ESAIM: PS 18 (2014) 400–417
DOI: 10.1051/ps/2014010

ESAIM: Probability and Statistics
www.esaim-ps.org

ADAPTIVE ESTIMATION OF A DENSITY FUNCTION USING BETA KERNELS

Karine Bertin[1] and Nicolas Klutchnikoff[2]

Abstract. In this paper we are interested in the estimation of a density − defined on a compact interval of $\mathbb{R}$ − from n independent and identically distributed observations. In order to avoid boundary effect, beta kernel estimators are used and we propose a procedure (inspired by Lepski's method) in order to select the bandwidth. Our procedure is proved to be adaptive in an asymptotically minimax framework. Our estimator is compared with both the cross-validation algorithm and the oracle estimator using simulated data.

Mathematics Subject Classification. 62G05, 62G07, 62G20.

Received May 15, 2013. Revised March 15, 2014.

1. Introduction

This paper deals with density estimation using Beta kernel estimators. In a first paper [2] – cited as (B-K) throughout this paper – the authors investigated the properties of beta kernel estimators of the density in an asymptotical minimax framework. Such estimators were first introduced by Chen [5, 6] in order to avoid the classical boundary effect which arises using classical kernels.

Different methods have been developed to solve the boundary bias problem. Let us briefly mention some of them. A classical and popular method is to reflect the data near the boundary in order to reduce the side effect (see for example [7, 21, 22]). Another popular method is to use "boundary kernels" (see [14, 16, 19], among others). Last, let us highlight the paper written by Zhang and Karunamuni [24]. In this paper, the authors investigate a method using the local polynomial fitting method.

The Beta kernel approach is another attempt to solve this problem. Given a sample $X_1, \ldots, X_n$, these asymmetric kernel estimators are defined by:

$$\tilde{f}_b(t) = \frac{1}{n} \sum_{k=1}^{n} K_{t,b}(X_k),$$

where b is a bandwidth parameter and the asymmetric kernel $K_{t,b}$ − linked with Beta distribution − is defined by equation (2.1) in Section 2.3. Let us mention that there exists a generalization of Beta kernels (and Gamma kernels), called *associated kernels* (see [1, 15]).

Keywords and phrases. Beta kernels, adaptive estimation, minimax rates, Hölder spaces.

[1] CIMFAV, Universidad de Valparaíso, Av. Pedro Montt, 2421 Valparaíso, Chile. karine.bertin@uv.cl

[2] CREST (ENSAI) et IRMA (UMR 7501 Université de Strasbourg et CNRS), Campus de Ker-Lann, Rue Blaise Pascal, BP 37203, 35172 BRUZ cedex, France. nicolas.klutchnikoff@ensai.fr

Beta kernels were successfully used in empirical applications: Renault and Scaillet [20] used such procedures in order to estimate recovery rate distributions in finance while Gustafsson *et al.* [11] used them to estimate operational loss distributions in risk management.

Bandwidth selection is the main problem in kernel density estimation. Several methods have been developed for choosing the bandwidth. In the case of Beta kernel estimators, few methods have been proposed.

A very popular one is the cross-validation. Such kind of procedure was studied in [12,23] in order to select an optimal bandwidth for kernel density estimates. But the main assumption of this paper is that kernels are of the classical form (*i.e.* obtained by translating and scaling a symmetric kernel). Nevertheless, the heuristic of this method does not require such an assumption and the cross validation approach was used in [3] in interaction with beta kernels. Bouezmarni and Rombouts [4] also proved that, in a multivariate framework, this procedure leads to an optimal choice (in a suitable sense) of the bandwidth among a finite family.

Another method consists in adopting a minimax approach in which one has to assume that the function to estimate belongs to a given class of functions. In (B-K), it is proved that, if the underlying density belongs to a Hölder space with smoothness parameter β less than 2 and if, moreover, the quality of estimation is measured in L^p loss (with $1 \leq p < 4$), then there exists a beta kernel estimator with properly chosen bandwidth which is optimal (in other words, it attains the minimax rate of convergence $n^{-\beta/(2\beta+1)}$). Unfortunately, as it is always the case in the minimax approach, the choice of the bandwidth depends on β which is unknown in practical situations.

The main goal of this work is to treat the smoothness of the density as a nuisance parameter and to find an adaptive procedure with respect to this parameter: we want to construct a data-driven procedure of estimation which attains the minimax rate of convergence without knowing the regularity β. This will furnish a new method for choosing the bandwidth for Beta kernel estimators.

Here, we present a procedure based on the well-known Lepski's approach (see Lepski [17]). This approach gives a general framework to construct adaptive estimators. Given a family of estimators indexed by a tuning parameter b which belongs to a finite collection of bandwidth B, the procedure consists in selecting $\hat{b} \in B$ using a data-driven criterion. We propose in this paper a Lepski-type procedure based on Beta kernel estimators which is a modification of the Lepski procedure.

In his paper, Lepski proved that his procedure, applied with classical kernel estimators, gives adaptive estimators. The key point of the proof is a concentration property of the estimators around the estimated density function (condition A3 in [17]). Obtaining a similar result for Beta kernel estimators is not trivial. This is proved in Proposition 3.2 which is the essential theoretical contribution of our paper.

We will concentrate our theoretical study to the cases where $p \in \{1, 2\}$. Let us explain this choice: first it is well-known that the L^1-loss is closely related to the total variation distance between probability measures and thus is of particular interest for density estimation [9]. On the other hand, one of our goals is to compare our procedure with the cross-validation method and the heuristic of this method requires the use the L^2-loss.

We prove that our procedure is adaptive in L^1 and L^2 losses. Moreover, we perform simulations and obtain that our procedure behaves similarly to the oracle estimator. Consequently, since the computer time of our method is much smaller than that of the cross validation, our procedure can be viewed as a competitive alternative in order to select a data-driven bandwidth for beta kernel estimators.

Our paper is organized as follows: in Section 2 we present the statistical model and the goals of our study. Section 3 is devoted to the presentation of our estimation procedure and to the statement of our results. Some simulations are performed in Section 4 where our procedure is compared with the cross-validation method and the *oracle* estimator. Section 5 is devoted to the proofs of the main results of our paper. Finally, Section 6 contains all auxiliary results and their proofs.

2. Statistical model

2.1. Density estimation

Let us suppose that $X_1, \ldots, X_n$ are n independent and identically distributed (i.i.d.) observations from a distribution $\mathbb{P}_f$ which admits a density f with respect to the Lebesgue measure on $[0,1]$. Our goal is to estimate with best possible accuracy the unknown density function f.

2.2. Assumption on the density

Following our first work (B-K), it will be supposed that the unknown density function f belongs to a ball of a Hölder space. The radius of this ball is denoted by L which is a positive known constant. Let us recall the definition of the Hölder class $\Sigma(\beta)$ which consists of all the density functions defined on $[0,1]$, m_β times differentiable and such that for all $(x,y) \in [0,1]^2$:

$$\left| f^{(m_\beta)}(x) - f^{(m_\beta)}(y) \right| \leq L|x-y|^{\beta - m_\beta},$$

where $m_\beta = \sup\{\ell \in \mathbb{N} : \ell < \beta\}$.

Now, we fix a small parameter $0 < \varepsilon_1 < 3/5$ and define:

$$\gamma_{0,1} = \varepsilon_1 \qquad \text{and} \qquad \gamma_{0,2} = \frac{1}{2}\,\frac{1+\varepsilon_1}{1-\varepsilon_1},$$

and, for $p = 1, 2$,

$$\mathscr{B}_p = [\gamma_{0,p}, 2].$$

Our procedure depends on ε_1 which is used to ensure a uniform behavior in concentration inequalities given in Proposition 3.2. In practice, this parameter will be chosen close to 0 to have $\mathscr{B}_p$ as large as possible (see Sect. 4 where $\varepsilon_1 = 10^{-4}$). Such a choice does not impact the performance of our procedure.

The main assumption of this paper is that there exists β in $\mathscr{B}_p$, the nuisance parameter, such that the unknown density f of the observations belongs to $\Sigma(\beta)$.

2.3. Minimax estimation

In order to measure the quality of an arbitrary estimator $\tilde{f}_n$ of the unknown density function f, we will introduce, for all $0 < \beta \leq 2$ its risk over $\Sigma(\beta)$ defined by:

$$R_{n,p}(\tilde{f}_n, \Sigma(\beta)) = \sup_{f \in \Sigma(\beta)} R_{n,p}(\tilde{f}_n, f)$$

where

$$R_{n,p}(\tilde{f}_n, f) = \left(\mathbb{E}_f \|\tilde{f}_n - f\|_p^p \right)^{\frac{1}{p}},$$

and $\mathbb{E}_f$ is the expectation with respect to $\mathbb{P}_f$. The minimax rate of convergence on $\Sigma(\beta)$ is defined as $r_{n,p}(\beta) = \inf_{\tilde{f}_n} R_n(\tilde{f}_n, \Sigma(\beta))$ where the infimum is taken over all the estimators. The asymptotic behaviour of $r_{n,p}(\beta)$ is well-known up to a multiplicative constant (see [13]) and is of order $\varphi_n(\beta) = n^{-\beta/(2\beta+1)}$ which does not depend on p.

This rate of convergence is achieved by beta kernel estimators (see B-K) with properly chosen bandwidth. Let us recall the definition of these estimators. Define, for all $(x,t,b) \in [0,1]^2 \times (0,1]$:

$$K_{t,b}(x) = \frac{x^{\frac{t}{b}}(1-x)^{\frac{1-t}{b}}}{B\left(\frac{t}{b}+1, \frac{1-t}{b}+1\right)} \tag{2.1}$$

where B is the standard Beta function, *i.e.*, if Γ denotes the Gamma function, then $B(u,v) = \Gamma(u)\Gamma(v)/\Gamma(u+v)$.

In (B-K), it is proved that there exist two positive constants $\varkappa_1$ and $\varkappa_2$ such that, for all $\beta \in (0,2]$, we have, for b small enough, the following control on the bias term:

$$\forall t \in [0,1], \quad \sup_{f \in \Sigma(\beta)} |\mathbb{E}_f(\tilde{f}_b(t)) - f(t)| \leq \varkappa_1 b^{\beta/2}, \tag{2.2}$$

and the following control on the stochastic term:

$$\left(\int_0^1 \mathbb{E}_f \left| \tilde{f}_b(t) - \mathbb{E}_f(\tilde{f}_b(t)) \right|^p \right)^{1/p} \leq \frac{\varkappa_2}{\sqrt{nb^{1/2}}}. \tag{2.3}$$

Using these inequalities, taking $b = b_n(\beta) = n^{-2/(2\beta+1)}$ and defining:

$$\hat{f}_\beta(t) = \tilde{f}_{b_n(\beta)}(t) = \frac{1}{n} \sum_{k=1}^{n} K_{t,b_n(\beta)}(X_k),$$

we obtain:

$$\sup_{f \in \Sigma(\beta)} \left(\mathbb{E}_f \|\hat{f}_\beta - f\|_p^p \right)^{\frac{1}{p}} \leq (\varkappa_1 + \varkappa_2)\varphi_n(\beta). \tag{2.4}$$

2.4. Adaptive estimation

The minimax strategy is not satisfactory in many practical situations since it furnishes a bandwidth that depends on the unknown regularity β of the estimated function. In this paper, we choose to adopt an adaptive point of view. In this framework, the quality of an estimator is measured simultaneously with several risks. Roughly speaking, the goal is to find a procedure of estimation $\hat{f}$ which achieves the rate $\varphi_n(\beta)$ over each considered functional space $\Sigma(\beta)$ simultaneously (*i.e.* for $\beta \in \mathscr{B}_p$). The precise result is given by Theorem 3.1.

In his paper, Lepski [17] gave a general construction for aggregating minimax estimators in order to obtain an adaptive procedure. This construction consists mainly in choosing (this choice is data-driven and measurable with respect to the observations) the *best* estimator in a given finite (even large) family with respect to a quite simple criterion.

Our procedure of estimation follows the main lines of this method and the main contribution of our paper is the derivation of new concentration inequalities for Beta kernel estimators.

3. Procedure of estimation and main result

The Lepski procedure, adapted to our framework, selects the largest bandwidth such that the beta kernel estimator constructed using this bandwidth is not significantly different from beta kernel estimators constructed with smaller bandwidth which correspond to smaller rates of convergence. Let us define our procedure of estimation in three steps.

3.1. Construction of a grid

Since the procedure is based on a pairwise comparison between estimators we have to construct a finite subset of all bandwidths which is sufficient to approximate all estimators (or, more precisely, all rates of convergence) in our family. To do so, let us consider a regular grid of $\mathscr{B}_p$. Set $K(n) = \lfloor \log n \rfloor^2$ and for all $k \in \mathscr{K}_n = \{0, \ldots, K(n)\}$ let us consider:

$$\gamma_k = \gamma_{0,p} + \frac{k}{K(n)}(2 - \gamma_{0,p}).$$

Let us remark that this choice of $K(n)$ leads to the following property: $\varphi_n(\gamma_k)/\varphi_n(\gamma_{k+1}) \to 1$ when $n \to +\infty$.

3.2. Admissible indexes

Let us define a set of *admissible* indexes as follows:

$$\mathscr{A} = \left\{ k \in \mathscr{K}_n : \forall \ell \leq k, \forall m \leq \ell,\ \|\hat{f}_{\gamma_m} - \hat{f}_{\gamma_\ell}\|_p \leq 2C_p^* \varphi_n(\gamma_m) \right\},$$

where

$$C_1^* = \varkappa_1 + \varkappa_2 + \varepsilon_2 \qquad \text{and} \qquad C_2^* = \varkappa_1 + \sqrt{2\varepsilon_2 + \varkappa_2^2} \tag{3.1}$$

with ε_2 is a constant in $(0,1)$ that can be chosen as small as we want and $\varkappa_1$ and $\varkappa_2$ are defined in equations (2.2) and (2.3). An index k which belongs to $\mathscr{A}$, and the corresponding estimator $\hat{f}_k$, are called *admissible*

3.3. Data-driven choice of our estimator

Next, let us define a data-driven index by $\hat{k} = \sup \mathscr{A}$. As $\hat{k}$ is the supremum of a finite and nonempty set (0 ever belongs to this set) it is measurable with respect to the observations. Finally, set $\hat{f} = \hat{f}_{\gamma_{\hat{k}}}$. In other words, $\hat{f}$ is the admissible estimator with the smallest standard deviation.

Our procedure is then designed so that the bias term of admissible estimators are well controlled. Indeed, the difference between the estimators (in the definition of the set of admissible indexes) is roughly the difference between the bias terms (if n large enough) of the corresponding estimators. So the procedure consists in choosing the estimator with the smallest deviation among the estimators with well-controlled bias.

3.4. Statement of the results

Equipped with these definitions we can state the main result of our paper which ensures that our procedure $\hat{f}$ achieves the minimax rate of convergence $\varphi_n(\beta)$ simultaneously over each functional Hölder space $\Sigma(\beta)$ for $\beta \in \mathscr{B}_p$.

Theorem 3.1. *For $p = 1, 2$, the estimator $\hat{f}$ satisfies*

$$\limsup_{n\to\infty} \sup_{\beta\in\mathscr{B}_p} \varphi_n^{-1}(\beta) R_{n,p}(\hat{f}, \Sigma(\beta)) < +\infty.$$

Proof of this result − given in Section 5 − is based on the following proposition which corresponds to the condition A3 in [17].

Proposition 3.2. *There exists an absolute constant $a > 0$, such that, for $p = 1, 2$ and all $\beta \in \mathscr{B}_p$, we have:*

$$\sup_{f\in\Sigma(\beta)} \sup_{\gamma_{0,p}\leq\gamma\leq\beta} \mathbb{P}_f \left(\|\hat{f}_\gamma - f\|_p > C_p^* \varphi_n(\gamma) \right) \leq \exp\left(-n^a\right).$$

This proposition gives a concentration inequality of the beta kernel estimator around f. This result is new and is the key-property to obtain adaptive estimators by Lepski method. Its proof is given in Section 5.2.

3.5. Comment

The main difference between our procedure and those presented in Lepski [17] consists in the definition of the set $\mathscr{A}$. The original Lepski's procedure defines a set $\widetilde{\mathscr{A}}$ in the following way:

$$\widetilde{\mathscr{A}} = \left\{ k \in \mathscr{K}_n : \forall \ell \leq k,\ \|\hat{f}_{\gamma_k} - \hat{f}_{\gamma_\ell}\|_p \leq C\varphi_n(\gamma_\ell) \right\},$$

where C is a positive constant. Unlike $\widetilde{\mathscr{A}}$, our set $\mathscr{A}$ consists of consecutive integers which implies that finding the supremum is more efficient. Indeed, whereas it is always necessary to perform all the comparisons in the standard Lepski's procedure, our procedure stops as soon as an integer does not belong to $\mathscr{A}$. Consequently the computer time can be drastically reduced in favorable cases.

4. Simulations

4.1. Presentation

The cross-validation is widely used to select the bandwidth in practical situations. As far as we know, this method is the only data-driven method that has been used with beta kernels (see [3,4]). Our goal in this section is to compare, on simulated data, our procedure with both cross-validation and the oracle defined by:

$$f^* = \hat{f}_{\gamma_{k^*}} \quad \text{where} \quad k^* = \underset{k=0,\dots,K_n}{\arg\min} \|\hat{f}_{\gamma_k} - f\|_2^2$$

if f is the density function to be estimated. Notice that our procedure depends on two tuning constants ε_1 and C_2^*. As explained above, we choose $\varepsilon_1 = 0.001$. In Section 4.2 we explain how to calibrate C_2^* in practice.

Since our goal here is to compare the accuracy of our procedure with respect to cross-validation, which is designed to minimize the L^2 risk, the results are presented only for $p = 2$. However the behaviour of the proposed procedure with respect to the oracle is similar whenever $p = 1$ or $p = 2$.

We consider three densities of probability with different behaviour with respect to the smoothness.

Firstly, f_1 is defined as a truncated gaussian density:

$$f_1(x) = c_1 \exp\left(-\frac{(x-0.5)^2}{0.3}\right) I_{[0,1]}(x)$$

where c_1 is a normalizing constant.

Secondly, let us define f_2 as follow:

$$f_2(x) = 1 + \cos(4\pi x).$$

Finally, set $\beta = 0.6$ and let us define f_3 as:

$$f_3(x) = 1 + 2\sum_{k=1}^{4} (-1)^{k+1} \left(\frac{1}{8^\beta} - \left|x - \frac{2k-1}{8}\right|^\beta\right) I_{[\frac{k-1}{4}, \frac{k}{4}]}(x).$$

Note that f_1 and f_2 are of regularity β for all $\beta \leq 2$ and f_3 is of regularity 0.6. Moreover the function f_2 is more difficult to estimate due to its oscillatory behaviour.

For each of these functions, we generate samples of size $n = 100, 200, 500$ and 1000. For each sample, with a density f_j, we compute $\hat{f}_j$ (the estimator obtained by our procedure), $\bar{f}_j$ (the estimator obtained by cross-validation) and f_j^* the oracle estimator. Then we compute the integrated squared error of these estimators. We replicate these simulations 200 times. The obtained values are represented in boxplots to visualize the performance of the estimators.

Figures 1–3 present these boxplots. For each figure, the boxplots BK, CV and O correspond to the integrated squared error (ISE) of respectively: (i) our procedure, (ii) the cross-validation procedure and (iii) the oracle estimator for different sample size values.

4.2. Calibration of the procedure

In our procedure, the choice of the tuning parameter C_2^* is very important and can be explicitly obtained. However, this constant is large and leads, in numerical simulations, to oversmoothing estimation.

In practice, in order to calibrate this constant, we adopt a similar strategy to that presented in [8].

We choose the constant C_2^* that leads to the best estimation for the beta distribution with parameters $(2, 2)$ (namely $B(2, 2)$) which is a classical symmetric and regular density with support $[0, 1]$.

More precisely, for a given size of sample $n \in \{100, 200, 500, 1000\}$, we simulate 200 samples of distribution $B(2, 2)$ of size n. Then, we consider a grid of constants from 0.05 to 1 by step 0.05. For each value in this grid,

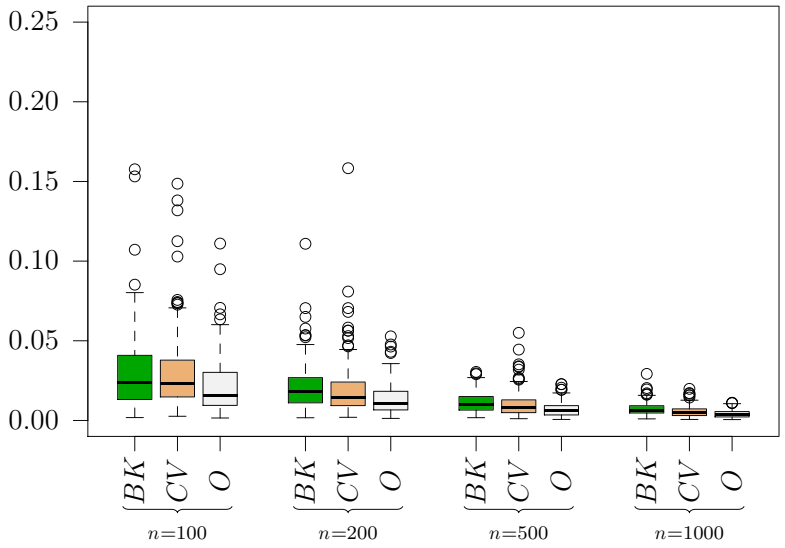

FIGURE 1. Boxplot of the ISE for f_1. BK: our procedure; CV: cross-validation procedure; O: oracle estimator.

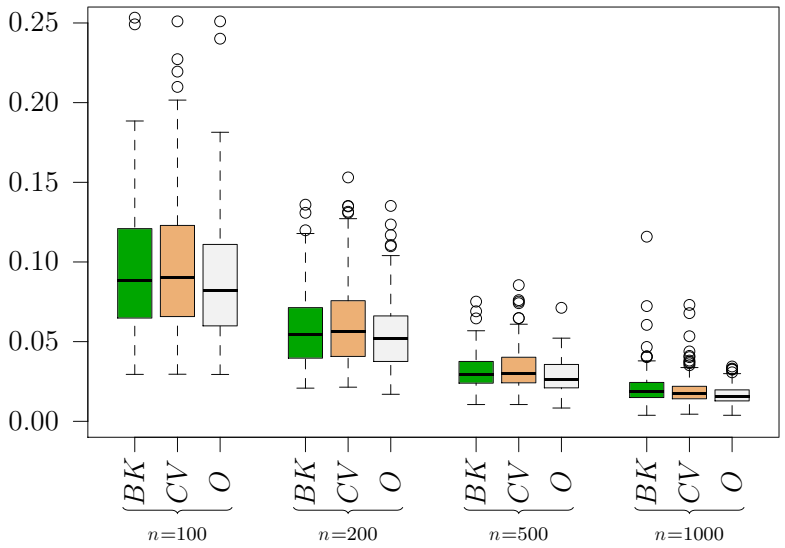

FIGURE 2. Boxplot of the ISE for f_2. BK: our procedure; CV: cross-validation procedure; O: oracle estimator.

we compute 200 estimators $\hat{f}$ based on the 200 samples and then we estimate the mean squared error. We select the value of the grid which minimizes this estimated mean squared error. We give the computed values for C_2^* (which depend on n) in the following table:

n	100	200	500	1000
C_2^*	0.5	0.5	0.6	0.65

Figure 4 represents the estimated MISE drawn for different values of the tuning constant for a 500-sample with distribution $B(2, 2)$.

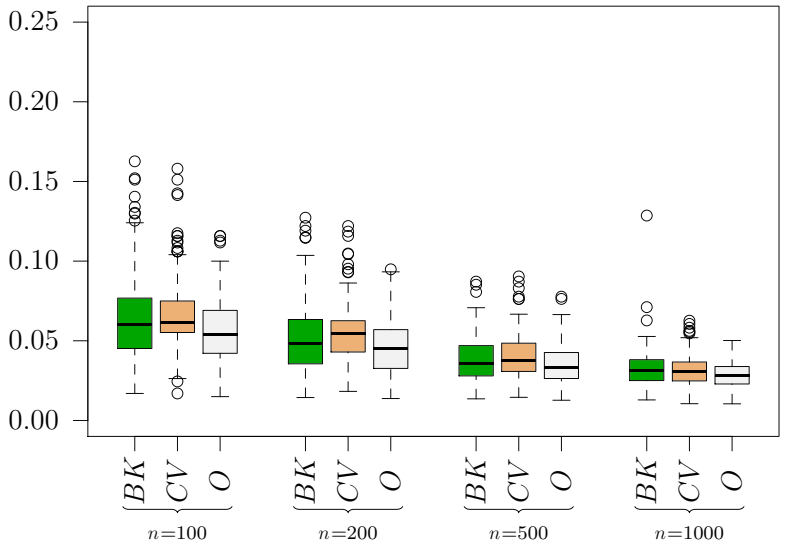

FIGURE 3. Boxplot of the ISE for f_3. BK: our procedure; CV: cross-validation procedure; O: oracle estimator.

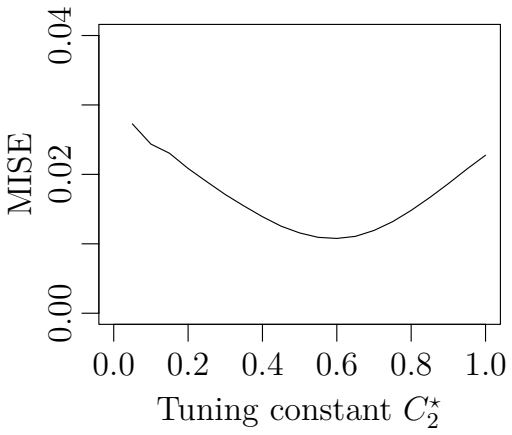

FIGURE 4. Estimated MISE (using a Monte-Carlo method with 200 replications) as a function of the tuning constant for a sample of size 500 with distribution $B(2,2)$.

4.3. Comments

Our procedure and cross validation behave similarly and give estimations close to the oracle estimator. For example, the median ratio between our procedure and the oracle is around 1.1 for a 500-sample of each of the three functions. As one can expect, quality of estimation improves with sample size for the three functions. The quality of estimation is better for function f_1 (which is a very regular function), than for the two other functions which are more irregular or oscillating.

The decisive advantage of our procedure is the computer time. In the following table, we put the running mean time (in seconds) for different sample sizes.

n	100	200	500	1000
Our procedure	0.035	0.064	0.183	0.329
Cross Validation	0.507	1.546	6.816	19.405

This advantage, combined with both good theoretical and practical behaviour, allow us to consider our estimator as a competitive alternative to the cross-validation for practical purpose.

5. Proof of main results

In the following, the letter C will denote a positive constant, the value of which may vary from line to line and may depend only on L.

5.1. Proof of Theorem 3.1

Set $p \in \{1,2\}$, $\beta \in \mathscr{B}_p$ and $f \in \Sigma(\beta)$. First, we need to introduce an index $k(\beta)$ such that $\gamma_{k(\beta)}$ is the "nearest" point of β in our grid. Let us define this index as:

$$k(\beta) = \left\lfloor \frac{\beta - \gamma_{0,p}}{2 - \gamma_{0,p}} K(n) \right\rfloor .$$

In particular the rates of convergence satisfy $\varphi_n(\beta)/\varphi_n(\gamma_{k(\beta)}) \to 1$ as $n \to \infty$. Our goal is to study the following quantity:

$$\mathbb{E}_f \left[\|\hat{f} - f\|_p^p \right] = \Delta_1(n) + \Delta_2(n),$$

where

$$\Delta_1(n) = \sum_{k=0}^{k(\beta)-1} \mathbb{E}_f \left[\|\hat{f}_{\gamma_k} - f\|_p^p I_{\{\hat{k}=k\}} \right] \text{ and } \Delta_2(n) = \mathbb{E}_f \left[\|\hat{f}_{\gamma_{\hat{k}}} - f\|_p^p I_{\{\hat{k} \geq k(\beta)\}} \right].$$

Study of $\Delta_1(n)$

Set $k < k(\beta)$. On the event $\{\hat{k} = k\}$, by definition of $\hat{k}$ we deduce that there exists $\ell < k(\beta)$ and $m \leq \ell$ such that:

$$\|\hat{f}_{\gamma_\ell} - \hat{f}_{\gamma_m}\|_p > 2C_p^* \varphi_n(\gamma_m).$$

This implies that

$$\{\hat{k} = k\} \subset \bigcup_{\ell \leq k(\beta)} \bigcup_{m \leq \ell} A_{\ell,m},$$

where

$$\begin{aligned} A_{\ell,m} &= \left\{ \|\hat{f}_{\gamma_\ell} - \hat{f}_{\gamma_m}\|_p > 2C_p^* \varphi_n(\gamma_m) \right\} \\ &\subset \left\{ \|\hat{f}_{\gamma_\ell} - f\|_p > C_p^* \varphi_n(\gamma_m) \right\} \cup \left\{ \|\hat{f}_{\gamma_m} - f\|_p > C_p^* \varphi_n(\gamma_m) \right\} \\ &\subset \left\{ \|\hat{f}_{\gamma_\ell} - f\|_p > C_p^* \varphi_n(\gamma_\ell) \right\} \cup \left\{ \|\hat{f}_{\gamma_m} - f\|_p > C_p^* \varphi_n(\gamma_m) \right\}. \end{aligned}$$

Thus we obtain:

$$\Delta_1(n) \leq \sum_{k<k(\beta)} \sum_{\ell \leq k(\beta)} \sum_{m \leq \ell} \mathbb{E}_f \left(\|\hat{f}_{\gamma_k} - f\|_p^p I_{A_{\ell,m}} \right).$$

Using Lemma 6.1, we have for n large enough:

$$\forall x \in [0,1], \quad K_{t,b_n(\gamma_k)}(x) \leq 2b_n^{-1}(\gamma_k).$$

This implies that, using Lemma 6.2, we obtain:

$$\|\hat{f}_{\gamma_\ell} - f\|_p \leq C b_n^{-1}(\gamma_k).$$

Thus, since $\gamma_{0,p} \leq \gamma_k$, we have:

$$\begin{aligned}
\Delta_1(n) &\leq C b_n^{-p}(\gamma_k) \sum_{k<k(\beta)} \sum_{\ell \leq k(\beta)} \sum_{m \leq \ell} \mathbb{P}_f\left(A_{\ell,m}\right) \\
&\leq C b_n^{-p}(\gamma_{0,p}) (\log n)^6 \sup_{\gamma_{0,p} \leq \gamma \leq \beta} \mathbb{P}_f\left(\|\hat{f}_\gamma - f\|_p > C_p^* \varphi_n(\gamma)\right) \\
&\leq C n^{2p} (\log n)^6 \sup_{\gamma_{0,p} \leq \gamma \leq \beta} \mathbb{P}_f\left(\|\hat{f}_\gamma - f\|_p > C_p^* \varphi_n(\gamma)\right).
\end{aligned}$$

Using Proposition 3.2 we obtain, for n large enough, that:

$$\Delta_1(n) \leq C \varphi_n^p(\beta).$$

Study of $\Delta_2(n)$

By construction of the procedure, we have, using that $\varphi_n(\beta)/\varphi_n(\gamma_{k(\beta)}) \to 1$ as n tends to infinity,

$$\begin{aligned}
\Delta_2(n) &\leq C \mathbb{E}_f\left[\left(\|\hat{f}_{\hat{k}} - \hat{f}_{\gamma_{k(\beta)}}\|_p^p + \|\hat{f}_{\gamma_{k(\beta)}} - f\|_p^p\right) I_{\{\hat{k} \geq k(\beta)\}}\right] \\
&\leq C\left(\varphi_n^p(\gamma_{k(\beta)}) + \mathbb{E}_f \|\hat{f}_{\gamma_{k(\beta)}} - f\|_p^p\right) \\
&\leq C\left(\varphi_n^p(\beta) + \mathbb{E}_f \|\hat{f}_{\gamma_{k(\beta)}} - f\|_p^p\right) \\
&\leq C\left(\varphi_n^p(\beta) + \left(b_n^{\beta/2}(\gamma_{k(\beta)}) + \frac{1}{n^{1/2} b_n^{1/4}(\gamma_{k(\beta)})}\right)^p\right) \\
&\leq C \varphi_n^p(\beta),
\end{aligned}$$

where the fourth line is a consequence of equations (2.2) and (2.3).

5.2. Proof of Proposition 3.2

In order to prove Proposition 3.2, we do not use the same techniques if $p = 1$ or $p = 2$. In the first case, we use a general concentration inequality for functions satisfying a *bounded difference assumption*. In the second case such a method fails and our proofs are based on classical Hoeffding's or Bernstein's inequalities.

First case $(p = 1)$

Our proof is based on a method used by Devroye and Lugosi [9] in order to obtain concentration inequalities for classical kernels in L^1-loss. Here, we adapt the proof in order to obtain our result. We have:

$$\begin{aligned}
P_n(\gamma) &= \mathbb{P}_f\left(\|\hat{f}_\gamma - f\|_1 \geq C_1^* \varphi_n(\gamma)\right) \\
&= \mathbb{P}_f\left(g(X_1, \ldots, X_n) \geq C_1^* \varphi_n(\gamma)\right) \\
&\leq \mathbb{P}_f\left(g(X_1, \ldots, X_n) - \mathbb{E}_f g(X_1, \ldots, X_n) \geq (C_1^* - \varkappa_1 - \varkappa_2)\varphi_n(\gamma)\right),
\end{aligned}$$

where g is defined in Lemma 6.4. Last inequality follows from the fact $\mathbb{E}_f g(X_1,\dots,X_n) \le (\varkappa_1+\varkappa_2)\varphi_n(\gamma)$ thanks to equation (2.4). Set $0<\varepsilon<1/4$. Using Lemmas 6.3 and 6.4, for $b=b_n(\gamma)$ small enough, we obtain:

$$\begin{aligned} P_n(\gamma) &\le \exp\left(-2\frac{(C_1^*-\varkappa_1-\varkappa_2)^2\varphi_n^2(\gamma)}{n^{-1}b^{-2\varepsilon}}\right)\\ &\le \exp\left(-\varepsilon_2^2 n b^{2\varepsilon}\varphi_n^2(\gamma)\right)\\ &\le \exp\left(-\varepsilon_2^2 n^{\frac{1-4\varepsilon}{2\gamma+1}}\right)\\ &\le \exp\left(-\varepsilon_2^2 n^{\frac{1-4\varepsilon}{5}}\right). \end{aligned}$$

Since $1-4\varepsilon>0$, Proposition 3.2 follows.

Second case ($p=2$)

Let us introduce some notations used throughout this proof. Let $\beta\in\mathscr{B}_2$, $f\in\Sigma(\beta)$ and $X_1,\dots,X_n$ are i.i.d. variables with density f. For all $b\in(0,1)$, we define:

$$\begin{aligned} \eta_{k,b}(t) &= K_{t,b}(X_k)-\mathbb{E}_f\left(K_{t,b}(X_k)\right)\\ Y_{k,k'}(b) &= \int_0^1 \eta_{k,b}(t)\eta_{k',b}(t)dt\\ Y_k(b) &= Y_{k,k}(b) = \|\eta_{k,b}\|_2^2 = \int_0^1 \eta_{k,b}^2(t)\mathrm{d}t. \end{aligned}$$

Our goal is now to bound the quantity

$$P_n(\gamma) = \mathbb{P}_f\left(\|\hat{f}_\gamma - f\|_2 \ge C_2^*\varphi_n(\gamma)\right).$$

We have, thanks to equation (2.2):

$$\begin{aligned} P_n(\gamma) &\le \mathbb{P}\left(\|\hat{f}_\gamma-\mathbb{E}\hat{f}_\gamma\|_2+\|\mathbb{E}\hat{f}_\gamma-f\|_2 > C_2^*\varphi_n(\gamma)\right)\\ &\le \mathbb{P}\left(\|\hat{f}_\gamma-\mathbb{E}\hat{f}_\gamma\|_2 > (C_2^*-\varkappa_1)\varphi_n(\gamma)\right)\\ &\le \mathbb{P}\left(\left\|\sum_{k=1}^n \eta_{k,b_n(\gamma)}(\cdot)\right\|_2^2 > (C_2^*-\varkappa_1)^2(n\varphi_n(\gamma))^2\right). \end{aligned}$$

In order to improve the readability, let us denote $b_n=b_n(\gamma)$, $\varphi_n=\varphi_n(\gamma)$, $Y_k=Y_k(b_n(\gamma))$ and $Y_{k,k'}=Y_{k,k'}(b_n(\gamma))$. Thanks to equation (2.3), one can deduce that:

$$\mathbb{E}_f Y_k \le \varkappa_2^2 n\varphi_n^2.$$

Then the following inequalities hold:

$$\begin{aligned} P_n(\gamma) &\le \mathbb{P}\left(\sum_{k=1}^n Y_k + \sum_{k\ne k'} Y_{k,k'} > (C_2^*-\varkappa_1)^2(n\varphi_n)^2\right)\\ &\le \mathbb{P}\left(\sum_{k=1}^n (Y_k-\mathbb{E}_f Y_k) + \sum_{k\ne k'} Y_{k,k'} > ((C_2^*-\varkappa_1)^2-\varkappa_2^2)(n\varphi_n)^2\right). \end{aligned}$$

Thus, we have:

$$P_n(\gamma) \leq \mathbb{P}\left(\sum_{k=1}^{n}(Y_k - \mathbb{E}_f Y_k) > \varepsilon_2(n\varphi_n)^2\right) + \mathbb{P}\left(\sum_{k\neq k'} Y_{k,k'} > \varepsilon_2(n\varphi_n)^2\right).$$

On the one hand, the first term can be bounded using Hoeffding's inequality and Lemma 6.5. Indeed we obtain:

$$\begin{aligned}\mathbb{P}\left(\sum_{k=1}^{n}(Y_k - \mathbb{E}_f Y_k) > \varepsilon_2(n\varphi_n)^2\right) &\leq \exp\left(-\frac{2\varepsilon_2^2(n\varphi_n)^4}{\varkappa_3^2 n b_n^{-2}\log^2 n}\right)\\ &\leq \exp\left(-\frac{Cn^{\frac{2\gamma-1}{2\gamma+1}}}{\log^2 n}\right)\\ &\leq \exp\left(-\frac{Cn^{\varepsilon_1}}{\log^2 n}\right). \qquad (5.1)\end{aligned}$$

On the other hand, the second term can be bounded using Bernstein type inequality for U-statistics see [10] combined with decoupling argument see [25] and Lemma 6.5. We have:

$$\begin{aligned}\mathbb{P}\left(\sum_{k\neq k'} Y_{k,k'} > \varepsilon_2(n\varphi_n)^2\right) &\leq C\exp\left(-\frac{\varepsilon_2^2}{C}\min\left[\frac{n^{1/2}}{(\log n)^{1/2}}, \frac{b_n^{-5/8}}{(\log n)^{1/2}}, \frac{n^{1/3}}{(\log n)^{2/3}}\right]\right)\\ &\leq C\exp\left(-\frac{\varepsilon_2^2}{C}\frac{n^{1/4}}{\log n}\right) \qquad (5.2)\end{aligned}$$

Combining equations (5.2) and (5.1), proposition follows.

6. Auxiliary results

Lemma 6.1 (Chen [6])**.** *There exists an absolute constant $c > 0$ such that, for all $b \in (0,1)$ and all $t, x \in [0,1]$, we have:*

$$K_{t,b}(x) \leq \min\left\{\frac{c}{\sqrt{bt(1-t)}}, (1+b)b^{-1}\right\}.$$

Lemma 6.2. *There exists an absolute positive constant Q that only depends on L such that for all $\beta \in (0,2]$:*

$$\sup_{f\in\Sigma(\beta)} \|f\|_\infty \leq Q.$$

This lemma can be easily proved using the mean value theorem.

Lemma 6.3. *Let $g : [0,1]^n \to \mathbb{R}$. Assume that for all i, there exists a constant c_i such that:*

$$\sup_{x_1,\ldots,x_n,x_i'} |g(x_1,\ldots,x_n) - g(x_1,\ldots,x_{i-1},x_i',x_{i+1},\ldots,x_n)| \leq c_i. \qquad (6.1)$$

Then for all random vector $(Z_1,\ldots,Z_n)$ in $[0,1]^n$ and $\eta > 0$, we have:

$$\mathbb{P}\left(g(Z_1,\ldots,Z_n) - \mathbb{E}g(Z_1,\ldots,Z_n) > \eta\right) \leq \exp\left(-2\frac{\eta^2}{\sum_{i=1}^n c_i^2}\right).$$

This Lemma is proved in McDiarmid [18] using martingales techniques.

Lemma 6.4. *Set $g : [0,1]^n \to \mathbb{R}$ defined by*

$$g(x_1,\dots,x_n) = \int_0^1 \left| \frac{1}{n}\sum_{k=1}^n K_{t,b}(x_k) - f(t) \right| \mathrm{d}t.$$

Set $0 < \varepsilon < 1/3$. Then, for b small enough the function g satisfies equation (6.1) with all c_i equal to $b^{-\varepsilon}/n$ up to a multiplicative constant.

Proof. Set $(x_1,\dots,x_n) \in [0,1]^n$. Set an index i and $x_i' \in [0,1]$. Let us also denote $(\tilde{x}_1,\dots,\tilde{x}_n)$ the vector defined by the following equations:

$$\tilde{x}_k = \begin{cases} x_k & \text{if } k \neq i \\ x_i' & \text{otherwise.} \end{cases}$$

We have:

$$\begin{aligned}
|g(x_1,\dots,x_n) - g(\tilde{x}_1,\dots,\tilde{x}_n)| &\leq \int_0^1 \left| \left| \frac{1}{n}\sum_k K_{t,b}(x_k) - f(t) \right| - \left| \frac{1}{n}\sum_k K_{t,b}(\tilde{x}_k) - f(t) \right| \right| \mathrm{d}t \\
&\leq \int_0^1 \left| \frac{1}{n}\sum_k \left(K_{t,b}(x_k) - K_{t,b}(\tilde{x}_k)\right) \right| \mathrm{d}t \\
&\leq \frac{1}{n}\int_0^1 |K_{t,b}(x_i) - K_{t,b}(x_i')| \,\mathrm{d}t \\
&\leq \frac{2}{n}\sup_{x\in[0,1]} \int_0^1 K_{t,b}(x)\mathrm{d}t \\
&\leq \frac{4}{n}\sup_{x\in[0,1/2]} \int_0^{1/2} K_{t,b}(x)\mathrm{d}t.
\end{aligned}$$

Thus, we have to control

$$I_b(x) = \int_0^{1/2} K_{t,b}(x)\mathrm{d}t.$$

Using Lemma 6.1, we obtain for all $0 < \varepsilon < 1$:

$$\begin{aligned}
I_b(x) &\leq (1+b)b^{-\varepsilon} + \int_{b^{1-\varepsilon}}^{1/2} K_{t,b}(x)\mathrm{d}t \\
&\leq (1+b)b^{-\varepsilon} + J_b(x)
\end{aligned}$$

where

$$\begin{aligned}
J_b(x) &= \int_{b^{1-\varepsilon}}^{1/2} K_{t,b}(t) \exp\left\{ \frac{t}{b}\log\left(1 + \frac{x-t}{t}\right) + \frac{1-t}{b}\log\left(1 - \frac{x-t}{1-t}\right)\right\} \mathrm{d}t \\
&\leq \int_{b^{1-\varepsilon}}^{1/2} K_{t,b}(t) \exp\left(-\frac{(t-x)^2}{6b}\right) \mathrm{d}t.
\end{aligned}$$

Last inequality follows from two classical inequalities:

$$\forall u \in (-1,+\infty), \quad \log(1+u) \leq u \quad \text{and} \quad \forall u \in (-1,+\infty), \quad \log(1-u) \leq -u - \frac{u^2}{6}.$$

Using Lemma 6.1 we obtain:

$$\begin{aligned} J_b(x) &\leq a \int_{b^{1-\varepsilon}}^{1/2} \exp\left(-\frac{(t-x)^2}{6b}\right) \frac{\mathrm{d}t}{\sqrt{bt}} \\ &\leq a(J_b^1(x) + J_b^2(x)) \end{aligned}$$

where, assuming $\varepsilon < 1/3$,

$$J_b^1(x) = \int_{b^{1-\varepsilon}}^{b^{2\varepsilon}} \exp\left(-\frac{(t-x)^2}{6b}\right) \frac{\mathrm{d}t}{\sqrt{bt}}$$

and

$$J_b^2(x) = \int_{b^{2\varepsilon}}^{1/2} \exp\left(-\frac{(t-x)^2}{6b}\right) \frac{\mathrm{d}t}{\sqrt{bt}}.$$

The quantity $J_b^1(x)$ can be written as follows:

$$J_b^1(x) = \int_{\frac{b^{1-\varepsilon}-x}{\sqrt{b}}}^{\frac{b^{2\varepsilon}-x}{\sqrt{b}}} \exp\left(-\frac{u^2}{6}\right) \frac{\mathrm{d}u}{\sqrt{x+\sqrt{b}u}}.$$

As the function under the integral sign (including the characteristic function of the segment which depends on b) tends to 0 almost everywhere and is bounded by $u \mapsto \exp(-u^2/6)/\sqrt{x}$ which is integrable, one can conclude that $J_b^1(x)$ goes to 0 with b.

The quantity $J_b^2(x)$ can be bounded as follows:

$$\begin{aligned} J_b^2(x) &= \int_{\frac{b^{2\varepsilon}-x}{\sqrt{b}}}^{\frac{1/2-x}{\sqrt{b}}} \exp\left(-\frac{u^2}{6}\right) \frac{\mathrm{d}u}{\sqrt{x+\sqrt{b}u}} \\ &\leq b^{-\varepsilon} \int_{\frac{b^{2\varepsilon}-x}{\sqrt{b}}}^{\frac{1/2-x}{\sqrt{b}}} \exp\left(-\frac{u^2}{6}\right) \mathrm{d}u \\ &\leq b^{-\varepsilon} \int_{\mathbb{R}} \exp\left(-\frac{u^2}{6}\right) \mathrm{d}u. \end{aligned}$$

Finally, for all $0 < \varepsilon < 1/3$, the quantity $I_b(x)$ is smaller than $b^{-\varepsilon}$ up to a multiplicative constant which does not depend on x. Lemma follows. □

Lemma 6.5. *There exists absolute positive constants $\varkappa_3$ and $\varkappa_4$ such that, for all $\beta \in \mathscr{B}_2$, $f \in \Sigma(\beta)$ and any sequence (b_n) that satisfies $n^{-2} \leq b_n \leq 1$, we have:*

- *Almost surely:*
$$Y_k(b_n) \leq \varkappa_3 b_n^{-1}(\log n) \quad \text{and} \quad |Y_{k,k'}(b_n)| \leq \varkappa_3 b_n^{-1}(\log n).$$
- *If $k \neq k'$:*
$$\mathbb{E}_f Y_{k,k'}(b_n) = 0, \qquad \mathbb{E}_f Y_{k,k'}^2(b_n) \leq \varkappa_4 b_n^{-3/4} \log n.$$

Proof. For the sake of simplicity, we will denote $Y_{k,k'} = Y_{k,k'}(b_n)$ and $Y_k = Y_k(b_n)$.

Firstly, let us remark that, using Lemma 6.2, we have:

$$|\eta_{k,b_n}(t)| \leq K_{t,b_n}(X_k) + Q.$$

Using this inequality combined with Lemma 6.1 we deduce the following bound on $\eta^2_{k,b_n}(t)$:

$$\eta^2_{k,b_n}(t) \le C \min\left\{b_n^{-2}, (b_n t(1-t))^{-1}\right\}.$$

Now, we obtain:

$$\begin{aligned} Y_k &= \int_0^1 \eta^2_{k,b_n}(t)\mathrm{d}t \\ &\le C \int_0^1 \min\left\{b_n^{-2}, (b_n t(1-t))^{-1}\right\} \mathrm{d}t \\ &\le C \int_0^{1/2} \min\left\{b_n^{-2}, (b_n t(1-t))^{-1}\right\} \mathrm{d}t \\ &\le C \int_0^{b_n} b_n^{-2}\mathrm{d}t + \int_{b_n}^{1/2} (b_n t(1-t))^{-1}\mathrm{d}t. \end{aligned}$$

Since $b_n \ge n^{-2}$, this leads to

$$\begin{aligned} Y_k &\le C\left(b_n^{-1} + b_n^{-1}(\log(b_n^{-1}) + \log 2)\right) \\ &\le C b_n^{-1}\log n. \end{aligned}$$

First inequality is proved.
The second inequality is quite simple because $|Y_{k,k'}| \le \sqrt{Y_k Y_{k'}}$.
The first equality $\mathbb{E}_f Y_{k,k'} = 0$ follows easily from the independence of the X_k's.
Finally, let us prove the third inequality:

$$\begin{aligned} \mathbb{E}_f Y^2_{k,k'} &= \int_0^1\int_0^1 \mathbb{E}_f\left(\eta_{k,b_n}(t)\eta_{k,b_n}(u)\right)\mathbb{E}_f\left(\eta_{k',b_n}(t)\eta_{k',b_n}(u)\right)\mathrm{d}t\mathrm{d}u \\ &= \int_0^1\int_0^1 \left(\mathbb{E}_f\left(\eta_{k,b_n}(t)\eta_{k,b_n}(u)\right)\right)^2 \mathrm{d}t\mathrm{d}u \\ &= \int_0^1\int_0^1 \left(\mathbb{E}_f\left(K_{t,b_n}(X_1)K_{u,b_n}(X_1)\right) - \mathbb{E}_f\left(K_{t,b_n}(X_1)\right)\mathbb{E}_f\left(K_{u,b_n}(X_1)\right)\right)^2 \mathrm{d}t\mathrm{d}u. \end{aligned} \tag{6.2}$$

Now we have to control the two following quantities:

$$\begin{cases} E_1 = \mathbb{E}_f\left(K_{t,b_n}(X_1)K_{u,b_n}(X_1)\right) \\ E_2 = \mathbb{E}_f\left(K_{t,b_n}(X_1)\right)\mathbb{E}_f\left(K_{u,b_n}(X_1)\right). \end{cases}$$

The control of E_2 is simple because $\mathbb{E}_f K_{t,b_n}(X_1) = \mathbb{E}f(\xi)$ where $\xi \sim K_{t,b_n}$. As f is bounded by Q we obtain:

$$E_2 \le Q^2.$$

The control of E_1 is quite complex. Firstly, we have

$$E_1 = \mathbb{E}[f(Y)]\frac{B\left(\frac{t+u}{b_n}+1, \frac{1-t+1-u}{b_n}+1\right)}{B\left(\frac{t}{b_n}+1, \frac{1-t}{b_n}+1\right)B\left(\frac{u}{b_n}+1, \frac{1-u}{b_n}+1\right)},$$

where the distribution of Y is of beta type with parameters $(t+u)/b_n+1$ and $(1-t+1-u)/b_n+1$.

Following (B-K), let us introduce the R function defined, for all $z \geq 0$ by:

$$R(z) = \frac{1}{\Gamma(z+1)} \left(\frac{z}{e}\right)^z \sqrt{2\pi z}.$$

Using our bound on f, we obtain:

$$\begin{aligned} E_1 &\leq Q \frac{\Gamma^2\left(\frac{1}{b_n}+2\right)\Gamma\left(\frac{t+u}{b_n}+1\right)\Gamma\left(\frac{1-t+1-u}{b_n}+1\right)}{\Gamma\left(\frac{2}{b_n}+2\right)\Gamma\left(\frac{t}{b_n}+1\right)\Gamma\left(\frac{u}{b_n}+1\right)\Gamma\left(\frac{1-t}{b_n}+1\right)\Gamma\left(\frac{1-u}{b_n}+1\right)} \\ &\leq \frac{Q\left(\frac{1}{b_n}+1\right)^2}{\sqrt{2\pi}\left(\frac{2}{b_n}+1\right)} \frac{\left(\frac{1}{b_n}\right)^{\frac{2}{b_n}}}{\left(\frac{2}{b_n}\right)^{\frac{2}{b_n}+\frac{1}{2}}} \tilde{R}(t,u,b_n) g(t,u) \exp\left\{\frac{1}{b_n} f(t,u)\right\}, \end{aligned}$$

where

$$\begin{aligned} \tilde{R}(t,u,b_n) &= \frac{R\left(\frac{2}{b_n}\right) R\left(\frac{t}{b_n}\right) R\left(\frac{u}{b_n}\right) R\left(\frac{1-t}{b_n}\right) R\left(\frac{1-u}{b_n}\right)}{R\left(\frac{t+u}{b_n}\right) R\left(\frac{1-t+1-u}{b_n}\right) R^2\left(\frac{1}{b_n}\right)}, \\ g(t,u) &= \left(\frac{(t+u)(1-t+1-u)}{tu(1-t)(1-u)}\right)^{1/2}, \\ f(t,u) &= (t+u)\log(t+u) + (1-t+1-u)\log(1-t+1-u) \\ &\quad - t\log(t) - u\log(u) - (1-t)\log(1-t) - (1-u)\log(1-u). \end{aligned}$$

Then using that R is an increasing function that tends to 1 to ∞, we deduce that for b_n small enough

$$E_1 \leq C b_n^{-\frac{1}{2}} g(t,u) \exp\left\{\frac{1}{b_n}\left(f(t,u) - 2\log 2\right)\right\}.$$

Now, let us consider the function

$$h : (t,u) \mapsto f(t,u) + \frac{1}{4}(t-u)^2 \left(\frac{1}{t+u} + \frac{1}{1-t+1-u}\right).$$

Since $\nabla h(t,u) = 0 \iff t = u$ and $h(t,t) = f(t,t) = 2\log 2$ it follows easily that:

$$\begin{aligned} f(t,u) &\leq 2\log 2 - \frac{1}{4}(t-u)^2 \left(\frac{1}{t+u} + \frac{1}{1-t+1-u}\right) \\ &\leq 2\log 2 - (t-u)^2. \end{aligned}$$

Using this result we obtain:

$$E_1 \leq C {b_n}^{-\frac{1}{2}} g(t,u) \exp\left\{-\frac{(t-u)^2}{b_n}\right\}.$$

Now, combining our bounds on E_1 and E_2 with equation (6.2) and applying Lemma 6.1 we obtain:

$$\mathbb{E}_f Y_{k,k'}^2 \leq C\left(a_n^2 b_n^{-4} + b_n^{-1}(I_1 + I_2) + Q^4\right),$$

where

$$I_1 = \int_{a_n}^{1-a_n} \int_{a_n}^{1-a_n} g^2(t,u) \exp\left\{-\frac{2(t-u)^2}{b_n}\right\} I_{\{|t-u|>c_n\}} \mathrm{d}t\mathrm{d}u$$

and

$$I_2 = \int_{a_n}^{1-a_n} \int_{a_n}^{1-a_n} g^2(t,u) \exp\left\{-\frac{2(t-u)^2}{b_n}\right\} I_{\{|t-u|<c_n\}} \mathrm{d}t\mathrm{d}u$$

where the sequences a_n and c_n tend to 0 as n tends to infinity and will be fixed later.

Firstly, let us bound I_1. We have, for n large enough:

$$\begin{aligned} I_1 \leq & \exp\left\{-\frac{2c_n^2}{b_n}\right\} \int_{a_n}^{1-a_n} \int_{a_n}^{1-a_n} g^2(t,u) \mathrm{d}t\mathrm{d}u \\ \leq & C \exp\left\{-\frac{2c_n^2}{b_n}\right\} (\log(a_n))^2. \end{aligned}$$

Secondly, in order to bound I_2 let us remark that for n large enough

$$I_2 = I_3 + I_4$$

where

$$I_3 = \int_{a_n}^{1/2} \int_{a_n}^{1/2+c_n} g^2(t,u) \exp\left\{-\frac{2(t-u)^2}{b_n}\right\} I_{\{|t-u|<c_n\}} \mathrm{d}t\mathrm{d}u$$

and

$$I_4 = \int_{1/2}^{1-a_n} \int_{1/2-c_n}^{1-a_n} g^2(t,u) \exp\left\{-\frac{2(t-u)^2}{b_n}\right\} I_{\{|t-u|<c_n\}} \mathrm{d}t\mathrm{d}u.$$

Moreover we have

$$\begin{aligned} I_3 \leq & 8 \int_{a_n}^{1/2} \int_{a_n}^{1/2+c_n} \frac{t+u}{tu} I_{\{|t-u|<c_n\}} (I_{t\geq u} + I_{t\leq u}) \mathrm{d}t\mathrm{d}u \\ \leq & C \int_{a_n}^{1/2+c_n} \frac{1}{t} \left(\int_{a_n}^{1/2+c_n} I_{\{|t-u|<c_n\}} \mathrm{d}u \right) \mathrm{d}t \\ \leq & C c_n \log(a_n). \end{aligned}$$

By symmetry, we have also $I_4 \leq C c_n \log(a_n)$. Finally, we obtain that

$$\mathbb{E}_f Y_{k,k'}^2 \leq C \left(a_n^2 b_n^{-4} + b_n^{-1} \exp\left\{-\frac{2c_n^2}{b_n}\right\} (\log(a_n))^2 + b_n^{-1} c_n \log(a_n) \right).$$

Now choosing $a_n = b_n^3$ and $c_n = b_n^{1/4}$, we obtain that for n large enough

$$\mathbb{E}_f Y_{k,k'}^2 \leq C b_n^{-3/4} \log n.$$

Lemma is then proved. □

Acknowledgements. The authors have been supported by Fondecyt project 1141258. Karine Bertin has been supported by the grant Anillo ACT-1112 CONICYT-PIA.

References

[1] B. Abdous and C.C. Kokonendji, Consistency and asymptotic normality for discrete associated-kernel estimator. *Afr. Diaspora J. Math.* **8** (2009) 63–70.

[2] K. Bertin and N. Klutchnikoff, Minimax properties of beta kernel estimators. *J. Statist. Plan. Inference* **141** (2011) 2287–2297.

[3] T. Bouezmarni and S. Van Bellegem, Nonparametric beta kernel estimator for long memory time series. Technical report (2009).

[4] T. Bouezmarni and J.V.K. Rombouts, Nonparametric density estimation for multivariate bounded data. *J. Statist. Plann. Inference* **140** (2010) 139–152.
[5] S.X. Chen, Beta kernel estimators for density functions. *Comput. Statist. Data Anal.* **31** (1999) 131–145.
[6] S.X. Chen, Beta kernel smoothers for regression curves. *Statist. Sinica* **10** (2000) 73–91.
[7] D.B.H. Cline and J.D. Hart, Kernel estimation of densities with discontinuities or discontinuous derivatives. *Statistics* **22** (1991) 69–84.
[8] I. Dattner and B. Reiser, Estimation of distribution functions in measurement error models. Technical report (2010).
[9] L. Devroye and G. Lugosi, Combinatorial methods in density estimation. *Springer Series in Statistics*. Springer-Verlag, New York (2001).
[10] E. Giné and R. Latała and J. Zinn, Exponential and moment inequalities for U-statistics. High dimensional probability, vol. II (Seattle, WA, 1999), Birkhäuser Boston, Boston, MA. *Progr. Probab.* **47** (2000) 13–38.
[11] J. Gustafsson, M. Hagmann, J.P. Nielsen and O. Scaillet, Local transformation kernel density estimation of loss distributions. *J. Bus. Econ. Statist.* **27** (2009) 161–175.
[12] P. Hall, Large sample optimality of least squares cross-validation in density estimation. *Ann. Statist.* **11** (1983) 1156–1174.
[13] I.A. Ibragimov and R.Z. Khas'minskiĭ, More on estimation of the density of a distribution. *Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI)* **108** 194, 198 (1981) 72–88.
[14] M.C. Jones, Simple boundary correction for kernel density estimation. *Statist. Comput.* **3** (1993) 135–146.
[15] C.C. Kokonendji and T.S. Kiessé, Discrete associated kernels method and extensions. *Statist. Methodol.* **8** (2011) 497–516.
[16] M. Lejeune and P. Sarda, Smooth estimators of distribution and density functions. *Comput. Statist. Data Anal.* **14** (1992) 457–471.
[17] O.V. Lepski, Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *Teor. Veroyatnost. i Primenen.* **36** (1991) 645–659.
[18] C. McDiarmid, On the method of bounded differences, in *Surveys in combinatorics (Norwich 1989)*, vol. 141 of *London Math. Soc. Lecture Note Ser.* Cambridge University Press, Cambridge (1989) 148–188.
[19] H.-G. Müller, Smooth optimum kernel estimators near endpoints. *Biometrika* **78** (1991) 521–530.
[20] O. Renault and O. Scaillet, On the way to recovery: A nonparametric bias free estimation of recovery rate densities. *J. Banking and Finance* **28** (2004) 2915–2931.
[21] E.F. Schuster, Incorporating support constraints into nonparametric estimators of densities. *Commun. Statist. — Theory Methods* **14** (1985) 1123–1136.
[22] B.W. Silverman, Density estimation for statistics and data analysis. *Monogr. Statist. Appl. Probability*. Chapman & Hall, London (1986).
[23] Ch.J. Stone, An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.* **12** (1984) 1285–1297.
[24] Sh. Zhang and R.J. Karunamuni, On kernel density estimation near endpoints. *J. Statist. Plann. Inference* **70** (1998) 301–316.
[25] H. Victor de la Peña and S.J. Montgomery-Smith, Decoupling inequalities for the tail probabilities of multivariate U-statistics. *Ann. Probab.* **23** (1995) 806–816.