NONPARAMETRIC STATISTICS FOR BIG DATA

Gilles Durrieu (Université Bretagne Sud)
Joint work with
Bernard Bercu (Université de Bordeaux) and Sami Capderou
(University of Geneva).

CYBERUS SUMMER SCHOOL

3 to 7 July 2023 - Online

Plan

Plan

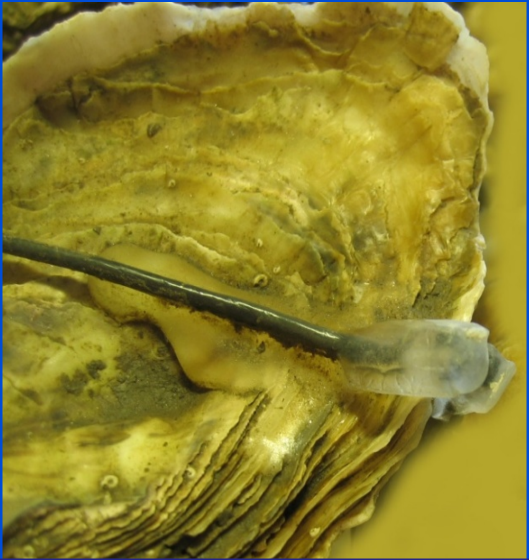
Some publications related to the subject:

- Senga-Kiesse T. and Durrieu G., *Statistics and Probability letters*, submitted.
- Bercu B., Capderou S. and Durrieu G. (2019) *Journal of Applied Statistics*, 46(1), 119-140.
- Bercu B., Capderou S., and Durrieu G. (2019) *Statistical Inference for Stochastic Processes*, 22(1), 17-40.
- Durrieu G., Grama I., Jaunatre K. and Tricot J.M. (2018) *Journal of Statistical Software*, 87,12, 1-20.
- Durrieu G., Grama I., Pham Q.K. and tricot J.M. (2015) *Extremes*, 18, 437-478.
- Durrieu G., Pham Q.K., Foltete A.S., Maxime V., Grama I., Le Tilly V., Duval H., Tricot J.M. and Sire O. (2016) *Environmental Monitoring and Assessment*, 188, 401-409.
- Durrieu G. and Briollais L. (2009) *Journal of American Statistical Association*, 104, 650-660.

Water quality and global warming effects

- Developing a procedure for monitoring the quality of water and measuring the global warming effects, based on the analysis of their behavior of bivalves at high frequency.

High frequency valvometry and Big Data



Electrodes

Electrodes

+

2 watts, Linux

Electrodes

+

2 watts, Linux

+

Solar panels

Electrodes

+

2 watts, Linux

+

Solar panels

+

$$\hat{m}_h(t) = \frac{\sum_{i=1}^{n} K\left(\frac{t-T_i}{h}\right) Y_i}{\sum_{i=1}^{n} K\left(\frac{t-T_i}{h}\right)}.$$

Statistical modeling

Electrodes

+

2 watts, Linux

+

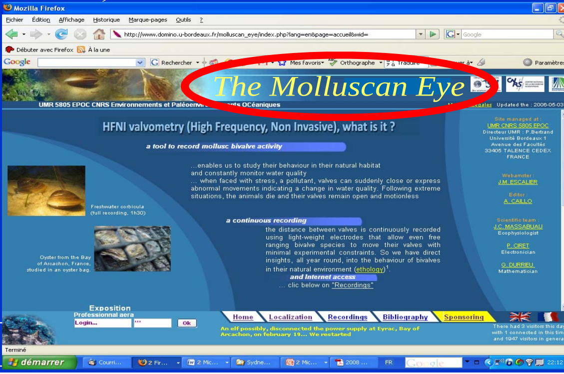
Solar panels

+

$$\hat{m}_h(t) = \frac{\sum_{i=1}^{n} K\left(\frac{t-T_i}{h}\right) Y_i}{\sum_{i=1}^{n} K\left(\frac{t-T_i}{h}\right)}.$$

Statistical modeling

+

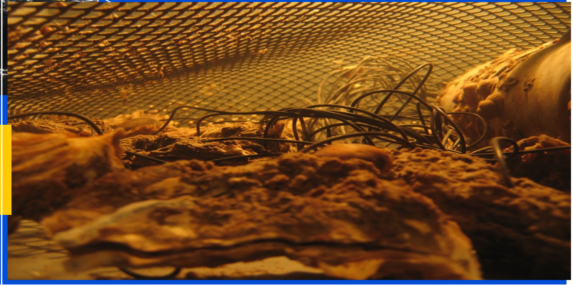


Website

Introduction

First Experimental site: LOCMARIAQUER Gulf of Morbihan - Atlantic Ocean

www.



Analyse de données et modélisation
LMBA UBS

Huîtres *Crassostrea Gigas* (n = 16)

Second experimental site: Havannah Canal in New Caledonia - Pacific Ocean

Near and far experimental sites

Different sites in the World

- France: Arcachon bay, Brest, Locmariaquer, Oléron, Lacq, New Caledonia
- Norway: Ny-Alesund (Spitzberg in Svalbard archipelago),
- Russia: Mourmansk,
- Spain: port of Santander,
- ...

Data

- Sampling frequency (10 Hz): one measurement every 0.1s, each animal is measured every 1.6 s ($N = 16$);
- 108,000 measurements for one oyster by day;
- $n = 1{,}728{,}000$ data points by day for the 16 oysters;
- 630,720,000 measurements/year

and so $12,614,400,000$ measurements/year for 20 sites

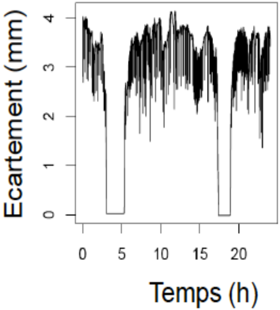
$+$

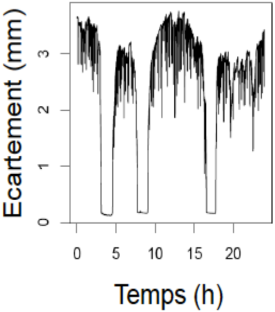
biological and environmental parameters: acetycholinesteras, EROD activity, vitellogenin, temperature, salinity, chlorophylle, mortality, animal growth.

Objectives

- Dealing with the data deluge,
- Model the animals behavior in their environment in order to detect environmental disturbances (such as pollution),
- Construction of mathematical indicators for monitoring water quality (pollution detection, climate change and global warming),
- Study the effect of global warming on a representative of marine fauna taken as a biosensor of the evolution of its environment,
- Setting up a database and automatic representation of data and results online.

Graphical representation of data



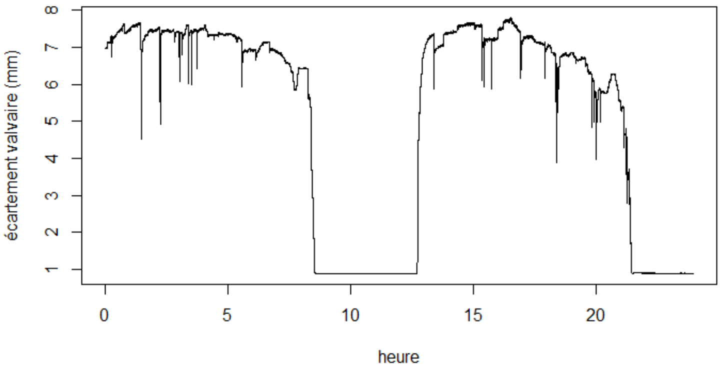


Introduction

Individu	Heure	Ouverture
11	0.0000011574	6.754
12	0.0000023148	5.436
13	0.0000034722	1.589
14	0.0000046296	6.356
15	0.0000057870	5.895
16	0.0000069444	4.754
1	0.0000081019	6.960

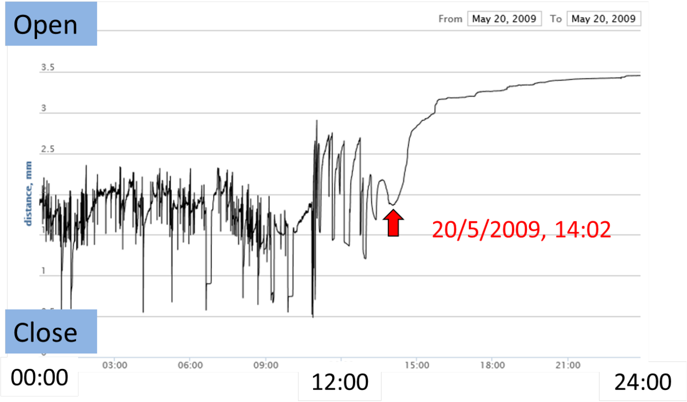
Introduction

Example of behavior



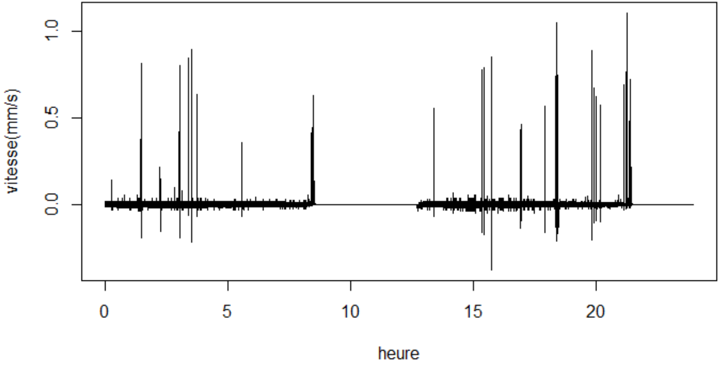
Introduction

Death



Introduction

Velocity of the valve opening/closing activity



Movement velocities as an indicator

- Environmental perturbations such as a pollution of global warming can affect the activity of biosensors and in particular the shells opening and closing velocities.
- A stressed animal due to the presence of pollution or environmental perturbations exhibits irregular and numerous microclosing and opening periods with changes in the velocities in comparison with the normal situation.

Movement velocities as an indicator

- Environmental perturbations such as a pollution of global warming can affect the activity of biosensors and in particular the shells opening and closing velocities.
- A stressed animal due to the presence of pollution or environmental perturbations exhibits irregular and numerous microclosing and opening periods with changes in the velocities in comparison with the normal situation.

Movement velocities as an indicator

- Environmental perturbations such as a pollution of global warming can affect the activity of biosensors and in particular the shells opening and closing velocities.
- A stressed animal due to the presence of pollution or environmental perturbations exhibits irregular and numerous microclosing and opening periods with changes in the velocities in comparison with the normal situation.

Plan

Random design regression

We consider the **nonparametric regression model** given, for all $n \geq 1$, by

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

where

- (X_n) (the time of the measurement) is a sequence of random variables **iid** with positive probability density function g,
- (ε_n) are unknown random errors **iid** independent of (X_n), such that $\mathbb{E}[\varepsilon_n] = 0$ et $\mathbb{E}[\varepsilon_n^2] = \sigma^2$,
- The regression function f and the density function g are unknown, bounded continuous, twice differentiable with bounded derivatives.

Objective

Estimation of the derivative f' of f.

Nadaraya-Watson estimator of f

- The **kernel** K is a positive symmetric bounded function, differentiable with bounded derivative.
- **The bandwidth** (h_n) is a sequence of positive real numbers, decreasing to zero, such that nh_n tends to infinity.

The **Nadaraya-Watson** estimator of f is given, for all $x \in \mathbb{R}$, by

$$f_n(x) = \frac{\sum_{k=1}^{n} Y_k K\left(\frac{x - X_k}{h_n}\right)}{\sum_{k=1}^{n} K\left(\frac{x - X_k}{h_n}\right)}.$$

Recursive Nadaraya-Watson estimator of f

The **recursive Nadaraya-Watson** estimator is given, for $x \in \mathbb{R}$, by

$$\widehat{f}_n(x) = \frac{\displaystyle\sum_{k=1}^{n} \frac{Y_k}{h_k} K\Big(\frac{x - X_k}{h_k}\Big)}{\displaystyle\sum_{k=1}^{n} \frac{1}{h_k} K\Big(\frac{x - X_k}{h_k}\Big)} = \frac{\widehat{h}_n(x)}{\widehat{g}_n(x)}$$

with

$$\widehat{g}_n(x) = \frac{1}{n} \sum_{k=1}^{n} \frac{1}{h_k} K\Big(\frac{x - X_k}{h_k}\Big),$$

$$\widehat{h}_n(x) = \frac{1}{n} \sum_{k=1}^{n} \frac{Y_k}{h_k} K\Big(\frac{x - X_k}{h_k}\Big).$$

Johnston and Wand-Jones alternative estimators of f

When *g* **is known**, the Johnston and Wand-Jones estimators are given, for all $x \in \mathbb{R}$, by

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

$$\breve{f}_n(x) = \frac{1}{n} \sum_{k=1}^{n} \frac{Y_k}{g(X_k)} \frac{1}{h_k} K\left(\frac{x - X_k}{h_k}\right).$$

Estimators of the closing and opening velocity

- $\widehat{f}_n(x) = \dfrac{\widehat{h}_n(x)}{\widehat{g}_n(x)}$,

- $\widetilde{f}_n(x) = \dfrac{\widehat{h}_n(x)}{g(x)}$,

- $\check{f}_n(x) = \dfrac{1}{n} \sum_{k=1}^{n} \dfrac{Y_k}{g(X_k)} \dfrac{1}{h_k} K\left(\dfrac{x - X_k}{h_k}\right)$.

Estimators of the closing and opening velocity

- $\widehat{f}'_n(x) = \dfrac{\widehat{h}'_n(x)}{\widehat{g}_n(x)} - \dfrac{\widehat{h}_n(x)\widehat{g}'_n(x)}{\widehat{g}^2_n(x)}$,

- $\widetilde{f}'_n(x) = \dfrac{\widehat{h}'_n(x)}{g(x)} - \dfrac{\widehat{h}_n(x)g'(x)}{g^2(x)}$,

- $\check{f}'_n(x) = \dfrac{1}{n}\sum_{k=1}^{n} \dfrac{Y_k}{g(X_k)} \dfrac{1}{h_k^2} K'\left(\dfrac{x - X_k}{h_k}\right)$.

Kernel assumptions

The kernel K is a positive symmetric bounded function, differentiable with bounded derivative, satisfying

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

$$\int_{\mathbb{R}} xK'(x)dx = -1, \qquad \int_{\mathbb{R}} x^2 K'(x)dx = 0,$$

$$\int_{\mathbb{R}} x^4 K(x)dx < \infty, \qquad \int_{\mathbb{R}} x^4 |K'(x)|dx < \infty.$$

Almost sure convergence

Theorem (Bercu, Capderou and Durrieu, 2019)

If $h_n = 1/n^{\alpha}$ *with* $0 < \alpha < 1/3$, *we have for any* $x \in \mathbb{R}$ *such that* $g(x) > 0$,

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

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

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

Asymptotic normality

Denote

$$\xi^2 = \int_{\mathbb{R}} \big(K'(y)\big)^2 \mathrm{d}y.$$

Theorem (Bercu, Capderou and Durrieu, 2019)

If (ε_n) *has a finite conditional moment of order* > 2 *and if* $h_n = 1/n^\alpha$ *with* $1/5 < \alpha < 1/3$, *we have for any* $x \in \mathbb{R}$ *such that* $g(x) > 0$,

$$\sqrt{nh_n^3}\big(\widehat{f}_n'(x) - f'(x)\big) \xrightarrow{\mathcal{D}} \mathcal{N}\Big(0, \frac{1}{1+3\alpha}\frac{\xi^2}{g(x)}\sigma^2\Big),$$

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

$$\sqrt{nh_n^3}\big(\check{f}_n'(x) - f'(x)\big) \xrightarrow{\mathcal{D}} \mathcal{N}\Big(0, \frac{1}{1+3\alpha}\frac{\xi^2}{g(x)}\big(\sigma^2 + f^2(x)\big)\Big).$$

Asymptotic variance

Triweight	$K(x) = \frac{35}{32}(1-x^2)^3 \mathrm{I}_{\{\vert x\vert \le 1\}}$	$\xi^2 = 3.18$
Biweight	$K(x) = \frac{15}{16}(1-x^2)^2 \mathrm{I}_{\{\vert x\vert \le 1\}}$	$\xi^2 = 2.14$
Cosine	$K(x) = \frac{\pi}{4} cos\left(\frac{\pi}{2}\, x\right) \mathrm{I}_{\{\vert x\vert \le 1\}}$	$\xi^2 = 1.52$
Epanechnikov	$K(x) = \frac{3}{4}(1-x^2) \mathrm{I}_{\{\vert x\vert \le 1\}}$	$\xi^2 = 1.5$
Gaussian	$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$	$\xi^2 = 0.14$

Conclusion: choice in the sense of the minimum asymptotic variance of the recursive Nadaraya-Watson estimator with Gaussien kernel.

Simulation

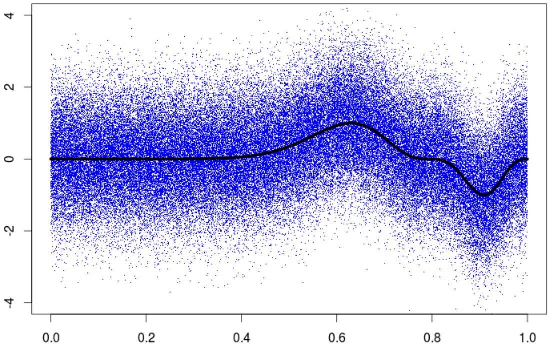
The data are generated from the nonparametric regression model for $k = 1, \ldots, n$ with $n = 10,000$

$$\boldsymbol{Y_k = f(X_k) + \varepsilon_k}$$

- The random observation (X_n) is a sequence of **iid** $\mathcal{U}([0,1])$,
- The source of variation (ε_n) is a sequence of **iid** $\mathcal{N}(0,1)$,
- The regression function f is given by

$$f(x) = \sin(2\pi x^3)^3.$$

Graphical representation

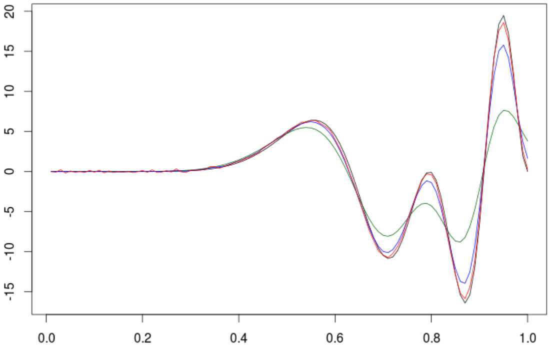


Almost sure convergence

The derivative of f is:

$$f'(x) = 18\pi x^2 \, cos(2\pi x^3)\,\sin(2\pi x^3)^2.$$

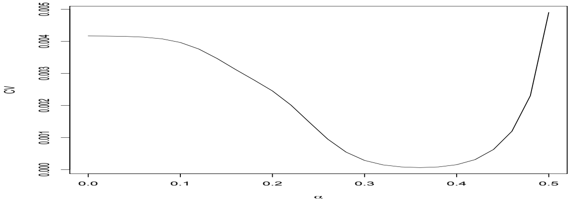
Representation of $\widehat{f'_n}$ estimator with Gaussian kernel and $\alpha = 0.1, 0.2, 0.3$.



Choice of α by cross validation method

$$CV(\alpha) = \frac{1}{n} \sum_{k=1}^{n} \left(\widehat{f'}_{(-k)}(X_k) - f'(X_k) \right)^2$$

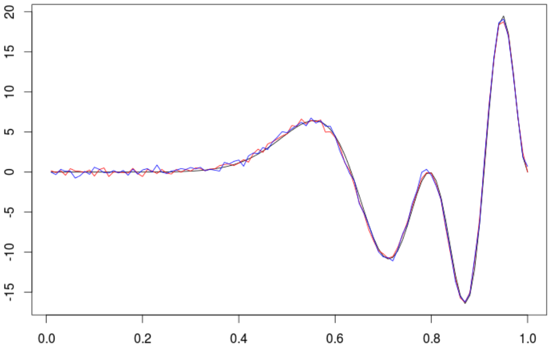
where $\widehat{f'}_{(-k)}(X_k)$ is the recursive Nadaraya-Watson estimator of $f'(X_k)$ determined with (X_k, Y_k) removed.



Choice of $\alpha_{CV} = 0.32$.

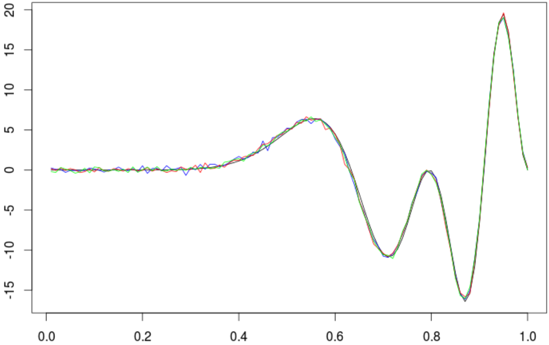
Almost sure convergence

Representation of $\widehat{f}'_n$ for Gaussian and Epanechnikov kernel



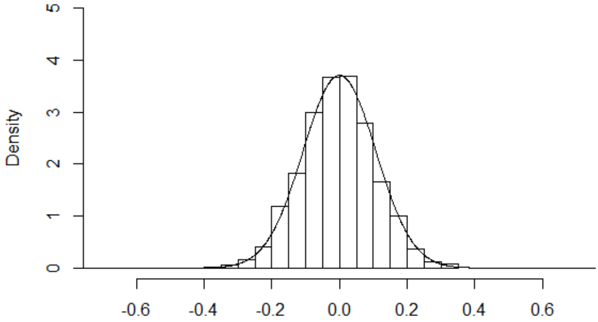
Almost sure convergence

Representation of the 3 estimators of the derivative f' of f.



Asymptotic normality

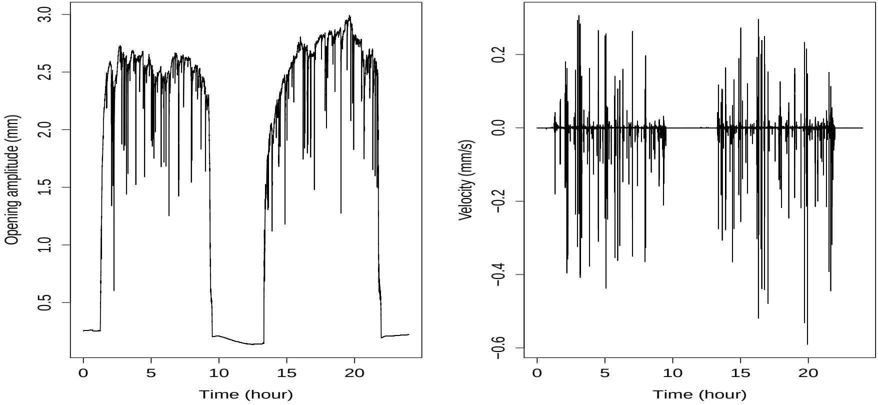
Recursive Nadaraya-Watson estimator $\widehat{f}_n'(x)$



Application in Brittany

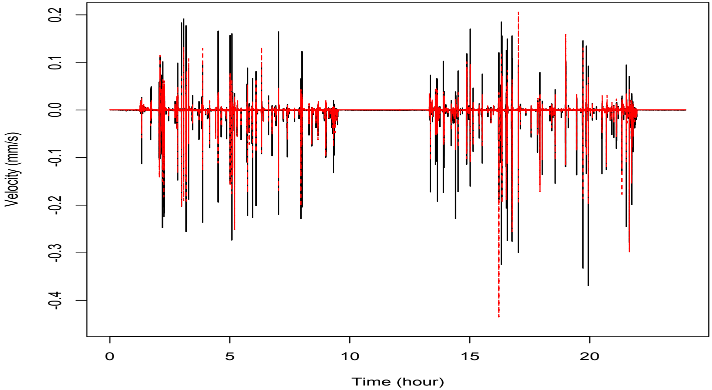
Locmariaquer site in the gulf of Morbihan.

Application in Brittany



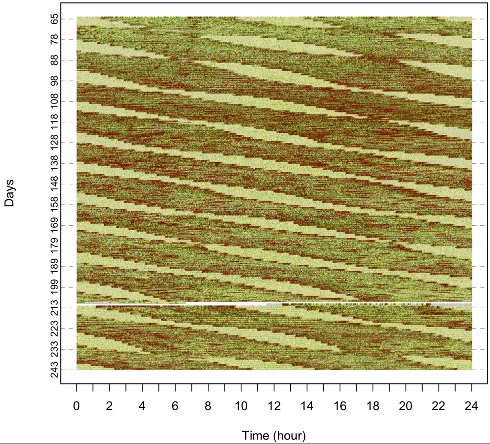
Application in Brittany

Representation of $\widehat{f}'_n(x)$ for one oyster.

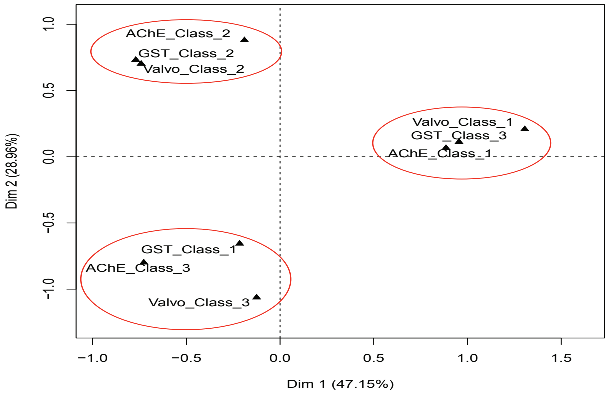


Application in Brittany

Representation of the opening and closing velocity estimator between the 4th of March and the 21th of August 2011 considering the 16 oysters in Locmariaquer.



Application in Brittany



Plan

Nonparametric fixed design regression

We consider the fixed design regression model given, for $n \geq 1$ and for $k = 1, \ldots, n$, by

$$\boldsymbol{Y_k = f(t_k) + \varepsilon_k}$$

where

- the times of measurement $\boldsymbol{t_k = k/n}$ are perfectly known,
- (ε_n) is the sequence of random error **iid** such that $\mathbb{E}[\varepsilon_n] = 0$ and $\mathbb{E}[\varepsilon_n^2] = \sigma^2$,
- the regression function f is bounded continuous, twice differentiable with bounded derivatives.

Objective

Estimation of the derivative f' of f.

Estimators

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

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

and its derivative f' by

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

Assumptions on the kernel

The kernel K is either the Gaussian kernel or a positive symmetric bounded function compactly supported, twice differentiable with bounded derivatives, such that

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

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

Almost sure convergence

Theorem (Bercu, Capderou and Durrieu, 2019)

Ifi $h_n = 1/n^{\alpha}$ *with* $0 < \alpha < 1/3$*, we have for any* $x \in]0,1[$

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

Asymptotic normality

Denote

$$\xi^2 = \int_{\mathbb{R}} (K'(y))^2 \mathrm{d}y.$$

Theorem (Bercu, Capderou and Durrieu, 2019)

We have as n tends to infinity for any $x \in]0, 1[$,

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

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

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

Concentration inequality

Denote

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

Theorem (Bercu, Capderou and Durrieu, 2019)

Assume that one can find a positive constant M such that, for all $1 \leq k \leq n$, $|\,Y_k\,| \leq M$ *a.s. Then, for any* $x \in]0,1[$ *and for any positive* $t > 0$,

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

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

Simulation

The data are generated from the regression model for $k = 1, \dots, n$ with $n = 10,000$ by

$$\boldsymbol{Y_k = f(t_k) + \varepsilon_k}$$

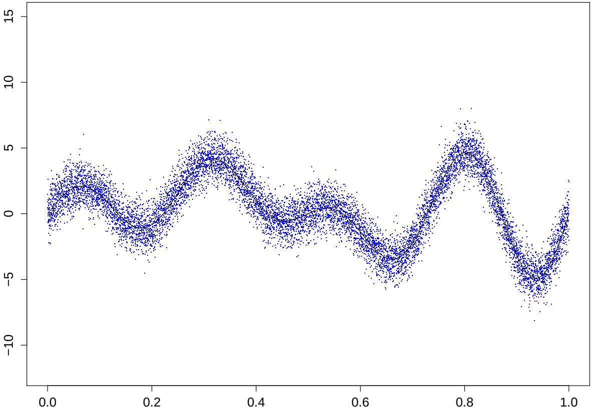
where

- The source of variation (ε_n) is a sequence of **iid** random variables $\mathcal{N}(0, 1)$,
- The regression function f is given by

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

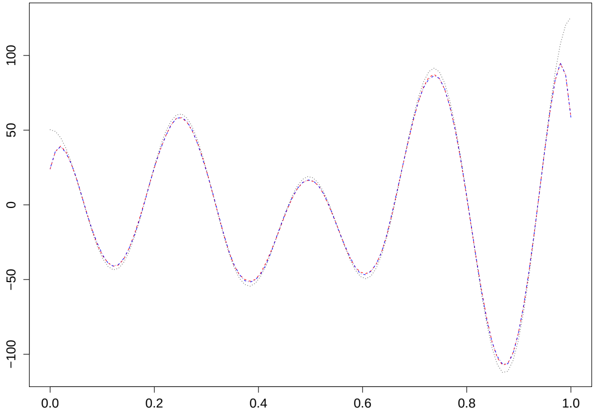
Simulation

Function f and $n = 10,000$ pairs of points (t_k, Y_k).



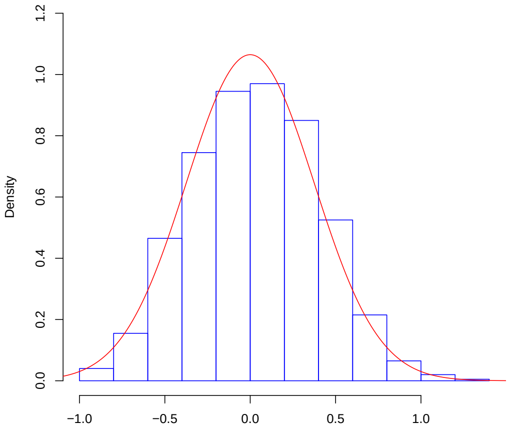
Almost sure convergence

Estimator of $\widehat{f}'_n$ using Gaussian kernel with $\alpha = 0.3$.



Asymptotic normality

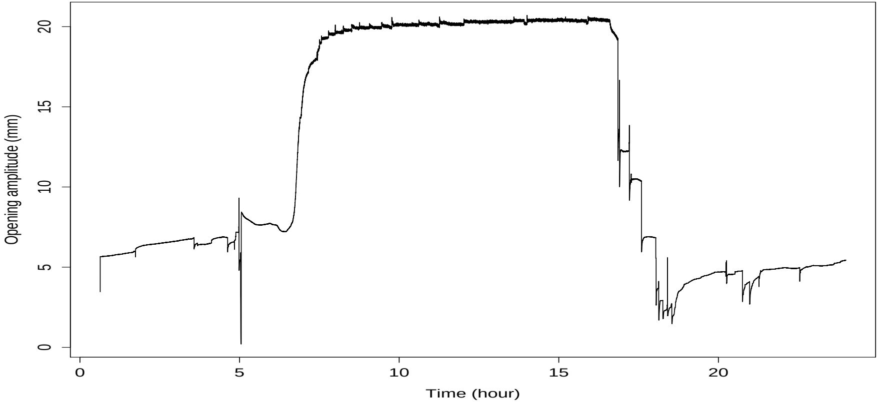
Asymptotic normality for $x = 0.7$ and $10,000$ replications.



Application in New Caledonia

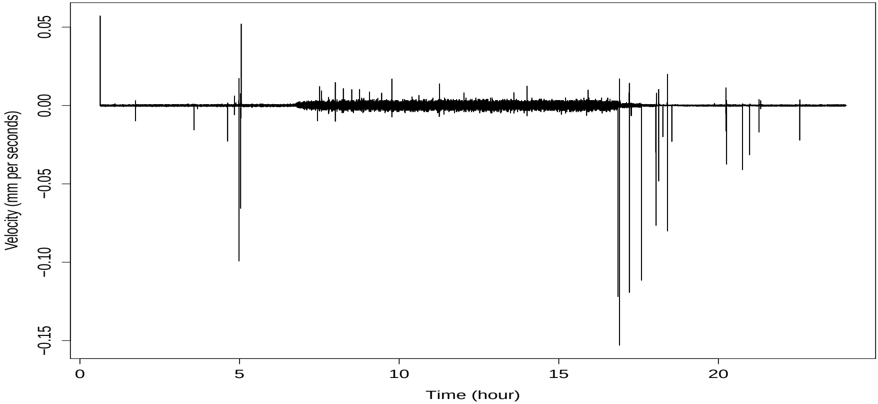
Application in New Caledonia

Representation of data for one giant clam



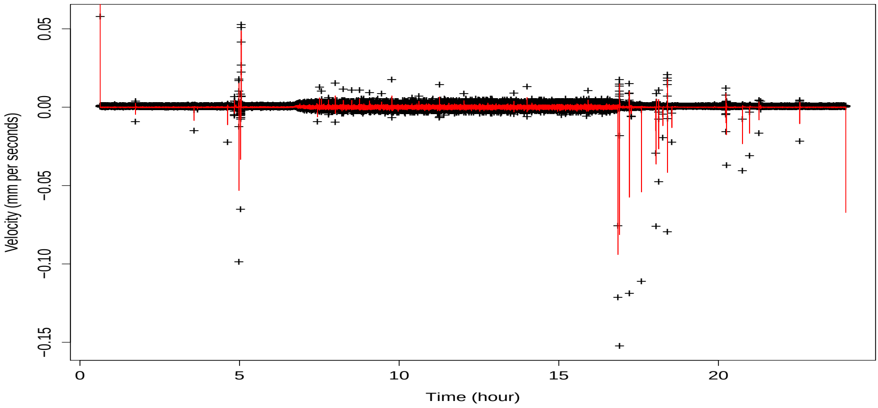
Application in New Caledonia

Representation of the opening and closing velocity of one giant clam.



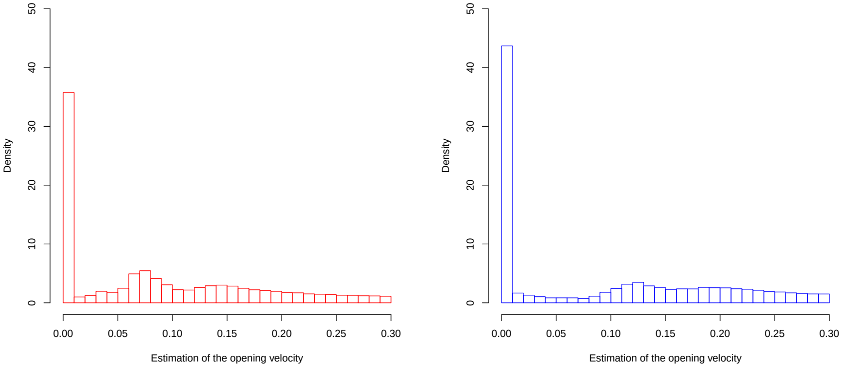
Application in New Caledonia

Representation of the velocity estimator of f'.



Application in New Caledonia

Histograms of the derivative estimators $\widehat{f}_n'(x)$: in red for the warmest period and in blue for the coldest period in New Caledonia



Plan

Conclusion

- 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 bivalves and we suggest that bivalves can be an interesting sentinel species.
- The combination of nonparametric statistical procedure with high-frequency valvometry data provides a new way for studying the behavior of bioindicators.

Multidisciplinary work with

Mathematicians, biologists and ecologists from:

- Faculté des Sciences et Sciences de l'Ingénieur (UFR SSI) of Université Bretagne Sud, Lorient and Vannes.
- l'Institut des Sciences Exactes et Appliquées of University of New Caledonia, Nouméa.
- University of Bordeaux, Bordeaux.

!!!! many thanks for your attention !!!!