Statistics Surveys
Vol. 14 (2020) 32–70
ISSN: 1935-7516
https://doi.org/10.1214/20-SS128

Flexible, boundary adapted, nonparametric methods for the estimation of univariate piecewise-smooth functions

Umberto Amato

Istituto di Scienze Applicate e Sistemi Intelligenti 'Eduardo Caianiello', Consiglio Nazionale delle Ricerche
Napoli, Italy
e-mail: umberto.amato@cnr.it

Anestis Antoniadis

Laboratoire Jean Kuntzmann, Department of Statistics, University Grenoble Alpes
Grenoble, France
Department of Statistical Sciences, University of Cape Town
Cape Town, South Africa
e-mail: Anestis.Antoniadis@univ-grenoble-alpes.fr

Italia De Feis

Istituto per le Applicazioni del Calcolo 'Mauro Picone', Consiglio Nazionale delle Ricerche
Napoli, Italy
e-mail: i.defeis@iac.cnr.it

Abstract: We present and compare some nonparametric estimation methods (wavelet and/or spline-based) designed to recover a one-dimensional piecewise-smooth regression function in both a fixed equidistant or not equidistant design regression model and a random design model.

Wavelet methods are known to be very competitive in terms of denoising and compression, due to the simultaneous localization property of a function in time and frequency. However, boundary assumptions, such as periodicity or symmetry, generate bias and artificial wiggles which degrade overall accuracy.

Simple methods have been proposed in the literature for reducing the bias at the boundaries. We introduce new ones based on adaptive combinations of two estimators. The underlying idea is to combine a highly accurate method for non-regular functions, e.g., wavelets, with one well behaved at boundaries, e.g., Splines or Local Polynomial. We provide some asymptotic optimal results supporting our approach. All the methods can handle data with a random design. We also sketch some generalization to the multidimensional setting.

To study the performance of the proposed approaches we have conducted an extensive set of simulations on synthetic data. An interesting regression analysis of two real data applications using these procedures unambiguously demonstrates their effectiveness.

MSC 2010 subject classifications: Primary 62H12, 60K35; secondary 62G08.

Keywords and phrases: Wavelets, boundary corrections, nonparametric regression, smoothing splines, thresholding, model selection, backfitting.

Received November 2019.

Contents

1 Introduction 33
2 Wavelets and nonparametric regression 36
2.1 Wavelet series expansions and discrete wavelet transform (a review) 36
2.2 Finite interval wavelet transform 39
2.3 Basis expansions 40
3 Boundary treatment in wavelet thresholding with equidistant design data 42
3.1 Polynomial wavelet regression (PWR) 43
3.2 Local polynomial wavelet regression (LPWR) 43
3.3 Proposed adaptive combination of two regression estimators . . . 44
3.3.1 The adaptive estimator 46
3.4 Some other methods: trend filtering 47
4 Proposals for handling the boundary problem in the general case . . . 48
4.1 Spline-wavelet adaptive combination 48
4.2 Spline-wavelet stacking 49
4.3 Matrix stacking regression of spline and wavelet bases 49
4.4 Adaptive regression mixing and aggregation 50
4.5 Greedy pursuit and best basis selection from multiple libraries . 52
4.6 Gaussian processes and stochastic partial differential equations . 53
5 Multidimensional problems 54
6 Numerical experiments 54
6.1 Methods 54
6.2 Test functions 56
6.3 Results 57
6.4 Real examples 62
7 Conclusions 65
A Appendix 66
Acknowledgements 67
References 67

1. Introduction

Suppose we are given data:

$$Y_i = f(x_i) + \sigma\varepsilon_i, \tag{1.1}$$

$1 \le i \le n$, where $x_i = (i-1)/(n-1)$, $\sigma > 0$ and the ε_i's are i.i.d. Gaussian $N(0,1)$ random errors. The function f is an unknown function of interest. We wish to estimate f globally and one can measure the performance of an estimate $\hat{f}$ by the expected global squared L_2 norm risk:

$$R(\hat{f}, f) = \mathbb{E} \int_0^1 (\hat{f}(x) - f(x))^2 dx,$$

the goal being to construct estimates that have "small" risk. In order to have some meaningful estimate according to this criterion, one must assume certain regularity conditions on the unknown function f, such as f belongs to some Hölder classes, Sobolev classes, Besov classes and so forth. When the regression function f is sufficiently smooth, efficient smoothing methods such as kernel, splines and basis expansions have received considerable attention in the nonparametric literature (see, for example, Green and Silverman (1993); Eubank (1999); Härdle (1990); Hastie (2003) and references given therein). In contrast, which is the case of this paper, when the unknown function f, mostly smooth, is suspected to have few discontinuities, sharp spikes and abrupt changes, wavelet methods are very popular. The application of wavelet theory to the field of statistical function estimation was pioneered in Donoho and Johnstone (1995); Donoho et al. (1995). The methodology includes a coherent set of procedures that are spatially adaptive and near optimal over a range of function spaces of inhomogeneous smoothness. Wavelet procedures achieve adaptivity through thresholding of the empirical wavelet coefficients. They enjoy excellent mean squared error properties when estimating functions that are only piecewise smooth and have near optimal convergence rates over large function classes. For example they attain optimal convergence rates for the L_2 risk when f is in a ball of a Besov space $\mathcal{B}^s_{p,r}$ for $p < 2$, which can not be achieved by any linear estimator. For a thorough review of wavelet methods in statistics the reader is referred to Antoniadis (2007).

Despite their considerable advantages, however, standard wavelet procedures have limitations. It might be noticed that the vast majority of wavelet-based regression estimation have been conducted within the setting that the design points are fixed and equally spaced to enable the application of the wavelet transform to a compactly supported signal. Moreover, equispaced design or not, it is customary to impose some boundary assumptions, such as periodicity or symmetry, on the regression function. Such assumptions are not always reasonable, and certain types of bias and artificial wiggles often arise in this context, particularly those due to edge or boundary effects, which detract from the global performance of the estimators and whose influence should be reduced or eliminated whenever possible.

To handle such boundary problems, at least in the equispaced design case, three types of approach are used in wavelet regression: one can either impose extra constraints on the function f, such as periodicity, symmetry or antisymmetry (Ogden, 1996), or construct specialized wavelets on a compact interval (Cohen, Daubechies and Vial, 1993) or combine low-order polynomial terms and wavelet basis (Oh, Naveau and Lee, 2001). In the first strategy, artificially large wavelet coefficients result when the extra conditions on the regression function f are not satisfied. For the second strategy, while theoretically appealing, implementation of a modified discrete wavelet transform is considerably more involved. In the third method called polynomial wavelet regression (PWR) the

idea is to estimate f with the sum of a set of (Coiflets) wavelet basis functions, $\hat{f}_{\mathrm{W}}$, and a low-order (global) polynomial, $\hat{f}_{\mathrm{P}}$:

$$\hat{f}_{\mathrm{PW}}(x) = \hat{f}_{\mathrm{P}}(x) + \hat{f}_{\mathrm{W}}(x),$$

where $\hat{f}_{\mathrm{PW}}$ is the PWR estimate for f. The hope is that, once $\hat{f}_{\mathrm{P}}(x)$ is removed from the data, the remaining signal hidden in the residuals can be well estimated using wavelet regression with, for example, periodic boundary assumption. The use of Coiflets allows, with appropriately chosen hyper-parameters, to prove both analytically and empirically that polynomial wavelet regression is superior to wavelet regression for functions of inhomogeneous smoothness. The use of PWR for resolving boundary problems works very well if $\hat{f}_{\mathrm{P}}(x)$ is able to remove the "non-periodicity" in the data. However, due to the global nature of $\hat{f}_{\mathrm{P}}(x)$ for those cases when f has complex boundary conditions or has some abrupt changing objects present near the boundaries, PWR does not work well. Oh and Lee (2005) therefore extend the PWR method to a method called LWPR by combining wavelet shrinkage with local polynomial regression which is known to possess excellent boundary properties. Note however, that no asymptotic analysis for the resulting estimator is given in their paper, but extensive simulation results provide some strong evidence that LWPR is effective in correcting boundary bias. Because of its performance and its simplicity we have chosen the last approach as a starting point to de-noise signals with irregular boundaries in the equispaced design case.

A closer look at the LWPR estimate shows that it can be considered as a linear combination of a local polynomial estimator and a wavelet regression estimator with equal coefficients. Our purpose is then to adopt a different approach by considering an adaptive combination of the two estimators, one based on stronger smoothing assumptions on the regression function f and well behaved at the boundaries and another one based on weaker assumptions. The adaptive choice of the weights will also allow us to get some asymptotic optimality results for the combined estimator.

A disadvantage of the above is that the method is limited to the simple equispaced dyadic case. In practice, however, there are many interesting applications in statistics where the samples are not equispaced and their size is not dyadic. It is therefore interesting to propose appropriate penalization methods to wavelet smoothing within the setting of non-equally spaced and non-dyadic design points that can handle efficiently the boundary problems. This will be studied in the second part of the paper. Let us just say that the idea of adaptively combining different regression procedures within a collection of regression procedures (e.g. kernel, spline, wavelet, local polynomial, etc.) will be explored in a context of ensemble learning by mixing or aggregation and compared to other wavelet regression procedures for random design univariate regression.

This paper is organized as follows. In Section 2 we describe the wavelet-based regression model with the basic concept of wavelets. We also present an aspect of wavelets described in Antoniadis and Fan (2001) crystallizing the penalized least squares approaches to wavelet nonparametric regression showing that they can be used to construct a set of basis functions over an arbitrary

compact interval and that linear combinations of such basis functions are able to estimate particular, usually jagged, regression functions better than spline bases. A detailed description of our boundary corrections procedures and their asymptotic properties in the equispaced case is presented in Section 3. The random design case is studied in Section 4, and the procedures are investigated via various simulation settings and real data application examples in Section 6. Some concluding remarks are given in Section 7.

2. Wavelets and nonparametric regression

We consider the regression problem stated in Equation (1.1) with a non-stochastic equidistant design $t_i = (i-1)/(n-1)$, $i = 1, \ldots, n$ of size $n = 2^J$ for some positive integer J, noise variables ϵ_i that are i.i.d. Gaussian $\mathcal{N}(0, \sigma^2)$ and with a potentially nonsmooth regression function f that may present a wide range of irregular effects. Wavelets provide smoothness characterization of several function spaces. Many traditional smoothness spaces, for example Hölder spaces, Sobolev spaces and Besov spaces, can be completely characterized by the sequence of wavelet coefficients (e.g., Meyer, 1993). In the present paper we will consider the problem of estimating the regression function either over a range of piecewise Hölder classes or through the sequence space representation of Besov spaces. A function in a piecewise Hölder class can be regarded as the superposition of a regular smooth function in a Hölder class and an irregular perturbation consisting of jump discontinuities. The (inhomogeneous) Besov spaces on the unit interval (e.g., Donoho and Johnstone, 1998) $\mathcal{B}^s_{p,r}([0,1])$, consist of functions that have a specific degree of smoothness in their derivatives. The parameter p can be viewed as a degree of function's inhomogeneity while s is a measure of its smoothness. Roughly speaking, the (not necessarily integer) parameter s indicates the number of function's (fractional) derivatives, where their existence is required in an L^p-sense; the additional parameter r is secondary in its role, allowing for additional fine tuning of the definition of the space. Assuming that f belongs either to piecewise Hölder class or a Besov space $\mathcal{B}^s_{p,r}([0,1])$ with $s + 1/p - 1/2 > 0$ (this condition ensures in particular that evaluation of f at a given point makes sense) enables to capture key characteristics of inhomogeneity in f and to exploit its sparse wavelet coefficients representation. Notice that Besov spaces contain not only the standard Hölder and Sobolev spaces but also the piecewise Hölder spaces with a finite number of discontinuous jumps (Meyer, 1993).

2.1. Wavelet series expansions and discrete wavelet transform (a review)

In this subsection, the wavelet transform and its implementation for discrete signals are briefly reviewed. The sole purpose of this review is to describe the tools which will be used later. We assume that we are working within an orthonormal basis generated by dilations and shifts of a compactly supported

scaling function, $\phi(t)$, and a compactly supported mother wavelet, $\psi(t)$, associated with an r-regular ($r \geq 0$) multi-resolution analysis of $L^2(\mathbb{R})$, the space of square integrable functions. The wavelet theory basically considers functions on the real line. When a finite interval $[0,1]$ is involved then an easy solution is to consider periodic wavelets. More precisely, let $\left(L^2[0,1], \langle \cdot, \cdot \rangle\right)$ be the space of squared-integrable functions on $[0,1]$ endowed with the inner product $\langle f, g \rangle = \int_0^1 f(t)g(t)dt$. Assuming that f is periodic, one may work with periodic wavelet bases on $[0,1]$ (e.g., Mallat, 2009, Section 7.5.1), letting

$$\phi_{jk}^{\text{per}}(t) = \sum_{l \in \mathbb{Z}} \phi_{jk}(t-l) \text{ and } \psi_{jk}^{\text{per}}(t) = \sum_{l \in \mathbb{Z}} \psi_{jk}(t-l),\ t \in [0,1],$$

where $\phi_{jk}(t) = 2^{j/2}\phi(2^j t - k)$ and $\psi_{jk}(t) = 2^{j/2}\psi(2^j t - k)$. For any given primary resolution level $j_0 \geq 0$, the collection

$$\{\phi_{j_0 k}^{\text{per}},\ k = 0, 1, \ldots, 2^{j_0} - 1;\ \psi_{jk}^{\text{per}},\ j \geq j_0;\ k = 0, 1, \ldots, 2^j - 1\}$$

is then an orthonormal basis of $L^2[0,1]$. The superscript "per" will be suppressed from the notation for convenience. Despite the poor behaviour of periodic wavelets near the boundaries, where they create high amplitude wavelet coefficients, they are commonly used because the numerical implementation is particularly simple. Therefore, for any $f \in L^2[0,1]$, we denote by $c_{j_0 k} = \langle f, \phi_{j_0 k} \rangle$, $k = 0, 1, \ldots, 2^{j_0} - 1$, the scaling coefficients and by $d_{jk} = \langle f, \psi_{jk} \rangle$, $j \geq j_0$, $k = 0, 1, \ldots, 2^j - 1$, the wavelet coefficients of f for the orthonormal periodic wavelet basis defined above; the function f is then expressed in the form

$$f(t) = \sum_{k=0}^{2^{j_0}-1} c_{j_0 k}\phi_{j_0 k}(t) + \sum_{j=j_0}^{\infty} \sum_{k=0}^{2^j - 1} d_{jk}\psi_{jk}(t),\ t \in [0,1].$$

The approximation space spanned by the scaling functions $\{\phi_{j_0 k},\ k = 0, 1, \ldots, 2^{j_0} - 1\}$ is usually denoted by V_{j_0} while the details space at scale j, spanned by $\{\psi_{jk},\ \ k = 0, 1, \ldots, 2^j - 1\}$, is usually denoted by W_j.

In statistical settings, we are more usually concerned with discretely sampled, rather than continuous, functions. It is then the wavelet analogy to the discrete Fourier transform which is of primary interest and this is referred to as the discrete wavelet transform (DWT). Given a vector of real values $\boldsymbol{e} = (e_1, \ldots, e_n)^T$, the discrete wavelet transform of $\boldsymbol{e}$ is given by $\boldsymbol{d} = W_{n \times n}\boldsymbol{e}$, where $\boldsymbol{d}$ is an $n \times 1$ vector comprising both discrete scaling coefficients, $s_{j_0 k}$, and discrete wavelet coefficients, w_{jk}, and $W_{n \times n}$ is an orthogonal $n \times n$ matrix associated with the orthonormal periodic wavelet basis chosen. In the following we will distinguish the blocks of $W_{n \times n}$ spanned by the scaling functions and the wavelets, respectively. The empirical coefficients $s_{j_0 k}$ and w_{jk} of $\boldsymbol{e}$ are given by

$$s_{j_0,k} \approx \frac{1}{\sqrt{n}} \sum_{i=1}^{n} e_i \phi_{j_0,k}(t_i),\ k = 0, \ldots, 2^{j_0} - 1$$

$$w_{j,k} \approx \frac{1}{\sqrt{n}} \sum_{i=1}^{n} e_i \psi_{j,k}(t_i),\ j = j_0, \ldots, J-1,\ k = 0, \ldots, 2^j - 1.$$

When $\boldsymbol{e}$ is a vector of function values $\boldsymbol{f} = (f(t_1), ..., f(t_n))^T$ at equally spaced points t_i, the corresponding empirical coefficients s_{j_0k} and w_{jk} are related to their continuous counterparts c_{j_0k} and d_{jk} (with an approximation error of order n^{-1}) via the relationships $s_{j_0k} \approx \sqrt{n}\, c_{j_0k}$ and $w_{jk} \approx \sqrt{n}\, d_{jk}$ (see for example Lemma 2 in Cai and Brown (1998)). Note that, because of orthogonality of $W_{n\times n}$, the inverse DWT (IDWT) is simply given by $\boldsymbol{f} = W_{n\times n}^{\mathrm{T}}\boldsymbol{d}$, where $W_{n\times n}^{\mathrm{T}}$ denotes the transpose of $W_{n\times n}$. If $n = 2^J$ for some positive integer J, the DWT and IDWT may be performed through a computationally fast algorithm (e.g., Mallat, 2009, Section 7.3.1) that requires only order n operations. Hereafter, the coarsest wavelet decomposition level j_0 will be chosen to be the closest integer to $\log_2(\log(n)) + 1$, as suggested in Antoniadis, Bigot and Sapatinas (2001).

Let us adopt a vector-matrix form of the nonparametric model given by equation (1.1):

$$\boldsymbol{Y} = \boldsymbol{f} + \boldsymbol{\epsilon},$$

for $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$, $\boldsymbol{f} = (f(t_1), \ldots, f(t_n))^T$ and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$. After applying a linear and orthogonal discrete wavelet transform $W_{n\times n}$, the above discretised model becomes

$$\boldsymbol{z} = \boldsymbol{\gamma} + \tilde{\boldsymbol{\epsilon}}, \tag{2.1}$$

where $\boldsymbol{z} = W_{n\times n}\boldsymbol{Y}$, $\boldsymbol{\gamma} = W_{n\times n}\boldsymbol{f}$ and $\tilde{\boldsymbol{\epsilon}} = W_{n\times n}\boldsymbol{\epsilon}$. The orthogonality of the DWT matrix $W_{n\times n}$ ensures that the transformed noise vector $\tilde{\boldsymbol{\epsilon}}$ is still distributed as a Gaussian white noise with variance $\sigma^2\boldsymbol{I}_n$. Hence, the representation of the model in the wavelet domain not only allows to retain the linear structure of the model but also to exploit in an efficient way the sparsity of the wavelet coefficients in the representation of the nonparametric component.

A key step in classical wavelet regression is to estimate the true wavelet coefficients $\boldsymbol{\gamma} = W_{n\times n}\boldsymbol{f}$ by thresholding the empirical wavelet coefficients $\boldsymbol{z} = W_{n\times n}\boldsymbol{Y}$. It is known that such wavelet thresholding estimators are special cases of penalized least-squares estimators (for example Antoniadis and Fan, 2001). That is, a thresholded estimator for $\boldsymbol{\gamma}$ can be obtained as the minimizer of

$$\|\boldsymbol{z} - \boldsymbol{\gamma}\|^2 + p_\lambda(\boldsymbol{\gamma}),$$

for some suitable penalty function p_λ with penalty parameter λ. Given a penalty function p_λ which is a nonnegative, nondecreasing and differentiable on $(0, \infty)$, the solution to the minimization of the above problem exists and is unique (Antoniadis and Fan, 2001).

When $p_\lambda(\cdot) = \lambda|\cdot|$ (ℓ_1 penalty) the corresponding estimator is obtained by the soft-thresholding operator (Antoniadis, 2007)

$$\Theta_{\mathrm{soft}}(u;\lambda) = \begin{cases} 0, & \text{if } |u| \le \lambda \\ r - \mathrm{sgn}(u)\lambda, & \text{if } |u| > \lambda. \end{cases} \tag{2.2}$$

Several methods exist to select an appropriate threshold value, λ, such as the SUREshrink wavelet regression procedure of Donoho et al. (1995), the cross-validation of Nason (1996), the universal threshold $\lambda(n) = \sigma\sqrt{2\log n}$ of Donoho

and Johnstone (1998) or the EbayesThresh procedure of Johnstone and Silverman (2005). A general review about some thresholding selection methods can be found in Antoniadis, Bigot and Sapatinas (2001).

2.2. Finite interval wavelet transform

To be able to perform the wavelet regression in $[0,1]$ by the discrete wavelet transform described above we have used periodic wavelets and scaling functions, which handle the boundaries by imposing periodicity of the regression function. When such an assumption is not satisfied artificially large wavelet coefficients may be created at the boundaries. A way to solve this problem is to use a particular wavelet basis for $L^2[0,1]$ developed by Cohen, Daubechies and Vial (1993) which is closely connected with Daubechies' orthonormal and compactly supported wavelet basis of $L^2(\mathbb{R})$. This is accomplished by defining special scaling and wavelet functions at the boundaries as linear combinations of original scaling and wavelet functions. This approach naturally preserves regularity and refinability of the wavelet system and maintains orthogonality, therefore an appropriate Discrete Wavelet transform (DWT) algorithm can be applied. In practice if M denotes the number of vanishing moments of the wavelet system (or equivalently $M-1$ is the maximum degree of exactly reproducible polynomials by the wavelet system), then M boundary left scaling and wavelet functions are defined at each scale j, ϕ_{jk}^L, ψ_{jk}^L, $k=0,\dots,M-1$, starting from the generating $\phi_k^L(t)$ and $\psi_k^L(t)$ as $\phi_{jk}^L(t)=2^{j/2}\phi_k^L(2^jt)$ and $\psi_{jk}^L(t)=2^{j/2}\psi_k^L(2^jt)$; analogously M boundary right wavelet and scaling functions are defined at each scale j, ϕ_{jk}^R, ψ_{jk}^R, $k=2^j-M,\dots,2^j-1$. Together with the 2^j-2M interior unaltered scaling and wavelet functions $\phi_{jk}(t)=2^{j/2}\phi(2^jt-k)$ and $\psi_{jk}(t)=2^{j/2}\psi(2^jt-k)$, $k=M,\dots,2^j-M-1$, they represent a full multiresolution analysis on the finite interval. From a DWT perspective special filters are introduced at the left and right of the interval depending on the scale j, allowing a more involved implementation of a modified Discrete Wavelet Transform.

Under such a framework, the regression function f can be represented by (index fi means finite interval)

$$f^{\mathrm{fi}}(t)=\sum_{k=0}^{M-1}c_{j_0k}\phi_{j_0k}^L(t)+\sum_{k=M}^{2^{j_0}-M-1}c_{j_0k}\phi_{j_0k}(t)+\sum_{k=2^{j_0}-M}^{2^{j_0}-1}c_{j_0k}\phi_{j_0k}^R(t)$$
$$+\sum_{j=j_0}^{\infty}\left(\sum_{k=0}^{M-1}d_{jk}\phi_{jk}^L(t)+\sum_{k=M}^{2^j-M-1}d_{jk}\phi_{jk}(t)+\sum_{k=2^j-M}^{2^j-1}d_{jk}\psi_{jk}^R(t)\right), \tag{2.3}$$

$t\in[0,1]$, provided that $2^{j_0}\geq 2M$ so that boundary scaling and wavelet functions (and corresponding filters) do not overlap.

After applying again the finite interval adapted discrete wavelet transform $W_{n\times n}^{\mathrm{fi}}$ the nonparametric regression problem can be written in matrix notation

as

$$\boldsymbol{z} = \boldsymbol{\gamma} + \tilde{\boldsymbol{\epsilon}}, \tag{2.4}$$

where $\boldsymbol{z} = W^{\mathrm{fi}}_{n\times n}\boldsymbol{Y}$, $\boldsymbol{\gamma} = W^{\mathrm{fi}}_{n\times n}\boldsymbol{f}$ is the vector of corresponding coefficients and $\tilde{\boldsymbol{\epsilon}} = W^{\mathrm{fi}}_{n\times n}\boldsymbol{\epsilon}$. And once again wavelet regression estimation can be achieved by least squares penalization, as for the periodic case.

2.3. Basis expansions

The wavelet regression methodology discussed in the previous subsections, is designed for treating dyadic samples of equispaced data. The application of a wavelet analysis to irregularly spaced samples, eventually random, say $\boldsymbol{T}_n = (T_1, \ldots, T_n)^T$, has been a subject of study for more than ten years. Most methods in the area work with a pre- and/or post-processing of the data in order to translate the problem into an equispaced one. Cai and Brown (1998) decompose the nonequispaced data into a warped wavelet basis and then project this decomposition onto a regular wavelet basis. Antoniadis and Pham (1998) implement a direct discretisation of a continuous wavelet analysis on the irregular grid to find numerical values for wavelet coefficients corresponding to regular basis functions. Kovac and Silverman (2000) interpolate the irregular observations in intermediate regular locations before starting the wavelet analysis. These and other methods require user-driven preprocessing, that might become difficult or even fail in case the data are "very" non-equidistant and are still affected by boundary problems. The idea is then to use wavelet basis functions evaluated on irregular grids as in Antoniadis and Fan (2001) and Wand and Ormerod (2011) as it is done for other functional bases such as for example B-splines, using a basis expansion based approximation for the nonparametric function f, which provides a way of handling nonequispaced designs.

When the nonparametric function f is supposed to be smooth one may use an approximation by its expansion on O'Sullivan splines basis functions $\{B_\ell\}_{\ell\in\mathbb{N}}$:

$$f(t) \approx \sum_{\ell=1}^{m} \alpha_\ell B_\ell(t), \tag{2.5}$$

where m is an appropriate truncation index that is allowed to increase to infinity with n. We assume that the B_ℓ are in canonical form (e.g., Wand and Ormerod, 2008, Section 4). Under reasonable smoothness assumptions, the regression function can be well approximated by the above expansion and its estimation is therefore equivalent to estimate the coefficient vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)^T$. Using the O'Sullivan basis construction described in Wand and Ormerod (2008) it is easy to compute the corresponding regression $n \times m$ matrix of the O'Sullivan basis functions evaluated at the $\boldsymbol{T}_n$ irregular grid, i.e.

$$\boldsymbol{B} = \begin{bmatrix} B_1(T_1) & \ldots & B_m(T_1) \\ \vdots & \ddots & \vdots \\ B_1(T_n) & \ldots & B_m(T_n) \end{bmatrix},$$

and adopt a vector-matrix form of the nonparametric regression model (1.1) to get:

$$\boldsymbol{Y} \approx \boldsymbol{B}\boldsymbol{\alpha} + \boldsymbol{\epsilon}, \tag{2.6}$$

for $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$. The vector of spline coefficients $\boldsymbol{\alpha}$ can be estimated by minimizing an objective function of the following form:

$$\tilde{f} = \underset{f \in W_2^s[0,1]}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(T_i))^2 + \lambda J(f) \right\}, \tag{2.7}$$

where the parameter λ is a smoothing parameter that controls the trade-off between fit and smoothness.

The standard estimation procedures assume a quadratic form in the spline coefficients for the penalty $J(f)$ (e.g., Silverman, 1985; Eilers and Marx, 1996; Wood, 2006). In this case, λ can be selected by minimisation of the generalised cross validation (GCV) score, the generalised Akaike's information criterion (AIC), or restricted maximum likelihood (REML) estimation, to name a few. The computational methods of Wood (2006) implemented in the R-package `mgcv` are available to estimate f minimising (2.7). Morever, it can be shown, when $s > 1$ (e.g., Du, Parmeter and Racine, 2013, Proposition 2.1), that under appropriate conditions on the design when it is fixed or on its distribution when it is random, if $\lambda \sim n^{-2\lfloor s \rfloor/(2\lfloor s \rfloor+1)}$, then the solution of (2.7) has the following asymptotic rates:

$$R(\tilde{f}, f) = O(n^{-2\lfloor s \rfloor/(2\lfloor s \rfloor+1)}).$$

Thus $\tilde{f}$ is consistent with convergence rates similar to those obtained in the equidistant fixed design case using local polynomials. However, other penalties can be used to impose some sparseness constraint on the coefficients.

When the nonparametric function is not smooth one can approximate it using instead wavelet bases, as in Antoniadis and Fan (2001) and Wand and Ormerod (2011). More precisely, we may use its expansion on wavelet basis functions $\{W_\ell\}_\ell$:

$$f(t) \approx \sum_{\ell=1}^{K} \gamma_\ell W_\ell(t), \tag{2.8}$$

where K is again an appropriate truncation index that is allowed to increase to infinity with n. Again, for f within some Besov ball, f can be well approximated by the above expansion and the estimation is therefore equivalent to estimate the wavelet coefficient vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_K)^T$. Similarly to the spline case, as alluded to in Antoniadis and Fan (2001) and implemented in Wand and Ormerod (2011), we can also define the design matrices containing wavelet basis functions evaluated at $\boldsymbol{T}$. We will denote again by $\boldsymbol{W}$ the corresponding wavelet regression $n \times K$ matrix of the wavelet basis functions evaluated at $\boldsymbol{T}$ (see the Appendix for a brief description of such a construction), i.e.,

$$\boldsymbol{W} = \begin{bmatrix} W_1(T_1) & \ldots & W_K(T_1) \\ \vdots & \ddots & \vdots \\ W_1(T_n) & \ldots & W_K(T_n) \end{bmatrix},$$

with a corresponding vector-matrix form given by

$$\boldsymbol{Y} \approx \boldsymbol{W}\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \tag{2.9}$$

for $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$. So, estimating f is equivalent to estimate the wavelet coefficient vector $\boldsymbol{\gamma}$. However, the fact that f belongs to a Besov space implies sparsity of the wavelet coefficients and therefore the wavelet vector $\boldsymbol{\gamma}$ is obtained by minimizing an objective function of the following form:

$$\frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{W}\boldsymbol{\gamma}\|_n^2 + \sum_{k=1}^{K} P_\lambda(|\gamma_k|), \tag{2.10}$$

using some efficient penalties P_λ in terms of estimation and model selection consistency as the ones discussed in Antoniadis (2007).

Directly optimising (2.10) can be tricky for a given penalty function, especially when the penalty is non convex. To tackle the optimisation, it is more convenient to use an iterative thresholding viewpoint with a thresholding function corresponding to the selected penalty (e.g., She, 2012; Daubechies, Defrise and Mol, 2004; Bredies, Lorenz and Reiterer, 2014)). The reader may also look in Antoniadis (2007) for a survey on the one-to-one correspondence between threshold functions and penalty functions. It can be shown, using Theorem 2.1 of She (2012), that, provided that the spectral norm of the design matrix $\boldsymbol{W}$ is not large, and whatever the starting value of $\boldsymbol{\gamma}$ is, the iterated thresholding estimates minimise (2.10). The condition on boundedness of the spectral norm of $\boldsymbol{W}$ is easily obtained by rescaling the vector of coefficients $\boldsymbol{\gamma}$ and the penalty parameter λ.

Using the results of Bunea, Lederer and She (2014), assuming that K grows to infinity at an appropriate rate in such a way that a compatibility condition holds for the design matrix $\boldsymbol{W}$ rescaled by some constant C_0, assuming a finite sparsity index less than n and a bounded entropy on the class of the irregular nonparametric regression functions, the penalized estimation produces estimates $\hat{f}$ that are consistent with optimal rates $R(\hat{f}, f) = O(n^{-2s/(2s+1)})$.

The above results on regression splines and regression wavelets, can be used to derive some new estimation procedures that tackle the boundary problem in a similar fashion as for the equidistant design case.

3. Boundary treatment in wavelet thresholding with equidistant design data

As noticed in the Introduction although classical wavelet regression (assuming periodic boundaries) provides reasonable fits far away from the boundaries, often artificial wiggles appear at the boundaries, one reason being that the wavelet transform filtering operations at the boundaries require values of the regression function outside its supported range. We will therefore review in this section first some of the existing methods to treat boundary problems in wavelet regression with fixed equidistant design, namely polynomial wavelet regression (PWR)

and local polynomial wavelet regression (LWPR), before proposing our adaptive combination of local polynomial and wavelet regression estimators with sound asymptotic properties.

3.1. Polynomial wavelet regression (PWR)

The polynomial wavelet regression estimator (PWR) was proposed by Oh, Naveau and Lee (2001). It is based on a combination of a wavelet based regression estimator $\hat{f}_{\mathrm{W}}$ and a low-order polynomial fit $\hat{f}_{\mathrm{P}}$. The resulting estimator of f, $\hat{f}_{\mathrm{PW}}$ is written as

$$\hat{f}_{\mathrm{PW}}(t)=\hat{f}_{\mathrm{P}}(t)+\hat{f}_{\mathrm{W}}(t)=\sum_{\ell=0}^{d}\hat{\beta}_\ell t^\ell+\sum_{k=0}^{2^{j_0}-1}\hat{c}_{j_0k}\phi_{j_0k}(t)+\sum_{j=j_0}^{J}\sum_{k=0}^{2^j-1}\hat{d}_{jk}\psi_{jk}(t),\ t\in[0,1], \tag{3.1}$$

where $\hat{f}_{\mathrm{P}}(t)=\sum_{\ell=0}^{d}\hat{\beta}_\ell t^\ell$ is a polynomial estimator of degree d. The use of $\hat{f}_{\mathrm{PW}}$ requires the choice of d as well as the threshold value λ used for estimating the wavelet coefficients in $\hat{f}_{\mathrm{W}}$. With appropriately chosen d and λ, it is demonstrated in Oh, Naveau and Lee (2001), both analytically and empirically, that $\hat{f}_{\mathrm{PW}}$ is superior to $\hat{f}_{\mathrm{W}}$. For this, it is desirable to maintain the orthogonality between the set of polynomial basis $\{t,\ldots,t^d\}$ and the wavelet basis. This means that the equations $\int t^\ell\psi(t)dt=\int t^\ell\phi(t)dt=0$ have to be satisfied for $\ell=1,\ldots,d$. Wavelets with such properties were constructed in Daubechies (1992) and named Coiflets. Hence, the use of Coiflets with at least $d+1$ vanishing moments implies that the polynomial regression term is orthogonal to the wavelet regression term. This orthogonality property allows Oh, Naveau and Lee (2001) to obtain some asymptotic results showing that the PWR estimators are competitive with other nonparametric procedures retaining the asymptotic optimality of the wavelet decomposition and reducing the edge effects.

On the practical side, several automatic methods for choosing both d and λ are proposed by Lee and Oh (2004). They are based on estimating values of d and λ that aim to minimize an estimate of the L^2-risk between f and $\hat{f}_{\mathrm{PW}}$. We only describe here one of the proposed methods. The interested reader is referred to Lee and Oh (2004) for a description of other approaches. To choose d a criterion similar to Mallow's C_p is used with the polynomial estimator $\hat{f}_{\mathrm{P}}$. Then the SUREshrink or the EBayesThresh wavelet regression procedure (e.g., Antoniadis, Bigot and Sapatinas, 2001) is applied to choose the λ that aims to minimize the risk between $f-\hat{f}_{\mathrm{P}}$ and $\hat{f}_{\mathrm{W}}$, where $\hat{f}_{\mathrm{W}}$ is obtained by applying ordinary wavelet regression to the polynomial residuals $Y_i-\hat{f}_{\mathrm{P}}(t_i)$. Note that due to the use of an appropriate Coiflet basis, $\hat{f}_{\mathrm{W}}$ is the same as the wavelet fit obtained on the original observations Y_i.

3.2. Local polynomial wavelet regression (LPWR)

The use of PWR for handling boundary problems works very well when the polynomial fit $\hat{f}_{\mathrm{P}}$ is able remove the non-periodicity in the data, so the remaining

signal can be well estimated using wavelet thresholding with periodic boundary conditions. However, due to the global nature of $\hat{f}_{\mathrm{P}}$, for those cases when f has complex boundary conditions or has some abrupt changes present near the boundaries, PWR does not work well. Lee and Oh (2004) proposed a new method which will also work well under such situations. The basic idea behind the proposed method is to introduce a local polynomial regression component to the wavelet shrinkage. Since local polynomial regression produces excellent boundary handling (Fan, 1992; Hastie and Loader, 1993), it is expected that the addition of this component to wavelet shrinkage will result in equally well boundary properties. Their local polynomial wavelet estimator can be written as

$$\hat{f}_{\mathrm{LPWR}}(t) = \hat{f}_{\mathrm{LP}}(t) + \hat{f}_{\mathrm{W}}(t). \tag{3.2}$$

The LPWR estimator is computed through an iterative algorithm inspired by a backfitting type algorithm. The following steps summarize the key points for finding the final local polynomial wavelet regression estimate, $\hat{f}_{\mathrm{LPWR}}$.

1. Select an initial estimate $\hat{f}^{(0)}$ for f and let $\hat{f}_{\mathrm{LPWR}} = \hat{f}^{(0)}$.
2. For $j = 1,\ 2, \ldots$ iterate the following steps:
   (a) Apply wavelet thresholding to the residuals $Y_i - \hat{f}_{\mathrm{LPWR}}(t_i)$ and obtain $\hat{f}_{\mathrm{W}}^{(j)}(t_i)$.
   (b) Estimate $\hat{f}_{\mathrm{LP}}^{(j)}$ by fitting a local polynomial regression to $Y_i - \hat{f}_{\mathrm{W}}^{(j)}(t_i)$.
3. Stop if $\hat{f}_{\mathrm{LPWR}}(t) = \hat{f}_{\mathrm{LP}}^{(j)}(t) + \hat{f}_{\mathrm{W}}^{(j)}(t)$ converges.

To use the above algorithm, one needs to choose the initial curve estimate $\hat{f}^{(0)}$ in Step 1 and the smoothing parameter for the local polynomial fit in Step 2(b). To do so, Oh and Lee use Friedman's super-smoother (available as `supsmu` in `R`), while for the smoothing parameter for computing the local polynomial estimator, they use cross-validation. The numerical experiences they provide in their simulation study, using their R-code `hybrid.r`, suggest that the above algorithm converges very quickly.

3.3. Proposed adaptive combination of two regression estimators

Inspired by the LPWR strategy, we are now proposing an adaptive combination of local polynomial and wavelet regression estimators with sound asymptotic properties.

The local polynomial estimator is based on stronger assumptions on the regression function than the wavelet regression one and thus, with appropriately chosen hyper-parameters, the asymptotic rates of convergence are different. We will adopt a linear combination of the two estimators where the weights are estimated by Stein's Unbiased Risk estimation (SURE) in such a way that the adaptive estimator retains the optimal rates of convergence. The technique used here is similar to the one adopted by Burman and Chaudhuri (2011) who combined a parametric estimator (rate n^{-1}) with a linear non-parametric estimator

to obtain an estimator with optimal rates. The main difference is that we do not need for one estimator to be parametric, as long as its rate is not faster than n^{-1}, and also in that we combine a linear estimator with a nonlinear one in what follows.

Let us first recall also that for any $s > 0$, the Sobolev space $W_2^s([0,1])$ with non-integer regularity index $s > 1/2$ is defined as the space of tempered distributions whose Fourier transforms are square integrable with respect to the measure $(1+|x|^2)^s$ on $[0,1]$ (Hörmandere, 1989, p. 240). When s is an integer, it coincides with the space of functions f having continuous derivatives of orders up to $s-1$, and a square integrable derivative of order s. By the Sobolev embedding results concerning Besov spaces on the interval, which can be found in, for example, Donoho et al. (1996), we have $B_{2,2}^s[0,1] \subset W_2^s[0,1] \subset W_2^{\lfloor s \rfloor}[0,1]$ for any $s > 1/2$, the inclusion being continuous.

A local polynomial estimator is linear in the data, and when the regression function is assumed to be within a Sobolev space $W_2^s[0,1]$ with $\lfloor s \rfloor \geq 2$, local polynomial regression smoothers, with an optimal choice of kernel and bandwidth, have nice sampling properties and high minimax efficiency. Our attention is focused on a fixed equidistant design. The global L^2-risk of any estimator $\hat{f}$ is then equivalent to the expected empirical risk of the estimator, that is

$$R_n(f,\hat{f}) = \frac{1}{n}\mathbb{E}\left(\sum_{i=1}^{n}(f(t_i)-\hat{f}(t_i))^2\right) = \mathbb{E}\|\boldsymbol{f}-\hat{\boldsymbol{f}}\|_n^2, \tag{3.3}$$

where $\|\cdot\|_n^2 = n^{-1}\|\cdot\|^2$ and $\|\cdot\|$ is the Euclidian norm of $\mathbb{R}^n$. It is then known (e.g., Fan, 1993) that a local polynomial estimator $\hat{f}_{\mathrm{LP}}$ with an Epanechnikov kernel and an optimal bandwidth h_n of the order $n^{-1/(2\lfloor s \rfloor+1)}$ is optimal in terms of rates of convergence with an optimal rate given by $r_1(n) := R_n(f,\hat{f}_{\mathrm{LP}}) = n^{-\frac{2\lfloor s \rfloor}{2\lfloor s \rfloor+1}}$.

If one assumes that the unknown regression function belongs instead to a Besov space $B_{2,2}^s[0,1] \subset W_2^s[0,1]$, then the best optimal rate is $n^{-\frac{2s}{2s+1}}$, which is smaller than $r_1(n)$ since $2\lfloor s \rfloor/(2\lfloor s \rfloor+1) < 2s/(2s+1)$. If $s > \lfloor s \rfloor$ then this rate, up to a logarithmic factor, can be only attained by a wavelet threshold estimator $\hat{f}_{\mathrm{W}}$. In fact if the optimal rate obtained by wavelet thresholding is denoted by $r_2(n) = n^{-\frac{2s}{2s+1}}\log n$, we obviously have $r_2(n)/r_1(n) \to 0$, as $n \to \infty$ which implies that $\|\boldsymbol{f}-\hat{\boldsymbol{f}}_{\mathrm{LP}}\|_n^2 = O_P(r_1(n))$ and $\|\boldsymbol{f}-\hat{\boldsymbol{f}}_{\mathrm{W}}\|_n^2 = O_P(r_2(n))$.

We would like now to use a combination estimator which decides on the basis of data which estimator to use among these two. Define the hybrid estimator

$$\hat{f}_\alpha = \alpha\hat{f}_{\mathrm{LP}} + (1-\alpha)\hat{f}_{\mathrm{W}}, \tag{3.4}$$

which can also be viewed as a smoothed version of a pretest-estimator where we test $f \in B_{2,2}^s[0,1]$ vs. $f \in W_2^{\lfloor s \rfloor}[0,1]\backslash B_{2,2}^s[0,1]$.

In order to focus on the main issue we will assume that the noise level in the regression problem (1.1) is known. Note that such a restriction is not too severe because one may robustly estimate σ using wavelet regression. Let us define

$\delta_n = \inf\{\|\boldsymbol{f} - \tilde{\boldsymbol{f}}\|_n^2; \tilde{f} \in B^s_{2,2}[0,1]\}$ which is attained at some point $g \in B^s_{2,2}[0,1]$ since $B^s_{2,2}[0,1]$ is in particular convex.

3.3.1. The adaptive estimator

The Euclidian empirical distance between the true regression function f and the hybrid estimator $\hat{f}_\alpha$ is $\|\boldsymbol{f} - \alpha\hat{\boldsymbol{f}}_{\mathrm{LP}} + (1-\alpha)\hat{\boldsymbol{f}}_{\mathrm{W}}\|^2$, which is minimized at

$$\alpha^\star = \frac{\langle \hat{\boldsymbol{f}}_{\mathrm{LP}} - \hat{\boldsymbol{f}}_{\mathrm{W}}, \boldsymbol{f} - \hat{\boldsymbol{f}}_{\mathrm{W}} \rangle}{\|\hat{\boldsymbol{f}}_{\mathrm{LP}} - \hat{\boldsymbol{f}}_{\mathrm{W}}\|^2}. \tag{3.5}$$

The approach chosen here is to treat this as a hyper-parameter and estimate it using Stein's unbiased risk estimation which is an unbiased estimate of the loss. By equation (3.3) we have $R_n(f, \hat{f}_\alpha) = \frac{1}{n}\mathbb{E}\left(\sum_{i=1}^n (f(t_i) - \hat{f}_\alpha(t_i))^2\right) = \mathbb{E}\|\boldsymbol{f} - \hat{\boldsymbol{f}}_\alpha\|_n^2 := \mathbb{E}(\mathcal{L}_n(\alpha))$. Now $\mathcal{L}_n(\alpha)$ is equal to

$$\mathcal{L}_n(\alpha) = \frac{1}{n}\sum_{i=1}^n \left(f(t_i)^2 + \hat{f}_\alpha(t_i)^2 - 2f(t_i)\hat{f}_\alpha(t_i)\right).$$

Direct optimization of $R_n(f, \hat{f}_\alpha)$ with respect to α is not feasible since the function f is unknown in the last term of the above expression. To proceed we need to derive an objective that substitutes for $R_n(f, \hat{f}_\alpha)$, and depends only on the noisy data. We now state a version of Stein's lemma with Gaussian errors that is useful in deriving an unbiased estimator of $R_n(f, \hat{f}_\alpha)$. A proof may be found for example in Blu and Luisier (2007).

Lemma 3.1. *Suppose that for all* $i = 1, \ldots, n$, $\mathbb{E}\left(\left|\frac{\partial \hat{f}_\alpha(t_i)}{\partial Y_i}\right|\right) < \infty$. *Then*

$$\mathbb{E}\left(\hat{f}_\alpha(t_i) f(t_i)\right) = \mathbb{E}\left(\hat{f}_\alpha(t_i) Y_i - \sigma^2 \frac{\partial \hat{f}_\alpha(t_i)}{\partial Y_i}\right).$$

Once the regression estimates are computed for some fixed bandwidth h_h and some fixed threshold λ_n of the right order, the optimum estimator $\hat{\alpha}$ is computed by minimizing an unbiased estimator of $\mathbb{E}(\mathcal{L}_n(\alpha))$ derived as follows using Lemma 3.1:

$$\widehat{\mathbb{E}(\mathcal{L}_n(\alpha))} = \frac{1}{n}\sum_{i=1}^n \left(Y_i^2 + \hat{f}_\alpha(t_i)^2 - 2Y_i\hat{f}_\alpha(t_i) + 2\sigma^2 \frac{\partial \hat{f}_\alpha(t_i)}{\partial Y_i}\right) - \sigma^2.$$

Both $\frac{\partial \hat{f}_{\mathrm{LP}}(t_i)}{\partial Y_i}$ and $\frac{\partial \hat{f}_{\mathrm{W}}(t_i)}{\partial Y_i}$ can be easily computed using results in Blu and Luisier (2007). We can now state the following theorem that outlines the asymptotic behavior of our hybrid estimator:

Theorem 3.1. *Suppose that* δ_n *tends to zero as* n *tends to infinty. Then*

- $\mathcal{L}_n(\alpha^\star) = \begin{cases} O_P(r_1(n)), & \text{if } \delta_n > r_1(n) \\ O_P(\delta_n), & \text{if } r_2(n) \le \delta_n \le r_1(n) \\ O_P(r_2(n)), & \text{if } \delta_n < r_2(n). \end{cases}$
- *if* $r_2(n)$ *is slower than* n^{-1}*, then* $\|\boldsymbol{f} - \hat{\boldsymbol{f}}_{\hat{\alpha}}\|_n^2 = \|\boldsymbol{f} - \hat{\boldsymbol{f}}_{\alpha^\star}\|_n^2(1 + o_P(1))$.

The proof relies upon some Lemmas very similar to Lemmas 5.1, 5.2 and 5.4 in Burman and Chaudhuri (2011) and it is given in the Appendix.

Remark 3.1. *The simple combination approach described above, even when using the SURE principle to estimate the mixing coefficient, may produce poor results in practice because it does not take into account the correlation induced by the fact that both nonparametric estimators are estimated on the same data set. Of course this poor behavior could be enhanced by using wavelets that are orthogonal to polynomials, like Coiflets. If one restricts attention to combinations that are linear in the estimators, computing the weights that mimic the least squares oracle weights may also be problematic in the non asymptotic sense since the resulting estimated coefficients, even when using the SURE principle, may overfit the data and therefore present a generalization error that can be poor. To attenuate this correlation problem one could use* stacking *regression for combining the estimates. Stacking was first presented by Wolpert (1992), who considered "neural networks", and extended to statistics by Breiman (1996), who considered "stacked regression" and provided some heuristic and numerical results to justify a method for combining estimators without suffering of the correlation problem and the generalization error. To avoid overfitting the weights are computed by minimizing a cross-validated squared error loss, where each estimator to be combined is estimated on the training data and the prediction errors are computed on the leaved-out test data. While this is not a problem for local polynomial estimators, it may be a problem when considering wavelet regression on nonequidistant data, since the DWT heavily relies upon equidistant data. However this idea will be pursued later in this paper, when considering regression on non-equidistant or random designs data.*

3.4. Some other methods: trend filtering

We give here a brief background of ℓ_1 trend filtering, presented as a nonparametric regression method that uses an ℓ_1-type penalty, and which is able to adapt to local differences in smoothness and achieve the minimax rate of convergence for weakly differentiable regression functions of bounded variation (Tibshirani, 2014). An implicit assumption with trend filtering is that the design points are evenly spaced. Assuming that $t_1 < t_2 < \cdots < t_n$ are unique and equally spaced, for a given integer $k \ge 0$, the kth order trend filtering estimate $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_n)$ of $(f(t_1), \ldots, f(t_n))$ is defined by a penalized least squares optimization problem

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^n}{\arg\min} \frac{1}{2}\|\boldsymbol{Y} - \beta\|_2^2 + \lambda \|D^{(k+1)}\boldsymbol{\beta}\|_1, \tag{3.6}$$

where $D^{(k+1)} \in \mathbb{R}^{(n-k-1)\times n}$ is the $k+1$ order finite difference matrix and $\lambda \ge 0$ is a tuning parameter. The constant factor $n^k/k!$ multiplying λ is unimpor-

tant, and can be absorbed into the tuning parameter λ, but it will facilitate comparisons in future sections.

When $k = 0$,

$$D^{(1)} = \begin{bmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{bmatrix} \in \mathbb{R}^{(n-1)\times n}, \tag{3.7}$$

and so $\|D^{(1)}\boldsymbol{\beta}\|_1 = \sum_{i=1}^{n-1} |\beta_i - \beta_{i+1}|$. Hence, the constant trend filtering is the same as one-dimensional total variation denoising (e.g., Rudin, Osher and Fatemi, 1992). When $k = 1$

$$D^{(2)} = \begin{bmatrix} 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \end{bmatrix} \in \mathbb{R}^{(n-2)\times n}. \tag{3.8}$$

In general, as described by Tibshirani (2014), $D^{(k+1)} = D^{(1)}D^{(k)}$.

Apart from requiring unique and equally spaced observations, (3.6) has one parameter per data point, no intercept, and the design matrix is the identity matrix. For a general k, the kth order trend filtering estimate has the structure of a kth order piecewise polynomial function, evaluated across the inputs $t_1, \ldots, t_n$. The knots in this piecewise polynomial are selected adaptively among $t_1, \ldots, t_n$, with a higher value of the tuning parameter λ (generally) corresponding to fewer knots. Taking a λ of the order $n^{1/(2k+3)}$ leads to an asymptotic rate faster than the minimax rate over Sobolev spaces (Tibshirani, 2014, Theorem 1).

4. Proposals for handling the boundary problem in the general case

4.1. Spline-wavelet adaptive combination

For a fixed nonequidistant design or even a random design one may still use the simple adaptive combination approach described in Subsection 3.3 applying the SURE principle to estimate the mixing coefficients with similar asymptotic results, since the rates of each estimator, based on splines or wavelet expansions (Subsection 2.3), are similar to those evoked for local polynomial estimators and wavelet thresholding estimators. Moreover, since no restriction is required on the design, "stacked regression" is also a possible option for combining spline and wavelet regression estimators without suffering of the correlation problem between the two estimates.

4.2. Spline-wavelet stacking

We will now describe with some extra details how stacked regression is performed with the two basis expansion estimators. To simplify notation, we will call $\hat{f}_1$ the optimal spline based estimator and $\hat{f}_2$ the optimal wavelet expansion estimator. Optimality here is regarded in an asymptotic sense, that is the regularization parameters for the splines penalty and for wavelet penalization are fixed but of the right asymptotic order. In practice however these hyperparameters are chosen in a data dependent way when computing the estimates. Stacked regression combines linearly the two estimators $\hat{f}_j$, $j = 1, 2$ to obtain $\hat{f}_{\text{stack}}$ given by

$$\hat{f}_{\text{stack}}(t) = \hat{\beta}_1 \hat{f}_1(t) + \hat{\beta}_2 \hat{f}_2(t), \quad t \in [0, 1], \tag{4.1}$$

where the estimator $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2)$ of the parameters in eq. (4.1) is obtained as

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^2}{\arg\min} \sum_{i=1}^{n} \left(Y_i - \beta_1 \hat{f}_1^{(-i)}(t_i) - \beta_2 \hat{f}_2^{(-i)}(t_i) \right)^2,$$

with $\hat{f}_j^{(-i)}(t_i)$ the leave-one-out estimate of the j type estimator at the design point t_i.

By using the cross-validated predictions stacked regression avoids giving unfairly high weight to models with higher complexity. There is a close connection between stacking and winner-takes-all model selection. If we restrict the minimization to weight vectors that have one unit weight and the rest zero, this leads to the model choice returned by the winner-takes-all based on the leave-one-out. Rather than choose a single model, stacking combines them with estimated optimal weights. This will often lead to better prediction, but less interpretability than the choice of only one of the models.

4.3. Matrix stacking regression of spline and wavelet bases

An alternative way of mixing wavelet and spline bases estimators is to stack the corresponding matrices of the bases.

Let $\boldsymbol{S}$ be the matrix of the spline basis and $\boldsymbol{W}$ the corresponding one for the wavelet basis (Subsection 2.3). Then the regression model writes up as

$$\boldsymbol{Y} \approx \alpha + \boldsymbol{S}\boldsymbol{\gamma}_S + \boldsymbol{W}\boldsymbol{\gamma}_W + \boldsymbol{\epsilon} = \boldsymbol{S}\boldsymbol{W}\boldsymbol{\gamma} + \boldsymbol{\epsilon},$$

with $\boldsymbol{S}\boldsymbol{W} = \begin{bmatrix} \mathbf{1}_N & \boldsymbol{S} & \boldsymbol{W} \end{bmatrix}$ being the matrix stacking $\boldsymbol{S}$ and $\boldsymbol{W}$ (and the intercept term) and $\boldsymbol{\gamma} = (\alpha, \boldsymbol{\gamma}_S, \boldsymbol{\gamma}_W)$ the set of intercept and spline and wavelet coefficients, respectively, to be estimated. As before a penalization term can be included in the regression resulting in the minimization of the following functional:

$$\min_{\boldsymbol{\gamma}} \frac{1}{2} \|\boldsymbol{Y} - \boldsymbol{S}\boldsymbol{W}\boldsymbol{\gamma}\| + \sum_{k=2}^{K_S + K_W + 1} P_\lambda(|\gamma_k|),$$

with K_S and K_W being the number of spline and wavelet coefficients, respectively. Inclusion of the spline term is intended to improve accuracy at boundaries.

4.4. Adaptive regression mixing and aggregation

In applications many more nonparametric regression procedures have been developed in the literature and studied both theoretically and through systematic numerical investigations. Since the model noise level and the true regression function are unknown, the task of identifying the best among several estimation procedures is typically very difficult. Therefore, there is an advantage if one considers a list of distinct nonparametric regression procedures so that at least one of them is optimal or well-behaving for the unknown underlying data generating process. For the goal of estimating the regression function, as is the focus in this paper, one approach is to combine the various estimation procedures by a proper weighting of estimates from them. An example of such an approach, when attention is restricted to estimators based on regularized linear expansions of the regression function in either spline bases or wavelet bases, is stacking regression proposed in the previous subsection. If the combination leads to a performance similar or close to the best method in each scenario of the underlying data generation process, the combined estimator or prediction can outperform all the candidate procedures in repeated applications across different scenarios of the data generation process.

Combining regression procedures has been studied and allows to prove various interesting theoretical properties. Oracle inequalities show that properly combining arbitrary regression procedures leads to a risk close to the best among a target class of combinations of the candidate estimators/predictions plus a minimax-rate optimal "price of combining" that reflects the largeness of the class of allowed combinations; see Chen and Yang (2010) for a literature review. Successes of combining different predictions in applications have prompted more interest.

It is not our purpose here to combine all existing nonparametric procedures. We will therefore restrict our attention to the class of regularized linear expansions of the regression function in either spline bases or wavelet bases, and two supplemental methods developed recently in the machine learning literature that may handle regression function with heterogeneous smoothness, namely spatially adaptive regression splines with accurate knot selection schemes and ℓ_1 trend filtering. After briefly describing each of these procedures in the following, we will combine them using an aggregation method developed by Yang (2001) and called Adaptive Regression by Mixing (ARM). Results from Yang (2001) and Catoni (2004) show that the combined regression estimator achieves the best performance offered by the candidates in an accumulated risk.

Spatially adaptive regression splines (SARS) Usually, good approximations of inhomogeneous functions by spline functions require a set of highly unevenly distributed knots. Zhou and Shen (2001) proposed an adaptive spline procedure based on a new knot selection scheme for nonparametric regression. The proposed procedure uses certain special local properties of spline function in knot selection and thus overcomes the knot confounding problem and high computational complexity in adaptive estimation encountered by spline proce-

dures based on the traditional knot addition and deletion method. To improve computational efficiency of the knot selection, for a spatially inhomogeneous regression function that is smooth in one region and nonsmooth in another region (for example the boundaries or the vicinity of change-points), one needs to insert more knots in areas where the regression function is less smooth. To achieve this task the authors use of a guided knot search, which makes the search more efficient. We will not describe the method further here but the interested reader is referred to Zhou and Shen (2001) for details and properties of their estimates. We are grateful to the authors for providing an implementation in `R` code of their algorithm (SARS).

ℓ_1 trend filtering When we reviewed trend filtering we have assumed that the design locations are implicitly evenly spaced; Ramdas and Tibshirani (2016) developed an algorithm to extend ℓ_1 trend filtering to irregularly spaced data using a specialized ADMM algorithm. Fortuitously, there is little that needs to be changed with the equidistant trend filtering problem when one moves from equispaced design points to arbitrary ones; the only difference is that the operator $D^{(k+1)}$ is replaced by $D^{(t,k+1)}$, which is adjusted for the uneven spacings present in the design. These adjusted difference operators are still banded with the same structure, and are still defined recursively. Under appropriate regularity conditions on the design, the resulting ℓ_1 trend filtering retains the same asymptotical optimality as for the equidistant design case.

We focus now on the adaptive regression by mixing (ARM). When the noise is normally distributed, the ARM uses least squares as a discrepancy measure in the core step to apportion the weights to each candidate, and leads to good theoretical results (Yang, 2001). We apply K nonparametric regression procedures on the data: procedure j yields an estimator $\hat{f}_j$. Denote the set of the K candidate methods by Γ. For simplicity, assume that n is even and that the data are sorted. The ARM algorithm is:

Step 1. Split the data into two parts $Z^{(1)} = \{(T_i, Y_i), i = 1, 3, \ldots, n-1\}$ and $Z^{(2)} = \{(T_i, Y_i), i = 2, 4, \ldots, n\}$.

Step 2. Based on $Z^{(1)}$, apply all the estimation procedures in Γ to get the regression estimates $\hat{f}_j$, $j = 1, \ldots, K$, and compute the mean squared error $\hat{d}_j = \frac{2}{n}\sum_{Z^{(1)}}(Y_i - \hat{f}_j(T_i))^2$ for each candidate procedure j.

Step 3. For each procedure $j \in \Gamma$, predict Y_i by $\hat{f}_j(T_i)$ for $Z^{(2)}$ and compute the overall measure of discrepancy $\hat{D}_j = \sum_{Z^{(2)}}(Y_i - \hat{f}_j(T_i))^2$.

Step 4. Compute the weight for procedure j as

$$\hat{W}_j = \frac{\hat{d}_j^{-n/2}\exp(-\hat{D}_j/\hat{d}_j)}{\sum_{k\in\Gamma}\hat{d}_k^{-n/2}\exp(-\hat{D}_k/\hat{d}_k)}. \tag{4.2}$$

Step 5. Let

$$\hat{f}_{ARM} = \sum_{j\in\Gamma}\hat{W}_j\hat{f}_j.$$

Assuming that K is finite and fixed and that each $\hat{f}_j$ is asymptotically optimal within the class of functions it is designed to estimate, the resulting estimate is asymptotically as good as the best estimate in Γ (Yang, 2003, Theorem 1). Such a result justifies therefore to consider such an estimator for correcting the boundary problem.

4.5. Greedy pursuit and best basis selection from multiple libraries

In this subsection we handle the boundary problems by approximating the regression function f by a finite linear combination of elements of a given dictionary D of functions. Here, by dictionary, we mean a union of several bases or collections of general waveforms from $L^2([0,1])$. One of the motivations for utilizing general overcomplete dictionaries rather than orthonormal systems is that it is not clear which orthonormal system, if any, is best for representing or approximating f. Thus, dictionaries are preferred and to manage the number of computations matching pursuits algorithms will be used with the aim to build "suboptimal yet good" finite approximations through a greedy selection of elements within the dictionary D.

We shall focus our attention to the BSML procedure recently proposed in Sklar et al. (2013). It selects basis functions adaptively from the union of multiple libraries, where each library consists of basis functions with similar forms and properties. Compared to using a single library, the advantage of using multiple libraries is that only relatively few basis functions need to be selected from each library to approximate the target function, particularly if the target function is spatially inhomogeneous and if the basis functions in different libraries capture different inhomogeneous features found in the true function. There are infinitely many choices of the libraries. Libraries may be selected from different families including Fourier, spline, radial, wavelet bases and so on. They may also be selected from different types within a family. We can have B-splines, truncated polynomials and reproducing kernel representers for the spline family. Within each type, we can specify different orders of basis, e.g., linear or cubic for polynomial splines. Wavelets of different types for wavelet families. Within each type, we can specify different filter numbers, resulting to more or less smooth wavelets.

To briefly describe the procedure, let us say that BSML starts with a null library $\mathcal{L}_0$, which contains all the basis functions that will be included in the model automatically. Let $m = |\mathcal{L}_0|$ and M be a pre-specified maximum number of basis functions we want to select (including those in $\mathcal{L}_0$). The number M is closeley related to the best rate of "suboptimal yet good" finite approximation through the greedy selection of elements within the dictionary union of multiple libraries. Basis functions are selected from L additional libraries that define the dictionary D one at a time. At each step k, denote the sequentially selected basis functions as b_k for $k = m+1, \ldots, M$. Let $\mathcal{B}_k = \{b_1, \ldots, b_k\}$ for $k = m, \ldots, M$, where $\mathcal{B}_m = \mathcal{L}_0$. Write "model $\mathcal{B}_k$" for "a linear combination of the basis functions in $\mathcal{B}_k$", and also $\mathcal{M}_k$ for the modeling procedure that includes both basis

functions selection and estimation steps. The BSML procedures utilize the generalized degrees of freedom (GDF) to measure the complexity of a modeling procedure. This approach can be computationally demanding, since the GDF needs to be estimated at every step of the forward selection process. At each forward selection step, a greedy search for only one basis function to add to the current model $\mathcal{M}_k$ is performed. For more details on the procedures the interested reader is refereed to the paper by Sklar et al. (2013).

To explore numerically the advantage of multiple libraries of basis functions using advanced model selection methods for treating the boundary problems in nonparametric regression for functions of heterogeneous smoothness, we have used a collection of `R` functions available in the `R` package `bsml` whose source can be downloaded from Yuedong Wang's web site.

4.6. Gaussian processes and stochastic partial differential equations

As suggested by one of the reviewers of our work, another possibility to derive flexible models which are practical to work with for estimating functions from noisy data are Gaussian processes (GP) and stochastic partial differential equations (SPDE). We would like to therefore add in this subsection some minor additional discussion concerning these two approaches.

A basic idea on how Gaussian Process models can be used for such a task is by formulating a Bayesian framework for regression. In order for the GP techniques to be of value in practice, one must chose between different mean and covariance functions in the light of the data at hand, reflecting any prior knowledge about the structure of the function of interest (a process that is referred to as *training* the GP model). The interested reader is referred to the book by Rasmussen and Williams (2005) for a comprehensive exposition to Gaussian Process regression models. These models can also be extended to handle piecewise-smooth functions with boundary constraints by adapting them for smoothing in the presence of change-points, which may be seen as more or less abrupt changes to the properties of the observed data. The paper by Osborne, Garnett and Roberts (2010) describes prior covariance functions for one-dimensional regression that model change-points and faults of many different types and gives a Bayesian solution to the smoothing of data from sources that may contain change-points and also some `MATLAB` code. We prefer to not pursue this approach here since training a GP model involves both model selection, or the discrete choice between different functional forms for mean and covariance functions as well as adaptation and estimation of the hyper-parameters of these functions which could be a disadvantage compared to the nonparametric methods discussed in our paper.

Concerning the second approach, the SPDE approach introduced by Lindgren and Rue (2008) and implemented in the `R-INLA` software package (Rue, Martino and Chopin, 2009) is also a flexible and efficient method to analyse data exhibiting complicated boundary constraints. Basically, the SPDE involves applying a differential operator D to a stochastic process, representing the structured dependence among observations, but this cannot be done in the same way as when

one applies D to a known function because the process is random and, in many cases, realizations of it will not be suitably differentiable. Moreover, available software implementations are difficult to customize without high-level technical knowledge, limiting application to only those models available in the software or specially requested from software developers. We may therefore prefer a presentation to SPDE smoothing that explicitly draw links with basis-penalty smoothing approaches. There is a direct correspondence between smoothing and stochastic processes, and works by Fahrmeir and Lang (2001), Lindgren and Rue (2008) and Yue et al. (2014) show how basis-penalty smoothers in a Bayesian framework can be interpreted within the SPDE paradigm. For practical purposes, one may use the results of Miller, Glennie and Seaton (2019) and their R-code, to better understand the implementation and theory behind the SPDE approach. However we believe that the results obtained using such an SPDE approach to smoothing data with typical boundary constrains addressed in our paper compare similarly.

5. Multidimensional problems

The present paper specifically addresses unidimensional functions. However many applications involve multidimensional problems, so that it is interesting to briefly consider a possible generalization to this setting.

It is out of the scope of the paper to present a full theoretical treatment and extensive experimentation. However we observe that all methods can be easily plugged in an additive framework by relying on a backfitting procedure. On the other side one of them is naturally suited for a multidimensional setting without the necessity of resorting a backfitting iteration. It is the case of matrix stacking regression (Section 4.3), where, analogously to the procedure of stacking matrices of different bases for the same dimension of data, these stacked matrices are further stacked across all dimensions, giving rise to a simultaneous estimate of coefficients of the series expansion across all dimensions. In this respect we observe that grouped penalization can also be invoked to achieve selection of dimensions and/or regression methods.

6. Numerical experiments

The present Section introduces numerical experiments worked out on some test functions by a selection of methods considered in the previous Section or available from the literature.

6.1. Methods

Throughout the paper, the following methods have been considered for comparison (see Tab. 1 for a summary):

SPL (Splines): the Spline expansion method discussed in Subsection 2.3 with the Minimax Concave Penalty function (MCP) and regularization parameter λ chosen by GCV. The degree of the Spline basis is 3 and the number of internal knots 11 (total knots 19). Splines are claimed to be accurate at boundaries.

WAV (Wavelets): the Wavelet expansion method discussed in Subsection 2.3 with the MCP and regularization parameter λ chosen by GCV. The Wavelet is generated from Daubechies Extremal Phase wavelets with 5 vanishing moments and maximum level 4. Due to periodicity of wavelets, the basis is not effective at boundaries for nonperiodic functions.

CDV (CDV Wavelets): the Wavelet expansion method discussed in Subsection 2.3 with the MCP and regularization parameter λ chosen by GCV, considering the wavelet basis constructed on the finite interval wavelets of Section 2.2. In this way the method is suited also for generally non-equispaced and non-dyadic grids. We set the filter number to 3 and the highest level to 4.

LPWR (Local Polynomial Wavelet Regression): the method presented in Subsection 3.2. It is claimed to take account of boundaries and therefore to improve accuracy there. LPWR works only for an equispaced and dyadic grid.

WHYBRID (Hybrid LPWR): it is an adapted version of LPWR (Subsection 3.2) where an MCP penalization term is considered for the regression of the residuals instead of wavelet thresholding. In this way the method is suited for not necessarily equispaced and dyadic grids. Moreover it is developed to improve accuracy of the solution at the boundaries.

WMESH (waveMesh; Haris, Simon and Shoiaje, 2018): it is based on a wavelet decomposition on a generally non-equispaced grid based on a linear interpolation scheme from wavelets on a regular grid to the actual data grid and a penalization functional given by l_1 norm; in addition a proximal gradient descent algorithm developed in Parikh (2014) is used to solve the corresponding optimization problem. The method is suited for generic non-equispaced grids and has no special treatment of boundary conditions.

TREND (Trend Filtering): the method described in Subsection 3.4 in its version adapted to generic non-equispaced grid. No special treatment of boundaries is present.

AC (Adaptive Combination): the method presented in Subsection 4.1 with basic regressors given by SARS, Trend Filtering and Wavelet series. By generalizing Eq. (3.4) the AC combination is obtained by unconstrained OLS with respect to weights w_1, w_2, w_3 in $\|\boldsymbol{Y} - w_1\hat{\boldsymbol{f}}_{\text{TREND}} - w_2\hat{\boldsymbol{f}}_{\text{W}} - w_3\hat{\boldsymbol{f}}_{\text{SARS}}\|_n^2$, with $\hat{\boldsymbol{f}}_{\text{TREND}}$, $\hat{\boldsymbol{f}}_{\text{W}}$ and $\hat{\boldsymbol{f}}_{\text{SARS}}$ being solutions obtained by Trend Filtering, Wavelet series and SARS, respectively. Inclusion of SARS among the regression methods is prone to achieve a better behavior at boundaries. The degree of the spline basis is 3. The wavelet basis is given by Coiflets with 3 vanishing moments and a maximum level of the transform of $\lceil \log_2 n \rceil - 2$; MCP is used for penalization.

TABLE 1

Methods considered for the numerical experiments. Column B indicates whether the method is specifically suited to improve accuracy at boundaries and column M if the method is naturally multidimensional (Y) or is strictly unidimensional (N) and can be applied to multidimensional problems only through a backfitting procedure in an additive regression setting. Finally the last column shows the source of the computational code.

Method	*Description*	*B*	*M*	*Software*
SPL	Spline basis	Y	N	Code by authors
WAV	Wavelet basis	N	N	Code by authors
CDV	Wavelets on the interval	Y	N	WAVELAB (Matlab)
LPWR	Local Polynomial Wavelet Regression	Y	N	code `hybrid.R`
WHYBRID	Hybrid LPWR	Y	N	Code by authors
WMESH	Wavelet basis and proximal gradient descent	N	N	code `waveMesh.R`
TREND	Trend Filtering	N	N	`R` package `glmgen`
AC	Adaptive Combination of Trend Filtering, SARS and Wavelet regression	Y	N	Code by authors
SSW	Stacked Splines and Wavelets	Y	N	Code by authors
MSR	Matrix Stacking Regression	Y	Y	Code by authors
ARM	Adaptive Regression by Mixing	N	N	`R` package `MuMIn`
BSML	Basis Selection Multiple Libraries	Y	N	`R` package `bsml`

SSW (Stacked Splines and Wavelets): the method described in Subsection 4.2. It works for generic grids. Spline and wavelet solutions are mixed, the former being included to improve accuracy at boundaries, see Eq. (4.1). The number of knots for the spline basis is $\lceil N/4 \rceil$ with lower and upper bounds 5 and 35, respectively. The wavelet family is Coiflet with 5 vanishing moments and 2 levels of the transform. MCP is used for penalization.

MSR (Matrix Stacking Regression with Splines and Wavelets): the method presented in Subsection 4.3, suited for generic grids and aimed at improving accuracy of the solution at the boundaries thanks to the inclusion of splines. The same hyperparameters as SPL and WAV are used for the spline and wavelet bases, respectively.

ARM (Adaptive Regression by Mixing): the method described in Subsection 4.4. We weight the solutions obtained by spline and wavelet basis, the former intended to improve accuracy at boundaries.

BSML (Basis Selection Multiple Libraries): the method discussed in Section 4.5. It is suited for a generic grid; inclusion of a spline library is intended to improve accuracy at the boundaries.

Several other methods and/or different variants have been included in the comparisons, especially mixing different regressors. They are not reported here because of poorer results. R-codes implementing the above methods as well as the simulations and the scripts to produce plots and tables are available as Supplementary Material.

6.2. Test functions

We consider three synthetic test functions representative of typical regularity and/or boundary conditions to have full control of the accuracy of results. All

functions have support in $[0,1]$ for simplicity. The data model is $y(x_i) = f(x_i) + \sigma\varepsilon_i$, $i = 1, \ldots, n$, $x_i \in [0,1]$, where ε_i are i.i.d. Gaussian and σ is such that the Signal to Noise Ratio (SNR) is 3.5, with SNR being defined as $\text{SNR} = \text{var}(y)/\sigma^2$. The test functions are defined as:

Sin: $f(x) = \sin(11x)$. It is a regular function discontinuous at boundaries ($|f(0) - f(1)| \approx 1$).

Irregular: It is a nonsmooth function with three discontinuities and 2 bumps:

$$f(x) = \sqrt{x(1-x)}\sin\left(\frac{1.6\pi}{x+0.2}\right) + 0.4\,\text{sgn}\,(x-0.13) + \\ + 0.7\,\text{sgn}\,(x-0.32)\,\text{sgn}(0.38-x) + \sum_{i=1}^{2} b_i \left(1 - \left|\frac{x-l_i}{w_i}\right|\right)_+,$$

with $b_1 = 0.43$, $b_2 = 0.42$, $l_1 = 0.65$, $l_2 = 0.91$, $w_1 = 0.03$, $w_2 = 0.015$. The function assumes different values at the boundaries ($|f(0) - f(1)| \approx 7$).

Heavisine: It was introduced in Buckheit and Donoho (1995) as a basic smooth sin function with two discontinuities at $x = 0.3$ and $x = 0.72$ but with coinciding values at the boundaries.

All the functions are normalized to have standard deviation 7.

We consider two different grid designs: equispaced and random, the latter generated by a uniform distribution in $[0,1]$. Finally we choose two different sample sizes to represent small and medium size ($n = 128$ and $n = 512$[1]).

In order to estimate accuracy of the different methodologies we rely on the Root Mean Square Error (RMSE) defined as

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}(x_i) - f(x_i)\right)^2}, \tag{6.1}$$

with $\hat{y}$ being the estimate obtained by any method.

To evaluate in particular accuracy at the boundaries, we also computed a specific RMSE at the boundaries (bRMSE) restricted to grid points close to the boundaries by a maximum distance $\delta = 0.1$.

Experiments were replicated 100 times and RMSE and bRMSE were averaged over the replicates.

6.3. Results

Tables 2 and 3, referring to the total error from experiments for the three test functions and the two types of grid with size 128 and 512, respectively, show that methods adapted to boundaries do improve overall accuracy of estimate.

We observe that LPWR was not included among competitors in the case of random grid. Moreover WHYBRID shows some instabilities for the random

[1]The choice of dyadic sample size is due to the fact that LPWR is subject to this constraint, all other methods being able to deal with any number of data points.

grids (the more with the larger sample size) that degrade performance; therefore it was not included in boxplots because of the corresponding high value outliers for $n = 512$. On some occasions iterations diverge.

TABLE 2

Total RMSE of the selected methods for the three chosen test functions (Sin, Irregular and Heavisine) and equispaced and random (drawn from a uniform distribution) grid of size 128. Results refer to an average of 100 replicates. Best values for each case are highlighted in bold.

	Equispaced grid			*Random uniform grid*		
	Sin	*Irregular*	*Heavisine*	*Sin*	*Irregular*	*Heavisine*
SPL	1.37	2.69	1.40	1.34	2.59	1.40
WAV	1.58	1.92	1.40	1.52	1.83	1.37
CDV	2.06	3.39	2.30	1.97	3.15	2.16
LPWR	1.09	2.26	1.27	—	—	—
WHYBRID	1.29	1.93	1.29	2.25	1.91	1.34
WMESH	1.73	2.53	2.00	1.76	2.13	1.91
TREND	3.69	3.16	3.68	1.82	2.21	1.90
AC	3.73	3.18	3.72	2.00	2.28	2.14
SSW	**1.04**	**1.87**	**1.20**	**1.01**	**1.77**	**1.22**
MSR	1.54	1.88	1.40	1.53	1.88	1.40
ARM	1.05	1.89	1.21	**1.01**	1.79	**1.22**
BSML	1.37	2.19	1.77	1.43	2.06	1.69

TABLE 3

Total RMSE of the selected methods for the three chosen test functions (Sin, Irregular and Heavisine) and equispaced and random (drawn from a uniform distribution) grid of size 512. Results refer to an average of 100 replicates. Best values for each case are highlighted in bold.

	Equispaced grid			*Random uniform grid*		
	Sin	*Irregular*	*Heavisine*	*Sin*	*Irregular*	*Heavisine*
SPL	0.66	2.49	0.87	0.65	2.45	0.86
WAV	0.96	**1.32**	1.09	0.92	1.28	1.06
CDV	1.69	3.19	1.94	1.65	3.13	1.89
LPWR	**0.52**	1.39	0.86	—	—	—
WHYBRID	0.74	1.35	0.85	2.15	1.31	0.95
WMESH	0.82	1.40	1.07	0.83	1.38	1.08
TREND	0.56	2.55	0.85	0.55	1.25	0.83
AC	0.75	2.83	0.99	0.72	**1.22**	1.02
SSW	0.54	1.46	**0.83**	**0.53**	1.41	**0.82**
MSR	0.76	1.42	0.92	0.74	1.37	0.93
ARM	0.54	1.46	**0.83**	**0.53**	1.41	**0.82**
BSML	0.77	1.51	1.13	0.78	1.44	1.11

Figures 1–2 present the corresponding boxplots that also show the variation of the performance over the repetitions and the outliers for $n = 128$ and equispaced grid and $n = 512$ and random uniform grid, respectively.

SSW is a clear winner among the methods because it is ranked first in 9 cases out of 12. The closest runner is ARM, which is ranked among the best three methods 5 times out of 12. In this respect we mention that accuracy of ARM is very close to SSW in general. Actually both methods share common steps: both are based on a mix of regressors (chosen the same); however ARM adopts a 2-fold Cross Validation (without repetitions) to train the regression, while

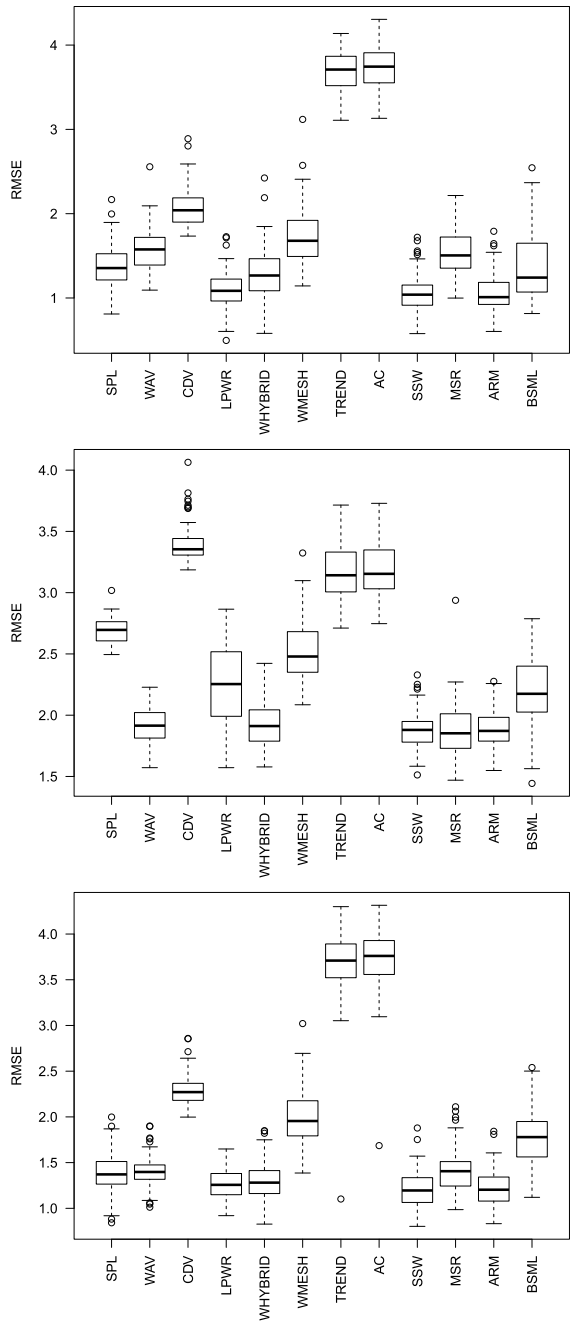

FIG 1. *RMSE for the Sin (upper panel), Irregular (middle) and Heavisine (lower) test functions in the case of equispaced grid of size 128.*

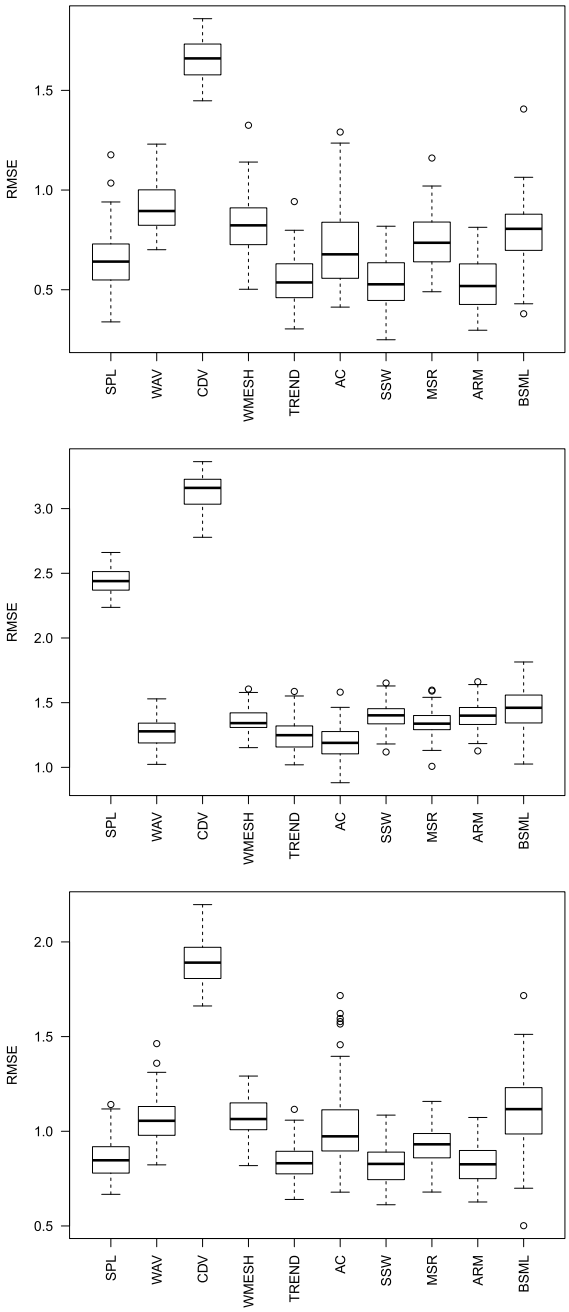

FIG 2. *RMSE for the Sin (upper panel), Irregular (middle) and Heavisine (lower) test functions in the case of random uniform grid of size 512.*

SSW relies on a full leave-one-out procedure. Moreover the mixing coefficients are estimated by (constrained) linear regression in SSW, while a closed formula (Eq. 4.2) is adopted for ARM.

WHYBRID and MSR also show quite good performances in general. All of them outperform Splines and Wavelets, that can be considered as the prototypes of methods for nonperiodic smooth and periodic irregular functions, respectively.

This is confirmed also when error at boundaries is considered (bRMSE, Tabs. 4 and 5).

TABLE 4

Boundary bRMSE of the selected methods for the three chosen test functions (Sin, Irregular and Heavisine) and equispaced and random (drawn from a uniform distribution) grid of size 128. Results refer to an average of 100 replicates. Best values for each case are highlighted in bold.

	Equispaced grid			*Random uniform grid*		
	Sin	*Irregular*	*Heavisine*	*Sin*	*Irregular*	*Heavisine*
SPL	1.59	2.48	1.45	1.52	2.41	1.48
WAV	2.24	2.08	1.11	1.94	1.91	1.22
CDV	2.54	3.35	2.79	2.56	3.20	2.72
LPWR	1.42	2.29	**1.09**	—	—	—
WHYBRID	1.37	1.86	1.22	2.89	1.78	1.17
WMESH	2.07	2.63	2.17	1.76	2.06	1.80
TREND	3.71	3.07	3.54	1.96	2.28	1.90
AC	3.74	3.08	3.57	2.23	2.36	2.31
SSW	**1.25**	**1.74**	1.12	**1.20**	**1.74**	**1.21**
MSR	1.74	1.82	1.18	1.67	1.87	1.41
ARM	1.29	1.78	1.17	1.24	1.79	1.26
BSML	1.55	2.15	1.55	1.64	2.13	1.57

TABLE 5

Boundary bRMSE of the selected methods for the three chosen test functions (Sin, Irregular and Heavisine) and equispaced and random (drawn from a uniform distribution) grid of size 512. Results refer to an average of 100 replicates. Best values for each case are highlighted in bold.

	Equispaced grid			*Random uniform grid*		
	Sin	*Irregular*	*Heavisine*	*Sin*	*Irregular*	*Heavisine*
SPL	0.83	2.05	0.87	0.79	1.94	0.78
WAV	1.56	1.49	1.09	1.47	1.36	0.80
CDV	2.28	3.33	1.94	2.30	3.23	2.57
LPWR	**0.66**	**1.28**	0.86	—	—	—
WHYBRID	0.73	1.35	0.85	3.62	1.39	0.75
WMESH	0.92	1.37	1.07	0.9	1.31	0.98
TREND	0.68	2.53	0.85	0.64	**1.18**	**0.66**
AC	0.96	2.81	0.99	0.90	1.21	0.98
SSW	0.69	**1.28**	**0.83**	**0.63**	1.24	0.69
MSR	0.88	1.34	0.92	0.84	1.28	0.79
ARM	0.70	1.29	**0.83**	0.65	1.24	0.69
BSML	0.93	1.57	1.13	0.93	1.48	0.94

Analyzing in greater details the results in the tables, we observe the good behaviour of LPWR and TREND in the case of equispaced and random grid, respectively (see in particular Figs. 1–2); both are 3rd in a virtual ranking among methods, after SSW and ARM. If we restrict our attention to the boundaries

(error bRMSE), we notice the good performance of LPWR (virtually ranked 2nd after SSW), that outperforms all other competitors in 3 cases out of 6. This confirms effectiveness of the method at the boundary, but at detriment of the performance in the middle. Finally we observe the good performances of Splines and Wavelets on specific functions, namely Sin for Splines and Irregular for Wavelets. In particular the latter is the first together with SSW in a virtual ranking among the methods for this function. Actually these functions satisfy the assumptions of the respective methods in terms of regularity and periodic conditions.

Finally we show in Fig. 3 the average of the retrieved functions over the 100 repetitions (therefore estimating the bias of the estimators) for an equispaced grid of size 512 and the best estimator (SSW). The estimate of the smooth parts of the functions is very good, including boundaries. Irregular parts as discontinuities or bumps are not satisfactory (in particular discontinuities are not reproduced in the Heavisine function).

Summarizing we can say that conventional methods based on single regressors and on splines or wavelets behave quite well for specific functions that meet the assumptions they are based on. However their accuracy quickly degrades when such requirements are not satisfied. SSW and ARM, that show some similarities in the procedures, are the best estimators. LPWR performs well at boundaries, for which it is designed, but not as well at the inner part of the function; in addition it can be applied only to equispaced, dyadic grids. Its upgraded version (WHYBRID), adapted to general grids, is degraded by some instabilities in the case of random grids. On the contrary a mix of regressors chosen individually with different properties and assumptions to satisfy does increase performances of the methods. This is particularly effective at boundaries, where inclusion of a well-behaved method there such as Splines or Local Polynomial regression is able to greatly improve performances not only at boundaries but also everywhere.

Our experiments did not show good performances from several other methods included in the present analysis and others not reported (PWR, CSRecSP based on Compressive sensing (Dai and Milenkovic, 2008), SARS, several variants of the methods introduced).

6.4. Real examples

In this Section we consider two real examples and compare the solutions obtained by the two best methods (SSW and ARM) with the classical Wavelet and Splines regression.

Motorcycle data It is the example introduced in Silverman (1985). It consists of acceleration data (in gravity units) over time after a motorcycle crash. It is an example of a quite smooth function (time spans a very short time, 57ms after the impact) with boundary values that can be considered almost similar. Fig. 4 shows the fit obtained by SPL, WAV, ARM and SSW. All solutions fit the data well, with Spline exhibiting a smoother behaviour.

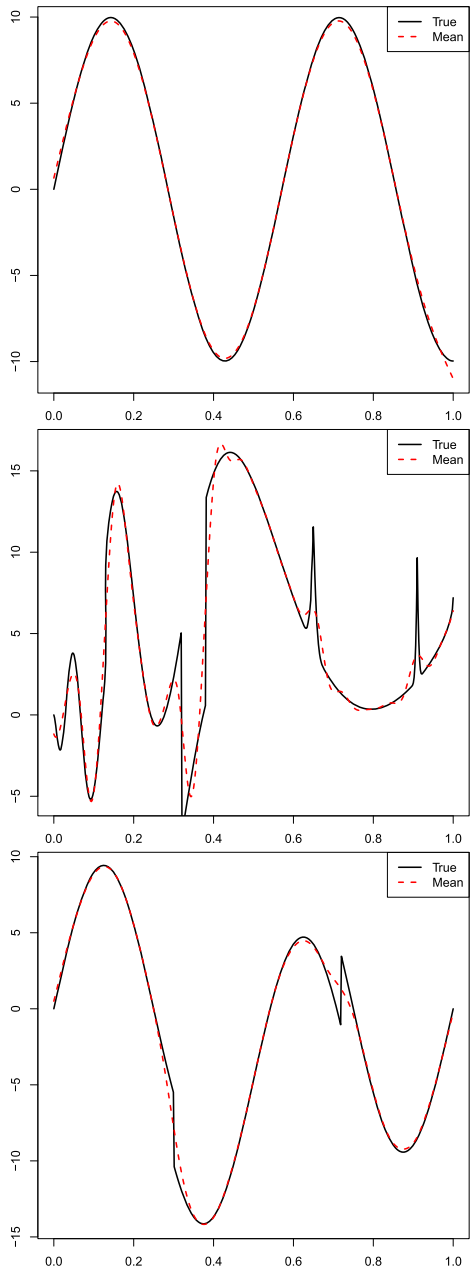

FIG 3. *Average of the estimator SSW over the repetitions for the Sin (upper panel), Irregular (middle) and Heavisine (lower) test functions in the case of equispaced grid of size 512. Continuous black line: synthetic function; dashed red line: average over repetitions.*

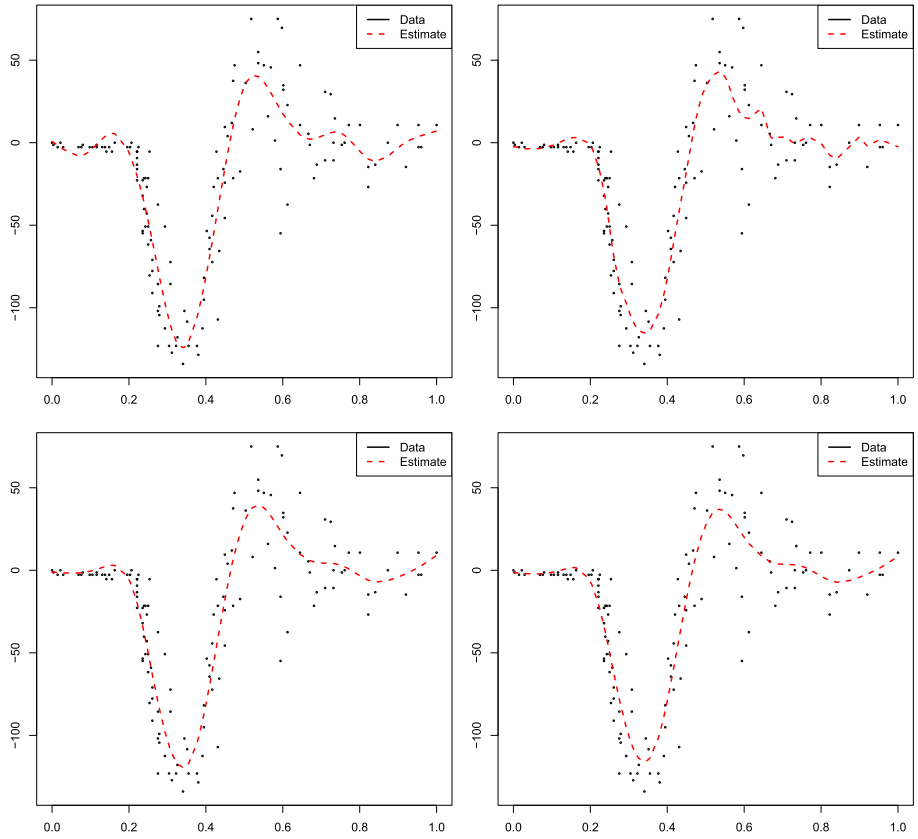

FIG 4. *Regression of the motorcycle data by SPL (upper left panel), WAV (upper right), ARM (lower left) and SSW (lower right). Black dots: data; red dashed line: the fit. Data of time were normalized to* $[0, 1]$.

Wool data This example is reported in Diggle (1991). The source of data is the Australian Wool Corporation, that recorded the price of the wool weekly from July 1976 to June 1984. The prices are the floor price, set from the Corporation, and the price actually paid for a particular grade (19μm nominal thickness in the data set), that comes out to be somewhat different. Data can be downloaded ad the link http://lib.stat.cmu.edu/datasets/diggle. The data, that in the time series show a seasonal trend, are complicated by a devaluation of the Australian dollar occurred in 1983 that generated a discontinuity. A consequence is that data are less regular; in addition the values at the boundaries are completely different. Fig. 5 shows the fit obtained by SPL, WAV, ARM and SSW. We notice incapability of a Wavelet regression to handle different boundaries, as expected. All other methods quite well reproduce the general behaviour of the function, with the Spline fit again being smoother and not able to reproduce some parts showing a sharper variability.

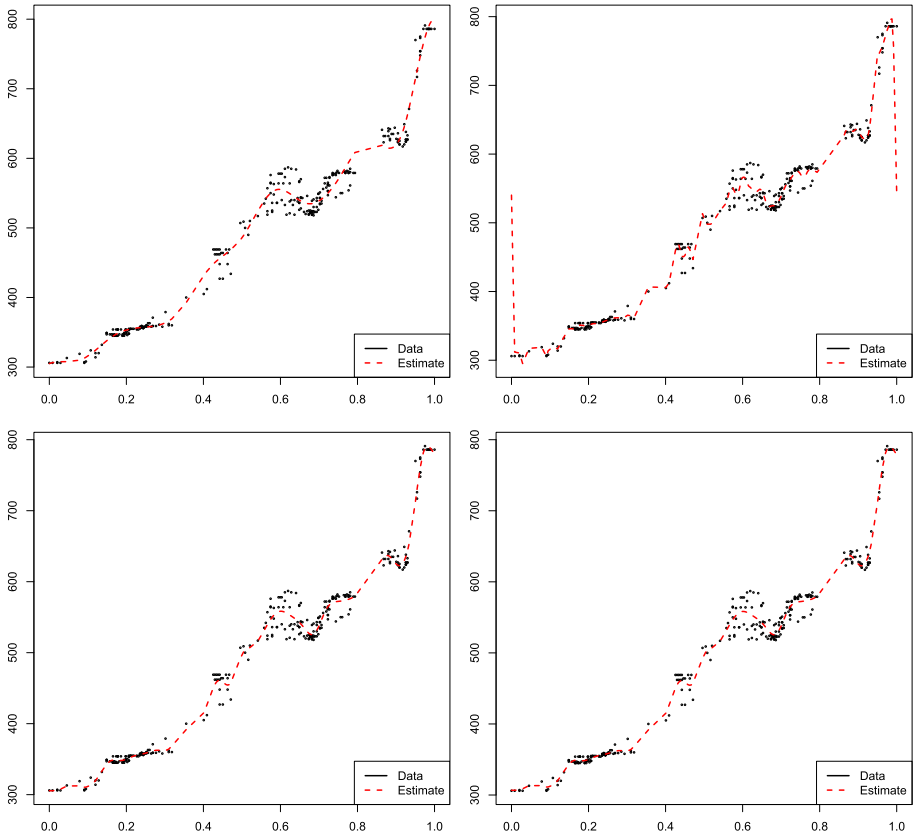

FIG 5. *Regression of the Wool data by SPL (upper left panel), WAV (upper right), ARM (lower left) and SSW (lower right). Black dots: data; red dashed line: the fit. Data of Retain Price in the abscissa were normalized to* $[0, 1]$.

7. Conclusions

This paper dealt with the problem of nonparametric regression of a univariate function when the distribution of nodes is equidistant, fixed or random. As well known, most common and effective methods introduce artifacts at the boundaries when their assumptions are not satisfied by actual data (e.g., periodic conditions). This is more pronounced with wavelets, that outperform splines and polynomial models from the theoretical point of view when regularity of functions is considered, but whose constraints on the boundaries are severely tight.

The present paper introduced some ideas about how to circumvent the boundary problem using as an example some of the recent and claimed most accurate nonparametric regression methods. These include in particular mixing or aggregation of models and methods based on libraries of bases. The key idea is

to include among methods at least one accurate also for nonsmooth functions, therefore wavelets, and one well behaved at boundaries, e.g., splines or polynomial regression. As a benefit, the current methods, particularly the ones based on wavelets, were also adapted to handle generic nodes, like nondyadic and nonequispaced ones.

Proofs of convergence for a class of mixing methods were given in the case of equispaced nodes. Experiments on simulated data have shown that such ideas improve accuracy of the fit non only at boundaries but also all over the domain of the function. Finally illustrations on two real applications were given.

The paper considered several methodologies as the basis of the ideas or simply as a comparison. For many of them, not reported for the sake of brevity, results came out to be poor. Among them we mention locally adaptive stacking, where the weights assigned to the different methodologies were depending on t. The idea was to weight locally the single regressions, e.g., favouring the more accurate ones where the function is irregular or the better appropriate at boundaries. We believe that this could be an interesting direction to investigate on.

An immediate extension of the methodologies concerns multivariate functions. In this respect all the proposed methodologies can be straightforwardly plugged in an additive framework by relying on backfitting. In addition one of them, namely MSR, is natively ready to be applied in a multidimensional setup without needing backfitting iterations.

Appendix A: Appendix

Proof of Theorem 3.1

First we look at the rate of convergence for the "oracle" choice α^* of the hyperparameter α. Using a proof similar to Lemmas 5.1, 5.2 and 5.4 in Burman and Chaudhuri (2011) it is easy to show that

$$\alpha^\star = \begin{cases} 1 + O_P((r_1(n)/\delta_n)^{1/2}), & \text{if } \delta_n > r_1(n) \\ O_P((\delta_n/r_1(n))^{1/2}), & \text{if } r_2(n) \le \delta_n \le r_1(n) \\ O_P((r_2(n)/r_1(n))^{1/2}), & \text{if } \delta_n < r_2(n). \end{cases}$$

Now,

$$\begin{aligned} \|\boldsymbol{f} - \hat{\boldsymbol{f}}_{\alpha^\star}\|_n &= \|\alpha^\star(\boldsymbol{f} - \hat{\boldsymbol{f}}_{\mathrm{LP}}) - (1-\alpha^\star)(\hat{\boldsymbol{f}}_{\mathrm{W}} - \boldsymbol{f})\|_n \\ &\le |\alpha^\star|\|\boldsymbol{f} - \hat{\boldsymbol{f}}_{\mathrm{LP}}\|_n + |1-\alpha^\star|\|\hat{\boldsymbol{f}}_{\mathrm{W}} - \boldsymbol{f}\|_n \\ &\le |\alpha^\star| O_P\left(r_1(n)^{1/2}\right) + |1-\alpha^\star|\{O_P\left(r_2(n)^{1/2}\right) + \delta_n^{1/2}\} \end{aligned}$$

and the first assertion of the Theorem follows from the asymptotic behavior of $\alpha^\star$ noted above.

By construction of the oracle estimator we also have

$$\|\boldsymbol{f} - \hat{\boldsymbol{f}}_{\hat{\alpha}}\|_n^2 - \|\boldsymbol{f} - \hat{\boldsymbol{f}}_{\alpha^\star}\|_n^2 = \|\hat{\boldsymbol{f}}_{\alpha^\star} - \hat{\boldsymbol{f}}_{\hat{\alpha}}\|_n^2 = (\hat{\alpha} - \alpha^\star)^2 \|\hat{\boldsymbol{f}}_{\mathrm{LP}} - \hat{\boldsymbol{f}}_{\mathrm{W}}\|_n^2.$$

Applying the rates for both the local polynomial estimator and the SUREshrink thresholded wavelet estimator and the derivation of $\hat{\alpha}$ using Stein's Lemma, it is easy to show, under the condition on δ_n stated in the Theorem, that $(\hat{\alpha} - \alpha^\star)^2 \|\hat{\boldsymbol{f}}_{\mathrm{LP}} - \hat{\boldsymbol{f}}_{\mathrm{W}}\|_n^2$ tends to zero as fast as $\|\boldsymbol{f} - \hat{\boldsymbol{f}}_{\alpha^\star}\|_n^2$ which implies the second assertion.

Acknowledgements

The authors thank the Editor and the two referees for their constructive comments and helpful suggestions, which improved the paper. Part of this work was completed while A. Antoniadis was visiting the Istituto per le Applicazioni del Calcolo "M. Picone", National Research Council, Naples, Italy, with the support of INDAM - Visiting Professors Program.

References

Antoniadis, A. (2007). Wavelet methods in statistics: some recent developments and their applications. *Statistics Surveys* **1** 16–55.

Antoniadis, A., Bigot, J. and Sapatinas, T. (2001). Wavelet Estimators in Nonparametric Regression: A Comparative Simulation Study. *Journal of Statistical Software* **6**. MR1111111

Antoniadis, A. and Fan, J. (2001). Regularization of Wavelet Approximations. *Journal of the American Statistical Association* **96** 939–967. MR1946364

Antoniadis, A. and Pham, D. T. (1998). Wavelet regression for random or irregular design. *Computational Statistics & Data Analysis* **28** 353–369.

Blu, T. and Luisier, F. (2007). The SURE-LET Approach to Image Denoising. *IEEE Transactions on Image Processing* **16** 2778–2786. MR2472419

Bredies, K., Lorenz, D. A. and Reiterer, S. (2014). Minimization of Non-smooth, Non-convex Functionals by Iterative Thresholding. *Journal of Optimization Theory and Applications* **165** 78–112. MR3327417

Breiman, L. (1996). Stacked regressions. *Machine Learning* **24** 49–64.

Buckheit, J. B. and Donoho, D. L. (1995). WaveLab and Reproducible Research. In *Wavelets and Statistics* (A. Antoniadis and G. Oppenheim, eds.). *Lecture Notes in Statistics* **103** 55–81. Springer New York. MR1364669

Bunea, F., Lederer, J. and She, Y. (2014). The Group Square-Root Lasso: Theoretical Properties and Fast Algorithms. *IEEE Transactions on Information Theory* **60** 1313–1325. MR3164977

Burman, P. and Chaudhuri, P. (2011). On a Hybrid Approach to Parametric and Nonparametric Regression. In *Nonparametric Statistical Methods and Related Topics* 233–256. WORLD SCIENTIFIC. MR2932489

Cai, T. T. and Brown, L. D. (1998). Wavelet shrinkage for nonequispaced samples. *The Annals of Statistics* **26** 1783–1799. MR1673278

Catoni, O. (2004). *Statistical Learning Theory and Stochastic Optimization.* Springer Berlin Heidelberg. MR2163920

Chen, L. and Yang, Y. (2010). Combining Statistical Procedures. In *High-Dimensional Data Analysis* 275–298. World Scientific Higher Education Press, China. MR2848207

Cohen, A., Daubechies, I. and Vial, P. (1993). Wavelets on the Interval and Fast Wavelet Transforms. *Applied and Computational Harmonic Analysis* **1** 54–81. MR1256527

Dai, W. and Milenkovic, O. (2008). Subspace Pursuit for Compressive Sensing: Closing the Gap Between Performance and Complexity. arXiv:0803.0811.

Daubechies, I. (1992). *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics. MR1162107

Daubechies, I., Defrise, M. and Mol, C. D. (2004). An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics* **57** 1413–1457. MR2077704

Diggle, P. J. (1991). *Time series, a biostatistical introduction*. Wiley. MR1055357

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to Unknown Smoothness via Wavelet Shrinkage. *Journal of the American Statistical Association* **90** 1200–1224. MR1379464

Donoho, D. L. and Johnstone, I. M. (1998). Minimax estimation via wavelet shrinkage. *The Annals of Statistics* **26** 879–921. MR1635414

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995). Wavelet Shrinkage: Asymptopia? *Journal of the Royal Statistical Society: Series B (Methodological)* **57** 301–337. MR1323344

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1996). Density estimation by wavelet thresholding. *The Annals of Statistics* **24** 508–539. MR1394974

Du, P., Parmeter, C. F. and Racine, J. S. (2013). Nonparametric kernel regression with multiple predictors and multiple shape constraints. *Statistica Sinica* **23** 1347–1371. MR3114717

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* **11** 89–121. MR1435485

Eubank, R. L. (1999). *Nonparametric Regression and Spline Smoothing, Second Edition*. CRC Press. MR1680784

Fahrmeir, L. and Lang, S. (2001). Bayesian inference for generalized additive mixed models based on Markov random field priors. *Journal of the Royal Statistical Society, Series C* **50** 201–220.

Fan, J. (1992). Design-adaptive Nonparametric Regression. *Journal of the American Statistical Association* **87** 998–1004. MR1209561

Fan, J. (1993). Local Linear Regression Smoothers and Their Minimax Efficiencies. *The Annals of Statistics* **21** 196–216. MR1212173

Green, P. J. and Silverman, B. W. (1993). *Nonparametric Regression and Generalized Linear Models*. Chapman and Hall/CRC. MR1270012

Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press. MR1161622

Haris, A., Simon, N. and Shoiaje, A. (2018). Wavelet regression and additive models for irregularly spaced data. In *Proc. 32nd Conference on Neural*

Information Processing Systems (NeurIPS 2018 1–11.
Hastie (2003). Maybe Hastie 2009? *journal*.
Hastie, T. and Loader, C. (1993). Local Regression: Automatic Kernel Carpentry. *Statistical Science* **8** 120–129.
Hörmandere, L. (1989). Continuity of pseedo-differential operators of type 1, 1. *Communications in Partial Differential Equations* **14** 231–243. MR976972
Johnstone, I. M. and Silverman, B. W. (2005). Empirical Bayes selection of wavelet thresholds. *The Annals of Statistics* **33** 1700–1752. MR2166560
Kovac, A. and Silverman, B. W. (2000). Extending the Scope of Wavelet Regression Methods by Coefficient-Dependent Thresholding. *Journal of the American Statistical Association* **95** 172–183.
Lee, T. C. M. and Oh, H.-S. (2004). Automatic polynomial wavelet regression. *Statistics and Computing* **14** 337–341. MR2062060
Lindgren, F. and Rue, H. (2008). On the second-order random walk model for irregular locations. *Scandinavian journal of statistics* **35** 691–700.
Mallat, S. (2009). *A Wavelet Tour of Signal Processing*. Elsevier.
Meyer, Y. (1993). *Wavelets and Operators. Cambridge Studies in Advanced Mathematics* **1**. Cambridge University Press. MR1228209
Miller, D. L., Glennie, R. and Seaton, A. E. (2019). Understanding the Stochastic Partial Differential Equation Approach to Smoothing. *Journal of Agricultural, Biological and Environmental Statistics*.
Nason, G. P. (1996). Wavelet Shrinkage Using Cross-Validation. *Journal of the Royal Statistical Society: Series B (Methodological)* **58** 463–479. MR1377845
Ogden, T. (1996). *Essential Wavelets for Statistical Applications and Data Analysis*. Birkhäuser. MR1420193
Oh, H.-S. and Lee, T. C. M. (2005). Hybrid local polynomial wavelet shrinkage: wavelet regression with automatic boundary adjustment. *Computational Statistics & Data Analysis* **48** 809–819. MR2133579
Oh, H.-S., Naveau, P. and Lee, G. (2001). Polynomial boundary treatment for wavelet regression. *Biometrika* **88** 291–298. MR1841277
Osborne, M. A., Garnett, R. and Roberts, S. J. (2010). Active Data Selection for Sensor Networks with Faults and Changepoints. In *Proceedings of the 2010 24th IEEE International Conference on Advanced Information Networking and Applications. AINA '10* 533–540. IEEE Computer Society, Washington, DC, USA.
Parikh, N. (2014). Proximal Algorithms. *Foundations and Trends in Optimization* **1** 127–239.
Ramdas, A. and Tibshirani, R. J. (2016). Fast and Flexible ADMM Algorithms for Trend Filtering. *Journal of Computational and Graphical Statistics* **25** 839–858. MR3533641
Rasmussen, C. E. and Williams, C. K. I. (2005). *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.
Rudin, L. I., Osher, S. and Fatemi, E. (1992). Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena* **60** 259–268. MR3363401

Rue, H., Martino, S. and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society, Series B* **71** 319–392.

She, Y. (2012). An iterative algorithm for fitting nonconvex penalized generalized linear models with grouped predictors. *Computational Statistics & Data Analysis* **56** 2976–2990.

Silverman, B. W. (1985). Some Aspects of the Spline Smoothing Approach to Non-Parametric Regression Curve Fitting. *Journal of the Royal Statistical Society: Series B (Methodological)* **47** 1–21. MR805063

Sklar, J. C., Wu, J., Meiring, W. and Wang, Y. (2013). Nonparametric Regression With Basis Selection From Multiple Libraries. *Technometrics* **55** 189–201. MR3176519

Tibshirani, R. J. (2014). Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics* **42** 285–323. MR3189487

Wand, M. P. and Ormerod, J. T. (2008). On sempiparametric regression with O'Sullivan penalized Splines. *Australian & New Zealand Journal of Statistics* **50** 179–198. MR2431193

Wand, M. P. and Ormerod, J. T. (2011). Penalized wavelets: Embedding wavelets into semiparametric regression. *Electronic Journal of Statistics* **5** 1654–1717. MR2870147

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks* **5** 241–259.

Wood, S. N. (2006). On confidence intervals for generalized additive models based on penalized regression splines. *Australian & New Zealand Journal of Statistics* **48** 445–464. MR2329279

Yang, Y. (2001). Adaptive Regression by Mixing. *Journal of the American Statistical Association* **96** 574–588. MR1946426

Yang, Y. (2003). Regression with Multiple Candidate Models: Selecting or Mixing? *Satistica Sinica* **13** 783–809. MR1997174

Yue, Y. R., Simpson, D., Lindgren, F. and Rue, H. (2014). Bayesian adaptive smoothing splines using stochastic differential equations. *Bayesian Analysis* **9** 397–424.

Zhou, S. and Shen, X. (2001). Spatially Adaptive Regression Splines and Accurate Knot Selection Schemes. *Journal of the American Statistical Association* **96** 247–259. MR1952735