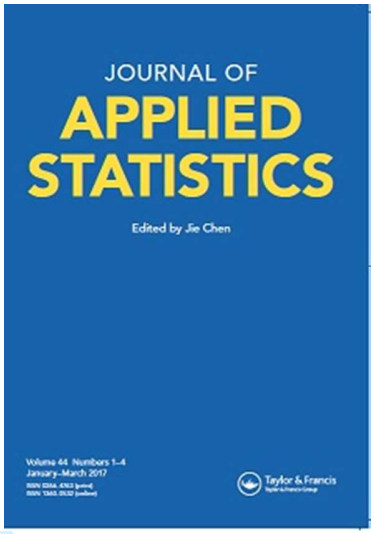

A nonparametric statistical procedure for the detection of marine pollution

Journal:	*Journal of Applied Statistics*
Manuscript ID	CJAS-2017-0334.R3
Manuscript Type:	Original Research Paper
Date Submitted by the Author:	20-Feb-2018
Complete List of Authors:	Bercu, Bernard; Universite de Bordeaux, Institut de Mathématiques de Bordeaux Capderou, Sami; Universite de Bordeaux, Institut de Mathématiques de Bordeaux Durrieu, Gilles; Universite de Bretagne-Sud, Laboratoire de Mathématiques de Bretagne Atlantique; Universite de la Nouvelle-Caledonie, Institut de Sciences Exactes et Appliquees
Keywords:	mathematical statistics, nonparametric estimation, fixed-design regression, Environmental statistics, data analysis
2010 Mathematics Subject Classification:	62G08

A nonparametric statistical procedure for the detection of marine pollution

Bernard Bercu[a], Sami Capderou[a] and Gilles Durrieu[b,c]

[a] Université de Bordeaux, Institut de Mathématiques de Bordeaux, UMR CNRS 5251, 351 Cours de la Libération, 33405 Talence, France;
[b] Université de Bretagne Sud, Laboratoire de Mathématiques de Bretagne Atlantique, UMR CNRS 6205, Campus de Tohannic, 56017 Vannes, France;
[c] Institut de Sciences Exactes et Appliquées, Université de la Nouvelle-Calédonie, 98800 Nouméa, Nouvelle-Calédonie, France.

ARTICLE HISTORY
Compiled February 23, 2018

ABSTRACT
This paper is devoted to the estimation of the derivative of the regression function in fixed-design nonparametric regression. We establish the almost sure convergence as well as the asymptotic normality of our estimate. We also provide concentration inequalities which are useful for small sample sizes. Numerical experiments on simulated data show that our nonparametric statistical procedure performs very well. We also illustrate our approach on high frequency environmental data for the study of marine pollution.

KEYWORDS
Environmental statistics and data analysis; mathematical statistics; nonparametric estimation; fixed-design regression

1. Introduction

Marine ecosystems are facing all over the world the impact of multiple stressors due to seashore pollution and climatic change. Water quality should be dealt as a major concern in our modern society. Marine pollution comes mostly from land based sources and this pollution can lead to the collapse of coastal ecosystems and have public health impacts.

In this context, there is an urgent need to elaborate real-time reliable sensors in order to supervise the water quality within a decision making process. Among these sensors, bio-indicators are increasingly used and are highly effective to reveal pollutions or perturbations in the aquatic systems. For example bivalve mollusks, such as oysters, mussels, and giant clams, are relevant sentinel organisms to evaluate water quality. These bio-indicators can also be considered as biological responses to climate change influencing water-quality and biological organisms. One of the main challenges is to determine how bio-indicators are affected by pollution and climate change.

The high frequency measurements of valve activity in bivalves using valvometry is a possible way to study their behaviors as well as a global analysis of possible pertur-

CONTACT B. Bercu. Email: bernard.bercu@u-bordeaux.fr

bations due to the environment. The idea of valvometry system is to use the bivalves ability to close its shell when exposed to a contaminant as an alarm signal [10]. Therefore, the continuous monitoring of the gap between the two valves is an efficient way to study their behavior when facing water pollution [13, 27]. This system measures at high frequency the bivalve shell movements. Nowadays, valvometric techniques allow autonomously long-term recordings of valve movements without interfering their normal behavior and produce massive volume of data.

Several statistical models have been developed to analyze this huge amount of data [3, 7, 11, 12]. The main objective is to propose optimized statistical procedures in order to detect the world's oceans changes. Here, we propose a statistical procedure to estimate the opening and closing velocities of bio-indicators. As a matter of fact, in an inhospitable environment, bio-indicators behavior is altered. Consequently, detecting changes of the closing and opening velocity can provide insights about the health of bio-indicators and so can be used as bio-indicators of the water quality.

Nonparametric regression processes were provided to modelize the opening/closing activity. The goal was to estimate the regression function governing the relationship between time and amplitude. We refer the reader to [9, 14, 33] for some excellent books on nonparametric kernel estimation of densities and regression functions. Here, we shall focus our attention on the estimation of the derivative of the regression function using a recursive version of the Nadaraya-Watson estimator [1, 18, 25, 26, 36]. In a recent work, [4] investigated the asymptotic properties of this estimate in a random-design regression framework. The recursive estimate is really powerful as soon as data come one by one. Indeed, each new value provides an additional information that can sharpen the estimation of the regression function. It is not necessary to compute again all the values using all the data. However, in most cases in valvometry, data are collected in one batch. We don't have access to the information until midnight where we obtain the last 24-hour measurements. The recursive nature of the estimation is then not necessary. Consequently, we propose an alternative strategy based on a fixed-design regression framework with a more efficient and easy to handle statistical procedure.

Another strategy for the recursive estimation of the regression function and of its derivatives is to make use of the local polynomial fitting approach [21, 34, 35]. The advantages of local polynomial fitting are twofold. The estimates are easily computable and they have nice asymptotic properties. However, we will see in the sequel that the main drawback of this approach is that it requires the kernels to be compactly supported. This is really important as the asymptotic variance of our estimates can be ten times smaller for the Gaussian kernel than for all other kernels with compact supports. Furthermore, a natural question is the extension of our results in a more general setting of dependent data. In all the previous literature on the estimation of the regression function for short-range dependent, long-range dependent or strong mixing data [2, 17, 18, 22], it is necessary to assume that the kernels are compactly supported. It seems really difficult to get rid of this assumption. This is the reason why we have chosen to restrict ourselves to the case of independent data and to favor our statistical application on the detection of marine pollution.

The paper is organized as follows. Section 2 describes the estimation of the first derivative of the regression function. It also describes the data sets compatible with the methodology developed in this paper. We establish in Section 3 the pointwise almost sure convergence, the asymptotic normality as well as concentration inequalities of our

estimate. Section 4 is devoted to numerical experiments in order to study the performance of our nonparametric statistical procedure. Section 5 presents an application for the survey of marine water using high-frequency valvometry. All the technical proofs are postponed to Appendices A, B and C.

2. Estimation of the velocity

In order to record the bivalve activity, each animal is equipped with two electrodes. The relationship between the distances of the two electrodes (Y_n) and the times of the measurement (X_n) is given by the nonparametric regression satisfying, for all $n \geqslant 1$,

$$Y_n = f(X_n) + \varepsilon_n \tag{1}$$

where (ε_n) is the random error sequence and (X_n) is a sequence of independent and identically distributed random variables. Our purpose is to estimate the derivative of the unknown regression function f which is directly associated with the velocity of the valve opening/closing activities of the animals. In [4], a statistical procedure based on the derivative of the recursive Nadaraya-Watson estimator was implemented. The regression function f was estimated, for any $x \in \mathbb{R}$, by

$$\widehat{f}_n(x) = \frac{\sum_{k=1}^{n} \frac{Y_k}{h_k} K\left(\frac{x - X_k}{h_k}\right)}{\sum_{k=1}^{n} \frac{1}{h_k} K\left(\frac{x - X_k}{h_k}\right)}, \tag{2}$$

where the kernel K is a chosen probability density function and the bandwidth (h_n) is a sequence of positive real numbers decreasing to zero.

As previously argued, the data are collected in one batch. Then, it is not necessary to make use of a recursive estimator of f. In addition to that, when we observe a group of 16 animals, one measurement is received every 0.1 seconds. So, the activity of one animal is measured every 1.6 seconds. Hence, the times of measurement (X_n) are perfectly known. Consequently, we are in a fixed design case since the differences between X_n and X_{n+1} are always the same. For all $k = 1, \ldots, n$, denote $t_k = k/n$ where n is the number of measurement taken over the whole day. Therefore, the model described in (1) can be rewritten, for all $k = 1, \ldots, n$, as

$$Y_k = f(t_k) + \varepsilon_k. \tag{3}$$

The nonparametric regression function f is estimated, for any $x \in]0, 1[$, by

$$\widehat{f}_n(x) = \frac{1}{nh_n} \sum_{k=1}^{n} Y_k K\left(\frac{x - t_k}{h_n}\right). \tag{4}$$

Since we are interested in the velocity, we will investigate the asymptotic behavior of the derivative of this estimator given, for any $x \in]0, 1[$, by

$$\widehat{f}_n'(x) = \frac{1}{nh_n^2} \sum_{k=1}^{n} Y_k K'\left(\frac{x - t_k}{h_n}\right). \tag{5}$$

3. Theoretical results

In order to investigate the asymptotic behavior of $\widehat{f}_n'(x)$, it is necessary to introduce several classical assumptions.

$(\mathcal{A}_1)$ The kernel K is either the Gaussian kernel or a positive symmetric bounded function compactly supported, twice differentiable with bounded derivatives, such that

$$\int_{\mathbb{R}} K(x)dx = 1, \qquad \int_{\mathbb{R}} K'(x)dx = 0, \qquad \int_{\mathbb{R}} xK'(x)dx = -1.$$

$(\mathcal{A}_2)$ The regression function f is bounded continuous, twice differentiable with bounded derivatives.

$(\mathcal{A}_3)$ The noise (ε_n) is a sequence of independent and identically distributed random variables with zero mean and finite positive variance σ^2.

Furthermore, the bandwidth (h_n) is a sequence of positive real numbers, decreasing to zero, such that nh_n tends to infinity. For the sake of simplicity, we shall make use of $h_n = 1/n^\alpha$ with $0 < \alpha < 1$.

Our first result on the almost sure convergence of our estimate is as follows.

Theorem 3.1. *Assume that* $(\mathcal{A}_1)$, $(\mathcal{A}_2)$ *and* $(\mathcal{A}_3)$ *hold. Then, for any* x *in* $]0,1[$, *as soon as* $\alpha < 1/3$, *we have the pointwise almost sure convergence*

$$\lim_{n\to\infty} \widehat{f}_n'(x) = f'(x) \qquad a.s. \tag{6}$$

Proof. The proof is given in Appendix A. □

Remark. A similar result was previously established by Gasser and Müller [15] for compactly supported kernels. However, our approach is different from the one of Gasser and Müller which relies on Hoeffding inequality. One can observe that in Theorem 2 of [15], it is necessary to assume that $2\alpha < 1 - 1/p$ where (ε_n) has a finite moment of order $p \geqslant 2$. Hence, for $p = 2$, the almost sure convergence given in [15] only holds in the more restrictive case $\alpha < 1/4$. Our approach relies on the law of iterated logarithm for weighted sums of independent random variables [16] and only finite variance is required for (ε_n).

Our second theoretical result is devoted to the asymptotic normality for $\widehat{f}_n'(x)$. Denote

$$\xi^2 = \int_{\mathbb{R}} \big(K'(x)\big)^2 dx.$$

Theorem 3.2. *Assume that* $(\mathcal{A}_1)$, $(\mathcal{A}_2)$ *and* $(\mathcal{A}_3)$ *hold. Then, for any* x *in* $]0,1[$, *we have as* n *tends to infinity the pointwise asymptotic normality*

$$\sqrt{nh_n^3}\big(\widehat{f}_n'(x) - \mathbb{E}\big[\widehat{f}_n'(x)\big]\big) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma^2\xi^2\right). \tag{7}$$

Furthermore, as soon as $1/5 < \alpha < 1/3$ *we also have as* n *tends to infinity for any* x *in* $]0,1[$,

$$\sqrt{nh_n^3}\big(\widehat{f}_n'(x) - f'(x)\big) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma^2\xi^2\right). \tag{8}$$

Proof. The proof is given in Appendix B. □

Remark. The asymptotic normality was previously investigated by Gasser and Müller [15] for compactly supported kernels. We shall see in the next Section that the asymptotic variance ξ^2 is ten times smaller for the Gaussian kernel than for all other kernels with compact supports. On the same vein, let $\widetilde{f}_n(x)$ be the recursive local polynomial estimator of $f(x)$ proposed by Vilar and Vilar [34] in the random-design framework. It was proven in [34] that

$$\sqrt{nh_n^3}\big(\widetilde{f}_n'(x) - f'(x)\big) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{(1-2\alpha)^2\zeta^2}{(1-\alpha)g(x)}\ \sigma^2\right)$$

where

$$\zeta^2 = \left(\int_{\mathbb{R}} x^2 K(x)dx\right)^{-2} \int_{\mathbb{R}} x^2\big(K(x)\big)^2 dx.$$

However, it was assumed in [34] that the kernel K is compactly supported. For example, for the Epanechnikov kernel, one can easily check that $\zeta^2 = 15/7 \simeq 2.1429$. However, it is easy to see that for the standard Gaussian kernel, $\xi^2 = 1/(4\sqrt{\pi}) \simeq 0.1410$. Consequently, our approach offers much more flexibility in the choice of the kernel K, and the asymptotic variance for $\widehat{f}_n'(x)$ is smaller than that of $\widetilde{f}_n'(x)$. We are also brought to the same conclusion in the case of dependent data [2, 17, 18, 22].

Our last result deals with concentration inequalities for our estimate, in the spirit of McDiarmid's inequality [23]. We wish to stress that this result is true for small or large sample sizes as it holds whatever the value of n is. Denote

$$\Lambda = \sup_{x\in\mathbb{R}} \big|K'(x)\big| \qquad \text{and} \qquad \zeta = \int_{\mathbb{R}} \big|K'(x)\big|dx.$$

Theorem 3.3. *Assume that* $(\mathcal{A}_1)$, $(\mathcal{A}_2)$ *and* $(\mathcal{A}_3)$ *hold. Moreover, assume that one can find a positive constant* M *such that, for all* $1 \leqslant k \leqslant n$, $|Y_k| \leqslant M$ *a.s. Then, for any* x *in* $]0,1[$ *and for any positive* t,

$$\mathbb{P}\Big(\big|\widehat{f}_n'(x) - \mathbb{E}\big[\widehat{f}_n'(x)\big]\big| \geqslant t\Big) \leqslant 2\exp\Big(-\frac{nh_n^2t^2}{2M^2\Lambda^2}\Big), \tag{9}$$

and

$$\mathbb{P}\Big(\Big|\int_{\mathbb{R}} \big|\widehat{f}_n'(x) - f'(x)\big|dx - \mathbb{E}\Big[\int_{\mathbb{R}} \big|\widehat{f}_n'(x) - f'(x)\big|dx\Big]\Big| \geqslant t\Big) \leqslant 2\exp\Big(-\frac{nh_n^2t^2}{2M^2\zeta^2}\Big). \tag{10}$$

Proof. The proof is given in Appendix C. □

4. Simulation results

This section is devoted to numerical experiments in order to evaluate the performances of our statistical estimation procedure and to assess more specifically the sensitivity of our approach when the assumptions of Theorems 3.1 and 3.2 are not satisfied. We shall also consider heavy-tailed distributions for the driven noise (ε_n) such as the well-known Pareto distribution. We have also simulated the short range dependance situation observed in Section 5 via an autoregressive process of order 1.

The data are generated by the nonparametric regression given, for all $k = 1, \ldots, n$, by

$$Y_k = f(t_k) + \varepsilon_k, \tag{11}$$

where $t_k = k/n$, the regression function f is defined, for all x in $[0, 1]$, by

$$f(x) = (x + 2)\sin(4\pi x^2) + 2\sin(8\pi x) \tag{12}$$

and the noise (ε_n) is sequence of independent identically distributed random variables. We implement our statistical procedure with sample size $n = 10\,000$ since we have large datasets in the application described in Sections 5 and 3.2. The simulated data associated with (11) are given in Figure 1. As illustrated on the bottom right of Figure 1 for the Pareto $\mathcal{P}(\theta)$ distribution, if the shape parameter θ is less than or equal to 2, then its variance is infinite. Consequently, the assumption of finite positive variance σ^2 given in $(\mathcal{A}_3)$ is not verified. We use the Gaussian kernel function

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{x^2}{2}\Big). \tag{13}$$

The derivative $f'(x)$ is given, for all x in $[0, 1]$, by

$$f'(x) = \sin(4\pi x^2) + 8\pi x(x + 2)\cos(4\pi x^2) + 16\pi \cos(8\pi x).$$

It is well-known that the choice of the bandwidth $h_n = 1/n^\alpha$ is crucial in nonparametric kernel estimation. In order to select an automatic choice of α, we use a standard cross validation method which leads to $\alpha = 0.3034$. After selecting α by cross validation, the left hand side of Figure 2 shows that the estimator $\widehat{f}_n'(x)$ approaches very well the true derivative $f'(x)$ for the $\mathcal{N}(0, 1)$ noise distribution, as well as for the Laplace $\mathcal{L}(0, 1)$ and Pareto $\mathcal{P}(3)$ noise distributions, in agreement with assumptions of Theorem 6. On the other hand, for the Pareto $\mathcal{P}(3/2)$ noise distribution, the estimate $\widehat{f}_n'(x)$ is quite far from the true derivative $f'(x)$. Furthermore, one can observe boundary effects for x close to 0 and 1. A wide literature is available on how to remove boundary effects, see e.g. the data reflection method given in [24].

In order to illustrate the pointwise asymptotic normality of our estimate, we implement a simulation study based on 2 000 realizations. We numerically check, for the $\mathcal{N}(0, 1)$ and the Pareto $\mathcal{P}(3)$ noise distributions, the asymptotic normality at point $x = 0.2$. One can see in in the first and second drawing of Figure 3 that

$$Z_n(x) = \sqrt{nh_n^3}\big(\widehat{f}_n'(x) - f'(x)\big)$$

is normally distributed and centered around 0. Furthermore, this distribution fits very

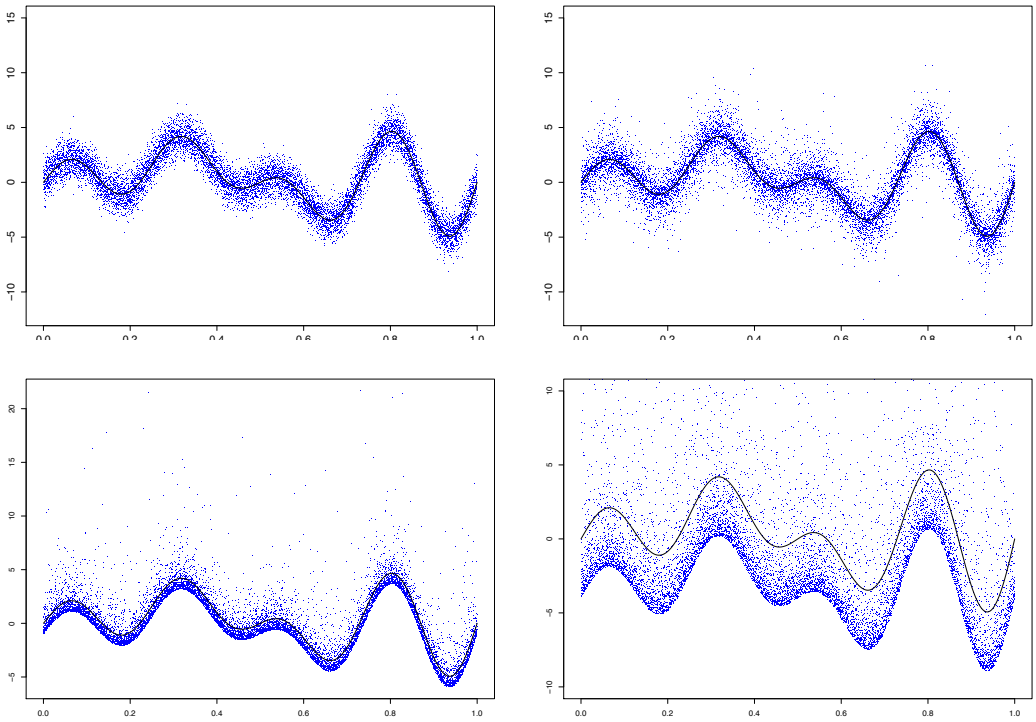

Figure 1.: Simulated data (t_k, Y_k) for $k = 1, \ldots, n$ and $n = 10\,000$. The noise (ε_k) has a $\mathcal{N}(0,1)$ distribution on the top left, a Laplace $\mathcal{L}(0,1)$ distribution on the top right and a Pareto $\mathcal{P}(\theta)$ distribution with shape parameter $\theta = 3$ on the bottom left and $\theta = 3/2$ on the bottom right. The solid line is the true function f given by (12).

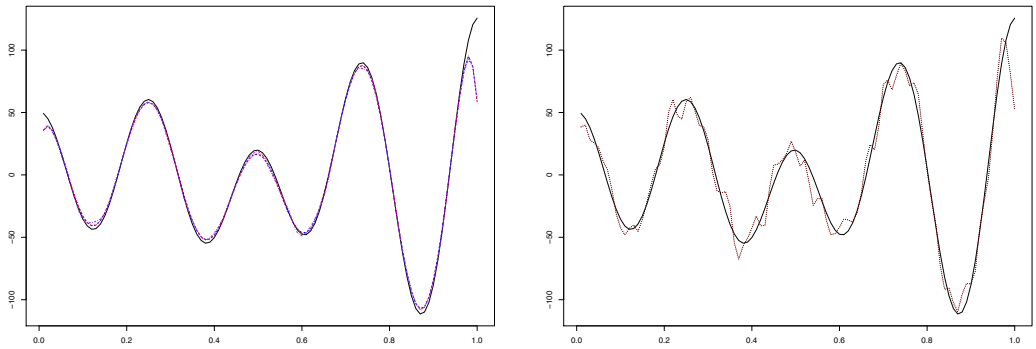

Figure 2.: Illustration of the pointwise almost sure convergence of $\widehat{f}_n'(x)$ for the $\mathcal{N}(0,1)$ (dashed line), the Laplace $\mathcal{L}(0,1)$ (dot-dashed line) and the Pareto $\mathcal{P}(3)$ (dotted line) noise distributions on the left and the Pareto $\mathcal{P}(3/2)$ (dotted line) noise distribution on the right. The solid lines represent the true derivative $f'(x)$.

well with the asymptotic $\mathcal{N}\big(0, \sigma^2\xi^2\big)$ distribution given in Theorem 3.2. It is not surprising to see in the last drawing of Figure 3 that the asymptotic normality is effectively not observed for the Pareto $\mathcal{P}(3/2)$ noise distribution. In order to illustrate the impact

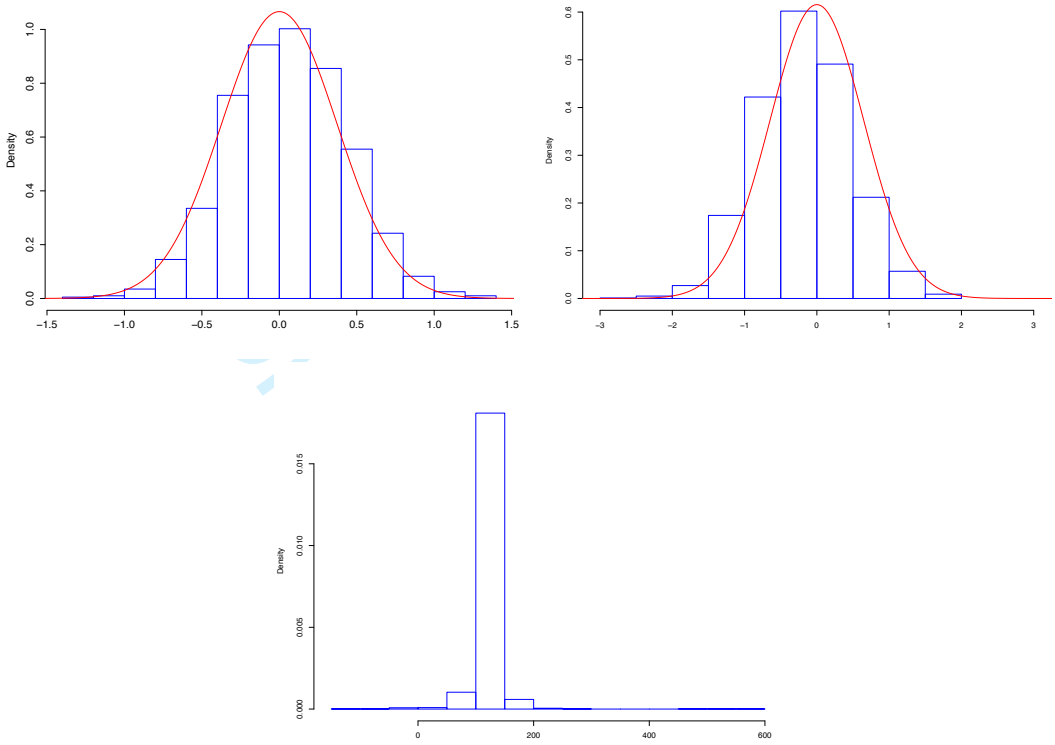

Figure 3.: Illustration of the asymptotic normality at point $x = 0.2$ for the $\mathcal{N}(0, 1)$, the Pareto $\mathcal{P}(3)$, and the Pareto $\mathcal{P}(3/2)$ noise distributions. The density curve represents the asymptotic $\mathcal{N}(0, \sigma^2\xi^2)$ distribution.

of ξ^2 in the asymptotic variance of $Z_n(x)$, we do the same simulation still with $\sigma^2 = 1$ but this time using different kernels. We choose the Epanechnikov, Cosine, Quartic and Triweight kernels, respectively given by

$$K(x) = \frac{3}{4}\big(1 - x^2\big)\mathrm{I}_{|x|\leqslant 1}, \qquad K(x) = \frac{\pi}{4}\cos\Big(\frac{\pi}{2}x\Big)\,\mathrm{I}_{|x|\leqslant 1},$$
$$K(x) = \frac{15}{16}\big(1 - x^2\big)^2\mathrm{I}_{|x|\leqslant 1}, \qquad K(x) = \frac{35}{32}\big(1 - x^2\big)^3\mathrm{I}_{|x|\leqslant 1.}$$

The corresponding values of ξ^2 are respectively 1.5, 1.522, 2.14 and 3.18 . In Figure 4, one can see that the behavior of the variance of $Z_n(x)$ is the same for all values of x in $]0, 1[$ and clearly depends on ξ^2. One can also observe that the smaller asymptotic variance is obtained with the Gaussian kernel.

To assess more specifically the sensitivity of our approach in the presence of autocorrelated errors, we have simulated the short range dependance observed in Section 5 using an autoregressive process of order 1. The data are still generated from the non-

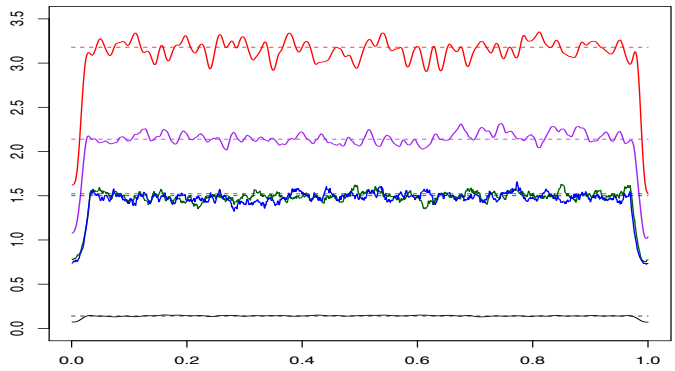

Figure 4.: Representation of the asymptotic variances of $Z_n(x)$ with different kernels. From top to bottom: Triweight, Quartic, Cosine, Epanechnikov and Gaussian. The horizontal dashed lines correspond to the values of ξ^2.

parametric regression model given by (11) with the same regression function given by (12). However, the random noise (ε_n) is autocorrelated. We have chosen a stationary autoregressive process of order 1, given by

$$\varepsilon_n = \rho\,\varepsilon_{n-1} + \xi_n$$

where the parameter $|\rho| < 1$ and the innovations (ξ_n) are independent random variables sharing the same $N(0,1)$ distribution. We have for any $k \geqslant 0$, $\mathrm{corr}(\varepsilon_n, \varepsilon_{n+k}) = \rho^k$. Consequently, the autocorrelation in the errors goes down geometrically as the distance between them goes up. We observed in Figure 5 that even for this autocorrelated noise, our simulations yields satisfactory results for the pointwise almost sure convergence to the true derivative $f'(x)$.

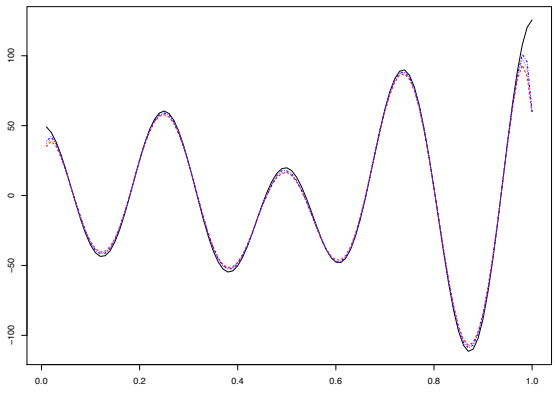

Figure 5.: Illustration of the almost sure convergence of $\widehat{f}_n'(x)$ with $\rho = 0.135$ (dashed), $\rho = 0.368$ (dotted) and $\rho = 0.607$ (dot-dashed). The solid line is true derivative $f'(x)$.

Finally, our simulations yield satisfactory results of our estimation even in the presence of short range dependence and heavy-tailed distribution.

5. High-frequency valvometry data

Our goal is now to propose a way to survey the marine water quality. The use of biological indicators is of great interest for the detection of seashore pollution. These bio-indicators can also be considered as biological responses to climate change influencing water-quality and biological organisms. The main challenge is to determine how bio-indicators are affected by pollution and climate change. For that purpose, we study bivalves activities using a high-frequency noninvasive valvometry electronic system developed by the EPOC team in Arcachon, France, see [6] and [32] as well as the website

http://molluscan-eye.epoc.u-bordeaux1.fr.

This electronic system records valve movements autonomously for a long period of time (at least one full year). Each animal is equipped with two light coils (sensors), of approximately 53mg each (unembedded), fixed on the edge of each valve. One of the coils emits a high-frequency, sinusoidal signal which is received by the other coil. The strength of the electric field produced between the two coils being proportional to the inverse of the distance between the point of measurement and the center of the transmitting coil, the distance between coils can be measured and the accuracy of the measurements is a few μm.

For each sixteen animals, one measurement is received every 0.1s (10 Hz). So, the activity of animals is measured every 1.6s and each day, we obtain 864 000 triplets of data: the times of the measurement, the distances between the two valves and the animal numbers. A first electronic card in a waterproof case next to the animals manages the electrodes and a second electronic card handles the data acquisition. The valvometry system uses a GSM/GPRS modem and a Linux operating system for the data storage, the internet access, and the data transmission. After each 24-hour period or any other programmed period of time, the data are transmitted to a workstation server and then inserted into a SQL database which is accessible with the software R or a text terminal.

Several valvometric systems have been installed around the world: southern lagoon of New Caledonia (South Pacific), Spain, Ny Alesund Svalbard at 1300 km from the north pole, the north east of Murmansk in Russia on the Barents sea and at several sites in France with various species. We focus our attention on the IORO reef in the Havannah channel (GPS coordinates 166°57′25″ E, 22°23′15″ S) in the southern lagoon of New Caledonia. The New Caledonia climatic regime is characterized by two main periods, an austral rainy summer (December-March) and a drier austral winter (May-October), each separated by transition phases. The water temperature closely follows air temperature changes. We study the opening/closing velocity of sixteen giant clams *Hippopus hippopus* placed in a single bag.

As argued in [11, 12], environmental perturbations such as a pollution can affect the activity of biosensors and in particular the shells opening and closing velocities. For instance, we observed that a stressed animals due to the presence of pollution or environmental perturbations exhibits irregular and numerous micro-closing and opening periods with changes in the velocities in comparison with the normal situation. Consequently, the movement velocities can be considered as an indicator of the animal reaction activity since its movements are associated with aquatic system perturbations.

Different statistical methods have been developed to analyse these high-frequency data [3, 7, 19, 28–30] among others and a nonparametric adaptive estimator was proposed for extreme tail probabilities and quantiles [11, 12]. In [3], a four-state stochastic process was considered to give inferences about oysters health and to provide some arguments about the healthiness of their environment. The authors especially delineate links between groups of oysters and features related to the survey of the environment in different experimental sites, such as environmental variations. These methods also exhibit a link between the tide and oysters behavior, as shown in [30]. In [7], a data-driven bandwidth choice for a kernel density estimator is investigated assuming the number of modes to be known and this methodology is illustrated using valvometry data. When considering these data, important intermittent activity at high frequency, with frequent and sudden "microclosing" events (meaning partial closures), at apparently random times and with random amplitudes is observed [29]. A fractal shot noise modeling is then proposed for quantifying and characterizing the behaviors of bioindicators directly exposed to their natural environment and exposed to changes in the water quality of their natural environment. Using valvometry data, a nonparametric quantile regression model was used to model *in situ* with accuracy giant clam growth rate behavior in relation to temperature [28] showing that the shell growth was significantly correlated to rising sea surface temperature. As the measurements were performed at 10 Hz for one among the sixteen animals, the developed platform for valvometry are able to measure the position of the opening of a mollusk shell. So, this system allows the bivalves to be studied in their natural environment with minimal experimental constraints with accuracy and the velocity measurements are enough accurate for our velocity estimations. An example of valves activity and opening/closing velocity recordings is depicted in Figure 6 for the 8th of July 2008. Because the data representation in Figures 6 and 7 are "noisy", we do not use the finite difference method to determine the velocity. We use instead the smooth nonparametric estimation of the derivative of the regression function described in Section 2.

We do not consider the fastest opening and closing velocities because the measurements every 1.6s do not allow us to estimate them with accuracy. Indeed, if an animal quickly closes or opens its shells, the 1.6 second gap between two measurements is too important to determine the velocity. To improve the higher velocities estimations, it is possible for a new study to use higher frequency data acquisition (for instance at 100 Hz). Here, we focus our attention on smaller velocities. Figure 7 shows the very good fit of fixed-design regression derivative $\widehat{f}_n'(x)$ to the observed velocities. We have chosen to restrict ourselves to opening/closing velocities smaller than 0.3 millimeters per second as higher velocities can be seen as statistically insignificant outliers.

In order to study the behavior of these bio-indicators with respect to the sea water temperature, we compare for the 16 bivalves the closing and opening velocities on two periods corresponding to the warmest period (from 10th of March to 2th of April 2008) and to the coldest period (July 2008). We focus here our attention on the opening velocities since the same trends are observed for the closing velocities.

There is some type of autocorrelation in the residuals confirmed by Box-Pierce, Ljung-Box and Durbin-Watson tests (p-value < 0.05). By examining the representation of the autocorrelation function given in Figure 8, the very first line (to the left) shows the correlation of residual with itself (lag 0) which is equal to 1. We observe also a "spike" at lag 1 followed by non significant values for lags past 1. As the autocorrelation function decreases at a geometric rate, we observed a short range dependence.

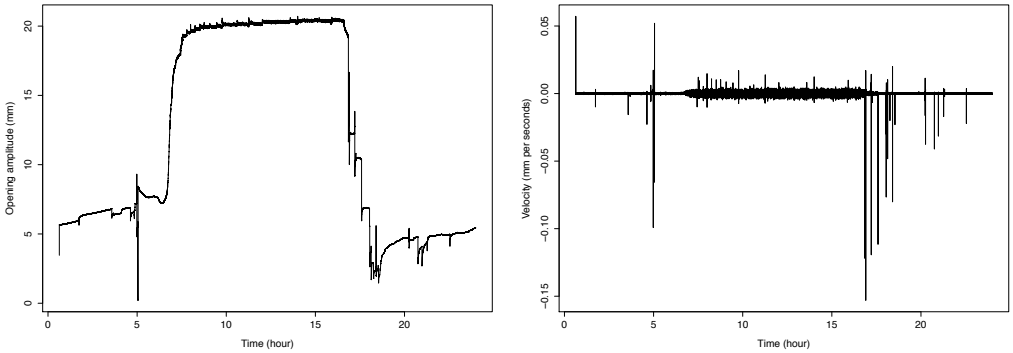

Figure 6.: A typical example of valvometric data for one giant clam the 8th of July 2008. On the left-hand side, relationship between the opening amplitude and the time of the experiment. On the right-hand side, the closing and opening velocity with respect to time.

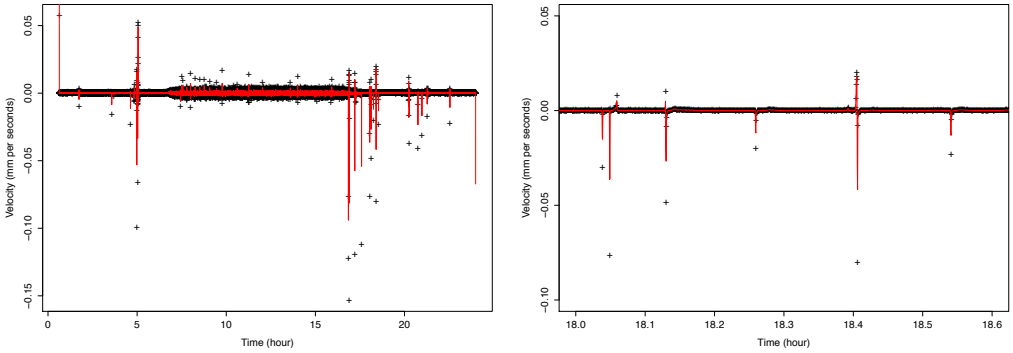

Figure 7.: On the left-hand side, the solid line displays for the 8th of July 2008, the estimated $f'(x)$ using estimator $\widehat{f}'_n(x)$ versus the time x and the points represent the observed velocities of valve openings and closings. A zoom between 6 p.m. and 6.6 p.m. is given on the right-hand side.

Due to the satisfactory results obtained in Section 4, we have chosen to restrict our-

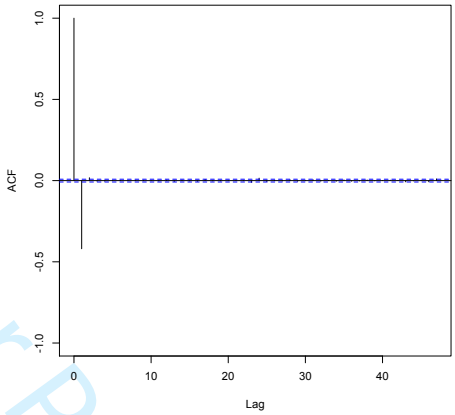

Figure 8.: ACF of the residuals of the model (1).

selves to the case of independent data and to favor our statistical application on the detection of marine pollution. Figures 9 and 10 reveal that the valve opening velocity distributions for the two periods are significantly different (pvalue < 0.0001 using a Kolmogorov-Smirnov test).

Figure 9 is not related to the central limit Theorem 3.2. This Figure represents the histogram of the opening velocities estimation given in (5) for the 16 different giant clams on the two selected periods. Figure 9 reveals that the valve velocity distributions for the two periods are significantly different (pvalue < 0.0001 using a Kolmogorov-Smirnov test). The Quantile-Quantile plot representation given in Figure 10 confirms graphically this result. At the highest temperatures for which it was reported that the animals were disturbed and at the limits of their thermal preference [28], the velocity distribution is different from that at the lowest temperatures. We observe on the left side of Figure 9 that the adductor muscle has a peak velocity around 0.07 mm/sec which does not exist for the coldest period on the right side of the Figure. At the warmest sea water temperature period (28°C in April 2008), the velocity distribution is clearly significantly different than at the coldest sea water temperature period (23°C in July 2008).

In conclusion, in the present context of globally increasing Sea Surface Temperature associated to climate change, our estimation indicates changes in the velocity distributions and so we are able to detect physiological impact of global warming and perturbations due to pollution. With tropical reefs around the world threatened by warming oceans, most research is focused on corals and fishes. Here, we show the effect of environmental conditions on giant clams and we suggest that the giant clam *Hippopus hippopus* could be an interesting sentinel species. The combination of non-parametric statistical procedure with high-frequency valvometry data provides a new way for studying the behavior of bio-indicators. Moreover, our approach highlights the idea that high-frequency noninvasive valvometry is of great interest for the study of

marine pollution or climatic change. The present work researches the effect of environmental conditions on giant clams in New Caledonia, focusing on a particular species that we believe to be more amenable than others to an online analysis of behavioral activity.

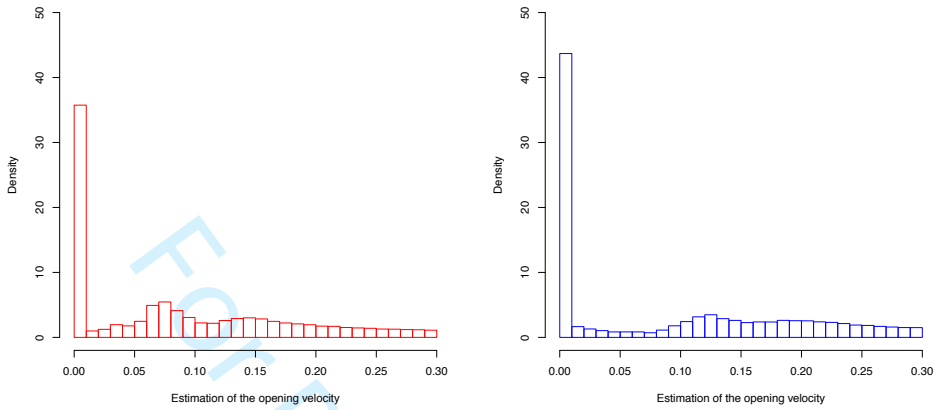

Figure 9.: Histograms of the derivative estimators $\widehat{f}_n'(x)$: On the left-hand side in red for the warmest period and on the right-hand side in blue for the coldest period in New Caledonia.

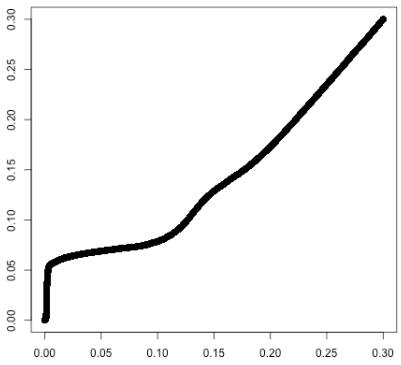

Figure 10.: Quantile-Quantile plot of the derivative estimator $\widehat{f}_n'(x)$ during the two selected periods: in abscissa the coldest period and in ordinate the warmest period in New Caledonia.

Acknowledgments. The authors would like to thank the editor and the two anonymous reviewers for their helpful comments and constructive suggestions which helped to improve the paper substantially.

References

[1] A. Amiri, *Recursive regression estimators with application to nonparametric prediction*, Journal of Nonparametric Statistics 24 (2012), pp. 169–186.

[2] P. Ango Nze, P. Bühlmann, and P. Doukhan, *Weak dependence beyond mixing and asymptotics for nonparametric regression*, Annals of Statistics 30 (2002), pp. 397–430.

[3] R. Azaïs, G. Coudret, and G. Durrieu, *A hidden renewal model for monitoring aquatic systems biosensors*, Environmetrics 25 (2014), pp. 189–199.

[4] B. Bercu, S. Capderou, and G. Durrieu, *Nonparametric recursive estimation of the derivative of the regression function with application to sea shores water quality*, Statistical Inference for Stochastic Processes 21 (2018), pp. 1-24.

[5] B. Bercu, B. Delyon, and E. Rio, *Concentration inequalities for sums and martingales*, SpringerBriefs in Mathematics, Springer, 2015.

[6] C. Chambon, A. Legeay, G. Durrieu, P. Gonzalez, P. Ciret, and J. C. Massabuaum *Influence of the parasite worm Polydora sp. on the behaviour of the oyster crassostrea gigas*, Marine Biology 152 (2007), pp. 329–338

[7] R. Coudret, G. Durrieu, and J. Saracco, *Comparison of kernel density estimators with assumption on number of modes*, Communication in Statistics - Simulation and Computation 44 (2015), pp. 196–216.

[8] L. Devroye, *Nonparametric functional estimation and related topics*, Exponential inequalities in nonparametric estimation, Springer (1991), pp. 31–44.

[9] L. Devroye and G. Lugosi, *Combinatorial methods in density estimation*, Springer Series in Statistics, Springer-Verlag, New York, 2001.

[10] F. G. Doherty, D. S. Cherry, and J. Cairns, *Valve closure responses of the asiatic clam corbicula fluminea exposed to cadmium and zinc*, Hydrobiologia 153 (1987), pp. 159–167.

[11] G. Durrieu, I. Grama, Q. K. Pham, and J. M. Tricot, *Nonparametric adaptive estimation of conditional probabilities of rare events and extreme quantiles*, Extremes 18 (2015), pp. 437–478.

[12] G. Durrieu, Q. K. Pham, A. S. Foltête, V. Maxime, I. Grama, V. Le Tilly, H. Duval, J. M. Tricot, C. B. Naceur, and O. Sire, *Dynamic extreme values modeling and monitoring by means of sea shores water quality biomarkers and valvometry*, Environmental Monitoring and Assessment 188 (2016), pp. 1–8.

[13] J. R. García-March, M. Á. S. Solsona, and A. García-Carrascosa, *Shell gaping behaviour of pinna nobilis l., 1758: circadian and circalunar rhythms revealed by in situ monitoring*, Marine Biology 153 (2008), pp. 689–698.

[14] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk, *A distribution-free theory of nonparametric regression*, Springer Series in Statistics, Springer-Verlag, New York, 2002.

[15] T. Gasser and H. G. Müller, *Estimating regression functions and their derivatives by the kernel method*, Scandinavian Journal of Statistics 111 (1984), pp. 171–185.

[16] P. Hall, *On iterated logarithm laws for linear arrays and nonparametric regression estimators*, Annals of Probability 19 (1991), pp. 740–757.

[17] B. E. Hansen, *Uniform convergence rates for kernel estimation with dependent data*, Econometric Theory 24 (2008), pp. 726–748.

[18] Y. Huang, X. Chen, and W. B. Wu, *Recursive nonparametric estimation for times series*, IEEE Transactions on Information Theory 60 (2014), pp. 1301–1312.

[19] L. J. Jou and C. M. Liao, *A dynamic artificial clam (corbicula fluminea) allows parcimony on-line measurement of waterborne metals*, Environmental Pollution 144 (2006), pp. 172–183.

[20] D. Li, and R. J. Tomkins, *The law of the logarithm for weighted sums of independent random variables*, J. Theoret. Probab. 16, (2003), pp. 519542.

[21] E. Masry, *Multivariate local polynomial regression for time series: uniform strong consistency and rates*, Journal of Time Series Analysis 17(1996), pp. 571–599.

[22] E. Masry and J. Fan, *Local polynomial estimation of regression functions for mixing processes*, Scandinavian Journal of Statistics 24 (1997), pp. 165179.

[23] C. McDiarmid, *On the method of bounded differences*, Surveys in combinatorics, London Math. Soc. Lecture Note Ser. Vol. 141, Cambridge Univ. Press, Cambridge, (1989), pp. 148–188.
[24] H. G. Müller, *On the boundary kernel method for non-parametric curve estimation near endpoints*, Scandinavian Journal of Statistics 20 (1993), pp. 313–328.
[25] E. A. Nadaraya, *On non-parametric estimates of density functions and regression curves*, Theory of Probability and Its Applications 10 (1965), pp. 186–190.
[26] E. A. Nadaraya, *On estimating regression*, Theory of Probability and Its Applications 9 (1964), pp. 141–142.
[27] H. Riisgard, J. Lassen, and C. Kittner, *Valve-gape response times in mussels (mytilus edulis): effects of laboratory preceding feeding conditions and in situ tidally induced variation in phytoplankton biomass*, Journal of shellfish researchn 25 (2006), pp. 901–911.
[28] C. Schwartzmann, G. Durrieu, M. Sow, P. Ciret, C. E. Lazareth, and J. C. Massabuau, *In situ giant clam growth rate behavior in relation to temperature*, Limnology and oceanography 56 (2011), pp. 1940–1951.
[29] F. G. Schmitt, M. De Rosa, G. Durrieu, M. Sow, P. Ciret, D. Tran, et al. (2011). *Statistical analysis of bivalve high frequency microclosing behavior: scaling properties and shot noise modeling*, International Journal of Bifurcation and Chaos 21(12) (2006), pp. 3565–3576.
[30] M. Sow, G. Durrieu, and L. Briollais, *Water quality assessment by means of HFNI valvometry and high-frequency data modeling*, Environmental Monitoring and Assessment 182 (2011), pp. 155–170.
[31] S. H. Sung, *A law of the single logarithm for weighted sums of i.i.d. random elements*, Statist. Probab. Lett. 79 (2009), pp. 1351–1357.
[32] D. Tran, P. Ciret, A. Ciutat, and G. Durrieu, *Estimation of potential and limits of bivalve closure response to detect contaminants: application to cadmium*, Environmental Toxicology and Chemistry 22 (2003), pp. 914–920.
[33] A. B. Tsybakov, *Introduction to nonparametric estimation*, Springer Series in Statistics, Springer-Verlag, New York, 2009.
[34] J. A. Vilar-Fernandez and J. M. Vilar-Fernandez, *Recursive estimation of regression functions by local polynomial fitting*, Annals of the Institute of Statistical Mathematics 50 (1998), pp. 729–754.
[35] J. A. Vilar-Fernandez and J. M. Vilar-Fernandez, *Recursive local polynomial regression under dependence conditions*, Test 9 (2000), pp. 209–232.
[36] G. S. Watson, *Smooth regression analysis*, Sankhya: The Indian Journal of Statistics, Series A, 26 (1964), pp. 359–372.

Appendix A. Proof of the almost sure convergence

In order to prove the pointwise almost sure convergence (6), we shall make use of the decomposition

$$\widehat{f}_n'(x) = \frac{1}{h_n^2}\left(A_n(x) + M_n(x)\right) \tag{A.1}$$

where $A_n(x)$ and $M_n(x)$ are given, for all x in $[0,1]$, by

$$A_n(x) = \frac{1}{n}\sum_{k=1}^{n} f(t_k)K'\left(\frac{x-t_k}{h_n}\right), \tag{A.2}$$

$$M_n(x) = \frac{1}{n}\sum_{k=1}^{n} \varepsilon_k K'\left(\frac{x-t_k}{h_n}\right). \tag{A.3}$$

First of all, we focus our attention on the asymptotic behavior of $A_n(x)$. By a Riemann sum approximation argument, we claim that for all x in $]0,1[$,

$$\lim_{n\to\infty} \frac{1}{h_n^2} A_n(x) = f'(x). \tag{A.4}$$

In order to prove convergence (A.4) denote, for any x and y in $[0,1]$,

$$a_n(x,y) = f(y)K'\left(\frac{x-y}{h_n}\right).$$

Since $t_k - t_{k-1} = 1/n$, we clearly have

$$A_n(x) = \sum_{k=1}^{n} (t_k - t_{k-1})\ a_n(x,t_k)$$

where $t_0 = 0$. One can easily notice that

$$A_n(x) - \int_0^1 a_n(x,y)dy = \sum_{k=1}^{n} \int_{t_{k-1}}^{t_k} \big(a_n(x,t_k) - a_n(x,y)\big)\mathrm{d}y$$

which ensures that

$$\left|A_n(x) - \int_0^1 a_n(x,y)dy\right| \leqslant \sum_{k=1}^{n} \int_{t_{k-1}}^{t_k} \big|a_n(x,t_k) - a_n(x,y)\big|\mathrm{d}y. \tag{A.5}$$

However, the function a_n is Lipschitz in the second variable. More precisely, for any fixed x in $[0,1]$, there exists some positive constant L such that, for all y and z in $[0,1]$,

$$\big|a_n(x,y) - a_n(x,z)\big| \leqslant \frac{L}{h_n}|y-z|. \tag{A.6}$$

As a matter of fact, f is a differentiable function on $[0,1]$ with bounded derivative, which implies that f is Lipschitz. By the same token, K' is also Lipschitz. Since f and K' are two bounded functions, the product a_n is Lipschitz in the second variable. Moreover, if K is a kernel with compact support, K' is also compactly supported. It means that, for any $x \neq y$ in $[0,1]$, and for n large enough, $a_n(x,y)$ vanishes. Hence, the Lipschitz constant for a_n does not depend on n. Furthermore, if K is the Gaussian kernel, it is well-known that

$$\sup_{t\in\mathbb{R}}\big|K'(t)\big| = \frac{1}{\sqrt{2\pi e}} \qquad \text{and} \qquad \sup_{t\in\mathbb{R}}\big|K''(t)\big| = \frac{1}{\sqrt{2\pi}}.$$

Consequently, for any all y and z in $[0,1]$,

$$\left|K'\left(\frac{x-y}{h_n}\right) - K'\left(\frac{x-z}{h_n}\right)\right| \leqslant \frac{1}{\sqrt{2\pi}}\frac{|y-z|}{h_n}$$

leading to

$$\left|a_n(x,y) - a_n(x,z)\right| \leqslant \frac{L_f|y-z|}{\sqrt{2\pi e}} + \frac{M_f}{\sqrt{2\pi}}\frac{|y-z|}{h_n} \tag{A.7}$$

where L_f is the Lipschitz constant for f and

$$M_f = \sup_{t\in\mathbb{R}} |f(t)|.$$

Therefore, (A.6) immediately follows from (A.7). Hereafter, we obtain from (A.5) and (A.6) that, for all x in $[0,1]$,

$$\begin{aligned}\left|A_n(x) - \int_0^1 a_n(x,y)dy\right| &\leqslant \frac{L}{h_n}\sum_{k=1}^{n}\int_{t_{k-1}}^{t_k}|t_k - y|dy,\\ &= \frac{L}{h_n}\sum_{k=1}^{n}\int_{t_{k-1}}^{t_k}(t_k - y)dy,\\ &= \frac{L}{2h_n}\sum_{k=1}^{n}(t_k - t_{k-1})^2,\end{aligned}$$

which implies that

$$\left|\frac{1}{h_n^2}A_n(x) - \frac{1}{h_n^2}B_n(x)\right| \leqslant \frac{L}{2nh_n^3} \quad \text{where} \quad B_n(x) = \int_0^1 a_n(x,y)dy. \tag{A.8}$$

Hence, as nh_n^3 goes to infinity as soon as $\alpha < 1/3$, the two sequences are equivalent. It remains to carefully investigate the asymptotic behavior of $B_n(x)$. The regression function f is bounded continuous and twice differentiable with bounded derivatives. Consequently, it follows from Taylor's formula with integral remainder that, for all x in $[0,1]$ and for any $t \in \mathbb{R}$,

$$f(x - h_n t) = f(x) - h_n t f'(x) + \int_{x-h_n t}^{x}\big(s - x + h_n t\big)f''(s)ds.$$

Therefore, $B_n(x)$ can be rewritten as

$$\begin{aligned}B_n(x) &= \int_0^1 f(y)K'\left(\frac{x-y}{h_n}\right)dy = h_n\int_{(x-1)\ell_n}^{x\ell_n} f(x - h_n t)K'(t)dt,\\ &= h_n f(x)\int_{(x-1)\ell_n}^{x\ell_n} K'(t)dt - h_n^2 f'(x)\int_{(x-1)\ell_n}^{x\ell_n} tK'(t)dt + R_n(x)\end{aligned} \tag{A.9}$$

where $\ell_n = 1/h_n$ and the remainder $R_n(x)$ is given by

$$R_n(x) = h_n\int_{(x-1)\ell_n}^{x\ell_n}\int_{x-h_n t}^{x}\big(s - x + h_n t\big)f''(s)K'(t)dsdt.$$

On the one hand, as the bandwidth h_n goes to zero, ℓ_n tends to infinity. Hence, for any x in $]0,1[$, $x\ell_n$ goes to $+\infty$, while $(x-1)\ell_n$ goes to $-\infty$. Consequently, we obtain

that

$$\lim_{n\to\infty}\int_{(x-1)\ell_n}^{x\ell_n} tK'(t)dt = \int_{\mathbb{R}} tK'(t)dt = -1. \tag{A.10}$$

On the other hand, we also have

$$\int_{(x-1)\ell_n}^{x\ell_n} K'(t)dt = K\big(x\,\ell_n\big) - K\big((x-1)\ell_n\big).$$

If K is a kernel with compact support, then for n large enough, this integral vanishes. In addition, if K is the Gaussian kernel, $K\big(x\,\ell_n\big) = o(h_n)$ and $K\big((x-1)\ell_n\big) = o(h_n)$. It ensures that

$$\int_{(x-1)\ell_n}^{x\ell_n} K'(t)dt = o(h_n). \tag{A.11}$$

Furthermore, if we denote

$$\zeta_f = \sup_{t\in\mathbb{R}} |f''(t)|,$$

the remainder $R_n(x)$ satisfies

$$|R_n(x)| \leqslant h_n\zeta_f \int_{(x-1)\ell_n}^{x\ell_n} \left(\int_{x-h_nt}^{x} \big(s - x + h_nt\big)ds\right) |K'(t)|dt,$$

which implies that

$$|R_n(x)| \leqslant \frac{h_n^3\zeta_f}{2}\int_{(x-1)\ell_n}^{x\ell_n} t^2|K'(t)|dt \leqslant \frac{h_n^3\zeta_f}{2}\int_{\mathbb{R}} t^2|K'(t)|dt.$$

Consequently,

$$\sup_{x\in[0,1]} |R_n(x)| = O(h_n^3). \tag{A.12}$$

Hence, we obtain from (A.9), (A.10), (A.11) and (A.12) that for all x in $]0,1[$,

$$\lim_{n\to\infty} \frac{1}{h_n^2} B_n(x) = f'(x). \tag{A.13}$$

Therefore, we deduce from (A.8) and (A.13) that for all x in $]0,1[$,

$$\lim_{n\to\infty} \frac{1}{h_n^2} A_n(x) = f'(x) \tag{A.14}$$

which is exactly what we wanted to prove. Hereafter, we focus our attention on the asymptotic behavior of $M_n(x)$ given in the right-hand side of (A.1). Our goal is to

show that

$$M_n(x) = o(h_n^2) \qquad \text{a.s.} \tag{A.15}$$

Using the same arguments as those in Section 4 of [16], we obtain from the law of iterated logarithm for weigthed sums of independent random variables that

$$\limsup_{n\to\infty} \frac{n\big|M_n(x)\big|}{\sqrt{2\Sigma_n(x)\log_2 \Sigma_n(x)}} = \sigma \qquad \text{a.s.} \tag{A.16}$$

where

$$\Sigma_n(x) = \sum_{k=1}^{n} \Big(K'\Big(\frac{x - t_k}{h_n}\Big)\Big)^2.$$

However, via the same reasoning as in (A.4),

$$\lim_{n\to\infty} \frac{\Sigma_n(x)}{nh_n} = \int_{\mathbb{R}} \big(K'(y)\big)^2\, dy = \xi^2. \tag{A.17}$$

Consequently, it follows from (A.16) and (A.17) that for n large enough

$$\big|M_n(x)\big| \leqslant 2\sigma\sqrt{\frac{h_n \log_2(nh_n)}{n}} \qquad \text{a.s.} \tag{A.18}$$

which clearly implies (A.15) as soon as $\alpha < 1/3$. One can observe that we need the more restrictive assumption $\alpha < 1/4$ if we use of the law of the single logarithm for weighted sums of independent random variables given by Li and Tomkins [20], [31]. Finally, we deduce from the decomposition (A.1) together with (A.4) and (A.15) that

$$\lim_{n\to\infty} \widehat{f}_n'(x) = f'(x) \qquad \text{a.s.}$$

which completes the proof of Theorem 3.1.

Appendix B. Proof of the asymptotic normality

We are now in the position to establish the pointwise asymptotic normality (7). It follows from (3) and (5) that

$$\sqrt{nh_n^3}\,\Big(\widehat{f}_n'(x) - \mathbb{E}[\widehat{f}_n'(x)]\Big) = C_n(x) \qquad \text{where} \qquad C_n(x) = \sum_{k=1}^{n} c_n(x, t_k)\varepsilon_k \tag{B.1}$$

and, for any x and y in $]0, 1[$, $c_n(x, y)$ stands for

$$c_n(x, y) = \frac{1}{\sqrt{nh_n}} K'\left(\frac{x - y}{h_n}\right).$$

Since (ε_n) is a sequence of independent random variables with mean zero and variance σ^2, we clearly have

$$\mathrm{Var}(C_n(x)) = \sigma^2 \sum_{k=1}^{n} (c_n(x,t_k))^2 = \frac{\sigma^2}{nh_n}\Sigma_n(x).$$

Hence, we obtain from (A.17) that

$$\lim_{n\to\infty} \mathrm{Var}(C_n(x)) = \sigma^2\xi^2. \tag{B.2}$$

Consequently, in order to apply Lindeberg-Feller's central limit theorem to the sequence $(C_n(x))$, it is only necessary to check that

$$\max_{1\leqslant k\leqslant n} (c_n(x,t_k))^2 = o\big(\mathrm{Var}(C_n(x))\big). \tag{B.3}$$

The kernel K is bounded, twice differentiable with bounded derivatives. Hence, there exists some positive constant M_K such that, for any x in $]0,1[$,

$$\max_{1\leqslant k\leqslant n} (c_n(x,t_k))^2 \leqslant \frac{M_K^2}{nh_n}.$$

It clearly implies (B.3) as nh_n goes to infinity as soon as n does. Therefore, (B.1) together with (B.2) leads to

$$\sqrt{nh_n^3}\big(\widehat{f}_n'(x) - \mathbb{E}[\widehat{f}_n'(x)]\big) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma^2\xi^2\right) \tag{B.4}$$

which is exactly convergence (7). It now remains to prove (8) replacing $\mathbb{E}[\widehat{f}_n'(x)]$ by $f'(x)$ in (B.4). We clearly have for all x in $]0,1[$

$$\sqrt{nh_n^3}\big(\widehat{f}_n'(x) - f'(x)\big) = \sqrt{nh_n^3}\big(\widehat{f}_n'(x) - \mathbb{E}[\widehat{f}_n'(x)]\big) + D_n(x)$$

where

$$D_n(x) = \sqrt{nh_n^3}\big(\mathbb{E}[\widehat{f}_n'(x)] - f'(x)\big).$$

Consequently, in order to prove (8), we only have to show that $D_n(x)$ goes to zero as n tends to infinity. We already saw in Appendix A that

$$\mathbb{E}[\widehat{f}_n'(x)] - f'(x) = \frac{1}{h_n^2}A_n(x) - f'(x) = \frac{1}{h_n^2}B_n(x) - f'(x) + \Delta_n(x) \tag{B.5}$$

where

$$\Delta_n(x) = \frac{1}{h_n^2}\big(A_n(x) - B_n(x)\big).$$

However, it follows from (A.8) that there exists some positive constant L such that

$$\big|\Delta_n(x)\big| \leqslant \frac{L}{2nh_n^3}.$$

Hence, as soon as, $0 < \alpha < 1/3$

$$\lim_{n\to\infty} \sqrt{nh_n^3}\Delta_n(x) = 0. \tag{B.6}$$

Furthermore, we deduce from the decomposition of $B_n(x)$ given in (A.9) that

$$\frac{1}{h_n^2}B_n(x) - f'(x) = \frac{1}{h_n}f(x)P_n(x) - f'(x)Q_n(x) + \frac{1}{h_n^2}R_n(x) \tag{B.7}$$

where

$$P_n(x) = \int_{(x-1)\ell_n}^{x\ell_n} K'(t)dt \qquad \text{and} \qquad Q_n(x) = 1 + \int_{(x-1)\ell_n}^{x\ell_n} tK'(t)dt.$$

On the one hand,

$$\sqrt{nh_n^3}\frac{1}{h_n}P_n(x) = \sqrt{nh_n}\big(K(x\ell_n) - K((x-1)\ell_n)\big).$$

Hence, if K is a kernel with compact support or if K is the Gaussian kernel, we clearly have for any x in $[0,1]$,

$$\lim_{n\to\infty} \sqrt{nh_n^3}\frac{1}{h_n}P_n(x) = 0. \tag{B.8}$$

On the other hand, an integration by parts for $Q_n(x)$ leads to

$$\begin{aligned} Q_n(x) &= 1 - \int_{(x-1)\ell_n}^{x\ell_n} K(t)dt + \Big[tK(t)\Big]_{(x-1)\ell_n}^{x\ell_n}, \\ &= \int_{\mathbb{R}} K(t)dt - \int_{(x-1)\ell_n}^{x\ell_n} K(t)dt + \Big[tK(t)\Big]_{(x-1)\ell_n}^{x\ell_n}, \\ &= \int_{x\ell_n}^{+\infty} K(t)dt + \int_{-\infty}^{(x-1)\ell_n} K(t)dt + \Big[tK(t)\Big]_{(x-1)\ell_n}^{x\ell_n}. \end{aligned}$$

If K is a kernel with compact support, then for n large enough, $Q_n(x)$ vanishes. In addition, if K is the Gaussian kernel, it follows from standard Gaussian calculation that

$$\int_{x\ell_n}^{+\infty} K(t)dt = \frac{1}{x\ell_n}K(x\ell_n)\big(1 + o(1)\big)$$

and

$$\int_{-\infty}^{(x-1)\ell_n} K(t)dt = -\frac{1}{(x-1)\ell_n}K((x-1)\ell_n)\big(1 + o(1)\big).$$

It implies that for any x in $[0,1]$,

$$\lim_{n\to\infty} \sqrt{nh_n^3} Q_n(x) = 0. \tag{B.9}$$

Moreover, we immediately infer from (A.12) that, as soon as $\alpha > 1/5$

$$\lim_{n\to\infty} \sqrt{nh_n^3} \frac{1}{h_n^2} R_n(x) = 0. \tag{B.10}$$

Finally, it follows from (B.7) together with (B.8), (B.9) and (B.10) that

$$\lim_{n\to\infty} \sqrt{nh_n^3} \left(\frac{1}{h_n^2} B_n(x) - f'(x) \right) = 0$$

which ensures that $D_n(x)$ goes to zero as n tends to infinity, completing the proof of Theorem 3.2. □

Appendix C. Proof of the concentration inequalities

We shall follow the same approach as [8] who was the first to make use of McDiarmid's inequality [23] in nonparametric density estimation. For any x in $[0,1]$, denote

$$\widehat{f}_n'(x) = \widehat{f}_n'(x, Y_1, \ldots, Y_n) = \frac{1}{nh_n^2} \sum_{k=1}^{n} Y_k K' \left(\frac{x - t_k}{h_n} \right).$$

Let $Z_1, \ldots, Z_n$ be a sequence such that $Z_i = Y_i$ for all $1 \leqslant i \leqslant n$, except for $i = k$. In addition, assume that $|Z_k| \leqslant M$ a.s. We clearly have

$$\begin{aligned} \left| \widehat{f}_n'(x, Y_1, \ldots, Y_n) - \widehat{f}_n'(x, Z_1, \ldots, Z_n) \right| &= \frac{1}{nh_n^2} |Y_k - Z_k| \left| K' \left(\frac{x - t_k}{h_n} \right) \right|, \\ &\leqslant \frac{2M\Lambda}{nh_n^2} \qquad \text{a.s.} \end{aligned}$$

Consequently, (9) immediately follows from McDiarmid's inequality [23], see also [9] or [5]. By the same token,

$$\left| \int_{\mathbb{R}} \left| \widehat{f}_n'(x, Y_1, \ldots, Y_n) - f'(x) \right| dx - \int_{\mathbb{R}} \left| \widehat{f}_n'(x, Z_1, \ldots, Z_n) - f'(x) \right| dx \right|$$

$$\begin{aligned} &\leqslant \int_{\mathbb{R}} \left| \widehat{f}_n'(x, Y_1, \ldots, Y_n) - \widehat{f}_n'(x, Z_1, \ldots, Z_n) \right| dx, \\ &\leqslant \frac{1}{nh_n^2} \int_{\mathbb{R}} |Y_k - Z_k| \left| K' \left(\frac{x - t_k}{h_n} \right) \right| dx, \\ &\leqslant \frac{2M\zeta}{nh_n^2} \qquad \text{a.s.} \end{aligned}$$

Finally, we obtain (10) once again from McDiarmid's inequality [23], which completes the proof of Theorem 3.3. □