the paper is as follows. After the description in Section 2 of the lifetimes model and some preliminaries on the projection spaces, we present in Section 3 the study of the estimator of the subdensity of the failure times based on a projection contrast function. Both convergence and adaptation results are given. This estimator is then used to estimate the hazard rate. Section 4 describes a direct adaptive procedure to estimate the hazard rate based on another projection contrast. Note that we consider here the case of both regular and non regular partitions of the set of estimation. Section 5 provides some examples and simulation results, together with comparisons with other estimators. Most proofs and technical lemmas are deferred to Section 6.

2. Notation and assumptions

2.1. **Model set-up.** We consider nonnegative i.i.d. random variables X_i^0, for $i = 1, \dots, n$ (lifetimes for the n subjects under study) with common continuous distribution function F^0, and $C_1, \dots, C_n$ i.i.d. nonnegative random variables ("censoring sequence") with common distribution function G, both sequences being independent. One classical problem when processing with lifetime data is the estimation of the hazard rate function or failure rate function h defined, if F^0 has density f^0 by

$$h(x) = \frac{d}{dx}H(x) = \frac{f^0(x)}{\bar{F}^0(x)}, \quad \text{for } F^0(x) < 1.$$

where $H = -\log(\bar{F}^0)$ is the cumulative hazard rate and $\bar{F}^0 = 1 - F^0$ is the survival function. In the setting of survival analysis data with random right censorship, the bivariate sample $(X_1, \delta_1), \dots, (X_n, \delta_n)$ is observed, where

$$X_i = X_i^0 \wedge C_i, \quad \delta_i = \mathbf{I}_{\{X_i^0 \le C_i\}}.$$

In other words, $\delta_i = 1$ indicates that the i-th subject's survival time is uncensored. We denote by f and F the common density and distribution function of the X_i's. Note that $\bar{F} = 1 - F = (1 - F^0)(1 - G)$. Moreover, we define the subdensity function ψ, the density of the uncensored data, as

$$\psi(x) = f^0(x)(1 - G(x)).$$

Consequently, the hazard rate function can also be expressed as

$$h(x) = \frac{\psi(x)}{1 - F(x)}, \quad \text{for } F(x) < 1.$$

Since the hazard rate function is not square integrable in general, estimates will only be computed over a bounded interval $[0, a]$ where $a < \tau_F = \sup\{x : F(x) < 1\}$. As mentioned in Antoniadis et al. (1999), it can be easily shown that $X_{(n)} \to \tau_F$ almost surely as $n \to \infty$ where $X_{(n)}$ is the n-th order statistic. So, this does not imply any practical restriction since, for estimation purpose, we can choose the bound a greater than the largest of the uncensored X_i's (see also Dölher and Rüschendorf (2002)). Without loss of generality, we set $a = 1$ in all the following (except in Section 5). Notice that this restriction is not necessary for the estimation of the subdensity itself and estimates can be constructed on the whole real line, but this would require another context and different assumptions. In our setting, the following condition is always satisfied

$$\exists c_F > 0, \ \ \forall x \in [0,1], \ \ c_F \leq 1 - F(x) < 1, \tag{2.1}$$

by taking $c_F = \inf_{x\in[0,1]}(1 - F(x)) = 1 - F(1)$.

2.2. **Description of the spaces of approximation.** In this section, the spaces $(S_m)_{m\in\mathcal{M}_n}$ considered in the sequel are described and their key properties are pointed out. The spaces will satisfy the following assumption:

($\mathcal{H}_1$) $(S_m)_{m\in\mathcal{M}_n}$ is a collection of finite-dimensional linear sub-spaces 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 satisfying:

$$\exists \Phi_0 > 0, \forall m \in \mathcal{M}_n, \forall t \in S_m, \|t\|_\infty \leq \Phi_0\sqrt{D_m}\|t\|. \tag{2.2}$$

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

An orthonormal basis of S_m is denoted by $(\varphi_\lambda)_{\lambda\in\Lambda_m}$ where $|\Lambda_m| = D_m$. It follows from Birgé and Massart (1997) that Property (2.2) in the context of ($\mathcal{H}_1$) is equivalent to

$$\exists \Phi_0 > 0, \|\sum_{\lambda\in\Lambda_m} \varphi_\lambda^2\|_\infty \leq \Phi_0^2 D_m. \tag{2.3}$$

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 $\mathcal{S}_n$: $\dim(\mathcal{S}_n) = N_n$ ($\forall m \in \mathcal{M}_n, D_m \leq N_n$).

We consider more precisely the following examples:

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

[P] *Regular piecewise polynomial spaces*: S_m is generated by $m(r+1)$ polynomials, $r+1$ polynomials of degree $0,1,\dots,r$ on each subinterval $[(j-1)/m, j/m]$, for $j = 1,\dots m$, $D_m = (r+1)m$, $m \in \mathcal{M}_n = \{1,2,\dots,[n/(r+1)]\}$. For example, consider the orthogonal collection in $\mathbb{L}^2([-1,1])$ of Legendre polynomials Q_k, where the degree of Q_k is equal to k, $|Q_k(x)| \leq 1, \forall x \in [-1,1]$, $Q_k(1) = 1$ and $\int_{-1}^1 Q_k^2(u)du = 2/(2k+1)$. Then the orthonormal basis is given by $\varphi_{j,k}(x) = \sqrt{m(2k+1)}Q_k(2mx - 2j+1)\mathbf{1}_{[(j-1)/m,j/m[}(x)$ for $j = 1,\dots,m$ and $k = 0,\dots,r$, with $D_m = (r+1)m$. In particular, the histogram basis corresponds to $r = 0$ and is simply defined by $\varphi_j(x) = \sqrt{D_m}\,\mathbf{1}_{[(j-1)/D_m,j/D_m]}(x)$ and $D_m = m$. We call dyadic collection of piecewise polynomials, and denote by [DP], the collection corresponding to dyadic subdivisions 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 (1994): S_m is generated by $\{\phi_{j_0,k}\,,\ \psi_{j,k}\,;\ k \in \mathbb{Z}\,,\ m \geq j \geq j_0\}$ for any fixed resolution j_0 and with $\phi_{j_0,k}(x) = \sqrt{2^{j_0}}\,\phi(2^{j_0}\,x - k)$ and $\psi_{j,k}(x) = \sqrt{2^j}\,\psi(2^j\,x - k)$ where ϕ and ψ denote respectively the *scaling function* and the *mother wavelet* on $[0,1]$ and are elements of the Hölder space $C^r\,, r \geq 0$. In this case, the multi-resolution analysis is said to be r regular. Moreover, the wavelet ψ has vanishing moments up to order r (see for example Daubechies, (1992)). Since ϕ and ψ are compactly supported on $[0,1]$, for any fixed j the sum over k is finite in the wavelet series, more precisely for a function $t \in S_m$,

$$t(x) = \sum_{k=0}^{2^{j_0}-1} a_{j_0,k}\,\phi_{j_0,k}(x) + \sum_{j=j_0}^{m}\sum_{k=0}^{2^j-1} b_{j,k}\,\psi_{j,k}(x).$$

Therefore, the generating basis is of cardinality $D_m = 2^{m+1}$ and $m \in \mathcal{M}_n = \{1,2,\dots,[\ln(n)/2]-1\}$.

All those 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 $(\mathcal{H}_2)$.

3. Analysis of the subdensity and of the resulting hazard rate estimators

3.1. **Definition of the estimator of the subdensity.** Consider the following contrast function

$$\gamma_n^{\psi}(t) = \|t\|^2 - \frac{2}{n}\sum_{i=1}^n \mathbf{1}_{\{X_i\leq 1\}}\delta_i t(X_i) \tag{3.1}$$

where t is a function of $\mathbb{L}^2([0,1])$, $\|t\|^2=\int_0^1 t^2(x)dx$. Let then $\hat\psi_m=\arg\min_{t\in S_m}\gamma_n^\psi(t)$. An explicit expression of the estimator follows from this definition by using the orthonormal basis $(\varphi_\lambda)_{\lambda\in\Lambda_m}$ of S_m described in $(\mathcal{H}_1)$:

$$\hat\psi_m=\sum_{\lambda\in\Lambda_m}\hat a_\lambda\varphi_\lambda \text{ with } \hat a_\lambda=\frac{1}{n}\sum_{i=1}^n \mathbf{I}_{\{X_i\le 1\}}\delta_i\varphi_\lambda(X_i). \tag{3.2}$$

We define also ψ_m as the orthogonal projection of ψ on S_m. We can write

$$\psi_m=\sum_{\lambda\in\Lambda_m}a_\lambda\varphi_\lambda \text{ with } a_\lambda=\int_0^1\varphi_\lambda(x)\psi(x)dx. \tag{3.3}$$

3.2. **Convergence results and adaptation.**

3.2.1. *Optimal rate of the estimator of the subdensity.* The rate of $\hat\psi_m$ is quite easy to derive. Indeed, it follows from Pythagoras theorem, (3.2) and (3.3) that

$$\begin{aligned}\|\psi-\hat\psi_m\|^2 &= \|\psi-\psi_m\|^2+\|\psi_m-\hat\psi_m\|^2=\|\psi-\psi_m\|^2+\sum_{\lambda\in\Lambda_m}(a_\lambda-\hat a_\lambda)^2\\ &= \|\psi-\psi_m\|^2+\sum_{\lambda\in\Lambda_m}\left(\frac{1}{n}\sum_{i=1}^n\mathbf{I}_{\{X_i\le 1\}}\delta_i\varphi_\lambda(X_i)-\int_0^1\psi(x)\varphi_\lambda(x)dx\right)^2.\end{aligned}$$

Therefore

$$\begin{aligned}\mathbb{E}(\|\psi-\hat\psi_m\|^2) &= \|\psi-\psi_m\|^2+\sum_{\lambda\in\Lambda_m}\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n\mathbf{I}_{\{X_i\le 1\}}\delta_i\varphi_\lambda(X_i)\right)\\ &= \|\psi-\psi_m\|^2+\frac{1}{n}\sum_{\lambda\in\Lambda_m}\mathrm{Var}\left(\mathbf{I}_{\{X_1\le 1\}}\delta_1\varphi_\lambda(X_1)\right)\\ &\le \|\psi-\psi_m\|^2+\frac{1}{n}\mathbb{E}\left[\left(\sum_{\lambda\in\Lambda_m}\varphi_\lambda^2(X_1)\right)\mathbf{I}_{\{X_1\le 1\}}\delta_1\right].\end{aligned}$$

Then by using (2.3), we obtain the following Proposition:

Proposition 3.1. *Let $\hat\psi_m=\arg\min_{t\in S_m}\gamma_n^\psi(t)$ where $\gamma_n^\psi(t)$ is defined by (3.1) and S_m is a D_m-dimensional linear space in a collection satisfying $(\mathcal{H}_1)$. Then*

$$\mathbb{E}(\|\psi-\hat\psi_m\|^2)\le\|\psi-\psi_m\|^2+\frac{\Phi_0^2D_m}{n}\int_0^1\psi(x)dx. \tag{3.4}$$

Inequality (3.4) gives the asymptotic rate for an estimator if we consider that ψ belongs to a Besov space.
Let us recall that the Besov space $\mathcal{B}_{\alpha,p,\infty}([0,1])$ is defined by:

$$\mathcal{B}_{\alpha,p,\infty}([0,1]) = \{f \in L_p([0,1]), \ \ |f|_{\alpha,p} := \sup_{t>0} t^{-\alpha}\omega_r(f,t)_p < +\infty\}$$

where $r = [\alpha] + 1$ ([.] denotes the integer part), and $\omega_r(f,t)_p$ is called the r-th modulus of smoothness of a function $f \in L_p(A)$ and is equal to:

$$\omega_r(f,t)_p = \sup_{0<h\le t} \|\Delta_h^r(f,.)\|_p([0,1-rh]), \ t \ge 0, \Delta_h^r(f,x) = \sum_{k=0}^{r} \binom{r}{k} (-1)^{r-k} f(x+kh).$$

Note that $|f|_{\alpha,p}$ is a semi-norm with usual associated norm $\|f\|_{\alpha,p} = \|f\|_p + |f|_{\alpha,p}$, $\|f\|_p = \left(\int |f|^p(x)dx\right)^{1/p}$. For details, we refer to DeVore and Lorentz (1993, p.54-57). The inclusion $\mathcal{B}_{\alpha,p,\infty}([0,1]) \subset \mathcal{B}_{\alpha,2,\infty}([0,1])$ for $p \ge 2$ justifies that we now restrict our attention to spaces $\mathcal{B}_{\alpha,2,\infty}([0,1])$, i.e. to square integrable functions with smoothness order α. Heuristically, a function in $\mathcal{B}_{\alpha,2,\infty}([0,1])$ can be seen as square integrable and $[\alpha]$-times differentiable with derivative of order $[\alpha]$ having a Lipschitz property of order $\alpha - [\alpha]$. Then the following (standard) rate is obtained:

Corollary 3.1. *Let $\hat{\psi}_m = \arg\min_{t\in S_m} \gamma_n^\psi(t)$ where $\gamma_n^\psi(t)$ is defined by (3.1) and S_m is a D_m-dimensional linear space in collection* [T], [P], *or* [W]. *Assume moreover that ψ belongs to $\mathcal{B}_{\alpha_\psi,2,\infty}([0,1])$ with $r > \alpha_\psi > 0$ and choose a model with $m = m_n$ such that $D_{m_n} = O(n^{1/(2\alpha_\psi+1)})$, then*

$$\mathbb{E}(\|\psi - \hat{\psi}_{m_n}\|^2) = O\left(n^{-\frac{2\alpha_\psi}{2\alpha_\psi+1}}\right). \tag{3.5}$$

Remark 3.1. The bound r on α_ψ stands for the regularity of the basis functions for collections [P] and [W]. For the trigonometric collection [T], no upper bound for the regularity α_ψ is required.

Proof. The result is a straightforward consequence of the results of DeVore and Lorentz (1993) and of Lemma 12 of Barron, Birgé and Massart (1999). They imply that, if $\psi \in \mathcal{B}_{\alpha_\psi,2,\infty}([0,1])$ for some $\alpha_\psi > 0$, then $\|\psi - \psi_m\|$ is of order $D_m^{-\alpha_\psi}$ in the three collections [T], [P] and [W]. Thus the minimum order in (3.4) is reached for a model S_{m_n} with $D_{m_n} = O([n^{1/(1+2\alpha_\psi)}])$, which is less than n for $\alpha_\psi > 0$. Then, we find the standard nonparametric rate of convergence $n^{-2\alpha_\psi/(1+2\alpha_\psi)}$. □

3.2.2. *Adaptive estimator of the subdensity.* The penalized estimator is defined in order to ensure an automatic choice of the dimension. Indeed, it follows from Corollary 3.1 that the optimal dimension depends on the unknown regularity α_ψ of the function to be estimated in the asymptotic setting, and more generally on the unknown constants involved in the squared-bias/variance terms. Then we define

$$\hat{m} = \arg \min_{m \in \mathcal{M}_n} [\gamma_n^\psi(\hat{\psi}_m) + \mathrm{pen}^\psi(m)] \tag{3.6}$$

where the penalty function pen^ψ is determined in order to lead to the choice of a "good" model. We easily derive the following result:

Theorem 3.1. *Let S_m be a D_m-dimensional linear space in a collection satisfying $(\mathcal{H}_1)$ and $(\mathcal{H}_2)$. Consider the estimator $\hat{\psi}_{\hat{m}}$ with $\hat{m}$ defined by (3.6) with*

$$\mathrm{pen}^\psi(m) = \kappa \Phi_0^2 \left(\int_0^1 \psi(x)dx \right) \frac{D_m}{n}$$

where κ is a universal constant ($\kappa \geq 4$ suits). Then $\hat{\psi}_{\hat{m}}$ satisfies

$$\mathbb{E}(\|\hat{\psi}_{\hat{m}} - \psi\|^2) \leq C_1 \inf_{m \in \mathcal{M}_n} \left(\|\psi - \psi_m\|^2 + \mathrm{pen}^\psi(m) \right) + \frac{C_2}{n}, \tag{3.7}$$

where C_1 is a constant depending on κ only and C_2 is a constant depending on Φ_0 and on $\int_0^1 \psi(x)dx$.

Therefore, the adaptive estimator automatically makes the squared-bias/variance compromise and from an asymptotic point of view, reaches the optimal rate, provided that the constant in the penalty is known. In practice, the constant in the penalty, denoted above by κ, is found by simulation experiments (see Section 5). Note that Inequality (3.7) is nevertheless non-asymptotic.

3.2.3. *Random penalization of the subdensity estimator.* The penalty given in Theorem 3.1 can not be used in practice since it depends on the unknown quantity $\int_0^1 \psi(x)dx$. A solution is to use that $\int_0^1 \psi(x)dx \leq 1$; it follows that Inequality (3.7) would hold for a penalty defined by $\mathrm{pen}^\psi(m) = \kappa \Phi_0^2 D_m/n$. This possibly results in overestimation of the penalty, in a way depending on the unknown function ψ. The alternative solution is to replace the unknown quantity by an estimator (rather than a bound), and to prove that the estimator of ψ built with this random penalty retains the adaptation property of the theoretical estimator. This is described in the following theorem:

Theorem 3.2. *Assume that the assumptions of Theorem 3.1 are satisfied. Consider the estimator $\hat\psi_{\hat m}$ with $\hat m$ defined by $\hat m = \arg\min_{m\in\mathcal{M}_n}[\gamma_n^\psi(\hat\psi_m) + \widehat{\mathrm{pen}}^\psi(m)]$ and*

$$\widehat{\mathrm{pen}}^\psi(m) = \kappa\Phi_0^2\left(\frac{1}{n}\sum_{i=1}^n \mathbf{I}_{\{X_i\le 1\}}\mathbf{I}_{\{\delta_i=1\}}\right)\frac{D_m}{n}$$

where κ is a universal constant ($\kappa \ge 8$ suits). Then $\hat\psi_{\hat m}$ satisfies

$$\mathbb{E}(\|\hat\psi_{\hat m} - \psi\|^2) \le \inf_{m\in\mathcal{M}_n} K_0\left[\|\psi-\psi_m\|^2 + \Phi_0^2\left(\int_0^1\psi(x)dx\right)\frac{D_m}{n}\right] + \frac{K_1}{n}, \tag{3.8}$$

where K_0 is a universal constant (depending on κ) and K_1 depends on ψ, Φ_0.

We can derive from Inequality (3.8) in Theorem 3.2 some adaptation result to unknown smoothness:

Proposition 3.2. *Consider the collection of models* [T], [DP] *or* [W], *with $r > \alpha_\psi > 0$. Assume that an estimator $\tilde\psi$ of ψ satisfies inequality (3.8) in Theorem 3.2 (respectively inequality (3.7) in Theorem 3.1). Let $L > 0$. Then*

$$\left(\sup_{\psi\in\mathbb{B}_{\alpha_\psi,2,\infty}(L)}\mathbb{E}\|\psi-\tilde\psi\|^2\right)^{\frac12} \le C(\alpha_\psi, L)n^{-\frac{\alpha_\psi}{2\alpha_\psi+1}} \tag{3.9}$$

where $\mathbb{B}_{\alpha_\psi,2,\infty}(L) = \{t \in \mathcal{B}_{\alpha_\psi,2,\infty}([0,1]), |t|_{\alpha_\psi,2} \le L\}$ where $C(\alpha_\psi, L)$ is a constant depending on α_ψ, L and also on ψ, Φ_0.

3.3. **Application to the estimation of the hazard rate.** An estimator of the hazard rate h is deduced from $\hat\psi_{\hat m}$ by setting

$$\tilde h_\psi = \frac{\hat\psi_{\hat m}}{1-\hat F_n} \quad \text{with } \hat F_n(x) = \frac{1}{n+1}\sum_{i=1}^n \mathbf{I}_{\{X_i\le x\}}. \tag{3.10}$$

Then by using the decomposition

$$\tilde h_\psi - h = \left(\frac{\hat\psi_{\hat m}-\psi}{1-\hat F_n} + \psi(\frac{1}{1-\hat F_n} - \frac{1}{1-F})\right) \tag{3.11}$$

we find (see Section 6) the bound:

$$\mathbb{E}\|\tilde h_\psi - h\|^2 \le \frac{2^4}{c_F^2}\,\mathbb{E}\|\hat\psi_{\hat m}-\psi\|^2 + \frac{C(c_F,\|\psi\|)}{n}, \tag{3.12}$$

where $C(c_F, \|\psi\|)$ is a constant depending on c_F (defined in (2.1)) and $\|\psi\|$. From Inequality (3.12), we deduce by using results (3.7) or (3.8) that $\tilde h_\psi$ is an adaptive estimator of h if the functions h and ψ have the same regularity $\alpha = \alpha_h = \alpha_\psi$, or if the distribution function G of the censoring time is smoother than the hazard rate h (or the density f^0). An analogous result is found (without adaptation) in Antoniadis et al. (1999). Here the following Proposition holds:

Proposition 3.3. *Consider the collection of models* [T], [DP] *or* [W], *with* $r > \alpha_h = \alpha_\psi > 0$ *and the estimator* $\tilde{h}_\psi$ *defined by (3.10). Let* $L > 0$. *Then*

$$\left(\sup_{h\in\mathbb{B}_{\alpha_h,2,\infty}(L)} \mathbb{E}\|h-\tilde{h}_\psi\|^2\right)^{\frac{1}{2}} \leq C(\alpha_h, L)n^{-\frac{\alpha_h}{2\alpha_h+1}} \tag{3.13}$$

where $\mathbb{B}_{\alpha_h,2,\infty}(L) = \{t \in \mathcal{B}_{\alpha_h,2,\infty}([0,1]), |t|_{\alpha_h,2} \leq L\}$ *where* $C(\alpha_h, L)$ *is a constant depending on* α_h, L *and also on* ψ, Φ_0 *and* c_F.

The rate in Proposition 3.3 is known from Huber and Mac Gibbon (2004) to be optimal in the minimax sense. But it must be pointed out also that if the index of regularity of h, α_h, is greater than the index of regularity of ψ, α_ψ, then the asymptotic rate of the estimator $\tilde{h}_\psi$ is given by $n^{-\alpha_\psi/(1+2\alpha_\psi)}$ instead of the optimal one $n^{-\alpha_h/(1+2\alpha_h)}$. Clearly, the procedure is very simple and it is shown in Section 5 that the estimate of ψ behaves well, but it is not completely satisfactory as an estimator of h, even if none of the indexes are required to be known for implementing the procedure. This is due to the fact that the reference bias term here is $\|\psi - \psi_m\|$ instead of $\|h-h_m\|$. This is the reason why another contrast may be chosen in order to estimate h.

4. Study of the direct estimator of the hazard rate

In this section, we define the estimator of h and its adaptive version. The global line of the study is the same as for estimating ψ, even if some additional difficulties appear.

4.1. **Definition of the estimator of the hazard rate.** A direct estimator of the hazard rate can be obtained by considering the following contrast function:

$$\gamma_n^h(t) = \|t\|^2 - \frac{2}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i\leq 1\}}\frac{\delta_i t(X_i)}{1-\hat{F}_n(X_i)} \tag{4.1}$$

where $\hat{F}_n(x) = (1/(n+1))\sum_{i=1}^n \mathbf{I}_{\{X_i \le x\}}$. An estimator of h can be defined by

$$\hat{h}_m = \arg\min_{t\in S_m} \gamma_n^h(t). \tag{4.2}$$

Then $\hat{h}_m$ can be expressed as

$$\hat{h}_m = \sum_{\lambda\in\Lambda_m} \hat{a}_\lambda \varphi_\lambda \quad \text{with} \quad \hat{a}_\lambda = \frac{1}{n}\sum_{i=1}^n \mathbf{I}_{\{X_i\le 1\}} \frac{\delta_i \varphi_\lambda(X_i)}{1-\hat{F}_n(X_i)}. \tag{4.3}$$

If h_m denotes the orthogonal projection of h on S_m, we find that

$$h_m = \sum_{\lambda\in\Lambda_m} a_\lambda\varphi_\lambda \quad \text{with } a_\lambda = \int_0^1 h(x)\varphi_\lambda(x)dx.$$

4.2. **Study of the quadratic risk of the hazard rate estimator $\hat{h}_m$.** The rate of convergence of $\hat{h}_m$ derived by considering the following decomposition of the contrast:

$$\gamma_n^h(t) - \gamma_n^h(s) \quad = \quad \|t-h\|^2 - \|s-h\|^2 - 2\nu_n'(t-s) - 2\,R_n(t-s), \tag{4.4}$$

with the centered empirical process defined by

$$\nu_n'(t) = \frac{1}{n}\sum_{i=1}^n \left(\mathbf{I}_{\{X_i\le 1\}} \frac{\delta_i t(X_i)}{(1-F(X_i))} - \int_0^1 t(x)h(x)dx \right) \tag{4.5}$$

and the residual term

$$R_n(t) = \frac{1}{n}\sum_{i=1}^n \mathbf{I}_{\{X_i\le 1\}} \frac{\delta_i t(X_i)(F(X_i)-\hat{F}_n(X_i))}{(1-F(X_i))(1-\hat{F}_n(X_i))}.$$

This leads to the following result:

Proposition 4.1. *Consider the estimator $\hat{h}_m$ given by (4.2) or (4.3) where S_m is a D_m-dimensional linear space in a collection satisfying $(\mathcal{H}_1)$. Then*

$$\mathbb{E}(\|\hat{h}_m - h\|^2) \le 7\|h_m - h\|^2 + K\Phi_0^2 \frac{D_m}{n}\left(\int_0^1 \frac{h(x)}{1-F(x)}\,dx + \frac{1}{c_F^6}\right) \tag{4.6}$$

where K is a numerical constant.

Equation (4.6) gives the usual terms of the squared-bias / variance term decomposition, up to the constants and therefore, the usual rate $n^{-2\alpha_h/(2\alpha_h+1)}$ when $h \in \mathcal{B}_{\alpha_h,2,\infty}([0,1])$ for a choice of D_m of order $n^{1/(2\alpha_h+1)}$.

Proof of Proposition 4.1. By taking $t = \hat{h}_m$ and $s = h_m$ in (4.4), it follows from $\gamma_n^h(\hat{h}_m) \le \gamma_n^h(h_m)$ that:

$$\|\hat{h}_m - h\|^2 \le \|h_m - h\|^2 + 2\nu_n'(\hat{h}_m - h_m) + 2R_n(\hat{h}_m - h_m). \tag{4.7}$$

For the centered empirical process, the standard bound follows:

$$\begin{aligned} 2\,\mathbb{E}(|\nu_n'(\hat{h}_m - h_m)|) &\le \frac{1}{8}\,\mathbb{E}(\|\hat{h}_m - h_m\|^2) + 8\,\mathbb{E}\left(\sup_{t\in S_m, \|t\|=1} (\nu_n'(t))^2\right) \\ &\le \frac{1}{4}\,\mathbb{E}(\|\hat{h}_m - h\|^2) + \frac{1}{4}\,\|h_m - h\|^2 + 8\Phi_0^2 \frac{D_m}{n}\int_0^1 \frac{h(x)}{(1-F(x))}dx. \quad (4.8) \end{aligned}$$

For the residual term R_n, consider the set

$$\Omega_{c_F} = \{\omega, \forall x \in [0,1], F(x) - \hat{F}_n(x) > -c_F/2\} \tag{4.9}$$

where c_F is defined by (2.1), on which $1 - \hat{F}_n(x) = 1 - F(x) + F(x) - \hat{F}_n(x) \ge c_F/2$. Note that $1 - \hat{F}_n(x) \ge 1/(n+1)$ otherwise. Therefore:

$$\begin{aligned} 2\mathbb{E}(|R_n(\hat{h}_m - h_m)|\mathbf{I}_{\Omega_{c_F}}) &\le \frac{4\Phi_0\sqrt{D_m}}{c_F^2}\mathbb{E}(\|\hat{h}_m - h_m\|\|\hat{F}_n - F\|_\infty) \\ &\le \frac{1}{8}\mathbb{E}(\|\hat{h}_m - h_m\|^2) + 32\frac{\Phi_0^2 D_m}{c_F^4}\mathbb{E}(\|\hat{F}_n - F\|_\infty^2) \\ &\le \frac{1}{4}\mathbb{E}(\|\hat{h}_m - h\|^2) + \frac{1}{4}\|h_m - h\|^2 + 32\Phi_0^2 c_F^{-4} C_1 \frac{D_m}{n}, \quad (4.10) \end{aligned}$$

where C_1 is defined in Lemma 6.1. On the other hand, on the complement of Ω_{c_F},

$$\begin{aligned} 2\mathbb{E}(|R_n(\hat{h}_m - h_m)|\mathbf{I}_{\Omega_{c_F}^c}) &\le \frac{2\Phi_0(n+1)\sqrt{D_m}}{c_F}\mathbb{E}(\|\hat{h}_m - h_m\|\|\hat{F}_n - F\|_\infty \mathbf{I}_{\{\|F-\hat{F}_n\|_\infty > c_F/2\}}) \\ &\le \frac{1}{8}\mathbb{E}(\|\hat{h}_m - h_m\|^2) + 2^7\frac{\Phi_0^2 D_m (n+1)^2}{c_F^6}\mathbb{E}(\|\hat{F}_n - F\|_\infty^6) \\ &\le \frac{1}{4}\mathbb{E}(\|\hat{h}_m - h\|^2) + \frac{1}{4}\|h_m - h\|^2 + 2^8\Phi_0^2 C_3 c_F^{-6}\frac{D_m}{n}, \quad (4.11) \end{aligned}$$

where C_3 is defined in Lemma 6.1. By gathering (4.8), (4.10) and (4.11) together in (4.7), the result follows. □

4.3. **Adaptation with theoretical and random penalization.** As in the case of the study of ψ, the decomposition (4.6) shows that $\hat{h}_m$ can reach the minimax usual rate if the dimension D_m is relevantly chosen as a function of n and of the unknown smoothness of the function h. Therefore, here again, we need to build an adaptive estimator and the model selection procedure is:

$$\hat{m} = \arg\min_{m\in\mathcal{M}_n} \left(\gamma_n^h(\hat{h}_m) + \text{pen}^h(m)\right). \tag{4.12}$$

Then by using decomposition (4.7) again, the following theorem can be proved:

Theorem 4.1. *Assume that* $\sup_{x\in[0,1]} f(x) \le f_1$ *and the* S_m*'s are* D_m*-dimensional linear spaces in collections* [T], [DP] *or* [W] *with* $|\mathcal{M}_n| \le n$ *and* $N_n \le n/(16 f_1 K_\varphi)$ *for* [DP] *and* [W] *and* $N_n \le \sqrt{n}/(4\sqrt{f_1})$ *for* [T], *where* K_φ *is a constant depending on the basis only. Then the estimator* $\hat{h}_{\hat{m}}$ *with* $\hat{m}$ *defined by (4.12) and*

$$\text{pen}^h(m) = \kappa\Phi_0^2 \left(\int_0^1 \frac{h(x)}{1-F(x)}dx\right)\frac{D_m}{n}$$

where κ *is a universal constant* ($\kappa > 16$ *suits) satisfies*

$$\mathbb{E}(\|\hat{h}_{\hat{m}} - h\|^2 \le \inf_{m\in\mathcal{M}_n} \left(7\|h-h_m\|^2 + 8\text{pen}^h(m)\right) + \frac{C\sqrt{\ln(n)}}{n}, \tag{4.13}$$

where C *is a constant depending on* f_1, Φ_0, $c_F = \inf_{x\in[0,1]}(1-F(x))$.

As for ψ, the penalty given in Theorem 4.1 can not be used in practice since it depends on the unknown quantity $\int_0^1 h(x)/(1-F(x))dx$. Therefore we replace this quantity by an estimator. We can then prove that the estimator of h built with this random penalty retains the adaptation property of the theoretical estimator. The idea is that $\int_0^1 h(x)/(1-F(x))dx$ is the second order moment of the independent random variables $\mathbf{I}_{\{\delta_i=1, X_i\le 1\}}/(1-F(X_i))$ and therefore can be estimated by

$$\hat{s}_2 = \frac{1}{n}\sum_{i=1}^n \frac{\mathbf{I}_{\{X_i\le 1\}}\delta_i}{(1-\hat{F}_n(X_i))^2}, \quad \hat{F}_n(x) = \frac{1}{n+1}\sum_{i=1}^n \mathbf{I}_{\{X_i\le x\}}. \tag{4.14}$$

The following result holds:

Corollary 4.1. *Under the assumptions of Theorem 4.1 but with* $\text{pen}^h(m)$ *replaced by*

$$\widehat{\text{pen}}^h(m) = \kappa\Phi_0^2\hat{s}_2\frac{D_m}{n}$$

where $\hat{s}_2$ *is defined by (4.14) and* κ *is a universal constant* ($\kappa > 16$ *suits), the estimator* $\hat{h}_{\hat{m}}$ *satisfies (4.13).*

Then the result given for the "ratio" estimator $\tilde{h}_\psi$ in Proposition 3.3 holds for $\tilde{h} = \hat{h}_{\hat{m}}$ defined in Theorem 4.1 without the restrictive condition $\alpha_h = \alpha_\psi$. Therefore this second estimator can automatically reach the optimal minimax rate in all cases. However, note that the condition $\alpha_h > 0$ for [DP] and [W] becomes $\alpha_h > 1/2$ for collection [T] since in that case, it follows from Theorem 4.1 that $N_n \leq O(\sqrt{n})$.

4.4. **Adaptive estimation of the hazard rate with a general collection of models.** Non regular models are of great interest to capture some variability in the regularity of the function to be estimated and are in particular applied for the detection of ruptures. Moreover, it is likely to be well suited for hazard rate estimation because of the frequent scarsity of the observations at the end of the time interval. Extension of the results to such a framework is possible under some mild restrictions. Some adaptation results are proved by Reynaud-Bouret (2002) for non regular histograms provided that knots are chosen in a set of cardinality less than $n/\ln^2(n)$. We can prove here a similar result, with a similar constraint, with the additional advantage that we consider a more general collection than histograms, namely piecewise polynomials. Let $K_n = N_n/(r+1)$ where N_n is the dimension of the largest space of the collection. We consider the set of knots $\Gamma = \{\ell/K_n, \quad \ell = 1, \ldots, K_n - 1\}$. A general piecewise polynomial model, non necessarily regular, is then defined by the maximal degree r of the polynomials and a set of knots

$$\{a_0 = 0, a_1, \ldots, a_\ell, a_{\ell+1} = 1\}$$

where $\{a_1, \ldots a_\ell\}$ is any subset of $\Gamma:$ its dimension is $D_m = (\ell+1)(r+1)$. This means that with a non regular collection, for any fixed dimension D_m, there is $\begin{pmatrix} K_n - 1 \\ \ell \end{pmatrix}$ associated models corresponding to the possible choices of the subset $\{a_1, \ldots a_\ell\}$ with $\ell = 1, \ldots, K_n - 1$. Therefore, the cardinality of the set $\mathcal{M}_n$ of all possible m is:

$$\sum_{\ell=1}^{K_n-1} \begin{pmatrix} K_n - 1 \\ \ell \end{pmatrix} = 2^{K_n-1} - 1 = \frac{1}{2}\exp(K_n \ln(2)) - 1\,.$$

Since K_n is of order n, it is exponentially great. In any case, it is much greater than the order $O(n)$ obtained in the regular case, when only one model per dimension is considered. The φ_λ's for $\lambda = (a_j, a_{j+1}; k) \in \Lambda_m$ are given by

$$\sqrt{\frac{2k+1}{a_j - a_{j-1}}} Q_k \left(\frac{2}{a_j - a_{j-1}} x - \frac{a_j + a_{j-1}}{a_j - a_{j-1}} \right) \mathbf{I}_{[a_{j-1}, a_j[}(x)$$

for $k = 0, 1, \ldots, r$ and $j = 0, \ldots, \ell + 1$, where Q_k denotes the kth Legendre polynomial. We call [GP] this general collection of piecewise polynomials.

Note that the key property (2.2) no longer holds for this collection. We can nevertheless prove the following result:

Theorem 4.2. *Let S_m be a D_m-dimensional linear space in the general collection of piecewise polynomials* [GP] *with $N_n \leq Kn/\ln(n)$ for a given constant K. Then the estimator $\hat{h}_{\hat{m}}$ with $\hat{m}$ defined by (4.12) and*

$$\text{pen}(m) = \kappa \sup_{x \in [0,1]} \left(\frac{h(x)}{1 - F(x)} \right) \frac{D_m(1 + \ln(n))}{n} \tag{4.15}$$

where κ is a universal constant satisfies

$$\mathbb{E}(\|\hat{h}_{\hat{m}} - h\|^2 \leq \inf_{m \in \mathcal{M}_n} \left(3\|h - h_m\|^2 + 4\text{pen}(m) \right) + \frac{K}{n}, \tag{4.16}$$

where K is a constant depending on h, F and r.

Note first that the additional $\ln(n)$ factor in the penalty implies a $\ln(n)$ factor loss in the resulting asymptotic rates which become of order $O((n/\ln(n))^{-2\alpha_h/(2\alpha_h+1)})$ in the best case. This is the price to pay for considering such a huge collection of models. Secondly, the unknown term in the penalty is now $\sup_{x\in[0,1]}(h(x)/(1-F(x))$ and must be replaced by an estimator as in the previous subsection. For instance choose $\sup_{x\in[0,1]}[\hat{h}_n(x)/(1 - \hat{F}_n(x))]$ where $\hat{h}_n$ is a given estimator of h on a well chosen regular space of piecewise polynomials and $\hat{F}_n$ is the empirical distribution of the data as defined above.

5. Simulations and examples

5.1. **Examples.** For simulated observations in $[0, a]$ (where a is the maximum of the observed X_i's), we compute one estimator of ψ and two estimators of h. For the sake of simplicity, we use here the trigonometric basis and this yields:

$$\hat{\psi}_D(x) = \frac{\hat{a}_1}{\sqrt{a}} + \sqrt{\frac{2}{a}} \sum_{k=1}^{[D/2]} \hat{a}_{2k} \cos\left(\frac{2\pi k x}{a}\right) + \sqrt{\frac{2}{a}} \sum_{k=1}^{[(D-1)/2]} \hat{a}_{2k+1} \sin\left(\frac{2\pi k x}{a}\right)$$

with

$$\hat{a}_1 = \frac{1}{n\sqrt{a}}\sum_{i=1}^{n}\delta_i, \hat{a}_{2k} = \frac{\sqrt{2}}{n\sqrt{a}}\sum_{i=1}^{n}\delta_i \cos\left(\frac{2k\pi X_i}{a}\right), \hat{a}_{2k+1} = \frac{\sqrt{2}}{n\sqrt{a}}\sum_{i=1}^{n}\delta_i \sin\left(\frac{2k\pi X_i}{a}\right).$$

Then we select $\hat{D}_\psi$ such that the penalized contrast, equal to $-\sum_{j=1}^{D}\hat{a}_j^2 + \kappa_\psi \hat{s}_2^\psi D/n$ is minimal. Here $\hat{s}_2^\psi = \sum_{i=1}^{n}\delta_i/n$ and κ_ψ is a well chosen constant. Given $\hat{\psi}_{\hat{D}_\psi} = \tilde{\psi}$, the estimator of h is defined as $\tilde{h}_\psi = \tilde{\psi}/(1-\hat{F})$ (cf. (3.10)) where we took $\hat{F}(x) = (1/(n+5))\sum_{i=1}^{n}\mathbf{1}_{\{X_i \le x\}}$. The other estimator is computed as

$$\hat{h}_D(x) = \frac{\hat{b}_1}{\sqrt{a}} + \sqrt{\frac{2}{a}}\sum_{k=1}^{[D/2]}\hat{b}_{2k}\cos\left(\frac{2\pi kx}{a}\right) + \sqrt{\frac{2}{a}}\sum_{k=1}^{[(D-1)/2]}\hat{b}_{2k+1}\sin\left(\frac{2\pi kx}{a}\right)$$

with

$$\hat{b}_1 = \frac{1}{n\sqrt{a}}\sum_{i=1}^{n}\frac{\delta_{(i)}}{1-\hat{F}(X_{(i)})}, \hat{b}_{2k} = \frac{\sqrt{2}}{n\sqrt{a}}\sum_{i=1}^{n}\frac{\delta_{(i)}}{1-\hat{F}(X_{(i)})}\cos\left(\frac{2k\pi X_{(i)}}{a}\right),$$

$$\hat{b}_{2k+1} = \frac{\sqrt{2}}{n\sqrt{a}}\sum_{i=1}^{n}\frac{\delta_{(i)}}{1-\hat{F}(X_{(i)})}\sin\left(\frac{2k\pi X_{(i)}}{a}\right),$$

where we took $\hat{F}(X_{(i)}) = \max(1.5, \min(1/(1-i/(n+5)), \sqrt{n})$. Here $X_{(i)}$ is the i-th order statistic for the sample $(X_1, \dots, X_n)$ and $\delta_{(i)}$ is the induced order statistic corresponding to $X_{(i)}$. Note that all denominators $1-i/(n+1)$ have been replaced by $1-i/(n+\log_2(n))$ for numerical reason. Then we select $\hat{D}_h$ such that the penalized contrast, equal to $-\sum_{j=1}^{D}\hat{b}_j^2 + \kappa_h \hat{s}_2 D/n$ is minimal. Here $\hat{s}_2 = (1/n)\sum_{i=1}^{n}\delta_{(i)}/(1-i/(n+\log_2(n)))^2$ and κ_h is a well chosen constant. Then $\tilde{h} = \hat{h}_{\hat{D}_h}$ is the other estimator of h.

Here the penalty is calibrated with $\kappa_\psi = \kappa_h = 0.5$ and following the findings of Birgé and Rozenholc (2002) for density estimation, we replace the term D/n in both penalties by $(D + \ln(D)^{2.5})/n$. Moreover, we removed the four (for $n = 200$) or five (for $n = 500$) largest values of the sample before computing the estimators and took $\hat{F}(x) = (1/(n+\log_2(n)))\sum_{i=1}^{n}\mathbf{1}_{\{X_i \le x\}}$ to avoid some instability problems due to the divisions.

For simulating the data, we consider two cases:

(a) The first set of simulations is called in the following the “Gamma case”. The X_i^0’s are generated from a Gamma distribution with shape parameter 5 and scale 1 and the independent C_i’s from an exponential distribution with mean 6.

(b) The second set is called "the bimodal case". The X_i^0's have a bimodal density defined by

$$f^0 = 0.8u + 0.2v$$

where u is the density of $\exp(Z/2)$ with $Z \sim \mathcal{N}(0,1)$ and $v = 0.17Z + 2$. The $C_i's$ are generated from an exponential distribution with mean 2.5.

Examples of estimation are given in Figures 1, 2 and 3. Figure 1 illustrates the interest of a relevant selection of D (D too small implies a curve too flat, D too great implies too much variance). Most plots for the estimation of h are usually truncated since it is well known that the estimation is often bad at the end of the interval. We kept the whole interval of estimation to illustrate it, but it is clear that a plot stopping at $x = 6$ for Figure 2 and at $x = 2$ for Figure 3 would look better (see Antoniadis et al. (1999, Fig. 1 and 2)) and avoid the problem of sparsity of the observations at the end of the interval. Moreover it appears that the estimators behave in very different ways, in particular at the end of the intervals.

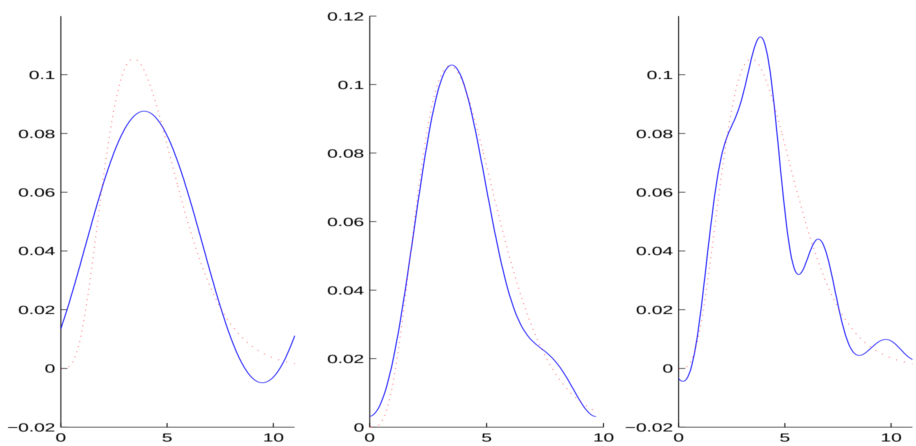

FIGURE 1. Estimation of ψ for $n = 500$ in model (a), with different dimensions of the projection space (left: $D = 3$, center: $D = 5$, right: $D = 13$). .....: true function ψ, ——: estimated ψ, $\hat{\psi}_D$.

5.2. **Simulation results.** The two examples (a) and (b) have been studied by Antoniadis et al. (1999) (wavelet estimator with selection of the coefficients by cross-validation) and Reynaud-Bouret (2002) (histogram and Fourier estimators of the

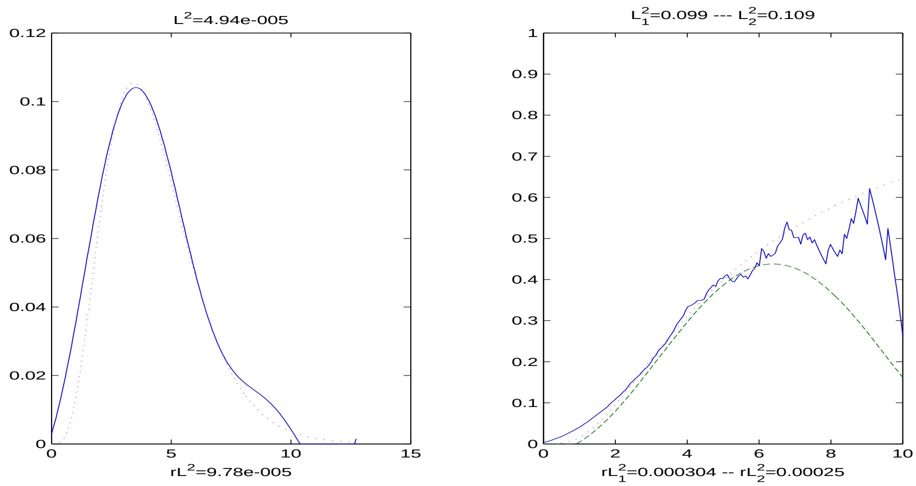

FIGURE 2. Estimation of ψ and h in the Gamma case (a).
Dotted: true function (left: ψ, right: h), continuous: $\tilde{\psi}$ (left) and $\tilde{h}_\psi$ (right), dashed: $\tilde{h}$ (right). L^2: MSE, rL^2: restricted MSE2 (for $\tilde{\psi}$ (left), or for $\tilde{h}_\psi$ (right, index 1) and for $\tilde{h}$ (right, index 2).

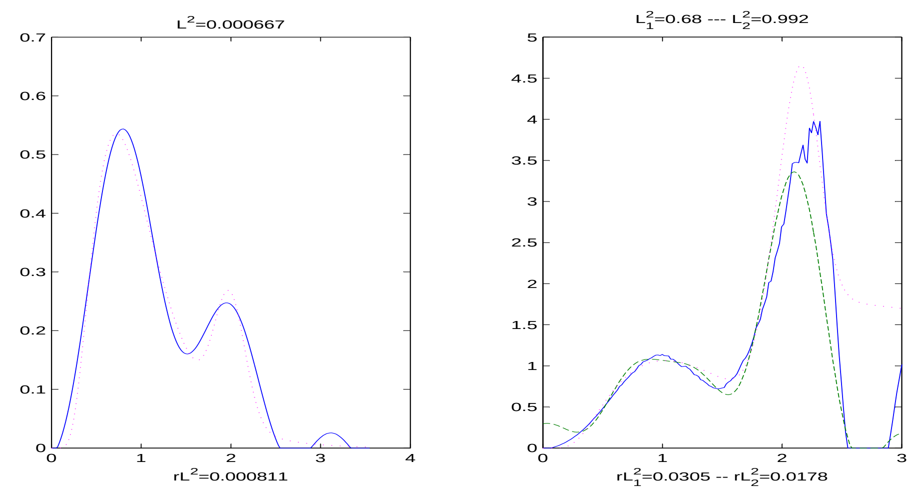

FIGURE 3. Estimation of ψ and h in the Bimodal case (b).
Dotted: true function (left: ψ, right: h), continuous: $\tilde{\psi}$ (left) and $\tilde{h}_\psi$ (right), dashed: $\tilde{h}$ (right). L^2: MSE, rL^2: restricted MSE2 (for $\tilde{\psi}$ (left), or for $\tilde{h}_\psi$ (right, index 1) and for $\tilde{h}$ (right, index 2).

Aalen intensity). Antoniadis et al. (1999) estimate both the subdensity and the hazard rate, whereas Reynaud-Bouret (2002) estimates h only. We compare our results to theirs in Table 1 for ψ and in Tables 2 and 3 for h.

The authors give the mean squared errors of their estimator computed over $T = 200$ replications of samples of size $n = 200$ and $n = 500$. The error is computed over K regularly spaced points t_k, $k = 1, \dots, K$, of the interval in which the X_i's fall ($[0, \max X_i]$), as the mean over the replications j of

$$\mathrm{MSE}_j = \frac{1}{K} \sum_{k=1}^{K} (h(t_k) - \hat{h}_j(t_k))^2$$

where $\hat{h}_j$ is the estimate of h for the sample number j, $j = 1, \dots, T$.

	Estimator of ψ of Antoniadis et al.				Our estimator of ψ, $\tilde{\psi}$.			
Model	Gamma		Bimodal		Gamma		Bimodal	
$n =$	200	500	200	500	200	500	200	500
MSE $\times 10^5$	18.5	6.7	369	263	17.3	8.52	340	130

TABLE 1. Results of Antoniadis et al. (1999, Table 2) and of our estimator for the estimation of ψ, $T = 200$ replications.

In order to take into account the sparsity of the observations at the end of the interval, ($\mathbb{P}(X^0 > 6) = 0.25$ in the Gamma case and $\mathbb{P}(X^0 > 2) = 0.16$ in the bimodal case), they 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. For $K = 64$, Table 1 compares our results to those of Antoniadis et al. (1999). Note that we consider only the global MSE here since we can see from Figures 2 and 3 that there is not any problem at the end of the interval for ψ. We can see that the results have the same orders, our estimator seems slightly better (three of our MSE values out of four are better with our estimator). A larger set of examples should be studied to get more conclusive evidence.

Reynaud-Bouret (2002) studies a Fourier strategy for an adaptive estimator based on a contrast different from ours and for the same kind of data. We expect to obtain results of the same order, but with a theoretically simpler tool. Her results and those of Antoniadis et al. (1999) are recalled in Table 2, while ours are given in Table 3.

	Estimator of Antoniadis et al.				Estimator of Reynaud-Bouret			
Model	Gamma		Bimodal		Gamma		Bimodal	
$n =$	200	500	200	500	200	500	200	500
MSE	0.112	0.0995	2.080	1.970	0.055	0.0579	1.259	1.122
MSE2	0.0025	0.0016	0.048	0.032	0.0032	0.0012	0.150	0.051

Table 2. Results of Antoniadis et al. (1999, Table 2) and of the Fourier strategy in Reynaud-Bouret (2002, Table 10) for the estimation of h, $T = 200$ replications

	$\tilde{h}_\psi$				$\tilde{h}$			
Model	Gamma		Bimodal		Gamma		Bimodal	
$n =$	200	500	200	500	200	500	200	500
MSE	0.0857	0.0900	0.902	0.706	0.0800	0.0986	1.117	1.140
MSE2	0.0023	0.0013	0.1068	0.0408	0.0091	0.0017	0.145	0.059

Table 3. Our results for our two estimators of h, $T = 200$ replications.

The first remark is that, surprisingly, $\tilde{h}_\psi$ is always slightly better than $\tilde{h}$. This is probably due to the fact that the ratio strategy works well for very regular functions. In term of the MSE, we can see that $\tilde{h}_\psi$ is always better than the estimator of Antoniadis et al. (1999), whereas in term of the MSE2, we have the same orders. This shows that our estimator is better with respect to the global interval and comparable when the end of the interval is cut. The comparison with Reynaud-Bouret (2002) shows that $\tilde{h}_\psi$ is better for both MSE and MSE2 for the second model and slightly worse for the first one, so that as expected, we have globally the same orders of errors. Note that the MSE2 given by Reynaud-Bouret (2002) for the Fourier Strategy are the lowest she obtains over all the strategies she experimented. Thus our estimator is superior to all her other strategies for estimating h in terms of MSE2. The other strategies used by Reynaud-Bouret (2002) may be studied in our case also, in particular by computing our projection estimator in the piecewise polynomial basis, with local selection of the degree on each bin. But this is beyond the scope of the present work.

As a conclusion, it appears that the estimator $\tilde{h}_\psi$ is a very good estimator, at least for the regular functions h considered here. Moreover $\tilde{h}$ obtains quite good results that may be improved by regularization of $\hat{F}$.

6. Proofs

6.1. **Preliminary results.**

6.1.1. *A useful Lemma.* First, we give a lemma used several times in the paper.

Lemma 6.1. *For all $k \in \mathbb{N}^*$, $\mathbb{E}\left(\|\hat{F}_n - F\|_\infty^{2k}\right) \le \frac{C_k}{n^k}$, with $C_k = 2^{k-1}(1 + 2\,k!)$ and where $\hat{F}_n(x) = n/(n+1)F_n(x)$ and $F_n(x)$ stands for the standard empirical distribution function of the X_i's.*

Proof. $\mathbb{E}\left(\|F_n - F\|_\infty^{2k}\right) \le 2k \int_0^{+\infty} x^{2k-1}\,\mathbb{P}(\|F_n - F\|_\infty > x)\,dx$. Now, it follows from Massart (1990) that $\forall \lambda > 0, \mathbb{P}(\sqrt{n}\|F_n - F\|_\infty \ge \lambda) \le 2e^{-2\lambda^2}$, where $F_n(x) = (1/n)\sum_{i=1}^n \mathbf{1}_{\{X_i \le x\}}$. This gives after successive changes of variables

$$\begin{aligned}
\mathbb{E}\left(\|F_n - F\|_\infty^{2k}\right) &\le 4k \int_0^{+\infty} x^{2k-1}\,e^{-2nx^2} dx = \frac{4k}{(2n)^k}\int_0^{+\infty} x^{2k-1}\,e^{-x^2}\,dx \\
&= \frac{2k}{(2n)^k}\int_0^{+\infty} x^{k-1}\,e^{-x}\,dx = \frac{2k}{(2n)^k}\Gamma(k) = \frac{2\,k!}{(2n)^k}.
\end{aligned}$$

Moreover,

$$\begin{aligned}
\mathbb{E}(\|\hat{F}_n - F\|_\infty^{2k}) &\le 2^{2k-1}\,\mathbb{E}\left(\|\hat{F}_n - F_n\|_\infty^{2k} + \|F_n - F\|_\infty^{2k}\right) \\
&\le 2^{2k-1}\left(\left(\frac{1}{n+1}\right)^{2k} + \frac{2\,k!}{(2n)^k}\right) \le 2^{2k-1}\,\frac{1 + 2\,k!}{(2n)^k}.
\end{aligned}$$

□

6.1.2. *Talagrand's Theorem.* Most of the proofs are based on the use of the following version of Talagrand's Inequality (see Talagrand (1996)):

Theorem 6.1. *Let $Z_1, \ldots, Z_n$ be i.i.d. random variables and $\nu_n(g)$ be defined by $\nu_n(g) = (1/n)\sum_{i=1}^n [g(Z_i) - \mathbb{E}(g(Z_i))]$ for g belonging to a countable class $\mathcal{G}$ of uniformly bounded measurable functions. Then for $\epsilon > 0$*

(6.1)
$$\mathbb{E}\left[\sup_{g\in\mathcal{G}} |\nu_n(g)|^2 - 2(1+2\epsilon)H^2\right]_+ \le \frac{6}{K_1}\left(\frac{v}{n}e^{-K_1\epsilon\frac{nH^2}{v}} + \frac{8M_1^2}{K_1 n^2 C^2(\epsilon)}e^{-\frac{K_1 C(\epsilon)\sqrt{\epsilon}}{\sqrt{2}}\frac{nH}{M_1}}\right),$$

with $C(\epsilon) = (\sqrt{1+\epsilon} - 1) \wedge 1$, K_1 *is a universal constant, and where*

$$\sup_{g\in\mathcal{G}} \|g\|_\infty \le M_1, \quad \mathbb{E}\left(\sup_{g\in\mathcal{G}} |\nu_n(g)|\right) \le H, \quad \sup_{g\in\mathcal{G}} \mathrm{Var}(g(X_1)) \le v.$$

6.2. **Proofs of Theorem 3.1 and 3.2.**

6.2.1. *Proof of a preliminary Lemma.* First, we prove the following lemma which is useful for the proofs of the first two Theorems:

Lemma 6.2. *Assume that* $(\mathcal{H}_1)$ *and* $(\mathcal{H}_2)$ *are fulfilled and denote by* $B_{m,m'}(0,1) = \{t \in S_m + S_{m'},\ \ \|t\| = 1\}$. *Let* $\nu_n(g_t)$ *be the centered empirical process with* $g_t(x,y) = \mathbf{I}_{\{y=1\}\cap\{x\le 1\}} t(x) - \int_0^1 t(x)\psi(x)dx$, *then for* $\epsilon > 0$

$$\mathbb{E}\left(\sup_{t\in B_{m,m'}(0,1)} \nu_n^2(g_t) - p^\psi(m,m')\right)_+ \le \frac{\kappa_1}{n}\left(e^{-\kappa_2\epsilon(D_m+D_{m'})} + \frac{e^{-\kappa_3\epsilon^{3/2}\sqrt{n}}}{C(\epsilon)^2}\right), \tag{6.2}$$

with $p^\psi(m,m') = 2(1+2\epsilon)\Phi_0^2 \int_0^1 \psi(x)dx\,(D_m + D_{m'})/n$ *and* $C(\epsilon) = (\sqrt{1+\epsilon} - 1) \wedge 1$. *The constants* κ_i *for* $i = 1,2,3$ *depend on* Φ_0, ψ *and* F.

We apply Talagrand's inequality (6.1) by taking $Z_i = (X_i, \delta_i)$ for $i = 1, \ldots, n$ and $g(x,y) = g_t(x,y)$ defined by $\nu_n(t) = \nu_n(g_t) := (1/n)\sum_{i=1}^n g_t(X_i, \delta_i)$. Usual density arguments show that this result can be applied to the class of functions $\mathcal{G} = \{g_t,\ t \in B_{m,m'}(0,1)\}$. Then we find for the present empirical process the following bounds: $\sup_{g\in\mathcal{G}} \|g\|_\infty = \sup_{t\in B_{m,m'}(0,1)} \|g_t\|_\infty \le \Phi_0\sqrt{D(m')} := M_1$ with $D(m')$ denoting the dimension of $S_m + S_{m'}$. Then $\sup_{g\in\mathcal{G}} \mathrm{Var}(g(X_1,\delta_1)) = \sup_{t\in B_{m,m'}(0,1)} \mathrm{Var}(g_t(X_1,\delta_1)) \le 1 := v$. Lastly,

$$\begin{aligned}
\mathbb{E}\left(\sup_{g\in\mathcal{G}} \nu_n^2(g)\right) &= \mathbb{E}\left(\sup_{t\in B_{m,m'}(0,1)} \nu_n^2(g_t)\right) \le \sum_{\lambda\in\Lambda_{m,m'}} \frac{1}{n}\mathrm{Var}(\mathbf{I}_{\{X_1\le 1\}}\delta_1\varphi_\lambda(X_1)) \\
&\le \frac{\Phi_0^2 D(m')}{n}\int_0^1 \psi(x)dx = C_1\frac{D(m')}{n} := H^2.
\end{aligned}$$

with the natural notation $\Lambda_{m,m'} = \Lambda_m \cup \Lambda_{m'}$. Then it follows from (6.1) that

$$\mathbb{E}\left(\sup_{t\in B_{m,m'}(0,1)} \nu_n^2(g_t) - p^\psi(m,m')\right) \le \kappa_1\left(\frac{1}{n}e^{-\kappa_2\epsilon D(m')} + \frac{1}{nC^2(\epsilon)}e^{-\kappa_3\epsilon^{3/2}\sqrt{n}}\right),$$

where κ_i for $i = 1,2,3$ are constants depending on K_1 and C_1 and $p^\psi(m,m') = 2(1+2\epsilon)C_1(D_m + D_{m'})/n$. □

6.2.2. *Proof of Theorem 3.1.* It follows from the definition of $\hat{\psi}_{\hat{m}}$ that: $\forall m \in \mathcal{M}_n$,

$$\gamma_n^\psi(\hat{\psi}_{\hat{m}}) + \mathrm{pen}^\psi(\hat{m}) \leq \gamma_n^\psi(\psi_m) + \mathrm{pen}^\psi(m). \tag{6.3}$$

Then, by using the decomposition $\gamma_n^\psi(t) - \gamma_n^\psi(s) = \|t-\psi\|^2 - \|s-\psi\|^2 - 2\nu_n(g_{t-s})$, it follows that

$$\gamma_n(\hat{\psi}_{\hat{m}}) - \gamma_n(\psi_m) = \|\hat{\psi}_{\hat{m}} - \psi\|^2 - \|\psi_m - \psi\|^2 - 2\nu_n(g_{\hat{\psi}_{\hat{m}}-\psi_m}),$$

where the process $\nu_n(g_t)$ is the same as previously. Then, by applying (6.3) and by noticing that $t \mapsto \nu_n(g_t)$ is linear, we get that:

$$\begin{aligned} \|\hat{\psi}_{\hat{m}} - \psi\|^2 &\leq \|\psi_m - \psi\|^2 + 2\nu_n(g_{\hat{\psi}_{\hat{m}}-\psi_m}) + \mathrm{pen}^\psi(m) - \mathrm{pen}^\psi(\hat{m}) \\ &\leq \|\psi_m - \psi\|^2 + 2\|\hat{\psi}_{\hat{m}} - \psi_m\| \sup_{t\in B_{m,\hat{m}}(0,1)} \nu_n(g_t) + \mathrm{pen}^\psi(m) - \mathrm{pen}^\psi(\hat{m}) \\ &\leq \|\psi_m - \psi\|^2 + \frac{1}{x}\|\hat{\psi}_{\hat{m}} - \psi_m\|^2 + x \sup_{t\in B_{m,\hat{m}}(0,1)} \nu_n^2(g_t) + \mathrm{pen}^\psi(m) - \mathrm{pen}^\psi(\hat{m}) \qquad (6.4) \end{aligned}$$

where we use that for a, b positive, $2ab \leq xa^2 + x^{-1}b^2$ for any positive x. We recall that $B_{m,\hat{m}}(0,1) = \{t \in S_m + S_{\hat{m}} \ / \ \|t\| \leq 1\}$. Note that the norm connection as described by (2.2) still holds for any element t of $S_m + S_{m'}$ as follows: $\|t\|_\infty \leq \Phi_0 \max(D_m, D_{m'})\|t\|$. Indeed, under $(\mathcal{H}_2)$, we restrict our attention to nested collections of models, so that $S_m + S_{\hat{m}}$ is equal to the largest of the two spaces. For a fixed integer m, we denote by $D(m')$ the dimension of $S_m + S_{m'}$, for all $m' \in \mathcal{M}_n$. Note that $D(m') = \max(D_m, D_{m'}) \leq D_m + D_{m'}$.
Let $p^\psi(m,m')$ be such that

$$xp^\psi(m,m') \leq \mathrm{pen}^\psi(m) + \mathrm{pen}^\psi(m') \quad \text{for all } m,\ m' \text{ in } \mathcal{M}_n. \tag{6.5}$$

We have $\|\hat{\psi}_{\hat{m}} - \psi_m\|^2 \leq (1+y^{-1})\|\hat{\psi}_{\hat{m}} - \psi\|^2 + (1+y)\|\psi_m - \psi_m\|^2$ for any positive y. Then by choosing $y = (x+1)/(x-1)$ and $x > 1$, we find that: $\forall m \in \mathcal{M}_n$,

$$\|\hat{\psi}_{\hat{m}} - \psi\|^2 \leq C_x^2\|\psi - \psi_m\|^2 + 2C_x \mathrm{pen}^\psi(m) + xC_x\left(\sup_{t\in B_{m,\hat{m}}(0,1)} \nu_n^2(g_t) - p^\psi(m,\hat{m})\right),$$

where $C_x = (x+1)/(x-1)$. Then if we prove
(6.6)

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

we have the following result, which proves the theorem: $\forall m \in \mathcal{M}_n$,

$$\mathbb{E}(\|\hat{\psi}_{\hat{m}} - \psi\|^2) \leq C_x^2 \|\psi - \psi_m\|^2 + 2C_x \text{pen}^\psi(m) + \frac{C}{n}.$$

Therefore by using equation (6.5) and the definition of $p^\psi(m, m')$ in Lemma 6.2, we choose $\text{pen}^\psi(m) = 2x(1+2\epsilon)\int_0^1 \psi(u)du(D_m/n)$. Inequality (6.6) is a straightforward consequence of Lemma 6.2 since

$$\sum_{m'\in\mathcal{M}_n} \mathbb{E}\left(\sup_{t\in B_{m,m'}(0,1)} \nu_n^2(g_t) - p^\psi(m,m')\right)_+ \leq \kappa_1 \left(\frac{\sum_{m'\in\mathcal{M}_n} e^{-\kappa_2 \epsilon D(m')}}{n} + \frac{|\mathcal{M}_n|}{n} e^{-\kappa_3 \epsilon^{3/2}\sqrt{n}}\right).$$

Then by taking $\epsilon = 1/2$ and assuming that $|\mathcal{M}_n| \leq n$ and since, under $(\mathcal{H}_2)$, $\sum_{m\in\mathcal{M}_n} e^{-aD_m} \leq \sum_{k=1}^n e^{-ka} < +\infty, \forall a > 0$, and this ensures (6.6). □

6.3. **Proof of Theorem 3.2.** Let

$$\Omega_b = \left\{ \left| \left(\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{X_i \leq 1\}}\delta_i \right) / \left(\int_0^1 \psi(u)du \right) - 1 \right| < b \right\}, \quad 0 < b < 1.$$

Then on Ω_b, the proof is quite similar to the proof of Theorem 3.1 and is omitted. It is based on the following inequalities:

$$\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{X_i\leq 1\}}\delta_i < (b+1)\int_0^1 \psi(u)du, \quad \int_0^1 \psi(u)du < \frac{1}{1-b}\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{X_i\leq 1\}}\delta_i .$$

We obtain for $C_x = (x+1)/(x-1)$, that $\forall m \in \mathcal{M}_n$,

$$\mathbb{E}\left(\|\hat{\psi}_{\hat{m}} - \psi\|^2 \mathbf{1}_{\Omega_b}\right) \leq C_x^2 \|\psi_m - \psi\|^2 + \frac{2C_x\Phi_0^2}{(1-b)}\left(\int_0^1 \psi(u)du\right)\frac{D_m}{n} + \frac{K}{n},$$

for $\widehat{\text{pen}}^\psi(m) = 2x(1+2\epsilon)/(1-b)\Phi_0^2((1/n)\sum_{i=1}^n \mathbf{1}_{\{X_i\leq 1\}}\delta_i)(D_m/n)$. Next we need to prove that

$$\mathbb{E}\left(\|\hat{\psi}_{\hat{m}} - \psi\|^2 \mathbf{1}_{\Omega_b^c}\right) \leq \frac{K'}{n}. \tag{6.7}$$

By analogy with (6.4), it follows from $\gamma_n^\psi(\hat{\psi}_{\hat{m}}) \leq \gamma_n^\psi(\psi_{\hat{m}})$ that $\|\hat{\psi}_{\hat{m}} - \psi\|^2 \leq \|\psi - \psi_{\hat{m}}\|^2 + 2\nu_n(g_{\hat{\psi}_{\hat{m}}-\psi_{\hat{m}}}) \leq \|\psi - \psi_{\hat{m}}\|^2 + (1/4)\|\hat{\psi}_{\hat{m}} - \psi_{\hat{m}}\|^2 + 4\sup_{t\in S_{\hat{m}}, \|t\|=1} \nu_n^2(g_t)$ that is

$$\|\hat{\psi}_{\hat{m}} - \psi\|^2 \leq 3\|\psi\|^2 + 8 \sup_{t\in S_{\hat{m}}, \|t\|=1} \nu_n^2(g_t). \tag{6.8}$$

Then $\sup_{t\in S_{\hat m},\|t\|=1}\nu_n^2(g_t) \le \sum_{m\in\mathcal{M}_n}\left(\sup_{t\in S_m,\|t\|=1}\nu_n^2(g_t)-\mathrm{pen}^\psi(m)\right)_+ + \mathrm{pen}^\psi(\hat m)$. We know by Lemma 6.2 that, for some well chosen κ in $\mathrm{pen}^\psi(m)$,

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

On the other hand, $\forall m\in\mathcal{M}_n$, $\mathrm{pen}^\psi(m)\le K'$, with $K'=\kappa\Phi_0^2\int_0^1\psi(u)\,du$, so that $\mathbb{E}\left(\mathrm{pen}^\psi(\hat m)\mathbf{I}_{\Omega_b^c}\right)\le K'\,\mathbb{P}(\Omega_b^c)$. Therefore

$$\mathbb{E}\left(\sup_{t\in S_{\hat m},\|t\|=1}\nu_n(g_t)^2\mathbf{I}_{\Omega_b^c}\right)\le\frac{K}{n}+K'\mathbb{P}(\Omega_b^c), \tag{6.9}$$

so that, by gathering (6.8) and (6.9), Inequality (6.7) holds provided that $\mathbb{P}(\Omega_b^c)\le 1/n$. Let $B=b\int_0^1\psi(u)du$, then by Tchebychev Inequality, we can write

$$\begin{aligned}\mathbb{P}(\Omega_b^c) &= \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n\mathbf{I}_{\{X_i\le 1\}}\delta_i-\int_0^1\psi(u)du\right|>B\right)\\ &\le \frac{1}{B^2}\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n\mathbf{I}_{\{X_i\le 1\}}\delta_i\right)=\frac{\left(\int_0^1\psi(u)du\right)^{-1}}{n\,b^2}\end{aligned}$$

and we prove the result. □

6.4. **Proof of Inequality (3.12).** It follows from (3.11) and the definition of Ω_{c_F} that

$$\|\tilde h_\psi-h\| \le \frac{2}{c_F}\|\hat\psi_{\hat m}-\psi\|+\frac{2}{c_F^2}\|\psi(\hat F_n-F)\|+(n+1)\left((\|\hat\psi_{\hat m}\|+\|\psi\|)+\frac{\|\psi\|}{c_F}\right)\mathbf{I}_{\Omega_{c_F}^c}.$$

By noting that

$$\|\hat\psi_{\hat m}\|^2 = \sum_{\lambda\in\Lambda_{\hat m}}\hat a_\lambda^2\le\frac{1}{n^2}\sum_{1\le i,j\le n}\sqrt{\sum_{\lambda\in\Lambda_{\hat m}}\varphi_\lambda^2(X_i)\sum_{\lambda\in\Lambda_{\hat m}}\varphi_\lambda^2(X_j)}\le\Phi_0^2D_{\hat m}\le\Phi_0^2 n,$$

and by taking the expectation

$$\begin{aligned}\mathbb{E}\|\tilde h_\psi-h\|^2 &\le 4\left(\frac{2}{c_F}\right)^2\mathbb{E}\|\hat\psi_{\hat m}-\psi\|^2+4\left(\frac{2}{c_F^2}\right)^2\|\psi\|^2\mathbb{E}(\|\hat F_n-F\|_\infty^2)\\ &\quad+\frac{4(n+1)^2}{c_F^2}(\Phi_0^2 n+4\|\psi\|^2)\mathbb{P}\left(\|F-\hat F_n\|_\infty>c_F/2\right)\end{aligned}$$

By applying Lemma 6.1, it follows that $\mathbb{E}(\|\hat F_n-F\|_\infty^2)\le C_1/n$ and that

$$\mathbb{P}\left(\|F-\hat F_n\|_\infty>c_F/2\right)\le\left(\frac{2}{c_F}\right)^8\mathbb{E}(\|\hat F_n-F\|_\infty^8)\le\left(\frac{2}{c_F}\right)^8\frac{C_4}{n^4}.\square$$

6.5. **Proof of Theorem 4.1 and of Corollary 4.1.** We apply Talagrand's inequality given by theorem 6.1 to $Z_i = (X_i, \delta_i)$, $i = 1, \dots, n$ and now $g(x,y) = k_t(x,y) = \mathbf{I}_{\{x\leq 1, y=1\}} t(x)/(1-F(x))$, so hereafter, $\nu'_n(t) = \nu_n(k_t)$. Then $\sup_{t\in B_{m,m'}(0,1)} \|k_t\|_\infty \leq (\Phi_0/c_F)\sqrt{D(m')} := M_1$, with $B_{m,m'}(0,1)$ defined as in Lemma 6.2 and

$$\sup_{t\in B_{m,m'}(0,1)} \mathrm{Var}(k_t(X_1,\delta_1)) \leq \sup_{t\in B_{m,m'}(0,1)} \frac{\mathbb{E}(t^2(X_1)\mathbf{I}_{\{X_1\leq 1\}})}{c_F^2} \leq \frac{f_1}{c_F^2} := v,$$

$$\begin{aligned}\mathbb{E}\left(\sup_{t\in B_{m,m'}(0,1)} [\nu'_n(t)]^2\right) &\leq \sum_{\lambda\in\Lambda_{m,m'}} \frac{1}{n}\mathrm{Var}\left(\mathbf{I}_{\{X_1\leq 1\}}\delta_1 \frac{\varphi_\lambda(X_1)}{1-F(X_1)}\right)\\ &\leq \frac{\Phi_0^2 D(m')}{n}\int_0^1 \frac{h(x)}{1-F(x)}dx := H^2.\end{aligned}$$

Then it follows from Theorem 6.1 (in an analogous manner as for Lemma 6.2) that (6.10)

$$\mathbb{E}\left(\sup_{t\in B_{m,m'}(0,1)} [\nu'_n(t)]^2 - p^h(m,m')\right) \leq \kappa'_1\left(\frac{1}{n}e^{-\kappa'_2\epsilon D(m')} + \frac{1}{nC^2(\epsilon)}e^{-\kappa'_3\epsilon^{3/2}\sqrt{n}}\right),$$

where κ'_i for $i = 1,2,3$ are constants depending on K_1 and f_1, c_F, Φ_0 and $\int_0^1 h(x)/(1-F(x))dx$, and

$$p^h(m,m') = 2(1+2\epsilon)\Phi_0^2\int_0^1 \frac{h(x)}{1-F(x)}dx\frac{D_m + D_{m'}}{n}. \tag{6.11}$$

As for the study of $\hat\psi_{\hat m}$, we can write that $\hat h_{\hat m}$ satisfies: $\forall m\in\mathcal{M}_n$,

$$\gamma_n^h(\hat h_{\hat m}) + \mathrm{pen}^h(\hat m) \leq \gamma_n^h(h_m) + \mathrm{pen}^h(m). \tag{6.12}$$

Then by using decomposition (4.4), it follows from (6.12) and from the definition of the process $\nu'_n(t) = \nu_n(k_t)$ that:

$$\begin{aligned}\|\hat h_{\hat m} - h\|^2 &\leq \|h_m - h\|^2 + 2\nu'_n(\hat h_{\hat m} - h_m) + \mathrm{pen}^h(m) - \mathrm{pen}^h(\hat m) + 2R_n(\hat h_{\hat m} - h_m)\\ &\leq \|h_m - h\|^2 + \frac{1}{8}\|\hat h_{\hat m} - h_m\|^2 + 8\sup_{t\in B_{m,\hat m}(0,1)}[\nu'_n(t)]^2\\ &\quad +\mathrm{pen}^h(m) - \mathrm{pen}^h(\hat m) + 2|R_n(\hat h_{\hat m} - h_m)|\mathbf{I}_{\Omega_{c_F}} + 2|R_n(\hat h_{\hat m} - h_m)|\mathbf{I}_{\Omega^c_{c_F}}\end{aligned} \tag{6.13}$$

where we recall that $B_{m,\hat m}(0,1) = \{t\in S_m + S_{\hat m}\ /\ \|t\|\leq 1\}$ and Ω_{c_F} is defined by (4.9). On $\Omega^c_{c_F}$, it follows from (4.11) that

$$2\mathbb{E}(|R_n(\hat h_{\hat m} - h_m)|\mathbf{I}_{\Omega^c_{c_F}}) \leq \frac{1}{4}\,\mathbb{E}(\|\hat h_{\hat m} - h\|^2) + \frac{1}{4}\,\|h_m - h\|^2 + \frac{2^{11}C_4\Phi_0^2}{nc_F^8}. \tag{6.14}$$

On the other hand, for $\nu_n"(t) = (1/n)\sum_{i=1}^n [t(X_i)\mathbf{1}_{\{X_i\le 1\}} - \mathbb{E}(t(X_i))]$ and $B_n(0,1) = \{t \in \mathcal{S}_n, \|t\| = 1\}$, we find

$$\begin{aligned}
2\,|R_n(\hat{h}_{\hat{m}} - h_m)|\mathbf{1}_{\Omega_{c_F}} &\le \frac{2}{c_F^2}\|\hat{h}_{\hat{m}} - h_m\|\|\hat{F}_n - F\|_\infty \sup_{t\in B_{m,\hat{m}}(0,1)} \frac{1}{n}\sum_{i=1}^n |t(X_i)|\mathbf{1}_{\{X_i\le 1\}} \\
&\le \frac{1}{8}\|\hat{h}_{\hat{m}} - h_m\|^2 + \frac{8}{c_F^4}\|\hat{F}_n - F\|_\infty^2 \sup_{t\in B_n(0,1)} \frac{1}{n}\sum_{i=1}^n t^2(X_i)\mathbf{1}_{\{X_i\le 1\}} \\
\le \frac{1}{4}\|\hat{h}_{\hat{m}} - h\|^2 &+ \frac{1}{4}\|h_m - h\|^2 + \frac{8f_1}{c_F^4}\|\hat{F}_n - F\|_\infty^2 \frac{8}{c_F^4}\|\hat{F}_n - F\|_\infty^2 \sup_{t\in B_n(0,1)} |\nu_n"(t^2)|.
\end{aligned}$$

Then it follows from Baraud (2002) (see also Baraud et al. (2001)) that

$$\mathbb{P}\left(\sup_{t\in B_n(0,1)} |\nu_n"(t^2)| \ge \rho\right) \le |\Lambda_n|^2 \exp\left(-\frac{n\rho^2}{4f_1 L(\varphi)}\right)$$

where $L(\varphi)$ is a quantity associated with the orthonormal basis of the largest space of the collection. Moreover, let $\mathcal{S}_n$ denote this space and $(\varphi_\lambda)_{\lambda\in\Lambda_n}$ denote its orthonormal basis, then $|\Lambda_n| = \dim(\mathcal{S}_n) := N_n$. We know from Baraud (2002) that $L(\varphi) \le K_\varphi N_n$ for the basis [DP] and [W] and $L(\varphi) \le N_n^2$ for the basis $[T]$. Consequently

$$\begin{aligned}
&\mathbb{E}\left(\sup_{t\in B_n(0,1)} [\nu"_n(t^2)]^2\right) \\
\le\ & 2\int_0^{\sqrt{\ln(n)}} x\mathbb{P}\left(\sup_{t\in B_n(0,1)} |\nu_n"(t^2)| \ge x\right) dx + 2\int_{\sqrt{\ln(n)}}^{+\infty} x\mathbb{P}\left(\sup_{t\in B_n(0,1)} |\nu_n"(t^2)| \ge x\right) dx \\
\le\ & \ln(n) + \frac{8f_1 N_n^2 L(\varphi)}{n}\int_{\frac{\sqrt{n\ln(n)}}{2\sqrt{f_1 L(\varphi)}}}^{+\infty} ue^{-u^2} du \le \ln(n) + 4f_1 N_n L(\varphi)\exp\left(-\frac{n\ln(n)}{4f_1 L(\varphi)}\right).
\end{aligned}$$

It follows that if $L(\varphi) \le n/(16f_1)$, which holds if $N_n \le n/(16f_1 K_\varphi)$ for [DP] and [W] and if $N_n \le \sqrt{n}/(4\sqrt{f_1})$ for [T], then

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

This leads to

$$\begin{aligned} & 2\mathbb{E}(|R_n(\hat{h}_{\hat{m}} - h_m)|\mathbf{I}_{\Omega_{c_F}}) \\ \leq\ & \frac{1}{4}\mathbb{E}(\|\hat{h}_{\hat{m}} - h\|^2) + \frac{1}{4}\|h_m - h\|^2 + \frac{8f_1C_2}{nc_F^4} + \frac{8}{c_F^4}\mathbb{E}^{1/2}(\|\hat{F}_n - F\|_\infty^4)\mathbb{E}^{1/2}(\sup_{t\in B_n(0,1)}|\nu_n"(t^2)|) \\ \leq\ & \frac{1}{4}\mathbb{E}(\|\hat{h}_{\hat{m}} - h\|^2) + \frac{1}{4}\|h_m - h\|^2 + \frac{8f_1C_2}{nc_F^4} + \frac{8C_4^{1/2}}{nc_F^4}\sqrt{2\ln(n)}\,. \end{aligned}$$

By gathering (6.13), (6.14) and the inequality above, we obtain

$$\begin{aligned} \frac{1}{4}\mathbb{E}(\|\hat{h}_{\hat{m}} - h\|^2) \ \leq\ & \frac{7}{4}\|h_m - h\|^2 + \mathrm{pen}^h(m) + 8\mathbb{E}(p^h(m,\hat{m}) - \mathrm{pen}^h(\hat{m})) \\ & +8\sum_{m'\in\mathcal{M}_n}\mathbb{E}\left([\nu_n'(t)]^2 - p^h(m,m')\right)_+ + \frac{K\sqrt{\ln(n)}}{n}. \end{aligned} \tag{6.15}$$

We take $pen^h(m)$ such that

$$8p^h(m,m') \leq \mathrm{pen}^h(m) + \mathrm{pen}^h(m') \quad \text{for all } m,\ m' \text{ in } \mathcal{M}_n. \tag{6.16}$$

A straightforward consequence of (6.10) is that the sum over $m \in \mathcal{M}_n$ is $O(1/n)$, so that Inequality (4.13) holds if $N_n \leq n/(16K_\varphi f_1)$ for [DP] and [W] and if $N_n^2 \leq n/(16f_1)$ for [T]. This ensures the result of Theorem 4.1.

The proof of Corollary 4.1 is similar to the proof of Theorem 3.2 with now

$$\widetilde{\Omega}_b = \left\{\left|\left(\frac{1}{n}\sum_{i=1}^n \frac{\mathbf{I}_{\{X_i\leq 1\}}\delta_i}{(1-\hat{F}_n(X_i))^2}\right) / \left(\int_0^1 \frac{h(x)}{1-F(x)}dx\right) - 1\right| < b\right\},$$

for $0 < b < 1$. The result holds because we can prove that $\mathbb{P}(\widetilde{\Omega}_b^c) \leq 1/n$, by iterated but simple use of Lemma 6.1. □

6.6. **Proof of Theorem 4.2.** The problem now is that the norm connection $\|t\|_\infty \leq \Phi_0\sqrt{D_m}\|t\|$ no longer holds for $t \in S_m$ but is replaced by $\|t\|_\infty \leq \Phi_0\sqrt{N_n}\|t\|$.

Preliminary 1. The first point is to take into account that, as the collection contains a great number of models, some weights are required in order to find convergent series. Indeed, note that, for $L_n = \ln(n)/(r+1)$,

$$\begin{aligned} \sum_{m\in\mathcal{M}_n} e^{-L_nD_m} &= \sum_{\ell=1}^{K_n-1}\binom{K_n-1}{\ell}e^{-L_n(\ell+1)(r+1)} = e^{-L_n(r+1)}\left[(1+e^{-L_n(r+1)})^{K_n-1} - 1\right] \\ &\leq \left(1+e^{-L_n(r+1)}\right)^{K_n} \leq \left(1+\frac{1}{n}\right)^n \leq e, \end{aligned}$$

by using that $K_n \leq n$. Taking $L_n = L$ simply leads in that case to an exponentially divergent sum.

Preliminary 2. We can prove that there exists a real function θ on $\mathcal{S}_n$ such that for all $t \in \mathcal{S}_n$ and $m \in \mathcal{M}_n$, $\|t_m\|_\infty \leq \theta(t)$, satisfying

$$|\theta(\bar{h}_n) - \theta(\hat{h}_n)| \leq \Phi\sqrt{N_n} \sup_{\lambda\in\Lambda_n} |\nu_n'(\varphi_\lambda) + R_n(\varphi_\lambda)|, \tag{6.17}$$

where $\bar{h}_n$ is the projection of h on $\mathcal{S}_n$ and $\hat{h}_n$ the projection estimator of h on $\mathcal{S}_n$. Indeed, we can simply choose $\theta(t) = (r+1)\|t\|_\infty$ and use Inequality (2.8) of Birgé and Massart (1997). Then $\sup_n \theta(\bar{h}_n) = \theta(h) = (r+1)\|h\mathbf{1}_{[0,1]}\|_\infty$ and

$$\begin{aligned}
&|\theta(\bar{h}_n) - \theta(\hat{h}_n)| = (r+1)|\|\bar{h}_n\|_\infty - \|\hat{h}_n\|_\infty| \leq (r+1)\|\bar{h}_n - \hat{h}_n\|_\infty \\
\leq\ & (r+1)\left\|\sum_{\lambda\in\Lambda_n}\left[< h,\varphi_\lambda > - \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{X_i\leq 1\}}\frac{\delta_i\varphi_\lambda(X_i)}{1-\hat{F}_n(X_i)}\right]\varphi_\lambda\right\|_\infty \\
\leq\ & (r+1)\left\|\sum_{\lambda\in\Lambda_n}(\nu_n'(\varphi_\lambda) + R_n(\varphi_\lambda))\varphi_\lambda\right\|_\infty \\
\leq\ & (r+1)\sup_{\lambda\in\Lambda_n}|\nu_n'(\varphi_\lambda)+R_n(\varphi_\lambda)|\left\|\sum_{\lambda\in\Lambda_n}\varphi_\lambda\right\|_\infty \leq (r+1)^2\sqrt{N_n}\sup_{\lambda\in\Lambda_n}|\nu_n'(\varphi_\lambda)+R_n(\varphi_\lambda)|
\end{aligned}$$

using (2.8) of Birgé and Massart (1997) again. Then we have (6.17) with $\Phi = (r+1)^2$.

Then we split the probability space $\Omega = \Omega_1 \cup \Omega_1^c$, with

$$\Omega_1 = \{|\theta(\hat{h}_n) - \theta(\bar{h}_n)| \leq \|h\mathbf{1}_{[0,1]}\|_\infty\}.$$

Then we bound $\mathbb{E}(\|h - \hat{h}_{\hat{m}}\|^2\mathbf{1}_{\Omega_1})$. Now, we proceed in two steps :

Step 1. On Ω_1,

$$\begin{aligned}
\sup_{m,m'}\|h_m - \hat{h}_{m'}\|_\infty &\leq \theta(\bar{h}_n) + \theta(\hat{h}_n) \leq |\theta(\hat{h}_n) - \theta(\bar{h}_n)| + 2\theta(\bar{h}_n) \\
&\leq \|h\mathbf{1}_{[0,1]}\|_\infty + 2\theta(h) := C(h).
\end{aligned}$$

$$\text{Let } W_n(m') = \left[\left(\sum_{t\in S_{m'}^*, 0\neq\|t-h_m\|_\infty\leq C(h)}\left|\nu_n'\left(\frac{t-h_m}{\|t-h_m\|\vee x(m')}\right)\right|\right)^2 - p(m,m')\right]_+$$

then we have

$$2|\nu_n'(\tilde{h}-h_m)| \leq \frac{1}{4}\|h_m - h\|^2 + \frac{1}{4}\|h-\hat{h}_{\hat{m}}\|^2 + \frac{1}{8}x(\hat{m})^2 + 8\sum_{m'\in\mathcal{M}_n} W_n(m') + 8p(m,\hat{m}),$$

with $x(m')^2 = 8\ln(n)(D_m + D'_m)/n$. Then we apply inequality (6.1) with $v = \|h\mathbf{I}_{[0,1]}\|_\infty/c_F$, $M_1 = C(h)/(c_F x(m'))$, $H^2 = [\sup_{x\in[0,1]} h(x)/(1-F(x))](D_m+D_{m'})/n$, and bounding $\sum_\lambda \mathbb{E}(\varphi_\lambda^2(X_1)/(1-F(X_1))^2)$ by $\sup_{x\in[0,1]}[h(x)/(1-F(x))]D(m')$ since $\int \varphi_\lambda^2(x)dx = 1$. By applying Theorem 6.1 again, this gives the bound

$$\mathbb{E}(W_n(m')\mathbf{I}_{\Omega_1}) \leq C_1\left(\frac{1}{n}e^{-C_2\epsilon D_{m'}} + \frac{1}{nC^2(\epsilon)}e^{-C_3C(\epsilon)\sqrt{\epsilon}\sqrt{\ln(n)}D_{m'}}\right),$$

where C_i, $i = 1,2,3$ are constants depending on $\|h\mathbf{I}_{[0,1]}\|_\infty$, $C(h)$, $\sup_{x\in[0,1]} h(x)/(1-F(x))$ and c_F for the choice: $p(m,m') = 2(1+2\epsilon)\sup_{x\in[0,1]} h(x)/(1-F(x))(D_m + D_{m'})/n$. Therefore, choosing $\epsilon = KL_n$ with $K > \max(1/C_2, 1/[(r+1)C_3^2])$ and $C(\epsilon) = 1$ we find that all terms are of order less than $(1/n)e^{-\ln(n)D_{m'}/(r+1)}$. We find a global order less than $1/n$ since we have checked in Preliminary 1 that $\sum_{m'\in\mathcal{M}_n} e^{-\ln(n)D_{m'}/(r+1)}$ is bounded, for a penalty given by (4.15). The other terms of (6.13) are bounded as previously.

Step 2. On the complement of Ω_1, we use (6.17). Since $\tilde{h}$ can be seen as the orthogonal projection of $\hat{h}_n$ on $S_{\hat{m}}$, we have

$$\begin{aligned}\|\tilde{h}\| &\leq \theta(\hat{h}_n) \leq \theta(\bar{h}_n) + |\theta(\hat{h}_n) - \theta(\bar{h}_n)| \leq \theta(h) + \Phi\sqrt{N_n}\sup_{\lambda\in\Lambda_n}|\nu'_n(\varphi_\lambda) + R_n(\varphi_\lambda)| \\ &\leq \theta(h) + C(\Phi, \Phi_0, c_F, \int_0^1 h(x)dx)nN_n\end{aligned}$$

using that $\|\varphi_\lambda\|_\infty \leq \Phi_0\sqrt{N_n}$ and some rough bounds as: $\|F - \hat{F}_n\|_\infty \leq 2$, $1/(1-\hat{F}_n(x)) \leq (n+1)$. Therefore

$$\begin{aligned}\mathbb{E}\left[\|h-\tilde{h}\|^2\mathbf{I}_{\Omega_1^c}\right] &\leq 2\|h\|^2\mathbb{P}(\Omega_1^c) + 2\mathbb{E}\left[\|\tilde{h}\|^2\mathbf{I}_{\Omega_1^c}\right] \\ &\leq \left[2\|h\|^2 + 4(\theta(h)^2 + C^2(\Phi, \Phi_0, c_F, \int_0^1 h(x)dx)n^2N_n^2)\right]\mathbb{P}(\Omega_1^c).\end{aligned}$$

Therefore, we need to prove that $\mathbb{P}(\Omega_1^c) \leq C/n^5$ for $N_n \leq n$. With obvious notations, we write

$$\begin{aligned}\mathbb{P}(\Omega_1^c) &= \mathbb{P}(|\theta(\hat{h}_n) - \theta(\bar{h}_n)| \geq \|h\mathbf{I}_{[0,1]}\|_\infty) \leq \mathbb{P}\left(\sup_{\lambda\in\Lambda_n}|\nu'_n(\varphi_\lambda) + R_n(\varphi_\lambda)| \geq \frac{\|h\mathbf{I}_{[0,1]}\|_\infty}{\Phi\sqrt{N_n}}\right) \\ &\leq \sum_{\lambda\in\Lambda_n}\mathbb{P}\left(|\nu'_n(\varphi_\lambda)| \geq \frac{\|h\mathbf{I}_{[0,1]}\|_\infty}{2\Phi\sqrt{N_n}}\right) + \mathbb{P}\left(|R_n(\varphi_\lambda)| \geq \frac{\|h\mathbf{I}_{[0,1]}\|_\infty}{2\Phi\sqrt{N_n}}\right).\end{aligned}$$

For the first term, we use Bernstein inequality: let $S_n = \sum_{i=1}^n Z_i$ and Z_i i.i.d., with $|Z_i| \leq B$, for all $i = 1, \dots, n$ and $\mathrm{Var}(Z_1) = \sigma^2$, $\mathbb{P}(S_n - \mathbb{E}(S_n) \geq n\eta) \leq \exp\left(-(n\eta^2/2)/(\sigma^2 + B\eta)\right)$. We apply it with $B = 2\Phi_0\sqrt{N_n}/c_F$ and $\sigma^2 = \sup_{x\in[0,1]}(h(x)/(1-F(x)))$ and we find

$$\mathbb{P}\left(|\nu_n'(\varphi_\lambda)| \geq \|h\mathbf{1}_{[0,1]}\|_\infty/(2\Phi\sqrt{N_n})\right) \leq 2\exp\left(-Kn/N_n\right) \leq 2n^{-6}$$

if $N_n \leq (K/6)(n/\ln(n))$. Here K is a constant depending on Φ, Φ_0, $\|h\mathbf{1}_{[0,1]}\|_\infty$, c_F and $\sup_{x\in[0,1]}(h(x)/(1-F(x)))$. The same kind of inequality is obtained for R_n by using again the decomposition involving Ω_{c_F}:

$$\mathbb{P}\left(|R_n(\varphi_\lambda)| \geq \|h\mathbf{1}_{[0,1]}\|_\infty/(2\Phi\sqrt{N_n})\right) \leq \mathbb{P}_1 + \mathbb{P}_2$$

where $\mathbb{P}_1$ is the probability of the intersection of event of interest with Ω_{c_F} and $\mathbb{P}_2$ with $\Omega_{c_F}^c$. For $\mathbb{P}_2$, it is easy to write $\mathbb{P}_2 \leq \mathbb{P}(\Omega_{c_F}^c) \leq \mathbb{P}(\|F - \hat{F}_n\|_\infty > c_F/2) \leq \left(2c_F^{-1}\right)^{12} C_6 n^{-6}$. On Ω_{c_F}, we write

$$\begin{aligned} R_n(\varphi_\lambda) &\leq \frac{2}{c_F^2}\|\hat{F}_n - F\|_\infty \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{X_i\leq 1\}}\delta_i|\varphi_\lambda(X_i)| \\ &\leq \frac{2}{c_F^2}\|\hat{F}_n - F\|_\infty \left[\tilde{\nu}_n(\varphi_\lambda) + \mathbb{E}(\mathbf{1}_{\{X_1\leq 1\}}\delta_1|\varphi_\lambda(X_1)|)\right] \\ &\leq \frac{2}{c_F^2}\left[\|\hat{F}_n - F\|_\infty \mathbb{E}(\mathbf{1}_{\{X_1\leq 1\}}\delta_1|\varphi_\lambda(X_1)|) + \frac{1}{2}\tilde{\nu}_n^2(\varphi_\lambda) + \frac{1}{2}\|\hat{F}_n - F\|_\infty^2\right], \end{aligned}$$

where $\tilde{\nu}_n(t) = (1/n)\sum_{i=1}^n \left(\mathbf{1}_{\{X_i\leq 1\}}\delta_i|\varphi_\lambda(X_i)| - \mathbb{E}(\mathbf{1}_{\{X_1\leq 1\}}\delta_1|\varphi_\lambda(X_1)|)\right)$. Then the result follows if $N_n \leq Kn/\ln(n)$ for some well chosen constant K, by using the exponential inequality recalled in Lemma 6.1 for the terms involving $\|F - \hat{F}_n\|_\infty$ and with Bernstein Inequality as previously for the terms involving $\tilde{\nu}_n$. □

Acknowledgements

The authors are very grateful to Yves Rozenholc for his precious help and his experience for the simulations and the practical implementation of the estimators.

References

[1] 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.

[2] Baraud, Y. (2002) Model selection for regression on a random design. *ESAIM P&S* **6**, 127-146.

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

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

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

[6] Birgé, L. and Massart, P. (1997). From model selection to adaptive estimation, in *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics* (D. Pollard, E. Torgersen and G. Yang, eds), 55-87, Springer-Verlag, New-York.

[7] Birgé, L. and Y. Rozenholc (2002). How many bins must be put in a regular histogram. Preprint of the LPMA no.721, http://www.proba.jussieu.fr/mathdoc/preprints/index.html.

[8] Blum, J.R. and Susarla, V. (1980). Maximal deviation theory of density and failure rate function estimates based on censored data. *Multivariate analysis, V*, (*Proc. Fifth Internat. Sympos., Univ. Pittsburgh, Pittsburgh, Pa., 1978*), 213-222. North-Holland, Amsterdam-New York.

[9] Brunel, E. and Comte, F. (2004). Adaptive estimation of hazard rate with censored data. Preprint MAP5 2004-16, University Paris 5, http://www.math-info.univ-paris5.fr/map5/.

[10] Daubechies, I. (1992). *Ten lectures on wavelets.* Philadelphia : Society for Industrial and Applied Mathematics.

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

[12] Diehl, S. and Stute, W. (1988). Kernel density and hazard function estimation in the presence of censoring. *J. Multivariate Anal.* **25**, 299-310.

[13] Döhler, S. and Rüschendorf, L. (2002). Adaptive estimation of hazard functions. *Probab. Math. Statist.* **22**, no. 2, 355–379.

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

[15] Huber, C. and MacGibbon, B. (2004). Lower bounds for estimating a hazard. *Advances in survival analysis,* 209–226, Handbook of Statist., 23 Elsevier, Amsterdam.-

[16] Kooperberg, C., Stone, C.J. and Truong, Y.K. (1995). The L_2 rate of convergence for hazard regression. *Scand. J. Statist.* **22**, 143-157.

[17] 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**, no. 3, 461-473.

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

[19] Mielniczuk, J. (1986). Some asymptotic properties of kernel estimators of a density function in case of censored data. *Ann. Statist.*, **14**, 766-773.

[20] Patil, P. N. (1997). Nonparametric hazard rate estimation by orthogonal wavelet method. *J. Statist. Plann. Inference* **60**, 153-168.

[21] Patil, P. N., Wells, M. T. and Marron, J. S. (1994). Kernel based estimation of ratio functions. *J. Nonparametric Statist.* **4**, 203-209.

[22] Ramlau-Hansen, H. (1983). Smoothing counting process intensities by means of kernel functions. *Ann. Statist.* **11**, 453-466.

[23] Reynaud-Bouret, P. (2002). Penalized projection estimators of the Aalen multiplicative intensity. Working paper.

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

[25] 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.

[26] Uzunoḡullari, U. and Wang, J.L. (1992). A comparison of hazard rate estimators for left truncated and right censored data. *Biometrika* **79**, 297-310.

[27] Wu, S.S. and Wells, M.T. (2003). Nonparametric estimation of hazard functions by wavelet methods. *Nonparametric Statistics* **15**, 187-203.

[28] Yandell, B. S. (1983). Nonparametric inference for rates with censored survival data. *Ann. Statist.* **11**, 1119-1135.