HAL
open science

Penalized wavelet monotone regression

Anestis Antoniadis, Jéremie Bigot, Irène Gijbels

To cite this version:

Anestis Antoniadis, Jéremie Bigot, Irène Gijbels. Penalized wavelet monotone regression. Statistics and Probability Letters, 2009, 77 (16), pp.1608. 10.1016/j.spl.2007.03.041 . hal-00580091

HAL Id: hal-00580091

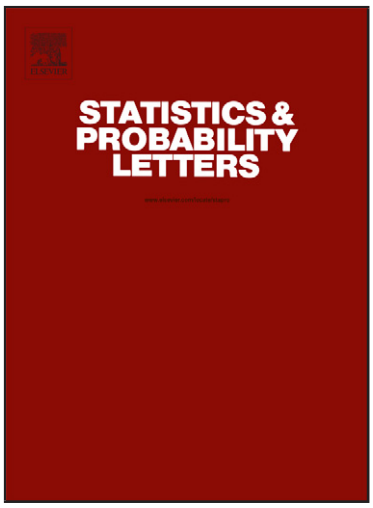
https://hal.science/hal-00580091v1

Submitted on 26 Mar 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Author's Accepted Manuscript



www.elsevier.com/locate/stapro

Penalized wavelet monotone regression

Anestis Antoniadis, Jéremie Bigot, Irène Gijbels

PII: S0167-7152(07)00136-8
DOI: doi:10.1016/j.spl.2007.03.041
Reference: STAPRO 4645

To appear in: *Statistics & Probability Letters*

Received date: 1 March 2005
Revised date: 1 March 2006
Accepted date: 20 March 2007

Cite this article as: Anestis Antoniadis, Jéremie Bigot and Irène Gijbels, Penalized wavelet monotone regression, *Statistics & Probability Letters* (2007), doi:10.1016/j.spl.2007.03.041

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting galley proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

Penalized wavelet monotone regression

Anestis Antoniadis,
Laboratoire IMAG-LMC, University Joseph Fourier,
BP 53, 38041 Grenoble Cedex 9, France;
Jéremie Bigot,
Department of Statistics, University Paul Sabatier,
Toulouse, France,
and
Irène Gijbels,
Department of Mathematics, Katholieke Universiteit Leuven,
Belgium.

December, 2005

Accepted manuscript

Abstract

In this paper we focus on nonparametric estimation of a constrained regression function using penalized wavelet regression techniques. This results into a convex optimization problem under linear constraints. Necessary and sufficient conditions for existence of a unique solution are discussed. The estimator is easily obtained via the dual formulation of the optimization problem. In particular we investigate a penalized wavelet monotone regression estimator. We establish the rate of convergence of this estimator, and illustrate its finite sample performance via a simulation study. We also compare its performance with that of a recently proposed constrained estimator. An illustration to some real data is given.

Key words and phrases: Besov spaces; Constrained curve fitting; Monotonicity; Splines; Wavelets; Wavelet nonparametric regression; Wavelet thresholding.
AMS 1991 subject classifications: Primary 62G07; secondary 65Dxx.

1 Introduction

Researchers in the physical and medical sciences are often interested in investigating an assumed monotonic relationship between an independent variable X and a dependent variable Y. Typical examples include the analysis of dose-response curves in pharmakinetics, growth curves in biology and many specific practical problems discussed in the literature cited below. Linear regression is usually too restrictive in these situations. Incorporating monotonicity constraints into the estimation of regression functions is then natural and dates back to the literature on isotonic regression. The isotonic regression model assumes that the expected value of Y is constant within disjoint regions of X and moreover that the mean levels within regions are nondecreasing in increasing X. An early exposition of this literature appears in Barlow, Bartholomew,

ACCEPTED MANUSCRIPT

Bremner & Brunk [3] and later in Robertson, Wright & Dykstra [22]. Consistency of monotonic regression is proved in Hanson, Pledger & Wright [12].

Smoother estimators of monotone regression can be found in Ramsay [21], Kelly and Rice [15], Mammen [18], Mammen and Thomas-Agnan [20], Hall and Huang [10] and Mammen, Marron, Turlach and Wand [19].

The previous literature on isotonic regression is based on determining the fitted values of the estimator on a finite set of points (usually the observed covariates) and uses a set of inequality constraints to impose restrictions on the value of the regression function at these points. The algorithms used to compute these estimators can be computationally intensive and involve a large set of inequality restrictions and require a special structure of the support. Moreover, as pointed out by Gijbels [9] many of the constrained estimates proposed in the literature reduce the smoothness of the estimator with which they started by using isotonic regression techniques or by projecting an unconstrained curve estimate onto a constrained subspace of regression functions. As a consequence the monotone estimates appear less smooth as the unconstrained estimates and often have jump discontinuities.

Series estimators provide a convenient alternative to the isotonic regression literature. Gallant [8] proposes the Fourier Flexible Form (FFF) estimator which is based on the trigonometric functions base. He identifies the set of restrictions on the coefficients of the FFF expansion that are sufficient to impose convexity. Monotonicity, however, cannot be easily imposed on the estimator. Especially convenient series estimators are those that are based on wavelets. The estimator used in this paper is a least-squares series estimator of the regression function based on wavelet basis functions. Given a grid of points on the support of the covariate the proposed estimator imposes restrictions on the values of the estimator at the grid points and then uses interpolation to compute the predicted values in any desired point on the support. This yields a finite number of constraints which are translated to linear inequality restrictions on the values of the coefficients of the wavelet functions. The regression function f is assumed to belong to a class of functions, $\mathcal{F}$, that satisfies certain regularity conditions.

As it is often the case in nonparametric regression when the regression function is assumed to belong to a large class of functions, the wavelet based least-squares estimator is inconsistent and a penalized or regularization procedure is a good remedy to provide a consistent estimator. Regularization is also a convenient framework for taking into account shape restrictions on the estimator and this is the approach chosen in this paper.

The remainder of the paper is structured as follows. Section 2 lays the foundations for wavelet based penalized estimation of a constrained regression function, in particular of a monotone regression function. Section 3 describes the optimization procedures for computing the estimator. Asymptotic properties of the estimator are established in Section 4. Section 5 presents a short Monte Carlo study on the efficiency and rate of convergence of the estimator and on a comparison with another estimator developed in the recent literature. Some concluding remarks are given in Section 6. Appendix A recalls some facts from convex optimization while Appendix B provides the proofs of the results.

Accepted manuscript

ACCEPTED MANUSCRIPT

2 Description of the estimator

The objective of this section is to describe a regression estimator that takes into account monotonicity constraints on the shape of the regression function but does not use functional form assumptions. This section focuses on the technical description of the estimator. A discussion on the asymptotic properties is left for the next section. For ease of presentation we restrict the discussion of the estimator to the one dimensional case. All results can be extended to a higher dimensional case but not without a considerable technical effort. In shape-restricted estimation, the curse of dimensionality has an additional effect: the number of constraints, needed to assure that the estimator satisfies certain restrictions, increases with dimensionality. An increase in the number of constraints in addition to the usual curse of dimensionality can make the problem intractable.

The following notations are used throughout the paper. For any n-dimensional vector $\mathbf{v} = (v_1, \ldots, v_n)^T$ with components v_i from $[0,1]$ and for any real-valued function f defined on $[0,1]$ define $\mathbf{f} = f(\mathbf{v}) = (f(v_1), \ldots, f(v_n))^T$. Furthermore, for a vector $\mathbf{v}$, $\mathbf{v} \leq 0$ means coordinate wise. Finally, for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $\langle \mathbf{x}, \mathbf{y} \rangle$ denotes the usual inner product in $\mathbb{R}^n$.

Assume that we have noisy observations of an unknown function $f : [0,1] \to \mathbb{R}$ at discrete time positions $x_i = \frac{i}{n}$, $i = 1, \ldots, n$:

$$y_i = f(x_i) + \sigma \epsilon_i, \; i = 1, \ldots, n, \tag{2.1}$$

where ϵ_i are i.i.d. normal variables with zero mean and variance 1, and σ is an unknown noise level parameter. The regression function $f(\cdot)$ is assumed to belong to a class of functions, $\mathcal{F}$, that satisfies certain regularity conditions. These conditions are further discussed later.

In an unconstrained setting, a least-squares estimator of the regression function in (2.1) is based on the empirical analogue of the expected value of a squared loss function and is the solution to the following problem

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2. \tag{2.2}$$

Assuming that n is a power of 2, let W_n be the discrete orthogonal wavelet transform matrix corresponding to some suitable wavelet basis of $L^2([0,1])$. Letting $\boldsymbol{\theta}$ be the vector of discrete wavelet coefficients of f, i.e. $\boldsymbol{\theta} = W_n \mathbf{f}$ and $\hat{\mathbf{d}} = W_n \mathbf{y}$ be the corresponding vector of noisy wavelet coefficients, an equivalent formulation of problem (2.2) is given by

$$\min_{\boldsymbol{\theta} \in \mathcal{W}} \frac{1}{2} \sum_{i=1}^{n} (\hat{d}_i - \theta_i)^2 = \min_{\boldsymbol{\theta} \in \mathcal{W}} \frac{1}{2} \|\boldsymbol{\theta} - \hat{\mathbf{d}}\|_2^2, \tag{2.3}$$

where $\mathcal{W}$ denotes the set of discrete wavelets coefficients of f when $f \in \mathcal{F}$. Recall that, for regular enough wavelets, if β_i denote the continuous wavelet coefficients of f, then the corresponding discrete wavelet coefficients θ_i of f are of the order $\sqrt{n}\beta_i$ (see e.g. Antoniadis and Fan [1]).

Often, in a nonparametric setup, the class $\mathcal{F}$ (and hence $\mathcal{W}$) is large and the above least-squares estimator is inconsistent. In order to obtain a consistent estimator one therefore must use some kind of regularization or thresholding. Antoniadis and Fan

Accepted manuscript

ACCEPTED MANUSCRIPT

[1] have shown that wavelet thresholding amounts to solve a penalized least-squares problem of the following form:

$$\min_{\theta\in\mathbb{R}^n}\left(\frac{1}{2}\|\boldsymbol{\theta}-\hat{\mathbf{d}}\|_2^2+\lambda\sum_{i=i_0}^{n}\psi(|\theta_i|)\right)\ , \tag{2.4}$$

where $\psi:\mathbb{R}\to\mathbb{R}^+$ is an appropriate penalty function depending on the method that is being used, and i_0 is the coarse level of smoothing. To simplify the notations, we will take hereafter $i_0=1$ without any loss of generality. We will concentrate on the following choice $\psi(|x|)=|x|$ which yields the so-called soft thresholding estimator:

$$\begin{aligned}\hat{\theta}_i &= 0 \text{ if } |\hat{d}_i|\le\lambda,\\ \hat{\theta}_i &= \hat{d}_i-\lambda\, sign(\hat{d}_i) \text{ if } |\hat{d}_i|>\lambda,\end{aligned}$$

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

Antoniadis and Fan [1] show that, for a large variety of function classes $\mathcal{F}$, controlling the rate at which the penalty λ/n converges to 0 as n tends to infinity, leads to estimators that achieve an optimal rate of convergence. The regularization procedure above is also a convenient framework for taking into account monotone restrictions on the estimator. Our task in what follows is to build such a penalized wavelet estimator of the regression function that satisfies the desired monotonicity constraint.

It should be mentioned that the form of the penalty function, i.e. $\psi(|x|)=|x|$ is not without importance. Indeed, this form will allow us to rely upon a simple dual formulation of the optimization problem. The equivalence between the original and dual optimization problem is carried by the convexity of the penalty function and by the fact that the absolute value of a wavelet coefficient is decomposed into the positive part and the minus-negative part. Extending the approach to a larger class of penalties is a challenge for future research.

Let $\hat{\mathbf{f}}=W_n^T\hat{\boldsymbol{\theta}}$ be the estimation of f at the design points. In a discrete setting, imposing shape constraints consists in determining a constrained estimate $\hat{\mathbf{f}}^c\in\mathbb{R}^n$ such that $\langle\hat{\mathbf{f}}^c,l^j\rangle\le 0$ for $j=1,\dots,m$, where the l^j's are appropriate vectors depending on the shape constraints. For instance with $m=n-1$, the choice $l_i^j=0$ if $i<j$ or $i>j+1$, $l_j^j=1$ and $l_{j+1}^j=-1$ would yield an increasing estimate in the discrete sense. These restrictions, resulting in m inequality constraints (with $m=(n-1)$) to be imposed on the estimator can be written in matrix notation as $D_m\hat{\mathbf{f}}^c(\boldsymbol{\gamma})\le 0$, where D_m denotes the $m\times n$ differentiation matrix given by

$$D_m=\begin{pmatrix}1 & -1 & 0 & 0 & \cdots & 0\\ 0 & 1 & -1 & 0 & \cdots & 0\\ \vdots & & \ddots & \ddots & & \vdots\\ 0 & \cdots & 0 & 1 & -1 & 0\\ 0 & \cdots & 0 & 0 & 1 & -1\end{pmatrix}$$

and where $\boldsymbol{\gamma}$ is the equidistant grid vector $(\frac{1}{n},\frac{2}{n},\dots,1)^T$.

For $\boldsymbol{\theta}\in\mathbb{R}^n$, let $\left(g_j(\boldsymbol{\theta})\right)_{j=1,\dots,n-1}$ be the components of the vector defined by $D_mW_n^T\boldsymbol{\theta}$. With the choice $\psi(|x|)=|x|$, wavelet regression under shape constraints can then be formulated as a convex optimization problem under linear constraints:

$$\begin{aligned}\min_{\boldsymbol{\theta}\in\mathbb{R}^n}\quad & h(\boldsymbol{\theta})=\min_{\boldsymbol{\theta}\in\mathbb{R}^n}\left(\frac{1}{2}\|\boldsymbol{\theta}-\hat{\mathbf{d}}\|_2^2+\lambda\sum_{i=1}^{n}|\theta_i|\right)\\ \text{s.t}\quad & g_i(\boldsymbol{\theta})\le 0,\ i=1,\dots,n-1\end{aligned} \tag{2.5}$$

Accepted manuscript

ACCEPTED MANUSCRIPT

Proposition 2.1 *The optimization problem (2.5) has a unique solution.*

Proof: see Appendix B. □

The next section explains how to construct the estimator.

3 Practical Implementation of the estimator

Note that the convex optimization problem that we consider in this paper is of the following form:

$$\begin{array}{ll} \min\limits_{\boldsymbol{\theta}\in\mathbb{R}^n} & \left(\frac{1}{2}\|\boldsymbol{\theta}-\mathbf{d}\|_2^2+\lambda\sum\limits_{i=1}^{n}|\theta_i|\right), \\ \text{s.t} & \Phi\boldsymbol{\theta}\leq 0, \end{array} \tag{3.1}$$

for some $\mathbf{d}\in\mathbb{R}^n$, and where Φ is an $m\times n$ real matrix. To retrieve our previous setting, take $\mathbf{d}=\hat{\mathbf{d}}$ and $\Phi=D_mW_n^T$. Since the objective function for problem (3.1) is not differentiable, the optimization results recalled in Appendix A, while useful, cannot be used directly. In this section, we show that the non-differentiable problem (3.1) falls into the category of *constrained non-smooth optimization problems* (CNSO) as described in Section 14.6 of Fletcher [7], and that necessary and sufficient conditions similar to the KKT equations of Appendix A can be derived.

3.1 Necessary and sufficient conditions for optimality

Following the notations in Fletcher [7], problem (3.1) can be written as a penalized constrained convex optimization problem:

$$\begin{array}{ll} \min\limits_{\boldsymbol{\theta}\in\mathbb{R}^n} & \{h(\boldsymbol{\theta})+k(c(\boldsymbol{\theta}))\}, \\ \text{s.t} & t(r(\boldsymbol{\theta}))\leq 0, \end{array} \tag{3.2}$$

with

$$\begin{aligned} r(\boldsymbol{\theta}) &= (g_1(\boldsymbol{\theta}),\ldots,g_m(\boldsymbol{\theta}))=\Phi\boldsymbol{\theta}\in\mathbb{R}^m, \\ t(r) &= \max_{i=1,\ldots,m} g_i(\boldsymbol{\theta}), \\ h(\boldsymbol{\theta}) &= \frac{1}{2}\|\boldsymbol{\theta}-\mathbf{d}\|_2^2, \\ c(\boldsymbol{\theta}) &= \lambda\boldsymbol{\theta}, \\ k(c(\boldsymbol{\theta})) &= \|c(\boldsymbol{\theta})\|_1 . \end{aligned}$$

Now, define the Lagrangian function as:

$$L(\boldsymbol{\theta},\boldsymbol{\mu},\mathbf{u},\pi)=h(\boldsymbol{\theta})+\boldsymbol{\mu}^Tc(\boldsymbol{\theta})+\pi\mathbf{u}^Tr(\boldsymbol{\theta}) .$$

For non-smooth convex functions, the notion of differentiability can be extended via the introduction of the *subdifferential*:

Accepted manuscript

ACCEPTED MANUSCRIPT

Definition 3.1 *If f is a convex function defined on a convex set K, then the subdifferential of f at $\boldsymbol{\theta}$ is defined by:*

$$\partial h(\boldsymbol{\theta}) = \{g : h(\boldsymbol{\theta}+\boldsymbol{\delta}) \geq h(\boldsymbol{\theta}) + \boldsymbol{\delta}^T g;\ \forall \boldsymbol{\theta}+\boldsymbol{\delta} \in K\}\ .$$

From Fletcher [7] p. 363, we have that for $\boldsymbol{\theta} \in \mathbb{R}^n$,

$$\begin{aligned}
\partial \max_{i=1,\ldots,n} \theta_i &= \{\boldsymbol{\mu} \in \mathbb{R}^n;\ \sum_{i=1}^n \mu_i = 1,\ \mu_i \geq 0 \text{ and } \theta_i < \max_{i=1,\ldots,n} \theta_i \Rightarrow \mu_i = 0\} \\
\partial \|\boldsymbol{\theta}\|_1 &= \{\boldsymbol{\mu} \in \mathbb{R}^n;\ |\mu_i| \leq 1, \text{ and } \theta_i \neq 0 \Rightarrow \mu_i = sign(\theta_i)\}\ .
\end{aligned}$$

From Fletcher [7], Theorem 14.6.1, we have that if $\boldsymbol{\theta}^*$ is a local minimizer of (3.1), then there exist $\boldsymbol{\mu} \in \partial k(\boldsymbol{\theta}^*)$, $\mathbf{u} \in \partial t(\boldsymbol{\theta}^*)$ and a real $\pi \geq 0$ such that:

$$\begin{aligned}
t(r(\boldsymbol{\theta}^*)) &\leq 0, \\
\pi t(r(\boldsymbol{\theta})) &= 0, \\
\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}^*, \boldsymbol{\mu}, \mathbf{u}, \pi) &= \nabla h(\boldsymbol{\theta}^*) + \boldsymbol{\mu}^T \nabla c(\boldsymbol{\theta}^*) + \pi \mathbf{u}^T \nabla r(\boldsymbol{\theta}^*).
\end{aligned}$$

In our setting, the above necessary conditions can be written as: there exists $\boldsymbol{\mu} \in \mathbb{R}^n$ and $\boldsymbol{\nu} \in \mathbb{R}^m$ such that

$$\begin{aligned}
\boldsymbol{\theta}^* - d + \lambda \boldsymbol{\mu} + \Phi^T \boldsymbol{\nu} &= 0, \\
g_j(\boldsymbol{\theta}^*) &\leq 0,\ j = 1, \ldots, m \\
\boldsymbol{\nu} &\geq 0, \\
g_j(\boldsymbol{\theta}^*) &< 0 \Rightarrow \nu_j = 0,\ j = 1, \ldots, m \\
|\mu_i| &\leq 1 \\
\theta_i^* &\neq 0 \Rightarrow \mu_i = sign(\theta_i^*),\ i = 1, \ldots, n\ .
\end{aligned}$$

Since $\nabla^2_{\boldsymbol{\theta}} L(\boldsymbol{\theta}, \boldsymbol{\mu}, \mathbf{u}, \pi) = I$ is positive definite for all $\boldsymbol{\theta} \in \mathbb{R}^n$, we have that the above conditions are also sufficient (see Theorem 14.6.3 in Fletcher [7]). Note that if we impose no shape constraints (i.e. $\boldsymbol{\nu} = 0$), the soft thresholding estimator satisfies the above necessary and sufficient conditions of optimality.

3.2 Dual formulation

For $\boldsymbol{\theta} \in \mathbb{R}^n$, let $\theta_i^+ = \max(\theta_i, 0)$ and $\theta_i^- = \max(-\theta_i, 0)$. Note that, $\theta_i^+ - \theta_i \geq 0$, $\theta_i^- + \theta_i \geq 0$ and $\theta_i^+ + \theta_i^- = |\theta_i|$. Let $\mathbf{e} \in \mathbb{R}^n$ be the unit vector with all entries equal to one, and consider the following constrained convex optimization problem:

$$\begin{aligned}
\min_{\boldsymbol{\theta}, \boldsymbol{\theta}^+, \boldsymbol{\theta}^- \in \mathbb{R}^n} \quad & \left(\frac{1}{2} \|\boldsymbol{\theta} - \mathbf{d}\|_2^2 + \lambda \mathbf{e}^T (\boldsymbol{\theta}^+ + \boldsymbol{\theta}^-) \right), \\
\text{s.t} \quad & \Phi \boldsymbol{\theta} \leq 0, \\
& \boldsymbol{\theta}^+ - \boldsymbol{\theta} \geq 0,\ \boldsymbol{\theta}^+ \geq 0, \\
& \boldsymbol{\theta}^- + \boldsymbol{\theta} \geq 0,\ \boldsymbol{\theta}^- \geq 0
\end{aligned} \tag{3.3}$$

Proposition 3.1 *The optimization problems (3.1) and (3.3) are equivalent in the sense that they yield the same unique solution $\hat{\boldsymbol{\theta}}^c \in \mathbb{R}^n$.*

Proof: see Appendix B. □

Accepted manuscript

ACCEPTED MANUSCRIPT

Proposition 3.2 *Problem (3.3) is a smooth convex optimization problem whose dual can be written as*

$$\begin{aligned} \max_{\boldsymbol{\theta},\boldsymbol{\mu},\boldsymbol{\nu}\in\mathbb{R}^n} \quad & L(\boldsymbol{\theta},\boldsymbol{\mu},\boldsymbol{\nu}) = \max_{\boldsymbol{\theta},\boldsymbol{\mu},\boldsymbol{\nu}\in\mathbb{R}^n} \left(\frac{1}{2}\|\boldsymbol{\theta}-\mathbf{d}\|_2^2 + \lambda\boldsymbol{\mu}^T\boldsymbol{\theta} + \boldsymbol{\nu}^T\Phi\boldsymbol{\theta}\right) , \\ s.t \quad & \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta},\boldsymbol{\mu},\boldsymbol{\nu}) = 0, \\ & -\mathbf{e} \le \boldsymbol{\mu} \le \mathbf{e}, \\ & \boldsymbol{\nu} \ge 0 \end{aligned} \tag{3.4}$$

Proof: see Appendix B. □

Note that the condition $\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta},\boldsymbol{\mu},\boldsymbol{\nu}) = 0$ implies that $\boldsymbol{\theta} = \mathbf{d} - \lambda\boldsymbol{\mu} - \Phi^T\boldsymbol{\nu}$. By eliminating $\boldsymbol{\theta}$, the dual problem (3.4) can then be written as:

$$\begin{aligned} \max_{\boldsymbol{\mu},\boldsymbol{\nu}\in\mathbb{R}^n} \quad & L(\boldsymbol{\mu},\boldsymbol{\nu}) = \max_{\boldsymbol{\mu},\boldsymbol{\nu}\in\mathbb{R}^n} \left(-\frac{1}{2}\lambda^2\boldsymbol{\mu}^T\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{\nu}^T\Phi\Phi^T\boldsymbol{\nu} + \lambda\boldsymbol{\mu}^T d - \lambda\boldsymbol{\mu}^T\Phi^T\boldsymbol{\nu} + \boldsymbol{\nu}^T\Phi\mathbf{d}\right) , \\ \text{s.t} \quad & -\mathbf{e} \le \boldsymbol{\mu} \le \mathbf{e}, \\ & \boldsymbol{\nu} \ge 0 \end{aligned}$$

Let $B = \begin{bmatrix}\lambda I & \Phi^T\end{bmatrix}$ (note that B is an $n \times (n+m)$ real matrix which may be singular). Hence, solving the dual problem (3.4) is equivalent to minimizing the following least-squares problem with box constraints:

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^{n+m}} \quad & \frac{1}{2}\|B\mathbf{x}-\mathbf{d}\|_2^2, \\ \text{s.t} \quad & -\mathbf{e} \le \boldsymbol{\mu} \le \mathbf{e}, \\ & \boldsymbol{\nu} \ge 0 , \end{aligned} \tag{3.5}$$

for $\mathbf{x}^T = [\boldsymbol{\mu}^T \ \boldsymbol{\nu}^T]$.

Hence, in our discrete setting, wavelet regression under shape constraints amounts to solving a least-squares minimization problem with bound constraints for which many algorithms are available.

4 Asymptotic properties

A large body of literature is devoted to establishing the consistency of wavelet thresholding estimators and to the rate at which they converge to the true parameter. Most of this literature treats the case where no shape restrictions are imposed on the estimator. Therefore, before using any of the results from this literature, one has to make sure that they remain intact under shape restrictions. The results presented in this section show that the monotonicity restricted estimator described in the previous section is consistent and that it achieves the same optimal rate of convergence as in the unrestricted estimation problem.

Let

$$R(\hat{f}_\lambda, f) = n^{-1}\sum_{i=1}^{n}\{\hat{f}_\lambda(x_i) - f(x_i)\}^2 ,$$

be the risk function of the unrestricted regularized wavelet estimator $\hat{f}_\lambda$. By translating the problem from the function space into the wavelet domain, Antoniadis and Fan [1]

Accepted manuscript

ACCEPTED MANUSCRIPT

have shown that the above estimator, with $\lambda = \sigma\sqrt{2\log n}$ nearly achieves the optimal rate of convergence for a large variety of function classes $\mathcal{F}$, a typical example of which are the Besov spaces.

Assume that the unknown signal f is in a Besov ball. Because of simple characterization of this space via the wavelet coefficients of its members, for $C > 0$, the Besov space ball $B^r_{p,q}(C)$ ($r > 0, 1 \leq p, q \leq \infty$) can be defined as

$$B^r_{p,q} = \left\{ f \in L_p : \sum_j \left(2^{j(r+1/2-1/p)} \|\boldsymbol{\beta}_{j\cdot}\|_p \right)^q < C \right\}, \tag{4.1}$$

where $\boldsymbol{\beta}_{j\cdot}$ is the vector of wavelet coefficients at the resolution level j. Here, r indicates the degree of smoothness of the underlying signal f. Note again that the wavelet coefficients $\boldsymbol{\beta}$ in the definition of the Besov space are continuous wavelet coefficients. They are approximately a factor of $n^{1/2}$ larger than the discrete wavelet coefficients $W_n\mathbf{f}$. This is equivalent to assuming that the noise level is of order $1/n$.

The following theorem evaluates the rate of convergence of the maximum risk of our estimator.

Theorem 4.1 *For the penalty function $\psi(x) = |x|$ and for $r > 1/2$, the maximum risk of the restricted monotone penalized least-squares estimator $\hat{f}^c$ over the Besov ball $B^r_{1,1}(C)$ is of rate $O(n^{-2r/(2r+1)} \log n)$ when the universal thresholding $\sigma\sqrt{2\log n}$ is used.*

The proof of this theorem is provided in Appendix B. The proof consists in computing the difference between the wavelet coefficients of the constrained and the unconstrained estimate and showing that it shrinks to zero at a faster rate than the optimal convergence rate of the unconstrained estimator.

Remark. For ease of presentation we stated the theorem for functions belonging to the Besov ball $B^r_{1,1}(C)$. However, using various embedding theorems for Besov spaces (see for example Corollary 9.2, page 124 of Härdle et al. [13]), under appropriate conditions on s, p and q the result is also true for more general Besov spaces $B^s_{p,q}$. We know, for example, that for any $\epsilon < 1/2$ and $q \geq 1$ we have $B^{r-\epsilon}_{1/(1-\epsilon),q} \subseteq B^r_{1,1}$ and the inclusion is continuous. From the proof of Theorem 4.1, it follows that the $\sum_i |\beta_i|$ is still uniformly bounded (the embedding is continuous) and the result follows again by applying results on the behavior of $R(\hat{f}, f)$ given by Antoniadis and Fan [1] but this time for the Besov ball $B^{r-\epsilon}_{1/(1-\epsilon),q}(C)$.

5 Monte-Carlo Simulations and a Real Example

We have designed a small scale simulation study for illustrating the performance of our penalized wavelet monotone smoother and for comparing it with another simple smoothing splines based monotone smoother developed recently by Zhang [25]. This estimator is also monotone in a discrete sense. The main idea of the smoothing splines smoother is to shift the monotonicity constraint on the underlying regression function to the positiveness or negativeness constraint on the associated derivative curve. The interested reader is referred to the paper by Zhang [25], where the smoothing parameter is chosen by generalized cross validation (GCV) and closed form formulas for the estimation are derived.

Accepted manuscript

ACCEPTED MANUSCRIPT

We investigate the regression model with an equidistant design on $[0,1]$, normally distributed errors, sample sizes $n = 128$ and $n = 256$ and signal to noise ratios (SNR) of 3 and 5. Signal-to-noise ratios are measured as $\mathrm{sd}(m(x))/\sigma$, where $\mathrm{sd}(m(x))$ is the estimated standard deviation of the regression function, $m(x_i)$ over the sample $i = 1, \ldots, n$ and σ is the true standard deviation of the noise in the data. The monotone regression functions that we consider are

$$m_1(x) = \frac{\exp(20(x-1/2))}{1+\exp(20(x-1/2))}$$

$$m_2(x) = x + \frac{1}{6\pi}\sin(6\pi x),$$

$$m_3(x) = u(x) + 0.1 I_{\{x>0.5\}}(x),$$

where

$$u(x) = \int_0^x \left(\frac{1}{2}\exp(-\frac{(t-0.2)^2}{2(0.05^2)}) + 1.5\exp(-\frac{(t-0.6)^2}{2*(0.1^2)})\right) dt$$

and

$$m_4(x) = xI_{\{x<0.3\}}(x) + 1.5xI_{\{x>0.6\}}(x) + 0.3I_{\{x>0.3\}}(x) + 2x * I_{\{x>0.8\}}(x)\ .$$

These functions correspond to, respectively, a function which changes several times from a strongly increasing part to a flat part, a function with a continuous "jump", a strictly increasing curve with a discontinuous jump; and an increasing function with linear parts and some flat parts. The different functions are displayed in Figure 1.

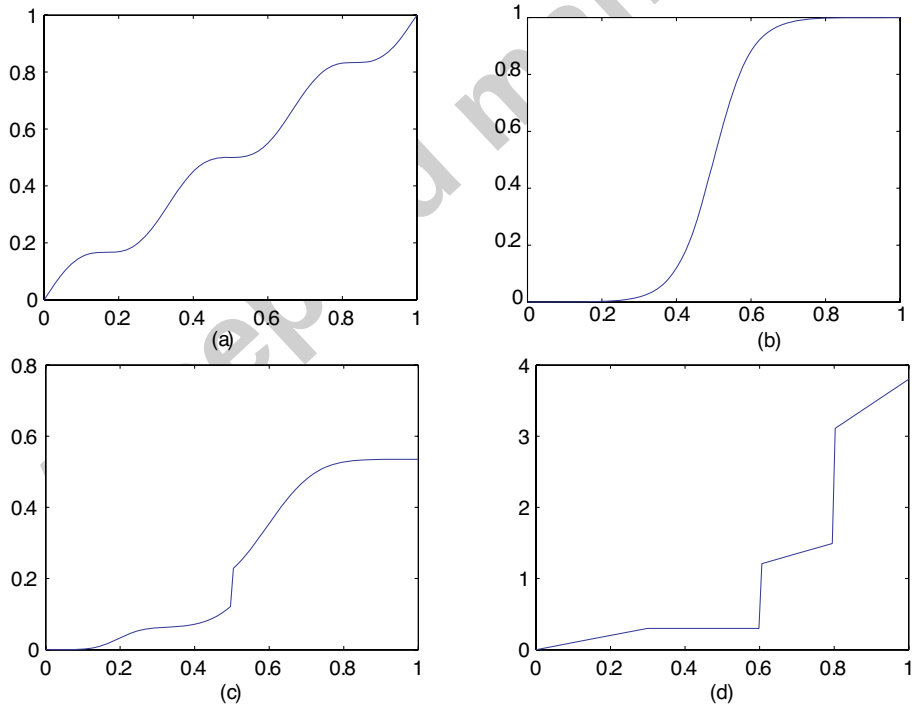


Figure 1: *Signals (a): m_1, (b): m_2, (c): m_3 and (d): m_4 used in the simulations.*

We have used 100 simulations runs, for each regression function, each sample size and each signal to noise ratio. For our penalized wavelet smoother estimates we used symlets of order 8, $i_0 = 3$ and the universal threshold $\lambda = \sigma\sqrt{2\log(n)}$ with a robust estimate of σ based on the median absolute deviation of the wavelet transform

ACCEPTED MANUSCRIPT

at the finest scale. Note that imposing monotonicity constraints tends to averaging the oscillating parts (pseudo-Gibbs phenomema) of the unconstrained estimate. For the monotone spline smoother we have used the matlab procedure implementing the smoother, available at `http://www.stat.nus.edu.sg/~zhangjt`. We have compared the constrained estimates for both methods (smoothing splines and penalized wavelet smoother). For the 100 simulations and each setting of the simulation design we calculate the pointwise mean squared error (MSE) for the two estimates $\hat{m}_w$ (wavelet estimator) and $\hat{m}_s$ (splines smoother). In the following, we present curves for the estimates and their MSE, where only results for the sample size $n = 128$, SNR=5 are displayed (see Figure 2 and Figure 3). The results corresponding to the cases $n = 256$ and SNR=3 are quite similar and available from the authors (see also Table 1 for results of the integrated MSE for all cases of the simulation design). The dashed curves in the figures correspond to the monotone splines smoother, whereas our wavelet estimator is plotted as a solid curve.

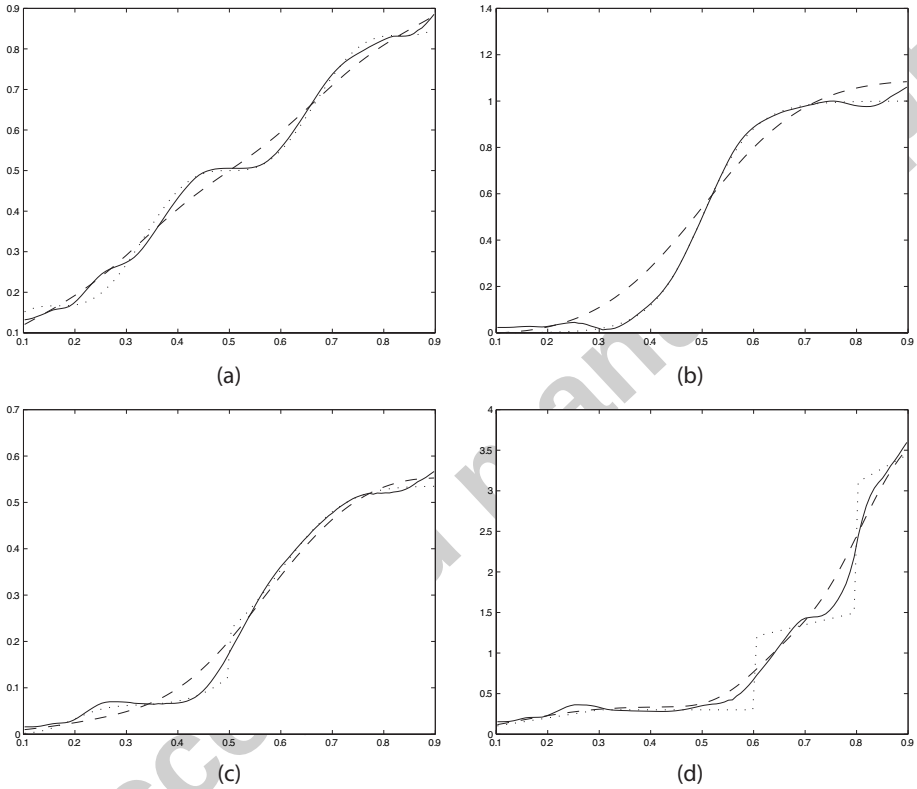


Figure 2: *Average estimates over 100 simulations for the 4 signals for a sample size $n = 128$ and a signal to noise ratio of 5. The true regresssion function is represented as a dotted curve, the estimate $\hat{m}_w$ as a solid curve and the estimate $\hat{m}_s$ as a dashed curve.*

As one can see, the monotone spline smoother has a tendency to oversmooth, leading to a large squared bias. The MSE comparison shows a substantial difference between the two estimates, mainly caused by the amount of oversmoothing of the spline smoother. The superiority of the wavelet smoother is also confirmed by the results displayed in Table 1.

As one can see from Table 1, the order of the integrated mean squared error for each estimator depends on the regression model under consideration, but whatever simulation setup is used the wavelet estimator outperforms the spline GCV-based monotone smoother.

ACCEPTED MANUSCRIPT

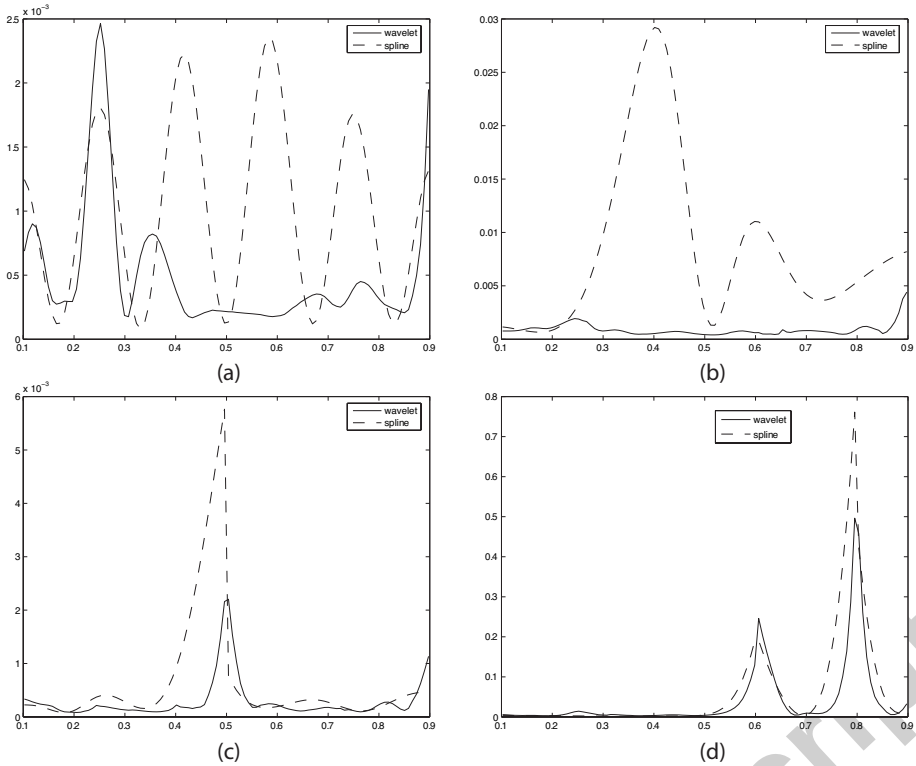


Figure 3: *Simulated mean squared error of the estimators $\hat{m}_w$ (solid curves) and $\hat{m}_s$ (dashed curves) for the four regression functions. The sample size is $n = 128$ and the signal to noise ratio is 5.*

		$\hat{m}_w$	$\hat{m}_s$
$m(x)$	SNR	IMSE	IMSE
	3	8.9 e-04	1.3 e-03
$m_1(x)$	5	5.3 e-04	1.0 e-03
	3	1.8 e-03	9.6 e-03
$m_2(x)$	5	8.2 e-04	8.5 e-03
	3	5.1 e-04	1.1 e-03
$m_3(x)$	5	2.8 e-04	6.9 e-04
	3	6.4 e-02	8.0 e-02
$m_4(x)$	5	4.0 e-02	6.7 e-02

Table 1: *Simulated integrated mean squared error of the estimators $\hat{m}_w$ and $\hat{m}_s$ for the four regression functions, based on samples of size $n = 128$.*

To end this section, we apply our methodology to a real example concerning monotone regression, namely the fuel consumption data that relate fuel efficiency (in miles per gallon) to engine output. These data have also been used by Mammen et al. [19] to illustrate a projection-based constrained smoothing procedure. A scatter plot of the data is displayed in Figure 4. The data are available at the *Statlib* internet repository at `http://lib.stat.cmu.edu/datasets/cars.data`.

Displayed in Figure 4 are also a spline smooth monotone fit (dashed curve) and a wavelet monotone fit (solid curve). The number of data points is 392 but with only 93 distincts x-coordinates. In order to apply our wavelet monotone estimate, we have used, preliminary to the estimation, the interpolation method to a fine regular grid of 128 points by Kovac and Silverman [17] (their `makegrid` procedure). The wavelet esti-

Accepted manuscript

ACCEPTED MANUSCRIPT

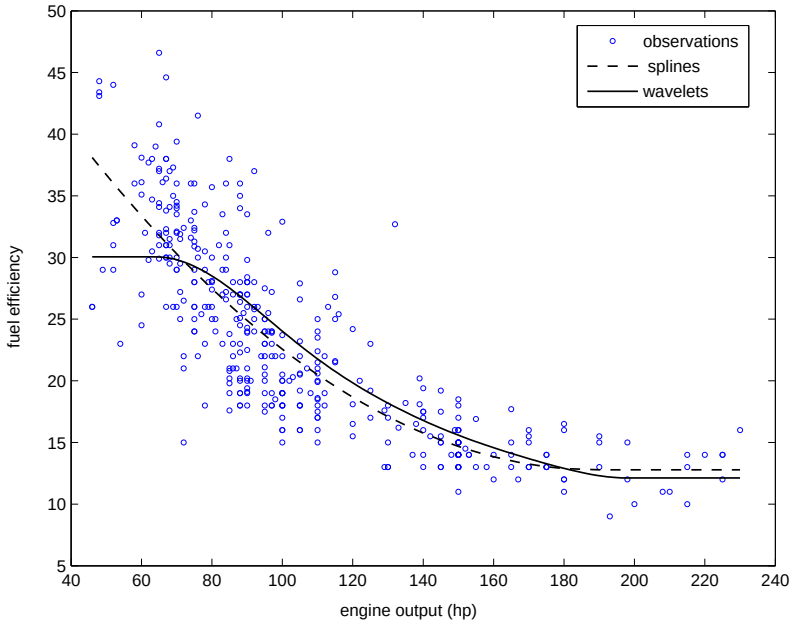


Figure 4: *Fuel efficiency versus Engine output.*

mator was then computed with symlets of order 8, $i_0 = 3$ and the universal threshold $\lambda = \sigma\sqrt{2\log(n)}$ with a robust estimate of σ based on the median absolute deviation of the wavelet transform at the finest scale. The spline monotone fit was produced by Zhang's monotone smoother using a smoothing parameter $4.06E - 011$ chosen by GCV. It can be seen that, overall, the difference between the two fits is small, except at the left end of the support, where many observations are available.

6 Concluding Remarks

A nonparametric monotone restricted estimator based on a penalized wavelet least-squares method is described in this paper. The benefits of the penalized wavelet framework are as follows. First, assumptions like monotonicity are easily incorporated into the estimator. Second, the estimator is computed using convex programming with linear inequalities constraints. Therefore, this estimator is easy to implement. Finally, the estimator behaves well even for functions that may present jumps or discontinuities.

A drawback of our estimation procedure is that it requires that observations are sampled on an equidistant grid and that the sample size is a power of 2. However, for a deterministic design, one may use several procedures that have been developed to relax these requirements without affecting the results obtained in this paper; such as, for example, the interpolation method of Hall and Turlach [11], the binning method of Antoniadis et al. [2], the transformation method of Cai and Brown [5], the isometric method of Sardy et al. [24], the interpolation method to a fine regular grid of Kovac and Silverman [17] and the penalized wavelet method of Antoniadis and Fan [1]. In principle, any of these methods can fit into the framework of this paper, with each interpolation method inducing a different choice of a function base for the series expansion.

For a random design one could use warped wavelets recently investigated by Kerky-

Accepted manuscript

ACCEPTED MANUSCRIPT

acharian and Picard [16]. It is shown there that for designs having a property of Muckenhoupt type, these new bases have a behavior quite similar to a regular wavelet basis, leading to estimation algorithms that mimick exactly the equi-spaced case. This would in principle enable us to prove that the associated constrained penalized procedure achieves rates of convergence which have been proved to be optimal in the uniform design case. However, it poses some challenges to extend this method to our setting and we intend to pursue this in a future work.

It should be mentioned that the general setup of Section 2 allows to treat more general constraint problems than the monotonicity constraint problem. Indeed, any restriction on the regression function that involves signing of any discrete functional of the wavelet coefficients of any order can be treated by our method. It is also possible to allow a number of such restrictions to apply simultaneously. For example, a regression function can be constrained to be both monotone and concave using the same technique.

The developed procedure produces an estimate that is monotone on the grid of design points. When the design is dense, i.e the grid can be as fine as one wishes (asymptotically), the approach is still feasible, but the algorithm can be computationally intensive since it involves then a very large set of inequality restrictions. But with a limited number of constraints, we still get a reasonable estimator. In order to get an estimated curve that satisfies the monotonicity constraints on the whole interval one would need to use a monotonicity preserving wavelet basis so that the estimator will satisfy the required restriction globally on the interval. An appropriate wavelet basis to insure this is a wavelet B-spline basis of fixed order, but such a basis is not anymore orthogonal and most of the derivations in the paper should be changed accordingly. While we believe that such a bi-orthogonal wavelet basis can fit into the framework of the paper and will lead to the same asymptotic rates, it involves hard theoretical work (insuring monotonicity, decomposing on this basis but reconstructing with the dual basis, extending the asymptotic results of Antoniadis and Fan [1] to this case). This is beyond the scope of the paper and subject of future research work.

Accepted manuscript

Appendix A

We recall here some classical results for convex optimization problems that we have used in Section 3. Further details can be found in the books by Rockafeller [23], Fletcher [7], and Boyd and Vandenberghe [4].

Consider the following optimization problem:

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & h(x) \\ \text{s.t} \quad & c_i(x) \le 0,\ i = 1,\dots,m \end{aligned} \tag{A.1}$$

where h and $c_1,\dots,c_m$ are convex and C^1 functions: $\mathbb{R}^n \to \mathbb{R}$. Let $C = \{x \in \mathbb{R}^n; c_i(x) \le 0,\ i = 1,\dots,m\}$. Assume that problem (A.1) has a solution $x^* = \arg\min_{x\in C} h(x)$ and let $p^* = h(x^*)$. A point $x \in C$ is said to be *strictly feasible* if $c_i(x) < 0$ for all $i = 1,\dots,m$. For $\mu \in \mathbb{R}^m$, define the Lagrangian function as:

$$L(x,\mu) = h(x) + \sum_{i=1}^{m} \mu_i c_i(x)\ .$$

ACCEPTED MANUSCRIPT

• **Karush-Kuhn-Tucker conditions**

If there exists a strictly feasible point $x \in C$ (Slater's condition), then there exist Lagrange multipliers $\mu^* \in \mathbb{R}^m$, such that x^*, μ^* satisfy the following system of equations, described as the Karush-Kuhn-Tucker (KKT) conditions:

$$\begin{aligned} \nabla_x L(x,\mu) &= 0, \\ c_i(x) &\leq 0,\ i = 1,\dots,m \\ \mu &\geq 0, \\ \mu_i c_i(x) &= 0,\ i = 1,\dots,m\ . \end{aligned}$$

Note that the KKT conditions also hold if the functions $c_i(x) \leq 0,\ i = 1,\dots,m$ are all linear constraints (see Theorem 9.1.1 in Fletcher [7]) without assuming Slater's condition. These conditions are also sufficient if Slater's condition holds.

• **Duality**

Define the Lagrange dual function as:

$$g(\mu) = \inf_{x \in \mathbb{R}^n} L(x,\mu)\ .$$

Note that g is a convex function. Then, for any $\mu \geq 0$ which is dual feasible i.e. $g(\mu) > -\infty$, we have that $g(\mu) \leq p^*$. Let μ^* denote a solution (if any) of the following convex optimization problem:

$$\begin{aligned} \max_{\mu \in \mathbb{R}^n} \quad & g(\mu) \\ \text{s.t} \quad & \mu \geq 0, \end{aligned} \tag{A.2}$$

and let $d^* = g(\mu^*)$. If there exists a strictly feasible point, then $d^* = p^*$ and x^* minimizes $L(x,\mu^*)$ over $x \in \mathbb{R}^n$, where μ^* denotes a solution (if any) of the dual problem (A.2). Note that the inequality $d^* \leq p^*$ holds when d^* and p^* are infinite. If $p^* = -\infty$, then $d^* = -\infty$ and so the Lagrange dual problem is infeasible. Conversely, if $d^* = \infty$, then $p^* = \infty$ i.e. the primal problem is infeasible.

Note that if x^* is a solution of the primal problem (A.1) and if x^*, μ^* satisfy the above KKT conditions, then under appropriate assumptions (e.g. if the functions $c_i(x) \leq 0,\ i = 1,\dots,m$ are all linear constraints, see Theorem 9.5.1 in Fletcher [7]) x^*, μ^* solves the dual problem:

$$\begin{aligned} \max_{(x,\mu) \in \mathbb{R}^n \times \mathbb{R}^m} \quad & L(x,\mu), \\ \text{s.t} \quad & \nabla_x L(x,\mu) = 0,\ \mu \geq 0, \end{aligned}$$

and $p^* = L(x^*,\mu^*)$.

Appendix B. Proofs

Proof of Proposition 2.1: Let $C = \{\theta \in \mathbb{R}^n; g_i(\theta) \leq 0,\ i = 1,\dots,n-1\}$. Since the g_i's are linear functions, C is a closed convex subset of $\mathbb{R}^n$. Since the objective function

Accepted manuscript

ACCEPTED MANUSCRIPT

h, defined in equation (2.5), is a continuous and nonnegative function on C, $h(C) = \{h(\theta); \theta \in C\}$ is a closed subset of $\mathbb{R}$ with a lower bound. Hence $h(C)$ has a minimum $h(\theta^c)$. The strict convexity of h implies the unicity of the minimum θ^c. $\square$.

Proof of Proposition 3.1: from Proposition 2.1 we have that problem (3.1) has a unique solution. The arguments in the proof of Proposition 2.1 can also be used to show that the convex optimization problem (3.3) has a unique solution. Let

$$\begin{aligned} f_1(\theta) &= \frac{1}{2}\|\theta - d\|_2^2 + \lambda \sum_{i=1}^{n} |\theta_i| \\ f_2(\theta, \theta^+, \theta^-) &= \frac{1}{2}\|\theta - d\|_2^2 + \lambda e^T(\theta^+ + \theta^-) \\ C_1 &= \{\theta \in \mathbb{R}^n; \Phi\theta \leq 0\} \\ C_2 &= \left\{(\theta, \theta^+, \theta^-) \in \mathbb{R}^{3n}; \Phi\theta \leq 0, \theta^+ - \theta \geq 0,\ \theta^+ \geq 0, \theta^- + \theta \geq 0,\ \theta^- \geq 0\right\}. \end{aligned}$$

Let $\theta_1 \in \mathbb{R}^n$ and $(\theta_2, \theta_2^+, \theta_2^-)$ be the unique minimizers of problems (3.1) and (3.3) respectively. Let $\theta_1^+ = \max(\theta_1, 0)$ and $\theta_1^- = \max(-\theta_1, 0)$ (the max is taken componentwise). Note that $(\theta_1, \theta_1^+, \theta_1^-) \in C_2$ and that for all $\theta \in C_1$, $f_1(\theta_1) = f_2(\theta_1, \theta_1^+, \theta_1^-) \leq f_1(\theta)$. Then, observe that for $(\theta, \theta^+, \theta^-) \in C_2$, $\theta^+ + \theta^- \geq \|\theta\|_1$ which implies that $f_2(\theta, \theta^+, \theta^-) \geq f_1(\theta)$. Hence, for all $(\theta, \theta^+, \theta^-) \in C_2$, $f_2(\theta, \theta^+, \theta^-) \geq f_2(\theta_1, \theta_1^+, \theta_1^-)$ which finally implies that $\theta_1 = \theta_2$ and completes the proof. $\square$

Proof of Proposition 3.2: by definition, the dual of problem (3.3) is:

$$\begin{aligned} \max_{\theta, \theta^+, \theta^- \in \mathbb{R}^n, \mu_+, \mu_-, \tau_+, \tau_-, \nu \in \mathbb{R}^n_+} \quad & \Big(\frac{1}{2}\|\theta - d\|_2^2 + \lambda e^T(\theta^+ + \theta^-) - \mu_+^T(\theta^+ - \theta) \\ & -\mu_-^T(\theta^- + \theta) - \tau_+^T\theta^+ - \tau_-^T\theta^- + \nu^T\Phi\theta\Big), \qquad \text{(B.1)} \\ \text{s.t} \quad & \theta - d + \mu_+ - \mu_- + \nu^T\Phi = 0, \\ & \lambda e - \mu_+ - \tau_+ = 0, \\ & \lambda e - \mu_- - \tau_- = 0 \end{aligned}$$

Since $\tau_+ \geq 0$ and $\tau_- \geq 0$, we must have $\mu_+ \leq \lambda e$ and $\mu_- \leq \lambda e$. By defining $\mu = \mu_+ - \mu_-$ and by eliminating τ_+ and τ_- in (B.1), we obtain the formulation (3.4) which completes the proof. $\square$

Proof of Theorem 4.1: Let $\hat{\boldsymbol{\theta}}^c = \mathbf{d} - \lambda\hat{\boldsymbol{\mu}}^c - \Phi^T\hat{\boldsymbol{\nu}}$ be the unique solution of problem (3.1) where $\begin{pmatrix}\hat{\boldsymbol{\mu}}^c \\ \hat{\boldsymbol{\nu}}\end{pmatrix} \in \mathbb{R}^{n+m}$ denotes the solution of problem (3.5). Let $\hat{\boldsymbol{\theta}} = \mathbf{d} - \lambda\hat{\mu}$ be the classical soft thresholding estimator which corresponds to the optimal solution if the problem (3.1) is unconstrained. Now, by the remark at the end of subsection 3.1 note that $-\mathbf{e} \leq \hat{\boldsymbol{\mu}} \leq \mathbf{e}$ which implies that $\|\lambda\hat{\boldsymbol{\mu}}^c + \Phi^T\hat{\boldsymbol{\nu}} - \mathbf{d}\|_2^2 \leq \|\lambda\hat{\boldsymbol{\mu}} - \mathbf{d}\|_2^2$ since $\begin{pmatrix}\hat{\boldsymbol{\mu}}^c \\ \hat{\boldsymbol{\nu}}\end{pmatrix}$ is a minimum for problem (3.5).

Now,

$$\|\hat{\boldsymbol{\theta}}^c - \boldsymbol{\theta}\|_2^2 = \|\boldsymbol{\theta}\|_2^2 + \|\hat{\boldsymbol{\theta}}^c\|_2^2 - 2\langle\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^c\rangle$$

and

$$\|\boldsymbol{\theta}\|_2^2 = \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2 - \|\hat{\boldsymbol{\theta}}^c\|_2^2 + 2\langle\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}\rangle.$$

ACCEPTED MANUSCRIPT

Hence, since $\|\hat{\boldsymbol{\theta}}^c\|_2^2 \leq \|\hat{\boldsymbol{\theta}}\|_2^2$, the following inequality holds:

$$\|\hat{\boldsymbol{\theta}}^c - \boldsymbol{\theta}\|_2^2 \leq \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2 + 2\langle \boldsymbol{\theta}, \hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}^c \rangle$$

If we assume that the function f satisfies the constraints of problem (3.1) namely $\Phi\boldsymbol{\theta} \leq 0$, then $\langle \boldsymbol{\theta}, \Phi^T \hat{\boldsymbol{v}} \rangle = (\Phi\boldsymbol{\theta})^T \hat{\boldsymbol{v}} \leq 0$ since $\hat{\boldsymbol{v}} \geq 0$. Hence, if f verifies the constraints that we want to impose, the following inequality holds

$$\|\hat{\boldsymbol{\theta}}^c - \boldsymbol{\theta}\|_2^2 \leq \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2 + 2\lambda \langle \boldsymbol{\theta}, \hat{\boldsymbol{\mu}}^c - \hat{\boldsymbol{\mu}} \rangle$$

Given that $-\mathbf{e} \leq \hat{\boldsymbol{\mu}} \leq \mathbf{e}$ and $-\mathbf{e} \leq \hat{\boldsymbol{\mu}}^c \leq \mathbf{e}$, we have that $|\langle \boldsymbol{\theta}, \hat{\boldsymbol{\mu}}^c - \hat{\boldsymbol{\mu}} \rangle| \leq 2\sum_{i=1}^n |\boldsymbol{\theta}_i|$. Recall that the empirical wavelet coefficients are such that $\theta_i \approx \sqrt{n}\beta_i$ where β_i is the corresponding continuous wavelet coefficient.

Hence, if f is in the Besov ball $B_{1,1}^r(C)$ then $\sum_{i=1}^n |\beta_i|$ is uniformly bounded and we obtain that for the classical universal threshold $\lambda = \sigma\sqrt{2\log(n)}$:

$$R(\hat{f}^c, f) = \frac{1}{n}\|\hat{\boldsymbol{\theta}}^c - \boldsymbol{\theta}\|_2^2 \leq R(\hat{f}, f) + 0(\frac{\sqrt{\log(n)}}{\sqrt{n}})$$

The result now follows using Theorem 4 of Antoniadis and Fan [1]. □

Acknowledgement. The authors wish to thank Dr Berwin Turlach for fruitful and helpful discussions on the topic. The authors are grateful to the reviewers for their remarks which led to an improvement of the paper.

Accepted manuscript

References

[1] Antoniadis, A. and Fan, J. (2001). Regularization of Wavelets Approximations (with discussion), *J. Amer. Statist. Assoc.*, **96**, No 455, 939-963.

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

[3] Barlow, R., Bartholomew, D., Bremner, J. and Brunk, H. (1972). *Statistical Inference under Order Restrictions*, John Wiley and Sons, London.

[4] Boyd, S. and Vandenberghe, L. (2004). Convex optimization, *Cambridge University Press*.

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

[6] Delouille, V. Simoens, J. and von Sachs, R. (2004). Smooth design-adapted wavelets for nonparametric stochastic regression, *J. Amer. Statist. Assoc.*, **99**, 643–658.

[7] Fletcher, R. (1987). *Practical methods of optimization*, John Wiley & Sons, Inc. New York, 2nd edition.

[8] Gallant, R. (1981). On the bias inßexible functional forms and an essential unbiased form : The fourier ßexible form, *Journal of Econometrics*, **15**, 211– 245.

ACCEPTED MANUSCRIPT

[9] Gijbels, I. (2005). Monotone regression, to appear in *Encyclopedia of Statistical Sciences, Second Edition*. Editors S. Kotz, N.L. Johnson, C.B. Read, N. Balakrishnan, and B. Vidakovic. Wiley, New York.

[10] Hall, P. and Huang L.-S. (2001). Nonparametric kernel regression subject to monotonicity constraints, *The Annals of Statistics*, **29**, 624–647.

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

[12] Hanson, D. L., Pledger, G. and Wright, F. T. (1973). On consistency in monotonic regression, *The Annals of Statistics*, **1**, 3, 401– 421.

[13] Härdle, W., Kerkyacharian, G., Picard, D., Tsybakov, A. (1998). *Wavelets, Approximation, and Statistical Applications*. Lecture Notes in Statistics, Vol. 129. Springer-Verlag, Berlin.

[14] He, X. and Shi, P. (1998). Monotone b-spline smoothing, *Journal of the American statistical Association*, **93**, 442, 643–650.

[15] Kelly, C. and Rice, J. (1990). Monotone smoothing with application to dose response curves and the assessment of synergism, *Biometrics*, **46**, 1071–1085.

[16] Kerkyacharian G. and Picard, D. (2003.) Regression in random design and warped wavelets. *Technical report*, Department of Mathematics, University Paris X, Nanterre, France.

[17] Kovac, A. and Silverman, B. W. (2000), Extending the scope of wavelet regression methods by coefficient-dependent thresholding, *J. Amer. Statist. Ass.*, **95**, 172–183.

[18] Mammen, E. (1991). Estimating a smooth monotone regression function, *The Annals of Statistics*, **19**, 724–740.

[19] Mammen, E. , Marron, J.S. , Turlach, B.A. and Wand, M.P. (2001). A general projection framework for constrained smoothing, *Statist. Sci.*, **16**, 232–248.

[20] Mammen, E. , Thomas-Agnan, C. (1999). Smoothing splines and shape restrictions, *Scand. J. Statist.* , **26**, 239–252.

[21] Ramsay, J. O. (1988). Monotone regression splines in action (with comments), *Statist. Sci.*, **3**, 425–461.

[22] Robertson, T., Wright, F. and Dykstra, R. (1988). *Order Restricted Statistical Inference*, Wiley Series in Probability and Mathematical Statistics , John Wiley and Sons.

[23] Rockafellar, R.T. (1969). Convex Analysis, *Princeton University Press, Princeton, N.J.*.

[24] Sardy, S., Percival, D. B., Bruce A., G., Gao, H.-Y. and Stuelzle, W. (1999). Wavelet shrinkage for unequally spaced data, *Statistics and Computing*, **9**, 65–75.

[25] Zhang, J.-T. (2004). A simple and efficient monotone smoother using smoothing splines, *Journal of Nonparametric Statistics*, **16**, 5, 779–796.

Accepted manuscript