Multivariate adaptive warped kernel estimation

Gaëlle Chagny, Thomas Laloë, Rémi Servien

MULTIVARIATE ADAPTIVE WARPED KERNEL ESTIMATION

GAËLLE CHAGNY[(1)], THOMAS LALOË[(2)], AND RÉMI SERVIEN[(3)]

ABSTRACT. We deal with the problem of nonparametric estimation of a multivariate regression function without any assumption on the compacity of the support of the random design. To tackle the problem, we propose to extend a "warping" device to the multivariate framework. An adaptive warped kernel estimator is first defined in the case of known design distribution and proved to be optimal in the oracle sense. Then, a general procedure is carried out: the marginal distributions of the design are estimated by the empirical cumulative distribution functions, and the dependence structure is built using a kernel estimation of the copula density. The copula density estimator is also studied and proved to be optimal in the oracle and in the minimax sense. The plug-in of this estimator in the regression function estimator provides a fully data-driven procedure. A numerical study illustrates the theoretical results.

(1) gaelle.chagny@univ-rouen.fr, LMRS, Université de Rouen Normandie et CNRS, UMR 6085, France.

(2) laloe@unice.fr, Université de Nice Sophia-Antipolis, Laboratoire J-A Dieudonné, Parc Valrose, 06108 Nice Cedex 02, France.

(3) remi.servien@inra.fr, INRA-ENVT, Université de Toulouse, UMR1331 Toxalim, F-31027 Toulouse, France.

1. INTRODUCTION

Let $(\mathbf{X}, Y)$ be a couple of random variables taking values on $\mathbb{R}^d \times \mathbb{R}$ such that

$$Y = r(\mathbf{X}) + \varepsilon, \tag{1}$$

with ε a centered real random variable with finite variance independent of $\mathbf{X} = (X_1, \ldots, X_d)$. Assume that we have an independent identically distributed (*i.i.d.* in the sequel) sample $(\mathbf{X}_i, Y_i)_{i=1\ldots n}$ distributed as $(\mathbf{X}, Y)$. The subject of the paper is the estimation of the multivariate regression function $r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ on a subset $A \subset \mathbb{R}^d$, not necessarily bounded with a warping device described below, that also requires the estimation of the dependence structure between the coordinates of $\mathbf{X}$.

Regression estimation is a classical problem in statistics, addressed in a significant number of research works frequently based on non-parametric methods such as kernel estimators (Nadaraya, 1964; Watson, 1964), local polynomial estimators (Fan and Gijbels, 1996), orthogonal series or spline estimators (Golubev and Nussbaum, 1992; Antoniadis et al., 1997; Efromovich, 1999; Baraud, 2002) and nearest neighbour-type estimators (Stute, 1984; Guyader and Hengartner, 2013). Among kernel methods, the most popular estimator is the well-known Nadaraya-Watson estimator, defined for model (1) by

$$\widehat{r}^{NW}(\mathbf{x}) = \frac{\sum_{i=1}^n Y_i K_{\mathbf{h}}(\mathbf{x} - \mathbf{X}_i)}{\sum_{i=1}^n K_{\mathbf{h}}(\mathbf{x} - \mathbf{X}_i)}, \tag{2}$$

Date: January 28, 2019.

where $\mathbf{h} = {}^t(h_1, \dots, h_d)$ is the so-called bandwidth of the kernel K, $K_{\mathbf{h}}(\mathbf{x}) = K_{1,h_1}(x_1)K_{2,h_2}(x_2)\dots K_{d,h_d}(x_d)$, with $K_{l,h_l}(x) = K_l(x/h_l)/h_l$ for $h_l > 0$, and $K_l : \mathbb{R} \to \mathbb{R}$ such that $\int_{\mathbb{R}} K_l(x)dx = 1$, $l = 1, \dots, d$.

A commonly shared assumption for regression analysis is that the support of $\mathbf{X}$ is a compact subset of $\mathbb{R}^d$ (Györfi et al., 2002; Guyader and Hengartner, 2013; Furer and Kohler, 2015). It could be very restrictive in some situations such as for example the estimation of the regression function on the level sets of the cumulative distribution function (c.d.f.) (Di Bernardino et al., 2015). Stone (1982) first conjecture that this assumption could be weakened. To our knowledge, Kohler et al. (2009) are the first who propose theoretical results with no boundedness assumption on the support of the design. The price to pay is to make a moment assumption on the design $\mathbf{X}$ (see Assumption (A4) in Kohler et al. 2009).

So far, "warped" estimators have been developed (Yang, 1981; Kerkyacharian and Picard, 2004) and require very few assumptions on the support of $\mathbf{X}$. If we assume, in a sake of clarity, that $d = 1$, the warped method is based on the introduction of the auxiliary function $g = r \circ F_{\mathbf{X}}^{-1}$, where $F_{\mathbf{X}} : x \in \mathbb{R} \mapsto \mathbb{P}(\mathbf{X} \leq x)$ is the c.d.f. of the design $\mathbf{X}$. First, an estimator $\hat{g}$ is proposed for g, and then, the regression r is estimated using $\hat{g} \circ \hat{F}$, where $\hat{F}$ is the empirical c.d.f. of $\mathbf{X}$. This strategy has already been applied in the regression setting using projection methods (Kerkyacharian and Picard, 2004; Pham Ngoc, 2009; Chagny, 2013) but also for other estimation problems (conditional density estimation, hazard rate estimation based on randomly right-censored data, and c.d.f. estimation from current-status data, see *e.g.* Chesneau and Willer 2015; Chagny 2015). If the warping device permits to weaken the assumptions on the design support, the warped estimator also depend on a unique bandwidth, for $d = 1$, whereas the ratio form of the kernel estimator (2) requires the selection of two smoothing parameters (one for the numerator, one for the denominator). In return, the c.d.f. $F_{\mathbf{X}}$ of $\mathbf{X}$ has to be estimated, but this can simply be done using its empirical counterpart. This does not deteriorate the optimal convergence rate, since the empirical c.d.f. converges at a parametric rate. A data-driven selection of the unique bandwidth involved in the resulting warped kernel estimator, in the spirit of Goldenshluger and Lepski (2011) leads to nonasymptotic risk bounds when $d = 1$ (Chagny, 2015). To our knowledge, this adaptive estimation has never been carried out for a ratio regression estimator, the only reference on this subject being Ngoc Bien (2014) who assumes that the design $\mathbf{X}$ has a known uniform distribution.

Nevertheless, the extension of the warped strategy to the multivariate framework is not trivial, and we propose to deal with this problem here. The key question is to take into account the dependence between the multivariate components of each $\mathbf{X}_i$. We propose to tackle this problem by using copulas, that permit to describe the dependence structure between random variables (Sklar, 1959; Jaworski et al., 2010). The price to pay is the additional estimation of the copula density of the design : the complete strategy requires the plug-in of such estimate in the final warped regression estimator. The results are obtained for random design distribution with possibly unbounded support, like in Kohler et al. (2009). However, we will see that the assumptions are not exactly the same : in particular, the warping device permits to avoid the moment conditions on X. Moreover, our results takes place in the field of nonasymptotic adaptive estimation, and the bandwidth of the kernel estimator we propose does not depend on the smoothness index of the target function, contrary to the one of Kohler et al. (2009).

The paper is thus organized as follows. We explain in Section 2 how to extend the Yang (1981) estimator to the multivariate design framework. For sake of clarity, we first concentrate on the simple toy case of known design distribution (Section 3): under mild assumptions, we

derive (i) a non-asymptotic oracle type inequality for an integrated criterion for a warped kernel estimator with a data-driven bandwidth selected with a Lepski-type method, and (ii) an optimal convergence rate over possibly anisotropic functional classes (Neumann, 2000; Kerkyacharian et al., 2001; Bertin, 2005). Then, a kernel copula estimate that also adapts automatically to the unknown smoothness of the design is exhibited and studied in Section 4. An oracle type inequality is also proved. Finally, warped regression estimation with unknown copula density is the subject of Section 5: as expected, the risk of the final estimate depends on the risks of both the copula estimator and the regression estimator with known design density. A simulation study is carried out in Section 6. Concluding remarks as well as perspectives for future works are given in Section 7 and all the proofs are gathered in Section 8. Throughout the paper, we pay a special attention to compare assumptions, methodology and results to the one of Kohler et al. (2009).

2. Multivariate warping strategy

If $d = 1$, the warping device is based on the transformation $F_{\mathbf{X}}(X_i)$ of the data X_i, $i = 1, \dots, n$. For $d > 1$, a natural extension is to use $F_l(X_{l,i})$, for $l = 1, \dots, d$ and $i = 1, \dots, n$, where F_l is the marginal c.d.f. of X_l. Let us introduce $\widetilde{F}_{\mathbf{X}} : \mathbf{x} = (x_l)_{l=1,\dots,d} \in \mathbb{R}^d \mapsto (F_1(x_1), \dots, F_d(x_d))$. Assume that $\widetilde{F}_{\mathbf{X}}^{-1} : \mathbf{u} \in [0,1]^d \mapsto (F_1^{-1}(u_1), \dots, F_d^{-1}(u_d))$ exists, and let

$$g = r \circ \widetilde{F}_{\mathbf{X}}^{-1},$$

in such a way that $r = g \circ \widetilde{F}_{\mathbf{X}}$. If we consider that the marginal variables X_l of $\mathbf{X}$ are independent, the estimator of Yang (1981) can immediately be adapted to the multivariate setting : we set

$$\hat{g}_{\perp} : \mathbf{u} \mapsto \sum_{i=1}^{n} Y_i K_{\mathbf{h}}(\mathbf{u} - \widetilde{F}_{\mathbf{X}}(\mathbf{X}_i)) \tag{3}$$

to estimate g, and it remains to compound by the empirical counterpart of $\widetilde{F}_{\mathbf{X}}$ to estimate r. However, a dependence between the coordinates $X_{l,i}$ of $\mathbf{X}_i$ generally appears. The usual model for this dependence using a copula C and the c.d.f $F_{\mathbf{X}}$ of $\mathbf{X}$ can be written

$$F_{\mathbf{X}}(\mathbf{x}) = C(F_1(x_1), \dots, F_d(x_d)) = C(\widetilde{F}_{\mathbf{X}}(\mathbf{x})). \tag{4}$$

Denoting the copula density by c, we have

$$c(\mathbf{u}) = \frac{\partial^d C}{\partial u_1 \dots \partial u_d}(\mathbf{u}), \quad \mathbf{u} \in [0;1]^d,$$

and the density $f_{\mathbf{X}}$ of $\mathbf{X}$ can be expressed as

$$f_{\mathbf{X}}(\mathbf{x}) = c(\widetilde{F}_{\mathbf{X}}(\mathbf{x})) \prod_{l=1}^{d} f_l(x_l), \quad \mathbf{x} = (x_l)_{l=1,\dots,d} \in \mathbb{R}^d,$$

where $(f_l)_{l=1,\dots,d}$ are the marginal densities of $\mathbf{X} = (X_1, \dots, X_d)$. It can then be proved that the previous estimator (3) estimates cg and not g (see the computation (8) below). As a consequence, we propose to set, as an estimator for g,

$$\widehat{g}_{\mathbf{h},\mathbf{b},\widehat{F}}(\mathbf{u}) = \frac{1}{n\widehat{c}_{\mathbf{b}}(\mathbf{u})} \sum_{i=1}^{n} Y_i K_{\mathbf{h}}(\mathbf{u} - \widehat{\widetilde{F}}_{\mathbf{X}}(\mathbf{X}_i)), \quad \mathbf{u} \in \widetilde{F}_{\mathbf{X}}(A),$$

where $\widehat{c}_{\mathbf{b}}$ is a kernel estimator of the copula density that will be defined later (see Section 4). We denote by $\widehat{\widetilde{F}}_{\mathbf{X}} : \mathbb{R}^d \to [0;1]^d$ the empirical multivariate marginal c.d.f.:

$$\widehat{\widetilde{F}}_{\mathbf{X}} = (\widehat{\widetilde{F}}_{\mathbf{X},1}, \dots, \widehat{\widetilde{F}}_{\mathbf{X},d}),\ \widehat{\widetilde{F}}_{\mathbf{X},l}(x_l) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{X_{l,i}\leq x_l},\ x_l \in \mathbb{R}, l \in \{1,\dots,d\}, \tag{5}$$

and finally set

$$\widehat{r}_{\mathbf{h},\mathbf{b},\widehat{F}}(\mathbf{x}) = \widehat{g}_{\mathbf{h},\mathbf{b},\widehat{F}} \circ \widehat{\widetilde{F}}_{\mathbf{X}}(\mathbf{x}) = \frac{1}{n\widehat{c}_{\mathbf{b}}(\widehat{\widetilde{F}}_{\mathbf{X}}(\mathbf{x}))} \sum_{i=1}^{n} Y_i K_{\mathbf{h}}(\widehat{\widetilde{F}}_{\mathbf{X}}(\mathbf{x}) - \widehat{\widetilde{F}}_{\mathbf{X}}(\mathbf{X}_i)), \quad \mathbf{x} \in A, \tag{6}$$

to rebuild our target function r from the data. In the sequel, we denote by $\|\cdot\|_{L^p(\Theta)}$ the classical L^p-norm on a set Θ.

3. The simple case of known design distribution

3.1. **Collection of kernel estimators.** For sake of clarity, we first consider the regression estimation problem with a known design distribution. In this section, the copula density c and the marginal c.d.f. $\widetilde{F}_{\mathbf{X}}$ are consequently considered to be known. Thus, (6) becomes

$$\widehat{r}_{\mathbf{h}}(\mathbf{x}) = \widehat{g}_{\mathbf{h}} \circ \widetilde{F}_{\mathbf{X}}(\mathbf{x}) = \frac{1}{nc(\widetilde{F}_{\mathbf{X}}(\mathbf{x}))} \sum_{i=1}^{n} Y_i K_{\mathbf{h}}(\widetilde{F}_{\mathbf{X}}(\mathbf{x}) - \widetilde{F}_{\mathbf{X}}(\mathbf{X}_i)), \quad \mathbf{x} \in A, \tag{7}$$

where we denote $\widehat{g}_{\mathbf{h}}(\mathbf{u}) = \sum_{i=1}^{n} Y_i K_{\mathbf{h}}(\mathbf{u} - \widetilde{F}_{\mathbf{X}}(\mathbf{X}_i))/(nc(\mathbf{u}))$, $\mathbf{u} \in [0,1]^d$. The following computation enlights the definitions (6) and (7) above. For any $\mathbf{u} \in \widetilde{F}_{\mathbf{X}}(A)$,

$$\begin{aligned}
\mathbb{E}[\widehat{g}_{\mathbf{h}}(\mathbf{u})] &= \mathbb{E}\left[\frac{Y K_{\mathbf{h}}(\mathbf{u} - \widetilde{F}_{\mathbf{X}}(\mathbf{X}))}{c(\mathbf{u})}\right], \\
&= \mathbb{E}\left[\frac{r(\mathbf{X}) K_{\mathbf{h}}(\mathbf{u} - \widetilde{F}_{\mathbf{X}}(\mathbf{X}))}{c(\mathbf{u})}\right], \\
&= \frac{1}{c(\mathbf{u})} \int_{\mathbb{R}^d} r(\mathbf{x}) K_{\mathbf{h}}(\mathbf{u} - \widetilde{F}_{\mathbf{X}}(\mathbf{x})) c(\widetilde{F}_{\mathbf{X}}(\mathbf{x})) \prod_{l=1}^{d} f_l(x_l) d\mathbf{x}, \\
&= \frac{1}{c(\mathbf{u})} \int_{[0,1]^d} g(\mathbf{u}') K_{\mathbf{h}}(\mathbf{u} - \mathbf{u}') c(\mathbf{u}') d\mathbf{u}', \\
&= \frac{K_{\mathbf{h}} \star (cg\mathbf{1}_{[0,1]^d})}{c}(\mathbf{u}).
\end{aligned} \tag{8}$$

where $\star$ is the convolution product. For small $\mathbf{h}$, the convolution product $K_{\mathbf{h}} \star (cg)\mathbf{1}_{[0,1]^d}$ is supposed to be closed to cg : this justifies that $\widehat{g}_{\mathbf{h}}$ is suitable to estimate g, and $\widehat{r}_{\mathbf{h}}$ suits well to recover the target r.

3.2. **Risk of the estimator with fixed bandwidth.** As in Kohler et al. (2009), we consider a global weighted integrated risk criterion, to study the properties of our estimator. Let $\|.\|_{f_{\mathbf{X}}}$ be the classical L^2-norm on the space of squared integrable functions with respect to the Lebesgue measure weighted by $f_{\mathbf{X}}$ on A: for any function t in this space,

$$\|t\|_{f_{\mathbf{X}}}^2 = \int_A t^2(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} = \int_{\widetilde{F}_{\mathbf{X}}(A)} t^2 \circ \widetilde{F}_{\mathbf{X}}^{-1}(\mathbf{u}) c(\mathbf{u}) d\mathbf{u}.$$

The mean integrated squared risk of the estimator $\widehat{r}_{\mathbf{h}}$ can thus be written

$$\mathbf{E}[\|\widehat{r}_{\mathbf{h}} - r\|^2_{f_{\mathbf{X}}}] = \mathbf{E}\left[\int_A \left(\widehat{r}_{\mathbf{h}}(\mathbf{x}) - r(\mathbf{x})\right)^2 f_{\mathbf{X}}(\mathbf{x})d\mathbf{x}\right] = \mathbf{E}\left[\int_{\widetilde{F}_{\mathbf{X}}(A)} \left(\widehat{g}_{\mathbf{h}}(\mathbf{u}) - g(\mathbf{u})\right)^2 c(\mathbf{u})d\mathbf{u}\right]$$

and, using a classical bias-variance decomposition, we have $\mathbf{E}[\|\widehat{r}_{\mathbf{h}} - r\|^2_{f_{\mathbf{X}}}] = B(\mathbf{h}) + V(\mathbf{h})$, where

$$\begin{aligned} B(\mathbf{h}) &= \int_{\widetilde{F}_{\mathbf{X}}(A)} c(\mathbf{u}) \left(\frac{K_{\mathbf{h}} \star (cg\mathbf{1}_{[0,1]^d})}{c}(\mathbf{u}) - g(\mathbf{u})\right)^2 d\mathbf{u}, \\ V(\mathbf{h}) &= \int_{\widetilde{F}_{\mathbf{X}}(A)} c(\mathbf{u}) \left(\widehat{g}_{\mathbf{h}}(\mathbf{u}) - \frac{K_{\mathbf{h}} \star (cg\mathbf{1}_{[0,1]^d})}{c}(\mathbf{u})\right)^2 d\mathbf{u}. \end{aligned} \tag{9}$$

To obtain upper-bounds for these two terms, we introduce the following assumptions.

$(H_{cg,\beta})$: The function $(cg)\mathbf{1}_{\widetilde{F}_{\mathbf{X}}(A)}$ belongs to an anisotropic Nikol'skiĭ ball $\mathcal{N}_2(\beta, L)$, with $L > 0$ and $\beta = (\beta_1, \ldots, \beta_d) \in (\mathbb{R}_+^*)^d$ (Nikol′skiĭ, 1975). This is the set of functions $f : \mathbb{R}^d \mapsto \mathbb{R}$ such that f admits derivatives with respect to x_l up to the order $\lfloor \beta_l \rfloor$ (where $\lfloor \beta_l \rfloor$ denotes the largest integer less than β_l), and

(i) for all $l \in \{1, \ldots, d\}$, $\|\partial^{\lfloor \beta_l \rfloor} f / (\partial x_l)^{\lfloor \beta_l \rfloor}\|_{L^2(\mathbb{R}^d)} \leq L$,

(ii) for all $l \in \{1, \ldots, d\}$ and $t \in \mathbb{R}$,

$$\int_{\mathbb{R}^d} \left| \frac{\partial^{\lfloor \beta_l \rfloor} f}{(\partial x_l)^{\lfloor \beta_l \rfloor}}(x_1, \ldots, x_{l-1}, x_l + t, x_{l+1}, \ldots, x_d) - \frac{\partial^{\lfloor \beta_l \rfloor} f}{(\partial x_l)^{\lfloor \beta_l \rfloor}}(\mathbf{x}) \right|^2 d\mathbf{x} \leq L^2 |t|^{2(\beta_l - \lfloor \beta_l \rfloor)}.$$

$(H_{K,\ell})$: The kernel K is of order $\ell \in (R_+)^d$, *i.e.*

(i) $\forall l \in \{1, \ldots, d\}$, $\forall k \in \{1, \ldots, \ell_l\}$, $\int_{\mathbb{R}^d} x_l^k K(\mathbf{x})d\mathbf{x} = 0$.

(ii) $\forall l \in \{1, \ldots, d\}$, $\int_{\mathbb{R}^d} (1 + x_l)^{\ell_l} K(\mathbf{x})d\mathbf{x} < \infty$.

$(H_{c,low})$: The copula density is lower bounded: $\exists m_c > 0$, $\forall \mathbf{u} \in \widetilde{F}_{\mathbf{X}}(A)$, $c(\mathbf{u}) \geq m_c$.

Assumptions $(H_{cg,\beta})$ and $(H_{K,\ell})$ are classical for nonparametric multivariate kernel estimation (Goldenshluger and Lepski, 2011; Comte and Lacour, 2013) and permit to control the bias term of the risk $B(\mathbf{h})$. Assumption $(H_{K,\ell})$ is not restrictive since a wide range of kernels could be chosen, and an assumption on the support and the bounds of the kernel is also necessary for Kohler et al. (2009) (see equation (7) of their paper). In $(H_{cg,\beta})$, the index β measures the smoothness of the function cg, and allows us to deal with possibly anisotropic regression functions (different smoothness according to the different direction can be considered), like assumption (A2) in Kohler et al. (2009). However, since they consider a pointwise criterion (local bandwidth choice), they rather choose Hölder spaces instead of Nikol'skiĭ spaces, which are designed for integrated risks (like our L^2-risk, see e.g. Tsybakov 2009) and global bandwidth selection purpose. A second difference lies in the use of this assumption. Although we assume that $(H_{cg,\beta})$ holds in the sequel, we will not assume the smoothness index β to be known. Its value is not required to compute our selected bandwidth (see Section 3.3), while Kohler et al. (2009) use it to choose the bandwidth of their kernel estimate (see equation (8) in their paper). The difficulty of $(H_{cg,\beta})$ is that this smoothness assumption is made directly on cg, and not on the targeted function r. It is for example satisfied if the two functions c and g separately belong to $\mathcal{N}_2(\beta, L')$ ($L' > 0$), for β such that each $\beta_l \leq 1$, $l \in \{1, \ldots, d\}$. The fact that the assumption is carried by the auxiliary function g and not r is classical in warped methods (Pham Ngoc, 2009; Chagny, 2015). Another solution is to consider weighted spaces: lots of

details can be found in Kerkyacharian and Picard (2004). Assumption $(H_{c,low})$ is specific to the warped method, which makes appear the copula density in the formula of the estimator. It is replaced by other assumptions in Kohler et al. (2009), see comments following Corollary 3.2. On $[0,1]^d$ (case of $A=\mathbb{R}^d$), it is verified for example for the Farlie-Gumbel-Morgenstern copula, for the Ali-Mikhail-Hacq copula with a parameter $\theta \in]-1,1[$ (Balakrishnan and Lai, 2009), or for the copula density of a design with independent marginals. For other copulas, it is possible to restrict the estimation set A to exclude problematic points : for example, for $d=2$, the points $(0,1)$, $(0,0)$ and $(1,0)$ are generally the ones which makes $(H_{c,low})$ not true. The choice $A=]\varepsilon,+\infty[^d$, for a fixed $\varepsilon>0$, (although still uncompact) sometimes permit to avoid the problem, and thus to consider other copula densities (the example $A\subset(\mathbb{R}_+)^d$ is related to the application of our method to level set estimation, see Di Bernardino et al. 2015). For example, for the Gumbel Type I bivariate with a parameter $\theta=1$ or the Frank copula, it is possible to choose $A=]\varepsilon,+\infty[^2$ for the case of nonnegative variables $(X_i)_i$. The proof of the following result can be found at Section 8.1.

Proposition 3.1. *Assume $(H_{c,low})$, $(H_{cg,\beta})$ and $(H_{K,\ell})$ for an index $\ell\in\mathbb{R}_+^d$ such that $\ell_j\geq\lfloor\beta_j\rfloor$. Then,*

$$\mathbf{E}[\|\widehat{r}_{\mathbf{h}}-r\|^2_{f_{\mathbf{X}}}]\leq\frac{1}{m_c}\left(L\sum_{l=1}^d h_l^{2\beta_l}+\|K\|^2_{L^2(\mathbb{R}^d)}\mathbb{E}[Y_1^2]\frac{1}{nh_1\dots h_d}\right).$$

This is a nonasymptotic bias-variance upper bound for the quadratic risk. The first term of the right-hand-side of the inequality of Proposition 3.1 is an upper-bound for the bias term $B(\mathbf{h})$. The second one bounds the variance term. Another choice would have been to kept $B(\mathbf{h})$ in the inequality (in this case, Assumptions $(H_{cg,\beta})$ and $(H_{K,\ell})$ are not required). Our choice permits to immediately deduce the following convergence rate, by computing the bandwidth that minimizes the right-hand-side of the inequality of Proposition 3.1, over all possible bandwidths $\mathbf{h}\in(\mathbb{R}_+^*)^d$ (see a brief proof in Section 8.2).

Corollary 3.1. *Under the same assumptions as Proposition 3.1, there exists a bandwidth $\mathbf{h}(\beta)$ such that*

$$\mathbf{E}[\|\widehat{r}_{\mathbf{h}(\beta)}-r\|^2_{f_{\mathbf{X}}}]=O\left(n^{-\frac{2\bar{\beta}}{2\bar{\beta}+d}}\right),$$

where $\bar{\beta}$ is the harmonic mean of $\beta_1,\dots,\beta_d$: $d\bar{\beta}^{-1}=\beta_1^{-1}+\dots+\beta_d^{-1}$.

Thus the usual convergence rate in multivariate nonparametric estimation can be achieved by our estimator, provided that its bandwidth is carefully chosen. Here, the bandwidth $\mathbf{h}(\beta)$ that minimizes the upper-bound of the inequality of Proposition 3.1 depends on the smoothness index β of the unknown function cg. This smoothness index is also unknown *a priori*. The challenge of adaptive estimation is to propose a data-driven choice that also leads to an estimator with the same optimal convergence rate.

3.3. **Estimator selection.** Let $\mathcal{H}_n\subset(\mathbb{R}_+^*)^d$ a finite bandwidth collection. We set

$$\widehat{B}(\mathbf{h})=\max_{\mathbf{h}'\in\mathcal{H}_n}\left\{\left\|\frac{K_{\mathbf{h}}\star(\widehat{cg}_{\mathbf{h}'}\mathbf{1}_{\widetilde{F}_{\mathbf{X}}(A)})}{c}\circ\widetilde{F}_{\mathbf{X}}-\widehat{r}_{\mathbf{h}',c}\right\|^2_{f_{\mathbf{X}}}-\widehat{V}(\mathbf{h}')\right\}_+ \tag{10}$$

with

$$\widehat{V}(\mathbf{h})=\kappa\frac{\sum_{i=1}^n Y_i^2}{\widehat{m}_c}\frac{1}{nh_1\dots h_d}, \tag{11}$$

where $\kappa > 0$ is a tuning constant and $\widehat{m}_c$ an estimator for m_c. We define

$$\widehat{\mathbf{h}} = \arg\min_{\mathbf{h}\in\mathcal{H}_n}\{\widehat{B}(\mathbf{h}) + \widehat{V}(\mathbf{h})\}, \tag{12}$$

and the final estimator $\widehat{r}_{\widehat{\mathbf{h}}}$. The criterion (12), inspired from Goldenshluger and Lepski (2011), is known to mimic the optimal "bias-variance" trade-off that has to be realized in a data-driven way. A short heuristic about the definition of the construction of the criterion could be found at the beginning of Section 8.3. It is a global criterion : we select the same bandwidth, whatever the estimation point is. This is one difference with Kohler et al. (2009), who propose local choices. Another difference is that our choice does not depend on the smoothness index of the target function.

We also introduce $\widetilde{\mathbf{h}} = \arg\min_{\mathbf{h}\in\mathcal{H}_n}\{\widetilde{B}(\mathbf{h}) + \widetilde{V}(\mathbf{h})\}$ with

$$\widetilde{B}(\mathbf{h}) = \max_{\mathbf{h}'\in\mathcal{H}_n}\left\{\left\|\frac{K_{\mathbf{h}} \star (c\widehat{g}_{\mathbf{h}'}\mathbf{1}_{\widetilde{F}_{\mathbf{X}}(A)})}{c} \circ \widetilde{F}_{\mathbf{X}} - \widehat{r}_{\mathbf{h}'}\right\|^2_{f_{\mathbf{X}}} - \widetilde{V}(\mathbf{h}')\right\}_+$$

and

$$\widetilde{V}(\mathbf{h}) = \kappa_0\frac{\mathbb{E}[Y_1^2]}{m_c}\frac{1}{nh_1\dots h_d}, \;\; \kappa_0 > 0.$$

We start with the study of the estimator $\widehat{r}_{\widetilde{h},c}$. The collection $\mathcal{H}_n$ is chosen such that

$$\begin{aligned} &\exists\alpha_0 > 0, \kappa_1 > 0, \sum_{\mathbf{h}\in\mathcal{H}_n}\frac{1}{h_1\cdots h_d} \leq \kappa_1 n^{\alpha_0} \\ &\text{and } \forall\kappa_1 > 0, \exists C_0 > 0, \sum_{\mathbf{h}\in\mathcal{H}_n}\exp\left(-\frac{\kappa_1}{h_1\cdots h_d}\right) \leq C_0. \end{aligned} \tag{13}$$

These assumptions are very common to derive such estimators (Comte and Lacour, 2013; Chagny, 2015). For example, $\mathcal{H}_n = \{k_1^{-1}\cdots k_d^{-1}, k_l \in \{1,\dots,\lfloor n^{1/r}\rfloor\}, l = 1,\dots,d\}$ satisfies them with $\alpha_0 = 2d/r$.

We also introduce additional assumptions:

- **(H_ε):** The noise ε is $p+2$ integrable, for some $p > 2\alpha_0$: $\mathbb{E}[|\varepsilon|^{2+p}] < \infty$.
- **$(H_{c,high})$:** The copula density is upper-bounded over $\widetilde{F}_{\mathbf{X}}(A)$: $\exists M_C > 0,\ \forall\mathbf{u} \in \widetilde{F}_{\mathbf{X}}(A),\ c(\mathbf{u}) \leq M_C$.

The assumption $(H_{c,high})$ is quite restrictive for copula density estimation if $A = \mathbb{R}^d$ (ie. $\widetilde{F}_{\mathbf{X}}(A) = [0,1]^d$). However, it is also required for copula density estimation (see Section 4), and it is classical for adaptive density estimation purpose. Moreover, the same upper-bound is assumed in Autin et al. (2010) on $[0,1]^d$. Assumption $(H_{c,high})$ is for example satisfied by the Frank copula density, the Farlie-Gumbel-Morgenstern copula, the copula density of a design with independent marginals... Assumption (H_ε) is classical in adaptive regression estimation, see e.g. Baraud (2002) and Chagny (2015). It is then possible to set the following upper bound, proved Section 8.3.

Theorem 3.1. *Assume that $\mathcal{H}_n$ satisfies* (13) *and assume also* (H_ε), $(H_{c,low})$ *and* $(H_{c,high})$. *Then there exist two constants c_1 et c_2 such that*

$$\mathbf{E}[\|\widehat{r}_{\widetilde{\mathbf{h}}} - r\|^2_{f_{\mathbf{X}}}] \leq c_1 \min_{\mathbf{h}\in\mathcal{H}_n} \left\{ \frac{1+\|K\|^2_{L^1([0,1]^d)}}{m_c} \left\| K_{\mathbf{h}} \star (cg\mathbf{1}_{\widetilde{F}_{\mathbf{X}}(A)}) - cg \right\|^2_{L^2(\widetilde{F}_{\mathbf{X}}(A))} + \|K\|^2_{L^2(\mathbb{R}^d)} \mathbb{E}[Y_1^2] \frac{1}{n m_c h_1 \dots h_d} \right\} + \frac{c_2}{n}.$$

This result is an oracle-type inequality which assesses that the selected estimator performs as well as the best estimator of the collection $(\widehat{r}_{\mathbf{h}})_{\mathbf{h}\in\mathcal{H}_n}$, up to multiplicative constants and a remainder term: it achieves the best bias-variance trade-off (see Proposition 3.1). No smoothness assumption is required to establish the result. If we add Assumptions $(H_{cg,\beta})$ and $(H_{K,\ell})$ (for an index $\ell \in \mathbb{R}^d_+$ such that $\ell_j \geq \lfloor \beta_j \rfloor$, $j = 1, \dots, d$) to the assumptions of Theorem 3.1, we obtain the same convergence rate as the one of Corollary 3.1 for the estimator $\widehat{r}_{\widetilde{\mathbf{h}}}$.

Corollary 3.2. *Under the same assumptions as Theorem 3.1, if we also assume that* $(H_{cg,\beta})$ *and* $(H_{K,\ell})$ *are fulfilled for an index $\ell \in \mathbb{R}^d_+$ such that $\ell_j \geq \lfloor \beta_j \rfloor$, we have*

$$\mathbf{E}[\|\widehat{r}_{\widetilde{\mathbf{h}}} - r\|^2_{f_X}] = O\left(n^{-\frac{2\bar{\beta}}{2\bar{\beta}+d}}\right),$$

where $\bar{\beta}$ is the harmonic mean of $\beta_1, \dots, \beta_d$: $d\bar{\beta}^{-1} = \beta_1^{-1} + \dots + \beta_d^{-1}$.

This result can be compared to Theorem 1 of Kohler et al. (2009): our estimate achieves the same convergence rate as their kernel estimate. As already indicated, the smoothness assumptions for the two results are similar. The main difference is that we do not need to know the smoothness index of the targeted function to compute our estimator, while they have to. This is what makes our result adaptive. The other assumptions to establish the respective results are specific to the chosen methodology : our assumptions $(H_{c,low})$ and $(H_{c,high})$ on the copula density are specific to the extension of the warping device to the multivariate setting, and permit to deal with unbounded support for the design. They are replaced by a moment assumption on the design, and a boundness assumption on the regression function in Kohler et al. (2009).

Notice that Theorem 3.1 and Corollary 3.2 cover the case of the estimator $\widehat{r}_{\widetilde{\mathbf{h}}}$, whose bandwidth $\widetilde{\mathbf{h}}$ is defined with a variance term that involves the unknown quantities $\mathbb{E}[Y_1^2]$ and m_c. We choose not to present the final results : to switch from $\widehat{r}_{\widetilde{\mathbf{h}}}$ to $\widehat{r}_{\widehat{\mathbf{h}}}$ it remains to replace the unknown expectation $\mathbb{E}[Y_1^2]$ by its empirical counterpart $\frac{1}{n}\sum_{i=1}^n Y_i^2$ and to change $\widetilde{V}(\mathbf{h})$ in $\widehat{V}(\mathbf{h})$. This is quite classical, and can be done for example like in Theorem 3.4 p.465 of Brunel and Comte (2005). It is more unusual to replace the lower bound for the copula m_c by an estimate $\widehat{m}_c$: this can nevertheless be done thanks to cumbersome computations, following for example the proof of Theorem 4.1 of Chagny et al. (2017). The oracle-type inequality that will be obtained is exactly the same as the one of Theorem 3.1, but will be valid only for a sample size n large enough. The convergence rate of Corollary 3.2 is unchanged. We do not go into details, to avoid burdening the text by adding two similar results and to avoid lengthening the proofs.

4. Copula density estimation

The estimator defined by (6) involves an estimator of the copula density c that was assumed to be known in the previous section, on $\widetilde{F}_{\mathbf{X}}(A)$. This section is devoted to the question of copula density estimation. Since it is an interesting question by itself, to be more general we perform the estimation on $[0,1]^d$ and not on $\widetilde{F}_{\mathbf{X}}(A)$, like in other papers that deal with copula density

estimation (Fermanian, 2005; Autin et al., 2010). However the results are the same if we restrict the risk, all L^p-norms involved in the method and the validity of the assumptions to $\widetilde{F}_{\mathbf{X}}(A)$.

An adaptive estimator based on wavelets is defined in Autin et al. (2010) but, to be consistent with the previous kernel regression estimator already chosen, we propose to use the kernel estimator defined by Fermanian (2005). Consider $\mathbf{b} = {}^t(b_1, \dots, b_d) \in (\mathbb{R}_+^*)^d$ a multivariate bandwidth, a kernel $W_{\mathbf{b}}(\mathbf{u}) = W_{1,b_1}(u_1)W_{2,b_2}(u_2)\dots W_{d,b_d}(u_d)$, with $W_{l,b_l}(u) = W_l(u/b_l)/b_l$ for $b_l > 0$, and $W_l : \mathbb{R} \to \mathbb{R}$ such that $\int_0^1 W_l(u)du = 1$, $l \in \{1, \dots, d\}$. Let us introduce

$$\widehat{c}_{\mathbf{b}}(\mathbf{u}) = \frac{1}{n}\sum_{i=1}^{n} W_{\mathbf{b}}(\mathbf{u} - \widehat{\widetilde{F}}_{\mathbf{X}}(\mathbf{X}_i)), \quad \mathbf{u} \in [0,1]. \tag{14}$$

The estimator is very close to the classical kernel density estimator, up to the warping of the data through the empirical c.d.f. Remark that if we replace the estimator $\widehat{\widetilde{F}}_{\mathbf{X}}$ in (14) by its target $\widetilde{F}_{\mathbf{X}}$, like in the previous section, then $\widehat{c}_{\mathbf{b}}(\mathbf{u})$ is the density estimator of the random vector $(F_1(X_1), \dots, F_d(X_d))$, with uniformly distributed marginal distributions. We easily obtain the following upper-bound for the risk of the copula density estimator when the marginal distributions are known:

$$\mathbb{E}\left[\|\widehat{c}_{\mathbf{b}} - c\|^2_{L^2([0,1]^d)}\right] \leq \|W_{\mathbf{b}} \star c - c\|^2_{L^2([0,1]^d)} + \frac{\|W\|^2_{L^2([0,1]^d)}}{nb_1\cdots b_d}. \tag{15}$$

The results of Fermanian (2005) are asymptotic. Since our goal is to prove nonasymptotic adaptive upper-bounds, the Goldenshluger-Lepski method allows us to select a bandwidth $\widehat{\mathbf{b}}$ among a finite collection $\mathcal{B}_n \subset (\mathbb{R}_+^*)^d$. The collection $\mathcal{B}_n$ should satisfy

$$\exists \alpha_1 > 0,\ \kappa_2 > 0,\ \sum_{\mathbf{b}\in\mathcal{B}_n} \frac{1}{b_1\cdots b_d} \leq \kappa_2 n^{\alpha_1}, \tag{16}$$

and one of the following constraints

$$|\mathcal{B}_n| \leq \ln(n), \ \text{ or } \ \forall \kappa_3 > 0,\ \exists C_0 > 0,\ \sum_{\mathbf{b}\in\mathcal{B}_n} \exp\left(-\frac{\kappa_3}{b_1\cdots b_d}\right) \leq C_0, \tag{17}$$

where $|\mathcal{B}_n|$ is the cardinal of the set $\mathcal{B}_n$. These assumptions are similar to (13). Let

$$\widehat{B}_c(\mathbf{b}) = \max_{\mathbf{b}'\in\mathcal{B}_n} \left\{\|W_{\mathbf{b}} \star \widehat{c}_{\mathbf{b}'} - \widehat{c}_{\mathbf{b}'}\|^2_{L^2([0,1]^d)} - V_c(\mathbf{b}')\right\}_+ \tag{18}$$

with

$$V_c(\mathbf{b}) = \kappa_c \frac{\|W\|^2_{L^1([0,1]^d)}\|W\|^2_{L^2([0,1]^d)}}{nb_1\cdots b_d}, \quad \kappa_c > 0, \tag{19}$$

like above for regression estimation, $\widehat{B}_c$ stands for an empirical counterpart of the bias term of the risk, and V_c has the same order as the variance term (compare to (15)).

An oracle-type inequality could be derived for the final copula density estimator $\widehat{c}_{\widehat{\mathbf{b}}}$, with $\widehat{\mathbf{b}} = \arg\min_{\mathbf{b}\in\mathcal{B}_n}\{\widehat{B}_c(\mathbf{b}) + V_c(\mathbf{b})\}$.

Proposition 4.1. *Assume ($H_{c,high}$) (on $[0,1]^d$), and assume that the marginal c.d.f. of the vector $\mathbf{X}$ are known. Then, there exist some nonnegative constants c_1 and c_2 such that*

$$\mathbb{E}\left[\left\|\widehat{c}_{\widehat{\mathbf{b}}} - c\right\|^2_{L^2([0,1]^d)}\right] \leq c_1 \min_{\mathbf{b}\in\mathcal{B}_n}\left\{\|W_{\mathbf{b}} \star c - c\|^2_{L^2([0,1]^d)} + \frac{\|W\|^2_{L^2([0,1]^d)}}{nb_1\cdots b_d}\right\} + \frac{c_2\ln(n)}{n}.$$

Note that the L^1-norm of the kernel does not appear in (15), but only in the variance term of the Goldenshluger-Lepski method, namely (19), for technical reasons (more details on the proof in Section 8.4 or in Section 3.4.2 of Comte 2015).

The logarithmic term in the upper-bound of the inequality can be avoided by assuming the second part of (17), instead of $|\mathcal{B}_n| \leq \ln(n)$. Like the tuning constant κ in $\widehat{V}$ (see (11)), the constant κ_c in (19) has to be calibrated. The bound that we obtain in the proof is unfortunately not accurate (this is a consequence of numerous technical upper bound, based on a concentration inequality), and cannot be used for practical purpose. The tuning of this parameter will be discussed below (see Section 6.2). Keep in mind for the following section that the same oracle inequality holds for an integrated risk on a smaller set than $[0,1]^d$, *e.g.* $\widetilde{F}_{\mathbf{X}}(A)$. In this case, it is enough to assume an upper-bound on the copula density on this set. Proposition 4.1 also permits to derive an adaptive convergence rate for our copula density estimator (even if its not the initial goal of the paper) : if the copula density c belongs to a Nikol'skiĭ ball $\mathcal{N}_2(\alpha, L')$ for $L' > 0$ and $\alpha = {}^t(\alpha_1, \ldots, \alpha_d) \in (\mathbb{R}_+^*)^d$, and if the kernel W is of order $\ell \in \mathbb{R}_+^d$ such that $\ell_j \geq \lfloor \alpha_j \rfloor$ for $j = 1, \ldots, d$, (see Assumption $(H_{K,\ell})$), then $\widehat{c}_{\widehat{\mathbf{b}}}$ automatically achieves the convergence rate $n^{-\frac{2\bar{\alpha}}{2\bar{\alpha}+d}}$ where $\bar{\alpha}$ is the harmonic mean of the components of α. Following Autin et al. (2010), this is also the lower bound for the minimax risk, and thus our estimator is minimax optimal (with no additional logarithm factor, comparing to Corollary 4.1 of Autin et al. 2010).

5. Plug-in regression estimate

Now we consider the general case of unknown copula density c to estimate the regression function r. The idea is to plug the kernel estimator $\widehat{c}_{\mathbf{b}}$ (defined by (14)) of c in (7) for a well-chosen bandwidth $\mathbf{b}$. We consider the case of fixed bandwidth, both for the regression and the copula estimators, this paves the way of future works about the fully data-driven estimator (with two selected bandwidth, see the concluding remarks below). Let us plug in $\hat{r}_{\mathbf{h}}$ the estimator $\widehat{c}_{\mathbf{b}}$: for any $\mathbf{b}, \mathbf{h} > 0$, under Assumption $(H_{c,low})$,

$$\widehat{r}_{\mathbf{h},\mathbf{b}}(\mathbf{x}) = \frac{1}{n\widehat{c}_{\mathbf{b}}(\widetilde{F}_{\mathbf{X}}(\mathbf{x}))} \sum_{i=1}^{n} Y_i K_{\mathbf{h}}(\widetilde{F}_{\mathbf{X}}(\mathbf{x}) - \widetilde{F}_{\mathbf{X}}(\mathbf{X}_i)) \mathbf{1}_{\widehat{c}_{\mathbf{b}}(\widetilde{F}_{\mathbf{X}}(\mathbf{x})) \geq m_c/2}, \quad \mathbf{x} \in A. \tag{20}$$

This means that $\widehat{r}_{\mathbf{h},\mathbf{b}}(\mathbf{x}) = ((c \times \widehat{g}_{\mathbf{h}})/\widehat{c}_{\mathbf{b}}) \circ \widetilde{F}_{\mathbf{X}}(\mathbf{x}) \mathbf{1}_{\widehat{c}_{\mathbf{b}}(\widetilde{F}_{\mathbf{X}}(\mathbf{x})) \geq m_c/2}$. To make the estimator fully computable, one needs to know the lower bound m_c of the copula: in practice it is possible to replace it by a lower bound of an estimator. As explained previously, to avoid making the proofs more technical and cumbersome, we choose to not consider the problem from a theoretical point of view.

We obtain the following upper-bound for our ratio estimator. Its risk has the order of magnitude of the worst risk between the risk of $\widehat{r}_{\mathbf{h}}$ and $\widehat{c}_{\mathbf{b}}$.

Proposition 5.1. *Assume $(H_{c,low})$ and $(H_{c,high})$. Then,*

$$\mathbf{E}[\|\widehat{r}_{\mathbf{h},\mathbf{b}} - r\|_{f_{\mathbf{X}}}^2] \leq \frac{4M_c}{m_c^2} \quad \{\, 2M_c \mathbf{E}[\|\widehat{r}_{\mathbf{h}} - r\|_{f_{\mathbf{X}}}^2] + (2\|g\|_{L^\infty(\widetilde{F}_{\mathbf{X}}(A))}^2 + \|g\|_{L^2(\widetilde{F}_{\mathbf{X}}(A))}^2) \mathbb{E}\left[\|\widehat{c}_{\mathbf{b}} - c\|_{\widetilde{F}_{\mathbf{X}}(A)}^2\right] \right\}.$$

The result is not surprising, and we cannot expect to obtain a sharper bound for the plug-in estimator. We thus have to add smoothness assumptions both on the regression function and on the copula density to derive the convergence rate of the plug-in estimator.

Finally, to obtain the fully computable estimator, one needs to replace the c.d.f. $\widetilde{F}_{\mathbf{X}}$ by its empirical counterpart introduced in (5). The switch is not a problem: the idea is that the empirical c.d.f. converges at a parametric rate, that does not deteriorate our slower nonparametric decrease of the risk. The multivariate setting does not change anything for the substitution compare to the univariate case. The scheme of the switching can now be considered as classical, since it has been widely detailed both by Kerkyacharian and Picard (2004) and Chagny (2015), but it significantly increases the length of the proofs. That is why, following many works about warped estimation (Chesneau and Willer 2015; Pham Ngoc 2009...), we do not give all the details.

6. Simulation study

In this section we illustrate the performance of our estimator with a simulation study, carried out with the free software R. The regression function that we consider is $r(x_1, x_2) = 1/\sqrt{x_1 x_2}$ (for $(x_1, x_2) \in \mathbb{R}^+ \times \mathbb{R}^+$, see Figure 1).

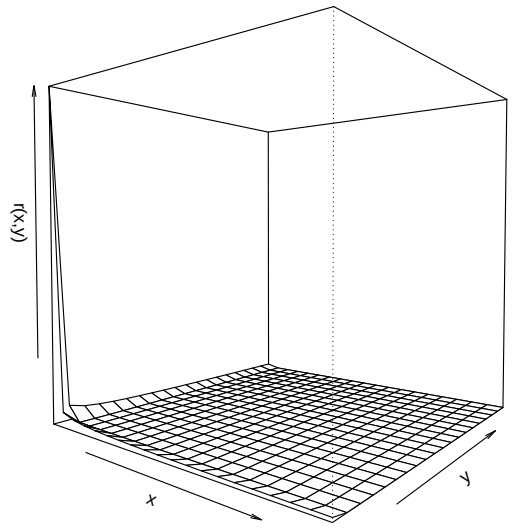

Figure 1. Regression function: $r(x_1, x_2) = 1/\sqrt{x_1 x_2}$ for $(x_1, x_2) \in \mathbb{R}^+ \times \mathbb{R}^+$.

To check the assumptions of the theoretical results, the design (X_1, X_2) is generated using a Frank Copula with parameter 10 and exponential marginals with mean 1 (see Figure 2). The support of the design distribution is thus unbounded, which is possible with our method, according to the theory. The case of bounded support for the design distribution is briefly investigated below, Section 6.3. The response variable is given by $Y = r(X_1, X_2) + \varepsilon$ with ε a Gaussian noise with mean 0 and standard deviation 0.025.

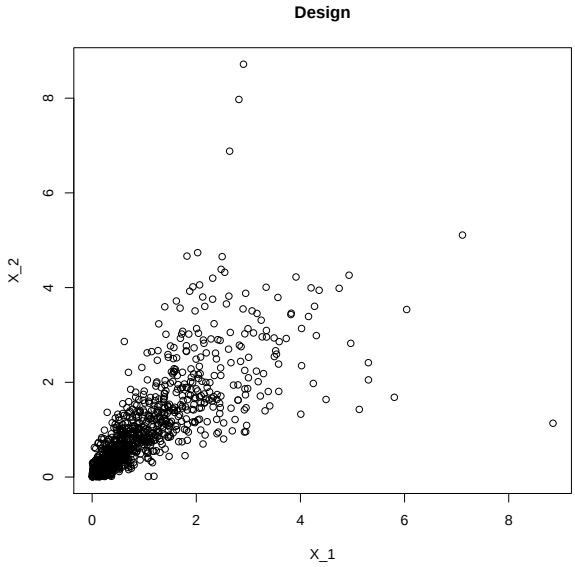

FIGURE 2. Illustration of the design : Frank copula with parameter 10 and exponential marginals.

To study the performances of our estimators, we use a Monte Carlo approximation, with 1000 iterations of independent samples (from the data used to compute an estimate $\widehat{r}$) of a relative L_2-risk, namely the Relative Mean Square Error (RMSE):

$$RMSE = \sum_{j=1}^{1000} \left(\frac{\widehat{r}(X_{j,1}, X_{j,2}) - r(X_{j,1}, X_{j,2})}{r(X_{j,1}, X_{j,2})} \right)^2 .$$

Finally, we confront our estimators with the classical Nadaraya-Watson kernel estimator with a cross-validation selected bandwidth (using the *npreg* function of the R package **np** (Hayfield and Racine, 2008)) that is not designed to deal with an unbounded design and the estimator proposed by Kohler et al. (2009).

6.1. **Impact of the estimation fo the marginal distributions of the design.** In this section we investigate how the estimation of the marginal design distribution, through the empirical c.d.f., affects the results. We compare the estimators $\widehat{r}_{\mathbf{h},\mathbf{b}}$ computed with the true marginal distributions and $\widehat{r}_{\mathbf{h},\mathbf{b},\widehat{F}}$ computed with the estimated c.d.f.. using the following bandwidths $h_1 = h_2 = b_1 = b_2 = (log(n)/n)^{0.5}$. Several bandwidths have been tested, and this choice "by hand" is a reasonable one among all the possibilities. We provide in Figure 3 the corresponding boxplots for sample sizes $n = 100$, 500 and 1000. For each sample size, 100 RMSE values (computed from independent samples) are plotted. The estimation of the marginal distributions is carried out using the classical empirical cumulative distribution function with the function *ecdf* of the software R.

The results are quite similar in both cases. Using the empirical counterpart $\widehat{\widetilde{F}}_{\mathbf{X}}$ instead of the true c.d.f. $\widetilde{F}_{\mathbf{X}}$ does not seem to affect the quality of the final estimator. This kind of results is not very surprising as the estimator $\widehat{\widetilde{F}}_{\mathbf{X}}$ is widely known to be a very good estimate of $\widetilde{F}_{\mathbf{X}}$. From now on, the marginal distributions are thus estimated for all the presented results.

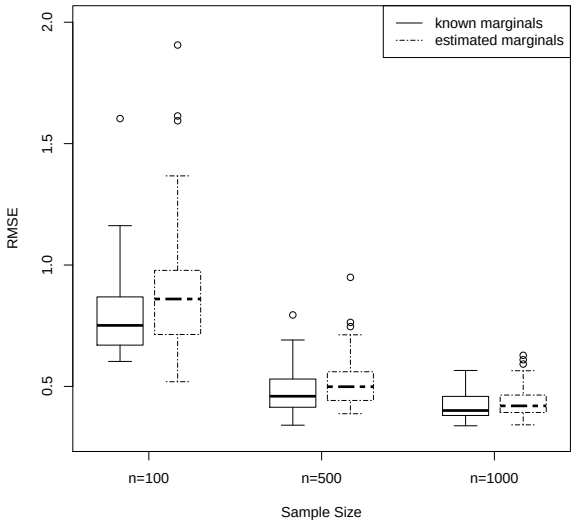

FIGURE 3. Effect of the marginal distributions estimation.

6.2. **Simulations with data-driven bandwidth for the copula estimator and fixed bandwidth for the regression estimator.** In this subsection, we confront our estimator (with and without bandwidth selection) to the well-known Nadaraya-Watson estimate and to the one proposed by Kohler et al. (2009).

The bandwidths of the regression estimator are chosen like in the previous subsection whereas there are two cases for the copula density estimator. For the estimator without bandwidth selection they are chosen like in the previous subsection but when we perform bandwidth selection, the applied methodology is the Goldenshuger-Lepski procedure detailed in Section 4. The L^2-norm involved in the approximation of the bias term (18) in the selection device is approximated by a Riemann sum over a regular grid of 50 points. As explained above (end of Section 4) the procedure also requires a value for the tuning constant κ_c involved in (19). Classically, we tune it once and for all, for each sample size. Following globally the scheme detailed by Bertin et al. (2016) (section 7.2), we study the evolution of the risk with respect to the constant, and choose a value that minimizes the risk. But, we take into account recent research by Lacour and Massart (2016) about the difficulty of optimal tuning of the Lepski methods. We just propose to select $\widetilde{\mathbf{b}} \in \arg\min_{\mathbf{b}\in\mathcal{B}_n}\{\widehat{B}_c(\mathbf{b}) + 2V_c(\mathbf{b})\}$ instead of $\widehat{\mathbf{b}}$, and to compute the new final estimate $\widehat{c}_{\widetilde{\mathbf{b}}}$. The reason are mainly technical, and we refer to Section 5 of Lacour and Massart (2016) for details. Figure 4 thus presents the calibration results (risk of $\widehat{c}_{\widetilde{\mathbf{b}}}$ with respect to the value of κ_c). Remark that the shape of the curve is the same with different regression functions and different design distributions.

The figure above assesses that the value of the constant is crucial : a too small or too large choice can lead to an increase of 50% of the RMSE. The selected values ($\kappa_c = 30$ for $n = 100$, $\kappa_c = 280$ for $n = 500$ and $\kappa_c = 680$ for $n = 1000$) are then used to compute the estimator $\widehat{c}_{\widetilde{\mathbf{b}}}$ and to evaluate its performances.

Once this calibration is made, we are in position to compare the risk of the 4 competing methods. This is carried out on Figure 5.

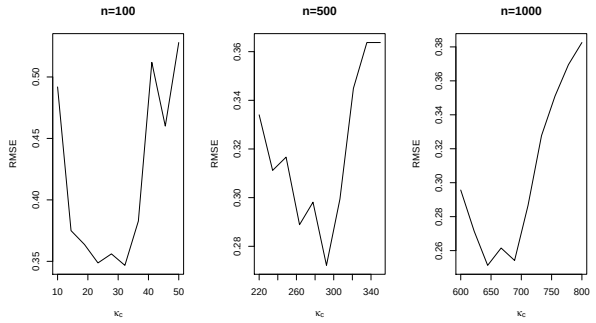

FIGURE 4. RMSE for $\widehat{c}_{\widetilde{\mathbf{b}}}$ with respect to the constant κ_c for different sample sizes n.

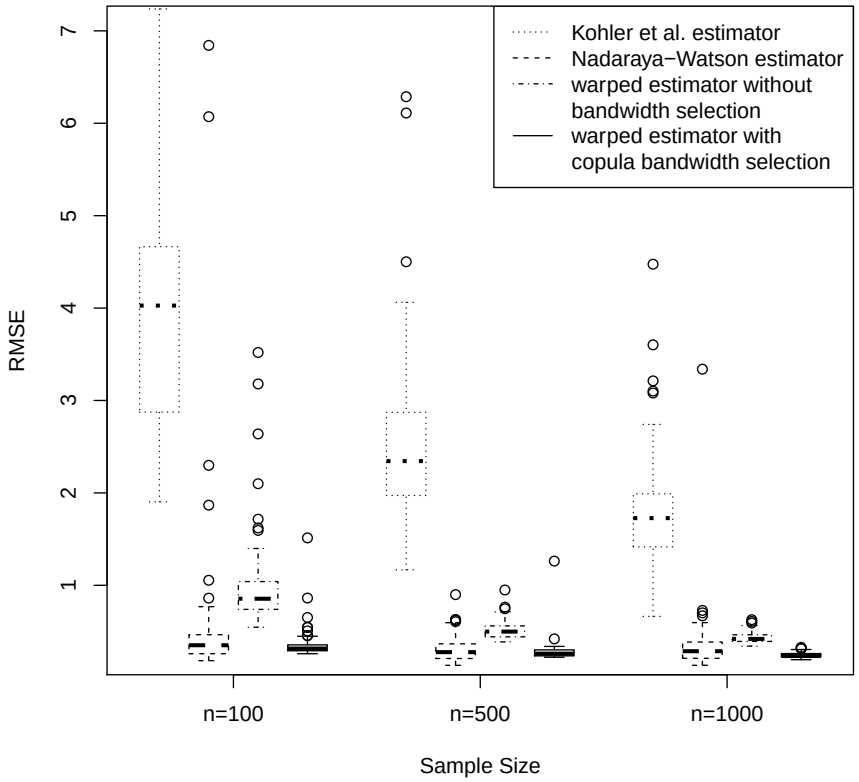

FIGURE 5. Comparison of the RMSE of $\widehat{r}_{\mathbf{h},\mathbf{b},\widehat{F}}$ and $\widehat{r}_{\mathbf{h},\widehat{\mathbf{b}},\widehat{F}}$ with the Nadaraya-Watson estimator and the estimator of Kohler et al. (2009).

First, we can see that the estimator of Kohler et al. (2009) seems to have lower performances than others. Then, the performances of the Nadaraya-Watson estimator are better than our approach without bandwidth selection but worse when we perform copula bandwidth selection. For example for $n = 100$, our estimator has a decrease of the median RMSE of 23% and, above

all, a variance divided by more than 100. This highlights the robustness of our estimator and the interest of the bandwidth selection step.

The simulations are implemented in R on a server with 140 cores, 400 Gbytes de Ram and a E5-2680 v2 @ 2.80Ghz processor. For a data sample of size 100, the computation of the Nadaraya-Watson estimate with a cross-validation selected bandwidth takes 32 seconds, the one of Kohler estimate is 9 seconds. Without any bandwidth selection, our estimate is computed in 7 seconds. By adding the selection step for the bandwidths of the copula density estimate, it requires 84 seconds.

6.3. **Case of bounded support for the design.** The above subsections illustrate the performances of our estimator for the case of an unbounded design. Let us now study the pertinence of our estimator when the design has a distribution with compact support. Here, the design is the same as previously (see Figure 2), but restricted to the square $[0,2] \times [0,2]$, and correctly normalized (see Figure 6). We consider here samples of size $n = 100$.

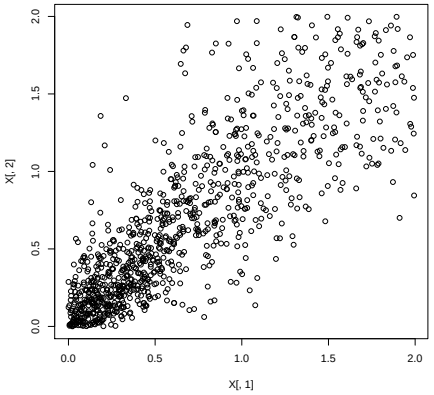

FIGURE 6. Illustration of the design : Frank copula with parameter 10 and exponential marginals truncated on the square $[0,2] \times [0,2]$.

The Figure 7 below shows the performances of the different methods with bounded or unbounded support.

These results highlight the importance of the bandwidth selection procedure and the robustness of our warped estimator to the case of bounded support : even in this case, our estimator has nearly the same median for the RMSE than the Nadaraya-Watson with a variance divided by 5.

7. Concluding remarks

The aim of the paper is to extend the so-called "warping" device to a multivariate framework, through the study of regression kernel estimation. When the design distribution is known, the extension of the method can be done and similar results as the ones obtained in the univariate framework (non-asymptotic risk bound and optimal convergence rate) are proved. When the design distribution is unknown, the challenge is to cope with the possible dependence structure between the coordinates of the design, and the extension can be done only through the additional estimation of the copula density. This can be done separately in an adaptive way, also with a kernel estimator. Section 5 paves the way for a future study of the plug-in estimator, which is

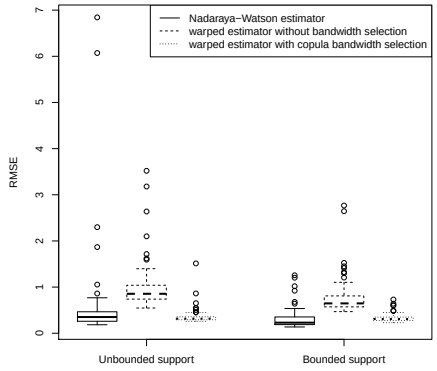

FIGURE 7. Impact of the compacity of the support of the design distribution on the performances of the different estimators (with $n = 100$).

out of the scope of the paper: the risk of the regression estimator with fixed bandwidth and after plug-in of the copula estimate (also with fixed bandwidth), depends both on the risk of the estimator with known distribution and on the risk of the copula estimator, which is not surprising.

The next step, out of the scope of the paper, is to propose a bandwidth selection rule for the regression estimate computed with the adaptive copula density estimate (that is with selected bandwidth). It requires to replace the copula density c in the Goldenshluger-Lepski estimation of the bias term of the risk (see (10)) by $\widehat{c}_{\widehat{\mathbf{b}}}$. This probably also implies a modification of the variance term (11) to penalize the plug-in, but is not straightforward. The difficulties are numerous, owing first to the problem of dependence (the regression estimate, the copula estimate, and the selected bandwidth depend on the design X_i) : it makes difficult to isolate the risk of the adaptive copula estimator from the risk of the regression estimator with known marginal distribution. A natural idea is to imagine that we have at our disposal an additional sample of the design, independent from the data. We can perhaps then conduct the study in the spirit of Bertin et al. (2016), who deal with similar questions for conditional density estimators (that involve the plug-in of marginal density estimates). Another way to tackle the problem could be to adapt very recent research that suggests to develop alternative selection algorithms (see for example Lacour et al. 2017; Nguyen 2018) to choose simultaneously the two bandwidths $\widehat{\mathbf{b}}$ and $\widehat{\mathbf{h}}$.

Finally notice that the results we obtain above for multivariate random design regression with additive error term can be extended to handle other multivariate estimation problems, such as regression estimation in the heteroskedastic model or cumulative distribution function estimation from data subject to interval censoring case 1, as it is proposed in Chagny (2015) for $d = 1$.

8. PROOFS

The main tool of the theoretical results is the following concentration inequality (Lacour, 2008).

Theorem 8.1 (Talagrand Inequality)**.** *Let $\mathcal{F}$ be a set of uniformly bounded functions, which have a countable dense sub-family for the infinite norm. Let $(V_1, \dots, V_n)$ be independent random*

variables and

$$Z = \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} \left(f(V_i) - \mathbb{E}\left[f(V_i)\right]\right) \right|.$$

Consider M_1, v, and H, such that

$$M_1 \geq \sup_{f\in\mathcal{F}} \|f\|_\infty, v \geq \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} Var(f(V_i)) \text{ and } H \geq \mathbb{E}[Z].$$

Then, for every $\delta > 1$, there exist numerical positive constants C_1, C_2, c_1 and c_2 such that

$$\mathbb{E}\left[\left(Z^2 - \delta H^2\right)_+\right] \leq C_1 \frac{v}{n}\exp\left(-c_1\frac{nH^2}{v}\right) + C_2\frac{M_1^2}{n^2}\exp\left(-c_2\frac{nH}{M_1}\right).$$

We will use several times the following standard convolution inequality, called the Young Inequality. Let $p, q \in [1;\infty)$ such $1/p + 1/q \geq 1$. If $s \in L^p(\mathbb{R}^d)$ and $t \in L^q(\mathbb{R}^d)$, then, $s \star t \in L^r(\mathbb{R})$ with $1/r = 1/p + 1/q - 1$, and

$$\|s \star t\|_{L^r(\mathbb{R}^d)} \leq \|s\|_{L^p(\mathbb{R}^d)}\|t\|_{L^q(\mathbb{R}^d)}. \tag{21}$$

8.1. **Proof of Proposition 3.1.** The variance term is

$$\begin{aligned} V(\mathbf{h}) &= \int_{\widetilde{F}_{\mathbf{X}}(A)} c(\mathbf{u})\mathrm{Var}\left(\widehat{g}_{\mathbf{h}}(\mathbf{u})\right) d\mathbf{u}, \\ &= \frac{1}{n}\int_{\widetilde{F}_{\mathbf{X}}(A)} \frac{1}{c(\mathbf{u})}\mathrm{Var}\left(YK_{\mathbf{h}}(\mathbf{u} - \widetilde{F}_{\mathbf{X}}(\mathbf{X}))\right) d\mathbf{u}, \\ &\leq \frac{1}{nm_c}\int_{\widetilde{F}_{\mathbf{X}}(A)} \mathbb{E}\left[Y^2K_{\mathbf{h}}^2(\mathbf{u} - \widetilde{F}_{\mathbf{X}}(\mathbf{X}))\right] d\mathbf{u}, \end{aligned}$$

using Assumption $(H_{c,low})$. But,

$$\mathbb{E}[Y^2K_{\mathbf{h}}^2(\mathbf{u} - \widetilde{F}_{\mathbf{X}}(\mathbf{X}))] = \mathbb{E}[r^2(\mathbf{X})K_{\mathbf{h}}^2(\mathbf{u} - \widetilde{F}_{\mathbf{X}}(\mathbf{X}))] + \mathbb{E}[\varepsilon^2]\mathbb{E}[K_{\mathbf{h}}^2(\mathbf{u} - \widetilde{F}_{\mathbf{X}}(\mathbf{X}))],$$

and

$$\begin{aligned} \int_{\widetilde{F}_{\mathbf{X}}(A)} \mathbb{E}\left[r^2(\mathbf{X})K_{\mathbf{h}}^2(\mathbf{u} - \widetilde{F}_{\mathbf{X}}(\mathbf{X}))\right] d\mathbf{u} &= \int_{\widetilde{F}_{\mathbf{X}}(A)} \left(\int_A r^2(\mathbf{x})K_{\mathbf{h}}^2(\mathbf{u} - \widetilde{F}_{\mathbf{X}}(\mathbf{x}))f_{\mathbf{X}}(\mathbf{x})d\mathbf{x}\right) d\mathbf{u}, \\ &= \int_A r^2(\mathbf{x})f_{\mathbf{X}}(\mathbf{x})\left(\int_{\widetilde{F}_{\mathbf{X}}(A)} K_{\mathbf{h}}^2(\mathbf{u} - \widetilde{F}_{\mathbf{X}}(\mathbf{x}))d\mathbf{u}\right) d\mathbf{x}, \\ &\leq \int_A r^2(\mathbf{x})f_{\mathbf{X}}(\mathbf{x})\frac{\|K\|^2}{h_1\dots h_d}d\mathbf{x} \leq \frac{\|K\|^2}{h_1\dots h_d}\mathbb{E}[r^2(\mathbf{X})] \end{aligned}$$

where $\|K\|=\|K\|_{L^2(\mathbb{R}^d)}$.

Similar computations lead to $\mathbb{E}[\varepsilon^2]\mathbb{E}[K_{\mathbf{h}}^2(\mathbf{u} - \widetilde{F}_{\mathbf{X}}(\mathbf{X}))] \leq \mathbb{E}[\varepsilon^2]\|K\|^2/(h_1\dots h_d)$. This proves that

$$V(\mathbf{h}) \leq \frac{\|K\|^2\mathbb{E}[Y_1^2]}{m_c}\frac{1}{nh_1\dots h_d}.$$

For the bias term, the result is classical, see *e.g.* Proposition 3 p.579 of Comte and Lacour (2013) (with $r_j = a_j = 0$).

8.2. **Proof of Corollary 3.1.** Let f be the multivariate function defined by

$$f \ : \ \mathbf{h}=(h_1,\dots,h_d)\in(\mathbb{R}_+^*)^d \longmapsto f(\mathbf{h})=C_1\sum_{l=1}^d h_l^{2\beta_l}+C_2\frac{1}{nh_1\dots h_d},$$

with $C_1=L$ and $C_2=\|K\|^2\mathbb{E}[Y_1^2]$. By studying f, we prove that it admits a unique minimum on $(\mathbb{R}_+^*)^d$. The function f is indeed differentiable, and admits a unique critical point $\mathbf{h}(\beta)$ (the one for which the gradient of f equals to 0) such that

$$\forall j\in\{1,\dots,d\},\ \ -\frac{C_2}{nh_1(\beta)\cdots h_d(\beta)}\frac{1}{h_j}+2C_1\beta_j h_j(\beta)^{2\beta_j-1}=0, \tag{22}$$

or equivalently

$$\forall j\in\{1,\dots,d\},\ \ h_j(\beta)=\left(\frac{C_2}{C_1\beta_j}\right)^{1/(2\beta_j)}(h_1(\beta)\cdots h_d(\beta))^{-1/(2\beta_j)}n^{-1/(2\beta_j)}.$$

By multiplying these d equalities, we get

$$h_1(\beta)\cdots h_d(\beta)=\prod_{j=1}^d\left(\frac{C_2}{C_1\beta_j}\right)^{1/(2\beta_j)}(h_1(\beta)\cdots h_d(\beta))^{-\sum_{j=1}^d 1/(2\beta_j)}n^{-\sum_{j=1}^d 1/(2\beta_j)}.$$

From this, we derive

$$h_1(\beta)\cdots h_d(\beta)=\left(\prod_{j=1}^d\left(\frac{C_2}{C_1\beta_j}\right)^{1/(2\beta_j)}\right)^{-\frac{2\bar\beta}{2\bar\beta+d}}n^{-\frac{d}{2\bar\beta+d}}.$$

This is sufficient to compute the associate convergence rate : indeed, we have,

$$C_2\frac{1}{nh_1(\beta)\cdots h_d(\beta)}=Cn^{-\frac{2\bar\beta}{2\bar\beta+d}},$$

for a constant C, and the second term of f, namely $C_1\sum_{l=1}^d h_l(\beta)^{2\beta_l}$ has the same order of magnitude (see (22)). Thus, we obtain the result of Corollary 3.1.

8.3. **Proof of Theorem 3.1.**

8.3.1. *Heuristic about the bandwidth selection method.* Let us briefly explain the ideas behind the Goldenshluger-Lepski method. Given the finite bandwidth collection $\mathcal{H}_n\subset(\mathbb{R}_+^*)^d$, the optimal choice is the one which minimizes the bias/variance trade off (see (9)), or the upper-bound of Proposition 3.1. However, since r is unknown, the variance and the bias term of the risk are unknown, and the smoothness index of r is likely to be unknown too. Thus the optimal bandwidth is unattainable. The idea of the method is to mimic the bias/variance decomposition of the result of Proposition 3.1, with empirical estimations. The simplest term is $\widehat{V}(\mathbf{h})$ (see (11)) that has the order of the variance term of the risk. The main specificity of the Goldenshluger-Lepski method is to provide an empirical counterpart for the bias term of the upper bound by comparing pair by pair several estimators. The proposition is to introduce auxiliary estimators that involve two kernels, $K_{\mathbf{h}}$ and $K_{\mathbf{h}'}$. In our framework, the bias term is $\|(K_{\mathbf{h}}\star(cg))/c)\circ\widetilde{F}_{\mathbf{X}}-r\|^2_{f_{\mathbf{X}}}$. Since r and cg are unknown, they are replaced by estimators with bandwidth $\mathbf{h}'$: this leads to $\|(K_{\mathbf{h}}\star(c\widehat{g}_{\mathbf{h}'}))/c)\circ\widetilde{F}_{\mathbf{X}}-\widehat{r}_{\mathbf{h}'}\|^2_{f_{\mathbf{X}}}$, which makes appear the auxiliary estimator of the method, $K_{\mathbf{h}}\star(c\widehat{g}_{\mathbf{h}'}))/c)\circ\widetilde{F}_{\mathbf{X}}$ in our framework. This adds a random part to a deterministic term, namely the bias : this random part should be corrected by substracting $\widehat{V}(\mathbf{h}')$. Finally, since any bandwidth $\mathbf{h}'$ of the collection could be chosen, we scan all the

collection. This leads to the definition of the bias estimate, $\widehat{B}(\mathbf{h})$ (see (10)). This explanation is obviously a rough heuristic, and we carefully show that the term $\widehat{B}$ has the order of the bias term (see Inequality (27) below). It remains to select the bandwidth $\widehat{\mathbf{h}}$ that minimizes the sum of these empirical $\widehat{B}(\mathbf{h})$ and $\widehat{V}(\mathbf{h})$, over all the possible bandwidths $\mathbf{h}$. We thus get (12).

8.3.2. *Proof of the result.* First, the loss function of the selected estimator can be written

$$\left\|\widehat{r}_{\widetilde{\mathbf{h}}}-r\right\|_{f_{\mathbf{X}}}^2=\int_{\widetilde{F}_{\mathbf{X}}(A)}\left(c\widehat{g}_{\widetilde{\mathbf{h}}}-cg\right)^2(\mathbf{u})\frac{1}{c(\mathbf{u})}d\mathbf{u}.$$

Let $\mathbf{h}\in\mathcal{H}_n$ be fixed. We introduce $K_{\widetilde{\mathbf{h}}}\star(c\widehat{g}_{\mathbf{h}}\mathbf{1}_{\widetilde{F}_{\mathbf{X}}(A)})$ and follow the decompositions of Theorem 4.2 in Comte (2015) to obtain

$$\left\|\widehat{r}_{\widetilde{\mathbf{h}}}-r\right\|_{f_{\mathbf{X}}}^2\leq 6\left(\widetilde{V}(\mathbf{h})+\widetilde{B}(\mathbf{h})\right)+3\left\|\widehat{r}_{\mathbf{h}}-r\right\|_{f_{\mathbf{X}}}^2.$$

By taking the expectation, the last term of the previous inequality is the risk of an estimator with fixed bandwidth, controlled by Proposition 3.1. It remains to bound $\widetilde{B}(\mathbf{h})$. Let us begin by splitting the norm involved in its definition. We have

$$\begin{aligned}\left\|\frac{K_{\mathbf{h}}\star(c\widehat{g}_{\mathbf{h}'}\mathbf{1}_{\widetilde{F}_{\mathbf{X}}(A)})}{c}\circ\widetilde{F}_{\mathbf{X}}-\widehat{r}_{\mathbf{h}',c}\right\|_{f_{\mathbf{X}}}^2 &= \int_{\widetilde{F}_{\mathbf{X}}(A)}\left(K_{\mathbf{h}}\star(c\widehat{g}_{\mathbf{h}'}\mathbf{1}_{\widetilde{F}_{\mathbf{X}}(A)})(\mathbf{u})-c(\mathbf{u})\widehat{g}_{\mathbf{h}'}(\mathbf{u})\right)^2\frac{d\mathbf{u}}{c(\mathbf{u})},\\ &\leq 3m_c^{-1}\sum_{l=1}^{3}T_{l,\mathbf{h},\mathbf{h}'},\end{aligned}$$

with

$$\begin{aligned}T_{1,\mathbf{h},\mathbf{h}'}&=\int_{\widetilde{F}_{\mathbf{X}}(A)}\left(K_{\mathbf{h}}\star(c\widehat{g}_{\mathbf{h}'}\mathbf{1}_{\widetilde{F}_{\mathbf{X}}(A)})(\mathbf{u})-K_{\mathbf{h}}\star K_{\mathbf{h}'}\star(cg\mathbf{1}_{\widetilde{F}_{\mathbf{X}}(A)})(\mathbf{u})\right)^2d\mathbf{u},\\ T_{2,\mathbf{h},\mathbf{h}'}&=\int_{\widetilde{F}_{\mathbf{X}}(A)}\left(K_{\mathbf{h}}\star K_{\mathbf{h}'}\star(cg\mathbf{1}_{\widetilde{F}_{\mathbf{X}}(A)})(\mathbf{u})-K_{\mathbf{h}'}\star(cg\mathbf{1}_{\widetilde{F}_{\mathbf{X}}(A)})(\mathbf{u})\right)^2d\mathbf{u},\\ T_{3,\mathbf{h},\mathbf{h}'}&=\int_{\widetilde{F}_{\mathbf{X}}(A)}\left(K_{\mathbf{h}'}\star(cg\mathbf{1}_{\widetilde{F}_{\mathbf{X}}(A)})(\mathbf{u})-c(\mathbf{u})\widehat{g}_{\mathbf{h}'}(\mathbf{u})\right)^2d\mathbf{u}.\end{aligned}$$

The proof has now some similarities with the proof of Theorem 1 of Chagny (2015), some easy calculations are thus omitted. We first apply the Young inequality (21) (with $r=2$, $p=1$, $q=2$) that leads to

$$\begin{aligned}T_{1,\mathbf{h},\mathbf{h}'}&\leq\|K\|_{L^1([0,1]^d)}\left\|c\widehat{g}_{\mathbf{h}'}-K_{\mathbf{h}'}\star(cg\mathbf{1}_{\widetilde{F}_{\mathbf{X}}(A)})\right\|_{L^2(\widetilde{F}_{\mathbf{X}}(A))}^2\\ T_{2,\mathbf{h},\mathbf{h}'}&\leq\|K\|_{L^1([0,1]^d)}\left\|K_{\mathbf{h}}\star(cg\mathbf{1}_{\widetilde{F}_{\mathbf{X}}(A)})-(cg)\right\|_{L^2(\widetilde{F}_{\mathbf{X}}(A))}^2.\end{aligned}$$

We thus obtain for any $\mathbf{h}\in\mathcal{H}_n$,

$$\begin{aligned}\widetilde{B}(\mathbf{h}) &\leq \frac{3}{m_c}(1+\|K\|_{L^1([0,1]^d)})\max_{\mathbf{h}'\in\mathcal{H}_n}\left\{\left\|c\widehat{g}_{\mathbf{h}'}-K_{\mathbf{h}'}\star(cg\mathbf{1}_{\widetilde{F}_{\mathbf{X}}(A)})\right\|_{L^2(\widetilde{F}_{\mathbf{X}}(A))}^2-\frac{m_c\widetilde{V}(\mathbf{h}')}{3(1+\|K\|_{L^1([0,1]^d)})}\right\}_+\\ &\quad+\frac{3}{m_c}\|K\|_{L^1([0,1]^d)}\left\|K_{\mathbf{h}}\star(cg\mathbf{1}_{\widetilde{F}_{\mathbf{X}}(A)})-(cg)\right\|_{L^2(\widetilde{F}_{\mathbf{X}}(A))}^2.\end{aligned}$$

We have that

$$\left\|c\widehat{g}_{\mathbf{h}'}-K_{\mathbf{h}'}\star(cg\mathbf{1}_{\widetilde{F}_{\mathbf{X}}(A)})\right\|_{L^2(\widetilde{F}_{\mathbf{X}}(A))}^2=\sup_{t\in\bar{S}(0,1)}\left(\langle c\widehat{g}_{\mathbf{h}'}-K_{\mathbf{h}'}\star(cg\mathbf{1}_{\widetilde{F}_{\mathbf{X}}(A)}),t\rangle_{\widetilde{F}_{\mathbf{X}}(A)}\right)^2=\nu_{n,\mathbf{h}'}(t),$$

with $\bar{S}(0,1)$ a dense countable subset of $\{t \in L^1(\widetilde{F}_{\mathbf{X}}(A)) \cap L^2(\widetilde{F}_{\mathbf{X}}(A)), \|t\|_{L^2(\widetilde{F}_{\mathbf{X}}(A))} = 1\}$, $\langle \cdot, \cdot \rangle_{\widetilde{F}_{\mathbf{X}}(A)}$ the usual scalar product on $L^2(\widetilde{F}_{\mathbf{X}}(A))$, and

$$\nu_{n,\mathbf{h}}(t) = \frac{1}{n}\sum_{i=1}^{n} \int_{\widetilde{F}_{\mathbf{X}}(A)} t(\mathbf{u}) \left(Y_i K_{\mathbf{h}} \left(\mathbf{u} - \widetilde{F}_{\mathbf{X}}(X_i) \right) - \mathbb{E}\left[Y_i K_{\mathbf{h}} \left(u - \widetilde{F}_{\mathbf{X}}(X_i) \right) \right] \right) d\mathbf{u}.$$

For any $t \in \bar{S}(0,1)$, we have that $\nu_{n,\mathbf{h}}(t)^2 \leq 3((\nu_{n,\mathbf{h}}(t)^{(1)})^2 + (\nu_{n,\mathbf{h}}(t)^{(2,1)})^2 + (\nu_{n,\mathbf{h}}(t)^{(2,2)})^2)$, with, for $l \in \{(1),(2,1),(2,2)\}$, $\nu^{(l)}_{n,\mathbf{h}}(t) = \frac{1}{n}\sum_{i=1}^n \varphi^{(l)}_{t,\mathbf{h},i} - \mathbb{E}[\varphi^{(l)}_{t,\mathbf{h},i}]$, and

$$\begin{aligned}
\varphi^{(1)}_{t,\mathbf{h},i} &= r(\mathbf{X}_i) \int_{\widetilde{F}_{\mathbf{X}}(A)} t(\mathbf{u}) K_{\mathbf{h}} \left(\mathbf{u} - \widetilde{F}_{\mathbf{X}}(X_i) \right) d\mathbf{u}, \\
\varphi^{(2,1)}_{t,\mathbf{h},i} &= \mathbf{1}_{|\varepsilon_i| \leq \kappa_n} \int_{\widetilde{F}_{\mathbf{X}}(A)} t(\mathbf{u}) K_{\mathbf{h}} \left(\mathbf{u} - \widetilde{F}_{\mathbf{X}}(X_i) \right) d\mathbf{u}, \\
\varphi^{(2,2)}_{t,\mathbf{h},i} &= \mathbf{1}_{|\varepsilon_i| > \kappa_n} \int_{\widetilde{F}_{\mathbf{X}}(A)} t(\mathbf{u}) K_{\mathbf{h}} \left(\mathbf{u} - \widetilde{F}_{\mathbf{X}}(X_i) \right) d\mathbf{u},
\end{aligned}$$

where $\kappa_n = c_0\sqrt{n}/\ln(n)$ is a quantity which plays a technical role in the proof (c_0 is a nonnegative constant). Writing $\mathbb{E}[Y_1^2] = \mathbb{E}[r^2(X_1)] + \mathbb{E}[\varepsilon_1^2]$, we thus split $\widetilde{V}(\mathbf{h}) = \widetilde{V}_1(\mathbf{h}) + \widetilde{V}_2(\mathbf{h})$, with $\widetilde{V}_1(\mathbf{h}) = \kappa_0 \mathbb{E}[r^2(X_1)]/(m_c n h_1 \dots h_d)$ and $\widetilde{V}_2(\mathbf{h}) = \kappa_0 \mathbb{E}[\varepsilon_1^2]/(m_c n h_1 \dots h_d)$, and consequently

$$\begin{aligned}
&\mathbb{E}\left[\max_{\mathbf{h}' \in \mathcal{H}_n} \left\{ \left\| c\widehat{g}_{\mathbf{h}'} - K_{\mathbf{h}'} \star (cg\mathbf{1}_{\widetilde{F}_{\mathbf{X}}(A)}) \right\|^2_{L^2(\widetilde{F}_{\mathbf{X}}(A))} - \frac{m_c \widetilde{V}(\mathbf{h}')}{3(1 + \|K\|_{L^1([0,1]^d)})} \right\}_+ \right] \\
&\leq 3 \sum_{\mathbf{h} \in \mathcal{H}_n} \left\{ \mathbb{E}\left[\left(\sup_{t \in \bar{S}(0,1)} \left(\nu^{(1)}_{n,\mathbf{h}}(t) \right)^2 - \frac{m_c \widetilde{V}_1(\mathbf{h})}{9(1 + \|K\|_{L^1([0,1]^d)})} \right)_+ \right] \right. \\
&\left. + \mathbb{E}\left[\left(\sup_{t \in \bar{S}(0,1)} \left(\nu^{(2,1)}_{n,\mathbf{h}}(t) \right)^2 - \frac{m_c \widetilde{V}_2(\mathbf{h})}{9(1 + \|K\|_{L^1([0,1]^d)})} \right)_+ \right] + \mathbb{E}\left[\sup_{t \in \bar{S}(0,1)} \left(\nu^{(2,2)}_{n,\mathbf{h}}(t) \right)^2 \right] \right\}. \quad (23)
\end{aligned}$$

Then we obtain, using (H_ε),

$$\sum_{\mathbf{h} \in \mathcal{H}_n} \mathbb{E}\left[\sup_{t \in \bar{S}(0,1)} \left(\nu^{(2,2)}_{n,\mathbf{h}}(t) \right)^2 \right] \leq \|K\|^2 \frac{1}{n\kappa_n^p} \mathbb{E}[|\varepsilon|^{2+p}] \sum_{\mathbf{h} \in \mathcal{H}_n} \frac{1}{h_1 \cdots h_d}. \quad (24)$$

For the other two terms in (23), we apply the Talagrand inequality (Theorem 8.1). This leads to

$$\begin{aligned}
\sum_{\mathbf{h} \in \mathcal{H}_n} \mathbb{E}\left[\left(\sup_{t \in \bar{S}(0,1)} \left(\nu^{(1)}_{n,\mathbf{h}}(t) \right)^2 - \delta_1 \frac{\mathbb{E}[r^2(X_1)]}{n h_1 \cdots h_d} \right)_+ \right] \;&\leq\; C \left(\frac{1}{n} \sum_{\mathbf{h} \in \mathcal{H}_n} \exp\left(-\frac{c_1}{h_1 \cdots h_d} \right) \right. \quad (25) \\
&\left. + \frac{1}{n^2} \exp(-c_2\sqrt{n}) \sum_{\mathbf{h} \in \mathcal{H}_n} \frac{1}{h_1 \cdots h_d} \right),
\end{aligned}$$

for constants δ_1, C, c_1 and c_2, and, for other constants δ_2, C, c_3 and c_4,

$$\sum_{\mathbf{h}\in\mathcal{H}_n}\mathbb{E}\left[\left(\sup_{t\in\bar{S}(0,1)}\left(\nu^{(2,1)}_{n,\mathbf{h}}(t)\right)^2-\delta_2\frac{\mathbb{E}[\varepsilon_1^2]}{nh_1\cdots h_d}\right)_+\right] \leq C\left(\frac{1}{n}\sum_{\mathbf{h}\in\mathcal{H}_n}\exp\left(-\frac{c_3}{h_1\cdots h_d}\right)\right. \tag{26}$$
$$\left.+\frac{\kappa_n^2}{n^2}\exp\left(-c_4\frac{\sqrt{n}}{\kappa_n}\right)\sum_{\mathbf{h}\in\mathcal{H}_n}\frac{1}{h_1\cdots h_d}\right).$$

The assumptions on the collection $\mathcal{H}_n$ permit to deduce that the right hand side of (24), (25), and (26) are less than C/n for a constant C. As soon as

$$\frac{m_c\widetilde{V}_1(\mathbf{h})}{9(1+\|K\|_{L^1([0,1]^d)})}\geq\delta_2\frac{\mathbb{E}[r^2(X_1)]}{nh_1\cdots h_d}\text{ and }\frac{m_c\widetilde{V}_2(\mathbf{h})}{9(1+\|K\|_{L^1([0,1]^d)})}\geq\delta_2\frac{\mathbb{E}[\varepsilon_1^2]}{nh_1\cdots h_d},$$

which is the case if κ_0 is large enough, (23) becomes

$$\mathbb{E}\left[\max_{\mathbf{h}'\in\mathcal{H}_n}\left\{\left\|c\widehat{g}_{\mathbf{h}'}-K_{\mathbf{h}'}\star(cg\mathbf{1}_{\widetilde{F}_{\mathbf{X}}(A)})\right\|^2_{L^2(\widetilde{F}_{\mathbf{X}}(A))}-\frac{m_c\widetilde{V}(\mathbf{h}')}{3(1+\|K\|_{L^1([0,1]^d)})}\right\}_+\right]\leq\frac{C}{n},$$

and consequently

$$\widetilde{B}(\mathbf{h}) \leq \frac{C}{n}+\frac{3}{m_c}\|K\|_{L^1([0,1]^d)}\left\|K_{\mathbf{h}}\star(cg\mathbf{1}_{\widetilde{F}_{\mathbf{X}}(A)})-(cg)\right\|^2_{L^2(\widetilde{F}_{\mathbf{X}}(A))}, \tag{27}$$

which ends the proof.

8.4. **Proof of Proposition 4.1.** As the previous one, this proof is based on oracle-type inequalities using Goldenshluger-Lepski method so we omit some detailed calculations. We first obtain, for any $\mathbf{b}\in\mathcal{B}_n$,

$$\left\|\widehat{c}_{\widehat{\mathbf{b}}}-c\right\|^2_{L^2([0,1]^d)}\leq 6\left(\widehat{B}_c(\mathbf{b})+V_c(\mathbf{b})\right)+3\left\|\widehat{c}_{\mathbf{b}}-c\right\|^2_{L^2([0,1]^d)}.$$

Taking into account the inequality (15), it remains to study $\widehat{B}_c$. Thanks to the convolution inequality (21), we get

$$\begin{aligned}\widehat{B}(\mathbf{b}) &\leq 3\left(\|W\|^2_{L^1([0,1]^d)}+1\right)\max_{\mathbf{b}'\in\mathcal{B}_n}\left(\|\widehat{c}_{\mathbf{b}'}-W_{\mathbf{b}'}\star c\|^2_{L^2([0,1]^d)}-\frac{V_c(\mathbf{b}')}{3(\|W\|^2_{L^1([0,1]^d)}+1)}\right)_+\\ &\quad+3\|W\|^2_{L^1([0,1]^d)}\|W_{\mathbf{b}}\star c-c\|^2_{L^2([0,1]^d)}.\end{aligned}$$

We roughly upper-bound

$$\begin{aligned}&\mathbb{E}\left[\max_{\mathbf{b}'\in\mathcal{B}_n}\left(\|\widehat{c}_{\mathbf{b}'}-W_{\mathbf{b}'}\star c\|^2_{L^2([0,1]^d)}-\frac{V_c(\mathbf{b}')}{3(\|W\|^2_{L^1([0,1]^d)}+1)}\right)_+\right]\\ &\leq\sum_{\mathbf{b}\in\mathcal{B}_n}\mathbb{E}\left[\left(\|\widehat{c}_{\mathbf{b}}-W_{\mathbf{b}}\star c\|^2_{L^2([0,1]^d)}-\frac{V_c(\mathbf{b})}{3(\|W\|^2_{L^1([0,1]^d)}+1)}\right)_+\right],\end{aligned}$$

and write, for any $\mathbf{b}\in\mathcal{B}_n$,

$$\|\widehat{c}_{\mathbf{b}}-W_{\mathbf{b}}\star c\|^2_{L^2([0,1]^d)}=\sup_{t\in S_c(0,1)}\langle\widehat{c}_{\mathbf{b}}-W_{\mathbf{b}}\star c,t\rangle_{L^2([0,1]^d)}=\sup_{t\in S_c(0,1)}\nu^2_{n,c}(t),$$

where $S_c(0,1)$ is a countable subset of the unit sphere of $L^2([0,1]^d)$ (the set of function t such that $\|t\|^2_{L^2([0,1]^d)}=1$), $\langle\cdot,\cdot\rangle_{L^2([0,1]^d)}$ is the scalar product of $L^2([0,1]^d)$, and $\nu_{n,c}(t)=$

$n^{-1}\sum_{i=1}^n \varphi_{t,c,\mathbf{b}}(X_i) - \mathbb{E}[\varphi_{t,c,\mathbf{b}}(X_i)]$, with $\varphi_{t,c,\mathbf{b}}(\mathbf{x}) = \int_{[0,1]^d} W_{\mathbf{b}}(\mathbf{u} - \widetilde{F}(\mathbf{x}))t(\mathbf{u})d\mathbf{u}$. We could now apply Theorem 8.1 to the centered empirical process $\nu_{n,c}$. It is not difficult to see that the following choice for the constants are possible:

$$M_{1,c} = \frac{\|W\|_{L^2([0,1]^d)}}{\sqrt{b_1\cdots b_d}}, \quad H_c^2 = \frac{\|W\|^2_{L^2([0,1]^d)}}{n(b_1\cdots b_d)}, \quad v_c = M_c\|W\|_{L^1([0,1]^d)}.$$

We only detail the computation of v_c: firstly $n^{-1}\sum_{i=1}^n \mathrm{Var}(\varphi_{t,c\mathbf{b}}(X_i)) = \mathrm{Var}(\varphi_{t,c\mathbf{b}}(X_1)) \leq \mathbb{E}[\varphi^2_{t,c\mathbf{b}}(X_1)]$. Then, denoting by $\check{W}_{\mathbf{b}}(x) = W_{\mathbf{b}}(-x)$, we use Assumption $(H_{c,high})$, and the Young inequality (21) with $p = 2$, $q = 1$, and $r = 2$:

$$\begin{aligned}\mathbb{E}\left[\varphi^2_{t,c,\mathbf{b}}(X_1)\right] &= \mathbb{E}\left[(t\star\check{W}_{\mathbf{b}})^2(\widetilde{F}_{\mathbf{X}}(X_1))\right] = \int_{\mathbb{R}^d}(t\star\check{W}_{\mathbf{b}})^2(\widetilde{F}_{\mathbf{X}}(\mathbf{x}))f_{\mathbf{X}}(\mathbf{x})d\mathbf{x},\\ &= \int_{[0,1]^d}(t\star\check{W}_{\mathbf{b}})^2(\mathbf{u})c(\mathbf{u})d\mathbf{u} \leq M_c\|t\star\check{W}_{\mathbf{b}}\|^2_{L^2([0,1]^d)} \leq M_c\|W\|^2_{L^1([0,1]^d)} := v_c,\end{aligned}$$

since $\|t\|^2_{L^2([0,1]^d)} = 1$. Using Theorem 8.1, we obtain, for any $\delta \geq 1$ and for some constants C, c_1, c_2 (that may change from line to line)

$$\begin{aligned}\sum_{\mathbf{b}\in\mathcal{B}_n}\mathbb{E}\left[\left(\sup_{t\in S_c(0,1)}\nu^2_{n,c}(t) - \delta\frac{\|W\|^2_{L^2([0,1]^d)}}{nb_1\cdots b_d}\right)_+\right] &\leq C\sum_{\mathbf{b}\in\mathcal{B}_n}\left\{\frac{1}{n}\exp\left(-c_1\frac{1}{b_1\cdots b_d}\right)\right.\\ &\qquad \left.\frac{1}{n^2}\exp(-c_2\sqrt{n})\frac{1}{b_1\cdots b_d}\right\},\\ &\leq C\left\{\frac{|\mathcal{B}_n|}{n} + \frac{1}{n^2}\exp(-c_2\sqrt{n})\sum_{\mathbf{b}\in\mathcal{B}_n}\frac{1}{b_1\cdots b_d}\right\},\\ &\leq \frac{C\ln(n)}{n},\end{aligned}$$

using the first part of (17) and then the constraint (16) on the collection $\mathcal{B}_\backslash$. If it is the second part of (17) which is assumed, then

$$\sum_{\mathbf{b}\in\mathcal{B}_n}\mathbb{E}\left[\left(\sup_{t\in S_c(0,1)}\nu^2_{n,c}(t) - \delta\frac{\|W\|^2_{L^2([0,1]^d)}}{nb_1\cdots b_d}\right)_+\right] \leq C\left\{\frac{1}{n} + \frac{1}{n^2}\exp(-c_2\sqrt{n})\sum_{\mathbf{b}\in\mathcal{B}_n}\frac{1}{b_1\cdots b_d}\right\} \leq \frac{C}{n}.$$

Thus, if $V_c(\mathbf{b})/(3(\|W\|^2_{L^1([0,1]^d)} + 1)) \geq \delta\frac{\|W\|^2_{L^2([0,1]^d)}}{nb_1\cdots b_d}$ which means that κ_c is large enough, we have proved the following inequality, which concludes the proof of Proposition 4.1 :

$$\widehat{B}(\mathbf{b}) \leq \frac{C}{n} + 3\|W\|^2_{L^1([0,1]^d)}\|W_{\mathbf{b}}\star c - c\|^2_{L^2([0,1]^d)}.$$

8.5. **Proof of Proposition 5.1.** We split the loss function $\|\widehat{r}_{\mathbf{h},\mathbf{b}} - r\|^2_{f_{\mathbf{X}}} = T_1 + T_2$, with

$$T_1 = \|(\widehat{r}_{\mathbf{h},\mathbf{b}} - r)\mathbf{1}_{\widehat{c}_{\mathbf{b}}\circ\widetilde{F}_{\mathbf{X}}\geq m_c/2}\|^2_{f_X}, \quad T_2 = \|r\mathbf{1}_{\widehat{c}_{\mathbf{b}}\circ\widetilde{F}_{\mathbf{X}}<m_c/2}\|^2_{f_X}.$$

First we have

$$T_1 = \left\|\left(\frac{c\times\widehat{g}_{\mathbf{h}}}{\widehat{c}_{\mathbf{b}}}\circ\widetilde{F}_{\mathbf{X}} - r\right)\mathbf{1}_{\widehat{c}_{\mathbf{b}}\circ\widetilde{F}_{\mathbf{X}}\geq m_c/2}\right\|^2_{f_X} = \left\|\left(\frac{c\times\widehat{g}_{\mathbf{h}}}{\widehat{c}_{\mathbf{b}}}\circ\widetilde{F}_{\mathbf{X}} - \frac{c\times g}{c}\circ\widetilde{F}_{\mathbf{X}}\right)\mathbf{1}_{\widehat{c}_{\mathbf{b}}\circ\widetilde{F}_{\mathbf{X}}\geq m_c/2}\right\|^2_{f_X}$$

and thus, $T_1 \leq 2(T_{1,1} + T_{1,2})$, with

$$T_{1,1} = \left\| \frac{c \times \widehat{g}_{\mathbf{h}} - c \times g}{\widehat{c}_{\mathbf{b}}} \circ \widetilde{F}_{\mathbf{X}} \mathbf{1}_{\widehat{c}_{\mathbf{b}} \circ \widetilde{F}_{\mathbf{X}} \geq m_c/2} \right\|_{f_X}^2,$$
$$T_{1,2} = \left\| (c \times g) \circ \widetilde{F}_{\mathbf{X}} \left(\frac{1}{\widehat{c}_{\mathbf{b}}} - \frac{1}{c} \right) \circ \widetilde{F}_{\mathbf{X}} \mathbf{1}_{\widehat{c}_{\mathbf{b}} \circ \widetilde{F}_{\mathbf{X}} \geq m_c/2} \right\|_{f_X}^2.$$

Then, with $(H_{c,high})$ and $(H_{c,low})$, $T_{1,1} \leq (4M_c^2/m_c^2)\|(\widehat{g}_{\mathbf{h}} - g) \circ \widetilde{F}_{\mathbf{X}}\|_{f_X}^2$ and similarly, adding a change of variables

$$T_{1,2} = \left\| g \frac{c - \widehat{c}_{\mathbf{b}}}{\widehat{c}_{\mathbf{b}}} c \mathbf{1}_{\widehat{c}_{\mathbf{b}} \geq m_c/2} \right\|_{L^2(\widetilde{F}_{\mathbf{X}}(A))}^2 \leq \frac{4M_c}{m_c^2} \|g\|_{L^\infty(\widetilde{F}_{\mathbf{X}}(A))}^2 \|c - \widehat{c}_{\mathbf{b}}\|_{L^2(\widetilde{F}_{\mathbf{X}}(A))}^2,$$

thus, $T_{1,2} \leq (4M_c/m_c^2)\|g\|_{L^\infty(\widetilde{F}_{\mathbf{X}}(A))}^2 \|c - \widehat{c}_{\mathbf{b}}\|_{L^2(\widetilde{F}_{\mathbf{X}}(A))}^2$.

Similar arguments lead to $T_2 \leq M_c \int_{\widetilde{F}_{\mathbf{X}}(A)} g^2(\mathbf{u}) \mathbb{P}(\widehat{c}_{\mathbf{b}}(\mathbf{u}) < m_c/2) d\mathbf{u}$. But, using $(H_{c,low})$,

$$\widehat{c}_{\mathbf{b}}(\mathbf{u}) \leq \frac{m_c}{2} \implies |\widehat{c}_{\mathbf{b}}(\mathbf{u}) - c(\mathbf{u})| \geq \frac{m_c}{2},$$

we deduce

$$\begin{aligned} \mathbb{E}[T_2] &\leq M_c \int_{\widetilde{F}_{\mathbf{X}}(A)} g^2(\mathbf{u}) \mathbb{P}(\widehat{c}_{\mathbf{b}}(\mathbf{u}) - c(\mathbf{u})| \geq m_c/2) d\mathbf{u}, \\ &\leq \frac{4M_c}{m_c^2} \int_{\widetilde{F}_{\mathbf{X}}(A)} g^2(\mathbf{u}) d\mathbf{u} \mathbb{E}\left[\|\widehat{c}_{\mathbf{b}} - c\|_{L^2(\widetilde{F}_{\mathbf{X}}(A))}^2 \right], \end{aligned}$$

by applying the Markov inequality. Gathering the bound for $T_{1,1}$, $T_{1,2}$ and T_2 concludes the proof.

Acknowledgements

We are very grateful to Patricia Reynaud-Bouret for constructive discussions. We thank the anonymous referees for carefully reading the manuscript and for numerous suggestions that improved the paper.

References

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

Autin, F., Le Pennec, E., and Tribouley, K. (2010). Thresholding methods to estimate copula density. *J. Multivariate Anal.*, 101(1):200–222.

Balakrishnan, N. and Lai, C. (2009). *Continuous Bivariate Distributions*. Springer Science+Business Media, New York.

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

Bertin, K. (2005). Sharp adaptive estimation in sup-norm for d-dimensional Hölder classes. *Mathematical Methods in Statistics*, 14:267–298.

Bertin, K., Lacour, C., and Rivoirard, V. (2016). Adaptive pointwise estimation of conditional density function. *Ann. Inst. Henri Poincaré Probab. Stat.*, 52(2):939–980.

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

Chagny, G. (2013). Penalization versus Goldenshluger-Lepski strategies in warped bases regression. *ESAIM Probab. Stat.*, 17:328–358.

Chagny, G. (2015). Adaptive warped kernel estimators. *Scandinavian Journal of Statistics*, 42(2):336–360.

Chagny, G., Comte, F., and Roche, A. (2017). Adaptive estimation of the hazard rate with multiplicative censoring. *J. Statist. Plann. Inference*, 184:25–47.

Chesneau, C. and Willer, T. (2015). Estimation of a cumulative distribution function under interval censoring "case 1" via warped wavelets. *Comm. Statist. Theory Methods*, 44(17):3680–3702.

Comte, F. (2015). *Estimation non-paramétrique*. Spartacus-IDH.

Comte, F. and Lacour, C. (2013). Anisotropic adaptive kernel deconvolution. *Ann. Inst. Henri Poincaré Probab. Stat.*, 49(2):569–609.

Di Bernardino, E., Laloë, T., and Servien, R. (2015). Estimating covariate functions associated to multivariate risks: a level set approach. *Metrika*, 78(5):497–526.

Efromovich, S. (1999). *Nonparametric curve estimation*. Springer Series in Statistics. Springer-Verlag, New York. Methods, theory, and applications.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications*, volume 66 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

Fermanian, J.-D. (2005). Goodness-of-fit tests for copulas. *J. Multivar. Anal.*, 95(1):119–152.

Furer, D. and Kohler, M. (2015). Smoothing spline regression estimation based on real and artificial data. *Metrika*, 78(6):711–746.

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

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

Guyader, A. and Hengartner, N. (2013). On the mutual nearest neighbors estimate in regression. *Journal of Machine Learning Research*, 14:2361–2376.

Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer Series in Statistics. Springer New York.

Hayfield, T. and Racine, J. S. (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software*, 27(5).

Jaworski, P., Durante, F., Härdle, W., and Rychlik, T. (2010). *Copula Theory and Its Applications: Proceedings of the Workshop Held in Warsaw, 25-26 September 2009*. Lecture Notes in Statistics. Springer Berlin Heidelberg.

Kerkyacharian, G., Lepski, O., and Picard, D. (2001). Nonlinear estimation in anisotropic multi-index denoising. *Probability Theory and Related Fields*, 121(2):137–170.

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

Kohler, M., Krzyzak, A., and Walk, H. (2009). Optimal global rates of convergence for nonparametric regression with unbounded data. *J. Statist. Plann. Inference*, 139(4):1286–1296.

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

Lacour, C. and Massart, P. (2016). Minimal penalty for Goldenshluger-Lepski method. *Stochastic Process. Appl.*, 126(12):3774–3789.

Lacour, C., Massart, P., and Rivoirard, V. (2017). Estimator selection: a new method with applications to kernel density estimation. *Sankhya A*, 79(2):298–335.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and its Applications*, 9:141–142.

Neumann, M. (2000). Multivariate wavelet thresholding in anisotropic function spaces. *Statistica Sinica*, 10:399–431.

Ngoc Bien, N. (2014). *Adaptation via des inéqualités d'oracle dans le modèle de régression avec design aléatoire*. PhD thesis, Université d'Aix-Marseille.

Nguyen, M.-L. J. (2018). Nonparametric method for space conditional density estimation in moderately large dimensions. *arXiv preprint arXiv:1801.06477*.

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.

Pham Ngoc, T. M. (2009). Regression in random design and Bayesian warped wavelets estimators. *Electron. J. Stat.*, 3:1084–1112.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8:229–231.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.*, 10(4):1040–1053.

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

Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

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

Yang, S.-S. (1981). Linear combination of concomitants of order statistics with application to testing and estimation. *Annals of the Institute of Statistical Mathematics*, 33(1):463–470.