Bernoulli **22**(4), 2016, 2177–2208
DOI: 10.3150/15-BEJ725

Integral approximation by kernel smoothing

BERNARD DELYON[1] and FRANÇOIS PORTIER[2]

[1]*Institut de recherches mathématiques de Rennes (IRMAR), Campus de Beaulieu, Université de Rennes 1, 35042 Rennes Cédex, France. E-mail: bernard.delyon@univ-rennes1.fr*
[2]*Institut de Statistique, Biostatistique et Sciences Actuarielles (ISBA), Université catholique de Louvain, Belgique. E-mail: francois.portier@gmail.com*

Let $(X_1, \dots, X_n)$ be an i.i.d. sequence of random variables in $\mathbb{R}^d$, $d \geq 1$. We show that, for any function $\varphi : \mathbb{R}^d \to \mathbb{R}$, under regularity conditions,

$$n^{1/2}\left(n^{-1}\sum_{i=1}^{n}\frac{\varphi(X_i)}{\widehat{f}(X_i)} - \int \varphi(x)\,dx\right) \xrightarrow{\mathbb{P}} 0,$$

where $\widehat{f}$ is the classical kernel estimator of the density of X_1. This result is striking because it speeds up traditional rates, in root n, derived from the central limit theorem when $\widehat{f} = f$. Although this paper highlights some applications, we mainly address theoretical issues related to the later result. We derive upper bounds for the rate of convergence in probability. These bounds depend on the regularity of the functions φ and f, the dimension d and the bandwidth of the kernel estimator $\widehat{f}$. Moreover, they are shown to be accurate since they are used as renormalizing sequences in two central limit theorems each reflecting different degrees of smoothness of φ. As an application to regression modelling with random design, we provide the asymptotic normality of the estimation of the linear functionals of a regression function. As a consequence of the above result, the asymptotic variance does not depend on the regression function. Finally, we debate the choice of the bandwidth for integral approximation and we highlight the good behavior of our procedure through simulations.

Keywords: central limit theorem; integral approximation; kernel smoothing; nonparametric regression

1. Introduction

Let $(X_1, \dots, X_n)$ be an i.i.d. sequence of random variables in $\mathbb{R}^d$, $d \geq 1$. We show that, for any function $\varphi : \mathbb{R}^d \to \mathbb{R}$, under regularity conditions,

$$n^{-1}\sum_{i=1}^{n}\frac{\varphi(X_i)}{\widehat{f}^{(i)}(X_i)} - \int \varphi(x)\,dx = o_{\mathbb{P}}\big(n^{-1/2}\big), \tag{1}$$

where $\widehat{f}^{(i)}$ is the classical leave-one-out kernel estimator of the density of X_1 say f, defined by

$$\widehat{f}^{(i)}(x) = \big((n-1)h^d\big)^{-1}\sum_{1\leq j\leq n, j\neq i} K\big(h^{-1}(x - X_j)\big) \qquad \text{for every } x \in \mathbb{R}^d,$$

where K is a d-dimensional kernel and where h, called the bandwidth, needs to be chosen and will certainly depend on n. Result (1) and the central limit theorem lead to the following reasoning: when estimating the integral of a function that is evaluated on a random grid (X_i), whether

f is known or not, using a kernel estimator of f provides better convergence rates than using f itself.

Result (1) certainly has some consequences in the field of integral approximation. In this area, many deterministic as well as random methods are available. Accuracy with respect to computational time is the usual trade-off that allows to compare them. The advantages of random over deterministic framework lie in their stability in high-dimensional settings. For a comprehensive comparison between both approaches, we refer to [8]. Among random methods, *importance sampling* is a widely used technique that basically reduces the variance of the classical Monte–Carlo integration through a good choice of the sampling distribution f, called the sampler. Estimators are unbiased having the form $n^{-1}\sum_{i=1}^{n}\varphi(X_i)/f(X_i)$ with $X_i \sim f$. Regarding the mean squared error (MSE), the optimal sampler f^* is unique and depends on φ (see Theorem 6.5 in [8], page 176). Among others, parametric [18] and nonparametric [25] studies focused on the estimation of the optimal sampler. Equation (1) indicates a new weighting of the observations $\varphi(X_1), \ldots, \varphi(X_n)$. Each weight $\widehat{f}^{(i)}(X_i)$ reflects how isolated is the point X_i among the sample. Therefore, our estimator takes into account this information by giving more weight to an isolated point. In summary our procedure, which is adaptive to the design points enjoys the following advantages:

- Faster than root n rates,
- one-step estimation based on a unique sample $(X_1, \ldots, X_n)$,
- each X_i drawn from f, possibly unknown.

To the best of our knowledge, when the design is not controlled, no such rates have been obtained.

In many semiparametric problems, it has been an important issue to construct root n estimators, possibly efficient [1], that rely on a kernel estimator of the nuisance parameter. Among others, it was addressed by Stone in [21] in the case of the estimation of a location parameter, by Robinson in [19] in the *partially linear regression model*, or by Härdle and Stoker in [16] studying the *single index model*. The result in equation (1), which would be seen as a superefficient estimator in the Le Cam's theory, cannot be linked actually to this theory since the quantity of interest $\int \varphi(x)\,dx$ does not depend on the distribution of X_1. As a result, the link between our work and the semiparametric literature relies mainly on the plug-in strategy we employed, by substituting the density f by a kernel estimator.

In this paper, we propose a comprehensive study of the convergence stated in equation (1). A similar result was originally stated by Vial in [24] (Chapter 7, equation (7.27)), as a lemma in the context of the *multiple index model*. To the best of our knowledge, this type of asymptotic result has not been addressed yet as a particular problem. Our theoretical aim is to extend result (1) by:

(A) Being more precise about the upper bounds: How does the dimension d, the window h, the regularity of φ and f, impact these bounds?

(B) Showing central limit theorems by specifying the regularity of φ.

To achieve this program, we need to introduce a corrected version of the estimate (1) for which the bias has been reduced. First, the corrected estimator is shown to have better rates of convergence than the initial one. Second, it is shown to be asymptotically normal with rates $nh^{d/2}$ in the case where φ is very regular, and with rates $(nh^{-1})^{1/2}$ in a special case in which φ jumps at

the boundary of its support. To compute the asymptotic distribution, we rely on the paper by Hall [12], where a central limit theorem for completely degenerate U-statistics has been obtained. An important point is that we have succeeded in proving our result with much weaker assumptions on the regularity of φ than on the regularity of f. For instance, equation (1) may hold even when φ has some jumps. However, the estimation of f is subject to the *curse of dimensionality*, that is, f is required to be smooth enough regarding the dimension of X_1.

Our aim is also to link equation (1) to nonparametric regression with random design, that is, the model $Y_i = g(X_i) + \sigma(X_i)e_i$ with g unknown and e_i i.i.d. with $e_i \perp\!\!\!\perp X_i$. In particular, we obtain the asymptotic normality for the estimators of the linear functionals of g. Thanks to the fast rates detailed previously, the asymptotic distribution does not depend on the function g.

The paper is organized as follows. Section 2 deals with technical issues related to equation (1). In particular, we examine the rates of convergence of (1) according to the choice of the bandwidth, the dimension and the regularity of the functions φ and f. Section 3 is dedicated to the convergence in distribution of our estimators. In Section 4, we show how to apply equation (1) to the problem of the estimation of the linear regression functionals. Finally, in Section 5, we give some simulations that compare our method with the traditional Monte–Carlo procedure for integration. The proofs and the technicalities are postponed in Section 6 at the end of the paper.

2. Rates of convergences faster than root *n*

In this section, we first provide upper bounds on the rates of convergence in probability of our estimators. Our main purpose is to show that rates faster than root n hold in a wide range of parameter settings for the estimation of $\int \varphi(x)\,dx$. Second, we argue that those faster than root n rates have no reason to hold when estimating other functionals of the type $f \mapsto \int T(x, f(x))\,dx$.

2.1. Main result

Let $Q \subset \mathbb{R}^d$ be the support of φ. The quantity $I(\varphi) = \int \varphi(x)\,dx$ is estimated by

$$\widehat{I}(\varphi) = n^{-1} \sum_{i=1}^{n} \frac{\varphi(X_i)}{\widehat{f}^{(i)}(X_i)}.$$

Actually, this estimator can be modified in such a way that the leading error term of its expansion vanishes asymptotically (see Remark 9 for more details). For that, we define $\widehat{v}^{(i)}(x)$ as

$$\widehat{v}^{(i)}(x) = \big((n-1)(n-2)\big)^{-1} \sum_{1 \le j \le n,\, j \ne i} \big(h^{-d} K\big(h^{-1}(x - X_j)\big) - \widehat{f}^{(i)}(x)\big)^2.$$

It is, up to a factor $(n-1)^{-1}$, the leave-one-out estimator of the variance of $h^{-d} K(h^{-1}(x - X_j))$. The corrected estimator is

$$\widehat{I}_c(\varphi) = n^{-1} \sum_{i=1}^{n} \frac{\varphi(X_i)}{\widehat{f}^{(i)}(X_i)} \left(1 - \frac{\widehat{v}^{(i)}(X_i)}{\widehat{f}^{(i)}(X_i)^2}\right).$$

To state our main result about the convergences of $\widehat{I}(\varphi)$ and $\widehat{I}_c(\varphi)$, we define the Nikolski class of functions $\mathcal{H}(s, M)$ of regularity $s = k + \alpha$, $k \in \mathbb{N}$, $0 < \alpha \leq 1$, with constant $M > 0$, as the set of bounded and k times differentiable functions φ whose all derivatives of order k satisfy [23]

$$\int \left(\varphi^{(l)}(x+u) - \varphi^{(l)}(x)\right)^2 dx \leq M|u|^{2\alpha}, \qquad l = (l_1, \ldots, l_d), \sum_{i=1}^{d} l_i \leq k,$$

where $|\cdot|$ stands for the Euclidean norm and the l_i's are natural integer. Be careful that k cannot be equal to s. We say that K is a kernel with order $r \in \mathbb{N}^*$ as soon as $K : \mathbb{R}^d \mapsto \mathbb{R}$ is bounded and satisfies

$$\int K(x)\,dx = 1, \qquad \int x^l K(x)\,dx = 0, \qquad l = (l_1, \ldots, l_d), 0 < \sum_{i=1}^{d} l_i \leq r - 1$$

with the notation $x^l = x_1^{l_1} \times \cdots \times x_d^{l_d}$. The following assumptions are needed to show our first result, they are discussed after the statement.

- (A1) For some $s > 0$ and $M > 0$, the support of φ is a compact set $Q \subset \mathbb{R}^d$ and φ is $\mathcal{H}(s, M)$ on $\mathbb{R}^d$.
- (A2) For some integer $r \geq 1$, the variable X_1 has a bounded density f on $\mathbb{R}^d$ such that its rth order derivatives are bounded.
- (A3) For every $x \in Q$, $f(x) \geq b > 0$.
- (A4) The kernel K has order r and $\int K(x)\,dx = 1$. Moreover, there exists $m_1 > 0$ and $m_2 > 0$ such that, for every $x \in \mathbb{R}^d$, $|K(x)| \leq m_1 \exp(-m_2|x|)$. In addition K is symmetric: $K(x) = K(-x)$.

The next theorem is proved in Section 6.

Theorem 1. *Under the assumptions* (A1) *to* (A2), *we have the following* $O_{\mathbb{P}}$ *estimates*

$$n^{1/2}\left(\widehat{I}(\varphi) - I(\varphi)\right) = O_{\mathbb{P}}\left(h^s + n^{1/2}h^r + n^{-1/2}h^{-d}\right), \tag{i}$$

$$n^{1/2}\left(\widehat{I}_c(\varphi) - I(\varphi)\right) = O_{\mathbb{P}}\left(h^s + n^{1/2}h^r + n^{-1/2}h^{-d/2} + n^{-1}h^{-3d/2}\right), \tag{ii}$$

which are valid if the sums inside the $O_{\mathbb{P}}$*'s tend to zero.*

Remark 1. Assumption (A2) about the smoothness of f is crucial to guarantee a rate faster than root n in Theorem 1. On the one hand, one needs $r > d$ to obtain such a rate in equation (i), on the other hand, $r > 3d/4$ suffices to get this rate in equation (ii). Otherwise there does not exist h such that the bounds in Theorem 1 go to 0. This phenomenon is often referred as the *curse of dimensionality*.

In equation (i) (resp., (ii)), when $h \propto n^{-\gamma}$, the best choice of γ depends on r and s; it balances two of the three (resp., four) terms while letting the other one(s) smaller. Precise rate acceleration for each situation is given in Table 1.

Table 1. Best acceleration of convergence rate in Theorem 1. Best rate acceleration $n^{-\beta}$ obtained with $h \propto n^{-\gamma}$

	β	γ
Equation (i)		
$2s \le r - d$	$\frac{s}{2(s+d)}$	$\frac{1}{2(s+d)}$
$0 < r - d \le 2s$	$\frac{(r-d)}{2(r+d)}$	$\frac{1}{r+d}$
Equation (ii)		
$d \le r - d/2 \le 2s$	$\frac{(r-d/2)}{2r+d}$	$\frac{1}{r+d/2}$
$d \le 2s \le r - d/2$	$\frac{s}{2s+d}$	$\frac{1}{2s+d}$
$r \le 3d/2$ and $0 < 4r - 3d \le 6s$	$\frac{4r-3d}{2(3d+2r)}$	$\frac{3}{3d+2r}$
$2s \le d$ and $6s \le 4r - 3d$	$\frac{2s}{2s+3d}$	$\frac{2}{2s+3d}$

As in many semiparametric problems (see, e.g., [16], Section 4.1), our estimator of f is suboptimal with respect to the density estimation problem (see [22]). Indeed, to achieve the optimal rates in density estimation one would need to take $h \propto n^{-1/(2r+d)}$ which would even prevent $n^{1/2}h^r$ to go to 0 in Theorem 1. A practical bandwidth selection is proposed Section 5.

Remark 2. Assumption (A2) prevents from bias problems in the estimation of f that may occur at the borders of Q. Indeed, if f jumps at the boundary of Q, then our estimate of f would be asymptotically biased and the rates provided in Theorem 1 would not hold. To get rid of this problem, if one knew the support of f, one could correct by hand the estimator as, for instance, in [17], or might use Beta kernels as detailed in [3].

Remark 3. Assumption (A3) basically says that f is separated from 0 on Q. The exponential bound on the kernel in assumption (A4) guarantees that f is estimated uniformly on Q (see [5]). This helps to control the random denominators $\widehat{f}^{(i)}(X_i)$'s in the expression of $\widehat{I}(\varphi)$ and $\widehat{I}_c(\varphi)$. In the context of Monte–Carlo procedures for integral approximation, assumptions (A2) and (A3) are not that restrictive because one can draw the X_i's from a distribution smooth enough and whose support contains the integration domain.

Remark 4. The use of leave-one-out estimators $\widehat{f}^{(i)}$ and $\widehat{v}^{(i)}$ in $\widehat{I}_c(\varphi)$ are not only justified by the simplification they involve in the proofs. It also leads to better convergence rates. Consider the term R_0 in the proof of equation (ii) in Theorem 1, when replacing the leave-one-out estimator of f by the classical one, R_0 remains a degenerate U-statistic but with nonzero diagonal terms. It is possible to show that these terms are leading terms of the resulting expansion. They imply a rate of convergence of order $n^{-1/2}h^{-d}$ which is larger than the rate we found for $\widehat{I}_c(\varphi)$.

However, concerning $\widehat{I}(\varphi)$, the leave-one-out estimator is not necessary to get (i). The leave-one-out estimator being indeed at a distance $O(n^{-1}h^{-d})$ from the ordinary one, the change would made a difference of order at most $n^{-1/2}h^{-d}$ in the left-hand side of (i), which already appears in the right-hand side of (i).

Remark 5. The function class $\mathcal{H}(s, M)$ contains two interesting sets of functions that provide different rates of convergence in Theorem 1. First, if φ is α-Hölder on $\mathbb{R}^d$ with Hölder constant M_1, and has bounded support, then φ is $\mathcal{H}(\alpha, M_1)$ on $\mathbb{R}^d$. Second, if the support of φ is a convex body (compact convex set with non-empty interior) and φ is α-Hölder (with constant M_1) *inside its support* (e.g., the indicator of a ball) then there exists $M_2 > 0$ such that φ is $\mathcal{H}(\min(\alpha, 1/2), M_2)$ on $\mathbb{R}^d$ (see Lemma 9 in the Section 6). Then, because the sum of two Nikolski functions is still Nikolski, the assumptions of Theorem 1 are valid for a wide range of integrand. Moreover, note that a loss of smoothness at the boundary of the support involves a loss in the rates of convergence (i) and (ii). More precisely, whatever the smoothness degree of φ inside its support, if continuity fails at the boundary, then the Nikolski regularity would be at most $1/2$ and, therefore, the rates acceleration in Theorem 1 could not exceed $h^{1/2}$. In Section 3, we study such an example and show a central limit theorem with such a rate.

Remark 6. The symmetry assumption in (A4) is actually superfluous, but simplifies the proof, because in this case we do not have to distinguish the convolution with $K(x)$ and the convolution with $K(-x)$.

2.2. On the generalization of Theorem 1

In view of the intriguing convergence rates stated in Theorem 1, one may be curious to know the behavior of our estimator when estimating more general functionals with the form

$$I_T = \int T\big(x, f(x)\big)\, dx,$$

where $T : \mathbb{R}^d \times \mathbb{R}^+ \to \mathbb{R}$. Following the same approach as previously, the estimator we consider is

$$\widehat{I}_T = n^{-1} \sum_{i=1}^{n} \frac{T(X_i, \widehat{f}^{(i)}(X_i))}{\widehat{f}^{(i)}(X_i)}.$$

It turns out that T given by $(x, y) \mapsto \varphi(x)$ is the only case for which the rates are faster than root n. For other functionals and a wide range of bandwidth, $\sqrt{n}(\widehat{I}_T - I_T)$ converges to a normal distribution. In view of the negative aspect of this result with respect to the statement of Theorem 1, we provide an informal calculation of the asymptotic law of $\sqrt{n}(\widehat{I}_T - I_T)$. We require that (A2) to (A4) hold and that $nh^{2r} \to 0$ and $nh^{2d} \to +\infty$ (the latter guarantees faster than root n rates in equation (i)). If $y \mapsto T(x, y)$ has a bounded (uniformly in x) second-order derivative, using a Taylor expansion with respect to the second coordinate of T (the first-order derivative of T with respect to the second coordinate is further denoted by $\partial_2 T$), we have

$$\begin{aligned} &n^{1/2}(\widehat{I}_T - I_T) \\ &\quad = n^{-1/2} \sum_{i=1}^{n} \left(\frac{T(X_i, f(X_i))}{\widehat{f}^{(i)}(X_i)} - I_T + \frac{\partial_2 T(X_i, f(X_i))(\widehat{f}^{(i)}(X_i) - f(X_i))}{\widehat{f}^{(i)}(X_i)} \right) + \widetilde{R}_2, \end{aligned}$$

where $\widetilde{R}_2$ can be treated by standard techniques of kernel estimation (see equations (12) and (17) for details), this gives that, with probability going to 1,

$$|\widetilde{R}_2| \le Cn^{-1/2} \sum_{i=1}^{n} \frac{(\widehat{f}^{(i)}(X_i) - f(X_i))^2}{\widehat{f}^{(i)}(X_i)} = O_{\mathbb{P}}\big(n^{1/2}h^{2r} + n^{-1/2}h^{-d}\big) = o_{\mathbb{P}}(1),$$

where $C > 0$ does not depend on n or h. Then we write

$$\sqrt{n}(\widehat{I}_T - I_T) = \widetilde{R}_0 + \widetilde{R}_1 + \widetilde{R}_2,$$

with

$$\widetilde{R}_0 = n^{-1/2} \sum_{i=1}^{n} \left(\frac{T(X_i, f(X_i))}{\widehat{f}^{(i)}(X_i)} - I_T - \frac{\partial_2 T(X_i, f(X_i)) f(X_i)}{\widehat{f}^{(i)}(X_i)} + \int \partial_2 T\big(x, f(x)\big) f(x)\, dx \right),$$

$$\widetilde{R}_1 = n^{-1/2} \sum_{i=1}^{n} \left(\partial_2 T\big(X_i, f(X_i)\big) - \int \partial_2 T\big(x, f(x)\big) f(x)\, dx \right).$$

If $x \mapsto T(x, f(x))$ and $x \mapsto \partial_2 T(x, f(x)) f(x)$ are Nikolski, applying Theorem 1 gives that $\widetilde{R}_0 = o_{\mathbb{P}}(1)$. As a consequence $\sqrt{n}(\widehat{I}_T - I_T) = o_{\mathbb{P}}(1)$ if and only if the variance of $\widetilde{R}_1$ is degenerate, that is equivalent to

$$\partial_2 T\big(X_i, f(X_i)\big) = c \qquad \text{a.s.}$$

If we want this to be true for a reasonably large class of distribution functions, it would imply

$$\partial_2 T(x, y) = c \qquad \text{for all } (x, y) \in \mathbb{R}^d \times \mathbb{R}^+,$$

for which the solutions have the form $T(x, y) = \varphi(x) + cy$.

3. Central limit theorem

In the previous section, we derived upper bounds on the convergence rates in probability under fairly general conditions. In this section, by being a little more specific about the regularity of φ, we are able to describe precisely the asymptotic distribution of $\widehat{I}_c(\varphi) - I(\varphi)$. Actually the approach is to decompose the latter quantity as a sum of a U-statistic U_n plus a martingale M_n with respect to the filtration $\{X_1, \dots, X_n\}$, plus a bias term B_n that is non-random (see the beginning of Section 6.2 for the definitions of U_n, M_n, B_n). Then existing results about the asymptotic behavior of completely degenerate U-statistics [12] and martingales [13] will help to derive the asymptotic distribution. We shall consider two cases. First, we present the case where φ is smooth enough so that the dominant term is U_n, and second we study an example where φ is not continuous at the boundary of its support. As a consequence, the dominant term is M_n.

For $\widehat{I}(\varphi) - I(\varphi)$, the situation is less interesting since for most of the choice of h a (non-random) bias term leads the asymptotic decomposition (see Remark 9).

3.1. Smooth case

The smooth case corresponds to situations where the functions f and φ are smooth enough, that is, $r > 3d/2$ and $2s > d$. This is highlighted by the assumptions on the bandwidth in the next theorem.

Theorem 2. *Under the assumptions* (A1) *to* (A4), *if* $nh^{2d} \to +\infty$, $nh^{r+d/2} \to 0$ *and* $nh^{2s+d} \to 0$, *the random variable* $nh^{d/2}(\widehat{I}_c(\varphi) - I(\varphi))$ *is asymptotically normally distributed with zero-mean and variance given by*

$$\int \left(\int \big(K(u+v) - K(v)\big) K(u)\, du \right)^2 dv \int \varphi(x)^2 f(x)^{-2}\, dx.$$

The assumptions on the bandwidth are not satisfied by the optimal bandwidths displayed in Table 1. This is, in fact, a presentation issue. Indeed we have chosen to make the bias term B_n vanish so that any optimal bandwidth that balances the bias and the variance is excluded. We could have proceeded the other way around, by stating that $nh^{d/2}(\widehat{I}_c(\varphi) - I(\varphi) - B_n)$ has the same limiting distribution as in Theorem 2, provided that $nh^{2d} \to +\infty$ and $nh^{2\min(r,s)+d} \to 0$. One can verify that this holds true for the optimal bandwidth given in the first line of Table 1 for equation (ii).

3.2. A non-smooth example

We are interested in the case where φ is not sufficiently regular so that M_n is no longer negligible with respect to U_n, that is, $nh^{2\min(r,s)+d}$ does not go to 0. This occurs whenever $s < d/2$. In this case the variance is hard to compute since it depends on the behavior of M_n and therefore on the rate of convergence of the kernel regularization of φ. Hence, a precise description cannot be provided by considering usual regularity classes, for example, Hölder, Nikolski or Sobolev since they only provide bounds on the rate of kernel regularization. For this reason, we consider a particular case where the function φ is Nikolski inside Q and vanishes outside. Typical functions we have in mind are the one that jump at the boundary of their support. Lemma 9 informs us that such functions are Nikoslki with regularity $1/2$. For $Q \subset \mathbb{R}^d$ compact and $x \in \partial Q$, we define

$$L_Q(x) = \iint \min\big(\langle z, u(x)\rangle, \langle z', u(x)\rangle\big)_+ K(z) K\big(z'\big)\, dz\, dz',$$

where $u(x)$ is the unit normal outer vector of Q at the point x. We need the following assumption in place of (A1).

(B1) For some $s > 1/2$ and $M > 0$, the support of φ is a convex body $Q \subset \mathbb{R}^d$ with $\mathcal{C}^2$ boundary and φ is $\mathcal{H}(s, M)$ on Q.

Theorem 3. *Under the assumptions* (A2) *to* (A4) *and* (B1), *if* $nh^{(3d+1)/2} \to +\infty$ *and* $nh^{2r-1} \to 0$ *the random variable* $(nh^{-1})^{1/2}(\widehat{I}_c(\varphi) - I(\varphi))$ *is asymptotically normally distributed with zero-mean and variance given by*

$$\int_{\partial Q} L_Q(x) \varphi(x)^2\, d\mathcal{H}^{d-1}(x),$$

where $\mathcal{H}^{d-1}$ stands for the $(d-1)$-dimensional Hausdorff measure.

4. Application to nonparametric regression

Equation (1) has applications in nonparametric regression with random design. Let

$$Y_i = g(X_i) + \sigma(X_i)e_i, \tag{3}$$

where (e_i) is an i.i.d. sequence of real random variables with mean 0 and unit variance, independent of the sequence (X_i), and $\sigma : \mathbb{R}^d \to \mathbb{R}$ and $g : \mathbb{R}^d \to \mathbb{R}$ are unknown functions. Let $Q \subset \mathbb{R}^d$ be a compact set and $L_2(Q)$ be the Hilbert space of squared-integrable functions on Q. Let $\psi \in L_2(Q)$ be extended to $\mathbb{R}^d$ by 0 outside of Q (ψ has compact support Q). The inner product in $L_2(Q)$ between the regression function g and ψ, is given by

$$c = \int g(x)\psi(x)\,dx,$$

note that if ψ belongs to a given basis of $L_2(Q)$, then c is a coordinate of g in this basis. Among typical applications, we can mention Fourier coefficients estimation for either nonparametric estimation (see, e.g., [14], Section 3.3), or location parameter estimation (see [11]). We also mention the link with the estimation of the index in the *single index model* (see [16]).

The estimation of the linear functionals of g is a typical semiparametric problem in the sense that it requires the nonparametric estimation of the density f of X_1 as a first step and then to use it in order to estimate a real parameter. To the best of our knowledge, in the case of a regression with unknown random design, estimators that achieve root n consistency have not been provided yet (see, e.g., [14] and the reference therein). Our approach is based on kernel estimates $\widehat{f}^{(i)}$ of the density of X_1 that are then plugged into the classical empirical estimator of the quantity $\mathbb{E}[Y\psi(X)f(X)^{-1}]$. We define the estimator

$$\widehat{c} = n^{-1}\sum_{i=1}^{n} \frac{Y_i\psi(X_i)}{\widehat{f}^{(i)}(X_i)},$$

to derive the asymptotic of $\sqrt{n}(\widehat{c}-c)$, we use model (3) to get the decomposition

$$\sqrt{n}(\widehat{c}-c) = A + B,$$

with

$$A = n^{-1/2}\sum_{i=1}^{n} \frac{\sigma(X_i)\psi(X_i)}{\widehat{f}^{(i)}(X_i)}e_i,$$

$$B = n^{-1/2}\sum_{i=1}^{n}\left(\frac{g(X_i)\psi(X_i)}{\widehat{f}^{(i)}(X_i)} - \int g(x)\psi(x)\,dx\right).$$

Roughly speaking, Theorem 1 provides that B is negligible with respect to A. As a result, A carries the weak convergence of $\sqrt{n}(\widehat{c}-c)$ and, therefore, the limiting distribution can be obtained

making full use of the independence between the X_i's and the e_i's. In order to achieve such a program, this assumption is needed.

(C1) For some $s > 0$ and $M > 0$, the support of ψ is a compact set $Q \subset \mathbb{R}^d$ and both ψ and g are $\mathcal{H}(s, M)$ on $\mathbb{R}^d$.

The following theorem is proved in Section 6.

Theorem 4. *Under the assumptions* (A2) *to* (A4), *and* (C1), *if* $n^{1/2}h^r \to 0$ *and* $n^{1/2}h^d \to +\infty$, *then the random variable* $n^{1/2}(\widehat{c} - c)$ *is asymptotically normally distributed with zero-mean and variance*

$$v = \mathrm{Var}\left(\frac{\sigma(X_1)\psi(X_1)}{f(X_1)}\right).$$

Remark 7. Let us compare $\widehat{c}$ with the appealing estimator

$$\widetilde{c} = n^{-1}\sum_{i=1}^{n}\frac{Y_i\psi(X_i)}{f(X_i)}$$

which requires the knowledge of f. First, if the signal is observed without noise, that is, $Y_i = g(X_i)$, then $n^{1/2}(\widehat{c} - c)$ goes to 0 in probability whereas $\widetilde{c}$ is asymptotically normal. Secondly, when there is some noise in the observed signal, meaning that $\sigma(X_1)$ is not 0, the comparison can be made regarding their asymptotic variances. Since we have

$$v \le \mathrm{Var}\big(n^{1/2}(\widetilde{c} - c)\big),$$

it is asymptotically more efficient to plug the nonparametric estimator of f than to use f directly.

Remark 8. The set Q reflects the domain where g is studied. Obviously, the more dense the X_i's in Q, the more stable the estimation. Nevertheless, it could happen that f vanishes on some point on Q and this is not taken into account by our framework. In such situations, one may adapt the estimation from the sample by ignoring the design points on which the estimated density takes too small values. The estimator $\widehat{c}$ might be replaced by

$$n^{-1}\sum_{i=1}^{n}\frac{Y_i\psi(X_i)}{\widehat{f}^{(i)}(X_i)}1_{\{\widehat{f}^{(i)}(X_i)>b\}},$$

where $b > 0$ will certainly depend on n. This method, often referred as *trimming*, has been employed in [16] and [4] and guarantees computational stability as well as theoretical properties. Even if such an approach is feasible here, it seems far beyond the scope of the article.

5. Simulations

In this section, we provide some insights about the implementation and the practical behavior of our integral approximation procedure. In particular, we propose an adaptive procedure that

selects the bandwidth for the kernel smoothing. While our theoretical study highlighted that our estimators suffers from the *curse of dimensionality* (see Remark 1), our simulation results confirm that the estimation accuracy of our methods diminishes when the dimension increases. In dimension 1, our procedure outperforms by far the Monte–Carlo method. In moderate sample size (from 200 to 5000) up to dimension 4, our method still realizes a significant improvement over the Monte–Carlo method. The simulations are conducted under fairly general design distributions that do not necessarily satisfy assumption (A2) (e.g., equation (7)).

5.1. Kernel choice

In the whole simulation study, our estimator of the density of the design is based on the kernel

$$K(x) = \tfrac{1}{2}c_d^{-1}(d+1)\big(d+2-(d+3)|x|\big)1_{|x|<1},$$
$$c_d = \frac{2\pi^{d/2}}{d\Gamma(d/2)},$$

where c_d is the volume of the unit ball in dimension d. This kernel is radial with order 3.

5.2. Bandwidth choice

One may follow [15] to select the optimal bandwidth by a plug-in method. It requires to optimize an asymptotic equivalent of the MSE with respect to h. In Section 3, we highlighted that the limiting distribution of $\widehat{I}(\varphi) - I(\varphi)$, and so the MSE, depends heavily on the degree of smoothness of φ. In practice, the regularity of φ is often unknown, as a result, we prefer a simulation–validation type strategy.

The idea is to pick the value h which gives the best result for the estimation of the integral $I(\widetilde{\varphi})$ of a test function $\widetilde{\varphi}$ which looks like φ, and for which $I(\widetilde{\varphi})$ is known. We choose this test function as

$$\widetilde{\varphi}(x) = n^{-1}\sum_{i=1}^{n}\frac{\varphi(X_i)}{\widehat{f}^{(i)}(X_i)}h_0^{-d}\widetilde{K}\left(\frac{x-X_i}{h_0}\right), \tag{4}$$

where $\widetilde{K}$ is simply the Epanechnikov kernel

$$\widetilde{K}(x) = \tfrac{1}{2}c_d^{-1}(d+2)\big(1-|x|^2\big)1_{|x|<1}. \tag{5}$$

Since we know that

$$I(\widetilde{\varphi}) = \int \widetilde{\varphi}(x)\,dx = n^{-1}\sum_{i=1}^{n}\frac{\varphi(X_i)}{\widehat{f}^{(i)}(X_i)},$$

we just take the value of h for which the estimate $\widehat{I}(\widetilde{\varphi})$ is closest to $I(\widetilde{\varphi})$; there is actually two values, one for $\widehat{I}(\widetilde{\varphi})$ and one for $\widehat{I}_c(\widetilde{\varphi})$. The smoothing parameter h_0 is chosen using the rule of

thumb given by

$$h_0 = \sigma \left(\frac{d 2^{d+5} \Gamma(d/2+3)}{(2d+1)n} \right)^{1/(4+d)}, \tag{6}$$

where σ^2 is the mean of the estimated variances of each component (see [20], Section 4.3.2). The density estimates $\widehat{f}^{(i)}(X_i)$ in (4) are computed with the same value h_0 and the same kernel.

We did not try to use a resampling method, thinking that it is better to have h adapted to the specific sample.

5.3. First model

In this model, f is a normal distribution

$$X_i \sim \mathcal{N}\left(\frac{1}{2}, \frac{1}{4} Id\right),$$

$$\varphi(x) = \prod_{k=1}^{d} 2 \sin(\pi x_k)^2 1_{0 \le x_k \le 1}.$$

The integral of φ is 1. Figure 1 shows simulations for different values of n and d, and using equations (4), (5) and (6) for the choice of h.

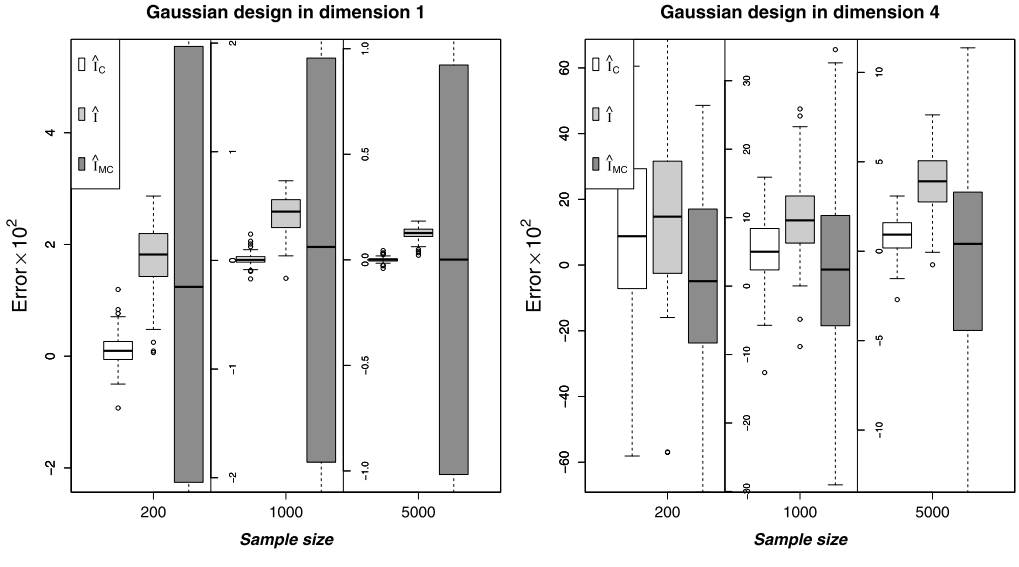

Figure 1. Each boxplot is based on 100 estimates $\widehat{I}_c(\varphi)$, $\widehat{I}(\varphi)$ and Monte–Carlo method noted $\widehat{I}_{\mathrm{MC}}$ for the first model with different values of n and d.

5.4. Second model

In this second model, the assumptions are not satisfied since the distribution is uniform over the unit cube, we have

$$X_i \sim \mathcal{U}\big([0,1]^d\big), \tag{7}$$

$$\varphi(x) = \prod_{k=1}^{d} 2\sin(\pi x_k)^2 1_{0\le x_k\le 1}. \tag{8}$$

In spite of the fact that (A2) is not any more satisfied, good results are still possible because φ cancels at the boundary of the cube. For the choice of h, we used equation (4), (5) but, it is important to constrain the function $\widetilde{\varphi}$ to have its support on the cube, and a way to do this is to remove the boundary terms out of (4) by choosing now

$$\begin{gathered}\widetilde{\varphi}(x) = |J|^{-1} \sum_{i\in J} \frac{\varphi(X_i)}{\widehat{f}^{(i)}(X_i)} h_0^{-d} \widetilde{K}\left(\frac{x - X_i}{h_0}\right),\\ J = \{i : h < X_{ij} < 1 - h,\, j = 1\cdots d\}.\end{gathered} \tag{9}$$

We could have done the other way around, use (4) and simulate uniformly extra points at distance less than h_0 of the cube, in order to cover the support of $\widetilde{\varphi}$. Figure 2 shows the results of the simulations for different values of n and d and using equations (9), (5) and (6) for the choice of h.

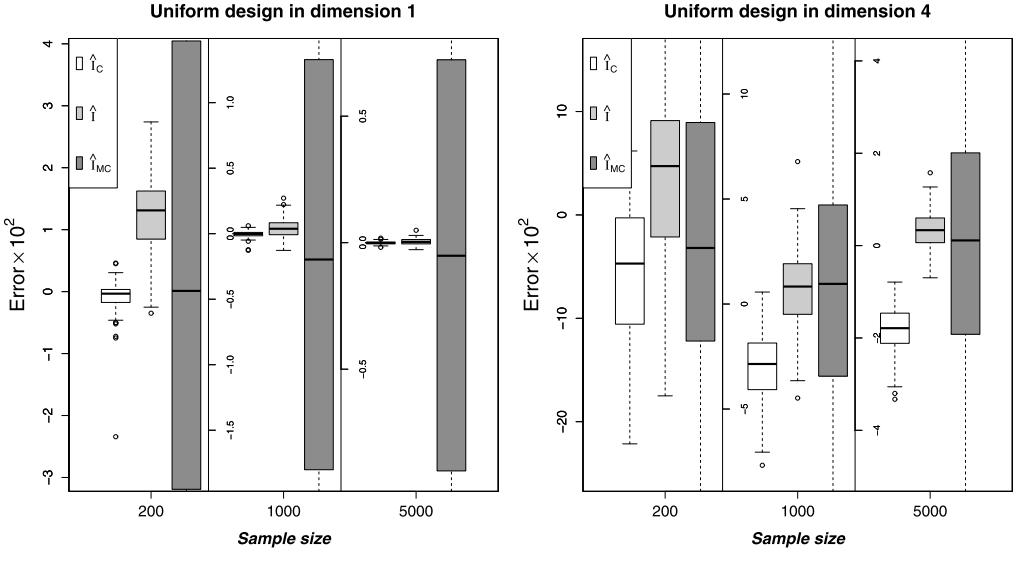

Figure 2. Each boxplot is based on 100 estimates $\widehat{I}_c(\varphi)$, $\widehat{I}(\varphi)$ and Monte–Carlo method noted $\widehat{I}_{\mathrm{MC}}$ for the second model with different values of n and d.

6. Proofs

Notation

The Euclidean norm, the L_p norm and the supremum norm are, respectively, denoted by $|\cdot|$, $\|\cdot\|_p$ and $\|\cdot\|_\infty$. We introduce $K_h(\cdot) = h^{-1}K(\cdot/h)$, and

$$K_{ij} = h^{-d}K\big(h^{-1}(X_i - X_j)\big),$$
$$\widehat{f_i} = \frac{1}{n-1}\sum_{1\le j\le n, j\neq i} K_{ij},$$
$$\widehat{v_i} = \frac{1}{(n-1)(n-2)}\sum_{1\le j\le n, j\neq i} (K_{ij} - \widehat{f_i})^2,$$

and for any function $g : \mathbb{R}^d \to \mathbb{R}$, we define

$$g_h(x) = \int g(x - hu)K(u)\,du,$$

and we put

$$\psi_q(x) = \frac{\varphi(x)}{f_h(x)^q}, \qquad q \in \mathbb{N},$$
$$\widetilde{\psi}(x) = \left(\varphi(x)\frac{f(x)}{f_h(x)^2}\right)_h.$$

6.1. Proof of Theorem 1

We start by showing (ii), then (i) will follow straightforwardly.

Proof of (ii). The following development reminiscent of the Taylor expansion

$$\frac{1}{\widehat{f_i}} = \frac{1}{f_h(X_i)} + \frac{f_h(X_i) - \widehat{f_i}}{f_h(X_i)^2} + \frac{(f_h(X_i) - \widehat{f_i})^2}{f_h(X_i)^3} + \frac{(f_h(X_i) - \widehat{f_i})^3}{\widehat{f_i} f_h(X_i)^3},$$

allows us to expand our estimator as a sum of many terms, where the density estimate $\widehat{f_i}$ is moved to the numerator, with the exception of the fourth one. We will show that this last term goes quickly to 0. For the linearised terms, this is very messy because the correct bound will be obtained by expanding also $\widehat{f_i}$ in those expressions. In order to sort out these terms, we borrow from Vial [24] the trick of making appear a degenerate U-statistic in such a development (by inserting the right quantity in R_0 below). More explicitly, recalling that

$$\widehat{I_c}(\varphi) - I(\varphi) = n^{-1}\sum_{i=1}^{n} \frac{\varphi(X_i)}{\widehat{f_i}}\left(1 - \frac{\widehat{v_i}}{\widehat{f_i^2}}\right) - I(\varphi),$$

we obtain

$$\widehat{I_c}(\varphi) - I(\varphi) = R_0 + R_1 + R_2 + R_3 + R_4 + R_5, \tag{10}$$

with (we underbrace terms which have been deliberately introduced and removed)

$$R_0 = n^{-1}\sum_{i=1}^{n}\bigl(\psi_1(X_i) - \psi_2(X_i)\widehat{f_i} + \underbrace{\widetilde{\psi}(X_i)} - \underbrace{\mathbb{E}\bigl[\psi_1(X_i)\bigr]}\bigr),$$

$$R_1 = \int \bigl(\underbrace{f(x)f_h(x)^{-1}} - 1\bigr)\varphi(x)\,dx,$$

$$R_2 = n^{-1}\sum_{i=1}^{n}\bigl(\psi_1(X_i) - \underbrace{\widetilde{\psi}(X_i)}\bigr),$$

$$R_3 = n^{-1}\sum_{i=1}^{n}\psi_3(X_i)\bigl(\bigl(f_h(X_i) - \widehat{f_i}\bigr)^2 - \underbrace{\widehat{v_i}}\bigr),$$

$$R_4 = n^{-1}\sum_{i=1}^{n}\frac{\psi_3(X_i)\widehat{v_i}}{\widehat{f_i}^3}\bigl(\underbrace{\widehat{f_i}^3} - f_h(X_i)^3\bigr),$$

$$R_5 = n^{-1}\sum_{i=1}^{n}\psi_3(X_i)\frac{(f_h(X_i) - \widehat{f_i})^3}{\widehat{f_i}},$$

where $\widehat{v_i}$ appears to be a centering term in R_3. We shall now compute bounds for each term separately.

Step 1. $\|n^{1/2}R_0\|_2 = O(n^{-1/2}h^{-d/2})$. Note that

$$R_0 = n^{-1}(n-1)^{-1}\sum_{i\neq j}\bigl(\mathbb{E}\bigl[u_{ij}|X_j\bigr] - u_{ij} + E\bigl[u_{ij}|X_i\bigr] - E[u_{ij}]\bigr),$$

with $u_{ij} = \psi_2(X_i)K_{ij}$, is a degenerate U-statistic. This is due to the fact that

$$\mathbb{E}[u_{ij}|X_i] = \psi_2(X_i)f_h(X_i) = \psi_1(X_i),$$
$$\mathbb{E}[u_{ij}|X_j] = (\psi_2 f)_h(X_j) = \widetilde{\psi}(X_j).$$

The $n(n-1)$ terms in the sum are all orthogonal with L_2 norm smaller than $\|u_{ij}\|_2$, hence

$$(n-1)^2 E\bigl[R_0^2\bigr] \leq \mathbb{E}\bigl[u_{12}^2\bigr] \leq \|\psi_2\|_\infty^2 E\bigl[K_{12}^2\bigr] \leq C_1 h^{-d},$$

because of equation (24) in Lemma 7.

Step 2. $n^{1/2}R_1 = O(n^{1/2}h^r)$. This is a consequence of equation (18) of Lemma 6, and from assumption (A3).

Step 3. $\|n^{1/2}R_2\|_2 = O(n^{1/2}h^r + h^s)$. We can rearrange the function $\psi_1(x) - \widetilde{\psi}(x)$ as

$$\psi_1(x) - \widetilde{\psi}(x) = \big(\psi_1(x) - \psi_{1h}(x)\big) + \big(\psi_{1h}(x) - \widetilde{\psi}(x)\big),$$

with

$$\begin{aligned} \big\|\psi_{1h}(x) - \widetilde{\psi}(x)\big\|_\infty &= \left\| \left(\psi_1(x) - \varphi(x)\frac{f(x)}{f_h(x)^2}\right)_h \right\|_\infty \\ &\le \left\| \psi_1(x) - \varphi(x)\frac{f(x)}{f_h(x)^2} \right\|_\infty \\ &= \left\| \frac{\varphi}{f_h^2}(f_h - f) \right\|_\infty \le Ch^r, \end{aligned} \tag{11}$$

for some constant C, where the last inequality follows from equation (18) in Lemma 6. Then we have

$$R_2 \le \left| n^{-1}\sum_{i=1}^n \big(\psi_{1h}(X_i) - \psi_1(X_i)\big) \right| + Ch^r,$$

and by spliting the mean and the variance of the first term we get

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n \big(\psi_1(X_i) - \psi_{1h}(X_i)\big)\right)^2\right] = \mathbb{E}\big[\psi_1(X_1) - \psi_{1h}(X_1)\big]^2 + \frac{1}{n}\operatorname{Var}\big(\psi_1(X_1) - \psi_{1h}(X_1)\big),$$

and we conclude by equations (19) and (20) of Lemma 6 (it is an easy exercise to show that ψ_1 is Nikolski with regularity $\min(r, s)$).

Step 4. $\|n^{1/2}R_3\|_2 = O(n^{-1/2}h^{-d/2})$. We first express R_3 as a U-statistic. Set

$$U_i = \big(f_h(X_i) - \widehat{f_i}\big)^2 - \widehat{v_i},$$

and rewrite R_3 as

$$R_3 = n^{-1}\sum_{i=1}^n \psi_3(X_i)U_i.$$

Consider a sequence of real numbers $(x_j)_{1\le j\le p}$ and set

$$\begin{aligned} m &= \frac{1}{p}\sum_{j=1}^p x_j, \\ v &= \frac{1}{p(p-1)}\sum_{j=1}^p (x_j - m)^2 = \frac{1}{p(p-1)}\sum_{j=1}^p \big(x_j^2 - m^2\big), \end{aligned}$$

then

$$m^2 - v = \left(1 + \frac{1}{p-1}\right)m^2 - \frac{1}{p(p-1)}\sum_{j=1}^{p} x_j^2 = \frac{2}{p(p-1)}\sum_{j<k} x_j x_k.$$

Applying this with $x_j = K_{ij} - f_h(X_i)$ (i is fixed) and $p = n-1$ we get

$$U_i = \frac{2}{(n-1)(n-2)}\sum_{j\neq i, k\neq i, j<k}\bigl(K_{ij} - f_h(X_i)\bigr)\bigl(K_{ik} - f_h(X_i)\bigr) = \frac{2}{(n-1)(n-2)}\sum_{j<k}\xi_{ij}\xi_{ik},$$

with

$$\begin{aligned}\xi_{ij} &= K_{ij} - f_h(X_i),\\ \xi_{ii} &= 0.\end{aligned}$$

Then

$$R_3 = \frac{2}{n(n-1)(n-2)}\sum_i\sum_{j<k}\psi_3(X_i)\xi_{ij}\xi_{ik}.$$

We are going to calculate $\mathbb{E}[R_3^2]$ by using the Efron–Stein inequality (Theorem 8) and the moment inequalities (23) to (25) for ξ_{ij} stated in Lemma 7; in particular, by (23), $\mathbb{E}[R_3^2] = \operatorname{Var}(R_3)$. Consider $R_3 = f(X_1, \ldots, X_n)$ as a function of the X_i's and define

$$\begin{aligned}R_3' &= f\bigl(X_1', X_2, \ldots, X_n\bigr),\\ \xi_{1i}' &= h^{-d}K\bigl(h^{-1}\bigl(X_1' - X_i\bigr)\bigr) - f_h\bigl(X_1'\bigr),\\ \xi_{i1}' &= h^{-d}K\bigl(h^{-1}\bigl(X_i - X_1'\bigr)\bigr) - f_h(X_i),\\ \xi_{ij}' &= \xi_{ij} \qquad \text{if } i\neq 1 \text{ and } j\neq 1,\end{aligned}$$

where X_1' is a copy of X_1 independent from the sample $(X_1, \ldots, X_n)$. Then by the Efron–Stein inequality (remember that $\xi_{ii} = 0$)

$$\|R_3\|_2 \leq \left(\frac{n}{2}\right)^{1/2}\bigl\|R_3 - R_3'\bigr\|_2,$$

which is of order

$$\begin{aligned}&n^{-5/2}\left\|\sum_{j<k}\bigl(\psi_3(X_1)\xi_{1j}\xi_{1k} - \psi_3\bigl(X_1'\bigr)\xi_{1j}'\xi_{1k}'\bigr) + \sum_i\sum_{1<k}\psi_3(X_i)\bigl(\xi_{i1} - \xi_{i1}'\bigr)\xi_{ik}\right\|_2\\ &\quad\leq n^{-5/2}\left(\left\|\sum_{j<k}\psi_3(X_1)\xi_{1j}\xi_{1k} - \psi_3\bigl(X_1'\bigr)\xi_{1j}'\xi_{1k}'\right\|_2 + \left\|\sum_{1<k}\sum_i\psi_3(X_i)\bigl(\xi_{i1} - \xi_{i1}'\bigr)\xi_{ik}\right\|_2\right)\\ &\quad= n^{-5/2}\bigl(\|T_1\|_2 + \|T_2\|_2\bigr).\end{aligned}$$

Noting that the terms in the first sum are orthogonal (by independence of ξ_{ij} and ξ_{ik} conditionally to X_i and (23)) we obtain

$$\begin{aligned}\|T_1\|_2 &= \frac{(n-1)^{1/2}(n-2)^{1/2}}{2^{1/2}}\big\|\psi_3(X_1)\xi_{12}\xi_{13} - \psi_3\big(X_1'\big)\xi_{12}'\xi_{13}'\big\|_2 \\ &\le 2^{1/2}n\|\psi_3\|_\infty\|\xi_{12}\xi_{13}\|_2 \\ &= 2^{1/2}n\|\psi_3\|_\infty\mathbb{E}\big[\mathbb{E}\big[\xi_{12}^2\xi_{13}^2|X_1\big]\big]^{1/2} \\ &= 2^{1/2}n\|\psi_3\|_\infty\big\|\mathbb{E}\big[\xi_{12}^2|X_1\big]\big\|_2 \\ &= O\big(nh^{-d}\big)\end{aligned}$$

by equation (24). Because the terms of the second sum are orthogonal whenever the values of k are different, we get

$$\|T_2\|_2 = (n-1)^{1/2}\Big\|\sum_i \psi_3(X_i)\big(\xi_{i1} - \xi_{i1}'\big)\xi_{i2}\Big\|_2.$$

By first developing and then using that X_1' is an independent copy of X_1, we obtain

$$\begin{aligned}\Big\|\sum_i \psi_3(X_i)\big(\xi_{i1} - \xi_{i1}'\big)\xi_{i2}\Big\|_2^2 &\le n\mathbb{E}\big[\psi_3(X_3)^2\big(\xi_{31} - \xi_{31}'\big)^2\xi_{32}^2\big] \\ &\quad + n^2\big|\mathbb{E}\big[\psi_3(X_3)\psi_3(X_4)\big(\xi_{31} - \xi_{31}'\big)\xi_{32}\big(\xi_{41} - \xi_{41}'\big)\xi_{42}\big]\big| \\ &\le \|\psi_3\|_\infty\big\{n\mathbb{E}\big[\big(\xi_{31} - \xi_{31}'\big)^2\xi_{32}^2\big] \\ &\quad + n^2\mathbb{E}\big[\big|\mathbb{E}\big[\big(\xi_{31} - \xi_{31}'\big)\xi_{32}\big(\xi_{41} - \xi_{41}'\big)\xi_{42}|X_3, X_4\big]\big|\big]\big\} \\ &= \|\psi_3\|_\infty\big\{2n\mathbb{E}\big[\xi_{31}^2\xi_{32}^2\big] + 2n^2\mathbb{E}\big[\big|\mathbb{E}[\xi_{31}\xi_{32}\xi_{41}\xi_{42}|X_3, X_4]\big|\big]\big\}.\end{aligned}$$

Then by equation (24), we have $\mathbb{E}[\xi_{31}^2\xi_{32}^2] = \mathbb{E}[\mathbb{E}[\xi_{31}^2|X_3]^2] \le C_1h^{-2d}$ and by equation (25), we get

$$\begin{aligned}\mathbb{E}\big[\big|\mathbb{E}[\xi_{31}\xi_{32}\xi_{41}\xi_{42}|X_3, X_4]\big|\big] &= \mathbb{E}\big[\mathbb{E}[\xi_{31}\xi_{41}|X_3, X_4]^2\big] \\ &\le 2\|f\|_\infty^2h^{-2d}\mathbb{E}\big[\widetilde{K}\big(h^{-1}(X_4 - X_3)\big)^2\big] + 2\|f\|_\infty^4 \\ &\le 2\|f\|_\infty^3h^{-d}\int \widetilde{K}(u)^2\,du + 2\|f\|_\infty^4,\end{aligned}$$

where $\widetilde{K}$ is defined in Lemma 7. Bringing everything together and because $nh^d \to \infty$, it holds that

$$\big\|n^{1/2}R_3\big\|_2 \le O\big(n^{-1}h^{-d} + n^{-1/2}h^{-d/2}\big) = O\big(n^{-1/2}h^{-d/2}\big).$$

Step 5. $n^{1/2}R_4 = O_{\mathbb{P}}(n^{-1}h^{-3d/2})$. We start with a lower bound for $\widehat{f_i}$ by proving the existence of $N(\omega)$ such that

$$\forall n \geq N(\omega), \forall i, \qquad \frac{b}{2} < \widehat{f_i} < 2\|f\|_\infty. \tag{12}$$

Notice that

$$\widehat{f_i} = \frac{n}{n-1}\left(\widehat{f}(X_i) - \frac{1}{nh^d}K(0)\right),$$

$$\widehat{f}(x) = \frac{1}{nh^d}\sum_{j=1}^{n} K\big(h^{-d}(x - X_j)\big),$$

due to the almost sure uniform convergence of $\widehat{f}$ to f (Theorem 1 in [5]) we have with probability 1 for n large enough

$$\frac{2b}{3} < \inf_{x\in Q}\widehat{f}(x) \leq \sup_{x\in Q}\widehat{f}(x) < \frac{3}{2}\|f\|_\infty,$$

and since assumption $nh^d \to \infty$, (12) follows. We can now compute the expectation of R_4 restricted to $\{n \geq N(\omega)\}$. Because $(a^3 - b^3) = (a-b)(a^2 + ab + b^2)$ for any real number a and b, and by the latter inequality, there exists a constant $C > 0$ which does not depend on n or h, such that

$$|R_4|1_{n>N(\omega)} \leq Cn^{-1}\sum_{i=1}^{n}\big|\widehat{f_i} - f_h(X_i)\big|\widehat{v}_i,$$

we have by the Cauchy–Schwarz inequality

$$\mathbb{E}\big[|R_4|1_{n>N(\omega)}\big] \leq C\mathbb{E}\big[\big(\widehat{f_1} - f_h(X_1)\big)^2\big]^{1/2}\mathbb{E}\big[\widehat{v}_1^2\big]^{1/2}. \tag{13}$$

Applying the fact that for any real number a, $\frac{1}{p}\sum_{j=1}^{p}(x_j - \overline{x})^2 \leq \frac{1}{p}\sum_{i=1}^{p}(x_j - a)^2$ to $x_j = K_{1j}$, $p = n-1$ and $a = f_h(X_1)$, we obtain that

$$\widehat{v}_1 \leq \frac{1}{(n-1)(n-2)}\sum_{j=2}^{n}\xi_{1j}^2,$$

then using (24)

$$\begin{aligned}\mathbb{E}\big[\widehat{v}_1^2\big] &\leq (n-1)^{-1}(n-2)^{-2}\mathbb{E}\big[\xi_{12}^4\big] + (n-1)^{-1}(n-2)^{-1}\mathbb{E}\big[\xi_{12}^2\xi_{13}^2\big] \\ &\leq C_1 n^{-3}h^{-3d} + C_1 n^{-2}h^{-2d} \\ &\leq O\big(n^{-2}h^{-2d}\big),\end{aligned} \tag{14}$$

because nh^d goes to infinity. On the other hand using equation (24) again,

$$\mathbb{E}\big[\big(\widehat{f}_1 - f_h(X_1)\big)^2\big] = \frac{1}{n-1}\mathbb{E}\big[\xi_{1i}^2\big] = O\big(n^{-1}h^{-d}\big). \tag{15}$$

Putting together (13), (14) and (15),

$$\mathbb{E}\big[|R_4|1_{n>N(\omega)}\big] = O\big(n^{-1}h^{-d}n^{-1/2}h^{-d/2}\big) = O\big(n^{-3/2}h^{-3d/2}\big).$$

In particular by Markov's inequality

$$\begin{aligned}\mathbb{P}\big(n^{3/2}h^{3d/2}|R_4| > A\big) &\le \mathbb{P}\big(n^{3/2}h^{3d/2}|R_4|1_{n>N(\omega)} > A\big) + \mathbb{P}\big(n \le N(\omega)\big)\\ &= A^{-1}O(1) + \mathbb{P}\big(n \le N(\omega)\big).\end{aligned}$$

This proves the boundedness in probability of $n^{3/2}h^{3d/2}|R_4|$.

Step 6. $n^{1/2}R_5 = O_{\mathbb{P}}(n^{-1}h^{-3d/2} + n^{-3/2}h^{-2d})$. Following (12) since

$$|R_5|1_{n>N(\omega)} \le 2b^{-3}\|\varphi\|_\infty n^{-1}\sum_{i=1}^n \big|\widehat{f}_i - f_h(X_i)\big|^3,$$

we can show the convergence in probability of the right-hand side term as in Step 5. We have indeed by the Rosenthal's inequality[1]

$$\begin{aligned}\mathbb{E}\left[n^{-1}\sum_{i=1}^n\big|\widehat{f}_i - f_h(X_i)\big|^p\right] &= (n-1)^{-p}\mathbb{E}\left[\left|\sum_{i=2}^n \xi_{1i}\right|^p\right]\\ &\le C_2 n^{-p}\big\{\big(n\mathbb{E}\big[\xi_{12}^2\big]\big)^{p/2} + n\mathbb{E}\big[|\xi_{12}|^p\big]\big\}\\ &\le C_1C_2\big\{n^{-p/2}h^{-pd/2} + n^{1-p}h^{-(p-1)d}\big\},\end{aligned} \tag{16}$$

where the latter inequality is due to equation (24). Hence, with $p = 3$

$$\mathbb{E}\big[|R_5|1_{n>N(\omega)}\big] \le C_1C_2\big\{n^{-3/2}h^{-3d/2} + n^{-2}h^{-2d}\big\}$$

and we conclude as in Step 5.

Putting together the steps 1 to 6, and taking into account, concerning R_5, that $n^{-3/2}h^{-2d} = (n^{-1/2}h^{-d/2})(n^{-1}h^{-3d/2})$, we obtain (ii).

Proof of (i). For (i), we use a shorter expansion which leads to an actually much simpler proof:

$$\frac{1}{\widehat{f}_i} = \frac{1}{f_h(X_i)} + \frac{f_h(X_i) - \widehat{f}_i}{f_h(X_i)^2} + \frac{(f_h(X_i) - \widehat{f}_i)^2}{\widehat{f}_i f_h(X_i)^2},$$

[1]For a martingale $(S_i, \mathcal{F}_i)_{i\in\mathbb{N}}$ and $2 \le p < +\infty$, we have $\mathbb{E}[|S_n|^p] \le C_2\{\mathbb{E}[(\sum_{i=1}^n \mathbb{E}[X_i^2|\mathcal{F}_{i-1}])^{p/2}] + \sum_{i=1}^n \mathbb{E}|X_i|^p\}$, where $X_i = S_i - S_{i-1}$ (see, e.g., [13], pp. 23–24).

and

$$\widehat{I}(\varphi) - I(\varphi) = R_0 + R_1 + R_2 + R_5',$$

with

$$R_5' = n^{-1} \sum_{i=1}^{n} \psi_2(X_i) \frac{(f_h(X_i) - \widehat{f}_i)^2}{\widehat{f}_i}.$$

The terms R_0, R_1 and R_2 have already been treated in the steps 1, 2 and 3 of the proof of (i). The term R_5' is bounded exactly as R_5 but now we use (16) with $p = 2$ instead of $p = 3$, to obtain

$$\mathbb{E}\big[\big|R_5'\big|1_{n>N(\omega)}\big) \le C_1 C_2 n^{-1} h^{-d}$$

and we get $n^{1/2}|R_5'| = O_{\mathbb{P}}(n^{-1/2}h^{-d})$. □

6.2. Proofs of Theorems 2 and 3

Let us define

$$M_n = n^{-1} \sum_{i=1}^{n} \psi_1(X_i) - \widetilde{\psi}(X_i) - \mathbb{E}\big[\psi_1(X_1) - \widetilde{\psi}(X_1)\big],$$

$$U_n = n^{-1}(n-1)^{-1} \sum_{i \neq j} c_{ij},$$

$$B_n = \mathbb{E}\big[\psi_1(X_1) - \widetilde{\psi}(X_1)\big] + \int \big(f(x) f_h(x)^{-1} - 1\big)\varphi(x)\, dx$$

with $c_{jk} = a_{jk} - b_{jk}$, and for $j \neq k$,

$$a_{jk} = \mathbb{E}\big[\psi_3(X_1)\xi_{1j}\xi_{1k}|X_j X_k\big],$$
$$b_{jk} = u_{jk} - \mathbb{E}\big[u_{jk}|X_j\big] - \mathbb{E}\big[u_{jk}|X_k\big] + \mathbb{E}[u_{jk}],$$

where u_{jk} has been defined at the beginning of step 3. Both proofs of Theorems 2 and 3 rely on the following lemma which turns Theorem 1 in a suitable way for weak convergence issues.

Lemma 5. *Under the assumptions of Theorem 1, we have*

$$\widehat{I}_c(\varphi) - I(\varphi) = B_n + U_n + M_n + O_{\mathbb{P}}\big(n^{-3/2}h^{-3d/2}\big).$$

Moreover, we have $B_n = O_{\mathbb{P}}(h^r)$, $U_n = O_{\mathbb{P}}(n^{-1}h^{-d/2})$ *and* $M_n = O_{\mathbb{P}}(n^{-1/2}(h^s + h^r))$.

Proof. By using the decomposition (10) and since $B_n + M_n = R_1 + R_2$, we have

$$\begin{aligned}\widehat{I}_c(\varphi) - I(\varphi) &= R_0 + R_1 + R_2 + R_3 + R_4 + R_5 \\ &= B_n + M_n + U_n + (R_0 + R_3 - U_n) + R_4 + R_5.\end{aligned}$$

We have already shown that $R_4 + R_5 = O_{\mathbb{P}}(n^{-3/2}h^{-3d/2} + n^{-2}h^{-2d})$ (this is exactly steps 5 and 6 of the proof of Theorem 1). By definition of U_n, we have

$$R_0 + R_3 - U_n = n^{-1}(n-1)^{-1}(n-2)^{-1}\sum_i \sum_{j\neq k}\big(\psi_3(X_i)\xi_{ij}\xi_{ik} - a_{jk}\big)$$

which is a completely degenerate U-statistic (R_3 is near to be completely degenerate and $a_{jk} = \mathbb{E}[\psi_3(X_1)\xi_{1j}\xi_{1k}|X_jX_k]$ appears as the good centering term). The order 2 moments of this quantity are of order $n^{-3}\mathbb{E}[\psi_3(X_1)^2\xi_{12}^2\xi_{13}^2] \propto n^{-3}h^{-2d}$. Hence, we have shown that $R_0 + R_3 - U_n = O_{\mathbb{P}}(n^{-3/2}h^{-d})$, which completes the first part of the proof. To obtain the bounds in probability, for U_n we just use step 1 and 4 of the proof of Theorem 1, for M_n we compute the L_2 norm as follows. We have

$$\begin{aligned}\|M_n\|_2 &= n^{-1/2}\big\|\psi_1(X_1) - \widetilde{\psi}(X_1)\big\|_2 \\ &\leq n^{-1/2}\big(\big\|\psi_1(X_1) - \psi_{1h}(X_1)\big\|_2 + \big\|\psi_{1h}(X_1) - \widetilde{\psi}(X_1)\big\|_2\big) \\ &\leq Cn^{-1/2}\big(h^s + h^r\big),\end{aligned}$$

for some constant C, where the last inequality is obtained using equation (11) for the term in the right and equation (20) in Lemma 6 for the term in the right. □

Remark 9. Under the assumption of Theorem 1, one may show that

$$\widehat{I}(\varphi) - I(\varphi) = \widehat{I}_c(\varphi) - I(\varphi) + n^{-1}(n-1)^{-2}\sum_{i,j}\psi_3(X_i)\xi_{ij}^2 + O_{\mathbb{P}}\big(\big(nh^d\big)^{-3/2}\big),$$

where the $O_{\mathbb{P}}$ comes from R_4 and the other remainder term corresponds to the diagonal term of the U-statistic R_3. This term equals $(n-1)^{-1}\mathbb{E}[\psi_2(X_1)(K_{12} - f_h(X_1))^2] = O(n^{-1}h^{-d})$ plus $o_{\mathbb{P}}(n^{-1}h^{-d/2})$, as a consequence, when h is such that $nh^{2(s+d)} \to 0$ and $nh^{r+d} \to 0$, the leading term of the decomposition is a constant.

6.2.1. *Proof of Theorem 2*

By Lemma 5 and the assumptions on h we have

$$\begin{aligned}nh^{d/2}(B_n + M_n + R_4 + R_5) &= O_{\mathbb{P}}\big(n^{3/2}h^{r+d/2} + n^{1/2}\big(h^{r+d/2} + h^{s+d/2}\big) + n^{-1/2}h^{-3d/2}\big) \\ &= o_{\mathbb{P}}(1).\end{aligned}$$

To derive the limiting distribution of $nh^{d/2}U_n$, we apply Theorem 1 in [12], quoted below (Theorem 11), with $H_n(X_j, X_k) = (n-1)^{-1}h^{d/2}(c_{jk} + c_{kj})$ where $c_{jk} = a_{jk} - b_{jk}$, has been defined at the beginning of Section 6.2. The asymptotic variance v_1 is the limit of the quantity

$\frac{n^2}{2}\mathbb{E}[H_n(X_1, X_2)^2]$ asymptotically equivalent to

$$h^d\big(\mathbb{E}\big[c_{12}^2\big] + \mathbb{E}[c_{12}c_{21}]\big).$$

To compute this easily, we introduce the function $\xi_i(x) = K_h(x - X_i) - f_h(x)$. First, use some algebra to obtain the formula $b_{12} = \psi_2(X_1)\xi_2(X_1) - \int \psi_2(x)\xi_2(x) f(x)\,dx$, then it follows that

$$\begin{aligned}
c_{12} &= a_{12} - b_{12} \\
&= \int \psi_3(x)\xi_1(x)\xi_2(x) f(x)\,dx - \psi_2(X_1)\xi_2(X_1) + \int \psi_2(x)\xi_2(x) f(x)\,dx \\
&= \int \big(\psi_3(x) f(x)\xi_2(x) - \psi_2(X_1)\xi_2(X_1)\big)K_h(x - X_1)\,dx \\
&\quad + \int \psi_3(x) f(x)\xi_2(x)\big(f_h(x) - f(x)\big)\,dx \\
&= \int \big(\psi_2(x)\xi_2(x) - \psi_2(X_1)\xi_2(X_1)\big)K_h(x - X_1)\,dx \\
&\quad + \int \psi_3(x)\xi_2(x)\big(f(x) - f_h(x)\big)K_h(x - X_1)\,dx + \int \psi_3(x) f(x)\xi_2(x)\big(f_h(x) - f(x)\big)\,dx \\
&= \int \big(\psi_2(x)K_h(x - X_2) - \psi_2(X_1)K_h(X_1 - X_2)\big)K_h(x - X_1)\,dx \\
&\quad + \int \big(-\big(\psi_2(x) f_h(x) - \psi_2(X_1) f_h(X_1)\big) + \psi_3(x)\xi_2(x)\big(f(x) - f_h(x)\big)\big)K_h(x - X_1)\,dx \\
&\quad + \int \psi_3(x) f(x)\xi_2(x)\big(f_h(x) - f(x)\big)\,dx.
\end{aligned}$$

Because K_h integrates to 1, it is not hard to see that the last two terms in the previous equation will be negligible in the computation of v_1. As a consequence, $h^d\mathbb{E}[c_{12}^2]$ has the same limit as

$$\begin{aligned}
&h^d \iint \left(\int \big(\psi_2(x)K_h(x - z) - \psi_2(y)K_h(y - z)\big)K_h(x - y)\,dx\right)^2 f(y) f(z)\,dy\,dz \\
&\quad = \iint \left(\int \big(\psi_2(y + hu)K(u + v) - \psi_2(y)K(v)\big)K(u)\,du\right)^2 f(y) f(y - hv)\,dy\,dv \\
&\quad = V_K \int \psi_2(y)^2 f(y)^2\,dy + o(1)
\end{aligned}$$

with $V_K = \int(\int(K(u + v) - K(v))K(u)\,du)^2\,dv$ and where the first equality follows from a change of variables and the last representation follows from the Lebesgue dominated theorem. Following the same steps as previously, we obtain an similar expression for $h^d\mathbb{E}[c_{12}c_{21}]$ and then

we get

$$v_1 = 2V_K \int \varphi(y)^2 f(y)^{-2}\, dy.$$

It remains to check the conditions of Theorem 11. Clearly, the computation of v_1 provides that $\mathbb{E}[H_n(X_1, X_2)^2] \approx n^{-2}$. We obtain similarly that $\mathbb{E}[H_n(X_1, X_2)^4] = O(n^{-5}h^{-d})$ and $\mathbb{E}[G_n(X_1, X_2)^2] = O(n^{-5}h^d)$ which implies the conditions of the theorem. □

6.2.2. *Proof of Theorem 3*

By (B1) and Lemma 9, there exists $M_2 > 0$ such that φ is $\mathcal{H}(\min(s, 1/2), M_2)$. Then we can apply Lemma 5 and by assumption on h, we obtain that

$$\begin{aligned}\big(nh^{-1}\big)^{1/2}(B_n + U_n + R_4 + R_5) &= O_{\mathbb{P}}\big(n^{1/2}h^{r-1/2} + n^{-1/2}h^{-(d+1)/2} + n^{-1}h^{-(3d+1)/2}\big)\\ &= o_{\mathbb{P}}(1).\end{aligned}$$

Since M_n is a sum of independent variables with zero-mean, we can apply the central limit theorem by checking the Lindeberg condition (see, e.g., [13], Chapter 3). Now we only have to compute the asymptotic variance v_2 defined as the limit of

$$\operatorname{Var}\big(h^{-1/2}\big(\psi_1(X_1) - \widetilde{\psi}(X_1)\big)\big) = h^{-1}\mathbb{E}\big[\big(\psi_1(X_1) - \widetilde{\psi}(X_1)\big)^2\big] - h^{-1}\mathbb{E}\big[\psi_1(X_1) - \widetilde{\psi}(X_1)\big]^2.$$

On the one hand, by equations (11) and (19), we have for some constant C

$$\begin{aligned}\big\|h^{-1/2}\big(\psi_{1h}(X_1) - \widetilde{\psi}(X_1)\big)\big\|_2 &\le Ch^{r-1/2},\\ h^{-1/2}\big|\mathbb{E}\big[\psi_1(X_1) - \psi_{1h}(X_1)\big]\big| &\le Ch^{r-1/2},\end{aligned}$$

as a consequence, we get

$$\operatorname{Var}\big(h^{-1/2}\big(\psi_1(X_1) - \widetilde{\psi}(X_1)\big)\big) = h^{-1}\big\|\psi_1(X_1) - \psi_{1h}(X_1)\big\|_2^2 + o(1).$$

On the other hand, for every $x \in Q$, we have

$$\begin{aligned}\psi_1(x) - \psi_{1h}(x) &= \int_{Q^c}\big(\psi_1(x) - \psi_1(y)\big)K_h(x-y)\,dy + \int_Q \big(\psi_1(x) - \psi_1(y)\big)K_h(x-y)\,dy\\ &= \psi_1(x)\int_{Q^c} K_h(x-y)\,dy + \int_Q \big(\psi_1(x) - \psi_1(y)\big)K_h(x-y)\,dy,\end{aligned}$$

where Q^c stands for the complement of the set Q in $\mathbb{R}^d$. Because ψ_1 is Nikolski with regularity $\min(s, r)$ inside Q, we use equation (20) of Lemma 6 to show that the L_2-norm of the right-hand side term is of order $h^{\min(s,r)}$. Clearly, since $\min(s, r) > 1/2$ we have

$$\operatorname{Var}\big(h^{-1/2}\big(\psi_1(X_1) - \widetilde{\psi}(X_1)\big)\big) = h^{-1}\left\|\psi_1(X_1)\int_{Q^c} K_h(X_1 - y)\,dy\right\|_2 + o(1)$$

and it remains to apply Lemma 10 to derive the stated limit.

6.3. Proof of the Theorem 4

By equation (3), we are interested in the asymptotic law of

$$n^{-1/2}\sum_{i=1}^{n}\frac{\sigma(X_i)\psi(X_i)}{\widehat{f}_i}e_i + n^{-1/2}\left(\sum_{i=1}^{n}\frac{g(X_i)\psi(X_i)}{\widehat{f}_i} - \int g(x)\psi(x)\,dx\right).$$

By Lemma 1, the right-hand side term goes to 0 in probability. For the other term, we use the decomposition $A_1 + A_2$, with

$$A_1 = n^{-1/2}\sum_{i=1}^{n}\frac{\sigma\psi(X_i)}{f(X_i)}e_i \quad \text{and} \quad A_2 = n^{-1/2}\sum_{i=1}^{n}\frac{\sigma\psi(X_i)(f(X_i)-\widehat{f}(X_i))}{\widehat{f}_i f(X_i)}e_i,$$

where $\sigma\psi(X_i) = \sigma(X_i)\psi(X_i)$. We define $\mathcal{F}$ as the σ-field generated by the set of random variables $\{X_1, X_2, \ldots\}$. We get

$$\mathbb{E}\big[A_2^2|\mathcal{F}\big] = n^{-1}\sum_{i=1}^{n}\frac{\sigma\psi(X_i)^2(f(X_i)-\widehat{f}_i)^2}{\widehat{f}_i^2 f(X_i)^2},$$

then, one has

$$\mathbb{E}\big[A_2^2|\mathcal{F}\big] \le \Big(b^2 \inf_i \widehat{f}_i^2\Big)^{-1}\|\sigma\psi\|_\infty^2 n^{-1}\sum_{i=1}^{n}\big(f(X_i)-\widehat{f}_i\big)^2.$$

For the term on the left, since $\sigma\psi$ has support Q we can use (12), that is for n large enough, it is bounded. For the right-hand side term, it follows that

$$n^{-1}\sum_{i=1}^{n}\big(f(X_i)-\widehat{f}_i\big)^2 \le 2n^{-1}\sum_{i=1}^{n}\big(f(X_i)-f_h(X_i)\big)^2 + 2n^{-1}\sum_{i=1}^{n}\big(f_h(X_i)-\widehat{f}_i\big)^2,$$

and then using equation (18) in Lemma 6 and (16) for $p=2$ we provide the bound

$$\left\|n^{-1}\sum_{i=1}^{n}\big(f(X_i)-\widehat{f}_i\big)^2\right\|_1 \le C\big(h^{2r} + n^{-1}h^{-d}\big) \tag{17}$$

for some $C > 0$. Therefore, we have shown that $\mathbb{E}[A_2^2|\mathcal{F}] \to 0$ in probability. Since for any $\varepsilon > 0$, $\mathbb{P}(|A_2| > \varepsilon|\mathcal{F}) \le \varepsilon^{-2}\mathbb{E}[A_2^2|\mathcal{F}]$, it remains to note that the sequence $\mathbb{P}(|A_2| > \varepsilon|\mathcal{F})$ is uniformly integrable to apply the Lebesgue domination theorem to get

$$\mathbb{P}(A_2 > \varepsilon) \longrightarrow 0.$$

To conclude, we apply the central limit theorem to A_1 and the statement follows.

6.4. Some lemmas

6.4.1. *Inequalities*

Lemma 6. *For any function $g : \mathbb{R}^d \to \mathbb{R}$, recall that $g_h(x) = \int g(x - hu)K(u)\,du$. Under assumptions* (A1), (A2) *and* (A4), *it holds that*

$$\|f_h - f\|_\infty \le C_K h^r \left\|f^{(r)}\right\|_\infty, \tag{18}$$

$$\left|\mathbb{E}\big[\varphi(X_1) - \varphi_h(X_1)\big]\right| \le C_K h^r \left\|f^{(r)}\right\|_\infty \int \left|\varphi(x)\right| dx, \tag{19}$$

$$\left\|\varphi_h(X_1) - \varphi(X_1)\right\|_2 \le C_K M h^s, \tag{20}$$

where C_K is a positive constant that depends K only.

Proof. We start by proving (19) and (20) assuming that (18) holds. For the mean: using Fubini's theorem, we have

$$\begin{aligned}\mathbb{E}\big[\varphi(X_1) - \varphi_h(X_1)\big] &= \int \big(\varphi(x) - \varphi_h(x)\big) f(x)\,dx \\ &= \int \varphi(x) f(x) - \varphi(x) f_h(x)\,dx,\end{aligned}$$

hence

$$\left|\mathbb{E}\big[\varphi(X_1) - \varphi_h(X_1)\big]\right| \le \left\|f(x) - f_h(x)\right\|_\infty \int \left|\varphi(x)\right| dx,$$

which by (18) gives

$$\left|\mathbb{E}\big[\varphi(X_1) - \varphi_h(X_1)\big]\right| \le C_K h^r \left\|f^{(r)}\right\|_\infty \int \left|\varphi(x)\right| dx.$$

This is (19). We turn now to (20):

$$\mathbb{E}\big[\big(\varphi_h(X_1) - \varphi(X_1)\big)^2\big] = \int \left(\int \big(\varphi(x - hu) - \varphi(x)\big) K(u)\,du\right)^2 f(x)\,dx. \tag{21}$$

We now use the Taylor formula with Lagrange remainder applied to $g(t) = \varphi(x - tu)$ with order k equal to the largest integer smaller than s:

$$\begin{aligned}\varphi(x - hu) - \varphi(x) &= \sum_{j=1}^{k-1} \frac{h^j}{j!} g^{(j)}(0) + \int_0^h g^{(k)}(t) \frac{(h-t)^{k-1}}{(n-1)!}\,dt \\ &= \sum_{j=1}^{k} \frac{h^j}{j!} g^{(j)}(0) + \int_0^h \big(g^{(k)}(t) - g^{(k)}(0)\big) \frac{(h-t)^{k-1}}{(n-1)!}\,dt.\end{aligned}$$

The first term is a polynomial in u which will vanish after insertion in (21) because K is orthogonal the first non-constant polynomial of degree $\leq r$. The second term is bounded as

$$\left|\int_0^h \bigl(g^{(k)}(t) - g^{(k)}(0)\bigr)\frac{(h-t)^{k-1}}{(k-1)!}\,dt\right| \leq |u|^k h^{k-1}\int_0^h \bigl|\varphi^{(k)}(x-tu) - \varphi^{(k)}(x)\bigr|\,dt.$$

Hence,

$$\begin{aligned}&\left|\int \bigl(\varphi(x-hu) - \varphi(x)\bigr)K(u)\,du\right| \\ &\quad\leq h^{k-1}\int_0^h \int \bigl|\varphi^{(k)}(x-tu) - \varphi^{(k)}(x)\bigr||u|^k K(u)\,du\,dt\end{aligned} \tag{22}$$

and by the generalized Minkowski inequality ([10] page 194)[2]

$$\begin{aligned}\bigl\|(\varphi_h - \varphi)(X_1)\bigr\|_2 &\leq h^{k-1}\int\left(\int \bigl|\varphi^{(k)}(x-tu) - \varphi^{(k)}(x)\bigr|^2 u^{2k}K(u)^2 1_{0\leq t\leq h} f(x)\,dx\right)^{1/2} du\,dt \\ &\leq Mh^{k-1}\int \bigl(|tu|^{2\alpha}|u|^{2k}K(u)^2\bigr)^{1/2}1_{0\leq t\leq h}\,du\,dt \\ &= M(1+\alpha)^{-1}h^{k+\alpha}\int \bigl(|u|^{2\alpha+2k}K(u)^2\bigr)^{1/2}\,du.\end{aligned}$$

This implies (20). Concerning (18), we use (22) with f and $k = r$ to get that

$$\bigl|f_h(x) - f(x)\bigr| \leq h^{r-1}\int_0^h \int \bigl|f^{(r)}(x+tu)\bigr||u|^r K(u)\,du\,dt,$$

the latter is bounded by a constant times h^r. □

The following lemma gives some bounds on the conditional moments of ξ_{12} that are useful in the proof of Theorem 1.

Lemma 7. *Let* $\xi_{ij} = K_{ij} - f_h(X_i)$, *under* (A1) *and* (A2)

$$\mathbb{E}[\xi_{12}|X_1] = 0, \tag{23}$$

$$\mathbb{E}\bigl[|\xi_{12}|^p|X_1\bigr] \leq 2^p\mathbb{E}\bigl[|K_{12}|^p|X_1\bigr] \leq C_1 h^{-(p-1)d}, \tag{24}$$

$$\bigl|\mathbb{E}[\xi_{13}\xi_{23}|X_1, X_2]\bigr| \leq \|f\|_\infty\bigl(h^{-d}\widetilde{K}\bigl(h^{-1}(X_2 - X_1)\bigr) + \|f\|_\infty\bigr), \tag{25}$$

with $\widetilde{K}(x) = \int |K(x-y)K(y)|\,dy$ *and* $C_1 > 0$.

[2]For any non-negative measurable function $g(\cdot,\cdot)$ on $\mathbb{R}^{k+d}$,

$$\left(\int\left(\int g(y,x)\,dy\right)^2 dx\right)^{1/2} \leq \int\left(\int g(y,x)^2\,dx\right)^{1/2} dy.$$

Proof. The first equation is trivial. For the second equation, the triangular inequality and the Jensen inequality provide

$$\mathbb{E}\big[|\xi_{12}|^p|X_1\big] \le 2^p\mathbb{E}\big[|K_{12}|^p|X_1\big] = 2^p h^{-(p-1)d}\int \big|K(u)\big|^p f(X_1 - hu)\,dx,$$

and the third one is derived by

$$\begin{aligned}\big|\mathbb{E}[\xi_{13}\xi_{23}|X_1, X_2]\big| &= \big|\mathbb{E}[\xi_{13}K_{23}|X_1, X_2]\big| \\ &= \left|\int \big(K_h(X_1 - x) - f_h(X_1)\big)K_h(X_2 - x)f(x)\,dx\right| \\ &= \left|\int \big(K_h(X_1 - X_2 + hu) - f_h(X_1)\big)K(u)f(X_2 - hu)\,du\right| \\ &\le \|f\|_\infty\big(h^{-d}\widetilde{K}\big(h^{-1}(X_2 - X_1)\big) + \|f\|_\infty\big).\end{aligned}$$

□

The Efron–Stein inequality helps to bound the L_2 moments of estimators. For the proof, we refer to the original paper [6] but also to [2].

Theorem 8 (Efron–Stein inequality). *Let $X_1, \dots, X_n$ be an i.i.d. sequence, X_1' be an independent copy of X_1 and f be a symmetric function of n variables, then*

$$\operatorname{Var}\big(f(X_1, \dots, X_n)\big) \le \frac{n}{2}\mathbb{E}\big[\big(f(X_1, \dots, X_n) - f\big(X_1', X_2, \dots, X_n\big)\big)^2\big].$$

6.4.2. *Measure results*

Lemma 9. *Let $s > 0$ and $M_1 > 0$, suppose that the support of φ is a convex body Q and that φ is $\mathcal{H}(s, M_1)$ on Q, then there exists $M > 0$ such that φ is $\mathcal{H}(\min(s, 1/2), M)$ on $\mathbb{R}^d$.*

Proof. We have

$$\begin{aligned}&\int \big|\varphi(x+u) - \varphi(x)\big|^2\,dx \\ &\quad= \int_{\{x\in Q, x+u\in Q\}} \big|\varphi(x+u) - \varphi(x)\big|^2\,dx + \int_{\{x\notin Q, x+u\in Q\}} \varphi(x+u)^2\,dx \\ &\qquad+ \int_{\{x\in Q, x+u\notin Q\}} \varphi(x)^2\,dx \\ &\quad= \int_{\{x\in Q, x+u\in Q\}} \big|\varphi(x+u) - \varphi(x)\big|^2\,dx + \int_Q \varphi(x)^2(1_{\{x-u\notin Q\}} + 1_{\{x+u\notin Q\}})\,dx \\ &\quad\le \int_{\{x\in Q, x+u\in Q\}} \big|\varphi(x+u) - \varphi(x)\big|^2\,dx + \|\varphi\|_\infty^2 \int 1_{\{\operatorname{dist}(x,\partial Q)\le|u|\}}\,dx \\ &\quad\le M_1|u|^{2s} + \|\varphi\|_\infty^2\xi_{d-1}(Q)|u|,\end{aligned}$$

where $\xi_{d-1}(S)$ is called a Quermassintegrale of Minkowski and dist stands for the Euclidean distance in $\mathbb{R}^d$. The last inequality follows from the fact that φ is $\mathcal{H}(s, M_1)$ on Q and by the Steiner's formula stated, for instance, in [9], Theorem 3.2.35, page 271. □

Lemma 10. *Under the assumption* (A4), *if Q is a compact set with $\mathcal{C}^2$ boundary and ψ is continuous*

$$\lim_{h\to 0} h^{-1} \int_Q \left(\int_{Q^c} K_h(x-y)\,dy \right)^2 \psi(x)\,dx = \int_{\partial Q} L_Q(x)\psi(x)\,d\mathcal{H}^{d-1}(x),$$

where

$$L_Q(x) = \iint \min\left(\langle z, u(x)\rangle, \langle z', u(x)\rangle\right)_+ K(z)K(z')\,dz\,dz',$$

and $\mathcal{H}^{d-1}$ stands for the $(d-1)$-dimensional Hausdorff measure, $u(x)$ is the normal outer vector of Q at the point x.

Proof. Let us start with an estimate of the integral over Q^c having a simpler dependency w.r.t. h. We define the function

$$\tau(x) = (1_{x\in Q} - 1_{x\notin Q})\,\mathrm{dist}(x, \partial Q).$$

This function is $\mathcal{C}^2$ in the neighborhood of ∂Q and its gradient $-u(x)$ is, for $x \in \partial Q$, the normal inner vector (since ∂Q is $\mathcal{C}^2$, using a local parametrization of Q, we are reduced to the case where ∂Q is a piece of hyperplane). Then

$$\begin{aligned}
\int_{Q^c} K_h(x-y)\,dy &= \int 1_{x+hz\in Q^c} K(z)\,dz \\
&= \int 1_{\tau(x+hz)\le 0} K(z)\,dz \\
&= \int 1_{\tau(x)-h\langle z,u(x)\rangle \le ah^2} K(z)\,dz,
\end{aligned}$$

where a actually depends on x and z but is smaller than a constant related to the curvature of ∂Q. Hence,

$$\begin{aligned}
&\left| \int_{Q^c} K_h(x-y)\,dy - \int 1_{\tau(x)-h\langle z,u(x)\rangle\le 0} K(z)\,dz \right| \\
&\quad \le \int |1_{\tau(x)-h\langle z,u(x)\rangle\le ah^2} - 1_{\tau(x)-h\langle z,u(x)\rangle\le 0}| K(z)\,dz \\
&\quad \le \int 1_{|\tau(x)-h\langle z,u(x)\rangle|\le |a|h^2} K(z)\,dz \\
&\quad \le a_0 h 1_{\tau(x)\le m_0 h}
\end{aligned}$$

for some a_0 and m_0, because the integration domain is a band of width $|a|h$. Hence,

$$\int_{Q^c} K_h(x-y)\,dy = \int 1_{\tau(x)\le h\langle z,u(x)\rangle} K(z)\,dz + a_1(x)h1_{\tau(x)\le m_0 h},$$

where a_1 is bounded. Since the second term has a $O(h^2)$ integral over Q, its contribution in the limit is negligible, and it suffices to prove that

$$\lim_{h\to 0} h^{-1}\int_Q \left(\int 1_{\tau(x)\le h\langle z,u(x)\rangle} K(z)\,dz\right)^2 \psi(x)\,dx = \int_{\partial Q} L_Q(x)\psi(x)\,d\mathcal{H}^{d-1}(x).$$

By setting

$$\varphi(x,t) = \left(\int 1_{0\le t\le\langle z,u(x)\rangle} K(z)\,dz\right)^2 \psi(x)1_{x\in Q},$$

the latter equality can be rewritten as

$$\lim_{h\to 0} h^{-1}\int \varphi\big(x,h^{-1}\tau(x)\big)\,dx = \int_{\partial Q} L_Q(x)\psi(x)\,d\mathcal{H}^{d-1}(x).$$

From Proposition 3, page 118 of [7], we have for any integrable function q and f Lipschitz with $\operatorname{essinf}|\nabla f| > 0$:

$$\int_{f\ge 0} q(x)\,dx = \int_0^\infty \left(\int_{f=s} \frac{q(x)}{|\nabla f(x)|}\,d\mathcal{H}^{d-1}(x)\right) ds$$

hence, with $f(x) = h^{-1}\tau(x)$ and $q(x) = \varphi(x, h^{-1}\tau(x))$, we obtain

$$h^{-1}\int \varphi\big(x,h^{-1}\tau(x)\big)\,dx = \int_0^\infty \left(\int_{\tau=hs} \varphi(x,s)\,d\mathcal{H}^{d-1}(x)\right) ds.$$

Letting $h\to 0$, we get

$$\begin{aligned}\lim_{h\to 0} h^{-1}\int \varphi\big(x,h^{-1}\tau(x)\big)\,dx &= \int_0^\infty \left(\int_{\partial Q} \varphi(x,s)\,d\mathcal{H}^{d-1}(x)\right) ds \\ &= \int_{\partial Q}\left(\int_0^\infty \varphi(x,s)\,ds\right) d\mathcal{H}^{d-1}(x).\end{aligned}$$

We can write $\int \varphi(x,s)\,ds$ as

$$\begin{aligned}\int_0^\infty \varphi(x,s)\,ds &= \psi(x)\iiint 1_{0\le s\le\langle z,u(x)\rangle} 1_{0\le s\le\langle z',u(x)\rangle} K(z)K\big(z'\big)\,dz\,dz'\,ds \\ &= \psi(x)\iint \min\big(\langle z,u(x)\rangle,\langle z',u(x)\rangle\big)_+ K(z)K\big(z'\big)\,dz\,dz'.\end{aligned}$$

□

6.4.3. *Weak convergence for degenerate U-statistics*

Theorem 11 (Hall (1984), [12]). *Let* $H_n : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, *with* H_n *symmetric, assume that* $\mathbb{E}[H_n(X_1, X_2)|X_1] = 0$ *and* $\mathbb{E}[H_n(X_1, X_2)^2] < +\infty$. *If*

$$\frac{\mathbb{E}[G_n(X_1, X_2)^2] + n^{-1}\mathbb{E}[H_n(X_1, X_2)^4]}{\mathbb{E}[H_n(X_1, X_2)^2]^2} \overset{n\to+\infty}{\longrightarrow} 0,$$

with $G_n(x, y) = \mathbb{E}[H_n(X_1, x)H_n(X_1, y)]$, *then* $\sum_{j<k} H(X_j, X_k)$ *is asymptotically normally distributed with zero mean and variance given by* $\frac{n^2}{2}\mathbb{E}[H(X_1, X_2)^2]$.

Acknowledgements

The authors would like to thank Céline Vial for helpful comments and advice on a latter version of this article.

Research supported by the Fonds de la Recherche Scientifique (FNRS) A4/5 FC 2779/2014–2017 No. 22342320.

References

[1] Bickel, P.J., Klaassen, C.A.J., Ritov, Y. and Wellner, J.A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. *Johns Hopkins Series in the Mathematical Sciences*. Baltimore, MD: Johns Hopkins Univ. Press. MR1245941

[2] Boucheron, S., Lugosi, G. and Bousquet, O. (2004). Concentration inequalities. In *Advanced Lectures on Machine Learning*. *Lecture Notes in Computer Science* **3176** 208–240. Berlin: Springer.

[3] Chen, S.X. (1999). Beta kernel estimators for density functions. *Comput. Statist. Data Anal.* **31** 131–145. MR1718494

[4] Delecroix, M., Hristache, M. and Patilea, V. (2006). On semiparametric M-estimation in single-index regression. *J. Statist. Plann. Inference* **136** 730–769. MR2181975

[5] Devroye, L.P. and Wagner, T.J. (1980). The strong uniform consistency of kernel density estimates. *J. Multivariate Anal.* **5** 59–77.

[6] Efron, B. and Stein, C. (1981). The jackknife estimate of variance. *Ann. Statist.* **9** 586–596. MR0615434

[7] Evans, L.C. and Gariepy, R.F. (1992). *Measure Theory and Fine Properties of Functions*. *Studies in Advanced Mathematics*. Boca Raton, FL: CRC Press. MR1158660

[8] Evans, M. and Swartz, T. (2000). *Approximating Integrals Via Monte Carlo and Deterministic Methods*. *Oxford Statistical Science Series*. Oxford: Oxford Univ. Press. MR1859163

[9] Federer, H. (1969). *Geometric Measure Theory*. *Die Grundlehren der Mathematischen Wissenschaften* **153**. New York: Springer. MR0257325

[10] Folland, G.B. (1999). *Real Analysis*: *Modern Techniques and Their Applications*, 2nd ed. *Pure and Applied Mathematics* (*New York*). New York: Wiley. MR1681462

[11] Gamboa, F., Loubes, J.-M. and Maza, E. (2007). Semi-parametric estimation of shifts. *Electron. J. Stat.* **1** 616–640. MR2369028

[12] Hall, P. (1984). Central limit theorem for integrated square error of multivariate nonparametric density estimators. *J. Multivariate Anal.* **14** 1–16. MR0734096

[13] Hall, P. and Heyde, C.C. (1980). *Martingale Limit Theory and Its Application*: *Probability and Mathematical Statistics*. New York: Academic Press. MR0624435
[14] Härdle, W. (1990). *Applied Nonparametric Regression*. *Econometric Society Monographs* **19**. Cambridge: Cambridge Univ. Press. MR1161622
[15] Härdle, W., Marron, J.S. and Tsybakov, A.B. (1992). Bandwidth choice for average derivative estimation. *J. Amer. Statist. Assoc.* **87** 218–226. MR1158640
[16] Härdle, W. and Stoker, T.M. (1989). Investigating smooth multiple regression by the method of average derivatives. *J. Amer. Statist. Assoc.* **84** 986–995. MR1134488
[17] Jones, M.C. (1993). Simple boundary correction for kernel density estimation. *Stat. Comput.* **3** 135–146.
[18] Oh, M.-S. and Berger, J.O. (1992). Adaptive importance sampling in Monte Carlo integration. *J. Stat. Comput. Simul.* **41** 143–168. MR1276184
[19] Robinson, P.M. (1988). Root-N-consistent semiparametric regression. *Econometrica* **56** 931–954. MR0951762
[20] Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. *Monographs on Statistics and Applied Probability*. London: Chapman & Hall. MR0848134
[21] Stone, C.J. (1975). Adaptive maximum likelihood estimators of a location parameter. *Ann. Statist.* **3** 267–284. MR0362669
[22] Stone, C.J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8** 1348–1360. MR0594650
[23] Tsybakov, A.B. (2009). *Introduction to Nonparametric Estimation*. *Springer Series in Statistics*. New York: Springer. MR2724359
[24] Vial, C. (2003). Deux contributions à l'étude semi-paramétrique d'un modèle de régression. Ph.D. thesis, Univ. Rennes.
[25] Zhang, P. (1996). Nonparametric importance sampling. *J. Amer. Statist. Assoc.* **91** 1245–1253. MR1424622

Received September 2014 and revised March 2015