



HAL
open science

Multivariate adaptive warped kernel estimation

Gaëlle Chagny, Thomas Laloë, Rémi Servien

► **To cite this version:**

Gaëlle Chagny, Thomas Laloë, Rémi Servien. Multivariate adaptive warped kernel estimation. 2017. hal-01616373v1

HAL Id: hal-01616373

<https://hal.science/hal-01616373v1>

Preprint submitted on 13 Oct 2017 (v1), last revised 1 Feb 2019 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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 compactness of the support of the random design, thanks to a “warping” device. 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 proved to be optimal in the oracle and in the minimax sense. The plug-in of these estimates in the regression function estimator provides a fully data-driven estimate. 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 $\mathbf{X} = (X_1, \dots, X_d)$. Assume that we have an independent identically distributed (*i.i.d.* in the sequel) sample $(\mathbf{X}_i, Y_i)_{i=1 \dots n}$ 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 $A \subset \mathbb{R}^d$, 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 countless number of research works frequently based on nonparametric 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), nearest neighbor-type estimators (Stute, 1984; Guyader and Hengartner, 2013)

Among kernel methods, the most popular estimator is the well-known Nadaraya-Watson estimate, defined for model (1) by

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

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 $K_l : \mathbb{R} \rightarrow \mathbb{R}$ 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). To weaken this assumption, Kohler et al. (2009) assume some smoothness properties on the regression function. It requires that the partial derivatives of the regression function r are k -Hölderian with a constant C (for further details see Definition 1 in Kohler et al. 2009).

In another hand, “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 estimates 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 this estimate 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 non-asymptotic 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 is 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 in the paper. 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 paper is thus organized as follows. We explain in Section 2 how to extend the Yang (1981) estimator to the multivariate design framework. For the sake of clarity, we then first concentrate on the simple toy case of known design distribution (Section 3): under mild assumptions, we derive 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 also an optimal convergence rate over possibly anisotropic functional classes. 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 illustrating the different methods 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.

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 $\tilde{F}_{\mathbf{X}} : \mathbf{x} = (x_l)_{l=1, \dots, d} \in \mathbb{R}^d \mapsto (F_1(x_1), \dots, F_d(x_d))$. Assume that $\tilde{F}_{\mathbf{X}}^{-1} : \mathbf{u} \in [0, 1]^d \mapsto (F_1^{-1}(u_1), \dots, F_d^{-1}(u_d))$ exists, and let

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

in such a way that $r = g \circ \tilde{F}_{\mathbf{X}}$. 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

$$\mathbf{u} \mapsto \sum_{i=1}^n Y_i K_{\mathbf{h}}(\mathbf{u} - \tilde{F}_{\mathbf{X}}(\mathbf{X}_i)) \quad (3)$$

to estimate g , and it remains to compound by the empirical counterpart of $\tilde{F}_{\mathbf{X}}$ 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

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

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

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

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 given by (3) estimates cg and not g (see the computation (8) below). As a consequence, we propose to set, as an estimator for g ,

$$\hat{g}_{\mathbf{h}}(\mathbf{u}) = \frac{1}{n\hat{c}(\mathbf{u})} \sum_{i=1}^n Y_i K_{\mathbf{h}}(\mathbf{u} - \hat{\tilde{F}}_{\mathbf{X}}(\mathbf{X}_i)), \quad \mathbf{u} \in [0, 1]^d,$$

where \hat{c} is an estimator of the copula density. We denote by $\hat{\tilde{F}}_{\mathbf{X}} : \mathbb{R}^d \rightarrow [0; 1]^d$ the empirical multivariate marginal c.d.f.:

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

and finally set

$$\hat{r}_{\mathbf{h}}(\mathbf{x}) = \hat{g}_{\mathbf{h}} \circ \hat{\tilde{F}}_{\mathbf{X}}(\mathbf{x}) = \frac{1}{n\hat{c}(\hat{\tilde{F}}_{\mathbf{X}}(\mathbf{x}))} \sum_{i=1}^n Y_i K_{\mathbf{h}}(\hat{\tilde{F}}_{\mathbf{X}}(\mathbf{x}) - \hat{\tilde{F}}_{\mathbf{X}}(\mathbf{X}_i)) \quad (6)$$

to rebuild our target function r from the data. In the sequel, we denote by $\|\cdot\|$ the (unweighted) L^2 -norm on $L^2(\mathbb{R}^d)$ and, more generally, 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 the 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. $\tilde{F}_{\mathbf{X}}$ are consequently considered to be known. Thus, (6) becomes

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

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

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

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 $\hat{g}_{\mathbf{h}}$ is suitable to estimate g , and $\hat{r}_{\mathbf{h}}$ suits well to recover the target r .

3.2. Risk of the estimator with fixed bandwidth. A global integrated criterion is considered to study the properties of our estimates. Let $\|\cdot\|_{f_{\mathbf{X}}}$ 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,

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

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

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

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

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

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

- $(H_{c,low})$: The copula density is lower bounded: $\exists m_C > 0, \forall \mathbf{u} \in [0, 1]^d, c(\mathbf{u}) \geq m_C$.
- $(H_{cg,\beta})$: The function $cg\mathbf{1}_{\tilde{F}_{\mathbf{X}}(A)}$ belongs to an anisotropic Nikol'skiĭ ball $\mathcal{N}_2(\beta, L)$, with $L > 0$ and $\beta = (\beta_1, \dots, \beta_d) \in (\mathbb{R}_+^*)^d$ (Nicol'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 $l \in \{1, \dots, d\}$, $\|\partial^{[\beta_l]} f / (\partial x_l)^{[\beta_l]}\|_{L^2(\mathbb{R}^d)} \leq L$,
 - (ii) for all $l \in \{1, \dots, d\}$ and $t \in \mathbb{R}$,

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

- $(H_{K,\ell})$: The kernel K is of order $\ell \in (R_+)^d$, *i.e.*
- (i) $\forall l \in \{1, \dots, d\}, \forall k \in \{1, \dots, \ell_l\}, \int_{\mathbb{R}^d} x_l^k K(\mathbf{x}) d\mathbf{x} = 0$.
 - (ii) $\forall l \in \{1, \dots, d\}, \int_{\mathbb{R}^d} (1 + x_l)^{\ell_l} K(\mathbf{x}) d\mathbf{x} < \infty$.

Note that Assumption $(H_{c,low})$ is verified for the Frank copula. For other copulas we restrict the estimation to $A =]\varepsilon, +\infty[^d$, for a fixed $\varepsilon > 0$, to exclude problematic points (see Balakrishnan and Lai 2009 for more details on other specific copula densities). Assumptions $(H_{cg,\beta})$ and $(H_{K,\ell})$ are classical for nonparametric multivariate kernel estimation (Goldenshluger and Lepski, 2011; Comte and Lacour, 2013) to control the bias term $B(\mathbf{h})$. Assumption $(H_{K,\ell})$ is not restrictive since a wide range of kernels could be chosen. In $(H_{cg,\beta})$, the index β measures the smoothness of the function cg . The difficulty 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 $\beta_l \leq 1, l \in \{1, \dots, d\}$. Furthermore, 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 Kerkycharian and Picard (2004). Using these assumptions we state the following proposition.

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,*

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

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 keep $B(\mathbf{h})$ in the inequality (in this case, Assumptions $(H_{cg,\beta})$ and $(H_{K,\ell})$ are not required). But now we could immediately deduce the following convergence rate.

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

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

where $\bar{\beta}$ is the harmonic mean of β_1, \dots, β_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 of the function cg . 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

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

with

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

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

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

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

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

and

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

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

$$\exists \alpha_0 > 0, \kappa_1 > 0, \sum_{\mathbf{h} \in \mathcal{H}_n} \frac{1}{h_1 \dots h_d} \leq \kappa_1 n^{\alpha_0} \quad (13)$$

$$\text{and } \forall \kappa_1 > 0, \exists C_0 > 0, \sum_{\mathbf{h} \in \mathcal{H}_n} \exp\left(-\frac{\kappa_1}{h_1 \dots h_d}\right) \leq C_0.$$

These assumptions are very common to derive such estimators (Comte and Lacour, 2013; Chagny, 2015). For example, $\mathcal{H}_n = \{k_1^{-1} \dots k_d^{-1}, k_l \in \{1, \dots, \lfloor n^{1/r} \rfloor\}, l = 1, \dots, d\}$ 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 $[0, 1]^d$: $\exists M_C > 0, \forall \mathbf{u} \in [0, 1]^d, c(\mathbf{u}) \leq M_C$.

The assumption $(H_{c,high})$ is quite restrictive for copula density estimation. However, it will also be required for copula density estimation below, and it is classical for adaptive density estimation purpose. It is satisfied also by the Frank copula density for example. Moreover, the same upper-bound is assumed in Autin et al. (2010). It is then possible to set the following upper bound.

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*

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

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). 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 $\widehat{r}_{\mathbf{h}}^-$. The difference is that the smoothness index β is not required to build $\widehat{r}_{\mathbf{h}}^-$: our selected estimator automatically adapts to the unknown smoothness of the function cg , as it is stated in the following corollary.

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*

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

where $\bar{\beta}$ is the harmonic mean of β_1, \dots, β_d : $d\bar{\beta}^{-1} = \beta_1^{-1} + \dots + \beta_d^{-1}$.

To switch from $\widehat{r}_{\mathbf{h}}^-$ to $\widehat{r}_{\mathbf{h}}$, it remains then to first replace the unknown expectation $\mathbb{E}[Y_1^2]$ by its empirical counterpart $\sum_{i=1}^n Y_i^2$ 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). We do not go into details, to avoid burdening the text, but also because the design distribution (and thus the copula) is supposed to be known in this section.

4. COPULA DENSITY ESTIMATION

The estimator defined by (6) involves an estimator \widehat{c} of the copula density c that was assumed to be known in the previous section. This section is devoted to the question of copula density estimation. 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 $W_l: \mathbb{R} \rightarrow \mathbb{R}$ such that $\int_0^1 W_l(u)du = 1$, $l \in \{1, \dots, d\}$. Let us introduce

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

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 $\widehat{F}_{\mathbf{X}}$ 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:

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

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

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

and one of the following constraints

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

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

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

with

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

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 $(H_{c,high})$, 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*

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

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 8.3 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).

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'skii ball $\mathcal{N}_2(\alpha, L')$ for $L' > 0$ and $\alpha = {}^t(\alpha_1, \dots, \alpha_d) \in (\mathbb{R}_+^*)^d$, and if the kernel W is of order $\ell \in \mathbb{R}_+^d$ such that $\ell_j \geq \lfloor \alpha_j \rfloor$ for $j = 1, \dots, d$, (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 estimates 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 $\widehat{c}_{\widehat{\mathbf{b}}}$ (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. Let us plug in $\hat{r}_{\mathbf{h}}$ the estimate $\hat{c}_{\mathbf{b}}$: for any $\mathbf{b}, \mathbf{h} > 0$, under Assumption $(H_{c,low})$,

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

This means that $\hat{r}_{\mathbf{h},\mathbf{b}}(\mathbf{x}) = ((c \times \hat{g}_{\mathbf{h}})/\hat{c}_{\mathbf{b}}) \circ \tilde{F}_{\mathbf{X}}(\mathbf{x}) \mathbf{1}_{\hat{c}_{\mathbf{b}}(\tilde{F}_{\mathbf{X}}(\mathbf{x})) \geq m_c/2}$. 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. To avoid making the proofs more technical and cumbersome, and since the minimum of classical copula density are very small (see the example of the Franck copula used for the simulation study), 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 $\hat{r}_{\mathbf{h}}$ and $\hat{c}_{\mathbf{b}}$.

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

$$\mathbf{E}[\|\hat{r}_{\mathbf{h},\mathbf{b}} - r\|_{f_{\mathbf{X}}}^2] \leq \frac{4M_c}{m_c^2} \left\{ 2M_c \mathbf{E}[\|\hat{r}_{\mathbf{h}} - r\|_{f_{\mathbf{X}}}^2] + (2\|g\|_{L^\infty(\tilde{F}_{\mathbf{X}}(A))}^2 + \|g\|_{L^2(\tilde{F}_{\mathbf{X}}(A))}^2) \mathbf{E} \left[\|\hat{c}_{\mathbf{b}} - c\|_{L^2([0,1]^d)}^2 \right] \right\}.$$

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 a fully computable estimator, one needs to replace the c.d.f. $\tilde{F}_{\mathbf{X}}$ 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 Kerkycharian 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).

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). Then 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.

To study the performances of our estimators \hat{r} ($\hat{r} = \hat{r}_{\mathbf{b}}, \hat{r}_{\hat{\mathbf{h}}}, \hat{r}_{\mathbf{h},\mathbf{b}}$) we compute a Monte Carlo approximation (with 1000 iterations of samples independent from the data sample used to compute \hat{r}) of a relative L_2 -risk, namely the Relative Mean Square Error (RMSE):

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

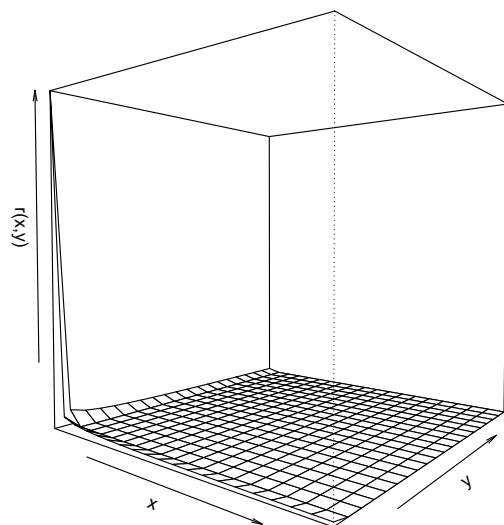


FIGURE 1. Regression function: $r(x_1, x_2) = 1/\sqrt{x_1 x_2}$ for $(x_1, x_2) \in \mathbb{R}^+ \times \mathbb{R}^+$.

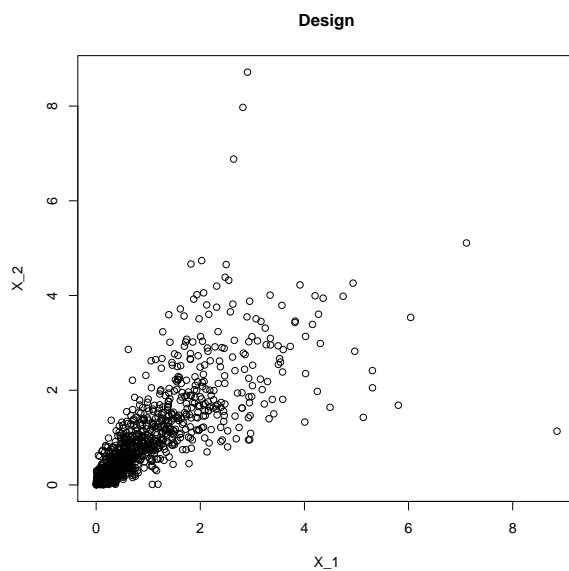


FIGURE 2. Illustration of the design : Frank copula with parameter 10 and exponential marginals.

Finally, we confront our estimators with the classical Nadaraya-Watson kernel estimator (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), that can deal with unbounded data.

6.1. Simulation with fixed bandwidths for the copula and the regression estimators.

In this section we compute the estimators $\hat{r}_{\mathbf{h},\mathbf{b}}$ and $\hat{c}_{\mathbf{b}}$ 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. Moreover, the minimum value for the Frank copula density we use is $m_c = 10e^{-20}/(1 - e^{-10}) \approx 10^{-8}$. We replace this value by $\hat{m}_c = 0$ in the definition of $\hat{r}_{\mathbf{h},\mathbf{b}}$. We provide corresponding boxplots for sample sizes $n = 100, 500$ and 1000 . For each sample size, 50 RMSE values (computed from independent samples) are plotted.

First of all we study the impact of the estimation of the marginal distributions of the design (*i.e.* the switching from $\tilde{F}_{\mathbf{X}}$ to $\hat{F}_{\mathbf{X}}$) on Figure 3. The estimation of the marginal distributions is carried out using the classical empirical cumulative distribution function with the function *ecdf* of the software R.

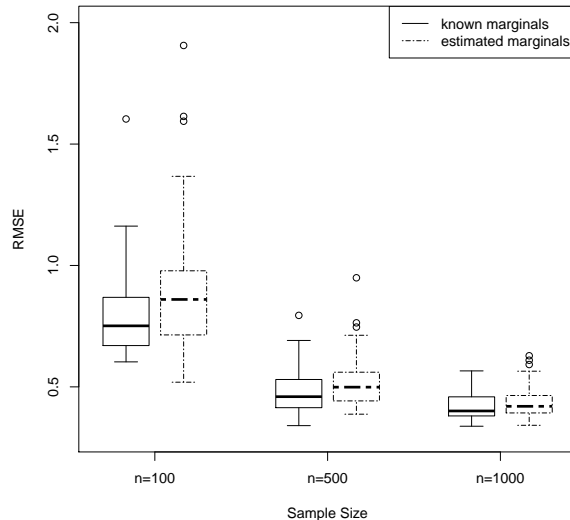


FIGURE 3. Effect of the marginal distributions estimation.

The results are similar in both cases. Using the empirical counterpart $\hat{F}_{\mathbf{X}}$ instead of the true c.d.f. $\tilde{F}_{\mathbf{X}}$ does not seem to affect the quality of the final estimator. This kind of results is not very surprising as the estimator $\hat{F}_{\mathbf{X}}$ is widely known to be a very good estimate of $\tilde{F}_{\mathbf{X}}$.

From now on, the marginal distributions are thus estimated for all the presented results. Then we confront our estimator to the well-known Nadaraya-Watson estimate and to the one proposed by Kohler et al. (2009). Results are gathered on Figure 4.

Our estimator seems to have better performances : the RMSE (mean and variance) is nearly divided by 5 compared to the ones of Kohler’s estimator and by 2 compared to the Nadaraya-Watson estimate. The global decreases of the RMSE with respect to the sample size n seems to be the same for the three estimates.

6.2. Simulations with data-driven bandwidth for the copula estimator and fixed bandwidth for the regression estimator.

In this subsection, the bandwidths of the copula

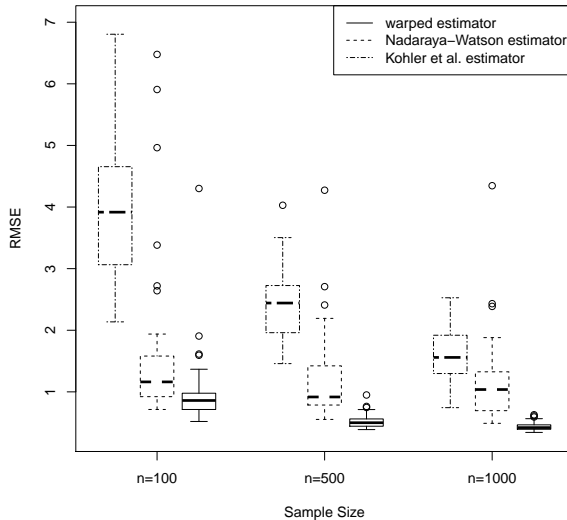


FIGURE 4. Comparison of the RMSE of our regression estimator $\hat{r}_{\mathbf{h}, \mathbf{b}}$ with the Nadaraya-Watson estimator and the estimator of Kohler et al. (2009).

density estimator are selected using the Goldenshuger-Lepski procedure detailed in Section 4 and the bandwidths of the regression estimator are still given by $h_1 = h_2 = (\log(n)/n)^{0.5}$. 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 $\tilde{\mathbf{b}} \in \arg \min_{\mathbf{b} \in \mathcal{B}_n} \{\hat{B}_c(\mathbf{b}) + 2V_c(\mathbf{b})\}$ instead of $\hat{\mathbf{b}}$, and to compute the new final estimate $\hat{c}_{\tilde{\mathbf{b}}}$. The reason are mainly technical, and we refer to Section 5 of Lacour and Massart (2016) for details. Figure 5 thus presents the calibration results (risk of $\hat{c}_{\tilde{\mathbf{b}}}$ 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 $\hat{c}_{\tilde{\mathbf{b}}}$ and to evaluate its performances. On Figure 6, we compare its risk to the risk of an estimator with the fixed bandwidths chosen in the previous section. The automatic selection of the bandwidth really improves the results : the RMSE is divided by two for the estimator with data-driven bandwidths for the copula.

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

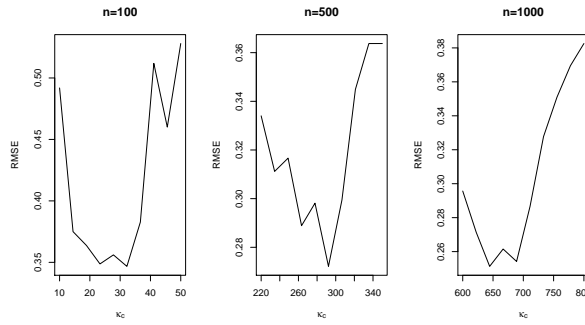


FIGURE 5. RMSE for $\hat{c}_{\mathbf{b}}$ with respect to the constant κ_c for different sample sizes n .

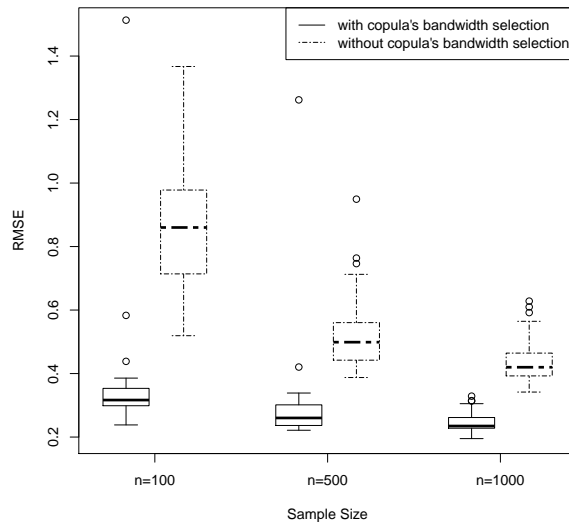


FIGURE 6. Evolution of the RMSE of the regression estimator with respect to the sample size with and without bandwidth selection for the copula plug-in estimate.

framework (non-asymptotic risk bound and optimal convergence rate) are proved: compare for example Theorem 3.1 and Corollary 3.2 to Theorem 1 and Corollary 1 of Chagny (2015). 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 kernel estimates. Section 5 paves the way for a future study of the plug-in estimator, which is 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 future challenge is to propose a data-driven selection method for the bandwidth of the regression estimator computed with an adaptive copula estimate. A reflexion is under way on determining if it is possible to replace c in the Goldenshluger-Lepski estimation of the

bias term of the risk (see (10)) by $\widehat{c}_{\mathbf{b}}$. This probably implies the modification of the variance term (11) to penalize the plug-in, but is not straightforward at all.

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*

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

Consider M_1 , v , and H , such that

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

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

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

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

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

8.1. Proof of Proposition 3.1. The variance term is

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

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

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

and

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

Similar computations lead to $\mathbb{E}[\varepsilon^2]\mathbb{E}[K_{\mathbf{h}}^2(\mathbf{u} - \tilde{F}_{\mathbf{X}}(\mathbf{X}))] \leq \mathbb{E}[\varepsilon^2]\|K\|^2/(h_1 \dots h_d)$. 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 Theorem 3.1. First, the loss function of the selected estimator can be written

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

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

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

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 $\tilde{B}(\mathbf{h})$. Let us begin by splitting the norm involved in its definition. We have

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

with

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

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

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

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

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

We have that

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

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

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

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 $l \in \{(1), (2, 1), (2, 2)\}$, $\nu_{n,\mathbf{h}}^{(l)}(t) = \frac{1}{n} \sum_{i=1}^n \varphi_{t,\mathbf{h},i}^{(l)} - \mathbb{E}[\varphi_{t,\mathbf{h},i}^{(l)}]$, and

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

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 $\tilde{V}(\mathbf{h}) = \tilde{V}_1(\mathbf{h}) + \tilde{V}_2(\mathbf{h})$, with $\tilde{V}_1(\mathbf{h}) = \kappa_0 \mathbb{E}[r^2(X_1)] / (m_c n h_1 \dots h_d)$ and $\tilde{V}_2(\mathbf{h}) = \kappa_0 \mathbb{E}[\varepsilon_1^2] / (m_c n h_1 \dots h_d)$, and consequently

$$\begin{aligned} & \mathbb{E} \left[\max_{\mathbf{h}' \in \mathcal{H}_n} \left\{ \left\| c \hat{g}_{\mathbf{h}'} - K_{\mathbf{h}'} \star (c g \mathbf{1}_{\tilde{F}_{\mathbf{X}}(A)}) \right\|_{L^2(\tilde{F}_{\mathbf{X}}(A))}^2 - \frac{m_c \tilde{V}(\mathbf{h}')}{3(1 + \|K\|_{L^1([0,1]^d)})} \right\}_+ \right] \\ & \leq 3 \sum_{\mathbf{h} \in \mathcal{H}_n} \left\{ \mathbb{E} \left[\left(\sup_{t \in \bar{S}(0,1)} \left(\nu_{n,\mathbf{h}}^{(1)}(t) \right)^2 - \frac{m_c \tilde{V}_1(\mathbf{h})}{9(1 + \|K\|_{L^1([0,1]^d)})} \right)_+ \right] \right. \\ & \left. + \mathbb{E} \left[\left(\sup_{t \in \bar{S}(0,1)} \left(\nu_{n,\mathbf{h}}^{(2,1)}(t) \right)^2 - \frac{m_c \tilde{V}_2(\mathbf{h})}{9(1 + \|K\|_{L^1([0,1]^d)})} \right)_+ \right] + \mathbb{E} \left[\sup_{t \in \bar{S}(0,1)} \left(\nu_{n,\mathbf{h}}^{(2,2)}(t) \right)^2 \right] \right\}. \quad (22) \end{aligned}$$

Then we obtain, using (H_ε) ,

$$\sum_{\mathbf{h} \in \mathcal{H}_n} \mathbb{E} \left[\sup_{t \in \bar{S}(0,1)} \left(\nu_{n,\mathbf{h}}^{(2,2)}(t) \right)^2 \right] \leq \|K\|^2 \frac{1}{n \kappa_n^p} \mathbb{E}[|\varepsilon|^2 + p] \sum_{\mathbf{h} \in \mathcal{H}_n} \frac{1}{h_1 \dots h_d}. \quad (23)$$

For the other two terms in (22), we apply the Talagrand inequality (Theorem 8.1). This leads to

$$\begin{aligned} \sum_{\mathbf{h} \in \mathcal{H}_n} \mathbb{E} \left[\left(\sup_{t \in \bar{S}(0,1)} \left(\nu_{n,\mathbf{h}}^{(1)}(t) \right)^2 - \delta_1 \frac{\mathbb{E}[r^2(X_1)]}{n h_1 \dots h_d} \right)_+ \right] & \leq C \left(\frac{1}{n} \sum_{\mathbf{h} \in \mathcal{H}_n} \exp \left(-\frac{c_1}{h_1 \dots h_d} \right) \right. \\ & \left. + \frac{1}{n^2} \exp(-c_2 \sqrt{n}) \sum_{\mathbf{h} \in \mathcal{H}_n} \frac{1}{h_1 \dots h_d} \right), \quad (24) \end{aligned}$$

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

$$\sum_{\mathbf{h} \in \mathcal{H}_n} \mathbb{E} \left[\left(\sup_{t \in \tilde{S}(0,1)} \left(\nu_{n,\mathbf{h}}^{(2,1)}(t) \right)^2 - \delta_2 \frac{\mathbb{E}[\varepsilon_1^2]}{nh_1 \cdots h_d} \right)_+ \right] \leq C \left(\frac{1}{n} \sum_{\mathbf{h} \in \mathcal{H}_n} \exp \left(-\frac{c_3}{h_1 \cdots h_d} \right) \right. \\ \left. + \frac{\kappa_n^2}{n^2} \exp \left(-c_4 \frac{\sqrt{n}}{\kappa_n} \right) \sum_{\mathbf{h} \in \mathcal{H}_n} \frac{1}{h_1 \cdots h_d} \right). \quad (25)$$

The assumptions on the collection \mathcal{H}_n permit to deduce that the right hand side of (23), (24), and (25) are less than C/n for a constant C . As soon as

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

which is the case if κ_0 is large enough, (22) becomes

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

and consequently

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

which ends the proof.

8.3. 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 $\mathbf{b} \in \mathcal{B}_n$,

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

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

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

We roughly upper-bound

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

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

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

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 $\|t\|_{L^2([0,1]^d)}^2 = 1$), $\langle \cdot, \cdot \rangle_{L^2([0,1]^d)}$ 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 $\varphi_{t,c,\mathbf{b}}(\mathbf{x}) = \int_{[0,1]^d} W_{\mathbf{b}}(\mathbf{u} - \tilde{F}(\mathbf{x}))t(\mathbf{u})d\mathbf{u}$. 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:

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

We only detail the computation of v_c : firstly $n^{-1} \sum_{i=1}^n \text{Var}(\varphi_{t,c,\mathbf{b}}(X_i)) = \text{Var}(\varphi_{t,c,\mathbf{b}}(X_1)) \leq \mathbb{E}[\varphi_{t,c,\mathbf{b}}^2(X_1)]$. 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 $r = 2$:

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

since $\|t\|_{L^2([0,1]^d)}^2 = 1$. 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)

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

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

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

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

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

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

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

First we have

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

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

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

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

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

thus, $T_{1,2} \leq (4M_c/m_c^2) \|g\|_{L^\infty(\widetilde{F}_{\mathbf{X}}(A))}^2 \|c - \widehat{c}_{\mathbf{b}}\|_{L^2([0,1]^d)}^2$.

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

$$\begin{aligned} \mathbb{E}[T_2] &\leq M_c \int_{\widetilde{F}_{\mathbf{X}}(A)} g^2(\mathbf{u}) \mathbb{P}(\widehat{c}_{\mathbf{b}}(\mathbf{u}) - c(\mathbf{u}) \geq m_c/2) d\mathbf{u}, \\ &\leq \frac{4M_c}{m_c^2} \int_{\widetilde{F}_{\mathbf{X}}(A)} g^2(\mathbf{u}) d\mathbf{u} \mathbb{E} \left[\|\widehat{c}_{\mathbf{b}} - c\|_{L^2([0,1]^d)}^2 \right], \end{aligned}$$

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

REFERENCES

- Antoniadis, A., Grégoire, G., and Vial, P. (1997). Random design wavelet curve smoothing. *Statist. Probab. Lett.*, 35(3):225–232.
- Autin, F., Le Penec, 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., 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.
- Kerkycharian, 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.
- Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and its Applications*, 9:141–142.
- Ngoc Bien, N. (2014). *Adaptation via des inégalités d’oracle dans le modèle de régression avec design aléatoire*. PhD thesis, Université d’Aix-Marseille.
- 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.

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