Computational Statistics (2022) 37:1621–1651
https://doi.org/10.1007/s00180-021-01174-4

ORIGINAL PAPER

Penalized wavelet estimation and robust denoising for irregular spaced data

Umberto Amato[1] · Anestis Antoniadis[2,3] · Italia De Feis[4] · Irène Gijbels[5]

Received: 22 March 2021 / Accepted: 10 November 2021 / Published online: 12 December 2021

Abstract

Nonparametric univariate regression via wavelets is usually implemented under the assumptions of dyadic sample size, equally spaced fixed sample points, and i.i.d. normal errors. In this work, we propose, study and compare some wavelet based nonparametric estimation methods designed to recover a one-dimensional regression function for data that not necessary possess the above requirements. These methods use appropriate regularizations by penalizing the decomposition of the unknown regression function on a wavelet basis of functions evaluated on the sampling design. Exploiting the sparsity of wavelet decompositions for signals belonging to homogeneous Besov spaces, we use some efficient proximal gradient descent algorithms, available in recent literature, for computing the estimates with fast computation times. Our wavelet based procedures, in both the standard and the robust regression case have favorable theoretical properties, thanks in large part to the separability nature of the (non convex) regularization they are based on. We establish asymptotic global

✉ Italia De Feis
i.defeis@iac.cnr.it

Umberto Amato
umberto.amato@cnr.it

Anestis Antoniadis
Anestis.Antoniadis@univ-grenoble-alpes.fr

Irène Gijbels
irene.gijbels@kuleuven.be

[1] Istituto di Scienze Applicate e Sistemi Intelligenti, Consiglio Nazionale delle Ricerche, Naples, Italy

[2] Laboratoire Jean Kuntzmann, Department of Statistics, University Grenoble Alpes, Grenoble, France

[3] Department of Statistical Sciences, University of Cape Town, Cape Town, South Africa

[4] Istituto per le Applicazioni del Calcolo "M. Picone", Consiglio Nazionale delle Ricerche, Naples, Italy

[5] Department of Mathematics and Leuven Statistics Research Center (LStat), KU Leuven, Leuven, Belgium

optimal rates of convergence under weak conditions. It is known that such rates are, in general, unattainable by smoothing splines or other linear nonparametric smoothers. Lastly, we present several experiments to examine the empirical performance of our procedures and their comparisons with other proposals available in the literature. An interesting regression analysis of some real data applications using these procedures unambiguously demonstrate their effectiveness.

Keywords Wavelets · Nonparametric regression · Proximal algorithms · Thresholding · Robust fitting

1 Introduction

Wavelets based methods work quickly, efficiently, and attain fast rates of convergence in the problem of nonparametric univariate regression over equally spaced, fixed sample points on an interval, usually [0, 1] with a dyadic number of data points and independent noise:

$$y_i = f(x_i) + \sigma\varepsilon_i, \quad i = 1, 2, \ldots, n. \tag{1.1}$$

In this model, the y_i are the observed data, $x_i = (i-1)/n$ are the sample points, and $\sigma\varepsilon_i$ is noise. The noise is assumed to be i.i.d. normal random variables in most cases. The function f is an unknown function of interest. We wish to estimate f and one can measure the performance of an estimate $\hat{f}$ by the expected global mean squared error,

$$R_n(\hat{f}, f) = \mathbb{E}\left\{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}(x_i) - f(x_i)\right)^2\right\},$$

the goal being to construct estimates that have "small" risk R_n. 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) and Härdle (1990) 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 with dyadic, equispaced data, was pioneered in Donoho and Johnstone (1995) and 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 of order $\log(n)n^{-2s/(2s+1)}$ for the mean squared error risk when f is in a ball of a Besov space $\mathcal{B}^s_{p,r}$ with a smoothness parameter s, which can not be achieved by any linear estimator. For a thorough review of wavelet methods in statistics and an important list of references 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 discrete wavelet transform (DWT) to a compactly supported signal. Since this orthogonal transform is a matrix with dyadic dimensions, these methods have generally required a dyadic number of data points.

When the data does not meet these requirements, various modifications have been proposed. If n is not dyadic, it can be extended to a dyadic integer by reflection or periodizing. For data that is not equispaced, Cai and Brown (1998) investigated wavelet methods on samples with fixed designs via an approximation approach. They showed that applying the methods devised for equispaced data directly to nonequispaced data may lead to suboptimal estimators. They then proposed a method that was adaptive and near optimal, but used the assumptions of i.i.d. noise and dyadic number of sample points. Work has also been done with sample points that are not only nonequispaced, but random as well. Cai and Brown (1999) show that good rates can be obtained for certain random design schemes using the DWT.

We will deal with fixed, equispaced or nonequispaced, sampling designs of a size that is not supposed to be a power of 2. In the nonequispaced case, the design points x_i are random design points drawn from a univariate distribution with a compactly supported density function f_X, that satisfies certain conditions (essentially no flat parts in the density, or ‘no holes’ in the design). However, our wavelet-based regression procedures are defined using a standard least squares loss function which is unbounded and can be very sensitive to the presence of outlying observations. Moreover, it is challenging to detect outliers with standard wavelet-based regression procedures, especially when denoising functions with abrupt changes. Thus, robust estimators for such regression procedures are in need, for such cases. Therefore, a further relaxation of the constraints in (1.1) will be made, this time on the noise level assumptions by assuming that there are either sub-Gaussian or even, sometime, heavy-tailed, when a presence of aberrant observations (outliers) of the response variable is suspected. We will then also address robust estimation, meaning that our procedures remain valid even when there are aberrant observations of the response variable. The literature on this line of work, as far as we can tell, is not extensive.

This paper is organized as follows. In Sect. 2 we describe the framework for the wavelet-based regression model with the basic concept of wavelets. A short description of the standard wavelet based regression procedures for the equispaced case via the DWT transform is also reviewed in this section. In Sect. 3 we 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 the procedures, some completely new, and their asymptotic properties in the general fixed design case is presented in several subsections of this section. The problem of robust nonparametric wavelet smoothing is studied in Sect. 4. All procedures are investigated via various simulation settings and real data application examples in relevant subsections throughout the paper. Some concluding remarks are given in Sect. 5. Technical proofs are relegated to an "Appendix".

The R codes and their description implementing the methods in the paper together with the driver scripts for simulations, plots and the analysis of real examples are made available in a compressed archive as supplementary material.

2 Wavelet regression for equispaced data

Let us first consider a particular univariate regression problem stated as

$$Y_i = f(t_i) + \epsilon_i, \quad i = 1, \ldots, n, \tag{2.1}$$

with a non-stochastic equidistant design $t_i = i/n$, $i = 1, \ldots, n$ 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 non-smooth function f, defined on [0, 1], that may present a wide range of irregular effects. There is not any loss of generality to assume that $\mu = \int_0^1 f(x)dx = 0$, since this parameter can always be estimated with a root-n rate by the empirical mean of the observations. We will therefore assume that the data has been centered.

2.1 A short background on wavelets

Wavelets denoising procedures have the ability to estimate inhomogeneous functions showing abrupt changes and discontinuities. Models for f in (2.1), that allow a wide range of irregular effects, are through the sequence space representation of (inhomogeneous) Besov spaces $\mathcal{B}^s_{d,q}([0, 1])$, $1 \le d \le \infty$, $1 \le q \le \infty$ (see e.g. Donoho et al. (1995)). To capture key characteristics of variations in f and to exploit its sparse wavelet coefficients representation, we will assume that f belongs to a "ball" of $\mathcal{B}^s_{d,q}([0, 1])$ of radius M with $1 \le q \le \infty$ and $s + 1/d - 1/2 > 0$. The last condition ensures in particular that evaluation of f at a given point makes sense.

To deal with the estimation of f in model (2.1) several wavelet based estimation procedures have been developed in the literature (see, e.g. Antoniadis (2007) for a review) and we will summarize in this subsection the main ingredients most relevant to our work.

Throughout this paper 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 ($r \ge 0$) multi-resolution analysis of $\left(L^2[0, 1], \langle\cdot, \cdot\rangle\right)$, the space of squared-integrable functions on [0, 1] endowed with the inner product $\langle f, g\rangle = \int_{[0,1]} f(t)g(t)\,dt$. For simplicity in exposition, we work with periodic wavelet bases on [0, 1] (see, e.g., Mallat (2009)), letting

$$\phi^{\mathrm{p}}_{jk}(t) = \sum_{l\in\mathbb{Z}} \phi_{jk}(t-l) \quad \text{and} \quad \psi^{\mathrm{p}}_{jk}(t) = \sum_{l\in\mathbb{Z}} \psi_{jk}(t-l), \quad \text{for} \quad t \in [0,1],$$

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

$$\{\phi^{\mathrm{p}}_{j_0 k},\ k = 0, 1, \ldots, 2^{j_0} - 1;\ \psi^{\mathrm{p}}_{jk},\ j \geq j_0;\ k = 0, 1, \ldots, 2^j - 1\}$$

is then an orthonormal basis of $L^2[0,1]$. The superscript “p” will be suppressed from the notation for convenience. 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.

Let us adopt a vector-matrix form for model (2.1),

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

with $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, $\mathbf{f} = (f(t_1), \ldots, f(t_n))^T$ and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$. After applying a linear and orthogonal wavelet transform, the discretized model becomes

$$\mathbf{Z} = \boldsymbol{\gamma} + \tilde{\boldsymbol{\epsilon}}. \tag{2.2}$$

where $\mathbf{Z} = W_{n\times n}\mathbf{Y}$, $\boldsymbol{\gamma} = W_{n\times n}\mathbf{f}$ 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 $\sigma^2\mathbb{I}_n$. Hence, the representation of the model in the wavelet domain allows to exploit in an efficient way the sparsity of the wavelet coefficients $\boldsymbol{\gamma}$ in the representation of the nonparametric component including the case of *hard sparsity*, meaning that $\boldsymbol{\gamma}$ has exactly K^* non-zero entries, or *soft sparsity*, based on the decay rate on the ordered entries of $\boldsymbol{\gamma}$ of wavelet coefficients is bounded above by say R.

By modeling the unknown signal with its DWT, we over-parameterized the linear model (2.2) derived from the observations. The prior knowledge on the sparsity of the wavelet coefficients of f, is used to regularize efficiently the over-parameterized linear model by minimizing with respect to $\boldsymbol{\gamma}$ a separable penalized least squares expression:

$$\frac{1}{n}\sum_{i=1}^{n}(Z_i - \boldsymbol{\gamma})^2 + \sum_{k=1}^{n} p_{\lambda_n}(|\gamma_k|),$$

where, for $\lambda > 0$, $p_\lambda(|\cdot|)$, is an appropriate penalty that preserves the sparsity (see e.g. Antoniadis and Fan (2001)). For example, using a SCAD penalty and an appropriate sequence of regularization parameters λ_n one can show, that $\hat{\mathbf{f}}$ has an optimal asymptotic mean-squared error rate

$$\frac{1}{n}\left\{\sum_{i=1}^{n}(f(t_i)-\hat{f}_i)^2\right\}^{1/2} := \|\hat{\mathbf{f}}-\mathbf{f}\|_{2,n} = \mathcal{O}_P\left((\log n)n^{-\frac{s}{2s+1}}\right).$$

Thus, for regularly spaced data with $n = 2^J$, wavelet regularizations provide consistent nonparametric estimators with nearly optimal rates.

3 General fixed design case

Our task again is to estimate an unknown function f defined on [0, 1], given noisy data y_i at sampling points $x_1, \ldots, x_n$:

$$y_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots, n.$$

We will now assume that the sampling points $x_1, \ldots, x_n$ are i.i.d realizations from an input distribution with a continuous density on [0, 1] which is bounded below by a constant $b_0 > 0$. However, the resulting $x_1, \ldots, x_n$ are regarded as deterministic points in [0, 1], so the only randomness involved is the sampling of the y_i's. When the underlying unknown signal f is smooth enough, several nonparametric approaches (local polynomials, kernel based procedures, spline smoothing or penalized splines) exist and their properties are well studied. However these methods are suboptimal when f belongs to some spaces of inhomogeneous smoothness such as the ball of the inhomogeneous Besov space $\mathcal{B}^s_{d,q}([0,1])$ we used before. We shortly reviewed in the previous subsection some efficient wavelet based procedures for estimating f. However, the task is not so easy when n is not necessarily a power of 2 and especially when we are also dealing with irregularly spaced data which is the realistic case when sampling points occur from a general input density.

Using the same notation as before, and once an r-regular multi-resolution analysis of $L^2[0,1]$ has been chosen, the unknown univariate function f has a derived wavelet basis expansion. For a given n, since f belongs to $\mathcal{B}^s_{d,q}(M)$, we can approximate the function f by its truncated expansion on the wavelet basis functions, say $\{W_\ell\}_\ell$:

$$f(x_i) \approx \sum_{\ell=1}^{K}\gamma_\ell W_\ell(x_i) \qquad \text{for} \quad i = 1, \ldots, n,$$

where K is an appropriate truncation index (depending on s) that is allowed to increase to infinity with n. Under our assumptions on f, the values of f on the design points are then well approximated by the above expansion and their estimation is therefore equivalent in estimating the (sparse) wavelet coefficient vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_K)^T$.

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 the design points. Using this construction, we denote by $\mathbf{W}$ the corresponding wavelet regression $n \times K$ matrix of the wavelet basis functions evaluated at the observed sampling points, i.e.

$$\mathbf{W} = \begin{bmatrix} W_1(x_1) & \dots & W_K(x_1) \\ \vdots & \ddots & \vdots \\ W_1(x_n) & \dots & W_K(x_n) \end{bmatrix}.$$

Adopt again a vector-matrix form of the regression model, given by equation (2.1) to get:

$$\mathbf{Y} \approx \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \tag{3.1}$$

for $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ and $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)^T$. The function f is therefore characterised by the K-dimensional sparse coefficient vector $\boldsymbol{\gamma}$. If the response Y is not centered, the intercept may be efficiently estimated by the empirical median or mean of the observations and therefore without any loss of generality we may assume that the columns of $\mathbf{W}$ are centered, which, in particular, means that the basis generating $\mathbf{W}$ doesn't involve scaling functions. Moreover, there is no loss of generality in assuming hereafter that $\mathbf{W}$ has its columns normalized, .i.e. $\|\mathbf{W}_i\|_2 = 1$, for $i = 1, \dots, K$.

Remark 1 To get an appropriate approximation of the function f and achieve optimal asymptotic rates for the resulting estimator, the index K will be allowed to be of the order $\mathcal{O}(n^q)$ with $q \in]0, 1]$. Therefore in the linear regression model (3.1) the matrix $\mathbf{W}$ may have less rows than columns. In this case, where $n \ll K$, problem (3.1) is severely underdetermined (and ill-posed), and hence it is challenging to obtain a meaningful solution. The sparsity of $\boldsymbol{\gamma}$, looking for an approximative solution with many zero entries is then essential for estimating the unknown regression function.

Under the assumption that $\boldsymbol{\gamma}$ is sparse in the sense that the number K^* of important nonzero elements of $\boldsymbol{\gamma}$ is small relative to n, we can estimate $\boldsymbol{\gamma}$ with several approaches proposed in the literature. Among them a popular method for realizing sparsity constraints is the Lasso (Tibshirani 1996). It leads to a convex but nonsmooth optimization problem:

$$\min_{\boldsymbol{\gamma} \in \mathbb{R}^K} \left\{ \frac{1}{2} \|\mathbf{Y} - \mathbf{W}\boldsymbol{\gamma}\|^2 + \lambda \|\boldsymbol{\gamma}\|_1 \right\}, \tag{3.2}$$

where $\|\cdot\|_1$ denotes the ℓ^1-norm of a vector, and $\lambda > 0$ is a regularization parameter. Since its introduction (Chen et al. 1998; Tibshirani 1996), problem (3.2) has gained immense popularity in many diverse disciplines, which can largely be attributed to the fact that it admits efficient numerical solution. Under certain regularity conditions (e.g., level wise restricted isometry property and level wise restricted eigenvalue condition) on the design matrix $\mathbf{W}$ and the sparsity level of the true vector $\boldsymbol{\gamma}$, it can produce models with good estimation and prediction accuracy [see e.g. Zhao and Yu (2006) and Wainwright (2009)]. However, Lasso tends to over-shrink large coefficients, which leads to biased estimates (Antoniadis and Fan 2001; Fan and Li 2001; Fan and Peng

2004) and this implies that there does not exist a sequence λ_n of the tuning parameter λ which can lead to both variable selection consistency and asymptotic normality [see Fan and Li (2001) and Zhao and Yu (2006)]. In other words this means that the Lasso estimates do not have the oracle property even when the minimum of the nonzero coefficients of $\boldsymbol{\gamma}$ is bounded below. A way to improve variable selection accuracy and gain oracle properties is by reducing the bias of Lasso via the adaptive Lasso procedure (Zou 2006) which solves the following weighted l_1 regularization problem for some $\alpha \in (0, 1)$:

$$\hat{\boldsymbol{\gamma}}(\lambda_n) = \arg\min_{\boldsymbol{\gamma}} \left\{ ||\mathbf{Y} - \mathbf{W}\boldsymbol{\gamma}||_2^2 + \lambda_n \sum_{j=1}^{n} |\hat{w}_j|^{-\alpha} |\gamma_j| \right\},$$

where $\hat{\mathbf{w}}$ is an estimator of $\boldsymbol{\gamma}$ (for example, the solution of the standard unweighted Lasso with regularisation parameter λ_n or a ridge estimate of $\boldsymbol{\gamma}$). A low-dimensional analysis in Zou (2006) shows that the adaptive Lasso solution can achieve the oracle property asymptotically. A high dimensional analysis of this procedure was given in Huang et al. (2008). For variable selection consistency and oracle properties to hold, the adaptive lasso requires strong non optimal conditions in terms of the minimal strength of nonzero components of the true wavelet coefficients $\boldsymbol{\gamma}$ and some extra conditions on the design matrix.

Remark 2 The restricted isometry condition (RIP) has been introduced first in the compressive sensing literature to analyze ℓ_1-regularized recovery of a sparse $\boldsymbol{\beta}$ from its random projection $X\boldsymbol{\beta}$ with i.i.d. $\mathcal{N}(0, 1)$ entries in X. While conditions for RIP are specialized to hold for random designs with specific covariance matrices, related conditions using sparse eigenvalues can be defined (for example, the sparse Riesz condition (SRC)) and lead to ℓ_2-norm estimation error bounds of the optimal order for ℓ_1-regularized estimators. However, all these conditions are just sufficient conditions for an exact signal recovery. The deterministic matrices derived from the DWT transform are generally incapable to satisfy a RIP or SRC condition. But some refinements of RIP and SRC introduced in the literature (see e.g. Raskutti et al. (2010)), such as, for example, the level wise restricted lower eigenvalues conditions are then very useful when using the DWT and provide ℓ_2-norm prediction error bounds of optimal order.

The requirements on the minimum value of the nonzero coefficients are not optimal. Nonconvex penalties such as the smoothly clipped absolute deviation (SCAD) penalty (Antoniadis and Fan 2001; Fan and Li 2001), the minimax concave penalty (MCP) (Zhang 2010a) and the capped ℓ_1 penalty (Zhang 2010b) were proposed to remedy these problems and achieve support recovery. These penalties are good penalty function since they all possess the three principles that a good penalty function should satisfy (see Antoniadis and Fan 2001): unbiasedness, in which there is no over-penalisation of large coefficients to avoid unnecessary modelling biases; sparsity, as the resulting penalised least-squares estimators should follow a thresholding rule such that insignificant coefficients should be set to zero to reduce model complexity; and continuity to avoid instability and large variability in model prediction. The interested reader is referred to Theorem 1 of Antoniadis and Fan (2001) which gives the necessary and sufficient conditions on a penalty function for the solution of the penalised least-squares problem to be thresholding, continuous and approximately unbiased for large values of their argument. In the subsections that follow we will therefore adopt such penalties.

3.1 A non-convex wavelet based approach

Our nonconvex approach for wavelet based denoising leads to the following optimization problem:

$$\min_{\boldsymbol{\gamma}\in\mathbb{R}^K} L(\boldsymbol{\gamma}) = \min_{\boldsymbol{\gamma}\in\mathbb{R}^K} \left\{ \frac{1}{2}\|\mathbf{W}\boldsymbol{\gamma} - \mathbf{Y}\|^2 + \sum_{i=1}^{K} \rho_{\lambda,\tau}(\gamma_i) \right\}, \tag{3.3}$$

where $\rho_{\lambda,\tau}$ is a nonconvex penalty, $\lambda > 0$ is a regularization parameter, and τ controls the degree of concavity of the penalty. We already mentioned important examples of such penalties in this section. The nonconvexity and nonsmoothess of the penalties $\rho_{\lambda,\tau}$ pose some challenge for the optimization problem (3.3), but their attractive theoretical properties and empirical successes justify our choice.

We focus on the most commonly used nonconvex penalties, i.e., SCAD, MCP, and capped ℓ_1 which is a linear approximation of the SCAD penalty [but similar results can be derived for some of the nonquadratic penalties discussed in Antoniadis et al. (2011)], for recovering the unknown function f; see Table 1 for explicit formulas (and the associated thresholding operators, to be defined later).

It is worth mentioning here that the penalties $\rho_{\lambda,\tau}(t)$ in the above table are all (μ, ξ)-*amenable* (see Loh and Wainwright (2015)) as soon as $\tau > 2$. Recall in particular, that if $\rho_{\lambda,\tau}(t)$ is (μ, ξ)-amenable, then $q_{\lambda,\tau}(t) := \lambda|t| - \rho_{\lambda,\tau}(t)$ is everywhere differentiable. Defining the vector version $q_{\lambda,\tau} : I\!R^K \to I\!R$ accordingly, it is easy to see that $\frac{\mu}{2}\|\boldsymbol{\gamma}\|_2^2 - q_{\lambda,\tau}(\boldsymbol{\gamma})$ is convex. A consequence of the above facts is that for any of the nonconvex penalties in Table 1, there exists at least one global minimizer to problem (3.3). This essentially follows from the boundedness of L from below by zero, and by the lower semi-continuity of each of the penalties.

We now recall the expressions of thresholding operators for the penalties in Table 1 which form the basis of many existing algorithms, e.g., coordinate descent, iterative thresholding and first-order proximal mappings, which may be found in the literature. We recall also here an unified derivation of the thresholding operators associated to these penalties. For any of the penalties $\rho(t)$ in Table 1 (the subscripts λ and τ are omitted for simplicity), define a function $g(t) : [0, \infty) \to \mathbb{R}^+ \cup \{0\}$ by

$$g(t) = \begin{cases} \frac{t}{2} + \frac{\rho(t)}{t}, & t \neq 0, \\ \liminf\limits_{t\to 0^+} g(t), & t = 0. \end{cases}$$

It is easy to show that $T^* = \inf_{t>0} g(t)$ is attained at some point $t^* \geq 0$ and that one has $(t^*, T^*) = (0, \lambda)$ for the capped-ℓ^1, SCAD and MCP.

Table 1 Nonconvex penalties $\rho_{\lambda,\tau}(t)$ and corresponding thresholding operators $S^{\rho}_{\lambda,\tau}(v)$

Penalty	$\rho_{\lambda,\tau}(t)$	$S^{\rho}_{\lambda,\tau}(v)$
capped-ℓ_1, $\tau > \frac{1}{2}$, Zhang (2010b)	$\begin{cases} \lambda^2\tau & \lvert t\rvert > \lambda\tau \\ \lambda\lvert t\rvert & \lvert t\rvert \le \lambda\tau \end{cases}$	$\begin{cases} 0 & \lvert v\rvert \le \lambda \\ \mathrm{sgn}(v)(\lvert v\rvert - \lambda) & \lambda < \lvert v\rvert < \lambda(\tau + \frac{1}{2}) \\ \mathrm{sgn}(v)(\lambda\tau \pm \frac{\lambda}{2}) & \lvert v\rvert = \lambda(\tau + \frac{1}{2}) \\ v & \lvert v\rvert > \lambda(\tau + \frac{1}{2}) \end{cases}$
SCAD, $\tau > 2$, Antoniadis and Fan (2001)	$\begin{cases} \frac{\lambda^2(\tau+1)}{2} & \lvert t\rvert > \lambda\tau \\ \frac{\lambda\tau\lvert t\rvert - \frac{1}{2}(t^2+\lambda^2)}{\tau-1} & \lambda < \lvert t\rvert \le \lambda\tau \\ \lambda\lvert t\rvert & \lvert t\rvert \le \lambda \end{cases}$	$\begin{cases} 0 & \lvert v\rvert \le \lambda \\ \mathrm{sgn}(v)(\lvert v\rvert - \lambda) & \lambda < \lvert v\rvert \le 2\lambda \\ \mathrm{sgn}(v)\frac{(\tau-1)\lvert v\rvert - \lambda\tau}{\tau-2} & 2\lambda < \lvert v\rvert \le \lambda\tau \\ v & \lvert v\rvert > \lambda\tau \end{cases}$
MCP, $\tau > 1$, Zhang (2010b)	$\begin{cases} \lambda(\lvert t\rvert - \frac{t^2}{2\lambda\tau}) & \lvert t\rvert < \tau\lambda \\ \frac{\lambda^2\tau}{2} & \lvert t\rvert \ge \tau\lambda \end{cases}$	$\begin{cases} 0 & \lvert v\rvert \le \lambda \\ \mathrm{sgn}(v)\frac{\tau(\lvert v\rvert - \lambda)}{\tau-1} & \lambda \le \lvert v\rvert \le \lambda\tau \\ v & \lvert v\rvert \ge \lambda\tau \end{cases}$

The thresholding operators S^ρ are defined by

$$S^\rho(v) = \arg\min_{u\in\mathbb{R}} \left((u-v)^2/2 + \rho(u)\right), \tag{3.4}$$

and can potentially be set-valued. Explicit expressions of the thresholding operators S^ρ associated with the three nonconvex penalties (capped-ℓ_1, SCAD and MCP) are as given in Table 1. Note that they are singled-valued, except at $v = \lambda(\tau + \frac{1}{2})$ for the capped-ℓ^1 penalty.

In order to obtain stationary points of the objective function (3.3), we can use the composite gradient descent algorithm (see Nesterov (2007)) and the fact that by our assumptions on f, the vector $\boldsymbol{\gamma}$ has a bounded ℓ_1 norm by M. Denoting $\bar{L}(\boldsymbol{\gamma}) := L(\boldsymbol{\gamma}) - q_{\lambda,\tau}(\boldsymbol{\gamma})$, we may rewrite the minimisation problem (3.3) as

$$\hat{\boldsymbol{\gamma}}_\lambda \in \arg\min_{\|\boldsymbol{\gamma}\|_1 \le M} \left\{\bar{L}(\boldsymbol{\gamma}) + \lambda\|\boldsymbol{\gamma}\|_1\right\}.$$

Then the composite gradient iterates are given by

$$\boldsymbol{\gamma}^{t+1} \in \arg\min_{\|\boldsymbol{\gamma}\|_1 \le M} \left\{\frac{1}{2}\left\|\boldsymbol{\gamma} - \left(\boldsymbol{\gamma}^t - \frac{\nabla\bar{L}(\boldsymbol{\gamma}^t)}{\eta}\right)\right\|_2^2 + \frac{\lambda}{\eta}\|\boldsymbol{\gamma}\|_1\right\}, \tag{3.5}$$

where η is a step-size parameter. Defining the soft-thresholding operator $S_{\lambda/\eta}(\beta)$ component-wise according to

$$S^j_{\lambda/\eta} := \operatorname{sgn}(\gamma_j)\left(|\gamma_j| - \frac{\lambda}{\eta}\right)_+,$$

a simple calculation shows that the iterates (3.5) take the form

$$\boldsymbol{\gamma}^{t+1} = S_{\lambda/\eta}\left(\boldsymbol{\gamma}^t - \frac{\nabla\bar{L}(\boldsymbol{\gamma}^t)}{\eta}\right).$$

Theorem 3 of Loh (2017) guarantees that the composite gradient descent algorithm will converge at a linear rate to the global minimizer as long as the initial point $\boldsymbol{\gamma}^0$ is chosen close enough to the true $\boldsymbol{\gamma}$, since L has an almost-everywhere bounded second derivative and since $q_{\lambda,\tau}$ is convex, as is the case for the penalties in Table 1.

We may now examine the finite sample properties and the empirical quadratic risk of the univariate estimator $\hat{\mathbf{f}} := \mathbf{W}\hat{\boldsymbol{\gamma}}$, where $\hat{\boldsymbol{\gamma}}$ is a stationary point of the objective function (3.3), close to the true wavelet coefficient vector $\boldsymbol{\gamma}$. Our first result below establishes near minimax convergence rates for the quadratic risk of the estimator. Its proof is given in the "Appendix".

Proposition 1 *Suppose $y_i = f^0(x_i) + \epsilon_i$ $(i = 1, \ldots, n)$ for i.i.d. mean zero, sub-Gaussian noise ϵ_i, where f^0 is the true centered signal f. Define the estimator $\hat{\mathbf{f}} := \mathbf{W}\hat{\boldsymbol{\gamma}}_\lambda = [\hat{f}(x_1), \ldots, \hat{f}(x_n)]^T$ where*

$$\hat{\boldsymbol{\gamma}}_\lambda \in \mathit{arg\ min}_{\|\boldsymbol{\gamma}\|_1 \le M} \left\{\bar{L}(\boldsymbol{\gamma}) + \lambda\|\boldsymbol{\gamma}\|_1\right\}.$$

for the $n \times K$-wavelet design matrix $\mathbf{W}$ associated to an r-regular orthogonal wavelet basis with $r > \max\{1, s\}$. Further, define $\mathbf{f}^0 = [f^0(x_1), \ldots, f^0(x_n)]^\top$. Assume that $f^0 \in B^s_{d,q}(M)$ and that the mother wavelet ψ, has r null moments and r continuous derivatives with $r > \max\{1, s\}$. Suppose $\lambda \geq c_1\sqrt{\eta^2 + 2\log K}$ for some $\eta > 0$ and some constant $c_1 > 0$ that may depend on M and the distribution of ϵ_i. Then, for sufficiently large K (specifically $K \geq c_1 n^{1/(2s+1)}$), with probability at least $1 - 2\exp(-\eta^2/2)$, we have

$$R_n(\hat{\mathbf{f}}, \mathbf{f}^0) \leq \mathcal{O}\left(\frac{\log K}{n}\right)^{\frac{2s}{2s+1}}.$$

3.2 An OEM perspective

We consider here a similar setting as the one in Antoniadis and Fan (2001). Since n is not necessarily a power of 2, and the design is not regular, we start by embedding the observed data points $\{x_1, \ldots, x_n\}$ into a fine equispaced grid by taking $N = 2^J$ dyadic grid points with J is a chosen resolution, which we denote t_ℓ, $\ell = 1, \ldots, N$ (since we are working on the unit interval [0, 1] this would mean taking $t_\ell = \ell/N$). The number N should be much larger than the sample size n (i.e $N \gg n$). Embedding here means that we are approximating the set of sampling points $\{x_1, \ldots, x_n\}$ with a subset of dyadic points at the scale J, that is we assume that $x_i \approx n_i/N$ for $i = 1, \ldots, N$ for some $1 \leq n_i \leq N$ (i.e. n_i/N is closest design point to the point x_i). It is easy to see that when N is much larger than n, and given that the sampling points x_i are i.i.d. realizations of a probability distribution on [0, 1] with a density that is continuous and bounded below by a strict positive constant, the approximation errors by moving non dyadic points to dyadic points is negligible. Let $\tilde{\mathbf{f}}_N$ be the underlying signal f collected at all dyadic points $\{\ell/N, \ell = 1, \ldots, N\}$, that is $\tilde{\mathbf{f}}_N = (f(t_1), \ldots, f(t_N))^T$ and let this time $\mathbf{W}$ be a given $N \times N$ orthogonal discrete wavelet transform (DWT). The wavelet transform of $\tilde{\mathbf{f}}_N$ is the vector $\boldsymbol{\beta}$ of discrete wavelet coefficients given by $\boldsymbol{\beta} = \mathbf{W}\tilde{\mathbf{f}}_N$. By orthonormality of the discrete wavelet transform we have $\tilde{\mathbf{f}}_N = \mathbf{W}^\top\boldsymbol{\beta}$.

Let $\mathbf{S}$ be the $n \times N$ sampling matrix which extracts the $[n_1, \ldots, n_n]$th entry from $\tilde{\mathbf{f}}_N$, then $\mathbf{f} = \mathbf{S}\tilde{\mathbf{f}}_N$ is the discrete signal at the sampling points. Denoting by $\mathbf{A} = \mathbf{S}\mathbf{W}^\top$ the observed data, up to a negligible approximation error, obeys the linear model

$$\mathbf{Y} = \mathbf{A}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{3.6}$$

where $\boldsymbol{\epsilon}$ is the noise vector which again is assumed to be independent of the entries of $\mathbf{A}$. The linear model derived from the observations is over-parameterized, but we have already noticed that for functions belonging to the Besov bodies adopted in this work, their wavelet coefficients are sparse. We can therefore estimate $\boldsymbol{\beta}$ by regularization.

We have already concisely reviewed in our previous subsection a coordinate descent Newton-type method that solves problem (3.6). In a similar spirit, for penalty functions such as the capped-ℓ_1, the MCP and the SCAD, Xiong et al. (2017) proposed a novel orthogonalizing expectation maximization (OEM) algorithm to solve the least squares problems in linear regression when the regularizers are separable over the variables.

Inspired by Daubechies et al. (2004) where the idea below was mentioned for the first time and by a recent work of Liu et al. (2020), we suggest a new perspective of OEM algorithm with emphasis on the capability of separating variables to avoid the computation of matrix inverse. With this, an algorithm is proposed to solve problem (3.6) and to apply it for wavelet denoising. The advantage again is that the minimization of (3.6) will be carried out by decomposing it as a sequence of univariate minimizations that then can be efficiently computed via the univariate thresholding operators reviewed above.

For a problem in (3.6) with a separable penalty $P_{\lambda,\tau}(\boldsymbol{\beta}) = 2\sum_{i=1}^{K} \rho_{\lambda,\tau}(\beta_i)$ denote by $\hat{\boldsymbol{\beta}}$ the final solution and introduce a function

$$\mathrm{RSS}(\boldsymbol{\beta}) = \|\mathbf{Y} - \mathbf{A}\boldsymbol{\beta}\|^2 + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top \mathbf{F}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + 2\sum_{i=1}^{K} \rho_{\lambda,\tau}(\beta_i),$$

where $\mathbf{F} = d\,\mathbf{I}_p - \mathbf{A}^\top\mathbf{A}$ and d is a positive integer such that $\mathbf{F}$ is positive semidefinite. Quite clearly, $\mathrm{RSS}(\boldsymbol{\beta})$ can also be minimized at $\hat{\boldsymbol{\beta}}$. On the other hand, by some simple algebraic computation, it is equally clear that, up to some additive constant

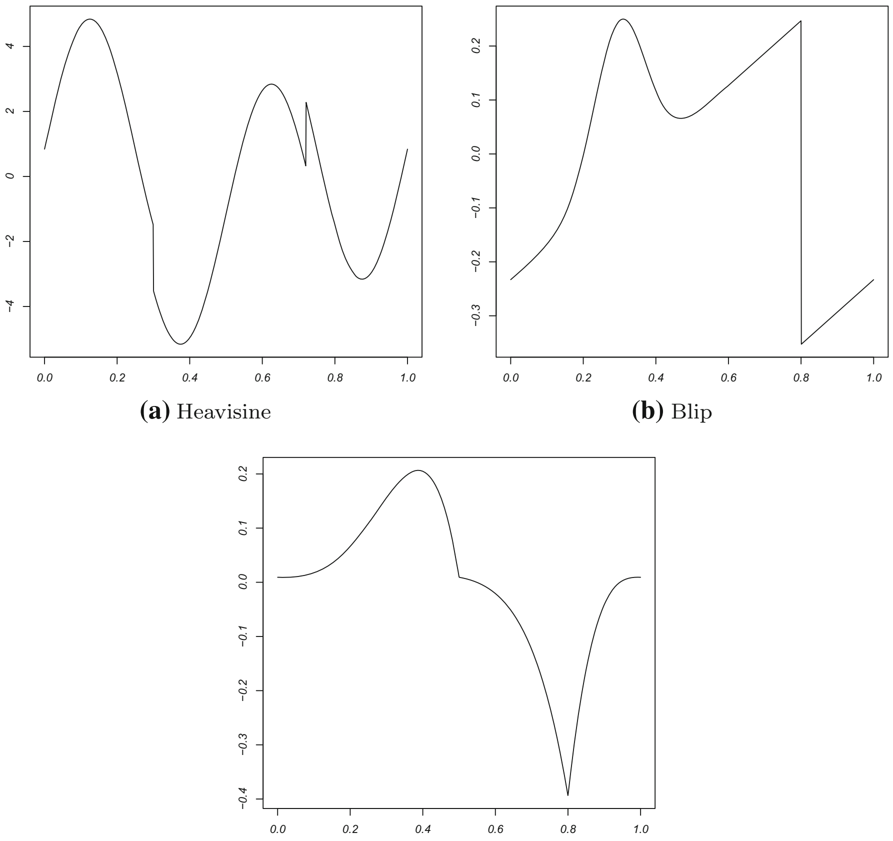
$$\mathrm{RSS}(\boldsymbol{\beta}) = d\sum_{i=1}^{K}\left(\beta_i - \tilde{\beta}_i\right)^2 + 2\sum_{i=1}^{K}\rho_{\lambda,\tau}(\beta_i),$$

where $\tilde{\beta}_i$ is the ith element of the dual $\tilde{\boldsymbol{\beta}} = \mathbf{A}^T\mathbf{Y}/d + \mathbf{F}\hat{\boldsymbol{\beta}}/d$. This gives rise to the following relationship:

$$\hat{\beta}_i = \arg\min_{\beta_i}\left[d\left(\beta_i - \tilde{\beta}_i\right)^2 + 2\rho_{\lambda,\tau}(\beta_i)\right], \quad i = 1, 2, \ldots, K$$

which involves only a single variable and can be done by applying component wise in an iterative manner the thresholding operator (3.4). From the discussion from Xiong et al. (2017), any choice of $d \geq \sigma_{\max}\left(\mathbf{A}^\top\mathbf{A}\right)$ will make the algorithm work in the sense that, under our assumptions on f and the regularity of the wavelet basis used in constructing $\mathbf{W}$, the thresholding iterations converge again to a stationary point of problem (3.6). We may, again, state the following proposition (very similar to Proposition 1 stated before), whose proof is omitted since it is a direct application of Theorem 4 of Antoniadis and Fan (2001).

Proposition 2 *Suppose $y_i = f^0(x_i) + \epsilon_i$ $(i = 1, \ldots, n)$ for i.i.d. mean zero, Gaussian noise ϵ_i with variance σ^2, where f^0 is the true centered signal f. Define the estimator $\hat{\mathbf{f}} := \mathbf{W}\hat{\boldsymbol{\gamma}}_\lambda = [\hat{f}(x_1), \ldots, \hat{f}(x_n)]^T$ where $\hat{\boldsymbol{\gamma}}_\lambda$ is the global separable minimizer obtained through the OEM thresholding iterations, with $\mathbf{W}$ the column normalized $n \times K$-wavelet design matrix associated to an r-regular orthogonal wavelet basis with $r > \max\{1, s\}$ (i.e. the mother wavelet ψ, has r null moments and r continuous derivatives with $r > \max\{1, s\}$). Suppose $\lambda = \mathcal{O}(\sqrt{\log n})$. Then the penalized least squares estimator $\hat{\mathbf{f}}$ over the Besov body $B^s_{p,q}(M)$ is of ℓ_2-rate $\mathcal{O}(n^{-2s/(2s+1)}\log n)$.*



(a) Heavisine

(b) Blip

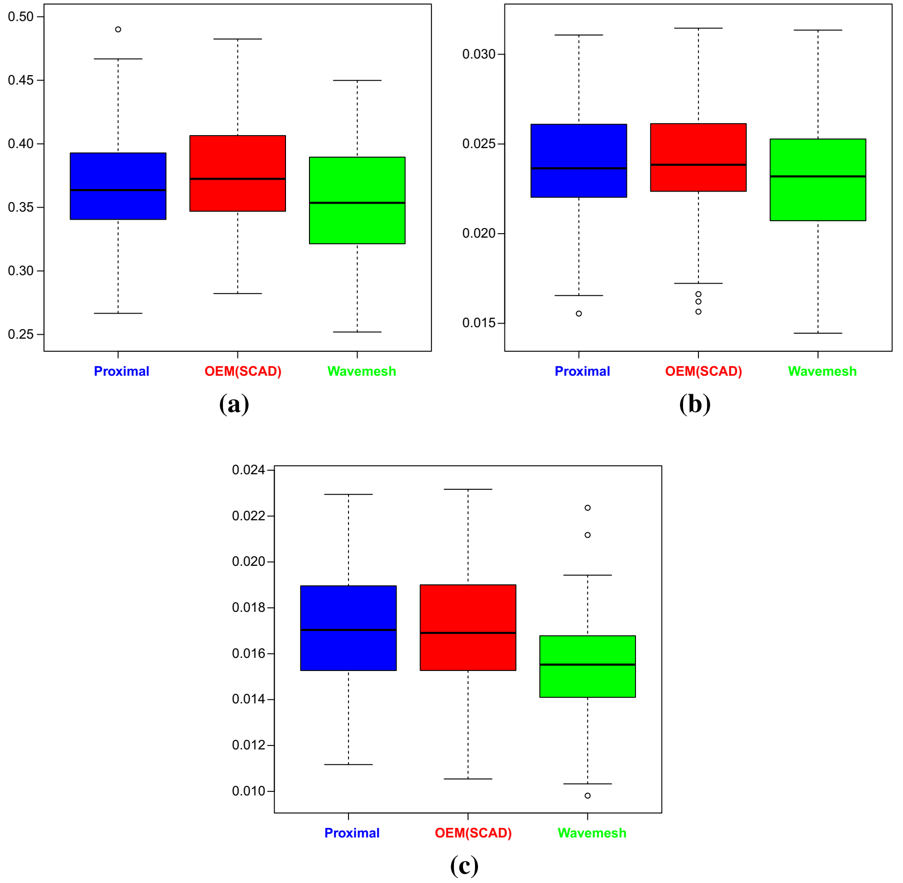
(c) Corner

Fig. 1 Testing functions used in the simulations

We may conclude that both versions of the univariate wavelet based estimators studied above, under appropriate assumptions, present optimal prediction performance. The next Sect. 3.3 reports an numerical comparison of their empirical performances, compared also to another recently developed wavelet based estimator, called Wavemesh (see Haris et al. (2018)).

3.3 Experimental comparison setup

The simulation setup to illustrate the above univariate wavelet based denoising procedures is as follows. Three testing functions, namely, Heavisine, Blip and Corner are used that represent a variety of function characteristics (see Fig. 1). A relatively low level of signal-to-noise ratio (SNR) equal to 4 is chosen, where SNR is defined as SNR=std$(f)/\sigma$. Two different sample sizes $n = 100$ and $n = 300$ are used. Nonequispaced sample points x_i, $i = 1, \ldots, n$, are generated according to, either a uniform distribution on [0, 1] (case (I)) or a Gaussian distribution $N(2, 1)$ whose values are



(a) **(b)** **(c)**

Fig. 2 Boxplots of the RMSE for each of the estimates when $n = 100$ and the design is uniform

normalized to have a range between [0, 1] (case (II)). Without any loss of generality the test functions evaluated at the design points are centered.

For each of the above experimental factor combinations, 50 data sets $\mathbf{Y}^h \in I\!R^n$, $h = 1, \ldots, 50$ are generated according to the model $Y_i = f(x_i) + \sigma\epsilon_i$, $i = 1, \ldots, n$, where ϵ_i's are i.i.d. standard normal distributed and σ is chosen to achieve a signal-to-noise ratio $SNR = 4$. For each simulated data set $\mathbf{Y}^h$, we applied each of the three methods of this subsection to estimate the test signal with optimal tuning parameters chosen from pre-specified grids. The root mean square error (RMSE) of the solution $\hat{\mathbf{f}}$ and the true signal $\mathbf{f}$ at the nonequispaced design points x_i is used to assess the quality of the estimates.

Boxplots of the RMSEs of the estimates obtained by the three methods for the design case (I) are presented in Figs. 2 and 3 for Heavisine, Blip, and Corner, respectively, for each sample size used.

When the sample size is small ($n = 100$) our proximal method is almost equivalent to the OEM based method and slightly worse than wavemesh. This is due to the fact

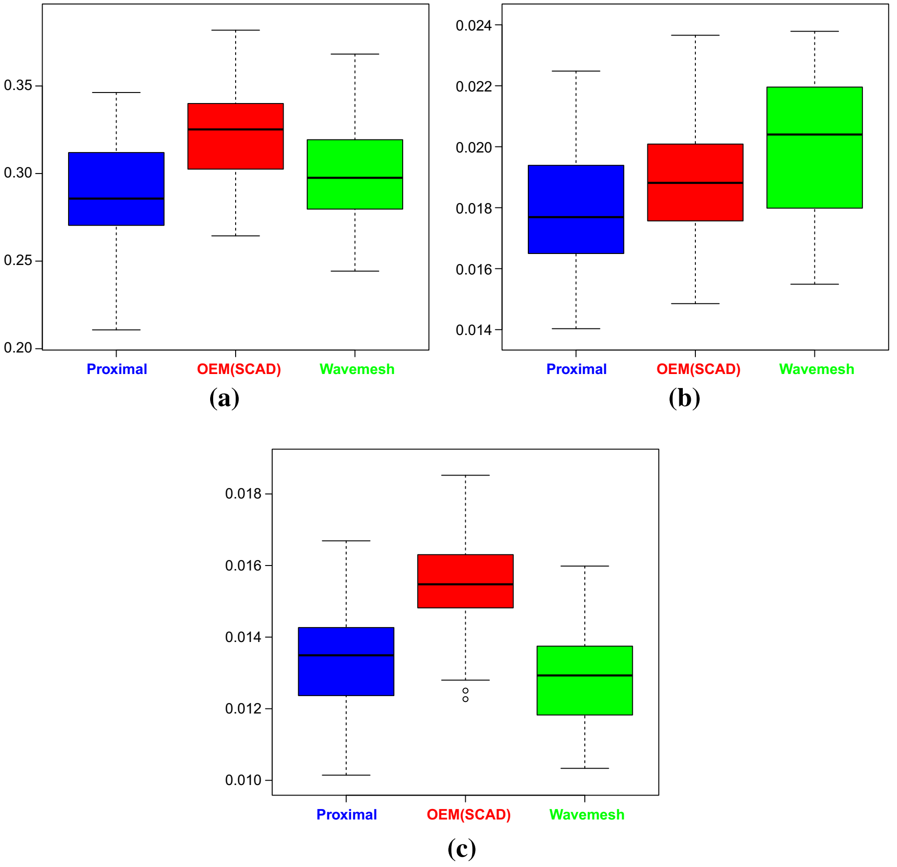


Fig. 3 Boxplots of the RMSE for each of the estimates when $n = 300$ and the design is uniform

that with such a sample size both proximal and OEM tend to over smooth the cusps and discontinuities present in the signals especially for the Heavisine function. The behaviour of the OEM based method is also expected since it is very close to the ROSE one-step estimator of Antoniadis and Fan (2001), whose optimal properties are mainly asymptotic.

When $n = 300$, all three methods have an excellent behavior. Contrary to the small sample case, our proximal estimator performs better than the two other methods for the signal Heavisine and Blip and is slightly worse than wavemesh for the signal Corner.

When the design grid is generated according to case (II), each of three methods of this subsection to estimate the test signal with optimal tuning parameters chosen from pre-specified grids is less sensitive to boundary effects since the tails of data within the interval [0, 1] are lighter by design.

Boxplots of the RMSEs of the estimates obtained by the three methods for the design case (II) are presented in Figs. 4 and 5 for Heavisine, Blip, and Corner, respectively, for each sample size used.

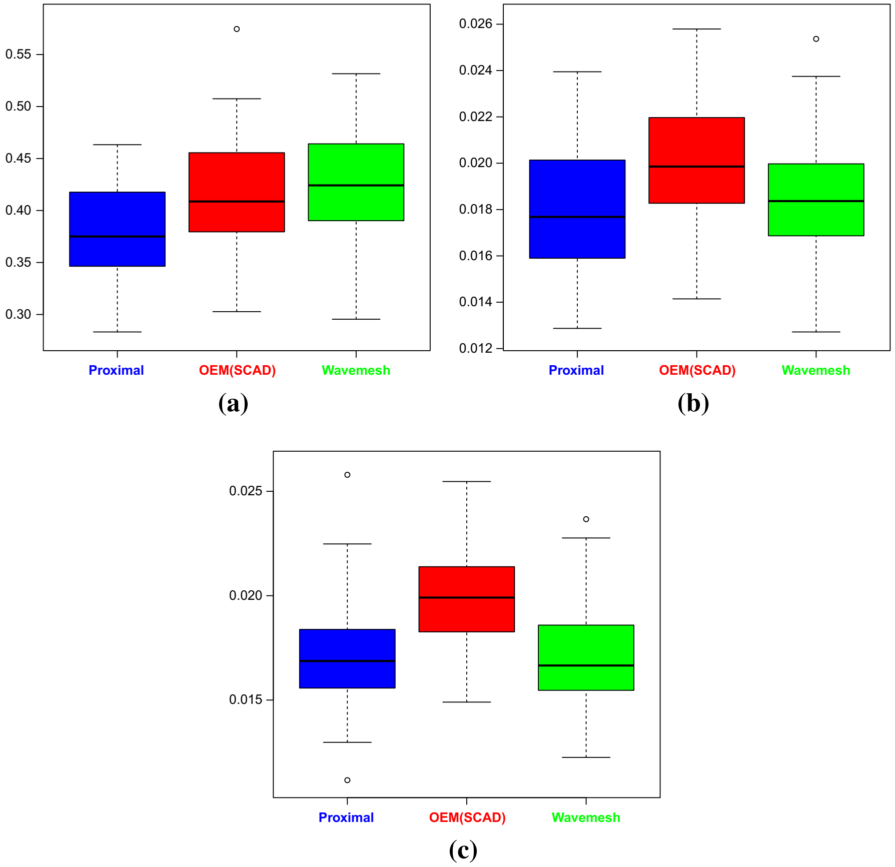


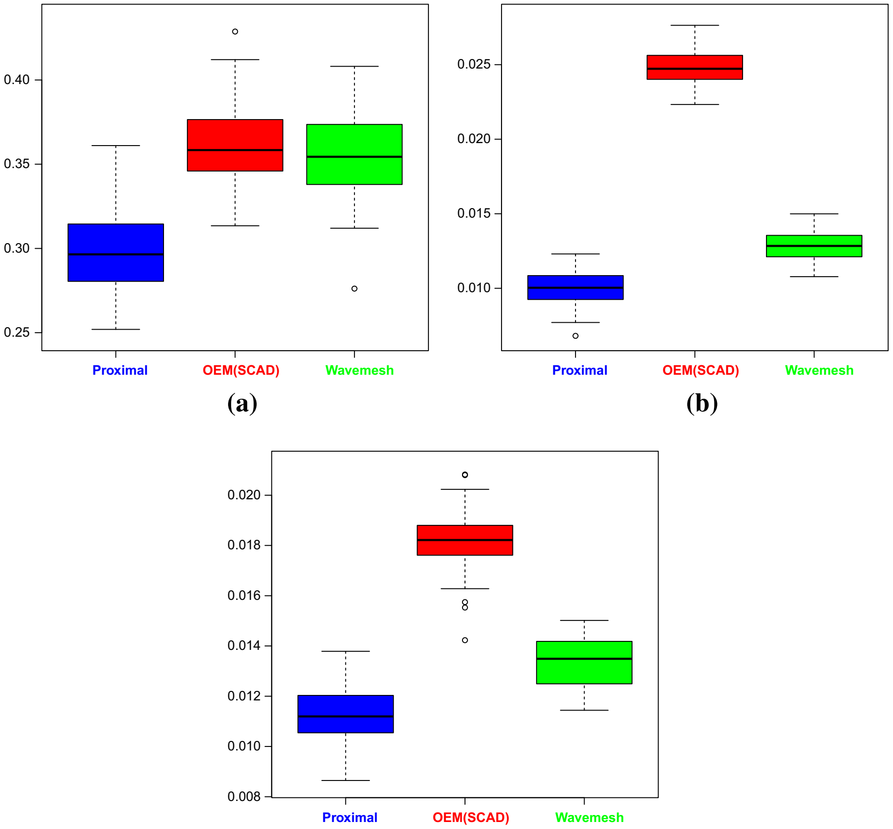
Fig. 4 Boxplots of the RMSE for each of the estimates when $n = 100$ for the case (II) design

When the sample size is small ($n = 100$) our proximal method is better than the the two other methods.

When $n = 300$, contrary to the small sample case, our proximal estimator performs always much better than the two other methods and wavemesh is better than OEM. It is worth to outlining that, in the average, the fastest procedure in CPU time is OEM followed by wavemesh and then proximal.

4 Robust wavelet smoothing

The nonparametric univariate regression procedures developed and studied in our previous subsections are well adapted to the cases of not uniformly spaced design points, sample sizes that are not a power of 2 and a standard Gaussian or Gaussian alike noise term, and mainly derived as penalized regression estimations using a penalized least squares loss function.The attraction of such procedures is that the unknown "inhomogeneous" regression function can then be estimated with optimal convergence



(a)

(b)

(c)

Fig. 5 Boxplots of the RMSE for each of the estimates when $n = 300$ for the case (II) design

rates. In practice however, some extreme observations may occur, and estimation using an unbounded loss function suffers from a lack of robustness, meaning that the estimated functions can be distorted by the outliers. Both the nonparametric function estimates themselves and the choice of the penalization parameters associated to them are affected. Indeed, when noise has a non-normal distribution (or a sub-Gaussian one as in Proposition 1), for instance, heavy-tailed distribution, even for data observed on a regular grid, the classical wavelet shrinkage is not efficient for estimating the true function. We may mention here some robust wavelet procedures that have been previously studied in the equidistant dyadic sample case. Bruce et al. (1994) proposed a robust wavelet transform to estimate wavelet coefficients at each resolution level which is robust to outliers. Kovac and Silverman (2000) developed a direct approach. Their approach can be summarized as identifying outliers with a classical robustness test, removing the outliers, and applying their thresholding procedure for unequally spaced data. However, their approach has a drawback in that information is lost by removing data. Sardy et al. (2001) proposed a robust wavelet de-noising estimator based on a robust M-type loss function. Their procedure is computationally intensive as it involves solving a non trivial nonlinear optimization problem. Averkamp and Houdré (2003) extend the minimax theory for wavelet thresholding to some broader classes of symmetric heavy-tailed distributions, but is only applicable if the distribution is known. Oh et al. (2007) studied a robust wavelet shrinkage algorithm that is based on the concept of pseudo data. However the selection of a robust threshold and the asymptotic properties of the resulting estimators were not addressed in their paper.

To address the lack of robustness, we introduce a new class of wavelet based robust estimators for performing estimation in sparse "inhomogeneous" univariate regression models. Our emphasis is on M-type penalized smoothing in a spirit much similar to the one adopted in our previous subsections. The class of estimators derived here are minimizers of a criterion that balances fidelity of a wavelet basis expansion to the data with sparseness/smoothness in the estimate. We will use for such a task, the methods that have been developed (Amato et al. 2021), which under high-dimensional sparse linear models, and for appropriate robust loss functions, are able in handling optimal, both computationally and asymptotically, predictions.

We will focus again at the linear wavelet regression model, stated, with a negligible approximation error, in (3.1):

$$\mathbf{Y} = \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)^T \in \mathbb{R}^n$ is the vector of responses, $\mathbf{W} \in I\!R^{n\times K}$ is the wavelet basis regression design matrix, and $\boldsymbol{\gamma} \in I\!R^K$ is the vector of wavelet regression coefficients. The difference here is that we assume that the noise vector $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)^T \in I\!R^n$ is having i.i.d components with a symmetric around zero distribution $F = F_\epsilon$ with a finite second moment which deviates from the Gaussian distribution, in the sense that it is a more or less heavy-tailed distribution producing outliers in the responses. As before the design points may be either deterministic or random. Therefore the parameter of interest $\mathbf{W}\boldsymbol{\gamma}$ in (3.1) is the mean of $\mathbf{Y}$ in the deterministic case or the conditional mean of $\mathbf{Y}$ on $\mathbf{X}$ in the random design case.

We will assume again that the unknown regression f to be recovered belongs to a Besov body $B^s_{d,q}(M)$ which implies in particular that the weak sparsity $K = K_n$, while allowed to tend to infinity with n is such that $n \geq c_0 K_n \log n$, for a sufficiently large constant c_0. By known information-theoretic results of Wainwright (2009), this type of lower bound is required for any method to recover the support of a K-sparse signal, hence is not a limiting restriction.

Since the error distribution F_ϵ deviates from the normal distribution, the standard unpenalized squared error loss function used in the previous subsections is typically changed to the log-likelihood $-\log f_\epsilon$, where f_ϵ is the density of F_ϵ, leading to penalised log-likelihood estimators. Unfortunately, in real life situations the error distribution F_ϵ is unknown and methods that adapt to many different distributions are needed. A large class of penalized M-estimators has been recently proposed by Amato et al. (2021). The methods in this class penalise an empirical loss function $\mathcal{L}_n$ in the following manner

$$\hat{\boldsymbol{\gamma}}(\lambda) = \underset{\boldsymbol{\gamma}\in I\!R^K, \|\boldsymbol{\gamma}\|_1 \leq R}{\arg\min} \left\{\mathcal{L}_n(\boldsymbol{\gamma}) + p_\lambda(\boldsymbol{\gamma})\right\}, \tag{4.1}$$

where $\mathcal{L}_n(\boldsymbol{\gamma}) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i - \mathbf{W}_i^T\boldsymbol{\gamma})$ for a chosen scaled loss ℓ and for a suitable penalty function p_λ. The side condition $\|\boldsymbol{\gamma}\|_1 \leq R$ for some positive number R guarantees the existence of local/global optima in cases when the loss and the penalty may be nonconvex. Such a condition is satisfied here by the fact that $f \in B^s_{p,q}(M)$ so the true regression vector $\boldsymbol{\gamma}$ is feasible for the optimisation problem (4.1).

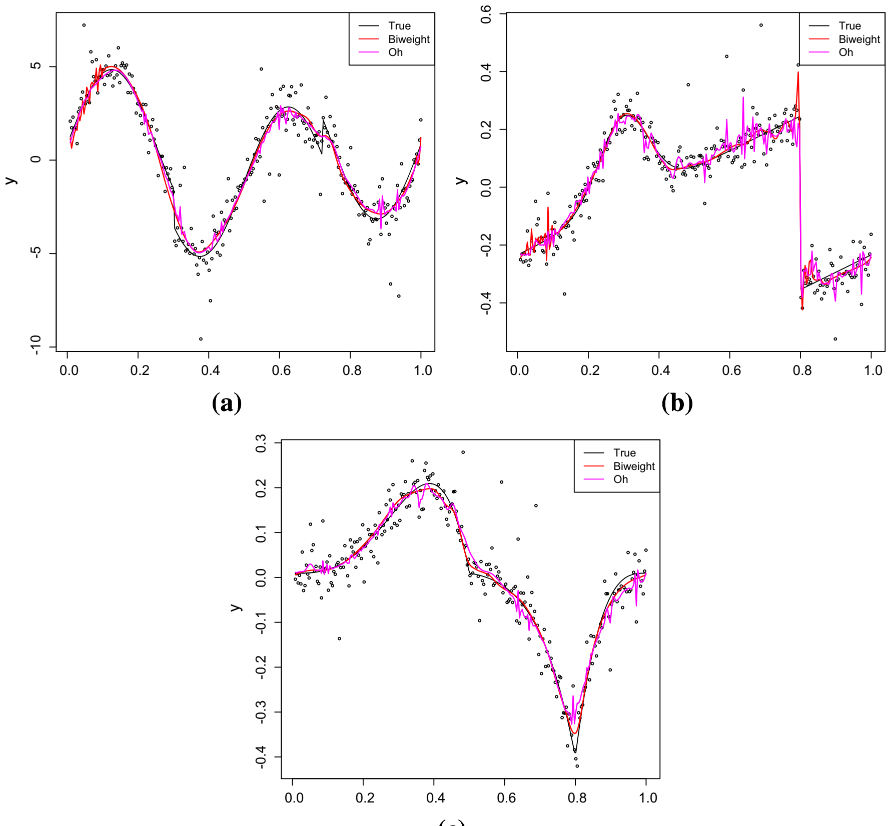
The loss functions that we will adopt in this subsection include Tukey's biweight and Welsh's loss (see Amato et al. (2021), Appendix 1). One could also use some other popular losses such as the least absolute loss or a Huber loss, but their comparison with Tukey's biweight and Welsh's loss made in Amato et al. (2021), favor the latter. An unifying property of these loss functions is that the derivative ℓ' is *bounded* and odd in each case which turns out to be an important property for robustness of the resulting estimator because it is lessening the contamination effect of gross outliers.

Since for the class $\mathcal{R}$ of robust loss functions that we consider also have $\ell''(u) > -\kappa$ for all $u \in I\!R$ and some constant $\kappa > 0$, and the density of the design is continuous and bounded below by a strict positive constant, the local RSC condition, already discussed in our Remark 2, for these loss functions holds with probability at least $1 - c\exp(-c_2 \log K_n)$ for some constant $c > 0$ as soon as K_n tends to infinity and $n > c_0\, K_n \, \log n$. Since the penalties used are μ-amenable, by Theorem 4.1 of Amato et al. (2021) the resulting wavelet regression estimator is asymptotically consistent. Moreover, the estimate can be again computed with a variant of the numerical algorithms described in Amato et al. (2021).

4.1 Some supplementary simulations

This subsection reports the results from a small simulation study designed to assess the practical performance of the two robust methods proposed above and compare them to the robust wavelet shrinkage method of Oh et al. (2007). Since the Oh procedure is only designed for a dyadic equispaced grid the simulations below are designed with an equispaced deterministic design with a sample size equal to 256. Note however, that our robust methods can be applied to any type of design which is not true for Oh et al's method.

The experimental set-up is as follows. The three test functions 'Heavisine', 'Blip' and 'Corner' that we have also used before were selected for this illustration. A sample size, $n = 256$, a dimension $K = \frac{n}{2}$ and a contaminated Gaussian mixture, $0.9\mathrm{N}(0, \sigma)+ 0.1\mathrm{N}\,(0, 4\sigma))$, as a noise distribution, were considered with a value of σ such that, when no outliers are present, a SNR of 4 is realized. Note here that such a distribution for the noise covers the assumptions of Theorem 4.1 of Amato et al. (2021). For each combination of test function and noise, 50 samples were generated. Then, for each generated sample, three regression estimates were obtained by applying, respectively, the robust estimator based on Tukey's Biweight loss function and penalized with an MCP penalty, the robust estimator based on Welsh's loss function and penalized with an MCP penalty and the robust method by Oh et al.. To scale the loss functions an estimate of the noise scale $\hat{\sigma}^2$ adopting a good robust estimate of the noise variance (see Amato et al. (2021)) is used. The optimal penalty is chosen over a fine grid of values. For each value of the penalty parameter on the grid, a robust model selection



(a) **(b)**

(c)

Fig. 6 Typical robust estimations using Biweight loss and Oh method for contaminated data of size $n = 256$

using Biweight or Welsh is performed and the resulting estimator is plugged into the BIC-type criterion, where the degrees of freedom are the number of nonzero estimated wavelet coefficients. The optimal penalty parameter and the corresponding selected model are the one minimizing the BIC criterion over the chosen grid. For Oh's procedure the modified universal thresholding is chosen according to the proposal in Oh et al. (2007).

The root mean-squared error, $RMSE = \sqrt{MSE}$, where the empirical mean square error is given by $MSE = \frac{1}{N} \sum \left\{ f_0(x_i) - \hat{f}(x_i) \right\}^2$, was calculated for each regression estimate. Figures 6 and 7 display some typical fits obtained by each of these methods for each of the test functions. The boxplots of the resulting RMSE for each scenario are displayed in Fig. 8.

The following major observations can be made. First, the proposed procedures, while less variable, are slightly worse than the Oh method for the Heavisine example, but remain very comparable. However, they both outperform the Oh method for the other two test functions. A likely explanation is that the Heavisine function has a less

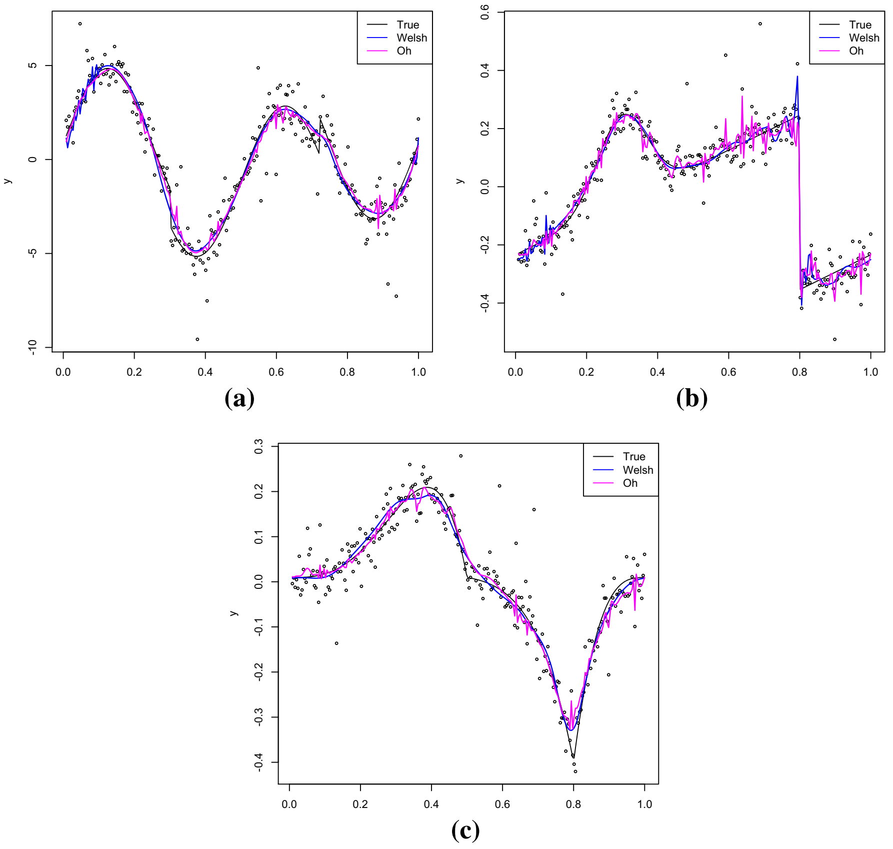


Fig. 7 Typical robust estimations using Welsh loss and Oh method for contaminated data of size $n = 256$

sparse representation in the wavelet domain compared to Blip and Corner. Having in mind that the other robust estimates can handle non dyadic and non regular designs they can be therefore considered superior.

4.2 Real data examples

The proposed univariate wavelet procedures were first applied to the glint dataset analyzed by Sardy et al. (2001). The data are radar glint observations from a target captured at $N = 512$ angles of an object in degrees. The signal contains a number of glint spikes, causing the apparent signal to behave erratically. From physical considerations, a good model for the true signal is a frequency oscillation about 0°. The data together with the regression estimates are displayed in Fig. 9. The median filter's estimate still shows a highly oscillating denoised signal. It can be seen that BiWeight is resistant to the adverse effects of outliers, notably at target angles near 5°, 90°, 140°, 200°, 320° and the range from 420° to 470°, whereas the nonrobust estimate remains jagged.

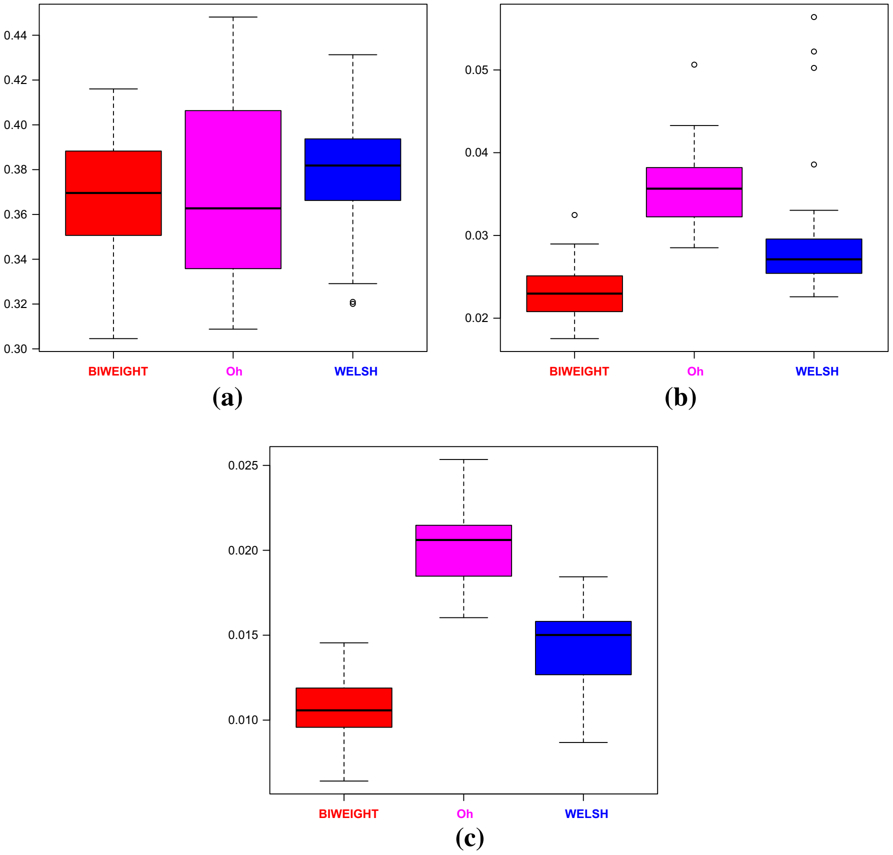


Fig. 8 RMSE of the 50 estimations using Biweight and Welsh loss and Oh method for contaminated data of size $n = 256$. **a** Heavisine; **b** Blip; **c** Corner

We have also applied the proposed methods to the balloon data used by Kovac and Silverman (2000). The data are radiation measurements from the sun, performed from a flight of a weather balloon. Since the measurement device was occasionally cut off from the sun, individual outliers and large outlier patches are introduced. We select every second or third observation from the 4984 observations to reduce the sample size to 2048. Although the data are very closely located to a regular grid, they are not exactly equally spaced. Therefore our wavelet procedures on an irregular design are appropriate to fit the data. Figure 10 shows the balloon data with a sample size of 2048 points and the estimates by a moving median filter with window size 14, the non robust MCP penalized wavelet procedure and the robust fitting based on MCP BiWeight, respectively. Note that the window size of the median filter is selected by eye to be "good". Increasing the window size does not help capturing the valley around 0.77 and decreasing the size makes the fit quite wiggly. The fit by the penalized MCP is still affected by the outliers. On the other hand, the robust fitting based MCP BiWeight reacts to the major pattern of the observations.

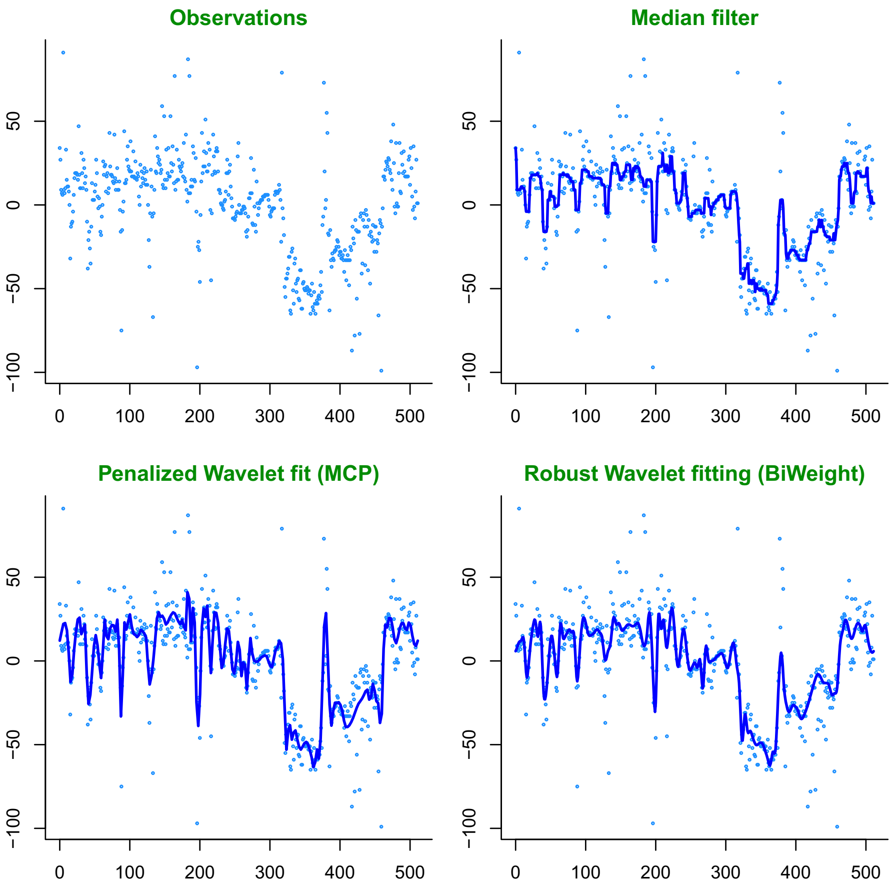


Fig. 9 Fittings for glint data

Finally we show two use cases arising in the Horizon 2020 MADEin4 project. The project aims at developing next generation metrology tools, machine learning methods and applications in support of Industry 4.0 high volume manufacturing in the semiconductor manufacturing industry. It is headed by STMicroelectronics, leader company in semiconductor industry.

The first use case concerns monitoring some moving holders that block silicon wafers into an equipment during the growth of semiconductors in a particular process step for semiconductor device fabrication, with the objective of predicting their malfunctioning. Monitoring occurs by collecting images of the holders and processing them. As result of this analysis, an indicator has been devised that measures how much an image is different from a reference one. Presence of artifacts on the holders generated during the process unduly increase the error indicator. Therefore it is needed to track behaviour of the indicator as a function of time, cleaned by these artifacts. Since measurements are not taken every day, a nonequispaced time series is generated. Figure 11 shows the sequence of data points and the fit obtained by a Median filter and

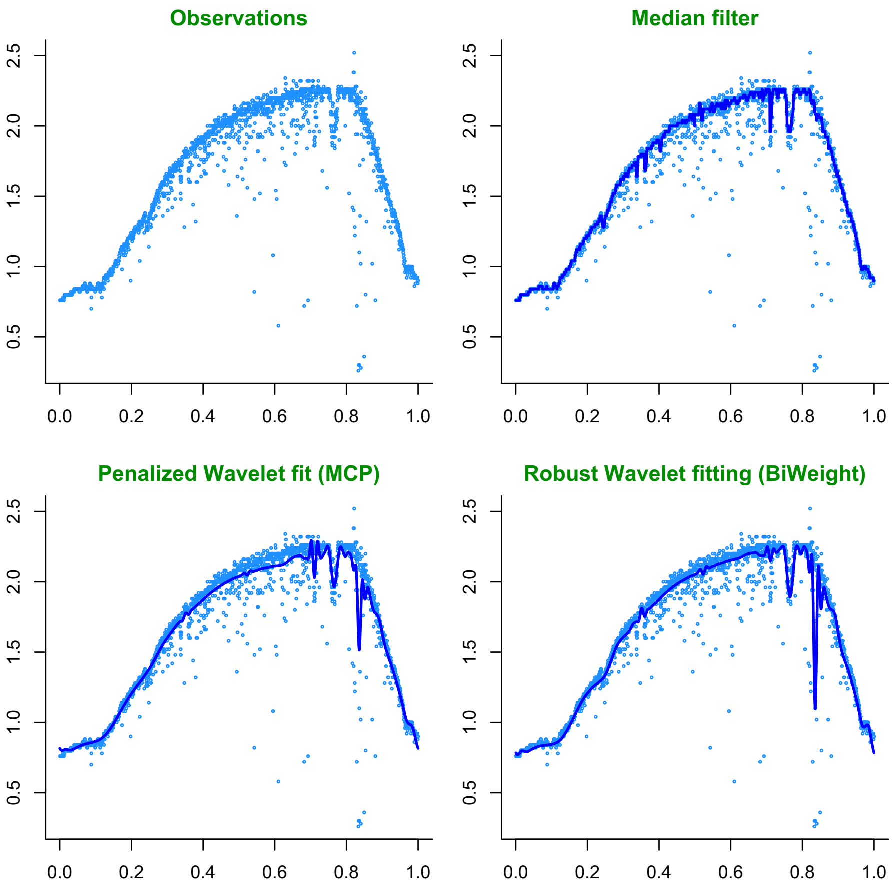


Fig. 10 Fittings for balloon data

robust fitting by Welsh and BiWeight loss based on MCP. All fits show to be unaffected by the large outlier with indicator above 9, with Welsh loss function yielding back a smoother function.

In the second use case on data provided by STMicroelectronics, physical parameters were measured and collected over time, during another particular process step of semiconductor device fabrication. Data have to be further processed in a Virtual Metrology setting. The generated time sequences are not equispaced in time. It is interesting to fit such sequences, removing noise eventually, and approximating on an equispaced grid, introducing least correlation than interpolation, so to have more flexibility in processing them and extracting specific Features. Figure 12 shows data for a typical sequence of one specific parameter. The length of the sequence is 1714 data. The figure also shows the fits obtained by Penalized wavelet estimation (MCP) and robust Wavelet fitting (with BiWeight and Welsh loss function based on MCP).

Indeed the data do not show outliers, therefore a nonrobust approach seems more suited to represent the data with a greater accuracy.

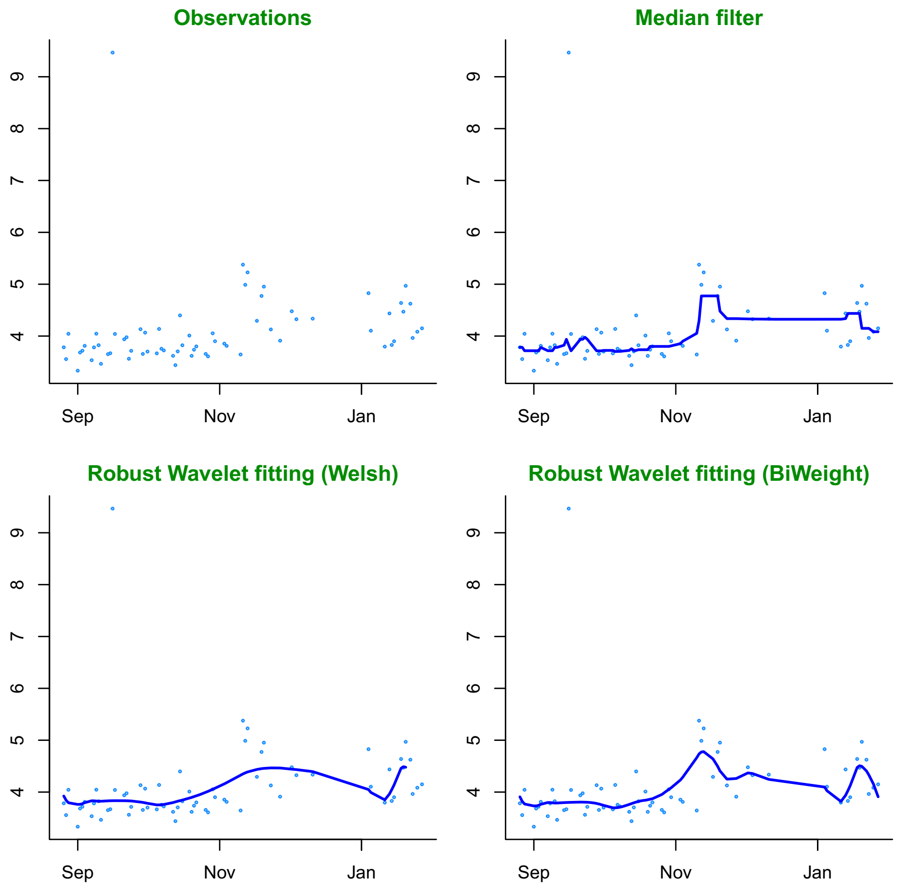


Fig. 11 Fits of the holder data arising in the process of growth of semiconductors

The results from all these real examples suggest that our methods possess promising empirical properties.

5 Conclusions and discussion

This paper introduced some wavelet-based method for nonparametric estimation of general univariate nonlinear regression models. The nonlinear regression functions in such models present a wide range of irregular effects in terms of smoothness and are approximated by appropriate truncated expansions on wavelet basis functions evaluated at the design sampling points. Unlike traditional methods the methods developed here do not require that the design points are uniformly spaced on the unit interval, nor do they require that the sample size is a power of 2. Adopting a matrix notation, we provide a unifying framework for achieving the estimation of the unknown regression

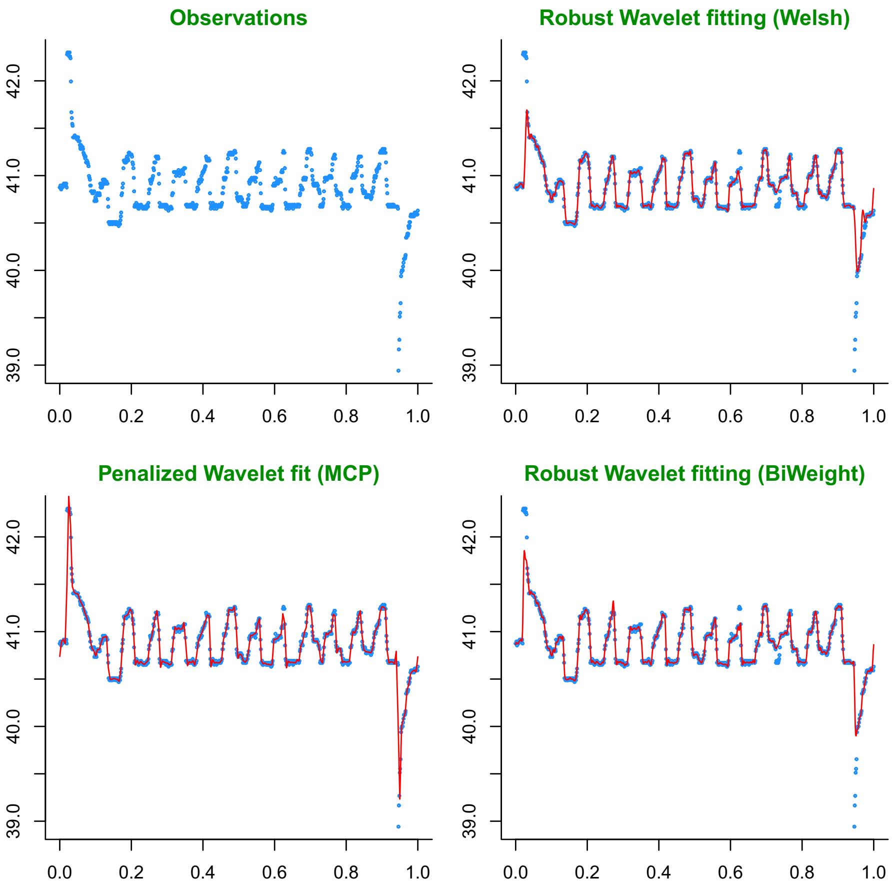


Fig. 12 Fits of a Parameter of Process data arising in the production of semiconductors

functions by adapting general algorithms for penalized linear models and provide a solid theoretical framework for establishing mean-squared estimation consistency.

Although in the paper we restrict to a univariate regression setting, the methodology presented can be extended to more flexible settings, which could be needed when dealing with more complex data, such as, for example, additive nonlinear regression models with irregular components. We would like to shortly discuss here a possible way to achieve it. To fix the ideas, assume that we now have at our disposal observations $(y_i, \mathbf{x}_i), i = 1, \ldots, n$, from an additive regression model of the form $Y_i = \mu + \sum_{j=1}^{p} f_j(X_i^j) + \varepsilon_i, i = 1, \ldots, n$, where $\mu \in I\!R$ is an overall mean parameter, each f_j is a univariate function, a mean-zero error vector $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ with independent components which is also independent of the covariates. Each of the p univariate functions f_j could then be approximated by their approximation in a wavelet basis as in this paper. By stacking all involved wavelet basis matrices in a big design matrix, the problem to be studied can be written as for the linear setting used in our paper and estimation could follow by adapting our methods using grouped penalization methods and extensions of the definition of proximal operators to vector-valued arguments. We postpone a detailed study of such an extension for future research.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00180-021-01174-4.

Acknowledgements We thank the associate editor and the three anonymous referees for their detailed and constructive comments on an earlier version of the article. U. Amato gratefully thanks P. Vasquez and her co-workers from STMicroelectronics for having provided data for the use cases on semiconductors and for their continuous support to the research. Part of this work was completed while A. Antoniadis and I. Gijbels were visiting the Istituto per le Applicazioni del Calcolo ‘M. Picone’, Italian National Research Council, Naples. I. Gijbels gratefully acknowledges financial support from the C16/20/002 project of the Research Fund KU Leuven, Belgium. I. De Feis acknowledges the INdAM-GNCS 2020 Project “Costruzione di metodi numerico/statistici basati su tecniche multiscala per il trattamento di segnali e immagini ad alta dimensionalità”.

Funding The work presented in this paper has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No. 826589, the MADEin4 project. The JU receives support from the European Union’s Horizon 2020 research and innovation programme and Netherlands, Belgium, Germany, France, Italy, Austria, Hungary, Romania, Sweden and Israel.

Code availability The R codes and their description implementing the methods in the paper together with the driver scripts for simulations, plots and the analysis of real examples are made available in a compressed archive as supplementary material.

Declarations

Conflict of interest Not applicable.

Availability of data and materials Upon request to the authors.

Appendix: Technical proofs

We present here the proof for Proposition 1. Recall that the estimator is given by $\hat{\mathbf{f}} := \mathbf{W}\hat{\boldsymbol{\gamma}}_\lambda = [\hat{f}(x_1), \ldots, \hat{f}(x_n)]^T$ where

$$\hat{\boldsymbol{\gamma}}_\lambda \in \arg\min_{\|\boldsymbol{\gamma}\|_1 \leq M} \left\{ \bar{L}(\boldsymbol{\gamma}) + \lambda \|\boldsymbol{\gamma}\|_1 \right\}.$$

for the $n \times K$-wavelet design matrix $\mathbf{W}$ associated to an r-regular orthogonal wavelet basis with $r > \max\{1, s\}$, where $\boldsymbol{\gamma}$ denotes the corresponding vector of wavelet coefficients f^0. A nice feature of the above optimization problem is that it is exactly the lasso problem with design matrix $\mathbf{W}$ and loss function the convex functional $\bar{L}$.

Proof of Proposition 1 We follow closely the results from van de Geer (2014). Let μ_ϵ satisfy

$$\mu_\epsilon \geq \|\mathbf{W}^\top \boldsymbol{\epsilon}\|_\infty / n,$$

where $\boldsymbol{\epsilon}$ is the noise vector.

Define for $\mu > \mu_\epsilon$,

$$\overline{\mu} = \mu + \mu_\epsilon, \quad \underline{\mu} = \mu - \mu_\epsilon,$$

and the stretching factor $\nu = \overline{\mu}/\underline{\mu}$. For any index set $S \subset \{1, \ldots, K\}$ and stretching factor ν, the compatibility constant $\eta^2(\nu, S)$ (see e.g. Definition 2.1 in van de Geer (2014)), is defined as follows:

$$\eta^2(\nu, S) = \min \left\{ n^{-1}|S| \|\mathbf{W}^\top \boldsymbol{\gamma}\|_2^2;\ \|\boldsymbol{\gamma}_S\|_1 = 1, \|\boldsymbol{\gamma}_{-S}\|_1 \leq \nu \right\}, \tag{5.1}$$

where $\boldsymbol{\gamma}_S$ is the vector $\boldsymbol{\gamma}$ with values equal to 0 for indices in S and similarly $\boldsymbol{\gamma}_{-S}$ is the vector $\boldsymbol{\gamma}$ with values equal to 0 for indices in S^c. Let $\sigma_W = \Lambda_{\min}(n^{-1}\mathbf{W}^\top \mathbf{W})$ the smallest eigenvalue of the nonnegative-definite design matrix $\mathbf{W}^\top \mathbf{W}$. By the regularity of the mother wavelet defining the wavelet regression matrix $\mathbf{W}$ and the fact that the continuous design density on [0, 1] is bounded from below, it follows that $\sigma_W > 0$, and we have, in particular, for any $S \subset \{1, \ldots, K\}$:

$$\begin{aligned} n^{-1}\ |S| \|\mathbf{W}^\top \boldsymbol{\gamma}\|_2^2 &\geq n^{-1} \sigma_W\ |S| \|\boldsymbol{\gamma}\|_2^2 \\ &= n^{-1}\ \sigma_W |S| \left\{ \|\boldsymbol{\gamma}_S\|_2^2 + \|\boldsymbol{\gamma}_{-S}\|_2^2 \right\} \\ &\geq n^{-1} \sigma_W |S| \left\{ \frac{\|\boldsymbol{\gamma}_S\|_1^2}{|S|} + \frac{\|\boldsymbol{\gamma}_{-S}\|_1^2}{K - |S|} \right\}, \end{aligned}$$

which gives us the strictly positive lower bound $n^{-1}\sigma_W$ for the compatibility constant $\eta^2(\nu, S)$. Let us take $\mu = 2\mu_\epsilon$. By Theorem 3.1 of van de Geer (2014) we therefore have:

$$\frac{1}{n} \left\| \mathbf{f}^0 - \hat{\mathbf{f}} \right\|_2^2 \leq \frac{1}{n} \|\mathbf{f}^0 - \mathbf{W}\boldsymbol{\gamma}\|_2^2 + \frac{9n|S|\mu_\epsilon^2}{\sigma_W}. \tag{5.2}$$

Since $\mathbf{W}^\top \boldsymbol{\epsilon}$ is a random vector, we need to update the previous inequality by somehow bounding above the required constant μ_ϵ. By concentration inequalities that hold for typical error distributions (Gaussian or sub-Gaussian), a value of order $\log K = n$ for the tuning parameter ensures that the condition $\mu_\epsilon \geq \|\mathbf{W}^\top \boldsymbol{\epsilon}\|_\infty / n$ holds with high probability. More precisely, for an r-regular orthonormal wavelet basis, for any $\delta > 0$, with probability at-least $1 - 2\exp(-\delta^2/2)$ we have $\sigma\sqrt{\delta^2 + 2\log n} \geq \|\mathbf{W}\boldsymbol{\epsilon}\|_\infty$.

Therefore we can take $\mu_\epsilon = n^{-1}\, c_1\sqrt{\delta^2 + 2\log n}$ for some large constant $c_1 > 0$ and therefore inequality (5.2) holds with such a value of μ_ϵ with probability at-least $1 - 2\exp(-\delta^2/2)$.

The proof is completed by noting that for functions within the assumed Besov body the approximation error $\frac{1}{n}\|\mathbf{f}^0 - \mathbf{W}\boldsymbol{\gamma}\|_2^2$ is of order $\mathcal{O}(2^{-Ks})$.

References

Amato U, Antoniadis A, De Feis I, Gijbels I (2021) Penalised robust estimators for sparse and high-dimensional linear models. Stat Methods Appl 30(1):1–48. https://doi.org/10.1007/s10260-020-00511-z

Antoniadis A (2007) Wavelet methods in statistics: some recent developments and their applications. Stat Surv 1:16–55. https://doi.org/10.1214/07-ss014

Antoniadis A, Fan J (2001) Regularization of wavelet approximations. J Am Stat Assoc 96(455):939–955. https://doi.org/10.1198/016214501753208942

Antoniadis A, Gijbels I, Nikolova M (2011) Penalized likelihood regression for generalized linear models with non-quadratic penalties. Ann Inst Stat Math 63(3):585–615. https://doi.org/10.1007/s10463-009-0242-4

Averkamp R, Houdré C (2003) Wavelets thresholding for non-necessary Gaussian noise: idealism. Ann Stat 31(1):110–151. https://doi.org/10.1214/aos/1046294459

Bruce A, Donoho D, Gao H, Martin D (1994) Denoising and robust non-linear wavelet analysis. SPIE Proc Wavelet Appl 2242:325–336. https://doi.org/10.1117/12.170036

Cai T, Brown L (1998) Wavelet shrinkage for nonequispaced samples. Ann Stat 26(5):1783–1799. https://doi.org/10.1214/aos/1024691357

Cai T, Brown L (1999) Wavelet estimation for samples with random uniform design. Stat Probab Lett 42(3):313–321. https://doi.org/10.1016/S0167-7152(98)00223-5

Chen S, Donoho D, Saunders M (1998) Atomic decomposition by basis pursuit. SIAM J Sci Comput 20(1):33–61. https://doi.org/10.1137/S003614450037906X

Daubechies I, Defrise M, De Mol C (2004) An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. Commun Pure Appl Math 57(11):1413–1457. https://doi.org/10.1002/cpa.20042

Donoho D, Johnstone I (1995) Adapting to unknown smoothness via wavelet shrinkage. J Am Stat Assoc 90(432):1200–1224. https://doi.org/10.1080/01621459.1995.10476626

Donoho D, Johnstone I, Kerkyacharian G, Picard D (1995) Wavelet shrinkage: asymptopia? J R Stat Soc Ser B (Methodol) 57(2):301–337. https://doi.org/10.1111/j.2517-6161.1995.tb02032.x

Eubank R (1999) Nonparametric regression and spline smoothing, 2nd edn. CRC Press, Cambridge. https://doi.org/10.1201/9781482273144

Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc 96:1348–1360. https://doi.org/10.1198/016214501753382273

Fan J, Peng H (2004) Nonconcave penalized likelihood with a diverging number of parameters. Ann Stat 32(3):928–961. https://doi.org/10.1214/009053604000000256

Green P, Silverman B (1993) Nonparametric regression and generalized linear models. Chapman and Hall/CRC, Cambridge. https://doi.org/10.1201/b15710

Härdle W (1990) Applied nonparametric regression. Cambridge University Press, Cambridge. https://doi.org/10.1017/ccol0521382483

Haris A, Shojaie A, Simon N (2018) Wavelet regression and additive models for irregularly spaced data. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in neural information processing systems, vol 31. Curran Associates Inc, Red Hook. URL: https://dl.acm.org/doi/abs/10.5555/3327546.3327573

Huang J, Ma S, Zhang C (2008) Adaptive lasso for sparse high-dimensional regression models. Stat Sin 18:1603–1618

Kovac A, Silverman B (2000) Extending the scope of wavelet regression methods by coefficient-dependent thresholding. J Am Stat Assoc 95(449):172–183. https://doi.org/10.1080/01621459.2000.10473912

Liu W, Tang Y, Wu X (2020) Separating variables to accelerate non-convex regularized optimization. Comput Stat Data Anal 147:1–18. https://doi.org/10.1016/j.csda.2020.106943

Loh P (2017) Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. Ann Stat 45(2):866–896. https://doi.org/10.1214/16-AOS1471

Loh P, Wainwright M (2015) Regularized m-estimators with nonconvexity: statistical and algorithmic theory for local optima. J Mach Learn Res 16:559–616

Mallat S (2009) A wavelet tour of signal processing. Elsevier, Amsterdam. https://doi.org/10.1016/b978-0-12-374370-1.x0001-8

Nesterov Y (2007) Gradient methods for minimizing composite objective function. Discussion paper 2007076, Center for Operations Research and Econometrics (CORE). Université Catholique de Louvain. https://doi.org/10.1007/s10107-006-0089

Oh H, Nychka D, Lee T (2007) The role of pseudo data for robust smoothing with application to wavelet regression. Biometrika 94(4):893–904. https://doi.org/10.1093/biomet/asm064

Raskutti G, Wainwright M, Yu B (2010) Restricted eigenvalue conditions for correlated Gaussian designs. J Mach Learn Res 11:2241–2259

Sardy S, Tseng P, Bruce A (2001) Robust wavelet thresholding. IEEE Trans Signal Process 49(6):1146–1152. https://doi.org/10.1109/ICASSP.1997.604599

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc Ser B (Methodol) 56:267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

van de Geer S (2014) Weakly decomposable regularization penalties and structured sparsity. Scand J Stat 41(1):72–86. https://doi.org/10.1111/sjos.12032

Wainwright M (2009) Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). IEEE Trans Inf Theory 55(5):2183–2202. https://doi.org/10.1109/TIT.2009.2016018

Wand M, Ormerod J (2011) Penalized wavelets: embedding wavelets into semiparametric regression. Electron J Stat 5:1654–1717. https://doi.org/10.1214/11-ejs652

Xiong S, Dai B, Qian P (2017) Achieving the oracle property of OEM with nonconvex penalties. Stat Theory Relat Fields 1(1):28–36. https://doi.org/10.1080/24754269.2017.1326079

Zhang C (2010) Nearly unbiased variable selection under minimax concave penalty. Ann Stat 38(2):894–942. https://doi.org/10.1214/09-AOS729

Zhang T (2010) Analysis of multi-stage convex relaxation for sparse regularization. J Mach Learn Res 11:1081–1107

Zhao P, Yu B (2006) On model selection consistency of lasso. J Mach Learn Res 7:2541–2563

Zou H (2006) The adaptive lasso and its oracle properties. J Am Stat Assoc 101(476):1418–1429. https://doi.org/10.1198/016214506000000735

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.