ADAPTIVE ESTIMATION OF THE CONDITIONAL DENSITY IN PRESENCE OF CENSORING

E. BRUNEL[(*),(1)], F. COMTE[(*),(2)] & C. LACOUR[(*),(3)]

Abstract. Consider an i.i.d. sample (X_i, Y_i), $i = 1, \dots, n$ of observations and denote by $\pi(x, y)$ the conditional density of Y_i given $X_i = x$. We provide an adaptive nonparametric strategy to estimate π. We prove that our estimator realizes a global squared-bias/variance compromise in a context of anisotropic function classes. We prove that our procedure can be adapted to positive censored random variables Y_i's, i.e. when only $Z_i = \inf(Y_i, C_i)$ and $\delta_i = \mathbb{1}_{\{Y_i \le C_i\}}$ are observed, for an i.i.d. censoring sequence $(C_i)_{1\le i\le n}$ independent of $(X_i, Y_i)_{1\le i\le n}$. Simulation experiments illustrate the method.

June 2007
AMS (2000) subject classification. 62N02, 62G07.
Keywords. Adaptive estimation. Censored data. Conditional density. Nonparametric methods.

1. Introduction

Consider an i.i.d. sample $(X_i, Y_i)_{1\le i\le n}$ of couples of random variables with common probability density function (pdf) $f_{(X,Y)}$. The marginal density of the X_i's is also denoted by f_X. Now, define the conditional density function of Y_i given $X_i = x$, for all real y and x such that $f_X(x) > 0$, by

$$\pi(x, y) = \frac{f_{(X,Y)}(x, y)}{f_X(x)}.$$

The aim of the paper is to provide a statistical strategy to recover π from the observations. A Nadaraya-Watson strategy building an estimator as the ratio of an estimator of $f_{(X,Y)}$ divided by an estimator of f_X is conceivable. But estimators resulting of such methods have the drawback of precisely involving a ratio, with a denominator which can be small. This is the reason why we provide rather a regression-type strategy based on a mean square contrast. Using tools developed for standard regression by Baraud et al. (2001), or for transition density estimation by Lacour (2007) in the Markov chain setting, we propose a simple adaptive strategy: we both build a collection of projection estimators on finite dimensional spaces and select, by penalization of the mean square contrast, the adequate space. Then we prove that the usual squared-bias/variance compromise is achieved, in a data driven way. It is worth mentioning that the projection spaces need not be the same in both x and y-directions, and thus, it allows to estimate functions π belonging to possibly anisotropic Besov spaces.

[(*)] MAP5, University Paris Descartes, France.
[(1)] email: elodie.brunel@math-info.univ-paris5.fr,
[(2)] email: fabienne.comte@univ-paris5.fr,
[(3)] email: claire.lacour@math-info.univ-paris5.fr.

We consider moreover the case of positive censored variables Y_i's, which is often encountered when survival times are under study. Let $(C_i)_{1\leq i\leq n}$ be an i.i.d. sequence of positive censoring variables, with cumulative distribution function (cdf) G, independent of $(X_i, Y_i)_{1\leq i\leq n}$. In this context, the observations are

$$(X_i, Z_i, \delta_i)_{1\leq i\leq n}, \quad Z_i = \inf(Y_i, C_i), \; \delta_i = \mathbb{1}_{\{Y_i\leq C_i\}}.$$

Then, we can provide both modified contrast and penalty following the transformation device proposed by Fan and Gijbels (1994) and also using tools for censoring correction taking advantage of Bitouzé et al. (1999)'s results as developed by Brunel and Comte (2005).

Kernel strategies involving local polynomials have been studied in the uncensored framework by De Gooijer and Zerom (2003) and Fan et al. (1996), see also Hyndman and Yao (2002), with possible dependence between the variables of a given sequence. In the context of a censored heteroscedastic regression model, Van Keilegom and Veraverbeke (2002) propose an estimation procedure to avoid the drawbacks of purely nonparametric kernel estimators. Indeed, its local building produces a bad behavior whenever the censoring is heavy in neighborhood of x. All these works have the pointwise feature of kernel estimators when we are in position to control a global integrated risk only. However, they require regularity conditions on the functions to be estimated much more restrictive than our flexible estimation tool. Moreover, the theoretical study of bandwidth selection is often omitted and the practical bandwidth chosen by empirical considerations, whereas we have at our disposal mathematical methods to prove the good theoretical properties of our adaptive estimator.

The plan of the paper is the following. First we describe in Section 2 the assumptions and the model collection. Then the estimators in both uncensored and censored contexts are defined and discussed in Section 3. The results are stated in Section 4, illustrated via simulations and examples in Section 5 and proved in Section 6.

2. Model and Assumptions

2.1. **Model.** To sum up, we consider two different frameworks. In both cases, the (X_i, Y_i), $i = 1, \dots n$ are independent and identically distributed couples of variables. We estimate π on a given compact set $A = A_1 \times A_2$ only. Moreover, we distinguish:

- the uncensored framework, where the variables (X_i, Y_i), $i = 1\dots, n$ are directly observed;
- the censored framework, where the censoring variables C_i are independent and identically distributed, with c.d.f. G and the sequences (X_i, Y_i) and (C_i), $i = 1\dots, n$ are independent. In this context, we observe either (X_i, Z_i, δ_i) where $Z_i = Y_i \wedge C_i$ and $\delta_i = \mathbb{1}_{\{Y_i\leq C_i\}}$. Only positive variables Y_i and C_i, $i = 1\dots, n$ are considered.

The aim of the paper is to estimate the conditional density $\pi(x, y)$ of Y_i given $X_i = x$ and to evaluate the price to pay from uncensored to censored case. Roughly speaking, what kind of additional constraint would we set to extend the results to censored data.

2.2. **Assumptions on the variables.** In all cases, we set the following usual assumptions

[$\mathcal{A}1$] The conditional density π belongs to the space of bounded and square integrable functions on $A = A_1 \times A_2$ denoted by $L^\infty \cap L^2(A)$

[$\mathcal{A}2$] The density f_X verifies $\|f_X\|_\infty := \sup_{x\in A_1} |f_X(x)| < \infty$ and there exists a positive real f_0 such that, for all x in A_1, $f_X(x) \geq f_0$.

Moreover, in the censored case, we suppose:

[$\mathcal{A}3$] For all $y \in A_2$, $1 - G(y) \geq c_G > 0$.

Note that the lower bound condition [$\mathcal{A}3$] is required only on the compact set A_2, which is a very mild assumption.

2.3. **Assumptions on the models.** In order to estimate π, we need to introduce a collection $\{S_m, m \in \mathcal{M}_n\}$ of projection spaces, that we call models. For each $m = (m_1, m_2)$, S_m is a space of functions with support in A defined by using two spaces: F_{m_1} and H_{m_2} which are subspaces of $(L^2 \cap L^\infty)(\mathbb{R})$ respectively spanned by two orthonormal bases $(\varphi_j^m)_{j\in J_m}$ with $|J_m| = D_{m_1}$ and $(\psi_k^m)_{k\in K_m}$ with $|K_m| = D_{m_2}$. For all j and all k, the supports of φ_j^m and ψ_k^m are respectively included in A_1 and A_2. Here j and k are not necessarily integers, it can be couples of integers as in the case of a piecewise polynomial space see Section 2.4. Then, we define

$$S_m = F_{m_1} \otimes H_{m_2} = \{t, \quad t(x, y) = \sum_{j\in J_m} \sum_{k\in K_m} a_{j,k}^m \varphi_j^m(x)\psi_k^m(y)\}$$

The assumptions on the models are the following:

[M1] For all m_2, $D_{m_2} \leq n^{1/2}$ and $\mathcal{D}_n := \max_{m\in\mathcal{M}_n} D_{m_1} \leq n^{1/2}/\log(n)$

[M2] There exist positive reals ϕ_1, ϕ_2 such that, for all u in F_{m_1}, $\|u\|_\infty^2 \leq \phi_1 D_{m_1} \int u^2$, and for all v in H_{m_2}, $\sup_{x\in A_2} |v(x)|^2 \leq \phi_2 D_{m_2} \int v^2$. By letting $\phi_0 = \sqrt{\phi_1\phi_2}$, that leads to

$$\forall t \in S_m \qquad \|t\|_\infty \leq \phi_0 \sqrt{D_{m_1} D_{m_2}} \|t\| \tag{1}$$

where $\|t\|^2 = \iint t^2(x, y)dxdy$ and $\|t\|_\infty = \sup_{(x,y)\in A_1\times A_2} |t(x, y)|$.

[M3]$D_{m_1} \leq D_{m_1'} \Rightarrow F_{m_1} \subset F_{m_1'}$ and $D_{m_2} \leq D_{m_2'} \Rightarrow H_{m_2} \subset H_{m_2'}$

The first assumption guarantees that $\dim S_m = D_{m_1} D_{m_2} \leq n$ where n is the number of observations. The condition [M2] implies a useful link between the L^2 norm and the infinite norm. The third assumption ensures that, for m and m' in $\mathcal{M}_n$, $S_m + S_{m'}$ is included in a model (since $S_m + S_{m'} \subset S_{m''}$ with $D_{m_1''} = \max(D_{m_1}, D_{m_1'})$ and $D_{m_2''} = \max(D_{m_2}, D_{m_2'})$). We denote by $\mathcal{S}$ the space with maximal dimension among the $(S_m)_{m\in\mathcal{M}_n}$. Thus for all m in $\mathcal{M}_n$, $S_m \subset \mathcal{S}$.

[M4] Nested model collection: $m_1 = m_2$, $H_{m_1} = F_{m_1}$ (i.e. $S_m = F_{m_1} \otimes F_{m_1}$) and $\mathcal{D}_n := \max_{m\in\mathcal{M}_n} D_{m_1} = \max_{m\in\mathcal{M}_n} D_{m_2} \leq n^{1/4}$.

This assumption is useful, even if a little restrictive, to deal with censored data.

2.4. **Examples of models.** We show here that Assumptions [M1]–[M3] are not too restrictive. Indeed, they are verified for the spaces F_{m_1} (and H_{m_2}) spanned by the following bases (see Barron et al. (1999)):

- Trigonometric basis: for $A_1 = [0, 1]$, $\mathrm{span}(\varphi_0, \ldots, \varphi_{m_1-1})$ with $\varphi_0 = \mathbb{1}_{[0,1]}$, $\varphi_{2j}(x) = \sqrt{2}\cos(2\pi jx)\ \mathbb{1}_{[0,1]}(x)$, $\varphi_{2j-1}(x) = \sqrt{2}\sin(2\pi jx)\mathbb{1}_{[0,1]}(x)$ for $j \geq 1$. For this model $D_{m_1} = m_1$ and $\phi_1 = 2$ hold.

- Histogram basis: for $A_1 = [0,1]$, $\text{span}(\varphi_1, \dots, \varphi_{2^{m_1}})$ with $\varphi_j = 2^{m_1/2} \mathbb{1}_{[(j-1)/2^{m_1}, j/2^{m_1}[}$ for $j = 1, \dots, 2^{m_1}$. Here $D_{m_1} = 2^{m_1}$, $\phi_1 = 1$.
- Regular piecewise polynomial basis: for $A_1 = [0,1]$, polynomials of degree $0, \dots, r$ (where r is fixed) on each interval $[(l-1)/2^D, l/2^D[, l = 1, \dots, 2^D$. In this case, $m_1 = (D, r)$, $J_m = \{j = (l,d),\ 1 \le l \le 2^D, 0 \le d \le r\}$, $D_{m_1} = (r+1)2^D$. We can put $\phi_1 = \sqrt{r+1}$.
- Regular wavelet basis: $\text{span}(\Psi_{lk}, l = -1, \dots, m_1, k \in \Lambda(l))$ where $\Psi_{-1,k}$ points out the translates of the father wavelet and $\Psi_{lk}(x) = 2^{l/2}\Psi(2^l x - k)$ where Ψ is the mother wavelet. We assume that the support of the wavelets is included in A_1 and that Ψ_{-1} belongs to the Sobolev space W_2^r.

3. Estimation procedure

3.1. **Definition of the contrast.** If no censoring occurs, the Y_i's are observed and we can choose the following contrast

$$\gamma_n^0(t) = \frac{1}{n}\sum_{i=1}^{n}[\int_{\mathbb{R}} t^2(X_i, y)dy - 2t(X_i, Y_i)]. \tag{2}$$

In fact, it is easy to explain the contrast since

$$\mathbb{E}\gamma_n^0(t) = \mathbb{E}\left[\int t^2(X_1, y)dy\right] - 2\,\mathbb{E}(t(X_1, Y_1)) = \|t - \pi\|_f^2 - \|\pi\|_f^2$$

where

$$\|t\|_f^2 = \iint t^2(x,y)f(x)dxdy.$$

Therefore $\gamma_n(t)$ is the empirical counterpart of $\|t-\pi\|_f^2 - \|\pi\|_f^2$ and its minimization comes down to minimize $\|t - \pi\|_f$. This contrast is new and its originality actually stands in the links with the regression-type contrasts, as we will see in the next subsection.

Now, let us take into account the fact that the Y_i's may be censored. We use a standard transformation of the data (see Fan and Gijbels (1994)) and introduce an empirical version of the weights

$$w_i = \begin{cases} 1 \text{ without censoring} \\ \dfrac{\delta_i}{\bar{G}(Z_i)} \text{ with censoring,} \end{cases} \tag{3}$$

where $\bar{G} = 1 - G$ is the survival function associated with the censoring variables, by choosing the contrast function

$$\gamma_n(t) = \frac{1}{n}\sum_{i=1}^{n}\left(\int_{\mathbb{R}} t^2(X_i, y)dy - 2\hat{w}_i t(X_i, Z_i)\right),\ \hat{w}_i = \begin{cases} 1 \text{ without censoring} \\ \dfrac{\delta_i}{\hat{\bar{G}}(Z_i)} \text{ with censoring.} \end{cases} \tag{4}$$

Here $\hat{\bar{G}}$ is the Kaplan Meier estimator of the c.d.f of the C_i's, modified in the way suggested by Lo et al. (1989), and defined by

$$\hat{\bar{G}}(y) = \prod_{Z_{(i)} \le y} \left(\frac{n-i+1}{n-i+2}\right)^{1-\delta_{(i)}}. \tag{5}$$

Note that $\widehat{\bar{G}}$ is built in order to satisfy the following useful property: $\widehat{\bar{G}}(y) \geq 1/(n+1)$, $\forall y$. Moreover, it follows from Lemma 1 in Section 6, that it is a very good estimator of $\bar{G}$ on the interval A_2 provided $A_2 \subsetneq [0, \tau]$ where $\tau = \sup\{y, G(y) < 1\}$, which condition is ensured by Assumption [$\mathcal{A}3$]. Therefore, we define

$$\hat{\pi}_m = \arg\min_{t \in S_m} \gamma_n(t), \tag{6}$$

in the sense explained in Section 3.2, and lastly, we set

$$\hat{m} = \arg\min_{m \in \mathcal{M}_n} \{\gamma_n(\hat{\pi}_m) + \mathrm{pen}(m)\} \tag{7}$$

where pen is a penalty function to be specified later. Note that $\gamma_n^0(t)$ and $\gamma_n(t)$ coincide if no censoring happens by definition of the weights $\hat{w}_i$. Then we can define $\tilde{\pi} = \hat{\pi}_{\hat{m}}$ and compute the empirical mean integrated squared error $\mathbb{E}\|\pi - \tilde{\pi}\|_n^2$ where $\|.\|_n$ is the empirical norm defined by

$$\|t\|_n = \left(\frac{1}{n}\sum_{i=1}^{n} \int_{\mathbb{R}} t^2(X_i, y)dy\right)^{1/2}. \tag{8}$$

This norm is the natural distance in this problem and we can notice that if t is deterministic with support included in $A_1 \times \mathbb{R}$

$$f_0\|t\|^2 \leq \mathbb{E}\|t\|_n^2 = \|t\|_f^2 \leq \|f\|_\infty \|t\|^2$$

and then the mean of this empirical norm is equivalent to the L^2 norm $\|.\|$.

3.2. **About the definition of the estimator.** We discuss now the definition of $\hat{\pi}_m$ given by (6). Let $t(x, y) = \sum_{j \in J_m} \sum_{k \in K_m} a_{j,k} \varphi_j^m(x) \psi_k^m(y)$ a function in S_m. Then,

$$\frac{\partial \gamma_n(t)}{\partial a_{j_0,k_0}} = 0 \Leftrightarrow \sum_{j \in J_m} a_{j,k_0} \frac{1}{n}\sum_{i=1}^{n} \varphi_j^m(X_i)\varphi_{j_0}^m(X_i) = \frac{1}{n}\sum_{i=1}^{n} \varphi_{j_0}^m(X_i)\hat{w}_i \psi_{k_0}^m(Z_i), \tag{9}$$

which implies that

$$\forall j_0 \forall k_0 \quad \frac{\partial \gamma_n(t)}{\partial a_{j_0,k_0}} = 0 \Leftrightarrow G_m A_m = \Upsilon_m,$$

where A_m denotes the matrix $(a_{j,k})_{j \in J_m, k \in K_m}$,

$$G_m = \left(\frac{1}{n}\sum_{i=1}^{n} \varphi_j^m(X_i)\varphi_l^m(X_i)\right)_{j,l \in J_m} \text{ and } \Upsilon_m = \left(\frac{1}{n}\sum_{i=1}^{n} \varphi_j^m(X_i)\hat{w}_i \psi_k^m(Z_i)\right)_{j \in J_m, k \in K_m}.$$

In fact, we cannot define a unique minimizer of the contrast $\gamma_n(t)$, since G_m is not necessarily invertible. For example, G_m is not invertible if there exists j_0 in J_m such that there is no observation in the support of φ_{j_0} (G_m has a null column). This phenomenon happens when localized bases (as histogram bases or piecewise polynomial bases) are used. However, the following proposition will still enable us to define an estimator:

Proposition 1.

$$\forall j_0 \forall k_0 \quad \frac{\partial \gamma_n(t)}{\partial a_{j_0,k_0}} = 0 \Leftrightarrow \forall y \quad (t(X_i, y))_{1 \leq i \leq n} = P_{\mathcal{W}}\left(\left(\sum_k \hat{w}_i \psi_k^m(Z_i)\psi_k^m(y)\right)_{1 \leq i \leq n}\right)$$

where $P_{\mathcal{W}}$ denotes the orthogonal projection on $\mathcal{W} = \{(t(X_i, y))_{1\le i\le n},\ t \in S_m\}$ with the euclidian scalar product $\langle\ .\ \rangle_{\mathbb{R}^n}$ in $\mathbb{R}^n$.

Thus the minimization of $\gamma_n(t)$ leads to a unique vector $(\hat{\pi}_m(X_i, y))_{1\le i\le n}$ defined as the projection of $(\sum_k \hat{w}_i\psi_k^m(Z_i)\psi_k^m(y))_{1\le i\le n}$ on $\mathcal{W}$. The associated function $\hat{\pi}_m(.,.)$ is not defined uniquely but we can choose a function $\hat{\pi}_m$ in S_m whose values at (X_i, y) are fixed according to Proposition 1. For the sake of simplicity, we use the notation (6). The underlying function is a theoretical tool but the estimator is actually the vector $(\hat{\pi}_m(X_i, y))_{1\le i\le n}$.

This remark leads to consider the risk defined with the empirical norm, as given by (8).

3.3. **Link with classical regression.** Let $\pi_m(x, y)$ be the orthogonal projection of π on $F_{m_1} \otimes H_{m_2}$. Then

$$\pi_m(x, y) = \sum_{k\in K_m} \pi_k^{F_{m_1}}(x)\psi_k^m(y)$$

where $\pi_k^{F_{m_1}}$ is the orthogonal projection of π_k on F_{m_1} and $\pi_k(x) = \mathbb{E}(\psi_k^m(Y_1)|X_1 = x)$. Then a natural way to estimate π is to estimate the functional coordinates $\pi_k^{F_{m_1}}(x)$, for which a set of regression equations arises by writing

$$\psi_k^m(Y_i) = \pi_k(X_i) + \varepsilon_{i,k},\ \text{where } \varepsilon_{i,k} = \psi_k^m(Y_i) - \mathbb{E}(\psi_k^m(Y_i)|X_i),\ i = 1, \ldots, n.$$

If no censoring occurs, the above equalities lead to well-known mean square contrast estimation of π_k via minimization over $u \in F_{m_1}$ of

$$\gamma_n^{(k)}(u) = \frac{1}{n}\sum_{i=1}^n [(\psi_k^m(Y_i))^2 - 2\psi_k^m(Y_i)u(X_i)].$$

Defining then

$$\hat{\pi}_k^{(m_1)} = \arg\min_{u\in F_{m_1}} \gamma_n^{(k)}(u)$$

we can propose as an estimator of π: $\sum_{k\in K_m} \hat{\pi}_k^{(m_1)}(x)\psi_k^m(y)$. Selecting the adequate $m = (m_1, m_2)$ via a penalized criterion is not tractable with such a formula.

Noticing that for $t \in F_{m_1} \otimes H_{m_2}$,

$$t(x, y) = \sum_{k\in K_m} t_k(x)\psi_k^m(y) \text{ where } t_k(x) = \int t(x, y)\psi_k^m(y)dy$$

and

$$\int \left(\sum_{k\in K_m} t_k(X_i)\psi_k(y)\right)^2 dy = \sum_{k,k'} t_k(X_i)t_{k'}(X_i)\int \psi_k^m(y)\psi_{k'}^m(y)dy = \sum_k t_k^2(X_i),$$

we find a link between $\gamma_n^{(k)}$ and our effective general contrast γ_n:

$$\begin{aligned}
\gamma_n^0(t) &= \frac{1}{n}\sum_{i=1}^n \left[\int \left(\sum_{k\in K_m} t_k(X_i)\psi_k^m(y)\right)^2 dy - 2\sum_{k\in K_m} t_k(X_i)\psi_k^m(Y_i)\right] \\
&= \sum_{k\in K_m} \gamma_n^{(k)}(t_k).
\end{aligned}$$

This shows how γ_n globalizes the procedure and allows model selection.

The same considerations hold in the censored case by replacing $\psi_k^m(Y_i)$ by $W_{i,k}$ or $\hat{W}_{i,k}$ where

$$W_{i,k} = w_i \psi_k^m(Z_i), \;\; \hat{W}_{i,k} = \hat{w}_i \psi_k^m(Z_i) \qquad \text{for } i \in \{1, \dots, n\}.$$

The regression equation becomes:

$$W_{i,k} = \pi_k(X_i) + \varepsilon_{i,k}, \text{ where } \varepsilon_{i,k} = w_i \psi_k(Y_i) - \mathbb{E}[w_i \psi_k(Y_i)|X_i], \tag{10}$$

and

$$\hat{W}_{i,k} = W_{i,k} + R_{i,k} = \pi_k(X_i) + \varepsilon_{i,k} + R_{i,k}, \tag{11}$$

where $R_{i,k} = \psi_k^m(Z_i)(\hat{w}_i - w_i)$, which is a negligible residual term. Note that the residual is null if there is no censoring.

4. Main results

4.1. **Uncensored case.** For a function h and a subspace S, let

$$d(h, S) = \inf_{g \in S} \|h - g\| = \inf_{g \in S} \left(\iint |h(x, y) - g(x, y)|^2 dx dy \right)^{1/2}.$$

With an inequality of Talagrand (1996), we can prove the following result in the uncensored case.

Theorem 1. *We consider the uncensored model described in Section 2.1 satisfying Assumptions* $[\mathcal{A}1]$–$[\mathcal{A}2]$*. We consider* $\tilde{\pi}$ *the estimator of the conditional density* π *described in Section 3 with models verifying Assumptions* [M1]–[M3] *and the following penalty:*

$$\text{pen}(m) = K_0 \|\pi\|_\infty \frac{D_{m_1} D_{m_2}}{n}, \tag{12}$$

where K_0 *is a numerical constant. Then*

$$\mathbb{E}\|\pi \mathbb{1}_A - \tilde{\pi}\|_n^2 \le C \inf_{m \in \mathcal{M}_n} \{d^2(\pi \mathbb{1}_A, S_m) + \text{pen}(m)\} + \frac{C'}{n} \tag{13}$$

where $C = \max(5\|f_X\|_\infty, 6)$ *and* C' *is a constant depending on* $\phi_1, \phi_2, \|\pi\|_\infty, f_0, \|f_X\|_\infty$.

The penalty (12) deserves some comments. First, the constant K_0 in the penalty is purely numerical and calibrated via simulations. On the other hand, the term $\|\pi\|_\infty$ is unknown. Note that, as inequality (13) holds for any penalty pen(.) such that $\text{pen}(m) \ge K_0 \|\pi\|_\infty D_{m_1} D_{m_2}/n$, any upper bound on $\|\pi\|_\infty$ gives a good result. In practice, $\|\pi\|_\infty$ is often replaced by an estimator $\|\hat{\pi}\|_\infty$ where $\hat{\pi}$ is an estimator of π. From a theoretical point of view, this makes the penalty random, and the procedure can be proved to give the same result as with the penalty above under some additional regularity constraints, see Birgé and Massart (1997) or Comte (2001).

We can deduce from Theorem 1 the rate of convergence of the risk. For that purpose, assume that π restricted to A belongs to the anisotropic Besov space on A with regularity $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$. Note that if π belongs to $B_{2,\infty}^{\boldsymbol{\alpha}}(\mathbb{R}^2)$, then π restricted to A belongs to $B_{2,\infty}^{\boldsymbol{\alpha}}(A)$. Let us recall the definition of $B_{2,\infty}^{\boldsymbol{\alpha}}(A)$. Let e_1 and e_2 be the canonical basis vectors in $\mathbb{R}^2$ and for $i = 1, 2$, $A^r_{h,i} = \{x \in \mathbb{R}^2; x, x + he_i, \dots, x + rhe_i \in A\}$. Next, for x in $A^r_{h,i}$, let

$$\Delta^r_{h,i} g(x) = \sum_{k=0}^{r} (-1)^{r-k} \binom{r}{k} g(x + khe_i)$$

the rth difference operator with step h. For $t > 0$, the directional moduli of smoothness are given by

$$\omega_{r_i,i}(g, t) = \sup_{|h| \le t} \left(\int_{A^{r_i}_{h,i}} |\Delta^{r_i}_{h,i} g(x)|^2 dx \right)^{1/2}.$$

We say that g is in the Besov space $B^{\boldsymbol{\alpha}}_{2,\infty}(A)$ if $\sup_{t>0} \sum_{i=1}^{2} t^{-\alpha_i} \omega_{r_i,i}(g, t) < \infty$ for r_i integers larger than α_i.

The estimation procedure described in the uncensored case allows an adaptation of the approximation space to each directional regularity. More precisely, assume for example that $\alpha_2 > \alpha_1$. Then the proof of Corollary 1 below shows that the estimator chooses a space of dimension $D_{m_2} = D_{m_1}^{\alpha_1/\alpha_2} < D_{m_1}$.

Corollary 1. *Assume that π restricted to A belongs to the anisotropic Besov space $B^{\boldsymbol{\alpha}}_{2,\infty}(A)$ with regularity $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$ such that $\alpha_1 > 1/2$ and $\alpha_2 > 1/2$. We consider the spaces described in Subsection 2.4 (with the regularity r of the polynomials and the wavelets larger than $\alpha_i - 1$). Then, under the assumptions of Theorem 1,*

$$\mathbb{E}\|\pi \mathbb{1}_A - \tilde{\pi}\|_n^2 = O(n^{-\frac{2\bar{\alpha}}{2\bar{\alpha}+2}}).$$

where $\bar{\alpha}$ is the harmonic mean of α_1 and α_2 (i.e. $2/\bar{\alpha} = 1/\alpha_1 + 1/\alpha_2$).

Thus we obtain the rate of convergence $n^{-\frac{2\bar{\alpha}}{2\bar{\alpha}+2}}$, which is optimal in the minimax sense (see Lacour (2007) for the lower bound).

Remark 1. The empirical norm is the more natural in this problem, but if we were interested in a L^2 control of the risk, we may modify the estimation procedure as follows:

$$\tilde{\pi}^* = \begin{cases} \tilde{\pi} & \text{if } \|\tilde{\pi}\| \le k_n \\ 0 & \text{else} \end{cases} \tag{14}$$

with $k_n = n^{2/3}$. We may prove in this framework a result similar to Theorem 1 but bounding $\mathbb{E}\|\tilde{\pi}^* - \pi \mathbb{1}_A\|^2$ instead of its empirical version, see Lacour (2007).

4.2. **Censored case.** In the censored case, the penalty coming out under the same assumptions as previously is

$$\text{pen}(m) = K_0(\|\pi\|_\infty / c_G)(D_{m_1} D_{m_2}/n), \tag{15}$$

but c_G is a quantity that cannot be easily estimated. Therefore, we use more restrictive assumptions to prove:

Theorem 2. *We consider the censored model described in Section 2.1 satisfying Assumptions [$\mathcal{A}$1]–[$\mathcal{A}$3]. We consider $\tilde{\pi}$ the estimator of the conditional density π described in*

	Example 1						Example 2					
n	200		500		2000		200		500		2000	
Censoring	0%	40 %	0%	40%	0%	40%	0%	40%	0%	40%	0%	40%
H	4.91	9.40	2.97	5.50	1.55	3.31	1.73	4.11	1.16	2.93	0.65	2.01
TP	2.70	4.76	2.16	3.19	1.97	2.98	1.43	2.19	1.31	2.17	0.91	1.79

TABLE 1. Monte-Carlo results (MISE $\times$ 100) for the estimator $\tilde{\pi}$, for $K =$ 500 replications and two bases: H for histograms and TP for Trigonometric polynomials. Examples 1 and 2.

Section 3 with (nested) models verifying Assumptions [M2]–[M4]. *We choose the following penalty:*

$$\text{pen}(m) = K_0 \frac{\phi_0^2}{f_0} \mathbb{E}\left(\frac{\delta_1}{\bar{G}^2(Z_1)}\right) \frac{D_{m_1} D_{m_2}}{n}, \tag{16}$$

where K_0 is a numerical constant. Then

$$\mathbb{E}\|\pi \mathbb{1}_A - \tilde{\pi}\|_n^2 \leq C \inf_{m \in \mathcal{M}_n} \{d^2(\pi \mathbb{1}_A, S_m) + \text{pen}(m)\} + \frac{C'}{n}$$

where $C = \max(5\|f_X\|_\infty, 6)$ and C' is a constant depending on $\phi_1, \phi_2, \|\pi\|_\infty, f_0$, $\|f_X\|_\infty$.

Here the penalty involves two quantities that should be estimated: f_0, the lower bound of f_X the density of the X_i's on the interval A_1 and the expectation $\mathbb{E}(\delta_1/\bar{G}^2(Z_1))$, see Section 5 for practical implementation.

5. SIMULATIONS AND EXAMPLES

5.1. **Simulation.** We study the estimation procedure by generating samples (X_i, Y_i) following four models:

- Example 1. Let $Y_i = b(X_i) + \varepsilon_i$, with ε_i i.i.d. $\mathcal{N}(0,1)$, X_i i.i.d. uniform $\mathcal{U}([0,1])$, $b(x) = 2x + 5$. We take $A = [0,1] \times [3, 9.5]$.
- Example 2. Let $Y_i = b(X_i) + \varepsilon_i$, with ε_i i.i.d. $\Gamma(4,1)$, X_i i.i.d. uniform $\mathcal{U}([0,1])$, $b(x) = 2x + 5$. We take $A = [0,1] \times [4, 15]$.
- Example 3. Let $Y_i = b(X_i) + \sigma(X_i)\varepsilon_i$, with ε_i i.i.d. $\Gamma(4,1)$, X_i i.i.d. uniform $\mathcal{U}([0,1])$, $b(x) = 2x^2 + 5$, $\sigma(x) = \sqrt{1.3 - |x|}$. We take $A = [0,1] \times [6, 14]$.
- Example 4. Given $X_i = x$, let Y_i follow the distribution $0.5\mathcal{N}(8-4x, 1) + 0.5\mathcal{N}(8+4x, 1)$. The X_i's are i.i.d. $\mathcal{U}([0,1])$. We take $A = [0,1] \times [2, 14]$.

The sets $A = A_1 \times A_2$ are fixed intervals, roughly calibrated with respect to each distribution. In the first three cases, the conditional density π is given by

$$\pi(x,y) = f_\varepsilon((y - b(x))/\sigma(x))/\sigma(x),$$

with $b(x) = 2x + 5$ (examples 1,2) or $b(x) = 2x^2 + 5$ (example 3) and $\sigma(x) \equiv 1$ (Examples 1 and 2) or $\sigma(x) = \sqrt{1.3 - |x|}$ (Example 3). For Example 4, we have $\pi(x,y) = 0.5\exp(-(y - 8 + 4x)^2/2)/\sqrt{2\pi} + 0.5\exp(-(y - 8 - 4x)^2/2)/\sqrt{2\pi}$.

	Example 3						Example 4					
n	200		500		2000		200		500		2000	
Censoring	0%	40 %	0%	40%	0%	40%	0%	20%	0%	20%	0%	20%
H	2.28	4.77	1.54	3.16	0.88	2.07	3.23	4.43	2.08	3.17	1.17	2.05
TP	1.29	3.44	1.21	2.36	1.01	1.90	5.45	6.77	5.04	5.60	2.72	3.34

TABLE 2. Monte-Carlo results (MISE $\times$ 100) for the estimator $\tilde{\pi}$, for $K =$ 500 replications and two bases: H for histograms and TP for trigonometric polynomials. Examples 3 and 4.

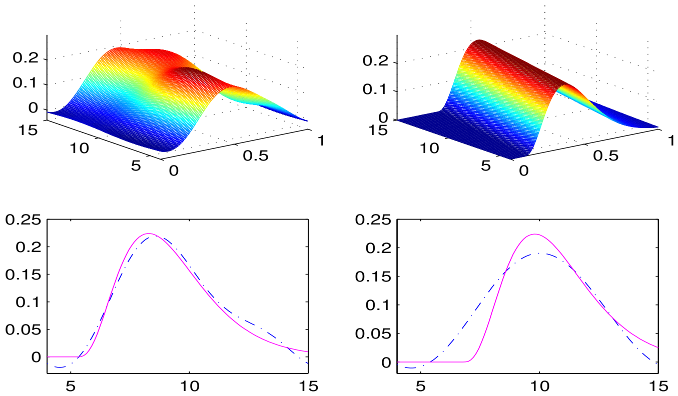

FIGURE 1. Plots of the estimated (top-left) with trigonometric basis and the true (top right) conditional density and $y \mapsto \pi(x, y)$ (full line), $\tilde{\pi}(x, y)$ (dashed dotted line) for $x = 0.15$ (bottom-left) and $x = 0.90$ (bottom-right) with $n = 2000$ observations in Example 2 and without censoring.

The penalty is chosen as follows:

$$0.5\|\hat{\pi}\|_\infty \left(\frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i}{\hat{\bar{G}}^2(Z_i)} \right) \frac{D_{m_1} D_{m_2}}{n}, \tag{17}$$

where $\|\hat{\pi}\|_\infty$ is preliminary estimated. We replace $\|\hat{\pi}\|_\infty$ by a bound equal to 0.4 in Examples 1, 3 and 4 and to 0.3 in Example 2. We mentioned in Section 4.1 that an upper bound on π could suit. If the data are uncensored, we recover the empirical version of (12) with constant K_0 calibrated as 0.5. Indeed, the term $(1/n) \sum_{i=1}^{n} \delta_i / \hat{\bar{G}}^2(Z_i)$ is equal to 1 and thus vanishes. In the case of censored data, this term stands for the empirical version of $\mathbb{E}(\delta_1/\bar{G}^2(Z_1))$ (see also Brunel and Comte (2005)). Then, we should follow formula (16) and estimate f_0. But, looking at (15) and for sake of robustness of the formula from uncensored to censored case, we make a compromise by keeping the factor $\|\hat{\pi}\|_\infty$.

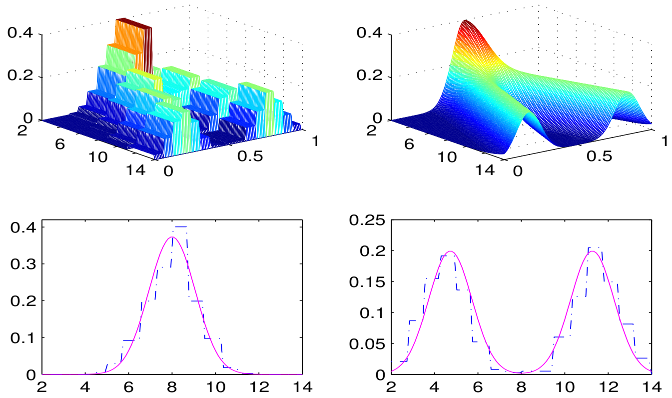

FIGURE 2. Plots of the estimated (top-left) with histogram basis and the true (top right) conditional density and $y \mapsto \pi(x, y)$ (full line), $\tilde{\pi}(x, y)$ (dashed dotted line) for $x = 0.1$ (bottom-left) and $x = 0.82$ (bottom-right), with $n = 2000$ observations in Example 4 and without censoring.

We compute the empirical MISE (Mean Integrated Squared Error) over $N = 500$ replications of the samples, by averaging over the paths $i = 1, \ldots, N$, the quantities

$$\frac{\ell(A_1)\ell(A_2)}{K^2} \sum_{k=1}^{K} (\tilde{\pi}^{(i)}(x_k, y_k) - \pi(x_k, y_k))^2,$$

where $\ell(A_i)$ is the length of the interval A_i, $i = 1, 2$, $(x_k)_{1 \le k \le K}$, $(y_k)_{1 \le k \le K}$ are subdivisions of A_1 and A_2 respectively, and $\tilde{\pi}^{(i)}$ is the estimate associated to the ith sample path. Note that we compute $\mathbb{L}^2$-type errors in both x- and y-directions instead of using the empirical norm in the x-direction: this is to allow comparison with other methods.

When censoring occurs, the C_i's are generated as exponential random variables $\mathcal{E}(c)$ with parameter c empirically adjusted to reach a given censoring rate (20% or 40%), namely $c = 6.97$ for Example 1 (40%), $c = 11.11$ for Example 2 (40%), $c = 10.25$ for Example 3 (40%), $c = 11.65$ for Example 4 (20%). Figures 1 and 2 illustrate the appearance of our estimates.

We experimented our method using histogram and trigonometric bases for all our examples, see Tables 1 and 2. We cannot pretend that one basis is much better than the others since the performances of the estimator depend on the example: sometimes, the fact that the histogram basis is localized is most useful to detect a change of modality, e.g. for Example 4; besides, the smoothness of the trigonometric basis gives better results. Example 3 checks that heteroscedasticity in the model is correctly handled, and Example 4 mimicks the real data illustration of Section 5.2 (change in modality). In all examples, censoring degrades the results. This is why only 20% of censoring is studied for example 4.

Monte Carlo experiment results are reported in Tables 1 and 2. As expected, censoring deteriorates the result, but surprisingly, even in smooth Gaussian cases, histograms seem to work very well and in particular when a change in modality occurs as in Example 4.

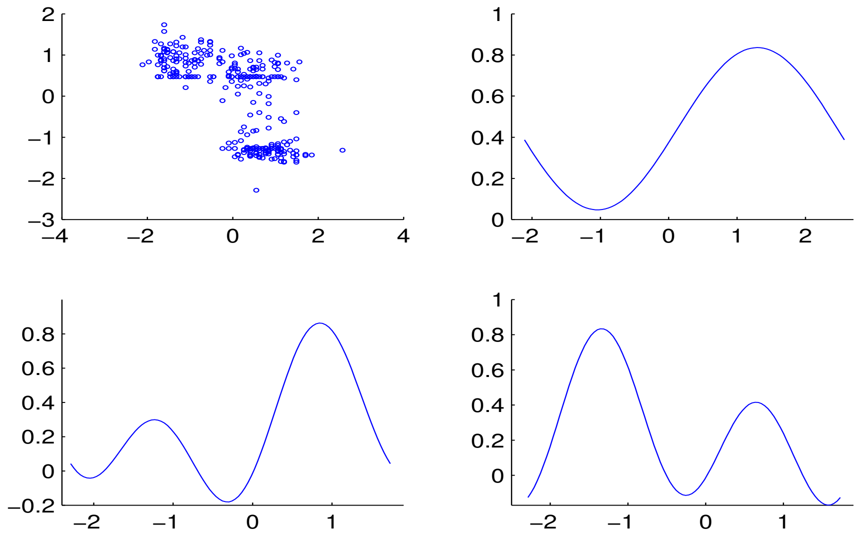

FIGURE 3. Plots of the Old Faithful geyser data (top right), estimators of the conditional density with $x \mapsto \pi(x, y)$ for $y = y_{25} = -1.31$ (top-right) and $y \mapsto \pi(x, y)$ for $x = x_5 = -1.92$ (bottom-left) and $x = x_{75} = 1.39$ (bottom-right).

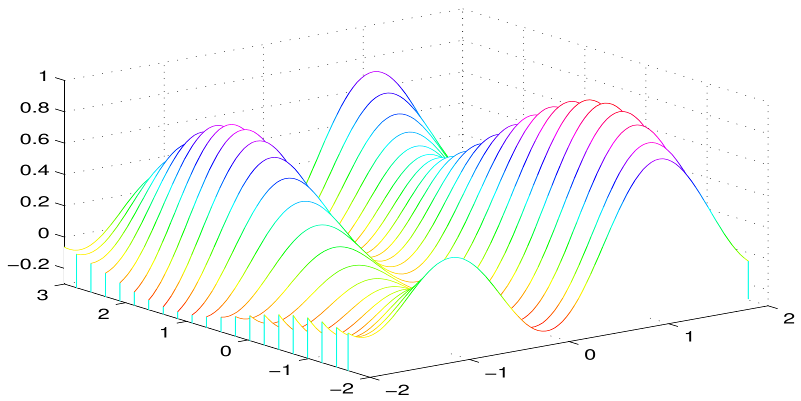

FIGURE 4. Two-dimensional plot of the estimate of the conditional density of eruption duration given the waiting time to the eruption for the Old Faithful geyser data using trigonometric polynomials.

5.2. **Example.** To compare our results with De Gooijer and Zerom (2003)'s and Hyndman and Yao (2002)'s, we provide the plot of our estimation results on the Old Faithful Geyser data, available from Azzalini and Bowman (1990): these are continuously collected data of 299 observations obtained between 1 and 15 August 1985, waiting time between the starts of successive eruptions and duration of the subsequent eruption for the infamous Old Faithful geyser in Yellowstone National Park, Wyoming, getting its name because of its punctuality and predictability. Over the years, the length of the interval has increased, which may be the result of earthquakes affecting subterranean water levels. This have made the earlier mathematical relationship between duration and interval between eruptions inaccurate. Nonparametric estimation of the conditional density of the duration time avoid this problem. The data have been transformed to be centered and with unit variance for comparison with the previous references. Our results are plotted in Figures 3 and 4. The estimate is computed using the penalty given by (17) with $\|\hat{\pi}\|_\infty$ estimated by 1 (see the picture to read the upper bound). We can see that our results are in accordance with previous ones: when the waiting time between eruptions is relatively short, the duration of the next eruption is long, whereas for a longer waiting time (longer than round about 70 minutes), the duration of the next eruption is a mixture of long and short durations. This results in a unimodal followed by a bimodal behavior of the conditional density of the duration of eruption given the waiting time between eruptions.

6. Proofs

6.1. **Proof of Proposition 1.**

Equality (9) yields, by multiplying by $\psi^m_{k_0}(y)$,

$$\sum_{j\in J_m} a_{j,k_0}\sum_{i=1}^n \varphi^m_j(X_i)\psi^m_{k_0}(y)\varphi^m_{j_0}(X_i) = \sum_{i=1}^n \varphi^m_{j_0}(X_i)\hat{w}_i\psi^m_{k_0}(Z_i)\psi^m_{k_0}(y).$$

Then, we sum over k_0 in K_m:

$$\sum_{i=1}^n t(X_i,y)\varphi^m_{j_0}(X_i) = \sum_{i=1}^n\sum_{k_0\in K_m}\hat{w}_i\psi^m_{k_0}(Z_i)\psi^m_{k_0}(y)\varphi^m_{j_0}(X_i).$$

If we multiply this equality by $a'_{j_0,k}\psi^m_k(y)$ and if we sum over $k\in K_m$ and $j_0\in J_m$, we obtain

$$\sum_{i=1}^n [t(X_i,y) - \sum_{k_0\in K_m}\hat{w}_i\psi^m_{k_0}(Z_i)\psi^m_{k_0}(y)]\sum_{k\in K_m}\sum_{j_0\in J_m} a'_{j_0,k}\varphi^m_{j_0}(X_i)\psi^m_k(y) = 0$$

i.e.
$$\sum_{i=1}^n [t(X_i,y) - \sum_{k_0\in K_m}\hat{w}_i\psi^m_{k_0}(Z_i)\psi^m_{k_0}(y)]u(X_i,y) = 0$$

for all u in S_m. So the vector $(t(X_i,y) - \sum_{k\in K_m}\hat{w}_i\psi^m_k(Z_i)\psi^m_k(y))_{1\le i\le n}$ is orthogonal to each vector in $\mathcal{W}$. Since $t(X_i,y)$ belongs to $\mathcal{W}$, the proposition is proved.

6.2. **Proof of Theorem 1 and 2.**

For ρ a real larger than 1, let

$$\Omega_\rho = \{\forall t\in\mathcal{S}\quad \|t\|_f^2 \le \rho\|t\|_n^2\}$$

We denote by π_m the orthogonal projection of π on S_m. Now,

$$\mathbb{E}\|\tilde{\pi} - \pi\mathbb{1}_A\|_n^2 = \mathbb{E}\left(\|\tilde{\pi} - \pi\mathbb{1}_A\|_n^2\mathbb{1}_{\Omega_\rho}\right) + \mathbb{E}\left(\|\tilde{\pi} - \pi\mathbb{1}_A\|_n^2\mathbb{1}_{\Omega_\rho^c}\right) \tag{18}$$

To bound the first term, we observe that for all s,t

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

where

$$\nu_n(t) = \frac{1}{n}\sum_{i=1}^{n}\left\{w_i t(X_i, Z_i) - \int_{\mathbb{R}} t(X_i, y)\pi(X_i, y)dy\right\},$$

$$R_n(t) = \frac{1}{n}\sum_{i=1}^{n} t(X_i, Z_i)\left[\hat{w}_i - w_i\right].$$

Since $\|t-\pi\|_n^2 = \|t - \pi\mathbb{1}_A\|_n^2 + \|\pi\mathbb{1}_{A^c}\|_n^2$, we can write

$$\gamma_n(t) - \gamma_n(s) = \|t - \pi\mathbb{1}_A\|_n^2 - \|s - \pi\mathbb{1}_A\|_n^2 - 2\nu_n(t-s) - 2R_n(t-s).$$

The definition of $\hat{m}$ gives, for some fixed $m \in \mathcal{M}_n$,

$$\gamma_n(\tilde{\pi}) + \mathrm{pen}(\hat{m}) \le \gamma_n(\pi_m) + \mathrm{pen}(m)$$

And then

$$\begin{aligned}
\|\tilde{\pi} - \pi\mathbb{1}_A\|_n^2 &\le \|\pi_m - \pi\mathbb{1}_A\|_n^2 + \mathrm{pen}(m) + 2R_n(\tilde{\pi} - \pi_m) \\
&\quad +2\nu_n(\tilde{\pi} - \pi_m) - \mathrm{pen}(\hat{m}) \\
&\le \|\pi_m - \pi\mathbb{1}_A\|_n^2 + \mathrm{pen}(m) + 2\|\tilde{\pi} - \pi_m\|_f \sup_{t\in B_f(\hat{m})} R_n(t) \\
&\quad +2\|\tilde{\pi} - \pi_m\|_f \sup_{t\in B_f(\hat{m})} \nu_n(t) - \mathrm{pen}(\hat{m})
\end{aligned}$$

where, for all m', $B_f(m') = \{t \in S_m + S_{m'}, \quad \|t\|_f = 1\}$. Let θ a real larger than 2ρ and $p(.,.)$ a function such that $2\theta p(m,m') \le \mathrm{pen}(m) + \mathrm{pen}(m')$. Then

$$\begin{aligned}
\|\tilde{\pi} - \pi\mathbb{1}_A\|_n^2\mathbb{1}_{\Omega_\rho} &\le \|\pi_m - \pi\mathbb{1}_A\|_n^2 + \frac{1}{\theta}\|\tilde{\pi} - \pi_m\|_f^2\mathbb{1}_{\Omega_\rho} + 2\mathrm{pen}(m) \\
&\quad +2\theta\sum_{m'\in\mathcal{M}_n}\left[\sup_{t\in B_f(m')} \nu_n^2(t) - p(m,m')\right]_+ \mathbb{1}_{\Omega_\rho} \\
&\quad +2\theta\sum_{m'\in\mathcal{M}_n}\sup_{t\in B_f(m')} R_n^2(t)\mathbb{1}_{\Omega_\rho}
\end{aligned} \tag{19}$$

But $\|\tilde{\pi} - \pi_m\|_f^2\mathbb{1}_{\Omega_\rho} \le \rho\|\tilde{\pi} - \pi_m\|_n^2\mathbb{1}_{\Omega_\rho} \le 2\rho\|\tilde{\pi} - \pi\mathbb{1}_A\|_n^2\mathbb{1}_{\Omega_\rho} + 2\rho\|\pi\mathbb{1}_A - \pi_m\|_n^2$.

Then, inequality (19) becomes

$$\begin{aligned}\|\tilde{\pi}-\pi\mathbb{1}_A\|_n^2\mathbb{1}_{\Omega_\rho}\left(1-\frac{2\rho}{\theta}\right) &\leq \left(1+\frac{2\rho}{\theta}\right)\|\pi_m-\pi\mathbb{1}_A\|_n^2+2\mathrm{pen}(m)\\ &\quad +2\theta\sum_{m'\in\mathcal{M}_n}\left[\sup_{t\in B_f(m')}\nu_n^2(t)-p(m,m')\right]_+\mathbb{1}_{\Omega_\rho}\\ &\quad +2\theta\sum_{m'\in\mathcal{M}_n}\sup_{t\in B_f(m')}R_n^2(t)\mathbb{1}_{\Omega_\rho}\end{aligned}$$

$$\begin{aligned}\text{so}\quad \mathbb{E}\left(\|\tilde{\pi}-\pi\mathbb{1}_A\|_n^2\mathbb{1}_{\Omega_\rho}\right) &\leq \frac{\theta+2\rho}{\theta-2\rho}\mathbb{E}\|\pi\mathbb{1}_A-\pi_m\|_n^2+\frac{2\theta}{\theta-2\rho}\mathrm{pen}(m)\\ &\quad +\frac{2\theta^2}{\theta-2\rho}\sum_{m'\in\mathcal{M}_n}\mathbb{E}\left(\left[\sup_{t\in B_f(m')}\nu_n^2(t)-p(m,m')\right]_+\mathbb{1}_{\Omega_\rho}\right)\\ &\quad +\frac{2\theta^2}{\theta-2\rho}\mathbb{E}\left(\left[\sup_{t\in B_f(\hat{m})}R_n^2(t)\right]_+\mathbb{1}_{\Omega_\rho}\right)\end{aligned}\tag{20}$$

We now use the following proposition:

Proposition 2. *Under the assumptions of Theorem 1, with* $p(m,m')=6\|\pi\|_\infty D(m,m')/n$ *where* $D(m,m')$ *denotes the dimension of* $S_m+S_{m'}$ *or under the assumptions of Theorem 2, with* $p(m,m')=6(\phi_0^2/f_0)\mathbb{E}(\delta_1/\bar{G}^2(Z_1))D(m,m')/n$*, there exists a constant* C_1 *such that*

$$\sum_{m'\in\mathcal{M}_n}\mathbb{E}\left(\left[\sup_{t\in B_f(m')}\nu_n^2(t)-p(m,m')\right]_+\mathbb{1}_{\Omega_\rho}\right)\leq\frac{C_1}{n}.\tag{21}$$

Moreover, we can prove that

$$\mathbb{E}\left(\left[\sup_{t\in B_f(\hat{m})}R_n^2(t))\right]_+\mathbb{1}_{\Omega_\rho}\right)\leq\mathbb{E}\left(\left[\sup_{t\in B_f(n)}R_n^2(t))\right]_+\mathbb{1}_{\Omega_\rho}\right)\leq\frac{C}{n}\tag{22}$$

where $B_f(n)$ is the unit ball of the largest space of the collection (which is nested as R_n appears only in the censored case).

To prove (22), let us define

$$\Omega_G=\{\omega,\forall y\in A_2,\hat{\bar{G}}(y)-G(y)>-c_G/2\}.\tag{23}$$

On Ω_G, $\hat{\bar{G}}(y)>c_G/2$ and $\hat{\bar{G}}(y)\geq 1/(n+1)$. Now we use the following key lemma, useful to control the probability of the uniform deviation of the estimator of the survival distribution function $\hat{\bar{G}}$:

Lemma 1. *For all* $k\in\mathbb{N}^*$*, there exists a constant* C_k *depending on* k *and* c_G *such that*

$$\mathbb{E}\left(\sup_{y\in A_2}|\hat{\bar{G}}(y)-\bar{G}(y)|^{2k}\right)\leq\frac{C_k}{n^k}.$$

This lemma is proved in Brunel and Comte (2005), see Lemma 6.1. Now write, for (φ_j, ψ_k) an orthonormal basis of the largest space of the collection,

$$\begin{aligned}
\mathbb{E}\left(\left[\sup_{t\in B_f(n)} R_n^2(t))\right]_+ \mathbb{1}_{\Omega_\rho}\mathbb{1}_{\Omega_G^c}\right) &\leq \frac{1}{f_0}\sum_{j,k}\mathbb{E}(R_n^2(\varphi_j\psi_k)\mathbb{1}_{\Omega_G^c}) \\
&\leq \frac{\phi_0^2\mathcal{D}_n^2}{c_G^2 f_0}\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n \frac{(\hat{\bar{G}}(Y_i)-\bar{G}(Y_i))^2\mathbb{1}_{A_2}(Y_i)}{\hat{\bar{G}}^2(Y_i)}\mathbb{1}_{\Omega_G^c}\right) \\
&\leq \frac{\phi_0^2\mathcal{D}_n^2}{c_G^2 f_0}(n+1)^2\mathbb{E}(\sup_{y\in A_2}|\hat{\bar{G}}(y)-\bar{G}(y)|^2\mathbb{1}_{\sup_{y\in A_2}|\hat{\bar{G}}(y)-\bar{G}(y)|>c_G/2}) \\
&\leq \frac{2\phi_0^2 n^3}{c_G^2 f_0}(2/c_G)^6\mathbb{E}(\sup_{y\in A_2}|\hat{\bar{G}}(y)-\bar{G}(y)|^8) \leq \frac{C}{n}.
\end{aligned}$$

Next, we need to study R_n on Ω_G. To this end, we write that:

$$\begin{aligned}
\mathbb{E}\left(\left[\sup_{t\in B_f(n)} R_n^2(t)\right]_+ \mathbb{1}_{\Omega_\rho}\mathbb{1}_{\Omega_G}\right) &\leq \frac{4}{c_G^4}\mathbb{E}\left[\sup_{y\in A_2}|\hat{\bar{G}}(y)-\bar{G}(y)|^2\sup_{t\in B_f(n)}\left(\frac{1}{n}\sum_{i=1}^n t^2(X_i,Y_i)\right)\right] \\
&\leq \frac{4}{c_G^4}\mathbb{E}\left(\sup_{y\in A_2}|\hat{\bar{G}}(y)-\bar{G}(y)|^2\right)\sup_{t\in B_f(n)}\mathbb{E}(t^2(X_1,Y_1)) \quad (24) \\
&\quad + \frac{4}{c_G^4}\mathbb{E}^{1/2}\left(\sup_{y\in A_2}|\hat{\bar{G}}(y)-\bar{G}(y)|^4\right)\mathbb{E}^{1/2}\left(\sup_{t\in B_f(n)}(\nu_n"(t^2))^2\right) \quad (25)
\end{aligned}$$

where

$$\nu_n"(t) = \frac{1}{n}\sum_{i=1}^n [t(X_i,Y_i)-\mathbb{E}(t(X_1,Y_1))].$$

As under $[\mathcal{A}1]$ and $[\mathcal{A}2]$, $f_{(X,Y)}$ is bounded on A, then $\sup_{t\in B_f(n)}\mathbb{E}(t^2(X_1,Y_1)) \leq F_1$ with $F_1 = \|\pi\|_\infty\|f_X\|_\infty$, and thus using Lemma 1 gives that (24) is of order $1/n$. Next, with Schwartz inequalities,

$$\begin{aligned}
\mathbb{E}\left(\sup_{t\in B_f(n)}(\nu_n"(t^2))^2\right) &\leq \frac{1}{n}\sum_{j,j',k,k'}\mathbb{E}(\varphi_j^2(X_1)\varphi_{j'}^2(X_1))\mathbb{E}(\psi_k^2(Y_1)\psi_{k'}^2(Y_1)) \\
&\leq \frac{\Phi_0^4\mathcal{D}_n^4}{n}.
\end{aligned}$$

It follows that this term is bounded if $\mathcal{D}_n \leq n^{1/4}$. This implies that (25) is also of order $1/n$. Gathering both terms (on Ω_G and Ω_G^c) gives Inequality (22).

Then, with $\theta = 3\rho$, inequalities (20) and (21) yield

$$\mathbb{E}\left(\|\tilde{\pi}-\pi\mathbb{1}_A\|_n^2\mathbb{1}_{\Omega_\rho}\right) \leq 5\|f\|_\infty\|\pi_m-\pi\mathbb{1}_A\|^2 + 6\mathrm{pen}(m) + \frac{9\rho C_1}{n} \quad (26)$$

The penalty term $\mathrm{pen}(m)$ has to verify $\mathrm{pen}(m)+\mathrm{pen}(m') \geq 36\rho AD(m,m')/n$ i.e. $36\rho A\dim(S_m+S_{m'})/n \leq \mathrm{pen}(m)+\mathrm{pen}(m')$, where $A = \|\pi\|_\infty$ in the uncensored case and $A = (\phi_0^2/f_0)\mathbb{E}(\delta_1/\bar{G}(Z_1))$ in the censored case. We choose $\rho = 3/2$ and so $\mathrm{pen}(m) = 54A(D_{m_1}D_{m_2})/n$.

To bound the second term in (18), we recall (see Section 3) that $(\hat{\pi}_{\hat{m}}(X_i, y))_{1\leq i\leq n}$ is the orthogonal projection of $(\sum_k \hat{W}_{i,k}\psi_k(y))_{1\leq i\leq n}$ on

$$\mathcal{W} = \{(t(X_i, y))_{1\leq i\leq n}, \quad t \in S_{\hat{m}}\}$$

where $\psi_k = \psi_k^{\hat{m}}$. Thus, since $P_{\mathcal{W}}$ denotes the orthogonal projection on $\mathcal{W}$, using (10)-(11)

$$\begin{aligned}
&(\hat{\pi}_{\hat{m}}(X_i, y))_{1\leq i\leq n} = P_{\mathcal{W}}((\sum_k \hat{W}_{i,k}\psi_k(y))_{1\leq i\leq n}) \\
=\ & P_{\mathcal{W}}((\sum_k \pi_k(X_i)\psi_k(y))_{1\leq i\leq n}) + P_{\mathcal{W}}((\sum_k \varepsilon_{i,k}\psi_k(y))_{1\leq i\leq n}) + P_{\mathcal{W}}((\sum_k R_{i,k}\psi_k(y))_{1\leq i\leq n}) \\
=\ & P_{\mathcal{W}}(\pi \mathbb{1}_A(X_i, y))_{1\leq i\leq n}) + P_{\mathcal{W}}((\sum_k \varepsilon_{i,k}\psi_k(y))_{1\leq i\leq n}) + P_{\mathcal{W}}((\sum_k R_{i,k}\psi_k(y))_{1\leq i\leq n})
\end{aligned}$$

We denote by $\|.\|_{\mathbb{R}^n}$ the Euclidean norm in $\mathbb{R}^n$, by X the vector $(X_i)_{1\leq i\leq n}$, by ε_k the vector $(\varepsilon_{i,k})_{1\leq i\leq n}$ and by R_k the vector $(R_{i,k})_{1\leq i\leq n}$. Thus

$$\begin{aligned}
&\|\pi \mathbb{1}_A - \hat{\pi}_{\hat{m}}\|_n^2 \\
&= \frac{1}{n}\int \|\pi \mathbb{1}_A(X, y) - P_{\mathcal{W}}(\pi \mathbb{1}_A(X, y)) - P_{\mathcal{W}}(\sum_k \varepsilon_k\psi_k(y)) - P_{\mathcal{W}}(\sum_k R_k\psi_k(y))\|_{\mathbb{R}^n}^2 dy \\
&= \frac{1}{n}\int \|\pi \mathbb{1}_A(X, y) - P_{\mathcal{W}}(\pi \mathbb{1}_A(X, y))\|_{\mathbb{R}^n}^2 dy + \frac{1}{n}\int \|P_{\mathcal{W}}(\sum_k(\varepsilon_k + R_k)\psi_k(y))\|_{\mathbb{R}^n}^2 dy \\
&\leq \frac{1}{n}\int \|\pi \mathbb{1}_A(X, y)\|_{\mathbb{R}^n}^2 dy + \frac{2}{n}\int \|\sum_k \varepsilon_k\psi_k(y)\|_{\mathbb{R}^n}^2 dy + \frac{2}{n}\int \|\sum_k R_k\psi_k(y)\|_{\mathbb{R}^n}^2 dy \\
&\leq \frac{1}{n}\sum_{i=1}^n \|\pi\|_\infty \int \pi(X_i, y)dy + \frac{2}{n}\sum_{i=1}^n \int [\sum_k \varepsilon_{i,k}\psi_k(y)]^2 dy + \frac{2}{n}\sum_{i=1}^n \int [\sum_k R_{i,k}\psi_k(y)]^2 dy \\
&\leq \|\pi\|_\infty + \frac{2}{n}\sum_{i=1}^n \sum_k \varepsilon_{i,k}^2 + \frac{2}{n}\sum_{i=1}^n \sum_k R_{i,k}^2.
\end{aligned}$$

But Assumption [M2] implies $\|\sum_{k\in K_{\hat{m}}} \psi_k^2\|_\infty \leq \phi_2 D_{\hat{m}_2}$. So, using (11),

$$\begin{aligned}
\varepsilon_{i,k}^2 &\leq 2(w_i\psi_k(Z_i))^2 + 2\mathbb{E}[w_i\psi_k(Z_i)|X_i]^2 \\
\text{and} \quad \sum_k \varepsilon_{i,k}^2 &\leq 2\sum_k \frac{\psi_k^2(Y_i)}{\bar{G}^2(Z_i)} + 2\mathbb{E}[\sum_k \frac{\psi_k^2(Y_i)}{\bar{G}^2(Z_i)}|X_i] \leq 4\frac{\phi_2 D_{\hat{m}_2}}{c_G^2}
\end{aligned}$$

On the other hand,

$$\sum_k R_{i,k}^2 \leq \sum_k \psi_k^2(Y_i)\frac{|\widehat{\bar{G}}(Y_i) - \bar{G}(Y_i)|^2 \mathbb{1}_{A_2}(Y_i)}{c_G^2 \widehat{\bar{G}}^2(Y_i)} \leq \frac{\phi_2 D_{\hat{m}_2}}{c_G^2}\frac{|\widehat{\bar{G}}(Y_i) - \bar{G}(Y_i)|^2 \mathbb{1}_{A_2}(Y_i)}{\widehat{\bar{G}}^2(Y_i)}.$$

Thus, it follows from the previous study (and in particular from Lemma 1), that the following inequality holds

$$\mathbb{E}^{1/2}\left[\left(\frac{1}{n}\sum_{i=1}^n \frac{|\widehat{\bar{G}}(Y_i) - \bar{G}(Y_i)|^2 \mathbb{1}_{A_2}(Y_i)}{\widehat{\bar{G}}^2(Y_i)}\right)^2\right] \leq \frac{\kappa}{n}. \tag{27}$$

Using (27), we obtain

$$\mathbb{E}\left(\|\pi\mathbb{1}_A - \hat{\pi}_{\hat{m}}\|_n^2 \mathbb{1}_{\Omega_\rho^c}\right) \le (\|\pi\|_\infty + 8\frac{\phi_2 n^{1/2}}{c_G^2})\mathbb{P}(\Omega_\rho^c) + \frac{2\kappa\phi_2}{c_G^2 n^{1/2}}\mathbb{P}^{1/2}(\Omega_\rho^c). \tag{28}$$

Now we will use the following proposition:

Proposition 3. *Let $\rho > 1$. Then, under the assumptions of Theorem 1, there exists $C_2 > 0$ such that $P(\Omega_\rho^c) \le \frac{C_2}{n^{3/2}}$.*

This proposition implies that $\mathbb{E}\left(\|\pi\mathbb{1}_A - \hat{\pi}_{\hat{m}}\|_n^2 \mathbb{1}_{\Omega_\rho^c}\right) \le \frac{C_3}{n}$.

Now we use (26) and we observe that this inequality holds for all m in $\mathcal{M}_n$, so

$$\mathbb{E}\|\tilde{\pi} - \pi\mathbb{1}_A\|_n^2 \le C \inf_{m\in\mathcal{M}_n} (\|\pi\mathbb{1}_A - \pi_m\|^2 + \text{pen}(m)) + \frac{C_4}{n}$$

with $C = \max(5\|f\|_\infty, 6)$.

6.3. **Proof of Proposition 2.** Let $\Gamma_i(t) = \delta_i t(X_i, Y_i)/\bar{G}(Y_i) - \int t(X_i, y)\pi(X_i, y)dy$. Then $\nu_n(t) = (1/n)\sum_{i=1}^n \Gamma_i(t)$. We use the following lemma:

Lemma 2. *(Talagrand (1996))*
Let $U_1, \dots, U_n$ i.i.d. variables and $(\zeta_t)_{t\in B}$ a set of functions and B is a unit ball of a finite dimensional subspace of $\mathbb{L}^2(A)$. Let $\nu_n(t) = (1/n)\sum_{i=1}^n[\zeta_t(U_i) - \mathbb{E}(\zeta_t(U_i))]$. We suppose that
(1) $\sup_{t\in B}\|\zeta_t\|_\infty \le M_1$, (2) $\mathbb{E}(\sup_{t\in B}|\nu_n(t)|) \le H$, (3) $\sup_{t\in B}\text{Var}[\zeta_t(U_1)] \le v$.
Then, there exists $K > 0$, $K_1 > 0$, $K_2 > 0$ such that

$$\mathbb{E}\left[\sup_{t\in B} \nu_n^2(t) - 6H^2\right]_+ \le K\left[\frac{v}{n}e^{-K_1\frac{nH^2}{v}} + \frac{M_1^2}{n^2}e^{-K_2\frac{nH}{M_1}}\right]$$

Here $\zeta_t(x, y, \delta) = \delta t(x, y)/\bar{G}(y) - \int t(x, u)\pi(x, u)du$ and $B = B_f(m')$. We now compute M_1, H and v.

(1)We recall that $S_m + S_{m'}$ is included in the model $S_{m''}$ with dimension $\max(D_{m_1}, D_{m_1'})$ $\max(D_{m_2}, D_{m_2'})$.

$$\begin{aligned} \sup_{t\in B}\|\zeta_t\|_\infty &\le \sup_{t\in B}\|t\|_\infty(1/c_G + 1) \\ &\le \phi_0(1/c_G + 1)\sqrt{\max(D_{m_1}, D_{m_1'})\max(D_{m_2}, D_{m_2'})}\|t\| \le \frac{(1/c_G + 1)\phi_0}{f_0}n^{1/2}. \end{aligned}$$

Then we set $M_1 = \frac{(1/c_G + 1)\phi_0}{f_0}n^{1/2}$.

$$(2)\ \text{Var}[\zeta_t(U_1)] = \mathbb{E}\left([\Gamma_1(t)]^2\right) \le \mathbb{E}\left[\frac{t^2(X_1, Y_1)}{\bar{G}(Y_1)}\right] \le \frac{1}{c_G}\mathbb{E}\left[t^2(X_1, Y_1)\right] \le \frac{\|\pi\|_\infty}{c_G}\|t\|_f^2.$$

Then $v = \frac{\|\pi\|_\infty}{c_G}$.

(3) Let $(\bar\varphi_j\otimes\psi_k)_{(j,k)\in\Lambda(m,m')}$ be an orthonormal basis of $(S_m+S_{m'},\|.\|_f)$.

$$\begin{aligned}\mathbb{E}(\sup_{t\in B}|\nu_n^2(t)|) &\le \sum_{j,k}\mathbb{E}(\nu_n^2(\bar\varphi_j\otimes\psi_k)) = \sum_{j,k}\frac{1}{n^2}\mathbb{E}\left[\left(\sum_{i=1}^n\Gamma_i(\bar\varphi_j\otimes\psi_k)\right)^2\right]\\ &= \sum_{j,k}\frac{1}{n}\mathbb{E}\left(\Gamma_1^2(\bar\varphi_j\otimes\psi_k)\right).\end{aligned}$$

In the uncensored case, we find

$$\begin{aligned}\mathbb{E}(\sup_{t\in B}|\nu_n^2(t)|) &\le \sum_{j,k}\frac{1}{n}\mathbb{E}\left(\bar\varphi_j^2(X_1)\psi_k^2(Y_1)\right) \le \frac{\|\pi\|_\infty}{n}\sum_{j,k}\int\bar\varphi_j^2(x)\psi_k^2(y)f(x)dxdy\\ &\le \frac{\|\pi\|_\infty}{n}D(m,m').\end{aligned}$$

In the censored case, the assumptions are more restrictive because we need to use the norm connection on $S_m+S_{m'}$, that is we need $S_m+S_{m'}$ to be a space of the collection. Indeed, in the general case above, the bound would be $\|\pi\|_\infty D(m,m')/(nc_G)$ and c_G can not be estimated. Whereas under [M4], we find

$$\begin{aligned}\mathbb{E}(\sup_{t\in B}|\nu_n^2(t)|) &\le \sum_{j,k}\frac{1}{nf_0}\mathbb{E}\left(\frac{\delta_1\varphi_j^2(X_1)\psi_k^2(Z_1)}{\bar G^2(Z_1)}\right) \le \frac{1}{nf_0}\mathbb{E}\left(\frac{\delta_1\|\sum_{j,k}\varphi_j^2\psi_k^2\|_\infty}{\bar G^2(Z_1)}\right)\\ &\le \mathbb{E}\left(\frac{\delta_1}{\bar G^2(Z_1)}\right)\frac{\phi_0^2}{f_0}\frac{D(m,m')}{n}.\end{aligned}$$

Then $\mathbb{E}(\sup_{t\in B}\nu_n^2(t))\le AD(m,m')/n$ and $H^2=AD(m,m')/n$ with $A=\|\pi\|_\infty$ in the uncensored case or $A=(\phi_0^2/f_0)\mathbb{E}(\delta_1/\bar G^2(Z_1))$ in the censored case.

According to Lemma 2, there exists $K'>0$, $K_1>0$, $K_2'>0$ such that

$$\mathbb{E}\left[\sup_{t\in B_f(m')}\nu_n^2(t)-6H^2\right]_+\le K'\left[\frac{1}{n}e^{-K_1D(m,m')}+\frac{1}{n}e^{-K_2'\sqrt{D(m,m')}}\right].$$

So, if $p(m,m')=6H^2=6AD(m,m')/n$,

$$\sum_{m'\in\mathcal{M}_n}\mathbb{E}\left[\sup_{t\in B_f(m')}\nu_n^2(t)-p(m,m')\right]_+\le\frac{K'}{n}\left[\sum_{m'\in\mathcal{M}_n}(e^{-K_1D(m,m')}+e^{-K_2'\sqrt{D(m,m')}})\right]\le\frac{A_1}{n},$$

and the result follows. □

6.4. **Proof of Proposition 3.** First we observe that

$$P(\Omega_\rho^c)\le P\left(\sup_{t\in\mathcal{B}}|\nu_n(t^2)|>1-1/\rho\right)$$

where $\nu_n(t)=\frac{1}{n}\sum_{i=1}^n\int[t(X_i,y)-\mathbb{E}(t(X_i,y))]dy$ and $\mathcal{B}=\{t\in\mathcal{S},\|t\|_f=1\}$. We denote by (φ_j,ψ_k) the orthonormal basis of $\mathcal{S}$, the set of maximal dimension of the collection.

But, if $t(x,y)=\sum_{j,k} a_{j,k}\varphi_j(x)\psi_k(y)$, then

$$\nu_n(t^2)=\sum_{j,j'}\sum_k a_{j,k}a_{j',k}\bar\nu_n(\varphi_j\varphi_{j'})$$

where

$$\bar\nu_n(u)=\frac{1}{n}\sum_{i=1}^n[u(X_i)-\mathbb{E}(u(X_i))]. \tag{29}$$

Let $b_j=(\sum_k a_{j,k}^2)^{1/2}$, then $|\nu_n(t^2)|\le\sum_{j,j'} b_jb_{j'}|\bar\nu_n(\varphi_j\varphi_{j'})|$ and, if $t\in\mathcal{B}$, $\sum_j b_j^2=\sum_j\sum_k a_{j,k}^2=\|t\|^2\le f_0^{-1}$.

Thus

$$\sup_{t\in\mathcal{B}}|\nu_n(t^2)|\le f_0^{-1}\sup_{\sum b_j^2=1}\sum_{j,l} b_jb_l|\bar\nu_n(\varphi_j\varphi_l)|.$$

Lemma 3. *Let $B_{j,l}=\|\varphi_j\varphi_l\|_\infty$ and $V_{j,l}=\|\varphi_j\varphi_l\|_2$. Let, for any symmetric matrix $(A_{j,l})$*

$$\bar\rho(A)=\sup_{\sum a_j^2=1}\sum_{j,l}|a_ja_l|A_{j,l}$$

and $L(\varphi)=\max\{\bar\rho^2(V),\bar\rho(B)\}$. Then, if [M2] is satisfied, $L(\varphi)\le\phi_1\mathcal{D}_n^2$.

This lemma is proved in Baraud et al. (2001).

Let $x=\dfrac{f_0^2(1-1/\rho)^2}{4\|f\|_\infty L(\varphi)}$ and $\Delta=\left\{\forall j\forall l\quad|\bar\nu_n(\varphi_j\varphi_l)|\le 4\left[B_{j,l}x+V_{j,l}\sqrt{2\|f\|_\infty x}\right]\right\}$. On Δ:

$$\begin{aligned}\sup_{t\in\mathcal{B}}|\nu_n(t^2)| &\le f_0^{-1}\sup_{\sum b_j^2=1}\sum_{j,l}b_jb_l\left[B_{j,l}x+V_{j,l}\sqrt{2\|f\|_\infty x}\right]\\ &\le f_0^{-1}\left[\bar\rho(B)x+\bar\rho(V)\sqrt{2\|f\|_\infty x}\right]\\ &\le (1-1/\rho)\left[\frac{f_0(1-1/\rho)}{4\|f\|_\infty}\frac{\bar\rho(B)}{L(\varphi)}+\frac{1}{\sqrt2}\left(\frac{\bar\rho^2(V)}{L(\varphi)}\right)^{1/2}\right]\\ &\le (1-1/\rho)\left[\frac14+\frac{1}{\sqrt2}\right]\le(1-1/\rho).\end{aligned}$$

Then $P\left(\sup_{t\in\mathcal{B}}|\nu_n(t^2)|>1-\dfrac{1}{\rho}\right)\le P(\Delta^c)$.

To bound $P(\bar\nu_n(\varphi_j\varphi_l)\ge B_{j,l}x+V_{j,l}\sqrt{2\|f\|_\infty x})$, we will apply the Bernstein inequality given in Birgé and Massart (1998) to the independent variables $U_i^{j,l}=\varphi_j(X_i)\varphi_l(X_i)$. We get

$$P(|\bar\nu_n(\varphi_j\varphi_l)|\ge B_{j,l}x+V_{j,l}\sqrt{2\|f\|_\infty x})\le 2e^{-nx}.$$

Given that $P(\Omega_\rho^c)\le P(\Delta^c)=\sum_{j,l}P\left(|\bar\nu_n(\varphi_j\varphi_l)|>B_{j,l}x+V_{j,l}\sqrt{2\|f\|_\infty x}\right)$,

$$P(\Omega_\rho^c)\ \le\ 2\mathcal{D}_n^2\exp\left\{-\frac{nf_0^2(1-1/\rho)^2}{40\|f\|_\infty L(\varphi)}\right\}\le 2n\exp\left\{-\frac{f_0^2(1-1/\rho)^2}{40\|f\|_\infty}\frac{n}{L(\varphi)}\right\}.$$

But $L(\varphi) \leq \phi_1 \mathcal{D}_n^2 \leq \phi_1 n/\log^2(n)$, so

$$P(\Omega_\rho^c) \leq 2n \exp\left\{-\frac{f_0^2(1-1/\rho)^2}{40\|f\|_\infty \phi_1}\log^2(n)\right\} \leq \frac{C}{n^{3/2}}. \tag{30}$$

6.5. **Proof of Corollary 1.** To control the bias term, we state the following lemma proved in Lacour (2007) and following from Hochmuth (2002) and Nikol′skiĭ (1975):

Lemma 4. *Let π_A belong to $B_{2,\infty}^{\boldsymbol{\alpha}}(A)$. We consider that S'_m is one of the following spaces on A:*

- *a space of piecewise polynomials of degrees bounded by $s_i > \alpha_i - 1$ ($i = 1, 2$) based on a partition with rectangles of sidelengthes $1/D_{m_1}$ and $1/D_{m_2}$,*
- *a linear span of $\{\phi_\lambda \psi_\mu, \lambda \in \cup_0^{m_1}\Lambda(j), \mu \in \cup_0^{m_2} M(k)\}$ where $\{\phi_\lambda\}$ and $\{\psi_\mu\}$ are orthonormal wavelet bases of respective regularities $s_1 > \alpha_1 - 1$ and $s_2 > \alpha_2 - 1$ (here $D_{m_i} = 2^{m_i}, i = 1, 2$),*
- *the space of trigonometric polynomials with degree smaller than D_{m_1} in the first direction and smaller than D_{m_2} in the second direction.*

Let π'_m be the orthogonal projection of π_A on S'_m. Then, there exists a positive constant C_0 such that

$$\left(\int_A |\pi_A - \pi'_m|^2\right)^{1/2} \leq C_0[D_{m_1}^{-\alpha_1} + D_{m_2}^{-\alpha_2}].$$

If we choose for S'_m the set of the restrictions to A of the functions of S_m and π_A the restriction of π to A, we can apply Lemma 4. But π'_m is also the restriction to A of π_m so that

$$\|\pi \mathbb{1}_A - \pi_m\| \leq C_0[D_{m_1}^{-\alpha_1} + D_{m_2}^{-\alpha_2}].$$

According to Theorem 1

$$\mathbb{E}\|\tilde{\pi} - \pi\mathbb{1}_A\|_n^2 \leq C'' \inf_{m \in \mathcal{M}_n}\left\{D_{m_1}^{-2\alpha_1} + D_{m_2}^{-2\alpha_2} + \frac{D_{m_1}D_{m_2}}{n}\right\}.$$

In particular, if m^* is such that $D_{m_1^*} = \lfloor n^{\frac{\alpha_2}{\alpha_1+\alpha_2+2\alpha_1\alpha_2}}\rfloor$ and $D_{m_2^*} = \lfloor (D_{m_1^*})^{\frac{\alpha_1}{\alpha_2}}\rfloor$ then

$$\mathbb{E}\|\tilde{\pi} - \pi\mathbb{1}_A\|_n^2 \leq C''' \left\{D_{m_1^*}^{-2\alpha_1} + \frac{D_{m_1^*}^{1+\alpha_1/\alpha_2}}{n}\right\} = O\left(n^{-\frac{2\alpha_1\alpha_2}{\alpha_1+\alpha_2+2\alpha_1\alpha_2}}\right).$$

But the harmonic mean of α_1 and α_2 is $\bar{\alpha} = 2\alpha_1\alpha_2/(\alpha_1+\alpha_2)$. Then $\mathbb{E}\|\tilde{\pi} - \pi\mathbb{1}_A\|_n^2 = O(n^{-\frac{2\bar{\alpha}}{2\bar{\alpha}+2}})$.

The condition $D_{m_1} \leq n^{1/2}/\log(n)$ allows this choice of m only if $\frac{\alpha_2}{\alpha_1+\alpha_2+2\alpha_1\alpha_2} < \frac{1}{2}$ i.e. if $\alpha_1 - \alpha_2 + 2\alpha_1\alpha_2 > 0$. In the same manner, the condition $\alpha_2 - \alpha_1 + 2\alpha_1\alpha_2 > 0$ must be verified. Both conditions hold if $\alpha_1 > 1/2$ and $\alpha_2 > 1/2$.

References

Azzalini, A. and Bowman, A. (1990). A look at some data on the old faithful geyser. *Applied Statistics*, 39(3):357–365.

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

Barron, A., Birgé, L., and Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413.

Birgé, L. and Massart, P. (1997). From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam*, pages 55–87. Springer, New York.

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

Bitouzé, D., Laurent, B., and Massart, P. (1999). A Dvoretzky-Kiefer-Wolfowitz type inequality for the Kaplan-Meier estimator. *Ann. Inst. H. Poincaré Probab. Statist.*, 35(6):735–763.

Brunel, E. and Comte, F. (2005). Penalized contrast estimation of density and hazard rate with censored data. *Sankhyā*, 67(3):441–475.

Comte, F. (2001). Adaptive estimation of the spectrum of a stationary Gaussian sequence. *Bernoulli*, 7(2):267–298.

De Gooijer, J. G. and Zerom, D. (2003). On conditional density estimation. *Statist. Neerlandica*, 57(2):159–176.

Fan, J. and Gijbels, I. (1994). Censored regression: local linear approximations and their applications. *J. Amer. Statist. Assoc.*, 89(426):560–570.

Fan, J., Yao, Q., and Tong, H. (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika*, 83(1):189–206.

Hochmuth, R. (2002). Wavelet characterizations for anisotropic Besov spaces. *Appl. Comput. Harmon. Anal.*, 12(2):179–208.

Hyndman, R. J. and Yao, Q. (2002). Nonparametric estimation and symmetry tests for conditional density functions. *J. Nonparametr. Stat.*, 14(3):259–278.

Lacour, C. (2007). Adaptive estimation of the transition density of a markov chain. *Ann. Inst. H. Poincaré Probab. Statist.*, 43(to appear):available online.

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(3):461–473.

Nikol′skiĭ, S. M. (1975). *Approximation of functions of several variables and imbedding theorems*. Springer-Verlag, New York. Translated from the Russian by John M. Danskin, Jr., Die Grundlehren der Mathematischen Wissenschaften, Band 205.

Talagrand, M. (1996). New concentration inequalities in product spaces. *Invent. Math.*, 126(3):505–563.

Van Keilegom, I. and Veraverbeke, N. (2002). Density and hazard estimation in censored regression models. *Bernoulli*, 8(5):607–625.