Nonparametric warped kernel estimators

Gaëlle Chagny

NONPARAMETRIC WARPED KERNEL ESTIMATORS

GAËLLE CHAGNY[A] *

Abstract. In this work, we propose a general method of adaptive nonparametric estimation, based on warped kernels. The aim is to estimate a real-valued function s from a sample of random variables (X, Y). We first deal with the auxiliary function $g = s \circ \phi_X$, for a bijective map ϕ_X depending on the distribution of the variable X: we consider a collection of kernel estimates built with the warped data $(\phi_X(X), Y)$. The data-driven selection of the best bandwidth is done with a model selection device in the spirit of Goldenshluger and Lepski (2011). This leads to an estimator $\hat{g}$ for the function g, which is then warped to estimate the target function s by $\hat{s} = \hat{g} \circ \hat{\phi}_X$ where $\hat{\phi}_X$ is an estimate for ϕ_X. The interest is twofold. From the practical point of view, the estimator can be computed easily and fastly, thanks to its simple explicit expression. From the theoretical point of view, the squared-bias/variance trade-off is realized: we derive non-asymptotic risk bounds. This general method permits to handle various problems such as additive and multiplicative regression, conditional density estimation, hazard rate estimation based on randomly right censored data, and cumulative distribution function estimation from current-status data.

Keywords: Adaptive estimator. Censored data. Bandwidth selection. Nonparametric estimation. Regression. Warped kernel.

AMS Subject Classification 2010: 62G05; 62G08; 62N02.

July 2012

1. Introduction

1.1. **Motivation.** Additive regression is one of the most studied model in nonparametric estimation. A huge variety of methods have been investigated, since the first kernel strategies initiated by Nadaraya (1964) and Watson (1964). Powerful techniques now enable to build estimators, which have adaptive properties in the sense that their Mean Integrated Squared Error (M.I.S.E.) automatically reaches the best possible rate associated with the unknown underlying smoothness of the regression function.

Moreover, the estimators built in this framework, from kernel to least-squares, are source of inspiration for several other functional estimation problems, such as multiplicative regression, conditional density estimation, hazard rate estimation based on randomly right censored data, and cumulative distribution function estimation from current-status data: the question of building estimators in the spirit of reweighted Nadaraya-Watson functions, or based on the minimization of regression-type criteria has received a lot of attention in the past decades.

The goal of this article is to propose a unified approach for functional estimation, which enjoys good adaptive theoretical performances, low computational complexity, and which permits to cover simultaneously all the aforementioned estimation problems. Therefore, the framework is

* Corresponding author. Email: gaelle.chagny@parisdescartes.fr

[A] *Laboratoire MAP5 (UMR CNRS 8145), Université Paris Descartes, Sorbonne Paris Cité, France.*

very general: we aim at recovering a real-valued function s on a borelian subset A of $\mathbb{R}$ or $\mathbb{R}^2$, from a data sample of observations distributed like a couple of real random variables (X, Y).

The main basic idea is to investigate thoroughly what we will call a "warping" method, following examples introduced successively by Yang (1981), Stute (1984, 1986) and more recently Kerkyacharian and Picard (2004): we first build a collection of warped-kernel estimators. Then, we address the problem of bandwidth selection by taking into account the recent Goldenshluger-Lepski method (shortened by the "GL" method from now on), detailed in Goldenshluger and Lepski (2011): we provide a totally data-driven procedure.

1.2. **Examples covered by the general framework.** To be more precise, we provide several examples illustrating the relevance of the general setting: we detail the couple (X, Y) and the target function s which can be handled with the warped-kernel strategy. We also mention nonparametric methods already studied in related problem. As we cannot reasonably make a review of all nonparametric estimation, we essentially focus on adaptive methods, in the sense that they do not require prior knowledge about the smoothness of the estimated function.

Hereafter, we assume that the variable X takes values in A, while we denote by B the analogous set for Y. Besides, X is supposed to have a density f_X with respect to the Lebesgue measure.

We first present two classical regression frameworks, before introducing two estimation problems based on lifetimes that can only be partly observed.

In the first two examples, we assume that A is an interval: $A = (a; b)$, with $-\infty \leq a < b \leq \infty$, and that the density f_X does not vanish on its support A.

Example E1: Additive random design regression. The function s to be estimated is the conditional expectation of Y given a value for X. Thus, we can write $Y = s(X) + \varepsilon$, where ε is a centered, square-integrable random variable, independent of X.

Example E2: Multiplicative regression. Here, we write $Y = \sigma(X)\varepsilon$, and the target function s is the volatility function σ^2. The assumptions are the following: ε is a centered random variable, independent of X, which satisfies $\mathbb{E}[\varepsilon^2] = 1$ and $\mathbb{E}[\varepsilon^4] < \infty$.

These estimation frameworks, mainly the additive one, are a famous subject of interest, and adaptive estimation is well developed. Autoregressive models are the setting the most related to Example E2. Historical methods are kernel strategies, initiated by Nadaraya (1964) and Watson (1964) (Example E1). Recall that their estimator is built as the ratio of a kernel estimator of the product sf_X divided by a kernel estimator of the density f_X. The data-driven choice of the bandwidth, leading to adaptive estimators, is for example studied more accurately by Fan and Gijbels (1992) (Example E1), Härdle and Tsybakov (1997) (Examples E1 and E2), and Neumann (1994) (Example E2) who provide asymptotic results, for methods also sometimes involving local polynomials.

At the same time, the expansion of s onto orthogonal bases has been used: among many authors, we can quote Golubev and Nussbaum (1992) for spline bases, Antoniadis et al. (1997) (Example E1) and Hoffmann (1999) (Example E2) for wavelet bases, and Efromovich (1999) (Fourier basis). Nonasymptotic risk bounds are first derived for this kind of estimates, usually by model selection, via the minimization of a penalized least-squares criterion (see Wegkamp 2003, Baraud 2002, and Birgé 2004 for Example E1, and Comte and Rozenholc 2002 for Example E2).

Among the aforementioned works, the ratio form of kernel estimates may be seen as a drawback, since it can be instable when a "hole" occurs in the data. Moreover, two bandwidths must

be selected, one for the numerator and one for the denominator (from the theoretical point of view, there are no reason to choose the same bandwidths). The recent Goldenshluger-Lepski method has only been investigated recently by Chichignoud (2011a,b), for specific frameworks (uniform design in additive regression, deterministic design and uniform noise in multiplicative regression). Moreover, least-squares contrasts provide explicit estimators under a matrix invertibility requirement (most of the time implicitly).

This motivates the investigation of "warping" methods to build estimators which satisfy powerful nonasymptotic risk bounds while being simple to implement.

For the third and fourth examples, related to reliability and survival analysis, X is a lifetime, and thus, A is equal to $\mathbb{R}_+$. Let also Z be a positive random variable of interest, which is unobserved.

Example E3: Interval censoring, Case 1. In this case, the couple (X,Y) is known as current-status data. The only knowledge about the survival time of interest Z is its current status at time of examination X. Thus, $Y = \mathbf{1}_{Z\leq X}$ indicates whether Z occurs before X or not. Such data naturally arise in infectious disease study for example, when the time Z of infection is unobserved, and a test is carried out at time X. The function s of interest is the cumulative distribution function of Z: $s = F_Z$. Hereafter, we assume that f_X is positive on its support $A = \mathbb{R}_+$.

In this model, not many investigations are concerned with adaptivity, since it is the unusual optimal rate of convergence ($n^{-1/3}$) that has diverted attention for the last decades. Thus, most results are devoted to the Non-Parametric Maximum Likelihood Estimator (NPMLE) (see Groeneboom and Wellner 1992; van de Geer 1993; Groeneboom 1995; Hudgens et al. 2007, for intance). Birgé (1999) built an histogram estimator, which is very simple but not adaptive. A review is provided by Jewell and van der Laan (2004). Adaptivity has been more recently discussed by Ma and Kosorok (2006), who selected the regularity parameter of the NPMLE and of a least-squares estimator, and by Brunel and Comte (2009) or Plancade (2011), who considered model selection devices to choose regression-type estimators which also require a matrix inversion.

Example E4: Hazard rate estimation from right censored-data. Here, X is the minimum of the variable of interest Z and of a censoring time C, which is a nonnegative random variable (like Z), supposed to be independent of the lifetime Z. We also know whether Z is censored or not. Then, we have $X = C \wedge Z$, and $Y = \mathbf{1}_{Z\leq C}$. Right censoring occurs when individuals, included in a clinical trial, are not observed until the end, for instance. The function s of interest is the hazard rate function:

$$s(x) = \frac{f_Z(x)}{1 - F_Z(x)}$$

which is the risk of death at time x, given that the patient is alive until x (also, this is the derivative of the log-survival function). We assume both that $F_Z(x) < 1$ and $F_C(x) < 1$ for all $x \in A = \mathbb{R}_+$.

We reasonably cannot quote all the estimators and results provided by the previous studies. Let us only recall that a lot of estimators use the well-known Kaplan-Meier estimate of the survival function (Kaplan and Meier, 1958), some others are based on the estimation of the cumulative hazard with the Nelson-Aalen function (Nelson, 1972), and others use a different decomposition of s, using the "subdensity" function. Different methods are proposed for each approach: kernel methods, with asymptotic results (Tanner and Wong, 1983; Müller and Wang, 1994; Patil, 1993), orthogonal-serie decomposition and model selection, leading both to

nonasymptotic and asymptotic risk bounds (Antoniadis et al., 1999; Brunel and Comte, 2005, 2008; Reynaud-Bouret, 2006; Akakpo and Durot, 2010).

A summary of these four frameworks can be found in Table 1. Finally, we extend at the end of the paper the method to a last example: the estimation of a conditional density, an example of bivariate functional estimation problem. The specifical setting and references will be discussed later.

1.3. **The "warping" method.** Let us present the sketch of the method briefly. Our goal is to estimate the real-valued function s on the set A, from n couples of observations $(X_i, Y_i)_{i\in\{1,\dots,n\}}$, distributed like (X, Y) according to Examples E1 to E4 previously described. We aim at providing adaptive kernel estimators in each of the previous framework, with simple expression.

To do so, we first consider an auxiliary function g defined by

$$g = s \circ \phi_X^{-1} \,:\, \phi_X(A) \to \mathbb{R}, \tag{1}$$

where ϕ_X^{-1} is the inverse of a function ϕ_X, which is called the "warping" function, and depends on the example:

- for Examples E1, E2, and E3, ϕ_X is the cumulative distribution function (c.d.f.) of the variable X. Since we assume that the density f_X is strictly nonnegative on A, ϕ_X admits an inverse and $\phi_X(A) = (0; 1)$.
- for Example E4, we consider for ϕ_X a function denoted by ϕ, which is a primitive of the survival function $1 - F_X$, that is to say $\phi : x \mapsto \int_0^x (1 - F_X(t))dt$. We also assume $F_X(t) < 1$ for all $t \geq 0$, thus ϕ admits an inverse too and $\phi_X(A) = \mathbb{R}_+$.

These choices will be justified below, see Identity (3).

In a first step, we will build a collection of classical kernel estimators for the auxiliary function g. In a second step, we will select $\hat{g}$, one of these kernel estimators, using the recent GL method, developed for density estimation initially (see Goldenshluger and Lepski 2011). Finally, according to Definition (1), we will define an estimator for the target s by

$$\hat{s} = \hat{g} \circ \phi_X \quad \text{or} \quad \hat{s} = \hat{g} \circ \hat{\phi}_X,$$

where $\hat{\phi}_X$ is an empirical counterpart for ϕ_X, since ϕ_X is unknown, in general.

This warping device, which we sum up in Table 1, has already been used in the regression framework (namely Example E1) by Stute (1984) who studied more accuratly the kernel estimator of Yang (1981), and more recently by Kerkyacharian and Picard (2004) and Pham Ngoc (2009), who adapted this to projection estimation.

The novelty of our contribution lies in the combination of the warping strategy and the GL method, to deal with several estimation settings simultaneously: our building estimators allows us to derive nonasymptotic adaptive results. Oracle-type inequalities are provided for the M.I.S.E., and convergence rates are deduced under regularity assumptions on the function g. Moreover, the simple expression of the estimates enables us to implement them easily, and to illustrate the theoretical results with promising simulation experiments.

Hereafter, for the sake of clarity, we focus on the case of known c.d.f. F_X, like in Pham Ngoc (2009): the case of an unknown c.d.f. F_X requires further technicalities, only due to the natural plug-in of an empirical version for ϕ_X, even if the theoretical results are similar. It has also been widely detailed in Chagny (2011, 2012), for warped-bases estimators of a regression function and of a conditional density respectively. Therefore, in the sequel, we prefer to concentrate on the wide range of examples which are covered by the method, from classical regression frameworks

Example	Target function	ϕ_X
E1 $Y = s(X) + \varepsilon$	s	F_X
E2 $Y = \sigma(X)\varepsilon$	σ^2	F_X
E3 $(X, \mathbf{1}_{Z \leq X})$	F_Z	F_X
E4 $(X = Z \wedge C, \mathbf{1}_{Z \leq C})$	$\frac{f_Z}{1-F_Z}$	$\phi : x \mapsto \int_0^x (1 - F_X(t))dt$

TABLE 1. Summary of the studied examples and of the "warping" function used in each case.

to bivariate functional estimation (conditional density), getting also through survival analysis problems.

1.4. **Organization of the paper.** We present in Section 2 the notations, the collection of kernel estimators and the data-driven bandwidth selection leading to a unique estimator. In Section 3, we investigate the performance of this estimator: we study its global risk, state our main results and comment them. Section 4 is devoted to a short simulation study, to illustrate the method in the Examples E1 and E4. We provide in Section 5 an extension of the method to conditional density estimation. Finally, the proofs are gathered in Section 6.

2. A general method of estimation

2.1. **Notations.** Throughout the article, we consider functions which are integrable with respect to the Lebesgue measure or to a weighted Lebesgue measure. For $0 < p \leq \infty$, we denote by $L^p(B)$ the set of the real-valued and measurable functions t on a borelian subset $B \subset \mathbb{R}$, such that the (quasi-)norm

$$\|t\|_{L^p(B)} = \begin{cases} \int_B |t(u)|^p du \text{ if } 0 < p \leq \infty \\ \sup_{u \in B} |t(u)| \text{ if } p = \infty \end{cases}$$

is finite. If $p = 2$, $\langle ., . \rangle_B$ is the usual scalar product of the Hilbert space $L^2(B)$. The following L^2- norm will also be useful, since it is the natural loss function of the problem:

$$\|t\|_{\phi'_X} = \int_A t^2(x)\phi'_X(x)dx, \tag{2}$$

and $L^2(A, \phi'_X)$ is the space of functions t for which the quantity (2) exists and is finite. This norm leads to another corresponding scalar product $\langle ., . \rangle_{\phi'_X}$. Notice besides that the following links hold between this space and the classical L^2-space previously defined: if t_1, t_2 belongs to $L^2(A, \phi'_X)$, we compute, using $F'_X = f_X$,

$$\|t_1 \circ \phi_X\|_{\phi'_X} = \|t_1\|_{L^2(\phi_X(A))}, \quad \langle t_1 \circ \phi_X, t_2 \circ \phi_X \rangle_{\phi'_X} = \langle t_1, t_2 \rangle_{\phi_X(A)}.$$

The convolution product of two functions t_1 and t_2 is $t_1 \star t_2 : x \mapsto \int_{\mathbb{R}} t_1(x - x')t_2(x')dx'$. Last, the notation x_+, for a real number x, means $\max(x, 0)$.

Hereafter, K is a kernel, that is a function such that $\int_{\mathbb{R}} K(u)du = 1$, and is assumed to belong to $L^2(\mathbb{R})$. We also denote by $\mathcal{H}_n$ a finite collection of nonnegative real numbers, the so-called bandwidths. Its cardinality may depend on the sample size n. Classically, for each $h \in \mathcal{H}_n$, K_h is the function $u \mapsto K(u/h)/h$. We easily get $\int_{\mathbb{R}} K_h(u)du = 1$, $\|K_h\|_{L^1(\mathbb{R})} = \|K\|_{L^1(\mathbb{R})}$, and finally $\|K_h\|_{L^2(\mathbb{R})} = \|K\|_{L^2(\mathbb{R})}/h$.

2.2. **Collection of warped kernel estimators.** Throughout the section, we fix a bandwidth $h \in \mathcal{H}_n$. We first deal with the transformed data $(\phi_X(X_i), Y_i)_{i\in\{1,\dots,n\}}$, to estimate the auxiliary function g defined by (1). The cornerstone of the method is that for all $u \in \phi_X(A)$,

$$\mathbb{E}\left[\theta(Y)K_h(u-\phi_X(X)]\right] = K_h \star \left(g\mathbf{1}_{\phi_X(A)}\right)(u), \quad \text{with } \theta(Y) = \begin{cases} Y \text{ for Examples E1, E3, and E4,} \\ Y^2 \text{ for Example E2.} \end{cases} \tag{3}$$

This identity explains the choices of the warping function ϕ_X, introduced in Section 1.3. For instance, let us prove it in Example E4. The proof of Equality (3) for the other cases is postponed to Section 6.1. We start with

$$\begin{aligned}
\mathbb{E}\left[YK_h(u-\phi_X(X))\right] &= \mathbb{E}\left[\mathbf{1}_{Z\le C}K_h(u-\phi(Z\wedge C)\right] = \mathbb{E}\left[\mathbf{1}_{Z\le C}K_h(u-\phi(Z)\right], \\
&= \int_{\mathbb{R}_+\times\mathbb{R}_+} \mathbf{1}_{z\le c}K_h\left(u-\phi(z)\right) f_C(c)f_Z(z)dzdc, \\
&= \int_{\mathbb{R}_+} K_h\left(u-\phi(z)\right) f_Z(z)\left(\int_{\mathbb{R}} \mathbf{1}_{x\le c}f_C(c)dc\right)dz, \\
&= \int_{\mathbb{R}_+} K_h\left(u-\phi(z)\right) f_Z(z)\left(1-F_C(z)\right)dz.
\end{aligned}$$

Then, we set $u' = \phi(z)$. The integral becomes

$$\mathbb{E}\left[YK_h(u-\phi_X(X))\right] = \int_{\mathbb{R}_+} K_h\left(u-u'\right) f_Z\circ\phi^{-1}(u')\left(1-F_C\right)\circ\phi^{-1}(u')\frac{du'}{(1-F_X)\circ\phi^{-1}(u')}.$$

Since Z and C are independent, $(1-F_X) = (1-F_C)(1-F_Z)$, and consequently

$$\begin{aligned}
\mathbb{E}\left[YK_h(u-\phi_X(X))\right] &= \int_{\mathbb{R}_+} K_h\left(u-u'\right) f_Z\circ\phi^{-1}(u')\frac{du'}{(1-F_Z)\circ\phi^{-1}(u')}, \\
&= \int_{\mathbb{R}_+} K_h\left(u-u'\right) s\circ\phi^{-1}(u')du', \\
&= \int_{\mathbb{R}_+} K_h\left(u-u'\right) g(u')du', \\
&= K_h \star \left(g\mathbf{1}_{\mathbb{R}_+}\right)(u).
\end{aligned}$$

A consequence of Equality (3) is that we define a natural estimator for g by

$$\forall u \in \phi_X(A),\ \ \hat{g}_h(u) = \frac{1}{n}\sum_{i=1}^{n} \theta(Y_i)K_h\left(u-\phi_X(X_i)\right).$$

Since the target function s can be written as $s = g\circ\phi_X$, we also set $\hat{s}_h = \hat{g}_h\circ\phi_X$. At this stage of the procedure, the interest lies in the simple expression of the estimators $\hat{s}_h$, $h \in \mathcal{H}_n$: it involves no ratio, one kernel only, and thus only one bandwidth to select.

2.3. **Bandwidth automatic-selection.** A collection of estimators $(\hat{s}_h)_{h\in\mathcal{H}_n}$ is available now, and classically, the next question is the choice of the bandwidth. As well as being data-driven, the selection should lead to an adaptive estimator: thus our problem is to build a statistical procedure that requires no prior knowledge on s but whose risk behaves almost like the minimum of the risk of the estimators in the collection, that is to say almost as the oracle bandwidth

$$\tilde{h} := \arg\min_{h\in\mathcal{H}_n} \mathbb{E}[\|s-\hat{s}_h\|^2_{\phi'_X}]. \tag{4}$$

In fact, the quadratic risk weighted by the derivative of the warping function ϕ_X is the natural criterion in our setting. Therefore, it requires that the function s belongs to $L^2(A,\phi'_X)$, and we assume it from now on:

- for Examples E1 to E3, $\phi'_X = f_X$, and this condition is fulfilled as soon as s is bounded on the set A,
- for Example E4, where ϕ'_X is the survival function of the variable Z, we can check that the integrability condition on s is verified for all classical distributions for C and Z in survival analysis (such as exponential, Weibull, Gamma...).

In order to explain what could be a "good" selection, we evaluate the performance of $\hat{s}_h$ for each h, by giving an upper-bound for its weighted risk, that is for the classical quadratic risk of g: $\mathbb{E}[\|\hat{s}_h - s\|^2_{\phi'_X}] = \mathbb{E}[\|\hat{g}_h - g\|^2_{L^2(\phi_X(A))}]$. For that purpose, we introduce the approximation of g by the kernel K_h, $g_h = K_h \star (g\mathbf{1}_{\phi_X(A)})$. It is first well known that

$$\mathbb{E}\left[\|\hat{s}_h - s\|^2_{\phi'_X}\right] = \|g-g_h\|^2_{L^2(\phi_X(A))} + \mathbb{E}\left[\|g_h-\hat{g}_h\|^2_{L^2(\phi_X(A))}\right], \tag{5}$$

since $\mathbb{E}[\hat{g}_h(u)] = g_h(u)$, thanks to (3). If the first term in the right-hand side of (5) decreases when h goes to zero, the opposite holds for the second term: in fact, we bound it as follows:

$$\mathbb{E}\left[\|g_h-\hat{g}_h\|^2_{L^2(\phi_X(A))}\right] = \mathbb{E}\left[\int_{\phi_X(A)} \left(\hat{g}_h(u) - \mathbb{E}\left[\hat{g}_h(u)\right]\right)^2 du\right] = \int_{\phi_X(A)} \mathrm{Var}\left(\hat{g}_h(u)\right) du,$$

and for each $u \in \phi_X(A)$,

$$\mathrm{Var}\left(\hat{g}_h(u)\right) = \frac{1}{n}\mathrm{Var}\left(\theta(Y_1)K_h\left(u-\phi_X(X_1)\right)\right) \leq \frac{1}{n}\mathbb{E}\left[\theta^2(Y_1)K_h^2\left(u-\phi_X(X_1)\right)\right].$$

Therefore, the variance-term grows when h decreases:

$$\mathbb{E}\left[\|g_h-\hat{g}_h\|^2_{L^2(\phi_X(A))}\right] \leq \mathbb{E}\left[\theta^2(Y_1)\right]\|K\|^2_{L^2(\mathbb{R})}\frac{1}{nh}.$$

Thus, we recover that choosing a bandwidth h which realizes a good compromise between the approximation term and the estimation (or variance) term leads to an estimator with small risk.

This aim can be achieved with the observed data only, with a method described in Goldenshluger and Lepski (2011). The idea is the following: if the bias and the variance term are unknown (since they depend on the unknown s), we replace them by empirical versions. We define first

$$\forall h\in\mathcal{H}_n,\ \ V(h) = \kappa\left(1+\|K\|^2_{L^1(\mathbb{R})}\right)\|K\|^2_{L^2(\mathbb{R})}\mathbb{E}\left[\theta^2(Y_1)\right]\frac{1}{nh}, \tag{6}$$

which corresponds to the upper-bound for the variance term. The constant κ is purely numerical, and its value will be specified in the proofs. Then, with a remark already used by Devroye

Example	s	ϕ_X	$\hat{s}(x)$
E1 $Y = s(X) + \varepsilon$	s	F_X	$\frac{1}{n}\sum_{i=1}^{n} Y_i K_{\hat{h}}(F_X(x) - F_X(X_i))$
E2 $Y = \sigma(X)\varepsilon$	σ^2	F_X	$\frac{1}{n}\sum_{i=1}^{n} Y_i^2 K_{\hat{h}}(F_X(x) - F_X(X_i))$
E3 $(X, \mathbf{1}_{Z\leq X})$	F_Z	F_X	$\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{Z_i\leq X_i} K_{\hat{h}}(F_X(x) - F_X(X_i))$
E4 $(X = Z \wedge C, \mathbf{1}_{Z\leq C})$	$\frac{f_Z}{1-F_Z}$	ϕ	$\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{Z_i\leq C_i} K_{\hat{h}}(\phi(x) - \phi(X_i))$

TABLE 2. Summary of the estimators in the four studied statistical examples, described in Section 1.2

(1989), we introduce the auxiliary estimators involving two kernels: $\hat{g}_{h,h'} = K_{h'} \star (\hat{g}_h \mathbf{1}_{\phi_X(A)})$, and, accordingly, $\hat{s}_{h,h'} = \hat{g}_{h,h'} \circ \phi_X$. We set

$$\forall h \in \mathcal{H}_n, \quad A(h) = \max_{h' \in \mathcal{H}_n} \left\{ \|\hat{s}_{h,h'} - \hat{s}_{h'}\|^2_{\phi'_X} - V(h') \right\}_+ . \tag{7}$$

It is shown in the proof that A has the same order as the bias-term (see Lemma 6). Then the selected bandwidth $\hat{h}$ and the corresponding warped kernel estimator are

$$\hat{h} = \arg\min_{h \in \mathcal{H}_n} \{A(h) + V(h)\}, \quad \hat{s} = \hat{s}_{\hat{h}}. \tag{8}$$

The formula of the estimators corresponding to each considered example are summarized in Table 2. Let us highlight the fact that the selected bandwidth $\hat{h}$ does not depend on the function s to be estimated: it is totally data-driven. Actually, in Examples E3 and E4, $\mathbb{E}[\theta^2(Y_1)]$ is bounded by 1, and can be replaced by 1 in the definition of V. For the two other examples (additive and multiplicative regression), this expectation can easily be replaced in practice and theory by the corresponding empirical mean (see Brunel and Comte 2005, proof of Theorem 3.4 p.465).

3. Theoretical results

3.1. **Assumptions and smoothness classes.** Now we are in position to state the result concerning the adaptive estimators built in the four examples.

To set nonasymptotic risk bound, we require only one or two assumptions, depending on the example we consider: one about the bandwidth collection, which should not be too large and one which is concerned with the distribution of the errors in the two regression settings (Examples E1 and E2).

Assumption (B_{α_0})**:**

(i) For any constant $\kappa_0 > 0$, there exists $C_0 > 0$, such that $\sum_{h\in\mathcal{H}_n} \exp\left(-\frac{\kappa_0}{h}\right) \leq C_0(\kappa_0)$,

(ii) There exists $\alpha_0 > 0$ such that $\sum_{h\in\mathcal{H}_n} \frac{1}{h} \leq k_0 n^{\alpha_0}$, for a constant $k_0 \geq 0$.

Assumption (M_p)**:** With α_0 fixed by Assumption (B_{α_0}),

(i) There exists $p > 2\alpha_0$, such that $\mathbb{E}[|\varepsilon_1|^{2+p}] < \infty$,

(ii) There exists $p > 4\alpha_0$, such that $\mathbb{E}[|\varepsilon_1|^{4+p}] < \infty$.

Remark 1. Classical collections of bandwidths satisfy Assumption (B_{α_0}). For instance:

(1) $\mathcal{H}_{n,1} = \left\{k^{-1},\ k = 1, \ldots, \varphi(n)\right\}$, for which Assumption (B_2) is fulfilled if $\varphi(n) = n$, or (B_1) if $\varphi(n) = \sqrt{n}$.

(2) $\mathcal{H}_{n,2} = \left\{2^{-k},\ k = 1, \ldots, [\ln(n)/\ln(2)]\right\}$, for which Assumption (B_1) is fulfilled.

Notice also that the smaller α_0 in Assumption (B_{α_0}), the less restrictive the integrability constraint p on the noise moments in Assumption (M_p).

To deduce rates of convergence from the nonasymptotic results, we will require the following additional assumption.

Assumption (K_l)**:** The kernel K is of order l, that is to say, for all $j \in \{1, \ldots, l+1\}$, the function $x \mapsto x^j K(x)$ is integrable, and for $1 \leq j \leq l$, $\int_{\mathbb{R}} x^j K(x)dx = 0$.

Moreover, for convergence rates, smoothness classes must be defined to quantify the bias term of the decomposition (5): $\|g_h - g\|^2$. For estimation in Examples E1 to E3, we consider functions t belonging to Hölder classes on an interval B, denoted by $\mathcal{H}(\beta, L, B)$, $\beta, L > 0$: this means that t admits derivatives up to order $[\beta]$ (where $[\beta]$ is the largest integer less than β), and

$$\forall x, x' \in B,\ \left|t^{[\beta]}(x) - t^{[\beta]}(x')\right| \leq L|x - x'|^{\beta-[\beta]}. \tag{9}$$

This property is relevant for the bound on the integrated bias on compact sets, but not on $\mathbb{R}_+$ as required for Example E4. The functional spaces associated to this case are Nikol'skii classes of functions, $\mathcal{N}_2(\beta, L)$: a function $t : \mathbb{R} \mapsto \mathbb{R}$ belongs to $\mathcal{N}_2(\beta, L)$, if its admits derivatives up to order $[\beta]$ and

$$\forall x \in \mathbb{R},\ \left\|\tau_x t^{[\beta]} - t^{[\beta]}\right\|_{L^2(\mathbb{R})} \leq L|x|^{\beta-[\beta]},$$

where τ_x is the translation operator by x. Both of these spaces are standard in kernel estimation: see Tsybakov (2009), Goldenshluger and Lepski (2011), and also Nikol′skiĭ (1975) for instance.

3.2. **Risk bounds.** We can prove the following result.

Theorem 1. *Assumption (B_{α_0}) is supposed to be fulfilled, for an $\alpha_0 > 0$ and also Assumption (M_p) for Examples E1 and E2. We also assume that the function s is bounded on the set A.*

Then there exist three constants c_l, $l = 1, 2, 3$, such that the following inequality holds for the estimator $\hat{s}$ defined by (8):

$$\mathbb{E}\left[\|\hat{s} - s\|^2_{\phi'_X}\right] \leq \min_{h\in\mathcal{H}_n}\left\{c_1\|s - s_h\|^2_{\phi'_X} + c_2\frac{\mathbb{E}\left[\theta^2(Y_1)\right]\|K\|^2_{L^2(\mathbb{R})}}{nh}\right\} + \frac{c_3}{n}. \tag{10}$$

The two constants c_1 and c_2 depend only on $\|K\|_{L^1(\mathbb{R})}$, and c_3 depends on $\|s\|_{L^\infty(A)}$, $\|K\|_{L^1(\mathbb{R})}$ and $\|K\|_{L^2(\mathbb{R})}$ in Examples E3 and E4, and also on $\mathbb{E}[\varepsilon_1^2]$, $\mathbb{E}[\varepsilon_1^{2+p}]$ and $\mathbb{E}[s^2(X_1)]$ for Example E1 or on $\mathbb{E}[\varepsilon_1^4]$, $\mathbb{E}[\varepsilon_1^{4+p}]$ and $\mathbb{E}[\sigma^4(X_1)]$ for Example E2.

Let us comment and discuss the result.

- **About the meaning of Inequality (10).** It is an oracle-type inequality: the selected bandwidth $\hat{h}$ is performing as well as the unknown oracle (4), up to some multiplicative constants c_1 and c_2, and up to a remainding term of order $1/n$, which is negligible. Actually, it follows from (10) that the adaptive estimators $\hat{s}$, in Examples E1 to E4, automatically make the squared-bias/variance compromise.
- **About the assumptions of Theorem 1.** The result holds for any sample size n and thus, is nonasymptotic. There is no assumption on the approximation properties of the kernel K, that is no assumption on its regularity and moments, contrary to most of asymptotic results for kernel estimators. This is the strength of the GL method, in the simple way we apply it. Especially in Example E1 (additive regression), the risk bound can thus be considered as an improvement of the results of Stute (1986), who provides asymptotic normality for a warped kernel estimate, with well-chosen kernel and bandsequence (but not adaptive).
- **About the case of unknown ϕ_X.** The previous method is applied in the general framework of unknown ϕ_X (that is to say the case of unknown F_X) by using a natural plug-in device: the c.d.f. F_X can be replaced by its empirical counterpart in all occurrences. Obviously, the adaptive result is the same, under stronger assumptions on the bandwidth collection $\mathcal{H}_n$. However the proof in this case requires much more technicalities than it may seem. Therefore we focus on the theoretical case of known F_X, to concentrate on the wide range of examples that the method covers. The substitution has already been widely detailed for regression estimation and conditional density estimation using warped bases: we refer the reader to Chagny (2011, 2012).

The "oracle-approach" also leads to convergence rate for the risk, under regularity assumptions for the auxiliary function g.

Corollary 1. *Let β and L be two nonnegative numbers. Assume that the function g satisfies $g(0) = g(1)$, and define $\tilde{g} = g\mathbf{1}_{[0;1]}$ on $\mathbb{R}$. Consider that $\tilde{g}$ belongs to Hölder class $\mathcal{H}(\beta, L)$, in Examples E1 to E3, or to Nikol'skii space $\mathcal{N}_2(\beta, L)$ in Example E4. Assume that (K_l) for $l = [\beta]$, and (B_{α_0}), for an $\alpha_0 > 0$, holds. Assume also that (M_p) is fulfilled for Examples E1 and E2. Then,*

$$\mathbb{E}\left[\|\hat{s} - s\|^2_{\phi'_X}\right] \leq Cn^{-\frac{2\beta}{2\beta+1}}, \tag{11}$$

where C is a constant which does not depend on n and β.

We recover the classical optimal rate in nonparametric estimation. Notice that the bounds (10) and (11) we provide are global ones: they hold for the MISE, with global bandwidth selection. Here, adaptation has no price: the rate of convergence is the one found for the bias term, without data-driven selection of the bandwidth, just by minimizing the right-hand side of (5). On the contrary, it is well known that adaptation costs a logarithm factor for pointwise selection. This explains why we focus on global selection, which is sufficient for our purpose (as it is shown in the previous theorem, and in the simulation study below).

4. Illustration

To illustrate the procedure, we focus only on two of the four examples: the classical additive regression (Example E1), and the estimation of c.d.f. under interval censoring case I. In each case, we propose to compare the warped kernel strategy, which we denote by WK in this section,

with another adaptive method: a regression type one, based on the minimization of a penalized least-squares contrast. We denote it by LS.

4.1. **Implementation of the warped-kernel estimators.** The theoretical study allows the choice of several kernels and bandwidth collection. For practical purpose, we consider the Gaussian kernel, $K : x \mapsto e^{-x^2/2}/\sqrt{2\pi}$, which satisfies Assumption (K_1). It has the advantage of having simple convolution-products:

$$\forall h, h' > 0, \quad K_h \star K_{h'} = K_{\sqrt{h^2+h'^2}}. \tag{12}$$

The experiment is conducted with the dyadic collection $\mathcal{H}_{n,2}$ defined by Remark 1. Notice that the larger collection $\mathcal{H}_{n,1}$ has also been tested: since it does not really improve the results but increases the computation time, we only keep the other collection. Besides, the simulations are performed in the general case of unknown ϕ_X, which equals F_X in Examples E1 and E3. We replace each of its occurrences by the empirical c.d.f. $\hat{F}_n = (1/n)\sum_{i=1}^n \mathbf{1}_{[X_i;\infty[}$. Therefore, the estimator is

$$\hat{s} \; : \; x \; \mapsto \; \frac{1}{n}\sum_{i=1}^{n} \theta(Y_i) K_{\hat{h}}(\hat{F}_n(x) - \hat{F}_n(X_i)).$$

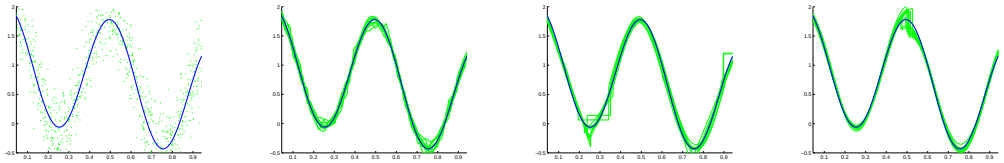
Then, the estimation procedure can be decomposed in some steps:

- Simulate a data sample (X_i, Y_i), $i = 1, \ldots, n$, fitting Example E1 or E3.
- Compute $V(h)$ and $A(h)$ for each $h \in \mathcal{H}_{n,1}$.
  - For $V(h)$: we have calibrated the numerical κ involved in (6). A lower bound for its theoretical value is provided by the proof. However, we keep in mind that this value is very pessimistic due to rough upper-bounds (for the sake of clarity). Thus, a practical calibration is required, like in most model selection devices. Since classical techniques such as the slope heuristic are not currently well developped for the GL method, we adjust κ prior to the comparison with the other estimates: we look at the quadratic risk with respect to the value of κ, and choose one of the values leading to reasonable risk.
  - For $A(h)$: thanks to (12), the auxiliary estimates are easily computed: $\hat{s}_{h,h'} = \hat{s}_{\sqrt{h^2+h'^2}}$. The L^2-norm is then approximated by a Riemann sum:

$$\|\hat{s}_{h,h'} - \hat{s}_{h'}\|^2_{\phi'_X} = \|\hat{g}_{h,h'} - \hat{g}_{h'}\|^2_{L^2(\phi_X(A))} \approx \frac{1}{K}\sum_{k=1}^{K} \left(\hat{g}_{h,h'}(u_k) - \hat{g}_{h'}(u_k)\right)^2,$$

    where $K = 50$, and $(u_k)_k$ are grid points evenly distributed across $\phi_X(A)$.
- Select $\hat{h}$ such that $A(h) + V(h)$ is minimum.
- Compute $\hat{s}_{\hat{h}}$.

4.2. **Example E1: additive regression.** We compare the warped kernel method (WK) with the strategy from Baraud (2002): the model selection device is designed with a penalized least-squares contrast, leading to an adaptive projection estimator, developed in an orthogonal basis of $L^2(A)$. The experiment is carried out with the Matlab toolbox FY3P, written by Yves Rozenholc, and available on his web page http://www.math-info.univ-paris5.fr/ rozen/. We choose a regular piecewise polynomial basis, with degrees chosen in an adaptive way. Since we use a kernel with only one vanishing moment, the comparison is fair if we consider polynomials with degrees equal to or less than 1, so that the bias of the least-squares estimator has the same order than the one of the warped-kernel estimate. We denote by LS1 the resulting estimator. However, as shown below, we will see that the warped-kernel generally outperforms the least-square, even if we use polynomials with degree at most 2 (LS2). We also experiment the Fourier basis, but the results are not as good as the polynomial basis. Thus, we do not mention the values.



(a) (b) (c) (d)

FIGURE 1. Estimation in Example E1, with true regression function s_3, design distribution $\gamma(4, 0.08)$, and $n = 1000$. (a) points: data $(X_i, Y_i)_i$, thick line: true function s_3. (b)-(c)-(d) beams of 20 estimators built from i.i.d. sample (thin lines) versus true function (thick line): warped kernel estimators (subplot (b)), least-squares estimator in piecewise polynomial bases with degree at most 1 (subplot (c)) or 2 (subplot (d)).

The procedure is applied for different regression functions, design and noise. We focus on the three following regression functions

$$\begin{aligned} s_1 &: x \mapsto x(x-1)(x-0.6) \\ s_2 &: x \mapsto -\exp(-200(x-0.1)^2) - \exp(-200(x-0.9)^2) + 1 \\ s_3 &: x \mapsto \cos(4\pi x) + \exp(-x^2) \end{aligned}$$

The influence of the design is explored through four distributions:

- $\mathcal{U}_{[0;1]}$, the uniform distribution on the interval $[0; 1]$,
- $\gamma(4, 0.08)$, the Gamma distribution, with parameters 4 and 0.08 (0.08 is the scale parameter),
- $\mathcal{N}(0.5, 0.01)$, the Gaussian distribution with mean 0.5 and variance 0.01,
- $\mathcal{BN}$ a bimodal Gaussian distribution, with density $x \mapsto c(\exp(-200(x-0.05)^2) + \exp(-200(x-0.95)^2))$ (c is a constant adjusted to obtain a density function).

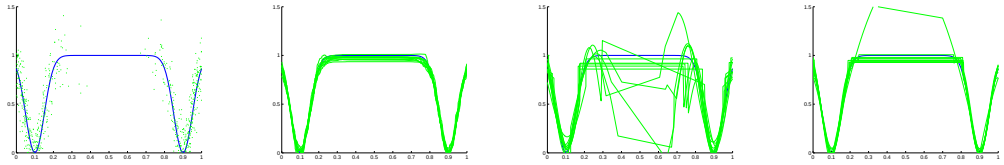
We also test the sensibility of the method to the noise distribution: contrary to the underlying design distribution, it does not seem to affect the results. Thus, we present the simulation for a Gaussian centered noise, with variance σ^2. We take into account the signal-to-noise ratio: therefore, the value of σ is chosen in each model such that the ratio of the variance of the signal ($\mathrm{Var}(s(X_1))$) over the variance of the noise ($\mathrm{Var}(\varepsilon_1)$) approximately equals 2.

Figures 1 and 2 plot the generated data-sets and the function to estimate, and illustrate the visual quality of the reconstruction: beams of estimators (WK, LS1, and LS2) are presented. Figure 1 shows a regular case, while Figure 2 aims at depicting the case where a hole occurs in the design density: the estimator built with warped kernel behaves still correctly, even if the data are very inhomogeneous.

We also perform a study of the risk which is reported in Table 3, for the sample size $n = 60, 200, 500$ and 1000. The MISE criterion is retained. To be more precise, it is computed over J sample replications, and the quadratic norm is approximated as follows:

$$MISE_j = \frac{b-a}{N} \sum_{k=1}^{N} \left(\tilde{s}(x_k) - s(x_k)\right)^2,$$

where $\tilde{s}$ stands for one of the estimators, b is the quantile of order 95% of the X_i and a is the quantile of order 5%. The $(x_k)_{k=1,\dots,N}$ are the sample points falling in $[a; b]$. Finally, the values displayed in Table 3 are the mean of the previous values for $j \in \{1, \dots, J = 200\}$. In 56% of the



(a) (b) (c) (d)

FIGURE 2. Estimation in Example E1, with true regression function s_2, design distribution $\mathcal{BN}$, and $n = 1000$. (a) points: data $(X_i, Y_i)_i$, thick line: true function s_2. (b)-(c)-(d) beams of 20 estimators built from i.i.d. sample (thin lines) versus true function (thick line): warped kernel estimators (subplot (b)), least-squares estimator in piecewise polynomial bases with degree at most 1 (subplot (c)) or 2 (subplot (d)).

examples, the risks of the warped-kernel estimator are smaller than the ones of the least-squares estimator, in piecewise polynomials basis with degrees at most 2 (LS2). Besides, if we consider the comparison with LS1, which is more fair as explained above, the WK estimators give better results in 77% of the cases.

4.3. **Example E3: Interval censoring, case 1.** The same comparison is carried out for the estimation of the c.d.f. under interval censoring. The adaptive least-squares estimate is provided by Brunel and Comte (2009), and the same Matlab toolbox is used for its implementation: in fact, in this statistical model, recall that the target function can be seen as a regression function: $s(x) = \mathbb{P}(Z \leq x) = \mathbb{E}[\mathbf{1}_{Z \leq x} | X = x] = \mathbb{E}[Y|X]$.

We consider different models for generating the data. We have calibrated the estimation set A, such that most of the data belong to this interval, as it is done in Brunel and Comte (2009). We shorten "follow the distribution" by the symbol "$\sim$".

- M1: $X \sim \mathcal{U}_{[0;1]}$, and $Z \sim \mathcal{U}_{[0;1]}$, $A = [0;1]$ (for instance, the target function is $F_Z : x \mapsto x$),
- M2: $X \sim \mathcal{U}_{[0;1]}$, and $Z \sim \chi_2(1)$ (Chi-squared distribution with 1 degree of freedom), $A = [0;1]$,
- M3: $X \sim \mathcal{E}(1)$ (exponential distribution with mean 1), and $Z \sim \chi_2(1)$, $A = [0;1.2]$,
- M4: $X \sim \beta(4,6)$ (Beta distribution of parameter (4,6)), $Z \sim \beta(4,8)$, $A = [0;0.5]$,
- M5: $X \sim \beta(4,6)$, $Z \sim \mathcal{E}(10)$ (exponential distribution with mean 0.1), $A = [0;0.5]$,
- M6: $X \sim \gamma(4, 0.08)$, $Z \sim \mathcal{E}(10)$, $A = [0, 0.5]$,
- M7: $X \sim \mathcal{E}(0.1)$, $Z \sim \gamma(4,3)$, $A = [1;23]$.

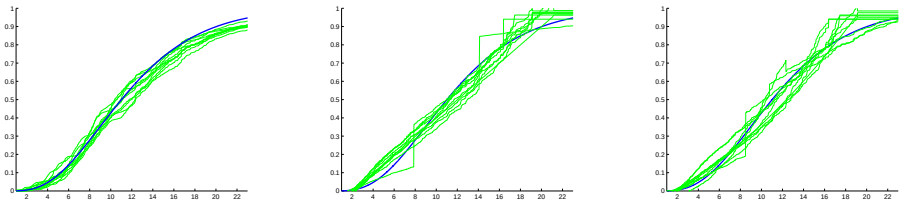
The first two models, and the fourth, were also used by Brunel and Comte (2009). All these models allow us to investigate thoroughly the sensibility of the method to the distribution of the examination time X, and to the range of the estimation interval.

Figure 3 shows the smoothness of warped-kernel estimates. We also explore the difference between the estimators by computing the MISE for the different models. Table 4 reveals that the warped-kernel estimates can advantageously be used as soon as the design X_i has not a uniform distribution: it always outperform the least-squares estimators in these cases.

To conclude, these results must be put into perspectives: for Example E1 as much as Example E3, more classes of functions and models should be studied to confirm the interest of the warped-kernel strategy, but it is beyond the scope of the paper. We just aim at illustrating that our estimators can stand comparison with other adaptive methods, in various models from the

s	X	σ	n= 60	200	500	1000	Method
s_1	$\mathcal{U}_{[0;1]}$	$\sqrt{.0006}$	0.3719	0.1341	0.1957	0.2454	WK
			0.3892	0.1293	0.0681	0.0446	LS2
	$\gamma(4, 0.08)$	5.10^{-5}	0.0052	0.0033	0.0004	0.0003	WK
			0.0097	0.004	0.0017	0.0012	LS2
	$\mathcal{N}(0.5, 0.01)$	0.011	0.0049	0.0020	0.0008	0.0005	WK
			0.0020	0.0012	0.0010	0.0008	LS2
	$B\mathcal{N}$	0.022	0.524	0.422	0.267	0.205	WK
			0.166	0.054	0.038	0.029	LS2
s_2	$\mathcal{U}_{[0;1]}$	0.17	16.35	6.791	3.51	0.837	WK
			33.212	2.058	0.691	0.407	LS2
	$\gamma(4, 0.08)$	0.08	1.885	0.354	0.204	0.147	WK
			4.047	0.801	0.552	0.429	LS2
	$\mathcal{N}(0.5, 0.01)$	0.01	0.0619	0.0186	0.0079	0.0006	WK
			0.0078	0.0014	0.0001	0.0001	LS2
	$B\mathcal{N}$	0.18	12.052	5.279	1.698	1.041	WK
			52.668	11.009	5.817	1.215	LS2
s_3	$\mathcal{U}_{[0;1]}$	0.35	28.03	10.55	4.63	2.747	WK
			125.055	45.298	12.607	5.713	LS1
			31.073	7.477	4.199	3.319	LS2
	$\gamma(4, 0.08)$	0.44	19.615	6.283	3.869	3.309	WK
			41.261	13.34	4.808	3.727	LS1
			23.213	5.549	2.059	0.86	LS2
	$\mathcal{N}(0.5, 0.01)$	0.44	6.341	2.452	1.28	0.861	WK
			10.453	3.961	2.098	1.078	LS1
			3.753	1.386	1.028	0.644	LS2
	$B\mathcal{N}$	0.32	44.381	13.618	9.637	7.928	WK
			182.525	58.787	24.229	12.317	LS1
			66.663	30.377	8.521	4.574	LS2

TABLE 3. Values of MISE $\times 1000$ averaged over 200 samples, for the estimators of the regression function (Example E1), built with the warped kernel method (WK) or the least-squares methods, with piecewise polynomials of degree at most 1 or 2 (LS1 or LS2).



(a) (b) (c)

FIGURE 3. Estimation in Example E3, in model M7, and $n = 1000$. (a)-(b)-(c) beams of 20 estimators built from i.i.d. sample (thin lines) versus true function (thick line): warped kernel estimators (subplot (a)), least-squares estimator in piecewise polynomial bases with degree at most 1 (subplot (b)) or 2 (subplot (c)).

Model	X	Z	[a;b]	n= 60	200	500	1000	Method
1	$\mathcal{U}_{[0;1]}$	$\mathcal{U}_{[0;1]}$	$[0;1]$	2.41	1.125	0.975	0.533	WK
				0.63	0.111	0.056	0.024	LS2
2	$\mathcal{U}_{[0;1]}$	$\chi_2(1)$	$[0;1]$	1.558	0.804	0.57	0.415	WK
				1.602	0.44	0.244	0.13	LS2
3	$\mathcal{E}(1)$	$\chi_2(1)$	$[0;1.2]$	1.285	0.614	0.243	0.247	WK
				2.385	0.893	0.651	0.365	LS2
4	$\mathcal{B}(4,6)$	$\mathcal{B}(4,8)$	$[0;0.5]$	0.423	0.236	0.09	0.094	WK
				0.449	0.271	0.117	0.105	LS2
5	$\mathcal{B}(4,6)$	$\mathcal{E}(10)$	$[0;0.5]$	0.388	0.229	0.119	0.103	WK
				0.467	0.261	0.13	0.095	LS2
6	$\gamma(4,0.08)$	$\mathcal{E}(10)$	$[0;0.5]$	0.424	0.166	0.102	0.069	WK
				0.698	0.286	0.162	0.095	LS2
7	$\mathcal{E}(0.1)$	$\gamma(4,3)$	$[1;23]$	14.955	5.145	3.973	2.113	WK
				19.825	11.797	9.738	5.898	LS2

TABLE 4. Values of MISE $\times 100$ averaged over 100 samples, for the estimators of the c.d.f. from current status data (Example E3) built with the warped kernel method (WK) or the least-squares methods, with piecewise polynomials of degree at most 1 or 2 (LS1 or LS2).

classical regression model to some estimation settings with censored data, while being simple and fast to implement.

5. Extension to the estimation of a conditional density

5.1. **Presentation.** From now on, we consider an extension of warped-kernel strategy to estimate a bivariate function, the conditional density. Assume again that we observe pairs $(X_i, Y_i)_{i\in\{1,\dots,n\}}$ of real random variables such as (X, Y), and denote by $f_{(X,Y)}$ a joint density of the couple. The relationship between the predictor X and the response Y can be thoroughly described by the conditional density,

$$\forall (x,y) \in A \times B, \;\; \pi(x,y) = \frac{f_{(X,Y)}(x,y)}{f_X(x)}$$

by assuming that f_X does not vanish on A.

Kernel estimators for π have been widely studied: typically, Nadaraya-Watson type estimates, such as "double-kernel" ratio estimators, with cross-validation methods to select bandwidths are studied from the asymptotic point of view (convergence rates and asymptotic normality are shown): among others, see Hyndman et al. (1996), Hyndman and Yao (2002), De Gooijer and Zerom (2003) or Fan and Yim (2004).

We propose to adopt again in this section a nonasymptotic setting, and to build a warped-kernel estimate for the conditional density π, which satisfies adaptive properties, like the recent estimates proposed by Brunel et al. (2007), Efromovich (2010), Akakpo and Lacour (2011), or Cohen and Le Pennec (2011), while having a simple expression.

To adapt the previous method, we only warp the first coordinate of π, by using the c.d.f. F_X of the design X as warping function:

$$g^{(cd)} : (u,y) \in [0;1] \times B \mapsto \pi\left(F_X^{-1}(u), y\right) \tag{13}$$

is the new auxiliary function to be estimated first. Let us introduce some notations. We consider two kernel functions $K^{(1)}$ and $K^{(2)}$, which are supposed to be squared-integrable on $\mathbb{R}$. From this point, $\mathcal{H}_n^{(cd)}$ denotes a set of bandwidth couples $(h_1, h_2) \in (\mathbb{R}_+^*)^2$. We set again $K_{h_l}^{(l)} : x \mapsto K^{(l)}(x/h_l)/h_l$ $(l = 1, 2)$, and denote by $\mathbb{K}(u,y) = K_{h_1}^{(1)} \otimes K_{h_2}^{(2)}(u,y)$ the product $K_{h_1}^{(1)}(u)K_{h_2}^{(2)}(y)$, for all real numbers u and y. The functional spaces and corresponding norm are also adapted to the bivariate setting, with the warping function ϕ_X equal to F_X. Particularly, $t \in L^2(A \times B, f_X)$ means

$$\|t\|_{f_X}^2 := \int_{\mathbb{R}^2} t(x,y) f_X(x) dxdy < \infty.$$

Finally, in this bivariate framework, the assumption (B_{α_0}) becomes:

Assumption $(B_{\alpha_0}^{(cd)})$:

(i) For any constant $\kappa_0 > 0$, there exists $C_0 > 0$, such that $\sum_{(h_1,h_2)\in\mathcal{H}_n} \exp\left(-\frac{\kappa_0}{h_1 h_2}\right) \leq C_0(\kappa_0)$,

(ii) There exists $\alpha_0 > 0$ such that $\sum_{h\in\mathcal{H}_n} \frac{1}{h_1 h_2} \leq k_0 n^{\alpha_0}$, for a constant $k_0 \geq 0$.

5.2. **Estimation and performance.** The cornerstone of the method in this new setting is to remark that the auxiliary $g^{(cd)}$ defined by (13) is the density of the transformed data $(F_X(X), Y)$. Thus, a collection of kernel estimators for $g^{(cd)}$ is

$$\forall (h_1,h_2) \in \mathcal{H}_n^{(cd)}, \;\; \hat{g}_{h_1,h_2}^{(cd)} : (u,y) \mapsto \frac{1}{n}\sum_{i=1}^n K_{h_1}^{(1)}\left(u - F_X(X_i)\right) K_{h_2}^{(2)}\left(y - Y_i\right),$$

and the analogous collection for π is $(\hat{\pi}_{h_1,h_2})_{(h_1,h_2)\in\mathcal{H}_n^{(cd)}}$, with $\hat{\pi}_{h_1,h_2}(x,y) = \hat{g}^{(cd)}_{h_1,h_2}(F_X(x),y)$. Stute (1986) studied similar estimators for a conditional distribution function. More recently, this collection has already been considered by Mehra et al. (2000), who compared it asymptotically to classical Nadaraya-Watson estimates.

The novelty which must be underlined here is the GL selection of the best bandwidths $(\hat{h}_1,\hat{h}_2)$: we set, in the same way as (6) and (7):

$$\forall (h_1,h_2)\in\mathcal{H}_n^{(cd)}, \begin{cases} V^{(cd)}(h_1,h_2) = \delta'(1+\|\mathbb{K}\|^2_{L^1(\mathbb{R})})\dfrac{\|\mathbb{K}\|_{L^2(\mathbb{R}^2)}}{nh_1h_2}, \\ A^{(cd)}(h_1,h_2) = \max\limits_{(h_1',h_2')\in\mathcal{H}_n}\left\{\|\hat{g}_{(h_1,h_2),(h_1',h_2')} - \hat{g}_{h_1',h_2'}\|^2 - V^{(cd)}(h_1',h_2')\right\}_+, \end{cases}$$

with

$$\hat{g}_{(h_1,h_2),(h_1',h_2')} \ : \ (u,y)\mapsto K^{(1)}_{h_1'}\otimes K^{(2)}_{h_2'} \star \left(\hat{g}_{h_1,h_2}\mathbf{1}_{[0;1]\times A_2}\right)(u,y).$$

To realize the bias-variance compromize, we define:

$$(\hat{h}_1,\hat{h}_2) = \arg\min_{(h_1,h_2)\in\mathcal{H}_n^{(cd)}}\{A^{(cd)}(h_1,h_2)+V^{(cd)}(h_1,h_2)\}.$$

We now set the oracle-type inequality, concerning the selected estimator $\hat{\pi}_{\hat{h}_1,\hat{h}_2}$.

Theorem 2. *Assume that π is bounded, and that Assumption $(B^{(cd)}_{\alpha_0})$ holds for the collection $\mathcal{H}_n^{(cd)}$. Then, there exists three constants κ_l, $l=1,2,3$ such that*

$$(14)\quad \mathbb{E}\left[\left\|\hat{\pi}_{\hat{h}_1,\hat{h}_2}-\pi\right\|^2_{f_X}\right] \le \min_{(h_1,h_2)\in\mathcal{H}_n^{(cd)}}\left\{c_1\left\|g^{(cd)}-g^{(cd)}_{h_1,h_2}\right\|^2_{f_X} + c_2\frac{\|K^{(1)}\times K^{(2)}\|^2_{L^2(\mathbb{R}^2)}}{nh_1h_2}\right\} + \frac{c_3}{n},$$

where $g^{(cd)}_{h_1,h_2} = (K^{(1)}_{h_1}\otimes K_{h_2})\star(g^{(cd)}\mathbf{1}_{[0;1]\times B})$, and where the κ_l depend on $\|K^{(1)}\times K^{(2)}\|_{L^1(\mathbb{R}^2)}$ $(l=1,2,3)$ and κ_3 additionally depends on $\|g^{(cd)}\|_{L^\infty([0;1]\times B)}$.

Therefore, the warped kernel strategy could be successfully adapted to a bivariate framework. The crucial choice of the bandwidth is performed automatically: thanks to the GL method, the optimal trade-off is reached. Therefore, we extend the results of Mehra et al. (2000) about warped kernel conditional density estimator.

Moreover, it should be mentioned that the estimator admits a simple expression, and can be consequently implemented with low complexity, like in the four univariate examples.

Finally, Inequality (14) can be used for derivation of adaptive optimal rates for conditional density estimation. For that purpose, we need to assume that the kernels $K^{(1)}$ and $K^{(2)}$ have vanishing moment property (such as Assumption (K_l)), and the convergence rate is established over anisotropic Hölder classes (defined for example in Section 2.4 of Comte and Lacour 2011). We do not detail this, since the main goal of this section is to show the adaptation of warped-bases strategy to establish nonasymptotic bound for conditional density estimation.

6. Proofs

We start with some useful results. The first one is a powerful concentration inequality, which permits to control the deviations of the supremum of an empirical process.

Lemma 3. *[Talagrand's Inequality] Let $\xi_1,\dots,\xi_n$ be i.i.d. random variables, and define $\nu_n(r) = \frac{1}{n}\sum_{i=1}^n r(\xi_i)-\mathbb{E}[r(\xi_i)]$, for r belonging to a countable class $\mathcal{R}$ of real-valued measurable functions.*

Then, for $\delta > 0$, there exist three constants c_l, $l = 1, 2, 3$, such that

$$\mathbb{E}\left[\left(\sup_{r\in\mathcal{R}}(\nu_n(r))^2 - c(\delta)H^2\right)_+\right] \leq c_1\left\{\frac{v}{n}\exp\left(-c_2\delta\frac{nH^2}{v}\right) + \frac{M_1^2}{C^2(\delta)n^2}\exp\left(-c_3C(\delta)\sqrt{\delta}\frac{nH}{M_1}\right)\right\},$$

with, $C(\delta) = (\sqrt{1+\delta}-1)\wedge 1$, $c(\delta) = 2(1+2\delta)$ and

$$\sup_{r\in\mathcal{R}}\|r\|_\infty \leq M_1,\ \mathbb{E}\left[\sup_{r\in\mathcal{R}}|\nu_n(r)|\right] \leq H,\ \text{and}\ \sup_{r\in\mathcal{R}} Var\left(r\left(\xi_1\right)\right) \leq v.$$

Inequality (3) is a classical consequence of the Talagrand Inequality given in Klein and Rio (2005): see for example Lemma 5 (page 812) in Lacour (2008).

Then, we state a lemma which will allow us to replace a L^2-norm by the supremum of an empirical process.

Lemma 4. *Let B be a borelian subset of $\mathbb{R}$ (or $\mathbb{R}^2$). Denote by $\tilde{S}_B(0,1)$ the set of functions $t \in L^1(B)\cap L^2(B)$ such that $\|t\|_{L^2(B)} = 1$. Then, for any function $v \in L^1(B)\cap L^2(B)$,*

$$\|v\|_{L^2(B)} = \sup_{t\in\tilde{S}_B(0,1)}\langle v, t\rangle_B.$$

Moreover, the supremum over $\tilde{S}_B(0,1)$ equals the supremum over a countable subset $\bar{S}_B(0,1)$ of $\tilde{S}_B(0,1)$.

Proof of Lemma 4. The Cauchy-Schwarz Inequality leads to

$$\sup_{t\in\tilde{S}_B(0,1)}\langle v, t\rangle_B \leq \sup_{t\in\tilde{S}_B(0,1)}\|v\|_{L^2(B)}\|t\|_{L^2(B)} = \|v\|_{L^2(B)}.$$

Besides, if we set $t = v/\|v\|_{L^2(B)}$, then t belongs to $\tilde{S}_B(0,1)$, and $\langle t, v\rangle_B = \|v\|_{L^2(B)}$. This ends the proof of the equality. Finally, we can replace $\tilde{S}_B(0,1)$ by one of its dense countable subset: such a set exists thanks to the separability of $L^2(\mathbb{R})$ (or $L^2(\mathbb{R}^2)$).

□

Finally, we recall a useful and standard property of the convolution product.

Lemma 5. *[Young Inequality] Let $p, q \in [1;\infty[$ such that $1/p + 1/q \geq 1$. If $u \in L^p(\mathbb{R})$ and $v \in L^q(\mathbb{R})$, then the convolution product $u\star v$ exists. Moreover, if r is defined by $1/r = 1/p+1/q-1$ then $u\star v \in L^r(\mathbb{R})$ and*

$$\|u\star v\|_{L^r(\mathbb{R})} \leq \|u\|_{L^p(\mathbb{R})}\|v\|_{L^q(\mathbb{R})}.$$

6.1. **Proof of Equality (3).** We compute $\mathbb{E}[\theta(Y)K_h(u-\phi_X(X))]$, for $u \in \phi_X(A)$, in Examples E1 to E3. Recall that we deal with Example E4 in Section 2.2 directly.

• **Example E1.** Since $Y = s(X) + \varepsilon$, X and ε are independent, and ε is centered,

$$\begin{aligned}\mathbb{E}\left[YK_h(u - F_X(X))\right] &= \mathbb{E}\left[s(X)K_h\left(u - F_X(X)\right)\right],\\ &= \int_A s(x)K_h(u - F_X(x))f_X(x)dx,\\ &= \int_0^1 g(u')K_h(u-u')du' = K_h\star g\mathbf{1}_{[0;1]}(u),\end{aligned}$$

by setting $u' = F_X(x)$ in the integral.

• **Example E2.** The computations are similar, with the specific properties or Example E2:

$$\begin{aligned}\mathbb{E}\left[Y^2K_h(u-F_X(X))\right] &= \mathbb{E}\left[\varepsilon^2\right]\mathbb{E}\left[\sigma^2(X)K_h\left(u-F_X(X)\right)\right],\\ &= \int_A \sigma^2(x)K_h(u-F_X(x))f_X(x)dx,\\ &= \int_0^1 g(u')K_h(u-u')du' = K_h\star\left(g\mathbf{1}_{[0;1]}\right)(u).\end{aligned}$$

• **Example E3.** Here, we obtain

$$\begin{aligned}\mathbb{E}\left[YK_h(u-F_X(X))\right] &= \mathbb{E}\left[\mathbf{1}_{Z\leq X}K_h\left(u-F_X(X)\right)\right],\\ &= \int_{A\times\mathbb{R}} \mathbf{1}_{z\leq x}K_h\left(u-F_X(x)\right)f_X(x)f_Z(z)dxdz,\\ &= \int_A K_h\left(u-F_X(x)\right)f_X(x)\left(\int_{\mathbb{R}}\mathbf{1}_{z\leq x}f_Z(z)dz\right)dx,\\ &= \int_A K_h\left(u-F_X(x)\right)f_X(x)F_Z(x)dx,\\ &= \int_0^1 K_h\left(u-u'\right)F_Z\circ F_X^{-1}(u')du' = \int_0^1 K_h\left(u-u'\right)g(u')du',\\ &= K_h\star\left(g\mathbf{1}_{[0;1]}\right)(u).\end{aligned}$$

The proof of Equality (3) is thus completed.

□

6.2. **Proof of Theorem 1.** Let $h\in\mathcal{H}_n$ be fixed. We begin with the following decomposition for the loss of the estimator $\tilde{s}=\hat{s}_{\hat{h}}$:

$$\begin{aligned}\left\|\hat{s}_{\hat{h}}-s\right\|^2_{\phi'_X} &= \left\|\hat{g}_{\hat{h}}-g\right\|^2_{L^2(\phi_X(A))},\\ &\leq 3\left\|\hat{g}_{\hat{h}}-\hat{g}_{h,\hat{h}}\right\|^2_{L^2(\phi_X(A))}+3\left\|\hat{g}_{h,\hat{h}}-\hat{g}_h\right\|^2_{L^2(\phi_X(A))}+3\left\|\hat{g}_h-g\right\|^2_{L^2(\phi_X(A))}.\end{aligned}$$

The definitions of $A(h)$ and $A(\hat{h})$ enable to write,

$$\begin{aligned}3\left\|\hat{g}_{\hat{h}}-\hat{g}_{h,\hat{h}}\right\|^2_{L^2(\phi_X(A))}+3\left\|\hat{g}_{h,\hat{h}}-\hat{g}_h\right\|^2_{L^2(\phi_X(A))} &\leq 3\left(A(h)+V\left(\hat{h}\right)\right)+3\left(A\left(\hat{h}\right)+V(h)\right),\\ &\leq 6\left(A(h)+V(h)\right),\end{aligned}$$

by using the definition of $\hat{h}$. Besides, we have already studied the bias-variance decomposition of $\hat{g}_h$ (see the beginning of Section 2.3):

$$\mathbb{E}\left[\|\hat{g}_h-g\|^2_{L^2(\phi_X(A))}\right]\leq\frac{\mathbb{E}[\theta^2(Y_1)]\|K\|^2_{L^2(\mathbb{R})}}{nh}+\|g_h-g\|^2_{L^2(\phi_X(A))}.$$

Thus,

$$\mathbb{E}\left[\left\|\hat{s}_{\hat{h}}-s\right\|^2_{\phi'_X}\right]\leq 6\mathbb{E}\left[A(h)\right]+6V(h)+\frac{\mathbb{E}[\theta^2(Y_1)]\|K\|^2_{L^2(\mathbb{R})}}{nh}+3\|g_h-g\|^2_{L^2(\phi_X(A))}. \tag{15}$$

Therefore, the remainding part of the proof follows from the lemma hereafter.

Lemma 6. *Let $h \in \mathcal{H}_n$ be fixed. Under the assumptions of Theorem 1, there exist a constant C_1 which depends on $\|K\|_{L^1(\mathbb{R})}$, and a constant C_2 which depends on $\|s\|_{L^\infty(A)}$, $\|K\|_{L^1(\mathbb{R})}$ and $\|K\|_{L^2(\mathbb{R})}$ in Examples E3 and E4, and also on $\mathbb{E}[\varepsilon_1^2]$, $\mathbb{E}[\varepsilon_1^{2+p}]$ and $\mathbb{E}[s^2(X_1)]$ in Example E1 or on $\mathbb{E}[\varepsilon_1^4]$, $\mathbb{E}[\varepsilon_1^{4+p}]$ and $\mathbb{E}[\sigma^4(X_1)]$ in Example E2, such that,*

$$\mathbb{E}\left[A(h)\right] \leq C_1 \|g_h - g\|^2_{L^2(\phi_X(A))} + \frac{C_2}{n}. \tag{16}$$

Applying Inequality (16) in (15) implies (10) by taking the infimum over $h \in \mathcal{H}_n$. This ends the proof of Theorem 1.

□

6.3. **Proof of Lemma 6.** To study $A(h)$, we introduce the auxiliary quantities $g_{h,h'} := K_{h'} \star (g_h \mathbf{1}_{\phi_X(A)}) = K_{h'} \star ((K_h \star g\mathbf{1}_{\phi_X(A)})\mathbf{1}_{\phi_X(A)})$, for any $h' \in \mathcal{H}_n$, and we first split

$$\|\hat{s}_{h,h'} - \hat{s}_{h'}\|^2_{f_X} = \|\hat{g}_{h,h'} - \hat{g}_{h'}\|^2_{L^2(\phi_X(A))} \leq 3\left(T_a + T_b + \|\hat{g}_{h'} - g_{h'}\|^2_{L^2(\phi_X(A))}\right), \tag{17}$$

where

$$T_a = \|\hat{g}_{h,h'} - g_{h,h'}\|^2_{L^2(\phi_X(A))}, \quad T_b = \|g_{h,h'} - g_{h'}\|^2_{L^2(\phi_X(A))}.$$

The first term can be bounded as follows, using Lemma 5, with $p = 2$, $q = 1$, and $r = 2$:

$$\begin{aligned} T_a &\leq \left\|K_h \star \left(\hat{g}_{h'}\mathbf{1}_{\phi_X(A)} - g_{h'}\mathbf{1}_{\phi_X(A)}\right)\right\|^2_{L^2(\mathbb{R})}, \\ &\leq \|K\|^2_{L^1(\mathbb{R})} \left\|\hat{g}_{h'}\mathbf{1}_{\phi_X(A)} - g_{h'}\mathbf{1}_{\phi_X(A)}\right\|^2_{L^2(\mathbb{R})} \\ &= \|K\|^2_{L^1(\mathbb{R})} \|\hat{g}_{h'} - g_{h'}\|^2_{L^2(\phi_X(A))}. \end{aligned}$$

In the same way,

$$T_b \leq \|K_{h'}\|^2_{L^1(\mathbb{R})} \|g_h - g\|^2_{L^2(\phi_X(A))}.$$

Therefore, decomposition (17) becomes:

$$\|\hat{s}_{h,h'} - \hat{s}_{h'}\|^2_{f_X} \leq 3\|K\|^2_{L^1(\mathbb{R})}\|g - g_h\|^2_{L^2(\phi_X(A))} + 3(1 + \|K\|^2_{L^1(\mathbb{R})}) \|\hat{g}_{h'} - g_{h'}\|^2_{L^2(\phi_X(A))}.$$

Now, we get back to the definition of $A(h)$ given by (7):

$$\begin{aligned} A(h) \leq 3 &\|K\|^2_{L^1(\mathbb{R})}\|g - g_h\|^2_{L^2(\phi_X(A))} \\ &+3(1 + \|K\|^2_{L^1(\mathbb{R})}) \max_{h' \in \mathcal{H}_n} \left(\|\hat{g}_{h'} - g_{h'}\|^2_{L^2(\phi_X(A))} - \frac{V(h')}{3(1 + \|K\|^2_{L^1(\mathbb{R})})}\right)_+. \end{aligned} \tag{18}$$

We apply Lemma 4:

$$\|\hat{g}_{h'} - g_{h'}\|_{L^2(\phi_X(A))} = \sup_{t \in \bar{S}(0,1)} \langle \hat{g}_{h'} - g_{h'}, t\rangle_{\phi_X(A)},$$

with $\bar{S}(0,1)$ a dense countable subset of $\tilde{S}(0,1) = \{t \in L^1(\phi_X(A)) \cap L^2(\phi_X(A)),\ \|t\|_{L^2(\phi_X(A))} = 1\}$. Now,

$$\begin{aligned} \langle \hat{g}_{h'} - g_{h'}, t\rangle_{\phi_X(A)} &= \frac{1}{n}\sum_{i=1}^n \int_{\phi_X(A)} \left\{\theta(Y_i)K_{h'}\left(u - F_X(X_i)\right) - \mathbb{E}\left[\theta(Y_i)K_{h'}\left(u - F_X(X_i)\right)\right]\right\} t(u)du \\ &= \nu_{n,h'}(t), \end{aligned}$$

where $\nu_{n,h'}$ is an empirical process. Thus, thanks to (18), it remains to bound the deviations of $\sup_{t\in\bar{S}(0,1)}\nu_{n,h'}^2(t)$. First, we have

$$\mathbb{E}\left[\max_{h'\in\mathcal{H}_n}\left(\sup_{t\in\bar{S}(0,1)}\nu_{n,h'}^2(t)-\frac{V(h')}{3(1+\|K\|_{L^1(\mathbb{R})}^2)}\right)_+\right]$$
$$\leq\sum_{h'\in\mathcal{H}_n}\mathbb{E}\left[\left(\sup_{t\in\bar{S}(0,1)}\nu_{n,h'}^2(t)-\frac{V(h')}{3(1+\|K\|_{L^1(\mathbb{R})}^2)}\right)_+\right].$$

Then, the conclusion results from the following lemma:

Lemma 7. *Under the assumptions of Theorem 1, there exists a constant C depending on $\|s\|_{L^\infty(A)}$, $\|K\|_{L^1(\mathbb{R})}$ and $\|K\|_{L^2(\mathbb{R})}$ in Examples E3 and E4, and also on $\mathbb{E}[\varepsilon_1^2]$, $\mathbb{E}[\varepsilon_1^{2+p}]$ and $\mathbb{E}[s^2(X_1)]$ in Example E1 or on $\mathbb{E}[\varepsilon_1^4]$, $\mathbb{E}[\varepsilon_1^{4+p}]$ and $\mathbb{E}[\sigma^4(X_1)]$ in Example E2, such that,*

$$\sum_{h\in\mathcal{H}_n}\mathbb{E}\left[\left(\sup_{t\in\bar{S}(0,1)}\nu_{n,h}^2(t)-\tilde{V}(h)\right)_+\right]\leq\frac{C}{n},$$

with $\tilde{V}(h)=\delta'\|K\|_{L^2(\mathbb{R})}\mathbb{E}[\theta(Y_1)^2]/(nh)$ for a purely numerical $\delta'>0$.

We choose the κ involved in the definition of V such that $\tilde{V}(h)\leq V(h)(1+\|K\|_{L^1(\mathbb{R})}^2)/3$. Thus, the proof is complete.

□

6.4. **Proof of Lemma 7.** We write the empirical process

$$\nu_{n,h}(t) = \frac{1}{n}\sum_{i=1}^n\psi_{t,h}(X_i,Y_i)-\mathbb{E}\left[\psi_{t,h}(X_i,Y_i)\right], \tag{19}$$
$$\text{with } \psi_{t,h}(X_i,Y_i)=\theta(Y_i)\int_{\phi_X(A)}K_h\left(u-F_X(X_i)\right)du.$$

The guiding idea is to apply Talagrand's Inequality (Lemma 3). However, we must distinguish different cases, depending on the example we deal with: for Examples E1 and E2, $\nu_{n,h}(t)$ is not bounded (due to $\theta(Y_i)$), and the inequality cannot be directly applied. On the opposite, in Examples E3 and E4, as we have already remarked, $\theta(Y_i)$ is bounded by 1 and the process is bounded. We detail the first example which is also the most technical, and we review briefly each of the others.

6.4.1. *Example E1.* Recall that $\phi_X=F_X$ and $\phi_X(A)=[0;1]$. We split the process $\nu_{n,h}$ into three parts, writing $\nu_{n,h}=\nu_{n,h}^{(1)}+\nu_{n,h}^{(2,1)}+\nu_{n,h}^{(2,2)}$, with, for $l=1,(2,1),(2,2)$,

$$\nu_{n,h}^{(l)}=\frac{1}{n}\sum_{i=1}^n\varphi_{t,h}^{(l)}(Z_i)-\mathbb{E}\left[\varphi_{t,h}^{(l)}(Z_i)\right],$$

$Z_i=X_i$ or (X_i,ε_i), and

$$\varphi_{t,h}^{(1)}:x\mapsto s(x)\textstyle\int_0^1K_h(u-F_X(x))t(u)du,$$
$$\varphi_{t,h}^{(2,1)}:(x,\varepsilon)\mapsto\varepsilon\mathbf{1}_{|\varepsilon|\leq\kappa_n}\textstyle\int_0^1K_h(u-F_X(x))t(u)du,$$
$$\varphi_{t,h}^{(2,2)}:(x,\varepsilon)\mapsto\varepsilon\mathbf{1}_{|\varepsilon|>\kappa_n}\textstyle\int_0^1K_h(u-F_X(x))t(u)du,$$

where we define, for a constant c which will be specified below,

$$\kappa_n = c\frac{\sqrt{n}}{\ln(n)}. \tag{20}$$

We apply Talagrand's Inequality to the first two bounded empirical processes, and bound roughly the last one. Thus, we split:

$$\begin{aligned}\sum_{h\in\mathcal{H}_n}\mathbb{E}\left[\left(\sup_{t\in\bar{S}(0,1)}\nu_{n,h}^2(t)-\tilde{V}(h)\right)_+\right] &\le 3\sum_{h\in\mathcal{H}_n}\left\{\mathbb{E}\left[\left(\sup_{t\in\bar{S}(0,1)}\left(\nu_{n,h}^{(1)}(t)\right)^2-\frac{\tilde{V}_1(h)}{3}\right)_+\right]\right.\\ &\quad+\mathbb{E}\left[\left(\sup_{t\in\bar{S}(0,1)}\left(\nu_{n,h}^{(2,1)}(t)\right)^2-\frac{\tilde{V}_2(h)}{3}\right)_+\right]\\ &\quad\left.+\,\mathbb{E}\left[\sup_{t\in\bar{S}(0,1)}\left(\nu_{n,h}^{(2,2)}(t)\right)^2\right]\right\},\end{aligned} \tag{21}$$

with the decomposition $\tilde{V}(h)=\tilde{V}_1(h)+\tilde{V}_2(h)$, and, denoting by $\delta''=\delta'/2$,

$$\tilde{V}_1(h)=3\delta''\frac{\|K\|_{L^2(\mathbb{R})}^2\mathbb{E}\left[s^2(X_1)\right]}{nh},$$
$$\tilde{V}_2(h)=3\delta''\frac{\|K\|_{L^2(\mathbb{R})}^2\mathbb{E}\left[\varepsilon_1^2\right]}{nh}.$$

Actually, recall that we have $\mathbb{E}[\theta^2(Y_1)]=\mathbb{E}[Y_1^2]=\mathbb{E}[s^2(X_1)]+\mathbb{E}[\varepsilon_1^2]$ here.

We now show that each of the three terms of the right hand-side of (21) is upper-bounded by a quantity of order $1/n$. This will end the proof.

- ***First term of (21).***

Let us begin with $\nu_{n,h}^{(1)}$. To do so, we compute $H^{(1)}$, $M^{(1)}$ and $v^{(1)}$, involved in Lemma 3.

- For $M^{(1)}$, let $t\in\bar{S}(0,1)$ and $x\in A$ be fixed:

$$\begin{aligned}\left|\varphi_{t,h}^{(1)}(x)\right| &\le |s(x)|\int_0^1|K_h(u-F_X(x))t(u)|\,du,\\ &\le |s(x)|\|K_h\|_{L^2(\mathbb{R})}\|t\|_{L^2(\phi_X(A))}=|s(x)|\frac{\|K\|_{L^2(\mathbb{R})}}{\sqrt{h}},\\ &\le \|s\|_{L^\infty(A)}\frac{\|K\|_{L^2(\mathbb{R})}}{\sqrt{h}}:=M^{(1)}.\end{aligned}$$

- For $H^{(1)}$, notice that

$$\nu_{n,h}^{(1)}(t)=\langle\hat{d}_h-g_h,t\rangle_{\phi_X(A)},\ \text{with}\ \hat{d}_h=\frac{1}{n}\sum_{i=1}^n s(X_i)K_h\left(.-F_X(X_i)\right).$$

Thus, thanks to Lemma 4, we obtain,

$$\begin{aligned}\mathbb{E}\left[\sup_{t\in\bar{S}(0,1)}\left(\nu_{n,h}^{(1)}(t)\right)^2\right] &= \mathbb{E}\left[\left\|\hat{d}_h-g_h\right\|_{L^2([0;1])}^2\right],\\ &= \int_0^1\mathrm{Var}\left(\hat{d}_h(u)\right)du,\ \text{since}\ g_h(u)=\mathbb{E}\left[\hat{d}_h(u)\right],\\ &\le \int_0^1\frac{1}{n}\mathbb{E}\left[s^2(X_1)K_h^2\left(u-F_X(X_1)\right)\right]du.\end{aligned}$$

Then, we use the same computation as the one done to bound the variance term in Section 2.3, and set $(H^{(1)})^2 = \|K\|^2_{L^2(\mathbb{R})}\mathbb{E}[s^2(X_1)]/(nh)$.

- For $v^{(1)}$, we also fix $t \in \bar{S}(0,1)$. Hereafter, if w is a real-valued function, $\check{w}$ is the function $x \mapsto w(-x)$. First,

$$\mathrm{Var}\left(\varphi^{(1)}_{t,h}(X_1)\right) \leq \mathbb{E}\left[\left(\varphi^{(1)}_t(X_1)\right)^2\right] \leq \|s\|^2_{L^\infty(A)}\mathbb{E}\left[\left(\int_0^1 K_h(u - F_X(X_1))t(u)du\right)^2\right],$$

and the expectation can be written

$$\begin{aligned}\mathbb{E}\left[\left(\int_0^1 K_h(u - F_X(X_1))t(u)du\right)^2\right] &= \mathbb{E}\left[\left(\check{K}_h * \left(t\mathbf{1}_{[0;1]}\right)\right)^2(F_X(X_1))\right],\\ &= \int_0^1 \left(\check{K}_h * \left(t\mathbf{1}_{[0;1]}\right)\right)^2(u)du,\\ &\leq \left\|\check{K}_h * \left(t\mathbf{1}_{[0;1]}\right)\right\|^2_{L^2(\mathbb{R})},\\ &\leq \left\|\check{K}_h\right\|^2_{L^1(\mathbb{R})}\|t\mathbf{1}_{[0;1]}\|^2_{L^2(\mathbb{R})} = \left\|\check{K}_h\right\|^2_{L^1(\mathbb{R})}\|t\|^2_{L^2([0;1])},\end{aligned}$$

thanks to Lemma 5. Therefore,

$$\mathrm{Var}\left(\varphi^{(1)}_t(X_1)\right) \leq \|s\|_{L^\infty(A)}\|K\|^2_{L^1(\mathbb{R})} := v^{(1)}.$$

Then, Lemma 3 gives, for $\delta > 0$,

$$\mathbb{E}\left[\left(\sup_{t\in\bar{S}(0,1)}\left(\nu^{(1)}_{n,h}(t)\right)^2 - 2(1+2\delta)\left(H^{(1)}\right)^2\right)_+\right] \leq k_1\left\{\frac{1}{n}\exp\left(-k_2\frac{1}{h}\right) + \frac{1}{n^2h}\exp\left(-k_3\sqrt{n}\right)\right\},$$

where k_1, k_2, k_3 are three constants which depend on $\mathbb{E}[s^2(X_1)]$, $\|s\|_{L^\infty(A)}$, $\|K\|_{L^1(\mathbb{R})}$ and $\|K\|_{L^2(\mathbb{R})}$. Assumption (B_{α_0}) leads to

$$\sum_{h\in\mathcal{H}_n}\mathbb{E}\left[\left(\sup_{t\in\bar{S}(0,1)}\left(\nu^{(1)}_{n,h}(t)\right)^2 - 2(1+2\delta)\|K\|^2_{L^2(\mathbb{R})}\mathbb{E}[s^2(X_1)]\frac{1}{nh}\right)_+\right] \leq \frac{C}{n},$$

with C a constant (which also depends on the previous quantities).

- ***Second term of (21).***

For the second empirical process $\nu^{(2,1)}_{n,h}$, the sketch of the proof is the same: similarly, we compute the quantities involved in the Talagrand Inequality,

$$M^{(2)} = \kappa_n\|K\|_{L^2(\mathbb{R})}\frac{1}{\sqrt{h}},\quad H^{(2)} = \|K\|_{L^2(\mathbb{R})}\left(\mathbb{E}[\varepsilon_1^2]\right)^{1/2}\frac{1}{\sqrt{nh}},\quad v^{(2)} = \|K\|^2_{L^1(\mathbb{R})}\mathbb{E}[\varepsilon_1^2],$$

and we obtain, by Lemma 3, for $\delta > 0$,

$$\mathbb{E}\left[\left(\sup_{t\in\bar{S}(0,1)}\left(\nu^{(2,1)}_{n,h}(t)\right)^2 - 2(1+2\delta)\left(H^{(2)}\right)^2\right)_+\right] \leq k_1\left\{\frac{1}{n}\exp\left(-k_2\frac{1}{h}\right) + \frac{\kappa_n^2}{n^2h}\exp\left(-k_3\frac{\sqrt{n}}{\kappa_n}\right)\right\},$$

where k_1, k_2, k_3 are three constants which depend on $\mathbb{E}[\varepsilon_1^2]$, $\|K\|_{L^1(\mathbb{R})}$ and $\|K\|_{L^2(\mathbb{R})}$. The first term of the right hand-side is like above. With the definition (20) of κ_n, the sum over $h \in \mathcal{H}_n$ of the second term of the upper bound can be written

$$\sum_{h\in\mathcal{H}_n}\frac{\kappa_n^2}{n^2h}\exp\left(-k_3\frac{\sqrt{n}}{\kappa_n}\right) = \frac{c^2}{n^{1+k_3/c}\ln^2(n)}\sum_{h\in\mathcal{H}_n}\frac{1}{h}.$$

Consequently, using Assumption (B_{α_0}) and choosing c in the definition of κ_n such that $c \leq k_3/\alpha_0$, we also obtain for a constant C,

$$\sum_{h\in\mathcal{H}_n} \mathbb{E}\left[\left(\sup_{t\in\bar{S}(0,1)} \left(\nu_{n,h}^{(2,1)}(t)\right)^2 - 2(1+2\delta)\|K\|_{L^2(\mathbb{R})}^2 \mathbb{E}[\varepsilon_1^2]\frac{1}{nh}\right)_+\right] \leq \frac{C}{n}.$$

- ***Third term of (21).***

The last empirical process is $\nu_{n,h}^{(2,2)}(t) = \int_0^1 t(u)\psi(u)du$, with

$$\psi(u) = \frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathbf{1}_{\{|\varepsilon_i|>\kappa_n\}} K_h\left(u - F_X(X_i)\right) - \mathbb{E}\left[\varepsilon_i \mathbf{1}_{\{|\varepsilon_i|>\kappa_n\}} K_h\left(u - F_X(X_i)\right)\right].$$

It is not bounded. Nevertheless, we use the Cauchy-Schwarz Inequality, and the equality $\|t\|_{L^2(\phi_X(A))} = 1$, for $t \in \bar{S}(0,1)$

$$\begin{aligned}
\mathbb{E}\left[\sup_{t\in\tilde{S}(0,1)} \left(\nu_{n,h}^{(2,2)}(t)\right)^2\right] &\leq \mathbb{E}\left[\int_0^1 \psi^2(u)du\right], \\
&\leq \frac{1}{n}\mathbb{E}\left[\varepsilon_1^2 \mathbf{1}_{\{|\varepsilon_1|>\kappa_n\}}\right] \mathbb{E}\left[\int_0^1 K_h^2\left(u - F_X(X_1)\right)du\right], \\
&\leq \frac{\|K\|_{L^2(\mathbb{R})}^2}{nh}\mathbb{E}\left[\varepsilon_1^2 \mathbf{1}_{\{|\varepsilon_1|>\kappa_n\}}\right] \leq \frac{\|K\|_{L^2(\mathbb{R})}^2 \kappa_n^{-p}}{nh}\mathbb{E}\left[\varepsilon_1^{2+p}\right].
\end{aligned}$$

Thus, there exists a constant k_1 which depends on $\|K\|_{L^2(\mathbb{R})}$ and $\mathbb{E}[\varepsilon_1^{2+p}]$,

$$\sum_{h\in\mathcal{H}_n} \mathbb{E}\left[\sup_{t\in\bar{S}(0,1)} \left(\nu_{n,h}^{(2,2)}(t)\right)^2\right] \leq k_1 \frac{\kappa_n^{-p}}{n}\sum_{h\in\mathcal{H}_n}\frac{1}{h} = c_1\kappa^{-p}\frac{\ln^p(n)}{n^{1+p/2}}\sum_{h\in\mathcal{H}_n}\frac{1}{h}.$$

The conclusion comes from Assumption (B_{α_0}), and the choice of $p \geq 2\alpha_0$.

□

6.4.2. *Examples E2 to E4.* For the multiplicative regression model E2, we split the process into two terms: $\nu_{n,h} = \nu_{n,h}^{(1)} + \nu_{n,h}^{(2)}$, with

$$\begin{aligned}
\nu_{n,h}^{(1)}(t) = \frac{1}{n}\sum_{i=1}^n \Bigg\{ &\sigma^2(X_i)\varepsilon_i^2 \mathbf{1}_{\{|\varepsilon_i|\leq\kappa_n\}} \int_0^1 K_h\left(u - F_X(X_i)\right)t(u)du \\
&- \mathbb{E}\left[\sigma^2(X_i)\varepsilon_i^2 \mathbf{1}_{\{|\varepsilon_i|\leq\kappa_n\}} \int_0^1 K_h\left(u - F_X(X_i)\right)t(u)du\right]\Bigg\}, \\
\nu_{n,h}^{(2)}(t) = \frac{1}{n}\sum_{i=1}^n \Bigg\{ &\sigma^2(X_i)\varepsilon_i^2 \mathbf{1}_{\{|\varepsilon_i|>\kappa_n\}} \int_0^1 K_h\left(u - F_X(X_i)\right)t(u)du \\
&- \mathbb{E}\left[\sigma^2(X_i)\varepsilon_i^2 \mathbf{1}_{\{|\varepsilon_i|>\kappa_n\}} \int_0^1 K_h\left(u - F_X(X_i)\right)t(u)du\right]\Bigg\},
\end{aligned}$$

where κ_n is still a constant for the proof, which equals $\sqrt{c\frac{\sqrt{n}}{\ln(n)}}$ and $c > 0$ is obtained by the computations, like in Example E1. We exactly recover the framework of this previous example: the deviations of the process $\nu_{n,h}^{(1)}$ are bounded thanks to Talagrand's Inequality of Lemma 3, and the second one is bounded in the same way as the process $\nu_{n,h}^{(2,2)}$ of the additive regression setting.

For Examples E3 and E4, there is no point in splitting the process (19), since it is already bounded (recall that $\theta(Y_1)$ is bounded by 1). Thus, we apply the concentration inequality.

Recall that $\phi_X(A) = \mathbb{R}_+$. In both of these cases, the quantity M_1 involved in the assumptions of Lemma 3 equals $M_1 = \|K\|_{L^2(\mathbb{R})}/\sqrt{h}$. Moreover, H^2 can be chosen as the upper-bound of the variance term of the estimator $\hat{g}_h$, that is $H^2 = \|K\|_{L^2(\mathbb{R})}/nh$. Finally, v equals $\|K\|_{L^1(\mathbb{R})}$ for Example E3, and $\|g\|_{L^\infty(\mathbb{R}_+)}\|K\|_{L^1(\mathbb{R})}$ for Example E4.

As an example, let us detail the computation of v in Example E4. Recall that $X = C \wedge Z$, $Y = \mathbf{1}_{Z\leq C}$, s is the hazard rate, and the warping ϕ_X is the function $x \mapsto \int_0^x (1 - F_X(t))dt$. Thus, denoting by f_C (respectively f_Z) a density of the variable C (respectively Z), and F_C (respectively F_Z) its c.d.f.,

$$\begin{aligned}
\mathrm{Var}\left(\varphi_{t,h}(X_1, Y_1)\right) &\leq \mathbb{E}\left[(\varphi_{t,h}(X_1, Y_1))^2\right], \\
&= \mathbb{E}\left[Y_1 \left(\int_{\mathbb{R}_+} K_h(u' - \phi(X_1))t(u')du'\right)^2\right], \\
&= \int_{\mathbb{R}_+\times\mathbb{R}} \mathbf{1}_{z\leq c}\left(\int_{\mathbb{R}_+} K_h(u' - \phi(z))t(u')du'\right)^2 f_C(c)f_Z(z)dzdc, \\
&= \int_{\mathbb{R}_+} \left(\int_{\mathbb{R}_+} K_h(u' - \phi(z))t(u')du'\right)^2 f_Z(z)\left(1 - F_C\right)(z)dz.
\end{aligned}$$

We set $z = \phi^{-1}(u)$. The integral becomes

$$\begin{aligned}
&\int_{\mathbb{R}_+} \left(\int_{\mathbb{R}_+} K_h(u' - \phi(z))t(u')du'\right)^2 f_Z(z)\left(1 - F_C\right)(z)dz \\
&= \int_{\mathbb{R}_+} \left(\int_{\mathbb{R}_+} K_h(u' - u)t(u')du'\right)^2 f_Z \circ \phi^{-1}(u)\left(1 - F_C\right) \circ \phi^{-1}(u)\frac{du}{\left((1 - F_X) \circ \phi^{-1}(u)\right)}.
\end{aligned}$$

Thanks to the same arguments as the ones used to prove Equality (3) in Section 2.2, we obtain:

$$\begin{aligned}
\mathrm{Var}\left(\varphi_{t,h}(X_1, Y_1)\right) &\leq \int_{\mathbb{R}_+} g(u)\left(\int_{\mathbb{R}_+} K_h(u' - u)t(u')du'\right)^2 du, \\
&= \int_{\mathbb{R}_+} g(u)\left(K_h * \left(t\mathbf{1}_{\mathbb{R}_+}\right)(u)\right)^2 du, \\
&\leq \|g\|_{L^\infty(\mathbb{R}_+)}\left\|\check{K}_h * \left(t\mathbf{1}_{\mathbb{R}_+}\right)\right\|_{L^2(\mathbb{R})}, \\
&\leq \|g\|_{L^\infty(\mathbb{R}_+)}\left\|\check{K}_h\right\|_{L^1(\mathbb{R})}\left\|\left(t\mathbf{1}_{\mathbb{R}_+}\right)\right\|_{L^2(\mathbb{R})}, \\
&= \|g\|_{L^\infty(\mathbb{R}_+)}\|K\|_{L^1(\mathbb{R})} := v.
\end{aligned}$$

Once we have the three quantities, we easily apply Lemma 3 and the proof is complete by using Assumption (B_{α_0}), like above (see the computations in Example E1).

□

6.5. **Proof of Corollary 1.** We must bound the bias term of the right hand-side of Inequality (10) (Theorem 1). Actually, if we prove that

$$\|s - s_h\|^2_{\phi'_X} \leq Ch^{2\beta},$$

where C is a constant, then the proof of the Corollary will be completed by computing the minimum which is involved in (10).

The beginning of the proof is the same for all the examples (E1 to E4). First,

$$\|s - s_h\|^2_{\phi'_X} = \|g - g_h\|^2_{L^2(\phi_X(A))} = \int_{\phi_X(A)} (g_h(u) - g(u))^2 \, du.$$

We then start with the definition of g_h: for $u \in \phi_X(A)$,

$$\begin{aligned} g_h(u) &= \frac{1}{h}\int_0^1 g(u')K\left(\frac{u-u'}{h}\right)du' = \int_{\frac{u-1}{h}}^{\frac{u}{h}} g(u-hz)K(z)dz, \\ &= \int_{\frac{u-1}{h}}^{\frac{u}{h}} \tilde{g}(u-hz)K(z)dz = \int_{\mathbb{R}} \tilde{g}(u-hz)K(z)dz. \end{aligned}$$

Thus, since $\int_{\mathbb{R}} K(u)du = 1$,

$$\tilde{g}_h(u) - g(u) = \int_{\mathbb{R}} K(z)\tilde{g}(u-hz)dz - \tilde{g}(u) = \int_{\mathbb{R}} K(z)\left[\tilde{g}(u-hz) - \tilde{g}(u)\right]dz.$$

Then we distinguish two cases:

6.5.1. *Examples E1 to E3.* In these cases, recall that $\phi_X(A) = [0;1]$. We use a Taylor-Lagrange formula for $\tilde{g}$: for $u \in [0;1]$, and $z \in \mathbb{R}$, there exists $\theta \in [0;1]$ such that

$$\tilde{g}(u-hz) - g(u) = -hz\tilde{g}'(u) + \frac{(-hz)^2}{2!}\tilde{g}''(u) + \cdots + \frac{(-hz)^{l-1}}{(l-1)!}\tilde{g}^{(l-1)}(u) + \frac{(-hz)^l}{l!}\tilde{g}^{(l)}(u-\theta hz).$$

With Assumption (K_l), we obtain

$$\|s - s_h\|^2_{\phi'_X} \leq \left(\int_{z\in\mathbb{R}} |K(z)|\frac{|hz|^l}{l!}\left\{\int_{u=0}^1 \left\{\tilde{g}^{(l)}(u-\theta hz) - \tilde{g}^{(l)}(u)\right\}^2 du\right\}^{1/2} dz\right)^2.$$

Since $\tilde{g}$ belongs to the Hölder space $\mathcal{H}(\beta, L)$,

$$\begin{aligned} \left[\int_{u=0}^1 \left\{\tilde{g}^{(l)}(u-\theta hz) - \tilde{g}^{(l)}(u)\right\}^2 du\right]^{1/2} &\leq \left[\int_{u=0}^1 L^2(\theta hu)^{2(\beta-l)}du\right]^{1/2}, \\ &= L|hz|^{\beta-l}, \end{aligned}$$

which enables us to conclude.

□

6.5.2. *Example E4.* Here, $\phi_X(A) = \mathbb{R}_+$. The idea is the same, but since we integrate over an unbounded subset, we choose a integrated remainding term in the Taylor formula:

$$\tilde{g}(u-hz) - \tilde{g}(u) = -hz\tilde{g}'(u) + \frac{(-hz)^2}{2!}\tilde{g}''(u) + \cdots + \frac{(-hz)^{l-1}}{(l-1)!}\tilde{g}^{(l-1)}(u) + \frac{(-hz)^l}{(l-1)!}\int_0^1 (1-\theta)^{l-1}\tilde{g}^{(l)}(u-\theta hz)d\theta.$$

The reasoning is then the same as in density estimation (see Tsybakov 2009 for details).

□

6.6. **Proof of Theorem 2.** The arguments and the sketch of the proof are exactly the same as those used to prove Theorem 1. We first obtain an equivalent of Inequality (15):

$$\mathbb{E}\left[\left\|\hat{\pi}_{\hat{h}_1,\hat{h}_2}-\pi\right\|_{f_X}^2\right] \leq 6\mathbb{E}\left[A^{(cd)}(h_1,h_2)\right]+6V^{(cd)}(h_1,h_2)+\frac{\|\mathbb{K}\|_{L^2(\mathbb{R}^2)}^2}{nh_1h_2}+3\|g_{h_1,h_2}^{(cd)}-g^{(cd)}\|_{L^2([0;1]\times B)}^2.$$

Similarly, we must bound $A^{(cd)}$: we use the splitting (17). Then, bounding $A^{(cd)}$ amounts to control the deviation of the following centered empirical process:

$$\begin{aligned}\nu_{n,h_1,h_2}^{(cd)}(t) &= \frac{1}{n}\sum_{i=1}^{n}\left\{\int_{(0;1)\times B} K_{h_1}^{(1)}\left(u-F_X(X_i)\right)K_{h_2}^{(2)}\left(y-Y_i\right)t(u,y)dudy-\right.\\ &\left.\mathbb{E}\left[\int_{(0;1)\times B} K_{h_1}^{(1)}\left(u-F_X(X_i)\right)K_{h_2}^{(2)}\left(y-Y_i\right)t(u,y)dudy\right]\right\}.\end{aligned}$$

Precisely, we apply the Talagrand Inequality (Lemma 3) with the following quantity:

$$M_1=\frac{\|\mathbb{K}\|_{L^2(\mathbb{R}^2)}}{\sqrt{h_1h_2}},\quad H^2=\frac{\|\mathbb{K}\|_{L^2(\mathbb{R}^2)}^2}{nh_1h_2},\quad v=\|g^{(cd)}\|_{L^\infty([0;1]\times A_2)}\|\mathbb{K}\|_{L^1(\mathbb{R}^2)}^2.$$

This proves that

$$\sum_{(h_1,h_2)\in\mathcal{H}_n}\mathbb{E}\left[\left(\sup_{t\in\bar{S}^{(cd)}(0,1)}\left(\nu_{n,h_1,h_2}^{(cd)}(t)\right)^2-\tilde{V}^{(cd)}(h_1,h_2)\right)_+\right]\leq\frac{C}{n},$$

which is the key point of the proof.

□

Acknowledgements

I would like to thank Fabienne Comte for a wealth of smart advice and carefully readings along this work.

References

Nathalie Akakpo and Cécile Durot. Histogram selection for possibly censored data. *Math. Methods Statist.*, 19(3):189–218, 2010.

Nathalie Akakpo and Claire Lacour. Inhomogeneous and anisotropic conditional density estimation from dependent data. *Electronic Journal of Statistics*, 5:1618–1653, 2011.

Anestis Antoniadis, Gérard Grégoire, and Pierre Vial. Random design wavelet curve smoothing. *Statist. Probab. Lett.*, 35(3):225–232, 1997.

Anestis Antoniadis, Gérard Grégoire, and Guy Nason. Density and hazard rate estimation for right-censored data by using wavelet methods. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 61(1): 63–84, 1999.

Yannick Baraud. Model selection for regression on a random design. *ESAIM Probab. Statist.*, 6: 127–146 (electronic), 2002.

Lucien Birgé. Interval censoring: a nonasymptotic point of view. *Math. Methods Statist.*, 8(3): 285–298, 1999.

Lucien Birgé. Model selection for Gaussian regression with random design. *Bernoulli*, 10(6): 1039–1051, 2004.

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

Elodie Brunel and Fabienne Comte. Adaptive estimation of hazard rate with censored data. *Comm. Statist. Theory Methods*, 37(8-10):1284–1305, 2008.

Elodie Brunel and Fabienne Comte. Cumulative distribution function estimation under interval censoring case 1. *Electron. J. Stat.*, 3:1–24, 2009.

Elodie Brunel, Fabienne Comte, and Claire Lacour. Adaptive estimation of the conditional density in the presence of censoring. *Sankhyā*, 69(4):734–763, 2007.

Gaëlle Chagny. Penalization versus goldenshluger-lepski strategies in warped bases regression. *ESAIM Probab. Statist.*, to appear (available online), 2011.

Gaëlle Chagny. Warped bases for conditional density estimation. *Submitted, hal-00641560, v2*, 2012.

Michaël Chichignoud. Minimax and minimax adaptive estimation in multiplicative regression : locally bayesian approach. *Prob. Theory and Relat. Fields*, to appear (available online), 2011a.

Michaël Chichignoud. Pointwise adaptive m-estimation in nonparametric regression. *Submitted, arXiv:1105.1646*, 2011b.

Serge Cohen and Erwan Le Pennec. Conditional density estimation by penalized likelihood model selection and applications. *Submitted arXiv:1103.2021*, 2011.

Fabienne Comte and Claire Lacour. Anisotropic adaptive kernel deconvolution. *Ann. Inst. H. Poincaré Probab. Statist.*, 2011.

Fabienne Comte and Yves Rozenholc. Adaptive estimation of mean and volatility functions in (auto-)regressive models. *Stochastic Process. Appl.*, 97(1):111–145, 2002.

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

Luc Devroye. The double kernel method in density estimation. *Ann. Inst. H. Poincaré Probab. Statist.*, 25(4):533–580, 1989.

Sam Efromovich. *Nonparametric curve estimation*. Methods, theory, and applications. Springer Series in Statistics. 1999. ISBN 0-387-98740-1.

Sam Efromovich. Oracle inequality for conditional density estimation and an actuarial example. *Ann. Inst. Statist. Math.*, 62(2):249–275, 2010.

Jianqing Fan and Irène Gijbels. Variable bandwidth and local linear regression smoothers. *Ann. Statist.*, 20(4):2008–2036, 1992.

Jianqing Fan and Tsz Ho Yim. A crossvalidation method for estimating conditional densities. *Biometrika*, 91(4):819–834, 2004.

Alexander Goldenshluger and Oleg Lepski. Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *Ann. Statist.*, 39(3):1608–1632, 2011.

GrigoriĭK K. Golubev and Michael Nussbaum. Adaptive spline estimates in a nonparametric regression model. *Teor. Veroyatnost. i Primenen.*, 37(3):554–561, 1992.

Piet Groeneboom. Nonparametric estimators for interval censoring problems. In *Analysis of censored data (Pune, 1994/1995)*, volume 27 of *IMS Lecture Notes Monogr. Ser.*, pages 105–128. Inst. Math. Statist., Hayward, CA, 1995.

Piet Groeneboom and Jon A. Wellner. *Information bounds and nonparametric maximum likelihood estimation*, volume 19 of *DMV Seminar*. Birkhäuser Verlag, Basel, 1992. ISBN 3-7643-2794-4.

Wolfgang Härdle and Alexander Tsybakov. Local polynomial estimators of the volatility function in nonparametric autoregression. *J. Econometrics*, 81(1):223–242, 1997.

Marc Hoffmann. On nonparametric estimation in nonlinear AR(1)-models. *Statist. Probab. Lett.*, 44(1):29–45, 1999.

Michael G. Hudgens, Marloes H. Maathuis, and Peter B. Gilbert. Nonparametric estimation of the joint distribution of a survival time subject to interval censoring and a continuous mark

variable. *Biometrics*, 63(2):372–380, 2007.

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

Rob J. Hyndman, David M. Bashtannyk, and Gary K. Grunwald. Estimating and visualizing conditional densities. *J. Comput. Graph. Statist.*, 5(4):315–336, 1996.

Nicholas P. Jewell and Mark van der Laan. Current status data: review, recent developments and open problems. In *Advances in survival analysis*, volume 23 of *Handbook of Statist.*, pages 625–642. Elsevier, Amsterdam, 2004.

Edward Kaplan and Paul Meier. Non parametric estimation from incomplete observations. *J Amer Statist Assoc*, 53(1):457–481, 1958.

Gérard Kerkyacharian and Dominique Picard. Regression in random design and warped wavelets. *Bernoulli*, 10(6):1053–1105, 2004.

Thierry Klein and Emmanuel Rio. Concentration around the mean for maxima of empirical processes. *Ann. Probab.*, 33(3):1060–1077, 2005.

Claire Lacour. Adaptive estimation of the transition density of a particular hidden Markov chain. *J. Multivariate Anal.*, 99(5):787–814, 2008.

Shuangge Ma and Michael R. Kosorok. Adaptive penalized M-estimation with current status data. *Ann. Inst. Statist. Math.*, 58(3):511–526, 2006.

K. L. Mehra, Y. S. Ramakrishnaiah, and P. Sashikala. Laws of iterated logarithm and related asymptotics for estimators of conditional density and mode. *Ann. Inst. Statist. Math.*, 52(4): 630–645, 2000.

Hans-Georg Müller and Jane-Ling Wang. Hazard rate estimation under random censoring with varying kernels and bandwidths. *Biometrics*, 50(1):61–76, 1994.

Elizbar Nadaraya. On estimating regression. *Theory of Probability and its Application*, 9(4): 141–142, 1964.

W. Nelson. Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14(4):945–966, 1972.

Michael H. Neumann. Fully data-driven nonparametric variance estimators. *Statistics*, 25(3): 189–212, 1994.

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

Prakash N. Patil. Bandwidth choice for nonparametric hazard rate estimation. *J. Statist. Plann. Inference*, 35(1):15–30, 1993.

Thanh Mai Pham Ngoc. Regression in random design and Bayesian warped wavelets estimators. *Electron. J. Stat.*, 3:1084–1112, 2009.

Sandra Plancade. Estimation of conditional cumulative distribution function from current status data. *Submitted arXiv:1110.5927*, 2011.

Patricia Reynaud-Bouret. Penalized projection estimators of the Aalen multiplicative intensity. *Bernoulli*, 12(4):633–661, 2006.

Winfried Stute. Asymptotic normality of nearest neighbor regression function estimates. *Ann. Statist.*, 12(3):917–926, 1984.

Winfried Stute. Conditional empirical processes. *Ann. Statist.*, 14(2):638–647, 1986.

Martin A. Tanner and Wing Hung Wong. The estimation of the hazard function from randomly censored data by the kernel method. *Ann. Statist.*, 11(3):989–993, 1983.

Alexandre B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009. ISBN 978-0-387-79051-0. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

Sara van de Geer. Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Ann. Statist.*, 21(1):14–44, 1993.

Geoffrey S. Watson. Smooth regression analysis. *Sankhyā Ser. A.*, 26(1):359–372, 1964.

Marten Wegkamp. Model selection in nonparametric regression. *Ann. Statist.*, 31(1):252–273, 2003.

Shie-Shien Yang. Linear functions of concomitants of order statistics with application to nonparametric estimation of a regression function. *J. Amer. Statist. Assoc.*, 76(375):658–662, 1981.