Annals of the Institute of Statistical Mathematics
https://doi.org/10.1007/s10463-021-00789-0

Nonparametric regression with warped wavelets and strong mixing processes

Luz M. Gómez[1] · Rogério F. Porto[2] · Pedro A. Morettin[1]

Received: 18 October 2019 / Revised: 9 October 2020 / Accepted: 4 December 2020

Abstract

We consider the situation of a univariate nonparametric regression where either the Gaussian error or the predictor follows a stationary strong mixing stochastic process and the other term follows an independent and identically distributed sequence. Also, we estimate the regression function by expanding it in a wavelet basis and applying a hard threshold to the coefficients. Since the observations of the predictor are unequally distant from each other, we work with wavelets warped by the density of the predictor variable. This choice enables us to retain some theoretical and computational properties of wavelets. We propose a unique estimator and show that some of its properties are the same for both model specifications. Specifically, in both cases the coefficients are unbiased and their variances decay at the same rate. Also the risk of the estimator, measured by the mean integrated square error is almost minimax and its maxiset remains unaltered. Simulations and an application illustrate the similarities and differences of the proposed estimator in both situations.

Keywords Nonparametric regression · Wavelet · Stationary process · α-mixing, Warped wavelets

1 Introduction

Nonparametric regression has received much attention in recent years and there is a huge literature on the subject. In the case of time series, which is our case, see Härdle et al. (1997) for a review. Other bibliographic references are Härdle (1990), Ruppert et al. (2003), Tsybakov (2009), Silverman (1986) and Wasserman (2006).

✉ Luz M. Gómez
lgomez928@gmail.com

[1] Institute of Mathematics and Statistics, University of São Paulo, Rua do Matão, 1010, São Paulo, SP CEP 05508-090, Brazil

[2] Bank of Brazil, SAUN Quadra 5, Lote B, Ed. Green Towers, Brasília, DF CEP 70742-010, Brazil

Published online: 13 March 2021

Usually methods include local polynomials, splines, kernels, orthogonal polynomials (like Chebyshev) and more recently wavelets. The seminal works on nonparametric regression via wavelets are Donoho and Johnstone (1994, 1995) and Donoho et al. (1995).

Most of the cases consider the situation of a regular design and independent and identically distributed (IID) normally distributed errors. The case of irregular fixed design and still IID normally distributed errors was considered by Cai and Brown (1998), while the case of irregular fixed design and stationary Gaussian errors was considered by Porto (2008). The case of uniform design and IID errors was treated by Cai and Brown (1999), and the case of correlated errors was considered by Porto et al. (2016). The case of general design was considered by Kerkyacharian and Picard (2004) and Porto et al. (2012) for the cases of IID errors and correlated errors, respectively. See also Hall and Turlach (1997), Antoniadis et al. (1997), Antoniadis and Fan (2001) and Delouille et al. (2001).

In this paper, we consider the situation of a univariate nonparametric regression where either the error or the predictor follows a stationary strong mixing stochastic process and the other term follows an IID sequence. This set up typically includes regression models where both the response and predictor variables are time series. Applications of these model specifications can be found, for instance, in signal processing, econometrics and finance. We leave the situation where both the error and the predictor are stationary strong mixing for a future research because we believe this would require substantially more mathematical derivations.

Strong-mixing stochastic processes that are well-known in applications include m-dependent processes, Gaussian processes with continuous and positive spectral densities, and first-order autoregressive processes with innovation random variables following, for instance, a normal, exponential, or uniform distribution. However, in this study, we restrict the error term of the regression model to be normally distributed and possible extensions may be published elsewhere.

Also, we estimate the regression function by expanding it in a wavelet basis and applying a hard threshold to the coefficients. Since the observations of the predictor are unequally distant from each other, we work with wavelets warped by the density of the predictor variable. This choice enables us to retain some theoretical and computational properties of wavelets.

We use the estimator proposed in Kerkyacharian and Picard (2004) and show that some of its properties are the same for both model specifications. Specifically, in both cases the coefficients are unbiased and their variances decay at the same rate. Also the risk of the estimator, measured by the mean integrated square error is almost minimax and its maxiset remains unaltered.

There are many works where either the error or the predictor follows a specific dependence condition. See the references in Chesneau (2013), for instance. However, a few of them consider both specifications in a unified way. The works we consider closer to ours are the papers by Baraud et al. (2001), Chesneau (2013), Chesneau (2014, Sect. 4.3), Li (2016), and Krebs (2018). The first uses β-mixing conditions and penalized least-squares estimator, while the others use α-mixing conditions and wavelet threshold estimators, except the last paper which uses a truncated least squares wavelet (and some others) estimator. Since β-mixing implies α

-mixing Bradley (2005), the first paper covers a narrower range of dependence conditions. All these papers evaluate the risk of the estimator by the mean integrated square error and assume the mixing conditions for the vector formed by the predictor and the predicted variables, which is equivalent to assume them for the vector formed by the predictor and the error term (Baraud et al. 2001, comments in Sect. 2). This assumption includes cases where both the error and the predictor may be α-mixing and independent.

We also consider α-mixing and wavelet thresholding, but contribute to the literature by requiring a somewhat weaker condition on the mixing coefficient, dealing only with Gaussian errors, and considering a wider range of risk functions. The price we pay is not to consider both the error and the predictor α-mixing.

The paper is organized as follows. After this introduction, Sect. 2 briefly exposes some necessary background concepts and definitions. In the sequence, Sect. 3 specifies the proposed model and the estimation method together with the assumptions needed to achive the desired theoretical results. In this section we also present a simple parametric model in order to give some intuition on the results. Next, in Sect. 4, we present the theoretical results. Simulations and an application, respectively, given in Sects. 5 and 6, illustrate the similarities and differences of the proposed estimator in both situations in practice. In Sect. 7 we collect some conclusions. Supplementary Material contain the proofs and additional simulation and application results.

2 Background

2.1 Wavelets and warped wavelets

Wavelets are functions localized in time and scale, which makes them ideal to analyze functions with discontinuities and fractal structure.

Consider an orthonormal wavelet basis generated from dilation and translation of a "father" wavelet ϕ (or scaling function) and a " mother" wavelet ψ. Let $N \in \mathbb{N}$, where $\mathbb{N}$ denotes the set of the natural numbers. We assume that both functions are compactly supported in $[0, N]$ and $[(1-N)/2, (1+N)/2]$ respectively, $\int_0^N \phi = 1$, $\int_0^N \psi = 0$ and ψ has $r \in \mathbb{N}$ vanishing moments. Let $j, k \in \mathbb{Z}$, where $\mathbb{Z}$ denotes the set of the integer numbers, and let

$$\phi_{j,k}(x) = 2^{j/2}\phi(2^j x - k) \quad \text{and} \quad \psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$$

so that $\psi_{j,k}$ has support $[2^{-j}((1-N)/2 + k), 2^{-j}((1+N)/2 + k)]$. For $x \in [a, b]$, with $a, b \in \mathbb{R}$, where $\mathbb{R}$ denotes the set of real numbers, let

$$\phi^p_{j,k}(x) = \sum_{l=-\infty}^{\infty} \phi_{j,k}(x-l) \quad \text{and} \quad \psi^p_{j,k}(x) = \sum_{l=-\infty}^{\infty} \psi_{j,k}(x-l)$$

denote the periodized wavelets, which we use henceforth, but with the superscript "p" suppressed, since it is a standard way of handling boundary conditions even if the signal is not regarded as periodic (see e.g. Ogden (1997), for details). Denote the scaling function $\phi_{0,k}$ by $\psi_{-1,k}$, such that the collection formed by $\{\psi_{j,k},\ j \geq -1,\ k = 0, \dots, 2^j - 1\}$, constitutes an orthonormal basis of $L_2[a, b]$, the space of square-integrable functions in $[a, b]$ (see e.g. Härdle et al. (1998), for details). Hereafter, we will only work in the set of real numbers, otherwise stated.

Denote the inner product by $\langle\cdot,\cdot\rangle$. For a given square-integrable function f on $[a, b]$, let

$$\beta_{j,k} = \left\langle f, \psi_{j,k} \right\rangle = \int_a^b f(x)\psi_{j,k}(x)\,dx.$$

The function f can be expanded into a wavelet series as

$$f(x) = \sum_{j=-1}^{\infty} \sum_{k=0}^{2^j-1} \beta_{j,k}\psi_{j,k}(x).$$

This expansion decomposes f into components with different resolutions. The coefficients $\beta_{-1,k}$ at the coarsest level capture the gross structure of the function f. The detail coefficients $\beta_{j,k}$, when $j \geq 0$, represent finer and finer structures in f as the resolution level j increases.

When x results from a random variable with distribution $G(\cdot)$, we may use it to warp the wavelet basis, which results in a non-orthogonal warped wavelet basis Kerkyacharian and Picard (2004). Provided that $f \circ G^{-1} \in L_2[a, b]$, we can expand the function f, in a mean square sense, as

$$f(x) = \sum_{j=-1}^{\infty} \sum_{k=0}^{2^j-1} \beta_{j,k}\psi_{j,k}(G(x)), \tag{1}$$

with

$$\beta_{j,k} = \int_a^b \psi_{j,k}(G(x))f(x)g(x)\,dx, \tag{2}$$

where g is the density associated to the distribution G, i.e., the derivative of G. Using a simple change of variables $y = G(x)$ at (1) and (2), we can write

$$f(G^{-1}(y)) = \sum_{j=-1}^{\infty} \sum_{k=0}^{2^j-1} \beta_{j,k}\psi_{j,k}(y), \tag{3}$$

with

$$\beta_{j,k} = \int_0^1 \psi_{j,k}(y)f(G^{-1}(y))\,dy. \tag{4}$$

Thus, $\beta_{j,k}$ is also the coefficient of the function $f(G^{-1}(\cdot))$ in the initial wavelet basis.

2.2 Besov and weighted Besov spaces

Until now, we have only mentioned $L_2[a, b]$, the space of functions that are square-integrable. This space can be generalized to $L_p[a, b]$, $0 < p \leq \infty$, the space of functions where $(\int_a^b |f(x)|^p)^{1/p}\, dx < \infty$, for $0 < p < \infty$, and $\operatorname{ess\,sup}_{x \in [a,b]} |f(x)| < \infty$, for $p = \infty$. Further generalizations result in other known function spaces, such as Hölder, Sobolev, and Besov spaces. Besov spaces is the most general since it includes the previous spaces (see e.g. Triebel (1992), for details).

For the definition of Besov spaces, let $\Delta_h f(x) = f(x+h) - f(x)$ and $\Delta_h^{N+1} f(x) = \Delta_h \Delta_h^N f(x)$, $N \in \mathbb{N}$. Also let the modulus of continuity ρ^N be given as

$$\rho^N(t, f, p) = \sup_{|h| \leq t} \left(\int_a^b |\Delta_h^N f(x)|^p\, dx \right)^{1/p}, \tag{5}$$

for $0 < p < \infty$, with the usual modification for $p = \infty$, and let us define the following (regular) Besov space:

$$B_{s,p,q} = \left\{ f : \left(\int_0^1 \left(\frac{\rho^N(t, f, p)}{t^s} \right)^q \frac{dt}{t} \right)^{1/q} < \infty \right\}. \tag{6}$$

The parameter s can be related to the number of derivatives of f, while p and q captures a number of smoothness features, including spatially inhomogeneous behaviors.

Now, when x results from a random variable with distribution $G(\cdot)$, we may define the weigthed Besov spaces. In order to do this, let $\Delta_h(G) f(x) = f(G^{-1}(G(x) + h)) - f(x)$ and $\Delta_h^{N+1}(G) f(x) = \Delta_h(G) \Delta_h^N(G) f(x)$, $N \in \mathbb{N}$. Also let the modulus of continuity $\tilde{\rho}^N(t, f, G, p)$ be given as the right-hand side of (5), but with $\Delta_h^N(G)$ instead of Δ_h^N. Thus, the weighted Besov space $B_{s,p,q}^G$ is defined as in (6), but with $\tilde{\rho}^N(t, f, G, p)$ in the place of $\rho^N(t, f, p)$.

Note that a weighted Besov space reduces to a regular Besov space when G is the uniform distribution. Besov spaces are convenient for us because they can be expressed in terms of wavelet coefficients. To see this, we first need the following definition Kerkyacharian and Picard (2004).

Definition 1 If $\mathcal{B}$ is the set of all intervals of $\mathbb{R}$ and if f is a measurable function, then, for any interval $I \subset \mathcal{B}$, the Hardy-Littlewood maximal function associated to f is

$$f^*(x) = \sup_{I \in \mathcal{B}, x \in I} \left(\frac{1}{|I|} \int_I |f(u)|\, du \right).$$

Thus, $f^*(x)$ is the maximum average value that f can have on intervals that contain the point x.

For a fixed interval I and any $1 < p < \infty$, a weight function $\omega \geq 0$ gives the bound

$$\int_I (f^*(x))^p \omega(x)\,dx \leq C \int_I |f(x)|^p \omega(x)\,dx,$$

where C is a constant independent of f, if and only if it satisfies the condition of the Theorem 2 of Muckenhoupt (1972). Let $a, b \geq 0$ and $p - 1 = p/q$, such that $1/p + 1/q = 1$. Since $\omega \geq 0$ and $p > 1$, then $a \leq b \Rightarrow a^{1/p} \leq b^{1/p}$, and we can rewrite his Theorem 2 as the following definition.

Definition 2 For $1 < p < \infty, 1/p + 1/q = 1$ and for any interval $I \subset \mathbb{R}$, a measurable function $\omega \geq 0$ is a Muckenhoupt weight (or belongs to the Muckenhoupt class A_p) if there exists a constant $0 < C < \infty$, that depends on p and ω but is independent of I, such that

$$\left(\frac{1}{|I|}\int_I \omega(x)\,dx\right)^{1/p}\left(\frac{1}{|I|}\int_I \omega(x)^{-q/p}\,dx\right)^{1/q} \leq C.$$

For $p = 1$, $\omega \geq 0$ belongs to the Muckenhoupt class A_1 if there exists $0 < C < \infty$ such that $\omega^*(x) \leq C\omega(x)$, almost everywhere, where ω^* is the Hardy-Littlewood maximal function. For $p = \infty$, define

$$A_\infty = \bigcup_{p \geq 1} A_p.$$

The Muckenhoupt classes form an increasing family as p increases as well. For $p = 1$, Theorem 5 of Muckenhoupt (1972) would require $\omega(x) = 0$ or $\omega(x) = \infty$, almost everywhere. Then, the definition for the class A_1 is modified to a limiting case, as above García-Cuerva and Rubio de Francia (1985). With this definition, it is easy to see that if ω is bounded from above and below, it belongs to A_1 and, thus, to any A_p, $p > 1$. For $p = \infty$, Theorem 3 of Muckenhoupt (1972) shows that $\omega(x) > 0$ or $\omega(x) = 0$, both for almost every x in I. In some sense, the Muckenhoupt class identifies how far ω is from a uniform weight that assign value similar to its inverse at each interval Kerkyacharian and Picard (2004).

In the case when $\omega(x) = [g(G^{-1}(x))]^{-1}$ is a Muckenhoupt weight that belongs to the class $A_p([a, b])$, for some $1 \leq p \leq \infty$, we can define a weighted Besov space $B_{s,p,q}(\omega) \equiv B^G_{s,p,q}$. Additionally, if the wavelet function ψ is compactly supported on [0, N] with $r > N$ vanishing moments then, by the Corollary 1 in Kerkyacharian and Picard (2004), for f written using warped wavelets, as in (1) and (2), we have that

$$\left(\int_0^1 \left(\frac{\tilde{\rho}^N(t, f, G, p)}{t^s}\right)^q \frac{dt}{t}\right)^{1/q} < \infty$$

implies that

$$\left(\sum_{j=-1}^{\infty}\left[2^{js}2^{j/2}\left(\sum_{k=1}^{2^j-1}|\beta_{j,k}|^p\omega(I_{j,k})\right)^{1/p}\right]^q\right)^{1/q}<\infty,$$

where $I_{j,k}=[k/2^j,(k+1)/2^j]$, with the usual modification if $q=\infty$.

Note that, when g is bounded from above and below, then $\omega(x)\in A_\infty$. When G is the uniform distribution and g is its density, this result reduces to the regular Besov space and the reverse implication is also valid.

3 Model specification and estimation

Consider a situation when we observe data $(X_1,Y_1),\dots,(X_n,Y_n)$, $n=2^J$, $J\in\mathbb{N}$, and we formulate the model

$$Y_i=f(X_i)+\epsilon_i, \tag{7}$$

$i=1,2,\dots,n$.

In this formulation, the function f is unknown but square integrable on its support [a, b], with $a,b\in\mathbb{R}$. The random variables X_i, $i=1,2,\dots,n$, have all the same known (or unknown) density g, which is compactly supported on the same interval [a, b] as the function f. The respective distribution function $G(x)=\int_a^x g(u)\,du$ is continuous and strictly monotone from [a, b] to [0, 1]. Its inverse $G^{-1}(x)$ is also continuous and strictly monotone. These conditions on the distribution function and its inverse imply that $G(G^{-1}(x))=x$ and $G^{-1}(G(x))=x$, for almost every $x\in[a,b]$.

The error terms ϵ_i are independent of X_t and normally distributed, with mean zero and variance $\sigma^2<\infty$, for $i,t=1,2,\dots,n$.

As a tool, we first consider a compactly supported orthonormal wavelet basis $\{\psi_{j,k}, j\geq -1, k=1,2,\dots,2^{j-1}\}$, where $\psi_{-1,k}$ denotes the scaling function. Now we warp the wavelet basis using the distribution function G, such that we can expand the function f as given by (1) and (2).

The setting exposed so far is very general and specific assumptions follow.

Explicit structures of dependence for the variables X_i and ϵ_i are given by the following assumption.

Assumption 1 Either $\{X_i, i\in\mathbb{Z}\}$ or $\{\epsilon_i, i\in\mathbb{Z}\}$ is a stationary strong mixing process, while the other is an IID sequence. Also, given $p>1$, there exists $c>p$, $c\in 2\mathbb{N}=\{0,2,4,6,\dots\}$, and $\delta>0$, such that

$$\sum_{h=1}^{\infty}(h+1)^{c-2}(\alpha_{X,h})^{\delta/(c+\delta)}<\infty,$$

where $\alpha_{X,h}$ is the strong mixing coefficient of $\{X_i, i\in\mathbb{Z}\}$, when this is the case. In the other case, when $\{\epsilon_i, i\in\mathbb{Z}\}$ is the stationary strong mixing process, $\alpha_{X,h}$ is replaced by $\alpha_{\epsilon,h}$.

An example of a process satisfying the Assumption 1 is the first order autoregressive process given by $\epsilon_i = \theta\epsilon_{i-1} + u_i$, where u_i are IID normally distributed random variables with zero mean and variance σ^2, for $i \in \mathbb{Z}$ and $\theta \in (0, 1)$. This statement is made precise in the Appendix.

In order to keep some properties of the initial wavelet basis into the warped wavelet basis, we need the following assumption.

Assumption 2 The function $\omega(x) = [g(G^{-1}(x))]^{-1}$ is a Muckenhoupt weight and belongs to the class $A_p([a, b])$, for some $p > 1$. In particular, if $0 < g < M < \infty$, then $\omega(x) \in A_\infty$.

In this situation, we estimate the wavelet coefficients $\beta_{j,k}$ by $\hat{\beta}_{j,k}$, as

$$\hat{\beta}_{j,k} = \frac{1}{n}\sum_{i=1}^{n} \psi_{j,k}(G(X_i))Y_i. \tag{8}$$

In practice we rarely know the function G and we use its empirical estimate $\hat{G}(x) = n^{-1}\sum_{i=1}^{n} I(X_i \leq x)$, where I is the indicator function. In order to avoid technicalities, our theoretical results consider mainly G. However, we present some very basic theoretical results using $\hat{G}$ and evaluate its use through simulations and find the results are consistent with theory.

Finally, we estimate f by the following hard thresholded (thus, nonlinear) estimator $\hat{f}$ due to Kerkyacharian and Picard (2004):

$$\hat{f}(x) = \sum_{j=-1}^{J_1}\sum_{k=0}^{2^j-1} \hat{\beta}_{j,k} I\left(|\hat{\beta}_{j,k}| \geq \kappa\sqrt{\frac{\log n}{n}}\right)\psi_{j,k}(G(x)), \tag{9}$$

for some $\kappa > 0$, where $2^{J_1} \leq C\min\{n_1, n_2\}$,

$$n_1 = \sqrt{\frac{n}{\log n}} \quad \text{and} \quad n_2 = \left(\frac{n^{(3p-2)/p}}{\log n}\right)^{(p+\delta)/(p+\delta-2)},$$

for some $\delta, C > 0$. In the definition of n_2, p is the index of the norm used in the risk function, as given in the next section. Hereafter, C denotes a constant that does not depends on n. Note that for all $\delta > 0$, if $p \geq 2$, then $n_1 < n_2$. This is also true if $1 < p < 2$ and $\delta \geq 1$. Thus, we should worry only if $1 < p < 2$ and $0 < \delta < 1$, in which case n_2 can be smaller than n_1.

3.1 Risk and maxiset

Given the estimators $\hat{\beta}_{j,k}$ and $\hat{f}$, respectively, in (8) and (9), we are interested in some of its properties. For the former, we may find its expected value, its variance and covariace as well as other statistics. For the latter, we may use its risk and maxiset, which have been widely used in nonparametric regression for study and comparison of estimators. In what follows, we briefly review these statistics.

First, note that the estimator $\hat{f}$ in (9) depends on the sample size n and we expect that some of its properties depend on n as well. However, for a given sample, the approximation of $\hat{f}$ to f can be quantified by a loss function $L(f,\hat{f})$, such as the L_p norm of the error given by

$$L_p(f,\hat{f}) = \|f - \hat{f}\|_p^p = \int_a^b |f(x) - \hat{f}(x)|^p \, dx,$$

where the integrated square error is a particular case when $p = 2$.

Since the data are generated by a probabilistic model, this quantity varies from sample to sample and then we may use the associated risk function

$$R(f,\hat{f}) = E\big(L_p(f,\hat{f})\big) = E\|f - \hat{f}\|_p^p,$$

where the mean integrated square error is a particular case when $p = 2$.

An estimator $\hat{f}$ is called minimax if it has the smallest possible maximum risk, in a given class of estimators, and thus, can be considered conservative. Sometimes the minimax risk can not be obtained but only its rate of decay. Nevertheless, given a rate of decay a_n and a constant $C > 0$, the quality of the estimator may also be evaluated by the associated maxiset

$$\begin{aligned} \text{Max } (\hat{f}, L_p, a_n)(C) &= \left\{ f \,:\, \sup_n R(f,\hat{f}) a_n^{-1} < C \right\} \\ &= \left\{ f \,:\, \sup_n E\|f - \hat{f}\|_p^p a_n^{-1} < C \right\}. \end{aligned}$$

Comparison of estimators based on maxisets can be considered less pessimistic than minimax comparisons since it uses only the rate of decay of the risk. Subproducts of maxisets are upper and lower bounds for minimax comparisons and we can deduce rates of convergence over other classes of functions just by proving their inclusion in the maxiset. However, while minimax risks can be evaluated for a given sample size, the notion of maxisets is of a pure asymptotic nature.

4 Theoretical results

In this section we present some results on the proposed estimator discussed at the previous section, in the cases when $\{\epsilon_i, i \in \mathbb{Z}\}$ is stationary strong mixing and $\{X_i, i \in \mathbb{Z}\}$ is IID, and vice-versa.

We begin with a prologue on a simple parametric model to give us some intuition, and then we discuss the nonparametric case.

4.1 Appraisal of a parametric case

Consider the model in (7) and the Assumption 1, but with the following simplifications. Let $f(X_i) = \beta X_i$, where the parameter β is unknown, and let $a, b \neq 0$.

Also, does not consider the specific condition on the strong mixing coefficient (the summation of Assumption 1) as well as the Assumption 2 as a whole.

In this simpler parametric case, we estimate the coefficient β by the least squares estimator $\hat{\beta}$ given as

$$\hat{\beta} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{t=1}^{n} X_t^2} = \frac{\sum_{i=1}^{n} X_i(\beta X_i + \epsilon_i)}{\sum_{t=1}^{n} X_t^2},$$

such that $E(\hat{\beta} - \beta) = 0$ and $\text{Var}\,(\hat{\beta} - \beta) = \text{ Var}\left(\sum_{i=1}^{n} X_i \epsilon_i (\sum_{t=1}^{n} X_t^2)^{-1}\right)$.

A little algebra shows that

$$\text{Var}\,(\hat{\beta} - \beta) = \sigma^2 \sum_{i=1}^{n} E\left(\frac{X_i^2}{(\sum_{t=1}^{n} X_t^2)^2}\right) + \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} E\left(\frac{X_i X_j}{(\sum_{t=1}^{n} X_t^2)^2}\right) E(\epsilon_i \epsilon_j). \tag{10}$$

When $\{X_i, i \in \mathbb{Z}\}$ is IID and $\{\epsilon_i, i \in \mathbb{Z}\}$ is stationary strong mixing, this variance is equal to

$$\sigma^2 n E\left(\frac{X_1^2}{(\sum_{t=1}^{n} X_t^2)^2}\right) + \left(E\left(\frac{X_1}{(\sum_{t=1}^{n} X_t^2)^2}\right)\right)^2 \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} E(\epsilon_i \epsilon_j). \tag{11}$$

When the opposite occurs, $\{X_i, i \in \mathbb{Z}\}$ is ergodic Bradley (2005), which is a property of a stationary sequence Davidson (1994, Sect. 13.4), and we have

$$\text{Var}\,(\hat{\beta} - \beta) = \sigma^2 n E\left(\frac{X_1^2}{(\sum_{t=1}^{n} X_t^2)^2}\right). \tag{12}$$

Since the density g is compactly supported in the interval $[a, b]$, let $m = \min(|a|, |b|)$ and $M = \max(|a|, |b|)$, such that we have the respective upper bounds from (11) and (12):

$$\text{Var}\,(\hat{\beta} - \beta) \leq \sigma^2 n \frac{M^2}{n^2 m^4} + \left(\frac{M}{n^2 m^4}\right)^2 n^2 \sigma^2 = O(n^{-1});$$
$$\text{Var}\,(\hat{\beta} - \beta) \leq \sigma^2 n \frac{M^2}{n^2 m^4} = O(n^{-1}).$$

Mixing conditions would impact mostly the second term of (10), but the boundedness condition on g, together with the interplay of IID conditions, turn results much more workable: when strong mixing brings difficulties, IID (or zero mean) conditions alleviate them. Were both processes mixing, more assumptions would be needed to achieve similar results with substantially more mathematical derivations.

4.2 The nonparametric case

In the nonparametric case, for a specific strong mixing condition and loss function, the estimator has the same risk and maxiset in both cases, in part due to the interplay between IID and strong mixing conditions, as illustrated with the parametric case.

Now, in the model (7), we consider the whole Assumptions 1 and 2 for the nonparametric case. Most of the results presented remains the same for both situations. Proofs are given in the Supplementary Material.

Some basic results about the estimators are given in the following proposition.

Proposition 1 *Consider the model* (7) *and the Assumption* 1*. Then, for any funcion* $f \in B_{s,p,q}(\omega)$ *the estimator* $\hat{\beta}_{j,k}$ *given by* (8) *and* (9) *satisfies:*

1. $E(\hat{\beta}_{j,k}) = \beta_{j,k}$;
2. $\text{Var}\,(\hat{\beta}_{j,k}) = O(n^{-1})$;
3. $\text{Cov}\,(\hat{\beta}_{j,k}, \hat{\beta}_{j',k'}) = O(n^{-1})$,

for any $-1 \le j, j' \le J_1$, $k, k' = 0, 1, \dots, 2^j - 1$, $j \neq j'$, $k \neq k'$, $p, q, r \ge 1$, *and* $1/p + 1/q + q/r = 1$*. Also* $2^{J_1} \le C \min\{n_1, n_2\}$*, where* n_1 *and* n_2 *are given just after Eq.* (9).

Note that the expected value and the order of decay of the variance and covariance of the estimator are the same in both cases considered. Specifically, when $\{\epsilon_i, i \in \mathbb{Z}\}$ is stationary strong mixing and $\{X_i, i \in \mathbb{Z}\}$ is IID, we have that

$$\text{Var}\,(\hat{\beta}_{j,k}) = n^{-1}\left(\int_0^1 f^2(G^{-1}(y))\psi_{j,k}^2(y)\,dy - \beta_{j,k}^2 + \sigma^2\right)$$

and

$$\begin{aligned}\text{Cov}\,(\hat{\beta}_{j,k}, \hat{\beta}_{j',k'}) =& n^{-1}\left(\int_0^1 f^2(G^{-1}(y))\psi_{j,k}(y)\psi_{j',k'}(y)\,dy - \beta_{j,k}\beta_{j',k'}\right.\\ &\left.+ \sigma^2 \int_a^b \psi_{j,k}(G(x))\psi_{j',k'}(G(x))g(x)\,dx\right).\end{aligned}$$

Since

$$|\ \text{Var}\,(\hat{\beta}_{j,k})| \le n^{-1}\left(\|f\|_\infty^2 + \sigma^2\right)2^j\|\psi\|_\infty^2$$

and

$$|\ \text{Cov}\,(\hat{\beta}_{j,k}, \hat{\beta}_{j',k'})| \le n^{-1}\left(\|f\|_\infty^2 + \sigma^2\right)2^{(j+j')/2}\|\psi\|_\infty^2,$$

an upper bound on the constants for the difference

$$\sqrt{|\operatorname{Var}(\hat{\beta}_{j,k})||\operatorname{Var}(\hat{\beta}_{j',k'})|} - |\operatorname{Cov}(\hat{\beta}_{j,k}, \hat{\beta}_{j',k'})|$$

is given by

$$\left(\|f\|_\infty^2 + \sigma^2\right)2^{j/2}\left(2^{j/2} - 2^{j'/2}\right)\|\psi\|_\infty^2.$$

In the other case, when $\{X_i, i \in \mathbb{Z}\}$ is stationary strong mixing and $\{\epsilon_i, i \in \mathbb{Z}\}$ is IID, we have that

$$|\operatorname{Var}(\hat{\beta}_{j,k})| \leq n^{-1}\left(||f||_\infty^2 + \sigma^2 - \beta_{j,k}^2 + \alpha_{W,0}^{1/r} C(f, j, p, q, \psi)\right),$$

where $C(f, j, p, q, \psi) = 8||f||_\infty 2^{(j/2)[(p-2)/p+(q-2)/q]}||\psi||_\infty^{(p-2)/p+(q-2)/q}$, and

$$\begin{aligned}|\operatorname{Cov}(\hat{\beta}_{j,k}, \hat{\beta}_{j',k'})| \leq & n^{-1}\alpha_{W,0}^{1/r} C(f, j, p, q, \psi) \\ & + n^{-1}\sigma^2 2^{(j+j')/2}\|\psi\|_\infty^2 \\ & + n^{-2} C(f, j, j', p, q, \psi) \sum_{t=1}^{n} \sum_{\substack{s=1 \\ s \neq t}}^{n} \alpha_{W,|s-t|}^{1/r},\end{aligned}$$

where $W_{i,j,k} = \psi_{j,k}(G(X_i))f(X_i)$, $\alpha_{W,h}$ is the strong mixing coefficient of $\{W_{i,j,k}, i \in \mathbb{Z}\}$, and

$$C(f, j, j', p, q, \psi) = 8||f||_\infty^2 2^{j(p-2)/(2p)+j'(q-2)/(2q)}||\psi||_\infty^{(p-2)/p+(q-2)/q}.$$

As in the previous case, an upper bound on the the difference between the constants for the variances and the covariance is given by

$$\|f\|_\infty^2 + \sigma^2\left(1 - 2^{(j+j')/2}\|\psi\|_\infty^2\right) - n^{-2} C(f, j, j', p, q, \psi) \sum_{t=1}^{n} \sum_{\substack{s=1 \\ s \neq t}}^{n} \alpha_{W,|s-t|}^{1/r}.$$

In a similar context Chesneau (2013) obtained the same result for the expected value of the wavelet coefficients, but no results are presented for their variance and covariance.

Additional basic results about the estimators, considering a class of losses and probability bounds are given in the following proposition.

Proposition 2 *Consider the model* (7) *and the Assumptions* 1 *and* 2. *Then, the estimator* $\hat{\beta}_{j,k}$ *given by* (8) *and* (9) *satisfies*:

1. $E\left(|\hat{\beta}_{j,k} - \beta_{j,k}|^{2p}\right) = O\left(\left(\frac{\log n}{n}\right)^p\right)$;

2. $P\left(|\hat{\beta}_{j,k} - \beta_{j,k}| \geq \kappa\sqrt{\frac{\log n}{n}} \right) = \min\left\{ O\left(\left(\frac{\log n}{n} \right)^p \right), O\left(\left(\frac{\log n}{n} \right)^2 \right) \right\}$, *when* $\{\epsilon_i, i \in \mathbb{Z}\}$ *is stationary strong mixing and* $\{X_i, i \in \mathbb{Z}\}$ *is IID*;
3. $P\left(|\hat{\beta}_{j,k} - \beta_{j,k}| \geq \kappa\sqrt{\frac{\log n}{n}} \right) = O(n^{-B/\|\psi\|_\infty})$, *when* $\{X_i, i \in \mathbb{Z}\}$ *is stationary strong mixing and* $\{\epsilon_i, i \in \mathbb{Z}\}$ *is IID*;

for some $B > 0$, $\kappa > 0$, *any* $p > 1$, $-1 \leq j \leq J_1$, *and* $k = 0, 1, \ldots, 2^j - 1$. *Also* $2^{J_1} \leq C\min\{n_1, n_2\}$, *where* n_1 *and* n_2 *are given just after Eq.* (9).

The rate of decay of the losses are the same for each case considered, but the upper bound on the probability decays differently when $\{X_i, i \in \mathbb{Z}\}$ is stationary strong mixing and $\{\epsilon_i, i \in \mathbb{Z}\}$ is IID than in the other way around.

We now present some theoretical results using $\hat{G}$, instead of G, as more usual in practice. For technical reasons, we do like in Kerkyacharian and Picard (2004) and assume we have a random vector $(X'_1, X'_2, \ldots, X'_n)$ that is independent but identically distributed as $(X_1, X_2, \ldots, X_n)$.

Proposition 3 *Consider the model* (7) *and the Assumption* 1. *Then, for any funcion* $f \in B_{s,p,q}(\omega)$ *the estimator* $\hat{\beta}^{@}_{j,k}$ *given by*

$$\hat{\beta}^{@}_{j,k} = \frac{1}{n}\sum_{i=1}^{n} \psi_{j,k}(\hat{G}(X_i))Y_i,$$

where $\hat{G}(x) = n^{-1}\sum_{i=1}^{n} I(X'_i \leq x)$, *and* $(X'_1, X'_2, \ldots, X'_n)$ *is independent, but has the same (joint) distribution of the random vector* $(X_1, X_2, \ldots, X_n)$, *satisfies*:

1. $E(\hat{\beta}^{@}_{j,k}) = \beta_{j,k} + O(n^{-1/2})$;
2. Var $(\hat{\beta}^{@}_{j,k}) = O(n^{-1})$;
3. Cov $(\hat{\beta}^{@}_{j,k}, \hat{\beta}_{j',k'}) = O(n^{-1})$,

for any $-1 \leq j, j' \leq J_1$, $k, k' = 0, 1, \ldots, 2^j - 1$, $j \neq j'$, $k \neq k'$, $p, q, r \geq 1$, *and* $1/p + 1/q + q/r = 1$. *Also* $2^{J_1} \leq C\min\{n_1, n_2\}$, *where* n_1 *and* n_2 *are given just after Eq.* (9).

Our main interest lies in properties of $\hat{f}$, the estimator of the function f. Thus we present a first result in the following theorem.

Theorem 1 *Consider the model* (7) *and Assumptions* 1 *and* 2. *Let* $I_{j,k}$ *be an interval in the real line indexed by* j *and* k, $p > 1$, $0 < q < p$, *and*

$$l_{q,\infty}(\nu) = \{f(x) = \sum_{j=-1}^{\infty}\sum_{k=0}^{2^j-1} \beta_{j,k}\psi_{j,k}(G(x)) : \sup_{t>0} t^q \nu\{(j,k) : |\beta_{j,k}| > t\} < \infty\},$$

where $\nu\{(j,k)\} = \|\psi_{j,k}(G(\cdot))\|_p^p$. *Then, for the estimator* $\hat{f}$, *given by* (9), *its associated maxiset can be written as*

$$Max\left(\hat{f}, L_p, \left(\frac{\log n}{n}\right)^{(p-q)/2}\right)(\infty)$$
$$= l_{q,\infty}(\nu) \cap \left\{ f(x) = \sum_{j=-1}^{\infty} \sum_{k=0}^{2^j-1} \beta_{j,k}\psi_{j,k}(G(x)) \ : \right.$$
$$\left. \sup_{l \geq 0} \left\| \sum_{j=-1}^{\infty} \sum_{k=0}^{2^j-1} \beta_{j,k}\psi_{j,k}(G(x)) \right\|_p^p 2^{l(p-q)} < \infty \right\},$$

when $\nu\{(j,k)\} = 2^{jp/2}\omega(I_{j,k})$, *where* ω *is a Muckenhoupt weight.*

This theorem says that the maxiset in both cases considered are the same. The second result states that the risk of the estimator, measured by the mean integrated square error, as well as by some other L_p error, is almost minimax in both cases.

Theorem 2 *Consider the model* (7) *and Assumptions* 1 *and* 2. *Let* $p > 1$, $\pi \geq p$, *and* $s \geq 1/2$. *Then, for any funcion* $f \in B_{s,\pi,\infty}(\omega)$, *the risk of the estimator* $\hat{f}$, *given by* (9) *is given by*

$$E\|f - \hat{f}\|_p^p = E\left(\int_a^b |f(x) - \hat{f}(x)|^p \, dx\right) = O\left(\left(\frac{\log n}{n}\right)^{sp/(1+2s)}\right).$$

This theorem in fact is a restatement of Theorem 2 in Kerkyacharian and Picard (2004) and its proof follows exactly the same steps of the original theorem, but using our Propostion 2 and our Theorem 1, under Assumptions 1 and 2.

The order of decay of the risk is proven to be almost minimax when $X_i = i/n$ and the errors are IID Donoho and Johnstone (1994, 1995). When $p = 2$, this rate of decay is the same obtained by Chesneau (2013) in a one-dimensional case and regular Besov spaces. Results in Baraud et al. (2001) are minimax but they consider β-mixing dependence and a penalized least-squares estimator, being different from our framework. This rate of decay is also found by Chesneau (2014, Sect. 4.3), with the clear advantage that both the error and the predictor may be α-mixing and the error term may follow distributions other than the Gaussian. However, the density of the predictor may not vanish, i.e., there must be no parts of the domain of the unknown function with very few observations. Similar rates are found by Li (2016), and Krebs (2018). However, while these papers show results only for the expected mean integrated square error, we allow a wider range of risk functions.

Theorems 1 and 2 state that it does not matter if either $\{X_i, i \in \mathbb{Z}\}$ is strong mixing and $\{\epsilon_i, i \in \mathbb{Z}\}$ is IID or vice-versa, the maxiset and risk of the estimator $\hat{f}$ remain the same.

5 Simulations

In this section we perform some simulations in order to evaluate the effect of the sample size n, the level of noise and choice of the wavelet on the estimators. We have used the package Wavethresh Nason (2016) in environment R R Core Team (2018).

We simulated data $(X_1, Y_1), \dots, (X_n, Y_n)$, for sample sizes $n = 128, 256, 512, 1024$ and 2048, from model (7), in the following two situations:

1. when the error $\{\epsilon_i, i = 1, \dots, n\}$ is stationary strong mixing and the predictor $\{X_i, i = 1, \dots, n\}$ is IID;
2. when the predictor $\{X_i, i = 1, \dots, n\}$ is stationary strong mixing and the error $\{\epsilon_i, i = 1, \dots, n\}$ is IID.

In the first situation, we simulated:

$$Y_i = f(X_i) + \epsilon_i \text{and} \epsilon_i = \theta\epsilon_{i-1} + u_i, \tag{13}$$

for $\theta = 0.2$ and 0.7, where X_i had a known density g, u_t had standard normal density, and both variables were IID and independent from each other, for $i, t = 1, \dots, n$.

In the second situation, we simulated:

$$Y_i = f(X_i) + \epsilon_i \quad \text{and} \quad X_i = \theta X_{i-1} + u_i, \tag{14}$$

for $\theta = 0.2$ and 0.7, where u_i had a known density h (implying a known density g for X_i), ϵ_t had standard normal density, and both variables were IID and independent from each other, for $i, t = 1, \dots, n$.

We considered the following three functions $f(x)$, $0 \le x \le 1$, representing different degrees of variability.

1. Sine: $f(x) = 0.2 + 0.6\sin(\pi x)$.
2. Heavisine: $f(x) = 4\sin(4\pi x) -$ sgn $(x - 0.3) -$ sgn $(0.72 - x)$.
3. Doppler: $f(x) = \sqrt{x(1-x)}\sin(2\pi(1+\delta)/(x+\delta))$, with $\delta = 0.05$.

The last two functions were studied by Donoho and Johnstone (1994), but we rescaled them so that $0.2 \le f(x) \le 0.8$, for every $0 \le x \le 1$, in the simulations.

We used the following two densities $h(z)$, both for $0 \le z \le b'$.

1. Uniform: $h(z) = 1/C_1$.
2. Sine: $h(z) = (1 + 0.2\sin(4\pi z))/C_2$.

Examples of these functions sampled at $n = 2048$ points from an stationary α-mixing predictor with autocorrelation coefficient equal to 0.2, $SNR = 7$ and Sine density, estimated using the Symmlet9 wavelets are shown in Fig. 1.

In the first situation, we used $b' = C_1 = C_2 = 1$ and $g \equiv h$. By Theorems 1 and 2 in Andrews (1983), specially its Remark 3, the error $\{\epsilon_i, i \in \mathbb{Z}\}$ is a stationary strong mixing process with an appropriate coefficient, satisfying Assumption 1.

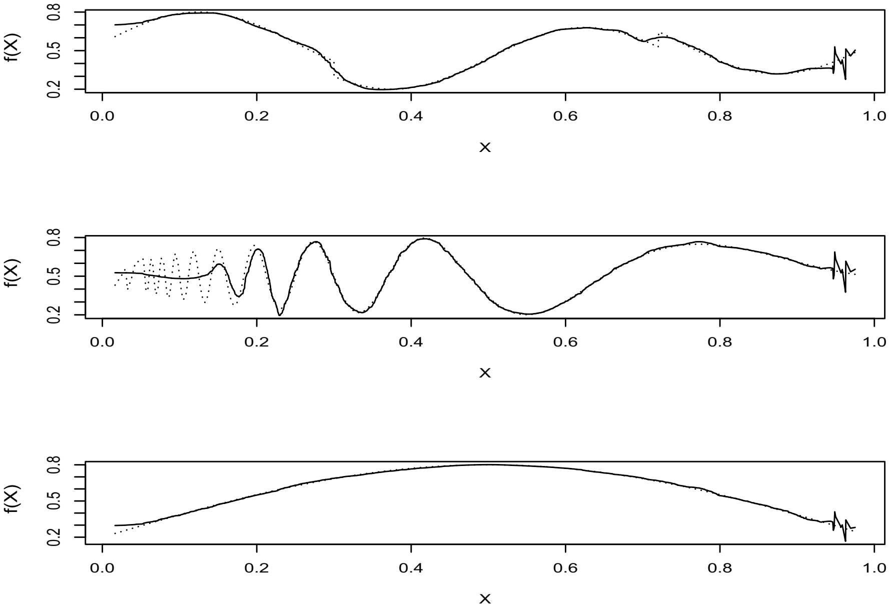

Fig. 1 Examples for the simulation study of the Heavisine, Doppler and Sine functions, respectively from the top, sampled at $n = 2048$ points following a Sine density, and stationary strong mixing predictor given by an autocorrelation coefficient equal to 0.2, using the Symmlet9 wavelets, and signal-to-noise ratio SNR = 7. Dotted line is the true noiseless function. Points are the estimated values from noisy data at predictor values

It is straightforward to see that Assumption 2, as well as the other conditions of model (7), are met for $a = 0$ and $b = 1$.

In the second situation, when $\theta = 0.2$, we used $b' = C_1 = 0.8$ and $C_2 = 0.8287914$; when $\theta = 0.7$, we used $b' = C_1 = 0.3$ and $C_2 = 0.3287914$. Note that, since $X_i = \sum_{j=0}^{\infty} \theta^j u_{i-j}$ and $0 \leq u_i \leq b'$, the process $\{X_i, i \in \mathbb{Z}\}$ is stationary and bounded in the interval $[a, b]$, where $a = 0$ and $b = 1$. By the same previous theorems in Andrews (1983), this process is (stationary) strong mixing with an appropriate coefficient, satisfying Assumption 1. Since the density g is different from the density h, we argue that Assumption 2 is met as follows. Each variable X_i is an infinite sum of independent but not identically distributed random variables $\theta^j u_{i-j}$. When u_i has an uniform density h, its mean and variance are finite, and its third central moment is zero, such that the Lyapunov condition is met for X_i. By the respective Central Limit Theorem, X_i has an asymptotic normal distribution. This distribution G, with density g, is asymptotically continuous, strictly monotonic, positive and bounded in [0, 1], which is sufficient to satisfy Assumption 2.

In this second situation, the samples X_i were obtained as follows: from $x_0 = 0$ we randomly generated a value u_i, from the known density h, by the acceptance-rejection method. Then, we set $x_i = \theta x_{i-1} + u_i$, for $i = 1, 2, \ldots, 3000$. Afterwards, the last n values were chosen, $n = 128$, 256, 512, 1024 and 2048. The respective values y_i were computed without the random error. Finally the random errors were added to the each respective value y_i. In the first situation, the samples were similarly obtained, but with the obvious modifications. Note that the design points and the errors are drawn anew in each simulation run.

For the estimator (9), we used the Daubechies' least-asymmetric wavelets with nine vanishing moments (Symmlet9) and coiflets with three vanishing moments (Coiflet3), with symmetric boundary conditions in both situations, and empirical distribution $\hat{G}(x)$. As in Chesneau and Willer (2007), we chose the constant $\kappa = 1$ and considered two levels of signal-to-noise ratio (SNR), SNR $= 1$ and SNR $= 7$, where

$$\text{SNR} = \frac{(n-1)^{-1}\sum_{i=1}^{n}(f(x_i) - \bar{f})^2}{\sigma^2},$$

and $\bar{f} = n^{-1}\sum_{i=1}^{n} f(x_i)$. In the estimator (9), thresholds were applied for levels $j \geq 0$, and J_1 was the greatest integer such that $2^{J_1} \leq \sqrt{n/\log n}$. In practice, SNR is almost not in control and our target here is to understand the behavior of the estimator under situations with low and high noise, as found in applications, and check if the main expected theoretical results still hold.

The quality of the estimator was assessed by the root mean square error (RMSE), which can be seen as an estimator of the square root of the L_2-risk, computed as the average value of 200 respective replications of

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(f(x_i) - \hat{f}(x_i)\right)^2}.$$

5.1 Simulations results

The main conclusion from our simulation study, which we detail in this section, is that the results are very similar in both situations: (i) when the error $\{\epsilon_i, i = 1, \dots, n\}$ is stationary strong mixing and the predictor $\{X_i, i = 1, \dots, n\}$ is IID; (ii) when the predictor $\{X_i, i = 1, \dots, n\}$ is stationary strong mixing and the error $\{\epsilon_i, i = 1, \dots, n\}$ is IID.

Using the root mean square error (RMSE), the simulations corroborate our main theorem.

The main results when the error $\{\epsilon_i, i = 1, \dots, n\}$ is stationary strong mixing and the predictor $\{X_i, i = 1, \dots, n\}$ is IID, are presented in the following Table 1.

Looking at Table 1, we see that the root mean square error (RMSE) decreases as the sample size increases, as predicted by the theory presented. As usual, results are better when the signal to noise ratio (SNR) is high than when SNR is low. Results for Symmlet9 are similar to those for Coiflet3 independent of the pair function-density and the SNR value.

In the other situation, when the predictor $\{X_i, i = 1, \dots, n\}$ is stationary strong mixing and the error $\{\epsilon_i, i = 1, \dots, n\}$ is IID, the results are presented in the following Table 2.

Table 1 Average of 200 root mean square error (RMSE) for the simulation study, with $Y_i = f(X_i) + \epsilon_i$, $\epsilon_i = 0.7\epsilon_{i-1} + u_i$, i.e., ϵ_i is stationary α-mixing, for each pair function-density, sample size n, signal-to-noise ratio (SNR) and wavelets Symmlet9 and Coiflet3

n	SNR = 1		SNR = 7	
	Symmlet9	Coiflet3	Symmlet9	Coiflet3
Sine-Uniform				
128	0.0749	0.0701	0.0340	0.0340
256	0.0530	0.0490	0.0243	0.0223
512	0.0428	0.0407	0.0174	0.0160
1024	0.0316	0.0298	0.0126	0.0116
2048	0.0275	0.0268	0.0103	0.0099
Sine-Sine				
128	0.0732	0.0693	0.0337	0.0337
256	0.0535	0.0490	0.0244	0.0223
512	0.0444	0.0419	0.0171	0.0163
1024	0.0311	0.0295	0.0121	0.0114
2048	0.0275	0.0266	0.0101	0.0099
Heavisine-Uniform				
128	0.0783	0.0773	0.0434	0.0468
256	0.0586	0.0552	0.0365	0.0342
512	0.0446	0.0425	0.0232	0.0226
1024	0.0342	0.0323	0.0194	0.0185
2048	0.0284	0.0278	0.0143	0.0140
Heavisine-Sine				
128	0.0773	0.0768	0.0438	0.0492
256	0.0605	0.0582	0.0392	0.0384
512	0.0454	0.0430	0.0229	0.0224
1024	0.0331	0.0320	0.0186	0.0185
2048	0.0280	0.0272	0.0140	0.0138
Doppler-Uniform				
128	0.1342	0.1304	0.1172	0.1167
256	0.1217	0.1260	0.1119	0.1194
512	0.0918	0.0936	0.0829	0.0858
1024	0.0858	0.0915	0.0812	0.0886
2048	0.0629	0.0645	0.0574	0.0604
Doppler-Sine				
128	0.1331	0.1303	0.1176	0.1176
256	0.1224	0.1241	0.1137	0.1170
512	0.1023	0.0987	0.0935	0.0910
1024	0.0978	0.0932	0.0940	0.0894
2048	0.0643	0.0654	0.0594	0.0620

Looking at Table 2, we see that the root mean square error (RMSE) decreases as the sample size increases, as predicted by the theory presented. As usual, results are better when the signal to noise ratio (SNR) is high than when SNR

Table 2 Average of 200 root mean square error (RMSE) for the simulation study, with $Y_i = f(X_i) + \epsilon_i$, $X_i = 0.7X_{i-1} + u_i$, i.e., X_i is stationary α-mixing, for each pair function-density, sample size n, signal-to-noise ratio (SNR) and wavelets Symmlet9 and Coiflet3

n	SNR = 1		SNR = 7	
	Symmlet9	Coiflet3	Symmlet9	Coiflet3
Sine-Uniform				
128	0.0235	0.0238	0.0186	0.0205
256	0.0172	0.0168	0.0138	0.0137
512	0.0128	0.0123	0.0091	0.0089
1024	0.0095	0.0104	0.0063	0.0079
2048	0.0078	0.0081	0.0043	0.0052
Sine-Sine				
128	0.0225	0.0232	0.0180	0.0197
256	0.0166	0.0169	0.0131	0.0136
512	0.0120	0.0119	0.0090	0.0089
1024	0.0091	0.0099	0.0062	0.0079
2048	0.0074	0.0079	0.0042	0.0051
Heavisine-Uniform				
128	0.0658	0.0656	0.0362	0.0429
256	0.0510	0.0461	0.0345	0.0301
512	0.0408	0.0387	0.0219	0.0218
1024	0.0304	0.0298	0.0179	0.0181
2048	0.0271	0.0269	0.0149	0.0157
Heavisine-Sine				
128	0.0641	0.0642	0.0359	0.0429
256	0.0511	0.0460	0.0327	0.0294
512	0.0401	0.0380	0.0213	0.0202
1024	0.0298	0.0285	0.0167	0.0168
2048	0.0263	0.0266	0.0136	0.0154
Doppler-Uniform				
128	0.0961	0.0928	0.0689	0.0700
256	0.0789	0.0754	0.0677	0.0625
512	0.0593	0.0591	0.0429	0.0462
1024	0.0505	0.0557	0.0419	0.0484
2048	0.0422	0.0401	0.0327	0.0306
Doppler-Sine				
128	0.0945	0.0926	0.0682	0.0704
256	0.0802	0.0753	0.0654	0.0628
512	0.0588	0.0588	0.0439	0.0452
1024	0.0506	0.0542	0.0405	0.0469
2048	0.0414	0.0407	0.0325	0.0313

is low. Results for Symmlet9 are similar to those for Coiflet3 independent of the pair function-density and the SNR value.

Similar results are shown in the Supplementary Material, for the case when the autocorrelation is weaker than in the previous simulations.

In both cases, and both situations, independent of the pair function-density, the estimates of the square root of the L_2-risk are very close and the biggest difference are driven by the sample size and the SNR.

Results other than the average, but still using the root mean square error (RMSE), are not much different from those previously reported. An example, for the Heavisine function sampled at points following a Sine density, and stationary strong mixing conditions given by an autocorrelation coefficient equal to 0.7, using the Symmlet9 wavelets, can be seen at Fig. 2. Similar to the previous results, we can see that the box-plots of root mean square errors (RMSE) get smaller as the sample size increases, and are a little narrower when SNR is high. We can also see that this behavior happens for both situations: when either the predictor or the error is stationary strong mixing. In the example shown, the results when the error is stationary strong mixing are a little better than when the predictor is stationary strong mixing. In other cases, the opposite can be true.

Comparing the results of the simulation studies, we conclude that they are qualitatively the same. When either the predictor or the error stationary is strong mixing, the root mean square error (RMSE) decreases as the sample size increases, as predicted by the theory.

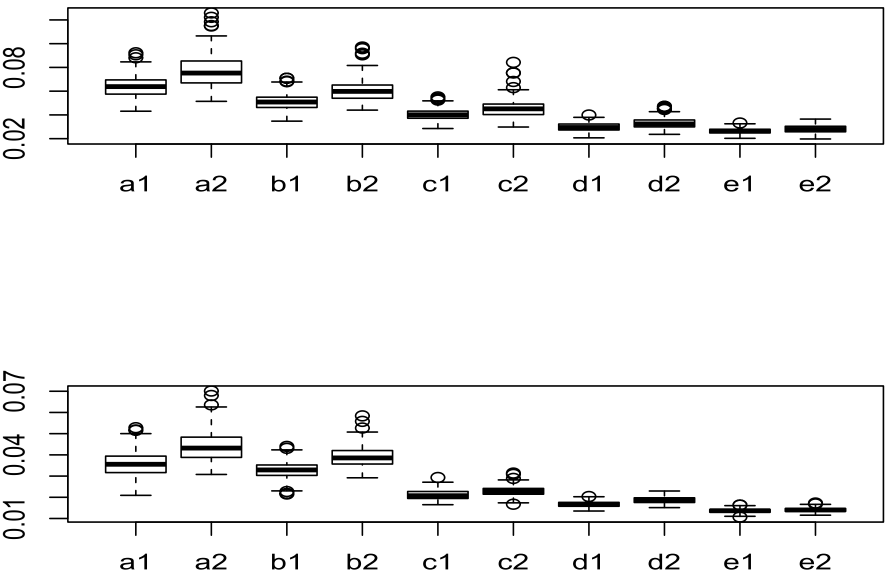

Fig. 2 Box-plots of 200 replications of the root mean square error (RMSE) for the simulation study of a Heavisine function sampled at points following a Sine density, and stationary strong mixing conditions given by an autocorrelation coefficient equal to 0.7, using the Symmlet9 wavelets, for two levels of signal-to-noise ratio (top row: SNR = 1; bottom row: SNR = 7) and five sample sizes (n: a = 128, b = 256, c = 512, d = 1024, e = 2048) when (1) predictor is stationary strong mixing and the error is IID; (2) error is stationary strong mixing and the predictor is IID

6 Application

A cryptocurrency is a digital cash that prevents double-spending by using a cryptographic ledger instead of a trusted institution like a bank or a financial institution. The first cryptocurrency invented was the Bitcoin Nakamoto (2008), but the website `coinmarketcap.com` records around 1500 cryptocurrencies traded on 10 thousand markets around the world. Unlike the usual foreign exchange market, trades occur continuously every day through the Internet using webpages or mobile applications, with very low cost and accessible for small investors. Then, cryptocurrencies are priced by traders like any other risky financial asset.

The Capital Asset Pricing Model (CAPM) is a financial pricing model independently proposed by Treynor (1961), Treynor (1962), Sharpe (1964), Lintner (1965), Mossin (1966), that describes a pricing relation for all risky assets. An introduction to the model can be found, for instance, in Van der Wijst (2013).

For the fitting of the CAPM we would need market returns such as those from a market index like CRIX Härdle and Trimborn (2015); Trimborn and Härdle (2016). However, one could argue that CRIX is very young and heavily weighted towards the Bitcoin.

A more general and empirically better model can be achieved through the Arbitrage Pricing Theory (APT) that, coming from a very different background and resting on different assumptions Ross (1976), postulate multifactor models. The price of this generality is paid in terms of uncertainty because, while CAPM's only common factor is the expected return on the market portfolio, APT does not specify what or how many factors to use. An intermediate treatment of these topics can be found, for instance, in Danthine and Donaldson (2014).

The fitting of multifactor models to cryptocurrencies is difficult because the driving factors are unknown for most of the currencies. Factors related to the market of cryptocurrencies seem to matter, while usual financial market factors like SP500 index Sovbetov (2018), traditional assets Lee et al. (2018), or foreign exchange currencies Baumöhl (2019) have weak or no correlation with cryptocurrencies.

Despite this weak correlation, we tried to investigate the presence of any nonlinear patter. We then obtain one factor from the first principal component of the returns of gold and the exchange rates between the United States Dollar and the Euro, the Japanese Yen and the British Pound. We call this factor the principal market factor (Prin 1).

Additionally, we consider a general nonlinear model that can be written as

$$R_{i,t} = f(R_{m,t}) + \epsilon_{i,t}, \tag{15}$$

where the function f can be known or unknown, and $R_{i,t}$ and $R_{m,t}$ are the respective returns of the asset i and of the principal market factor, at time t. The error term $\epsilon_{i,t}$ has zero mean and determines the idiosyncratic (diversifiable, unsystematic) risk of asset i.

When f is unknown, we may estimate it from nonparametric techniques as done by Péter Erdös and Ormos (2011) and Gómez-González and

Sanabria-Buenaventura (2014), for instance. This may be specially true with the recent advent of cryptocurrencies.

Thus we fit models (15) to a set of cryptocurrecies, with unknown *f*, as a benchmark model that possibly give insights for some nonlinear parametric models.

In order to build Prin 1, we used data from the Federal Reserve Economic Data (FRED), obtained through the website https://fred.stlouisfed.org/series/CODE, by replacing `CODE` by `GOLDAMGBD228NLBM`, `DEXUSEU`, `DEXJPUS` and `DEXUSUK` for the series of gold, Euro, the Japanese Yen and the British Pound, respectively.

We got data of log returns for six cryptocurrencies as listed in Table 3.

Log returns were calculated from daily closing prices, from the respective time periods at Table 3, sampled at each seven days, so that we had six cryptocurrencies with 64 7-days log returns. Three other cryptocurrencies are similar to those six and were not selected for this study. The other 11 cryptocurrencies were not studied because their sample sizes were smaller than 64 points.

For each cryptocurrency, we calculated 7-days log returns of Prin 1 during the respective time period. The behavior of Prin 1 log returns differ accordingly to their periods. Plots of the log returns time series are shown in Fig. 3 for each studied cryptocurrency together with the correspondent Prin 1 log returns.

Dependences in the time series of log returns were considered when autocorrelograms and partial autocorrelograms showed lags different from zero at the 5% level of significance.

Correspondent Prin 1 log returns of Bitcoin, Ethereum, Ripple and NEM have autocorrelograms with only the first lag different from zero at the 5% level of significance and the residuals of the fitted functions are white noise. In the case of Stellar, the correspondent Prin 1 log returns are white noise but the autocorrelogram of the residuals of the fitted function show a 13th significant lag. Finally, for Thether, the correspondent Prin 1 log returns and the residuals of the fitted function are both white noise.

Then, for all these cases we considered that Assumption 1 was satisfied, with Thether acting as an special case.

Estimates of the functions *f*, using model (15) for each of the six cryptocurrencies' log return (Y) as a function of CRIX log return (X), are shown in Fig. 4.

Table 3 Six cryptocurrencies and their approximate market capitalization in the end of March 2018 in billions of United States dollars

Number	Symbol	Name	Market Cap	n	Period
1	BTC	Bitcoin	126.6	64	2016-06-29–2018-03-09
2	ETH	Ethereum	41.0	64	2016-07-19–2018-03-28
3	XRP	Ripple	21.9	64	2016-07-19–2018-03-28
8	XLM	Stellar	4.4	64	2016-07-13–2018-03-22
14	XEM	NEM	2.4	64	2016-07-19–2018-03-28
15	USDT	Tether	2.3	64	2016-07-13–2018-03-22
Total			142.6		

Sample size n corresponds to 7-days log returns during the respective period, in year-month-day format

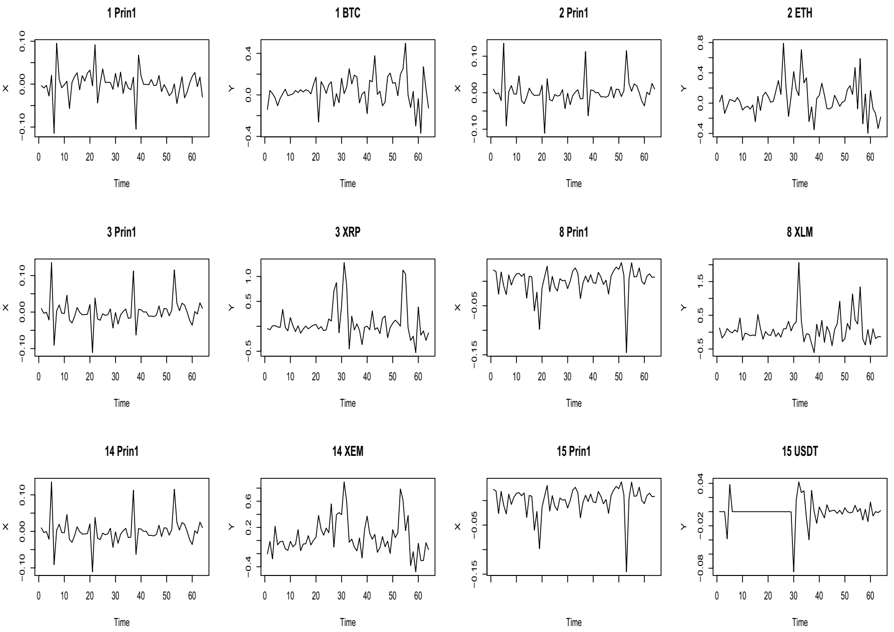

Fig. 3 Time series of 7-days log returns of Prin 1 and six studied cryptocurrencies. They are plotted in pairs, over the matching period of Prin 1 and each studied cryptocurrency

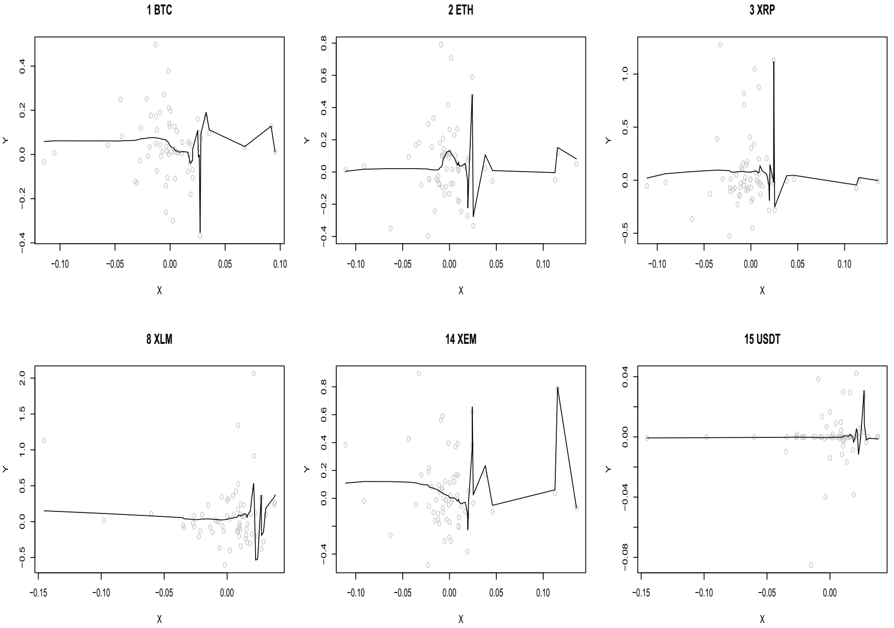

Fig. 4 Estimates of the functions *f* (full curves), using model (15) for the six cryptocurrencies log return (*Y*) as a function of Prin 1 log return (*X*). Pairs of observations *X* and *Y* are plotted as gray circles

Estimation was done using the package Wavethresh Nason (2016) in environment R R Core Team (2018). We considered Daubechies' least asymmetric orthonormal compactly supported wavelets with nine vanishing moments, symmetric boundary conditions, and empirical distribution $\hat{G}(x)$. The estimator (9) was exactly the same used in the simulations. Residuals from model fitting are clearly not normal by Shapiro-Wilk tests. Normal quantile-quantile plots show the financial stylized fact of heavy tails.

In Fig. 4, we may discard the fitted functions at their extremities and consider only the part where data are not sparse, roughly corresponding to the range $[-0.03, 0.02]$.

In this range, returns of Bitcoin and Ethereum are weak, but present some small nonlinear behaviors as a function of the returns of Prin 1. However, the log returns of NEM are pretty linear of those of Prin 1. Finally, log returns of Ripple, Stellar and Thether are not correlated with the returns of Prin 1.

7 Conclusion

The interest of this paper was to obtain convergence rates for the risk of estimators in nonparametric regression models, when either the error or the predictor is a stationary α-mixing process. Since the design is irregular, warped wavelets Kerkyacharian and Picard (2004) are used instead of usual wavelets, which are appropriate for equally spaced designs. Results were obtained under several assumptions on the process of the error and of the predictor, the regularity of the function f to be estimated and on the density of the design. The errors were assumed normal, since our theoretical results rely on some specific properties. However, we believe the results may hold for a wider class of densities by using a more general assumption like Assumption **H1** in Chesneau (2013). An apparently theoretically harder improvement would be to consider both the error and the predictor stationary strong mixing. We leave these improvements for future research.

Through simulations we assessed the behavior of the proposed estimators for finite samples. The study shows that we can expect to obtain good results in practice even for a moderate sample size.

When the procedures using warped wavelets were applied to real data, we obtained good results, as expected from the simulation study. In the application, we used the results of these paper to apply the same estimation procedure to models that show stationary strong mixing conditions sometimes in the error term and, other times, in the predictor component. The warped wavelet estimators seem good at the non sparse range of the data and show some nonlinearities for some cases.

Supplementary Information The online version of this article (https://doi.org/10.1007/s10463-021-00789-0) contains supplementary material, which is available to authorized users.

Acknowledgements The authors thank two anonymous referees for suggestions that greatly improved the text. PAM acknowledges the partial support of FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo), grant 2018/04654-9 and LMG acknowledges the partial support of FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo), grant 2019/23078-1.

References

Andrews DWK (1983). *First order autoregressive processes and strong mixing*. Tech. Rep. 664, Cowles Foundation for Research in Economics at Yale University, New Haven, Connecticut.

Antoniadis, A., Fan, J. (2001). Regularization of wavelet approximations. *Journal of the American Statistical Association, 96*(455), 939–967.

Antoniadis, A., Grégoire, G., Vial, P. (1997). Random design wavelet curve smoothing. *Statistics and Probability Letters, 35*(3), 225–232.

Baraud, Y., Comte, F., Viennet, G. (2001). Adaptive estimation in autoregression or β-mixing regression via model selection. *The Annals of Statistics, 29*(3), 839–875.

Baumöhl, E. (2019). Are cryptocurrencies connected to forex? A quantile cross-spectral approach. *Finance Research Letters, 29,* 363–372.

Bradley, R. C. (2005). Basic properties of strong mixing conditions. a survey and some open questions. *Probability Surveys, 2,* 107–144. https://doi.org/10.1214/154957805100000104.

Cai, T. T., Brown, L. (1998). Wavelet shrinkage for nonequispaced samples. *The Annals of Statistics, 26*(5), 1783–1799.

Cai, T. T., Brown, L. (1999). Wavelet estimation for samples with random uniform design. *Statistics and Probability Letters, 42,* 313–321.

Chesneau, C. (2013). On the adaptive wavelet estimation of a multidimensional regression function under α -mixing dependence: Beyond the standard assumptions on the noise. *Commentationes Mathematicae Universitatis Carolinae, 4,* 527–556.

Chesneau, C. (2014). A general result on the mean integrated squared error of the hard thresholding wavelet estimator under α-mixing dependence. *Journal of Probability and Statistics, 2014,* 1–12. https://doi.org/10.1155/2014/403764.

Chesneau, C., Willer, T. (2007). Numerical performances of a warped wavelet estimation procedure for regression in random design. https://hal.archives-ouvertes.fr/hal-00133831/document

Danthine, J., Donaldson, J. (2014). *Intermediate financial theory*. Amsterdam: Academic Press Advanced Finance, Elsevier Science.

Davidson, J. (1994). *Stochastic limit theory*. New York: Oxford University Press.

Delouille, V., Franke, J., von Sachs, R. (2001). Nonparametric stochastic regression with design-adapted wavelets. *Sankhyā:The Indian Journal of Statistics*, *63,* 328–366. (**series A, Pt. 3, Special issue on Wavelets**).

Donoho, D. L., Johnstone, I. M. (1994). Ideal spatial adaptation via wavelet shrinkage. *Biometrika, 81,* 425–455.

Donoho, D. L., Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association, 90,* 1200–1224.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., Picard, D. (1995). Wavelet shrinkage: Asymptopia? *Journal of the Royal Statistical Society Series B Statistical Methodology, 57*, 301–369.

García-Cuerva, J., Rubio de Francia, J. L. (1985). *Weighted norm inequalities and related topics*. Mathematics Studies: North-Holland.

Gómez-González, J. E., Sanabria-Buenaventura, E. M. (2014). Non-parametric and semi-parametric asset pricing: An application to the Colombian stock exchange. *Economic Systems, 38*(2), 261–268.

Hall, P., Turlach, B. A. (1997). Interpolation methods for nonlinear wavelet regression with irregularly spaced design. *The Annals of Statistics, 25*(5), 1912–1925.

Härdle, W. (1990). Applied nonparametric regression. no. 19 in Econometric society monographs, Cambridge University Press, Cambridge, UK.

Härdle, W., Trimborn, S. (2015). CRIX or evaluating Blockchain based currencies. Report no. 42/2015, Mathematisches Forschungsinstitut Oberwolfach, http://crix.hu-berlin.de/data/preliminary_OWR_2015_42.pdf, meeting on The Mathematics and Statistics of Quantitative Risk Management, organized by Richard Davis, Paul Embrechts, Thomas Mikosch and Andrew Patton, 20–26 September 2015.

Härdle, W., Lütkepohl, H., Chen, R. (1997). A review of nonparametric time series analysis. *International Statistical Review, 65*(1), 49–72.

Härdle, W., Kerkyacharian, G., Picard, D., Tsybakov, A. (1998). *Wavelets, approximation, and statistical applications. lecture notes in statistics* (Vol. 129). New York: Springer.

Kerkyacharian, G., Picard, D. (2004). Regression in random design and warped wavelets. *Bernoulli, 10*(6), 1053–1105.

Krebs, J. T. N. (2018). Non-parametric regression for spatially dependent data with wavelets. *Statistics, 52*(6), 1270–1308. https://doi.org/10.1080/02331888.2018.1506924.

Lee, D. K. C., Guo, L., Wang, Y. (2018). Cryptocurrency: A new investment opportunity? *The Journal of Alternative Investments, 20*(3), 16–40.

Li, L. (2016). Nonparametric regression on random fields with random design using wavelet method. *Statistical Inference for Stochastic Processes, 19*(1), 51–69. https://doi.org/10.1007/s11203-015-9119-8.

Lintner, J. (1965). The valuation of risky assets and the selection of risky investments in stock portfolios and capital budgets. *The Review of Economics and Statistics, 47*(1), 13–37.

Mossin, J. (1966). Equilibrium in a capital asset market. *Econometrica, 34*(4), 768–783. https://doi.org/10.2307/1910098.

Muckenhoupt, B. (1972). Weighted norm inequalities for the Hardy maximal function. *Transactions of the American Mathematical Society, 165,* 207–226.

Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system. Originally posted to the cryptography mailing list The Cryptography Mailing, on November 1st, 2008.

Nason, G. (2016). wavethresh: Wavelets statistics and transforms. http://CRAN.R-project.org/package=wavethresh, r package version 4.6.8.

Ogden, R. T. (1997). *Essential wavelets for statistical applications and data analysis* (1st ed.). Boston: Birkhäuser.

Péter Erdös, D. Z., Ormos, Mihály. (2011). Non-parametric and semi-parametric asset pricing. *Economic Modelling, 28*(3), 1150–1162.

Porto, R., Morettin, P., Percival, D., Aubin, E. (2016). Wavelet shrinkage for regression models with random design and correlated errors. *Brazilian Journal of Probability and Statistics, 30*(4), 614–652. https://doi.org/10.1214/15-BJPS296.

Porto, R.F. (2008). Regressão não-paramétrica com erros correlacionados via ondaletas. PhD thesis, University of São Paulo.

Porto, R. F., Morettin, P. A., Aubin, E. C. Q. (2012). Regression with autocorrelated errors using design-adapted Haar wavelets. *Journal of Time Series Econometrics, 4*(1).

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/.

Ross, S. A. (1976). The arbitrage theory of capital asset pricing. *Journal of Economic Theory, 13*(3), 341–360.

Ruppert, D., Wand, M., Carroll, R. (2003). Semiparametric regression. Cambridge series in statistical and probabilistic mathematics, Cambridge University Press, https://books.google.com.br/books?id=Y4uEvXFP2voC.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance, 19*(3), 425–442.

Silverman, B. (1986). Density estimation for statistics and data analysis. Chapman & Hall/CRC Monographs on statistics & applied probability, Taylor & Francis, https://books.google.com.br/books?id=e-xsrjsL7WkC.

Sovbetov, Y. (2018). Factors influencing cryptocurrency prices: Evidence from bitcoin, ethereum, dash, litcoin, and monero. *Journal of Economics and Financial Analysis, 2*(22), 1–27.

Treynor, J.L. (1961). Market value, time, and risk, edited version of an unpublished personal communication to professor John Virgil Lintner, available at SSRN: https://ssrn.com/abstract=2600356 or http://dx.doi.org/10.2139/ssrn.2600356.

Treynor, J.L. (1962). Toward a theory of market value of risky assets, edited version of an unpublished manuscript, available at SSRN: https://ssrn.com/abstract=628187 or http://dx.doi.org/10.2139/ssrn.628187.

Triebel, H. (1992). Theory of function spaces II. monographs in mathematics, Springer Basel, https://books.google.com.br/books?id=Nf3LPOf0Q3kC.

Trimborn, S., & Härdle, W. (2016). CRIX an index for blockchain based currencies. SFB 649 Discussion Paper 2016-021, Humboldt-Universität zu Berlin, Germany, http://crix.hu-berlin.de/data/CRIXpaper.pdf, sFB 649 Economic Risk.

Tsybakov, A. (2009). *Introduction to nonparametric estimation*. Springer series in statistics. New York: Springer.

Wasserman, L. (2006). *All of nonparametric statistics*. Springer texts in statistics. New York: Springer.

Van der Wijst, D. (2013). *Finance*: *a quantitative introduction*. New York: Cambridge University Press.