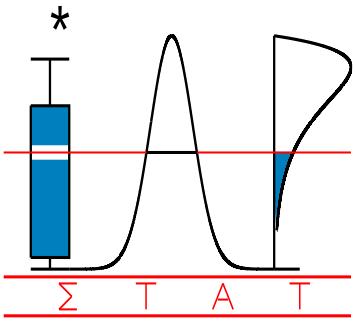
T E C H N I C A L

R E P O R T

0426

WAVELET KERNEL PENALIZED ESTIMATION FOR NON-EQUISPACED DESIGN REGRESSION

AMATO U., ANTONIADIS A. and PENSKY M.



I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

Wavelet kernel penalized estimation for non-equispaced design regression

Umberto Amato
Istituto per le Applicazioni del Calcolo 'M. Picone' CNR - Sezione di Napoli
Via Pietro Castellino 111, 80131 Napoli , Italy.
Anestis Antoniadis,
Laboratoire IMAG-LMC, University Joseph Fourier,
BP 53, 38041 Grenoble Cedex 9, France.
Marianna Pensky[*]
Department of Statistics , University of Central Florida,
Orlando, FL 32816 -1364, USA.

Abstract

The paper considers regression problems with univariate design points. The design points are irregular and no assumptions on their distribution are imposed. The regression function is retrieved by a wavelet based reproducing kernel Hilbert space (RKHS) technique with the penalty equal to the sum of blockwise RKHS norms. In order to simplify numerical optimization, the problem is replaced by an equivalent quadratic minimization problem with an additional penalty term. The computational algorithm is described in detail and is implemented with both the sets of simulated and real data. Comparison with existing methods showed that the technique suggested in the paper does not oversmooth the function and is superior in terms of the mean squared error. It is also demonstrated that under additional assumptions on design points the method achieves asymptotic optimality in a wide range of Besov spaces.

Key words: Reproducing kernel, wavelet decomposition, penalization, Besov spaces, smoothing splines ANOVA, entropy.

[*]Corresponding author. E-mail: mpensky@pegasus.cc.ucf.edu

1 Introduction

Consider the regression problem $y_i = f(x_i) + \epsilon_i$, $i = 1, \ldots, n$, where the x_i's are univariate design points, the ϵ_i's are i.i.d. mean-zero random errors with variance σ^2 and f is an unknown compactly supported regression function to be estimated. In a classical parametric regression analysis, f is assumed to be of the form $f(x, \boldsymbol{\beta})$, which is known up to the parameters $\boldsymbol{\beta}$, which are to be estimated from the data. In such a case the dimension of the model space, i.e. the number of unknown parameters, is presumably much smaller than the sample size n. The estimator is judged in terms of prediction accuracy and interpretability. However, parametric models often incur model bias. To avoid this, an alternative approach to estimation is to allow f to vary in a high (possibly infinite) dimensional function space, leading to various nonparametric estimation methods. A popular approach to the nonparametric estimation of f is via the minimization of a penalized least squares functional and the methods developed in this paper can be casted into this setting.

Given a wavelet type expansion of f, we define an estimation procedure for f closely related to soft wavelet thresholding. It is now known (see Antoniadis and Fan (2001)) that separable penalized least-squares methods with appropriate additive penalties provide a unified framework for many seemingly different wavelet thresholding rules in different nonparametric function estimation contexts and enables one to systematically study a class of wavelet estimators simultaneously. In this work, we consider and study a class of non-separable wavelet estimators for the nonparametric regression problem using a penalized least-squares approach with non-additive penalties. The penalties are chosen in order to control the smoothness of the resulting estimator. For this, we focus on semi-norm penalties. Such estimations issues are well-known in the literature and have been studied by several authors such as Wahba (1990), Green and Silverman (1994), Wahba et al. (1995) in the general nonparametric setting of the smoothing spline framework. We take for penalty, a weighted sum of wavelet details spaces norms. Asymptotically, the penalty will be equivalent to a Besov semi-norm. We will investigate several choices of the penalty and show that, in all cases, the minimization problem has a solution. We will point out the cases where it is possible to give a direct explicit solution of the optimization program and, in the other cases, we will provide an approximation of the exact case. Our approach provides a unified framework for several recent proposals of

wavelet thresholding rules and gives an alternative interpretation of the penalty term in separable penalized least-squares methods with an additive penalty being the L_1 norm of the wavelet coefficients: it is the sum of component norms.

The regularization procedure that we propose is inspired by the sparse kernel selection approach in Gunn and Kandola (2002) and the COSSO (COmponent Selection and Smoothing Operator) approach for fitting smoothing spline ANOVA models recently proposed by Lin and Zhang (2003). However, the motivation of our method as well as the setting is different and relates to the penalized least-squares method for wavelet regression developed by Antoniadis and Fan (2001).

Some background on wavelets, reproducing kernel Hilbert spaces and the general methodology that we have adopted are introduced in Section 2. In Section 3 we present our estimator and we show that we can reach an asymptotic optimal rate of convergence, provided that we know the regularity of the function we try to estimate. We also present there a computational algorithm and briefly discuss the choice of the tuning parameter. Simulations and a real example analysis are given in Section 4, where we compare our method with other popular nonparametric regression methods. The proofs of the main results are given in the Appendix.

2 Wavelet series expansions and wavelet kernels

We briefly recall first some relevant facts about wavelet series expansions, the discrete wavelet transform, and the class of (inhomogeneous) Besov spaces on the unit interval that we need further.

2.1 Wavelet series expansions

Let $L_2([0,1))$ denote the Hilbert space of $\mathbb{Z}$-periodic real-valued functions on $\mathbb{R}$ that are square integrable over the one-dimensional torus group, parameterized by $[0,1)$, with scalar product

$$\langle f, g \rangle = \int_{[0,1)} f(x)g(x)dx,$$

and associated norm

$$\|f\| := \sqrt{\langle f, f \rangle}.$$

Let $G_{-1} = \{-1\} \times \{0\}$, $G_0 = \{0\} \times \{0, 1\}$ and for each integer $L \geq 1$ let $G_L = \{L\} \times \{k \in \{0, \ldots, 2^L\};\ k/2 \notin \mathbb{Z}\}$.

We assume the reader is familiar with the concept of an orthonormal wavelet basis and associated multiresolution analysis. We construct an orthonormal wavelet basis for $L_2([0, 1))$ by periodizing an orthonormal basis for $L_2(\mathbb{R})$ generated by dilations and translations of a compactly supported scaling function, ϕ, and a compactly supported mother wavelet, ψ, associated with an r-regular ($r \geq 0$) multiresolution analysis of $L^2(\mathbb{R})$. The resulting orthonormal basis provides an orthogonal decomposition

$$L_2([0, 1)) = V_0 \oplus W_0 \oplus W_1 \oplus \cdots,$$

where V_0 consists of constant functions (spanned by $\phi_{0,0} = \psi_{-1,0}$) and W_j is a 2^j-dimensional space spanned by wavelets indexed by G_j. For $L \geq 1$ define

$$V_L = V_0 \oplus \oplus_{j=0}^{L-1} W_j.$$

The space V_L has an orthonormal wavelet basis comprising the constant function $\psi_{-1,0} = \mathbf{1}$ together with the orthonormal wavelet basis functions for each space W_j, $j < L$. It also has an orthonormal scaling function basis obtained by translating by $\{k \in \{0, \ldots, 2^L\};\ k/2 \notin \mathbb{Z}\}$ the periodized scaling function ϕ scaled by 2^{-j}. For any $f \in L^2([0, 1])$, and any integer $j_0 \geq 0$, we denote by $u_{j_0 k} = \langle f, \phi_{j_0 k} \rangle$ ($k = 0, 1, \ldots, 2^{j_0} - 1$) the scaling coefficients and by $w_{jk} = \langle f, \psi_{jk} \rangle$ ($j \geq j_0;\ k = 0, 1, \ldots, 2^j - 1$) the wavelet coefficients of f for the orthonormal periodic wavelet basis defined above; the function f is then expressed in the form

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

The whole set of indexes pairs (j, k) that describes all wavelets in this basis will be denoted by $G = \cup_{j \geq -1} G_j$. With such a notation, any function $f \in L_2([0, 1))$ admits therefore the infinite wavelet expansion

$$f = \sum_{g \in G} f_g \psi_g$$

where ψ_g is the wavelet basis function indexed by $g \in G$ and f_g is the corresponding expansion coefficient.

The scaling function expansion coefficients of a function $f \in V_L$ are represented by an element $s_L(f) \in R^{G_L}$ and the wavelet expansion coefficients of f are obtained by

$W_L(s_L) \in R^{G_L}$ where

$$W_L : R^{G_L} \to R^{G_L}$$

is the L-level discrete wavelet transform. The discrete wavelet transform is orthogonal and invertible and its inverse can be efficiently implemented with $O(2^L)$ operations using Mallat's algorithms (Mallat (1999)). For a nice account of the DWT and IDWT in terms of filter operators we refer to, for example, Nason (1998).

To end this subsection, we also mention that if $f \in L_2([0,1))$ is continuous then

$$\lim_{L\to\infty} \max |2^{L/2} s_L(f) - f|_{G_L}| = 0. \tag{1}$$

For detailed expositions of the mathematical aspects of wavelets we refer to, for example, Meyer (1992), Daubechies (1992) and Mallat (1999), while comprehensive expositions and reviews on wavelets applications in statistical settings are given in, for example, Härdle et al. (1998), Vidakovic (1999), Abramovich *et al.* (2000) and Antoniadis *et al.* (2001).

2.2 Reproducing Kernel Hilbert Spaces

Regularization in Hilbert spaces is an approximation framework that is theoretically well founded. Reproducing Kernel Hilbert Spaces (RKHS) provide a unified context for regularization in a wide variety of statistical modelling and function estimation problems. In this subsection we briefly review a few important facts about RKHS, that we are going to use later on.

A RKHS is a Hilbert space (Aronszajn (1950)) in which all the point evaluations are bounded linear functionals. Letting $\mathcal{H}$ be a Hilbert space of functions on some domain $\mathcal{T}$, this means, that for every $t \in \mathcal{T}$ there exists an element $\eta_t \in \mathcal{H}$, such that

$$f(t) = \langle \eta_t, f \rangle, \quad \forall f \in \mathcal{H},$$

where $\langle \cdot, \cdot \rangle$ is the inner product in $\mathcal{H}$. Let $s, t \in \mathcal{T}$ and set $\langle \eta_s, \eta_t \rangle = K(s,t)$. Then $K(s,t)$ is positive definite on $\mathcal{T} \times \mathcal{T}$, that is, for any distinct points $t_1, \ldots, t_n \in \mathcal{T}$, the $n \times n$ matrix with j,k-entry $K(t_j, t_k)$ is positive definite. The kernel K is called the reproducing kernel (RK) for $\mathcal{H}$. It is a theorem that $\eta_t = K(t, \cdot)$ and therefore $\langle K(s,\cdot), K(t,\cdot) \rangle = K(s,t)$, this being the origin of the term "reproducing kernel".

The famous Moore-Aronszajn theorem (Aronszajn (1950)) states that for every positive definite function $K(\cdot,\cdot)$ on $\mathcal{T} \times \mathcal{T}$, there exists a unique RKHS and vice versa.

The Hilbert space associated with K can be constructed as containing all finite linear combinations of the form $\sum_j a_j K(t_j, \cdot)$, and their limits under the norm induced by the inner product $\langle K(s,\cdot), K(t,\cdot)\rangle = K(s,t)$. Note that absolutely nothing has been said about $\mathcal{T}$.

Remark 2.1 *Tensor sums and products of RK's are RK's, which allow the building up of rather general spaces on rather general domains. Closed subspaces of RKHS are also RKHS, and the RK's can be obtained by e. g. projecting the representers of evaluation in $\mathcal{H}$ onto the subspace. The above apply to any domain on which it is possible to define a positive definite function, a matrix being a special case when $\mathcal{T}$ has only a countable or finite number of points.*

We are now ready to recall a (very special case of a) general lemma about optimization problems in RKHS (see Kimeldorf and Wahba (1971)).

Lemma 2.1 (Representer Theorem) *Given a set of observations $\{(y_i, t_i);\ i = 1, 2, \ldots, n\}$, where y_i is a real number and $t_i \in \mathcal{T}$, and given K and (possibly) given some particular functions $\{\Phi_1, \ldots \Phi_M\}$ on $\mathcal{T}$, find f of the form $f(s) = \sum_{d=1}^{M} a_d \Phi_d(s) + h(s)$ where $h \in \mathcal{H}_K$ to minimize*

$$\mathcal{I}(f, \mathbf{y}) := \sum_{i=1}^{n} \mathcal{C}(y_i, f(t_i)) + \lambda^2 \|h\|^2_{\mathcal{H}_K}, \tag{2}$$

where $\mathcal{C}$ is a convex function of f. Assuming that the minimizer of $\mathcal{C}(y_i, f(t_i))$ in the span of the Φ_d's is unique, the minimizer of $\mathcal{I}(f, \mathbf{y})$ has a representation of the form:

$$f(s) = \sum_{d=1}^{M} a_d \Phi_d(s) + \sum_{i=1}^{n} c_i K(t_i, s). \tag{3}$$

The coefficient vectors $\mathbf{a} = (a_1, \ldots, a_M)^T$ and $\mathbf{c} = (c_1, \ldots, c_n)^T$ are found numerically by substituting (3) into the first term in (2). The minimization generally has to be done numerically by an iterative descent method, except in the case that $\mathcal{C}$ is quadratic in f, in which case a linear system has to be solved.

When $K(\cdot, \cdot)$ is a smooth function of its arguments and n is large, it has been found that excellent approximations to the minimizer of (2) for various $\mathcal{C}$ can be found with functions of the form:

$$f(s) = \sum_{d=1}^{M} a_d \Phi_d(s) + \sum_{j=1}^{L} c_j K(t_j^*, s),$$

where the $t_1^*, \ldots, t_L^*$ are a relatively small subset of $t_1, \ldots, t_n$, thus reducing the computational load. The $t_1^*, \ldots, t_L^*$ may be chosen in various ways (see for example Lin et. al. (2000)), as a random subset, by clustering the t_i's and selecting from each cluster, or by a greedy algorithm, depending on the problem.

2.3 Wavelet-based Norms

We now define a class of wavelet-based Hilbert spaces. For any function

$$\Gamma : G \to [0, \infty)$$

define the Hilbert space

$$\mathcal{H}_\Gamma = \{f \in L_2([0,1)) : \sum_{g \in G} \Gamma(g) |f_g|^2 < \infty\},$$

with scalar product

$$\langle f, h \rangle_\Gamma = \sum_{g \in G} f_g h_g \Gamma(g),$$

and associated norm $\|\cdot\|_\Gamma$. Clearly, since G_L is a finite subset of G, we have $V_L \subset \mathcal{H}_\Gamma$ for every $L \geq 0$. Moreover, for any $f \in \mathcal{H}_\Gamma$,

$$\lim_{L \to \infty} \|f - \pi_L(f)\|_\Gamma = 0, \tag{4}$$

where π_L denotes orthogonal projection of $L_2([0,1))$ onto its closed subspace V_L.

We now construct wavelet-based norms that yield reproducing kernel Hilbert spaces whose functions are continuous. First construct an orthonormal wavelet basis of continuous compactly supported wavelets. Second, construct a function $\Gamma : G \to [0, \infty)$ such that

$$\sum_{j \geq 0} 2^{j/2} \Gamma_j^{-1/2} = B_1 < \infty \tag{5}$$

where

$$\Gamma_j = \min_{g \in G_{j+1}} |\Gamma(g)|.$$

To show that $\mathcal{H}_\Gamma$ is a RKHS choose $M > 0$ and $B_2 > 0$ such that for any $x \in [0,1)$ and for any $j \geq 0$ there are at most M wavelet basis functions indexed by elements in G_{j+1} that are nonzero at x, and such that the maximum modulus of the wavelet basis functions indexed by G_{j+1} is $\leq B_2 2^{j/2}$. This is obviously possible since the wavelets that we are

considering are compactly supported and periodic. Define $\gamma := MB_2\sqrt{B_1}$. Let $f \in \mathcal{H}_\Gamma$ and define

$$f_j := \max_{g \in G_{j+1}} |f_g|.$$

Then the Schwartz inequality implies

$$|f(x)| \leq MB_2 \sum_{j \geq 0} f_j 2^{j/2} \leq \gamma \|f\|_\Gamma.$$

Therefore $\mathcal{H}_\Gamma$ is a RKHS and since it has a dense subspace of continuous functions and γ is independent of x, all the functions in $\mathcal{H}_\Gamma$ are continuous. The corresponding reproducing kernels are

$$K^\Gamma(x, \cdot) = \sum_{g \in G} \frac{\psi_g(x)}{\Gamma(g)} \psi_g, \quad x \in [0, 1).$$

Note also that by Remark 2.1 and by definition of the index set G, the kernel K defined above can also be written as a sum of the reproducing kernels

$$K_j^\Gamma(s, t) = \sum_{k=0}^{2^j - 1} \frac{\psi_{j,k}(s)}{\Gamma((j,k))} \psi_{j,k}(t),$$

which means that the RKHS $\mathcal{H}_\Gamma$, can be decomposed into a direct sum of wavelet RKHS's as

$$\mathcal{H}_\Gamma = V_0 \oplus \oplus_{j \geq 0} \mathcal{W}_{j,\Gamma}, \tag{6}$$

where each "detail" space is the RKHS associated to the kernel K_j^Γ, i.e. is the RKHS spanned by a set of wavelets of scale j. Note also that when Γ is only a function of j in G_j, then the kernel K may be viewed as a weighted, Γ_j^{-1}, infinite linear sum of kernels

$$K_j(s, t) = \sum_{k=0}^{2^j - 1} \psi_{j,k}(s) \psi_{j,k}(t).$$

Remark 2.2 *By the representer Theorem if the set $X = \{t_i; i = 1, \ldots, n\}$ is such that the restrictions of functions in $\mathcal{H}_\Gamma$ spans $\mathbb{R}^n$, the solution to the minimization problem*

$$f_\lambda = \text{argmin}_{h \in \mathcal{H}_\Gamma} \mathcal{I}(h, \mathbf{y})$$

where

$$\mathcal{I}(h, \mathbf{y}) = \sum_{i=1}^{n} (h(t_i) - y_i)^2 + \lambda^2 \|h\|_{\mathcal{H}_\Gamma}^2,$$

can be written explicitly with the functions $K^\Gamma(t_i, \cdot)$*, i .e.*

$$f_\lambda(x) = \sum_{i=1}^{n} u_i K^\Gamma(t_i, x),$$

where the vector of coefficients $\mathbf{u} = (u_1, \ldots, u_n)^T$ *is given by*

$$\mathbf{u} = (K^\Gamma + \lambda^2 I_n)^{-1}(Y),$$

and K^Γ *denotes, with some abuse of notation, the* $n \times n$ *Gram matrix* $(K^\Gamma(t_i, t_j))$*.*

As noticed before, the representation (6) involves an infinite decomposition of the detail space, and despite the fact that we are dealing with compactly supported functions the computational complexity for computing f_λ is high, especially for large samples. In this situation we will have to be content with approximations. The basic idea is simple. Instead of using the infinite decomposition $\oplus_{j=0}^{\infty}\mathcal{W}_{j,\Gamma}$ in (6), we truncate it up to a maximum resolution J. By the properties of the wavelet basis, the resulting nested sequence of finite dimensional subspaces $\mathcal{H}_{J,\Gamma} = V_0 \oplus \oplus_{j=0}^{J}\mathcal{W}_{j,\Gamma}$ defines a multiresolution analysis of $\mathcal{H}_\Gamma$ and we can then compute approximations to f_λ by choosing a resolution level J and restricting the functional $\mathcal{I}(h, \mathbf{y})$ to $\mathcal{H}_{J,\Gamma}$. More precisely, for J sufficiently large, using equations (1) and (4) an approximation $f_{J,\lambda}$ of f_λ may be computed by solving instead the minimization problem in the finite dimensional approximation space $\mathcal{H}_{J,\Gamma}$ defined by the truncated kernel

$$K_J^\Gamma(x, y) = \sum_{g \in \cup_{0 \le j \le J} G_j} \frac{\psi_g(x)}{\Gamma(g)}\psi_g(y), \quad x, y \in [0, 1).$$

Indeed, denoting by K_J^Γ the corresponding Gram matrix and by $\mathbf{u}_J$ the corresponding coefficients, and expressing

$$\mathbf{u} - \mathbf{u}_J = (K_J^\Gamma + \lambda^2 I_n)^{-1}(K_J^\Gamma - K^\Gamma)(K^\Gamma + \lambda^2 I)^{-1} f|_X,$$

yields

$$\max |u_i - u_{i,J}| \le \|(K^\Gamma + \lambda^2 I_n)^{-1}\|_\infty \, \|(K_J^\Gamma + \lambda^2 I_n)^{-1}\|_\infty \|K^\Gamma - K_J^\Gamma\|_\infty \max_{x \in X} |f(x)|,$$

where $\|A\|_\infty := \max_{y \in X} \sum_{x \in X} |A(x, y)|$ denotes the ℓ^∞ operator norm of a matrix indexed by X. Since the maximum and minimum eigenvalues of the matrices K_J^Γ satisfy

$$\mu_{\max}(K_J^\Gamma) \le \mu_{\max}(K_{J+1}^\Gamma) \cdots \to \mu_{\max}(K^\Gamma),$$

and

$$\mu_{\min}(K_J^\Gamma) \geq \mu_{\min}(K_{J+1}^\Gamma) \cdots \to \mu_{\min}(K^\Gamma) > 0,$$

it follows that $\|(K_J^\Gamma + \lambda^2 I_n)^{-1}\|_\infty$ are uniformely bounded in J. Furthermore, by (4),

$$\lim_{J\to\infty} \|K^\Gamma - K_J^\Gamma\|_\infty = 0.$$

Therefore

$$\lim_{J\to\infty} \max_i |u_i - u_{i,J}| = 0 \tag{7}$$

and

$$f_{J,\lambda} \to f_\lambda,$$

as $J \to \infty$.

To end this section, note that, in the above definition of wavelet-based RKHS norm, if $s > 1/2$ and $\Gamma(j,k)$ equals 2^{2js} on G_j, then $\mathcal{H}_\Gamma$ equals the Sobolev space $B_{2,2}^s([0,1))$ of index s whenever the wavelet basis functions are of regularity r greater than or equal to s. In this case the wavelet-based norms are equivalent to the standard Sobolev norms.

3 Wavelet kernel penalized estimation

The wavelet-based reproducing kernel Hilbert spaces defined in the previous sections and the general results on RKHS allows us to define a penalized least square wavelet procedure for estimating the values of the unknown regression function f at *the design points*.

To simplify the analysis we will assume hereafter that the weight function Γ is only a function of j in G_j, and therefore the wavelet-based reproducing kernel K^Γ defined in the previous section is a weighted, Γ_j^{-1}, linear sum of wavelet tensor product kernels

$$K_j(s,t) = \sum_{k=0}^{2^j-1} \psi_{j,k}(s)\psi_{j,k}(t).$$

Denoting $\mathcal{W}_j$ the RKHS associated to K_j (a classical detail space at scale j), one has $\mathcal{W}_j^\Gamma = \Gamma_j^{-1}\mathcal{W}_j$ and the function space $\mathcal{H}_\Gamma$ can be written as

$$\mathcal{H}_\Gamma = \{1\} \oplus \oplus_{j\geq 0} \Gamma_j^{-1}\mathcal{W}_j, \tag{8}$$

since V_0 is also the subspace of $L^2([0,1))$ spanned by the constant functions on $[0,1)$. By the orthogonality of the wavelet basis and the scalar product in $\mathcal{H}_\Gamma$ note also that in the above decomposition the subspaces $\mathcal{W}_j^\Gamma$ are orthogonal subspaces of $\mathcal{H}_\Gamma$.

In the following we will denote by $P_j f$ (resp. $P_j^\Gamma f$) the orthogonal projection of f onto $\mathcal{W}_j$ (resp. the orthogonal projection of f onto $\mathcal{W}_j^\Gamma$). In analogy with a traditional smoothing spline ANOVA type procedure one way to estimate f could be to find $f \in \mathcal{H}_\Gamma$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n}\{y_i - f(x_i)\}^2 + \lambda^2 \sum_{j\geq 0} \theta_j^{-1} \|P_j^\Gamma f\|_\Gamma^2, \tag{9}$$

where $\theta_j \geq 0$. If $\theta_j = 0$, then the minimizer is taken to satisfy $\|P_j^\Gamma f\|_\Gamma^2 = 0$, using the convention $0/0 = 0$. The smoothing parameter λ is confounded with the θ's, but it is usually included in the setup for computational purposes. Note also that by Remark 2.2, for $J \simeq \log_2(n)$ the above minimum is in fact reached in $\mathcal{H}_{J,\Gamma}$ and the minimisation problem (8) can be restated as : find $f \in \mathcal{H}_{J,\Gamma}$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n}\{y_i - f(x_i)\}^2 + \lambda^2 \sum_{0\leq j\leq J} \theta_j^{-1} \|P_j^\Gamma f\|_{J,\Gamma}^2. \tag{10}$$

The first term in the above expression discourages the lack of fit of f to the data, the second term penalizes the roughness of f and the smoothing parameter λ controls the trade-off between the two conflicting goals. Such an estimation procedure is controled by a quadratic penalty and as such produces linear estimates that have good rates for smooth functions only. We could propose instead finding $f \in \mathcal{H}_{J,\Gamma}$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n}\{y_i - f(x_i)\}^2 + \lambda^2 R_J(f), \text{ with } R_J(f) = \sum_{j=0}^{J} \|P_j^\Gamma f\|_{J,\Gamma}. \tag{11}$$

The penalty term in (11) is a sum of wavelet-based RKHS norms, instead of the squared RKHS norm employed in (10). Note that $R_J(f)$ is not a norm in $\mathcal{H}_{J,\Gamma}$ but a pseudo-norm in the following sense: $R_J(f) \geq 0$, $R_J(cf) = |c|R_J(f)$, $R_J(f+h) \leq R_J(f) + R_J(h)$ for any $f, h \in \mathcal{H}_{J,\Gamma}$, and, $R_J(f) > 0$ for any non constant $f \in \mathcal{H}_{J,\Gamma}$. Moreover

$$\sum_{j=0}^{J} \|P_j^\Gamma f\|_{\mathcal{H}_{J,\Gamma}}^2 \leq R_J(f)^2 \leq J \sum_{j=0}^{J} \|P_j^\Gamma f\|_{\mathcal{H}_{J,\Gamma}}^2. \tag{12}$$

Another difference between the procedure defined by (11) and the one defined by (10) is that there is only one smoothing parameter λ instead of multiple smoothing parameters θ's.

Classical level dependent wavelet soft thresholding in wavelet regression with an equidistant design can be seen as a special case of the procedure defined by (11). Indeed,

assuming that $n = 2^J$, wavelet thresholding can be seen as a penalized separable least-squares procedure that looks for the minimum of

$$\|\mathbf{y} - W_n\boldsymbol{\beta}\|_n^2 + \lambda^2 \sum_{j=0}^{J} \sum_{k=0}^{2^j-1} 2^{js}|\beta_{j,k}|,$$

where W_n denotes the J-level discrete wavelet transform, $\|\cdot\|_n$ is the Euclidian norm of $\mathbb{R}^n$ and $\boldsymbol{\beta}$ denotes the vector of wavelet coefficients of f. Considering that for each $(j,k) \in G_j$, the tensor product $\psi_{j,k}(s)\psi_{j,k}(t)$ defines a wavelet kernel and denoting by $\mathcal{W}_{j,k}$ the corresponding (one-dimensional) RKHS, one can see that penalizing the l_1 norm of wavelet coefficients is equivalent to (12) with $R_J(f) = \sum_{j=0}^{J}\sum_{k=0}^{2^j-1} \|P_{j,k}^{\Gamma} f\|_{\mathcal{W}_{j,k}}$ with $\Gamma(j,k) = 2^{2js}$, thus interpreting the penalty in standard wavelet thresholding regression as the sum of the norm of function components. An interpretation of this sort has been also suggested in Canu et al. (2003) within the context of SVM.

In the case of an equidistant design, the estimator resulting from the procedure (11) is obtained by minimizing with respect to $\boldsymbol{\beta}$ the expression

$$\|W_n^T\mathbf{y} - \boldsymbol{\beta}\|_n^2 + \lambda^2 \sum_{j=0}^{J} \sqrt{\Gamma_j} \sqrt{\sum_{k=0}^{2^j-1} \beta_{j,k}^2}. \tag{13}$$

For each $(j,k) \in G_j$, simple calculations show that the solution is given by solving the nonlinear equations

$$\beta_{j,k}(1 + \frac{\lambda^2\sqrt{\Gamma_j}}{\|\beta j, \cdot\|})^2 = d_{j,k}$$

where $d_{j,k}$ denote the (j,k) empirical wavelet coefficient, and it is easy to see that

$$\beta_{j,k} = d_{j,k}\left(1 - \lambda^2\sqrt{\Gamma_j}\Big/(2\|d_{j,\cdot}\|)\right)_+.$$

The procedure defined by Eq. (13) leads therefore to a group level-by-level wavelet thresholding of all empirical wavelet coefficients within each scale. However, such a procedure may not be optimal for non homogeneous functions and the resulting reconstruction is often over-smoothed. Hall et al (1999), Cai (1999) and Cai and Silverman (2001) considered block thresholding for wavelet function estimation for equispaced designs which thresholds empirical wavelet coefficients in groups within each scale rather that individually with a goal to increase estimation precision by using information about neighboring coefficients. More precisely, let us partition all wavelets coefficients at level j

into blocks T_{jm} of length L_{jm}. All blockwise procedures suggested so far in the literature try to mimic the benchmark

$$\hat{\beta}_{jk} = \frac{L_{jm}^{-1}\sum_{\ell\in T_{jm}}\beta_{j\ell}^2}{L_{jm}^{-1}\sum_{\ell\in T_{jm}}\beta_{j\ell}^2 + \sigma n^{-1}} d_{jk}. \tag{14}$$

The suggested length of the block for blocks of identical length is $\approx (\log n)^{1+\delta}$ with $\delta \geq 0$. It seems therefore natural, at least in the case of a regular design, to consider an optimization problem similar to the one suggested in expression (13) by minimizing instead the following functional:

$$\|W^T\mathbf{y} - \boldsymbol{\beta}\|_n^2 + \lambda^2 \sum_{j=0}^{J}\sum_m \sqrt{\Gamma_{jm}}\sqrt{\sum_{s\in T_{jm}}\beta_{j,s}^2}. \tag{15}$$

Taking derivatives with respect to β_{jk} we obtain:

$$\beta_{jk} = d_{jk}\frac{\sqrt{\sum_{s\in T_{jm}}\beta_{j,s}^2}}{\sqrt{\sum_{s\in T_{jm}}\beta_{j,s}^2} + \lambda^2\sqrt{\Gamma_{jm}}/2} \tag{16}$$

mimicking eq. (14) when $\frac{\lambda^2\sqrt{\Gamma_{jm}}}{\sqrt{L_{jm}}} \simeq n^{-1/2}$.

In our case we do not have an equidistant design and our coefficients are not wavelet coefficients. However, the above remarks suggest to define an optimization problem which essentially mimics expression (15) under our general RKHS setup. Hence, define

$$K_{jm}(s,t) = \sum_{k\in T_{jm}} \psi_{j,k}(s)\psi_{j,k}(t),$$

$$K_{jm}^{\Gamma}(s,t) = \Gamma_{jm}^{-1}K_{jm}(s,t),$$

$$K^{\Gamma}(s,t) = \sum_{j\geq 0}\sum_m K_{jm}^{\Gamma}(s,t),$$

and

$$K_J^{\Gamma}(s,t) = \sum_{j=0}^{J}\sum_m K_{jm}^{\Gamma}(s,t),$$

and let $\mathcal{H}_{j,m}$, $\mathcal{H}_{\Gamma,j,m}$, $\mathcal{H}_{\Gamma}$, and $\mathcal{H}_{\Gamma,J}$ the corresponding reproducing kernel Hilbert spaces. Note that K^{Γ} and K_J^{Γ} are the same as before if $\Gamma_{jm} \equiv \Gamma_j$ for all m at the resolution level j. All results of Section 2.2 remain valid with these new kernels and we finally propose finding $f \in \mathcal{H}_{J,\Gamma}$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n}\{y_i - f(x_i)\}^2 + \lambda^2 R_J(f), \text{ with } R_J(f) = \sum_{j=0}^{J}\sum_m \|P_{jm}f\|_{\mathcal{H}_{\Gamma,j,m}}. \tag{17}$$

The penalty term in (17) is again a sum of wavelet-based RKHS norms and a pseudo-norm. Inequality (12) is still valid just with a different constant than J in the right hand side.

Remark 3.1 *Under the above set up, the following expansion of the unknown regression function f holds:*

$$f(t) = \sum_{j=0}^{\infty} \sum_{i=1}^{n} c_i \sum_{m} \Gamma_{jm}^{-1} K_{jm}(t_i, t), \tag{18}$$

and we obtain

$$\|P_{jm}^{\Gamma} f\|_{\mathcal{H}_{\Gamma,j,m}}^2 = \Gamma_{jm} \|P_{jm} f\|_{\mathcal{H}_{j,m}}^2.$$

Hence, in the case of equispaced design, the penalty R_J in (17) leads to solution (16). The difference between $\|P_j^{\Gamma} f\|_{\mathcal{H}_{\Gamma,j}}$ and $\|P_{jm}^{\Gamma} f\|_{\mathcal{H}_{\Gamma,j,m}}$ is that the former involves all wavelet coefficients, while the latter involves only coefficients of the block $T_{j,m}$, namely,

$$\beta_{j,k} = d_{j,k} \left(1 - \lambda^2 \sqrt{\Gamma_j} \Big/ (2\|d_{j,m,\cdot}\|)\right)_+$$

where $\|d_{j,m,\cdot}\| = \sqrt{\sum_{s \in T_{jm}} d_{j,s}^2}$.

3.1 Existence of a solution and asymptotic properties

The existence of the estimate obtained by the penalization procedure (17) is guaranteed by the following theorem.

Theorem 3.1 *Let $\mathcal{H}_{J,\Gamma}$ be the wavelet-based RKHS of functions over $[0, 1)$ defined at the end of the previous subsection and consider its decomposition*

$$\mathcal{H}_{J,\Gamma} = V_0 \oplus \oplus_{j=0}^{J} \sum_{m} \Gamma_{jm}^{-1} \mathcal{W}_{j,m} := V_0 \oplus \oplus_{j=0}^{J} \mathcal{W}_{j,\Gamma}.$$

Then there exists a minimizer of (17) in $\mathcal{H}_{J,\Gamma}$.

The uniqueness of the solution to (17) is not addressed in the above theorem but should follow under mild conditions on the design. We will not pursue this question here, and the developments that follow do not depend on the uniqueness of the estimate.

In the literature on nonparametric estimation by wavelet methods, one often considers the class of Besov spaces $B_{p,q}^s([0, 1))$ a short description of which we have given in the

Appendix. Recall that these spaces refer to functions on $[0,1)$ with "smoothness" s. Using regular enough wavelets, the wavelet coefficients $\beta_{j,k}$ of a function f that lies in a unit ball of $B^s_{p,q}([0,1))$ satisfy

$$\left(\sum_{j=0}^{J} 2^{j((2s+1)\frac{p}{2}-1)\frac{q}{p}} \left\{ \sum_{k=0}^{2^j-1} |\beta_{j,k}|^p \right\}^{\frac{q}{p}} \right)^{\frac{1}{q}} \leq 1, \tag{19}$$

which is equivalent, when $J=\infty$ to the Besov semi-norm. Assume that $s \geq \max(1/q - 1/2, 1/p + 1/2)$, $p,q \geq 1$. Denote $\rho = 2/(2s+1)$ and let $J < \infty$. Assume also that the unknown regression function is such that the sequence of its wavelets coefficients $\boldsymbol{\beta}_J = \{\beta_g, g \in \cup_{j=0}^{J} G_j\}$ satisfies (19). Take $\Gamma_j = 2^{\mu j}$. Then, in order $f \in \mathcal{H}_{\Gamma,J}$ condition (5) should hold which is ensured by $\mu > 1$. The penalty $R_J(f)$ is finite whenever $\mu < 2(s-1/p)$. The following theorem shows that the estimator defined above has a rate of convergence $n^{-(2-\rho)/4}$ if the tuning parameter λ is chosen appropriately and the noise process is Gaussian.

Theorem 3.2 *Consider the regression model $Y_i = f_0(x_i) + \epsilon_i$, $i = 1, \ldots, n$ where x_i's are given deterministic points in $[0,1)$, and the ϵ_i's are independent $N(0,\sigma^2)$ noise variables. Assume that $s \geq \max(1/q - 1/2, 1/p + 1/2)$, $p,q \geq 1$, and let $J = \log_2 n$. Assume that f_0 is such that the sequence of its wavelets coefficients satisfies (19). Take $\Gamma_{jm} = \Gamma_j = 2^{\mu j}$ with $1 < \mu < 2(s-1/p)$. Consider the estimator $\hat{\mathbf{f}}$ of the values of the unknown regression function at the design points as defined by (17). Then (i) if f_0 is not a constant, and $\lambda_n^{-1} = O_P(n^{(2-\rho)/4}) R_J^{(1-\rho)/2}(f_0)$, we have $\frac{1}{n}\|\mathbf{f} - \mathbf{f}_0\|_n = O_P(\lambda_n) R_J^{1/2}(f_0)$; (ii) if f_0 is constant, we have $\frac{1}{n}\|\hat{\mathbf{f}} - \mathbf{f}_0\|_n = O_P(\max\{(n\lambda_n)^{-2/3}, n^{-1/2}\})$.*

As one can see the resulting estimator attains the asymptotic minimax rate over the appropriate functional class without an extra $\log n$ factor that is usual for soft wavelet thresholding rules. As noted by Cai (2001) the extra $\log n$ factor is a drawback of separable penalized least-squares estimators and arises through the need to guard against "false positive" about the presence of true significant wavelet coefficients (corresponding to irregularities of the regression function f_0). As a result, standard soft thresholded estimators are often oversmoothed. The problem is unavoidable for separable estimators, since decisions about individual terms are based on a relatively low level of information. Therefore there are true benefits to consider more general penalties than those used in (10).

Note also that for the particular case of a Sobolev space $B^s_{2,2}([0,1))$ with index $s = 2$ Theorem 3.2 leads to the same asymptotic rates, under of course a different setup and with different constants, of the COSSO estimates in functional ANOVA models proved by Lin and Zhang (2003).

Finally, note that Theorem 3.2 is related to the optimal behavior of the estimator at the design points without any further assumptions on their distribution. If however we want to prove consistency of our penalized estimator in the integrated mean squared error it is natural to require some additional assumptions on the design points.

Introduce the matrix Ψ the i-th row of which contains values of $\psi_{j,k}$ at the point x_i. Assume that the wavelets $\psi_{j,k}$ are regular enough and that the function f_0 is periodic, i.e. $f_0(0) = f_0(1)$. Under such an assumption we can now state:

Theorem 3.3 *Consider the regression model $Y_i = f_0(x_i) + \epsilon_i$, $i = 1, \ldots, n$ where x_i's are given deterministic points in $[0,1)$ such that the lowest eigenvalue of matrix $\Psi^T\Psi$ is bounded away from zero by a constant independent of n, and the ϵ_i's are independent $N(0,\sigma^2)$ noise variables. Assume that f_0 is nonconstant and periodic and also that the assumptions of Theorem 3.2 hold. Under the additional assumption that the empirical wavelet coefficients of the penalized estimator satisfy (19), the penalized estimator $\hat{f}_n$ is weakly consistent in the integrated mean squared error with a rate of order $O_P(n^{-\frac{2s}{2s+1}})$.*

Regularity of the wavelets is needed to control the interpolation error as in Antoniadis (1996). The periodicity of f_0 is needed because we are using periodic wavelets in the interval. Note however that such assumption is not needed if we use more general wavelets on the interval (see Daubechies (1992)).

What follows are some useful lemmas for the practical implementation of our estimator. The next result shows that the solution to (17) is finite-dimensional and the estimate can be computed directly from (17) by linear programming techniques.

Lemma 3.1 *Let $\hat{f}_J = \hat{b} + \sum_{j=0}^{J} \hat{f}_j$ be a minimizer of (17), with $\hat{f}_J \in \mathcal{W}_{j,\Gamma}$. Then $\hat{f}_J \in span\{K_j^\Gamma(t_i, \cdot),\ i = 1, \ldots, n\}$, where K_j^Γ is the reproducing kernel of the space $\mathcal{W}_{j,\Gamma}$.*

However, following the suggestion of Antoniadis and Fan (2001) for solving penalized problems with an l_1 penalty, it is possible to give an equivalent formulation of (17) that is much easier to compute in practice. Consider the problem of finding $\boldsymbol{\theta} = \{\theta_{j,m}, j =$

$0, \ldots, J; m = 1, \ldots, M_j\}$ where M_j denotes the number of blocks at scale j and $f \in \mathcal{H}_{\Gamma,J}$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n}\{y_i - f(x_i)\}^2 + \lambda_0 \sum_{0\le j\le J}\sum_{m\in M_j} \theta_{jm}^{-1}\|P_{jm}^{\Gamma} f\|_{\Gamma,j,m}^2 + \nu \sum_{j=0}^{J}\sum_{m\in M_j}\theta_{jm}, \tag{20}$$

subject to $\theta_{jm} \ge 0,\ j = 0, \ldots, J; m = 1, \ldots, M_j$, where λ_0 is a constant that can be fixed to any positive value and $\nu = \nu_n$ is a smoothing parameter. Fix λ_0 at some value. Then

Lemma 3.2 *Set $\nu = \lambda^4/(4\lambda_0)$. (i) If $\hat{f}$ minimizes (17), set $\hat{\theta}_{jm} = \lambda_0^{1/2}\lambda^{-1/2}\|P_{jm}^{\Gamma}\hat{f}\|_{\Gamma,J,m}$, then the pair $(\hat{\boldsymbol{\theta}}, \hat{f})$ minimizes (20). (ii) On the other hand, if a pair $(\hat{\boldsymbol{\theta}}, \hat{f})$ minimizes (20), then $\hat{f}$ minimizes (17).*

The form of (20) is very similar to the one of (10) with multiple smoothing parameters, except that there is an additional penalty on the θ's. Notice that there is only one smoothing parameter ν in (20). The θ's are part of the estimate, rather than free smoothing parameters. The additional penalty on θ's in (20) makes it possible to have some θ's be zeros, giving rise to zero block detail components in the estimate, thus producing a sparse kernel estimate in the sense of Gunn and Kandola (2002).

3.2 Algorithm and penalty choice

In what follows, we shall use an iterative optimization algorithm. On each step of iteration, for some fixed values of $\boldsymbol{\theta}$ we shall minimize (20) with respect to f, and then for this choice of f we shall minimize (20) with respect to $\boldsymbol{\theta}$. For any fixed $\boldsymbol{\theta}$, the function f minimizing (20) is given by the representer Theorem which in the case of multiple smoothing parameters $\theta_{j,m}$ suggests f of the following form (see Wahba (1990), Section 10.1 and recall our discussion in remark 2.2)

$$f(x) = b + \sum_{i=1}^{n} c_i \sum_{j=0}^{J}\sum_{m\in M_j}\theta_{j,m}K_{jm}^{\Gamma}(t_i, x),$$

where $\mathbf{c} = (c_1, \ldots, c_n)^T \in \mathbb{R}^n$, $b \in \mathbb{R}$ and K_{jm}^{Γ} is the reproducing kernel of $\mathcal{W}_{\Gamma,j,m}$. Assume equal block sizes and with some abuse of notations, let K_{jm}^{Γ} also stand for the $n \times n$ matrix $\{K_{jm}^{\Gamma}(t_i, t_\ell)\}$, $i = 1, \ldots, n$, $\ell = 1, \ldots, n$. Let $K_{\boldsymbol{\theta}}^{\Gamma}$ also stand for the matrix $\sum_{j=0}^{J}\sum_m \theta_{jm}K_{jm}^{\Gamma}$, and let $\mathbf{1}_r$ be the column vector consisting of r ones. Then we can

write $\mathbf{f} = K^{\Gamma}_{\boldsymbol{\theta}}\mathbf{c} + b\mathbf{1}_n$, and (20) can be expressed as

$$\frac{1}{n}\left\|\mathbf{y} - \sum_{j=0}^{J}\sum_{m\in M_j}\theta_{jm}K^{\Gamma}_{jm}\mathbf{c} - b\mathbf{1}_n\right\|_n^2 + \lambda_0\mathbf{c}^T K^{\Gamma}_{\boldsymbol{\theta}}\mathbf{c} + \nu\sum_{j=0}^{J}\sum_{m\in M_j}\theta_{j,m}, \tag{21}$$

where $\theta_{j,m} \geq 0$, $j = 0, \ldots, J; m = 1, \ldots, M_j$.

The form (21) turns out to be similar to the sparse kernel selection approach in Gunn and Kandola (2002). They used a different reproducing kernel and put penalty on all components including the constant b. They motivated their method by noting that the form of the penalty on the θ's in (21) tends to give sparse solutions for θ's, and gave empirical evidence to support the insight. Our method is motivated from a different formulation which relates to the standard wavelet thresholding estimation procedure.

If θ's are fixed, then (21) can be written as

$$\min_{\mathbf{c},b} \|\mathbf{y} - K^{\Gamma}_{\boldsymbol{\theta}}\mathbf{c} - b\mathbf{1}_n\|_n^2 + n\lambda_0\mathbf{c}^T K^{\Gamma}_{\boldsymbol{\theta}}\mathbf{c}, \tag{22}$$

which is a quadratic minimization problem that can be solved by linear methods. On the other hand, if $\mathbf{c}$ and b were fixed, denote $\mathbf{d}_{jm} = K^{\Gamma}_{jm}\mathbf{c}$, and let D be the $n \times (\sum_j M_j)$ matrix with the (j,m)th column being $\mathbf{d}_{jm}$. Simple calculation shows that the vector $\boldsymbol{\theta}$ that minimizes (21) is the solution to

$$\min_{\boldsymbol{\theta}} \|\mathbf{z} - D\boldsymbol{\theta}\|_n^2 + n\nu\sum_{j=0}^{J}\sum_{m\in M_j}\theta_{jm}, \tag{23}$$

where $\mathbf{z} = \mathbf{y} - (1/2)n\lambda_0\mathbf{c} - b\mathbf{1}_n$. Therefore a reasonable scheme would be to iterate between (22) and (23). In each iteration (21) is decreased. Notice that (23) is equivalent to

$$\min_{\boldsymbol{\theta}} \|\mathbf{z} - D\boldsymbol{\theta}\|_n^2 \text{ subject to } \theta_{j,m} \geq 0,\ \sum_{j=0}^{J}\sum_{m\in M_j}\theta_{jm} \leq M, \tag{24}$$

for some $M \geq 0$. If the algorithm that iterates between (22) and (23) converges, then the solution is also a fixed point for the algorithm that iterates between (22) and (24) for a fixed M. We prefer to iterate between (22) and (24) for computational considerations. Analysis of such a type algorithm and its convergence properties can be found in Karlovitz (1970) and has been recently used by Donoho *et al.* (2004) for stable recovery of sparse overcomplete representations in presence of noise.

It can take a large number of iterations for the algorithm to converge. In applications, though, we do not really need an exact solution. By starting from a simpler estimate such as the one obtained by penalized least squares with quadratic penalties on the coefficients in a spirit close to that of Antoniadis (1996) or Amato and Vuza (1997) and applying a limited number of iterations of our algorithm, we get what we view as an iterative improvement on the wavelet thresholded estimator. This motivates us to consider the following one step update procedure:

1. Initialization: Fix $\theta_{j,m} = 1$, $j = 0, \ldots, J; m = 1, \ldots, M_j$.

2. Solve for $\mathbf{c}$ and b with (22).

3. For the $\mathbf{c}$ and b obtained in step 2, solve for $\boldsymbol{\theta}$ with the (24).

4. With the new $\boldsymbol{\theta}$, solve for $\mathbf{c}$ and b with (22).

This one step update procedure has the flavor of the one step maximum likelihood procedure, in which one step Newton-Raphson algorithm is applied to a good initial estimator and which is as efficient as the fully iterated maximum likelihood. A discussion of one step procedure and fully iterated procedure can be found in Antoniadis and Fan (2001).

To end this section let us stress that the performance of the penalized least-squares estimator depends on the regularization parameter λ, the chosen resolution J and the "complexity" parameter M. The choice of these parameters obviously involves an arbitrary decision. In the present context we prefer regarding the free choice of the resoluiton index J as an advantage of the model rather than a problem to be solved. By varying the resolution level up to a maximum value of $\log_2 n$, features of the data arising on different scales can be explored. As for the smoothing parameter λ a convenient way to get a data based estimate of it is by using generalized cross validation as proposed by Craven and Wahba (1979) for choosing smoothing parameters in smoothing splines algorithms and which is widely used for the selection of smoothing parameters in penalized least squares regression. However the minimization problem stated in (11) is not quadratic and it is not obvious how such a method may be applied. Tibshirani (1996) proposed a GCV-type criterion for choosing the tuning parameter for the LASSO through a ridge estimate approximation. This approximation is particularly easy to understand in light of the form (10): fix the θ_{jm}'s at their estimated values $\hat{\theta}_{jm}$'s, and calculate GCV for the

corresponding ridge regression. Of course, this approximation ignores some variability in the estimation process but the simulation study in Tibshirani (1996) suggests that it is a useful approximation. This motivates why we have used the GCV score or CV with five or ten fold cross-validation for the penalized least squares in (10) when θ's are fixed at the solution. As for the choice of M, which depends on the size of the blocks, we used the golden section search for minimizing GCV or CV with respect to M.

3.3 Bayesian interpretation

It is well known (see e.g. Vidakovic (1999)) that optimization problem usually allows Bayesian interpretation with an appropriate choice of priors. It follows from the fact that errors $\epsilon_i,\ i = 1, \cdots, n$, are iid normal and from representation (18) of f that the pdf of $\mathbf{y}$ given $\mathbf{c}$ and $\mathbf{\Gamma}$ is

$$p(\mathbf{y}|\mathbf{c}, \mathbf{\Gamma}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2} \left(\mathbf{y} - \sum_{j,m} \frac{K_{j,m}}{\Gamma_{j,m}} \mathbf{c} \right)^T \left(\mathbf{y} - \sum_{j,m} \frac{K_{j,m}}{\Gamma_{j,m}} \mathbf{c} \right) \right\}. \tag{25}$$

Impose the following proper prior on the parameter vector $\mathbf{c}$

$$p(\mathbf{c}|\mathbf{\Gamma}, \varrho) = C_1(\mathbf{\Gamma}) \varrho^n \exp\left\{ -\varrho \sum_{j,m} \sqrt{\frac{\mathbf{c}^T K_{j,m} \mathbf{c}}{\Gamma_{j,m}}} \right\}. \tag{26}$$

where $C_1(\mathbf{\Gamma})$ is the coefficient depending on $\mathbf{\Gamma}$ only such that (26) integrates to one. Since the posterior pdf of $\mathbf{c}$ given $\mathbf{y}$ is proportional to the joint pdf of $\mathbf{c}$ and $\mathbf{y}$, the Bayesian estimator of $\mathbf{c}$ based on the posterior mode is the solution of the following minimization problem

$$\frac{1}{n} \left(\mathbf{y} - \sum_{j,m} \frac{K_{j,m}}{\Gamma_{j,m}} \mathbf{c} \right)^T \left(\mathbf{y} - \sum_{j,m} \frac{K_{j,m}}{\Gamma_{j,m}} \mathbf{c} \right) + \frac{2\sigma^2 \varrho}{n} \sum_{j,m} \sqrt{\frac{\mathbf{c}^T K_{j,m} \mathbf{c}}{\Gamma_{j,m}}}$$

which is equivalent to (17) with $\lambda^2 = 2\sigma^2 \varrho / n$.

As it was mentioned before, problem (17) is hard to treat computationally. For this reason, model (25)–(26) can be replaced by the hierarchical Bayesian model by assuming that the pdf of $\mathbf{c}$ depends on a random vector $\boldsymbol{\alpha}$, namely,

$$p(\mathbf{c}|\mathbf{\Gamma}, \boldsymbol{\alpha}, \beta) = \sqrt{\det(A)} (\beta/\pi)^{n/2} \exp\left\{ -\beta \sum_{j,m} \frac{\mathbf{c}^T K_{j,m} \mathbf{c}}{\Gamma_{j,m} \alpha_{j,m}^2} \right\} \tag{27}$$

where

$$A = \sum_{j,m} \frac{\mathbf{c}^T K_{j,m} \mathbf{c}}{\Gamma_{j,m} \alpha_{j,m}^2}.$$

Now, let the pdf of vector $\boldsymbol{\alpha}$ be of the form

$$p(\boldsymbol{\alpha}|\boldsymbol{\Gamma}, \gamma) = C_2(\boldsymbol{\Gamma}, \gamma)(\det(A))^{-1/2} \exp\left\{-\gamma \sum_{j,m} \alpha_{j,m}^2\right\}. \tag{28}$$

The joint pdf of $\mathbf{c}$ and $\boldsymbol{\alpha}$ is obtained by multiplying (27) and (28). Integrating $\boldsymbol{\alpha}$ out of the joint pdf with the help of formula 3.325 of Gradshteyn and Ryzhik (1980) we ensure that the expression for the marginal pdf of $\mathbf{c}$ coincides with (26).

Note that the joint pdf of $\mathbf{c}$ and $\boldsymbol{\alpha}$ given $\mathbf{y}$ is proportional to the product of (25), (27) and (28). The Bayesian estimator based on the posterior mode is again obtained by minimizing the expression

$$\frac{1}{n}\left(\mathbf{y} - \sum_{j,m} \frac{K_{j,m}}{\Gamma_{j,m}}\mathbf{c}\right)^T \left(\mathbf{y} - \sum_{j,m} \frac{K_{j,m}}{\Gamma_{j,m}}\mathbf{c}\right) + \frac{2\sigma^2\beta}{n} \sum_{j,m} \frac{\mathbf{c}^T K_{j,m} \mathbf{c}}{\alpha_{j,m}^2 \Gamma_{j,m}} + \frac{2\sigma^2\gamma}{n} \sum_{j,m} \alpha_{j,m}^2$$

which coincides with (20) when $\lambda_0 = 2\sigma^2\beta/n$, $\theta_{j,m} = \alpha_{j,m}^2$ and $\nu = 2\sigma^2\gamma/n$.

4 Simulations and a real example

In this section we study the empirical performance of our estimator in terms of estimation accuracy and model selection. Our estimate is compared with the Kovac and Silverman (2000) wavelet term-by-term thresholding procedure (SK for short) for denoising functions sampled at nonequispaced design points. Recall that the Kovac and Silverman (2000) procedure relies upon a linear interpolation transformation R to the observed data vector y that maps it to a new vector of size 2^J ($2^{J-1} < n \leq 2^J$), corresponding to a new design with equispaced points. After the transformation, the new vector is multivariate normal with mean Rf and covariance that is assumed to have a finite bandwidth so that the computational complexity of their algorithm is of order n. For the SK procedure a term-by-term estimator with soft-thresholding and Stein's unbiased risk estimation policy (Sure Shrink) was considered as it is implemented in the R-package `Wavethresh3` (Nason (1998)). It is detailed in Kovac and Silverman (2000). Additionally, for both estimators (ours and SK's) the lowest level of detail coefficients used was set at 3 for all simulations. The wavelets used were Daubechies extremal phase wavelets with 5 vanishing moments.

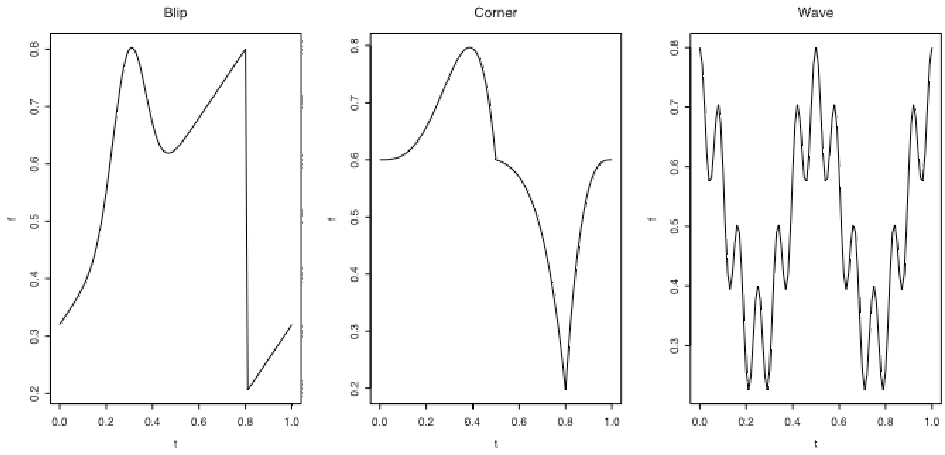


Figure 4.1: The three test functions used in the simulations.

Since we deal with compactly supported wavelets the numerical algorithm for the kernel computations is based on Daubechies cascade procedures (Daubechies (1992)). More precisely, the cascade algorithm computes the values of wavelets and scaling functions at dyadic points. In order to evaluate the entries of the kernel matrices K_j^Γ we have computed the values of the wavelets on a fine grid of dyadic points and stored them in a table. Values of scaling functions and wavelet functions at arbitrary points, necessary for the evaluation of the kernels K_j^Γ, were then computed by interpolation or by considering the value at the closest point on the tabulated grid.

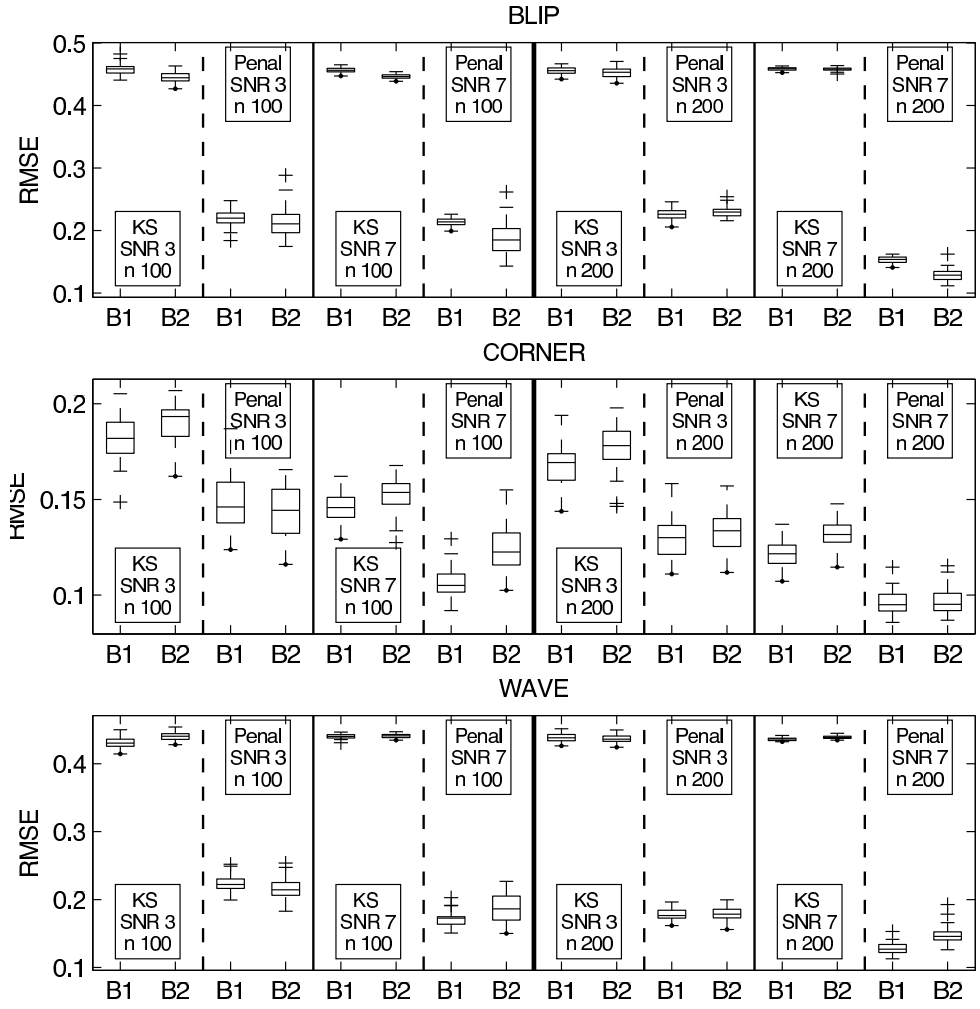


Figure 4.2: Graphical display (boxplots) of the results given in Table 1. B1: Beta(0.9,1.1) design; B2: Beta(2,2) design.

4.1 Simulation comparison

The estimators were applied to simulated data sets of varying size, sample point placement, signal to noise ratio and test function. The values of n used were $n = 100$ and 200. Three test functions were used that represent a variety of function characteristics. These are well known functions in the wavelet literature (see Antoniadis et al. (2001)) and are displayed in Figure 4.1. The `Blip` function presents a discontinuity and does not really satisfy our assumptions. The `Corner` and `Wave` functions are typical functions for which our algorithm is well suited. The test functions have been scaled so they all have a standard deviation of 1. Nonequispaced placement of the sample points was done in

BLIP

Grid	SNR	n	SK	Penalization	n	SK	Penalization
Beta(0.9,1.1)	3	100	0.210 (0.008)	0.048 (0.005)	200	0.208 (0.005)	0.051 (0.004)
Beta(0.9,1.1)	7	100	0.208 (0.004)	0.046 (0.002)	200	0.210 (0.002)	0.023 (0.002)
Beta(2,2)	3	100	0.197 (0.008)	0.05 (0.01)	200	0.205 (0.007)	0.053 (0.004)
Beta(2,2)	7	100	0.199 (0.003)	0.036 (0.009)	200	0.210 (0.002)	0.017 (0.003)
CORNER							
Beta(0.9,1.1)	3	100	0.033 (0.004)	0.023 (0.005)	200	0.028 (0.003)	0.017 (0.003)
Beta(0.9,1.1)	7	100	0.021 (0.003)	0.011 (0.002)	200	0.015 (0.002)	0.009 (0.001)
Beta(2,2)	3	100	0.036 (0.004)	0.021 (0.004)	200	0.032 (0.004)	0.018 (0.003)
Beta(2,2)	7	100	0.023 (0.003)	0.016 (0.003)	200	0.017 (0.002)	0.009 (0.001)
WAVE							
Beta(0.9,1.1)	3	100	0.186 (0.007)	0.050 (0.005)	200	0.192 (0.005)	0.032 (0.003)
Beta(0.9,1.1)	7	100	0.193 (0.003)	0.029 (0.004)	200	0.190 (0.002)	0.016 (0.002)
Beta(2,2)	3	100	0.194 (0.005)	0.047 (0.006)	200	0.191 (0.005)	0.032 (0.003)
Beta(2,2)	7	100	0.194 (0.003)	0.035 (0.007)	200	0.193 (0.002)	0.022 (0.004)

Table 1: Root mean square error of the SK and penalization methods for the synthetic data sets of FIg. 4.1 with 50 repetitions. Root mean square error is shown together with its estimated standard deviation for grids Beta(0.9,1.1) and Beta(2,2), Signal-to-Noise ratio (SNR) 3 and 7, sample size n 100 and 200.

a variety of methods. They include placing the points on the interval [0,1] uniformly, or distributed as a Beta(9/10,11/10), Beta(1/2,1/2), or Beta(2,2) random variable. Next, independent, i.i.d. standard Gaussian noise was added to the test signals to give signal to noise ratios of 3 and 7 (low and high).

Some of the results of these simulations are presented with Table 1 and boxplots are depicted in Figure 4.2. There, the Beta(9/10, 11/10) distribution or the Beta(2,2) where used to place the sample points and a signal to noise ratio of 3 or 7 is used. 50 data sets were generated for each sample size range $n = 100$ or 200. As can be seen in Table 1 (data assumed to be i.i.d.), our block penalized procedure outperforms the SK estimator in all cases. We noticed that SK SureShrink based procedure includes sometimes visually unpleasant artifacts. Moreover, the SK procedure is designed for relatively large sample sizes, so that $n = 100$ is almost the lowest possible size it can handle. Consequently, for

the `Blip` and the `Wave` test functions, the precision of the SK method does not improve even when SNR grows significantly. The latter is due to the bias introduced in SK by pre-processing of the data. The most interesting comparisons are for the `Wave` test example where the benefits of our block penalization are the most impressive.

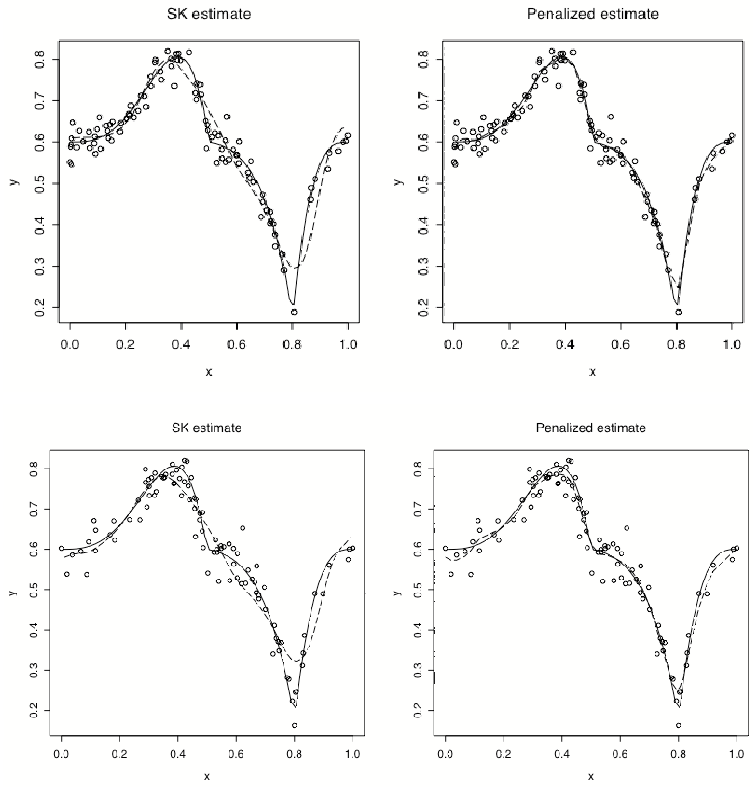


Figure 4.3: Results of one realization of simulations for the `Corner` test function. Solid line represents the true function; circles the input data and dashed line the reconstructed functions (left panel the SK estimate; right panel the penalized estimate). Top figure: Beta(9/10,11/10) design; bottom: Beta(2,2) design.

Figure 4.3 shows typical reconstructions for the `Corner` test function using these two estimators. We have used a signal to noise ratio of 5 and the Beta(9/10,11/10) and Beta(2,2) distributions for sample point placement. In Figure 4.3, the number of data points is 100. Independent, identical mean zero Gaussian noise has been added to the test signal in the same fashion as for the simulations reported earlier.

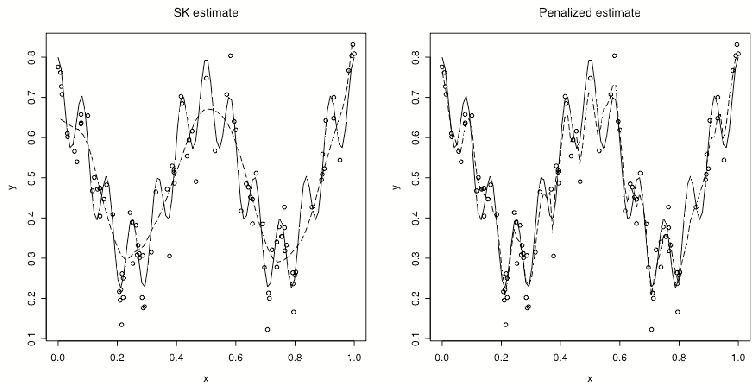


Figure 4.4: Results of one realization of simulations for the `Wave` test function. Solid line represents the true function; circles the input data and dashed line the reconstructed functions (left panel the SK estimate; right panel the penalized estimate). 100 points with Beta(2,2) distribution and a SNR of 3 were used for simulating the data.

More surprising are the results obtained for the `Wave` test function displayed in the panels of Figure 4.4. Here again sample points were drawn according to a $Beta(2, 2)$ distribution and a Gaussian noise was added to the true function with a signal-to-noise ratio of 3. As it can be seen the SK estimate produces a smooth fit similar to the one expected for spline smoothing while our penalized estimate tracks very well the oscillations of the `Wave` function.

4.2 Two real data examples

In this subsection, we describe a comparison of the estimators on two real data sets, the ethanol and the motorcycle data set. Figure 4.5 contains the ethanol data set collected by Brinkman (1981), and it is easy to see that the data points are not equispaced. The set contains $n = 88$ observations each consisting of three measurements: the concentration of NO and NO2 emissions from a single-cylinder engine, the engines equivalence ratio, and the engines compression ratio. In the example below we shall analyze only the concentration of NO and NO2 and the equivalence ratio, and we scale the data to the interval $[0, 1]$.

The data are analyzed by two different wavelet procedures in both of which the data are assumed to follow the model $y_i = f(x_i) + \epsilon_i$. Wavelets with five vanishing moments are employed with four levels of detail coefficients starting at level 3 subjected

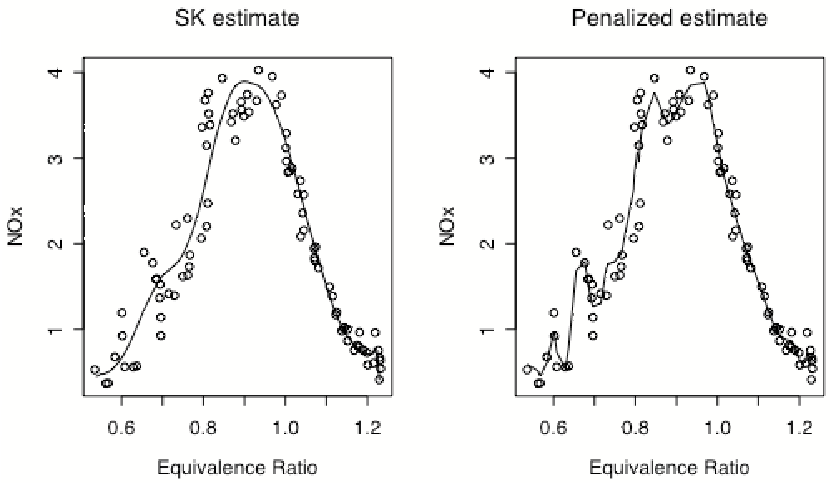


Figure 4.5: Measurement of Exhaust from Burning Ethanol.

to thresholding. The first method is based on the SK term-by-term method VisuShrink. The noise is assumed to be i.i.d. normal and the estimator of its level is based only on the wavelet coefficients of the highest resolution level which are normalized as in Kovac and Silverman (2000) to take into account the unequal variances of the coefficients after the transformation by R. The second method is the penalized block estimation suggested above with blocks of length 3 and hyperparameters determined by five-fold CV. The results of the application of the first and the second methods are displayed in the left and right part of Figure 4.5, respectively. As one can see, the penalized estimates are not as smooth as the SK estimator since it detects some bumps that are not present in the SK estimate, suggesting that our penalized estimator is consistent with an assumption of homoscedastic errors while the apparent smoothness of the SK scatterplot could be only explained by the assumption of nonconstant variance of the measurement process.

The second example deals with the so-called motorcycle data (see Silverman (1985) for a complete description of this data set). The experiments are designed to test crash helmets, and the data consists of the time in seconds as a design variable and head acceleration in g as a response. Again, in this example, the design is not equispaced and the errors are well known to be heteroscedastic. We apply the same two procedures to this data as for the ethanol data set. The results are plotted in Figure 4.6. As one can see in this example, the penalized estimator is much less affected by the sparseness of the design points around the maximum, suggesting that the estimator is more robust

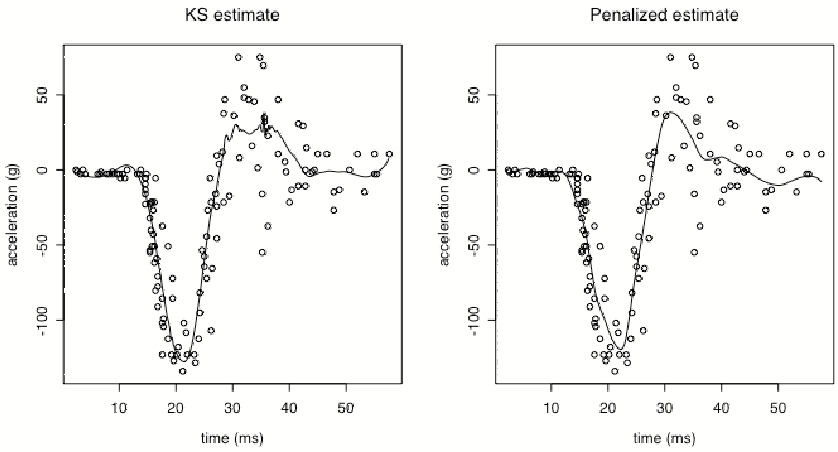


Figure 4.6: Head acceleration versus time after simulated impact (motrocycle data).

to heteroscedasticity than the SK procedure since it detects some bumps that are not present in the SK estimate.

5 Comments and discussion

In the present paper we study regression problems with nonequispaced design points. The method suggested in the paper requires neither pre-processing of the data by interpolation or similar technique, nor the knowledge of the distribution of the design points. For this reason, the method works really well even in the case when the distribution of the design points deviates far from the uniform. When the estimation error is calculated at the design points only, the method achieves optimal convergence rates in Besov spaces no matter how irregular the design is (see Theorem 3.2). In order to obtain asymptotic optimality in the L^2 metric, an extra assumption on the design points should be imposed, namely, the density of the design points should be bounded away from zero (see Theorem 3.3). The estimator demonstrates excellent computational properties. Extensive simulations show that in terms of the square root of the mean integrated squared error it is several times superior to the estimator suggested by Kovac and Silverman (2000).

The procedure suggested above involves parameters λ, J and M, the optimal selection of which depend on the Besov ball $B^s_{p,q}[0,1]$. An adaptive choice of these parameters is a topic of future investigation. One of the possibilities in this direction is extension of

Bayesian techniques. As we have shown in Section 3.3, minimizing the penalized error is equivalent to constructing Bayesian estimator based on posterior mode and special selection of priors. Furthermore, Bayesian methods can be replaced by empirical Bayesian procedures where unknown parameters are elicitated by maximizing empirical likelihood. It will however be necessary to prove that the estimators based on this adaptive choice of λ and J still attain optimal convergence rates. Another possible extension of the method above is introduction of various other penalties corresponding to other choices of prior distributions as well as construction of Bayesian estimators based on posterior means and medians.

Acknowledgements. This research was supported by funds from the IAP research network nr P5/24 of the Belgian Government (Federal Office for Scientific, Technical and Cultural Affairs), the EC-HPRN-CT-2002-00286 Breaking Complexity network, the National Science Foundation (NSF) of United States, grant DMS-0004173, the CNR-CNRS project "Advanced statistical methods for data analysis and applications" and the Italian Space Agency which are gratefully acknowledged. M. Pensky thanks the LMC-IMAG department of the Université Joseph Fourier, Grenoble and both, A. Antoniadis and M. Pensky, thank the Istituto per le Applicazioni del Calcolo CNR Naples, for their hospitality and research facilities provided. The authors would also like to thank A. Iouditski, R. Hildebrand (LMC) and E. Loubes (CNRS, Toulouse) for helpful discussions.

Appendix

Besov Spaces on the Unit Interval

The (inhomogeneous) Besov spaces on the unit interval, $B^s_{\rho_1,\rho_2}([0,1])$, consist of functions that have a specific degree of smoothness in their derivatives. The parameter ρ_1 can be viewed as a degree of function's inhomogeneity while s is a measure of its smoothness. Roughly speaking, the (not necessarily integer) parameter s indicates the number of function's derivatives, where their existence is required in an L^{ρ_1}-sense; the additional parameter ρ_2 is secondary in its role, allowing for additional fine tuning of the definition of the space.

More specifically, let the rth difference of a function $f(t)$ be

$$\Delta_h^{(r)} f(t) = \sum_{k=0}^{r} \binom{r}{k} (-1)^k f(t + kh),$$

and let the rth modulus of smoothness of $f(t) \in L^{\rho_1}[0,1]$ be

$$\nu_{r,\rho_1}(f;t) = \sup_{h \le t}(||\Delta_h^{(r)} f||_{L^{\rho_1}[0,1-rh]}).$$

Then the Besov seminorm of index (s, ρ_1, ρ_2) is defined for $r > s$, where $1 \le \rho_1, \rho_2 \le \infty$, by

$$|f|_{B^s_{\rho_1,\rho_2}} = \left[\int_0^1 \left\{\frac{\nu_{r,\rho_1}(f;h)}{h^s}\right\}^{\rho_2} \frac{dh}{h}\right]^{1/\rho_2}, \qquad \text{if} \qquad 1 \le \rho_2 < \infty,$$

and by

$$|f|_{B^s_{\rho_1,\infty}} = \sup_{0<h<1} \left\{\frac{\nu_{r,\rho_1}(f;h)}{h^s}\right\}.$$

The Besov norm is then defined as

$$||f||_{B^s_{\rho_1,\rho_2}} = ||f||_{L^{\rho_1}} + |f|_{B^s_{\rho_1,\rho_2}}$$

and the Besov space on $[0,1]$, $B^s_{\rho_1,\rho_2}([0,1])$, is the class of functions $f : [0,1] \to \mathbb{R}$ satisfying $f(t) \in L^{\rho_1}[0,1]$ and $|f|_{B^s_{\rho_1,\rho_2}} < \infty$, i.e. satisfying $||f||_{B^s_{\rho_1,\rho_2}} < \infty$. The Besov classes include, in particular, the well-known Hilbert-Sobolev ($H_2^s[0,1]$, $s = 1, 2, \ldots$) and Hölder ($C^s[0,1]$, $s > 0$) spaces of smooth functions ($B^s_{2,2}([0,1])$ and $B^s_{\infty,\infty}([0,1])$ respectively), but in addition less-traditional spaces, like the space of bounded-variation, sandwiched between $B^1_{1,1}[0,1]$ and $B^1_{1,\infty}[0,1]$. The latter functions are of statistical interest because they allow for better models of spatial inhomogeneity (see, for example, Meyer (1992); Donoho *et al.* (1995)).

The Besov norm for the function f is related to a sequence space norm on the wavelet coefficients of the function. As noted in Section 2.1, confining attention to the resolution and spatial indices $j \ge j_0$ and $k = 0, 1, \ldots, 2^j - 1$ respectively, and denoting by $s' = s + 1/2 - 1/\rho_1$, the sequence space norm is given by

$$\begin{aligned} ||w||_{b^s_{\rho_1,\rho_2}} &= ||u_{j_0}||_{\rho_1} + \left\{\sum_{j=j_0}^{\infty} 2^{js'\rho_2} ||w_j||_{\rho_1}^{\rho_2}\right\}^{1/\rho_2}, \quad \text{if} \quad 1 \le \rho_2 < \infty, \\ ||w||_{b^s_{\rho_1,\infty}} &= ||u_{j_0}||_{\rho_1} + \sup_{j \ge j_0}\left\{2^{js'} ||w_j||_{\rho_1}\right\}, \end{aligned}$$

where

$$||u_{j_0}||_{\rho_1}^{\rho_1} = \sum_{k=0}^{2^{j_0}-1} |u_{j_0k}|^{\rho_1} \quad \text{and} \quad ||w_j||_{\rho_1}^{\rho_1} = \sum_{k=0}^{2^j-1} |w_{jk}|^{\rho_1}.$$

If the mother wavelet ψ is of regularity $r > 0$, it can be shown that the corresponding orthonormal periodic wavelet basis defined in Section 2.1 is an unconditional basis for the Besov spaces $B^s_{\rho_1,\rho_2}([0,1])$ for $0 < s < r$, $1 \leq \rho_1, \rho_2 \leq \infty$. In other words, we have

$$K_1||f||_{B^s_{\rho_1,\rho_2}} \leq ||w||_{b^s_{\rho_1,\rho_2}} \leq K_2||f||_{B^s_{\rho_1,\rho_2}},$$

where K_1 and K_2 are constants, not depending on f. Therefore the Besov norm of the function f is equivalent to the corresponding sequence space norm defined above; this allows one to characterize Besov spaces in terms of wavelet coefficients (see, for example, Meyer (1992); Donoho *et al.* (1995)). For a more detailed study on (inhomogeneous) Besov spaces we refer to, for example, DeVore and Popov (1988), Triebel (1983) and Meyer (1992).

Entropy

The rate of convergence of the estimator in Theorem 3.2 is derived from the entropy of sets in Besov balls. We will not go into many details here, but mainly recall the basic definitions and properties of entropy of such sets that will allow us to get the appropriate rates. A good reference about entropy and nonparametric estimation is the monograph of van de Geer (2000).

Let T be a subset of a metric space. For $\delta > 0$, the δ-covering number $N(\delta, T)$ is the minimal number of balls with radius $\delta > 0$ that is necessary to cover T. The δ-entropy of T is then defined by $H(\delta, T) = \log N(\delta, T)$.

In the situation we are looking, we essentially need entropies of subsets of $\mathbb{R}^n$ endowed with the normalized Euclidian norm $\|\cdot\|_n$. Let $0 < \rho < 2$. If $\mathcal{A}_n = \{\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^T; \sum_{j=1}^n |\alpha_j|^\rho \leq 1\}$ by Lemma 4 of Loubes and van de Geer (2002) there exists a constant A, depending only on ρ, such that

$$H(\delta, \mathcal{A}_n) \leq A\delta^{-\frac{2\rho}{2-\rho}}(\log n + \log\frac{1}{\delta}).$$

Such a result yields already a bound of the entropy of Besov balls when the α_j's denote wavelet coefficients. However, in Besov spaces, coefficients at higher levels tend to be smaller, i.e. there is more structure than can be described by the "roughness" parameter

ρ. As a result it turns out that Besov spaces have entropies without logarithmic factors. More precisely, let $\mathcal{B}^s_{p,q}$ be the set of coefficients $\{\alpha_{j,k}\}$ that satisfy

$$\left(\sum_{j=0}^{J} 2^{j((2s+1)\frac{p}{2}-1)\frac{q}{p}} \left\{ \sum_{k=0}^{2^j-1} |\alpha_{j,k}|^p \right\}^{\frac{q}{p}} \right)^{\frac{1}{q}} \leq 1, \tag{29}$$

where the $\alpha_{j,k}$'s are wavelet coefficients of the appropriate wavelet basis. Considering $\mathcal{B}^s_{p,q}$ as a subset of the Euclidian space $\mathbb{R}^{2^J-2}$, with Euclidian norm $\|\cdot\|$, an entropy bound without logarithmic factors can be found in Birman and Solomjak (1967) for the case of Sobolev spaces $\mathcal{B}^s_{2,2}$, and in Birgé and Massart (2000) and Kerkacharian and Picard (2003) for general Besov spaces. It is shown there that for $p \geq 1$ and $\rho = 2/(2s+1) < p$ the δ-entropy for the L_∞ norm of a Besov ball with radius 1 in $B^s_{p,q}([0,1))$ is of the order $\delta^{-\frac{1}{s}}$, with lower and upper constant bounds that depend only on p, q and s, provided that $s > 1/p - 1/q$, i.e.

$$a\delta^{-\frac{1}{s}} \leq H(\delta, \mathcal{B}^s_{p,q}) \leq A\delta^{-\frac{1}{s}}, \quad \delta > 0. \tag{30}$$

To end this section and for completeness we re-state also here Theorem 10.2 of van de Geer (2000) since our proof of Theorem 3.2 follows by this.

Lemma 5.1 (Theorem 10.2 of van de Geer) *Consider the regression model $Y_i = f_0(z_i) + \epsilon_i$, $i = 1, \ldots, n$, where f_0 lies in a given subset Θ of the set of all real-valued functions on $[0,1]$, $z_1, \ldots, z_n$ are given points in $[0,1]$ and $\epsilon_1, \ldots, \epsilon_n$ are independent $N(0, \sigma^2)$ measurement errors. Let $R : \Theta \to [0, \infty[$ be a pseudo-norm on Θ and define the penalized least-squares estimator of f_0 by $\hat{f} = \operatorname{argmin}_{f \in \Theta} \|\mathbf{y} - \mathbf{f}\|_n^2 + \nu_n^2 R(f)$. If*

$$H(\delta, \{\frac{f - f_0}{R(f) + R(f_0)};\ f \in \Theta,\ R(f) = R(f_0) > 0\}) \leq A\delta^{-\eta},$$

for all $\delta > 0$, $n \geq 1$ and some $A > 0$ and $0 < \eta < 2$, then

- *If $R(f_0) > 0$ and $\nu_n^{-1} = O_P(n^{1/(2+\eta)}) R^{(2-\eta)/(4+2\eta)}(f_0)$, then*

$$\|\hat{\mathbf{f}} - \mathbf{f}_0\|_n = O_P(\nu_n) R^{1/2}(f_0).$$

- *If $R(f_0) = 0$ and $\nu_n^{-1} = O_P(n^{1/(2+\eta)}) R^{(2-\eta)/(4+2\eta)}(f_0)$, then*

$$\|\hat{\mathbf{f}} - \mathbf{f}_0\|_n = O_P(n^{-1/(2-\eta)}) \nu_n^{-2\eta/(2-\eta)}.$$

Proofs

PROOF OF THEOREM 3.1. Consider the following decomposition of $\mathcal{H}_{J,\Gamma}$:

$$\mathcal{H}_{J,\Gamma} = V_0 \oplus \mathcal{V}_{J,\Gamma},$$

where $\mathcal{V}_{J,\Gamma} = \oplus_{j=0}^{J} \sum_m \mathcal{W}_{j,m,\Gamma}$. Denote by $A(f)$ the functional to be minimized in (17). It is easy to see that $A(f)$ is convex and continuous. By inequality (12) we have $R_J(f) \geq \|f\|_{\mathcal{H}_{J,\Gamma}}$ for any $f \in \mathcal{V}_{J,\Gamma}$. Let K_J^Γ be the reproducing kernel of $\mathcal{V}_{J,\Gamma}$ and $\langle \cdot, \cdot \rangle_J$ be the inner product of $\mathcal{V}_{J,\Gamma}$. Denote by $e_n = \max_{i=1}^n (K_J^\Gamma)^{1/2}(t_i, t_i)$. By the properties of the weighting function Γ and of the reproducing kernel, we have, for any $f \in \mathcal{V}_{J,\Gamma}$ and any $i = 1, \ldots, n$,

$$|f(t_i)| \leq |\langle f, K_J^\Gamma(t_i, \cdot)\rangle_J| \leq e_n \|f\|_{\mathcal{H}_{J,\Gamma}}.$$

The set

$$D = \{f \in \mathcal{H}_{J,\Gamma};\ f = b + f_1, \text{ with } b \in V_0, f_1 \in \mathcal{V}_{J,\Gamma}, R_J(f) \leq v, |b| \leq v^{1/2} + (e_n + 1)v\},$$

where $v = \max_i\{y_i^2 + |y_i| + 1\}$, is obviously closed, convex and bounded. Therefore by Theorem 4 of Tapia and Thompson (1978), there exist a minimizer $\bar{f}$ of (17) in D and $A(\bar{f}) \leq A(0) < v$.

On the other hand, for any $f \in \mathcal{H}_{J,\Gamma}$ with $R_J(f) > v$, clearly $A(f) \geq R_J(f) > v$; for any $f \in \mathcal{H}_{J,\Gamma}$, $f = b + f_1$, with $b \in V_0$, $f_1 \in \mathcal{V}_{J,\Gamma}$, $R_J(f) \leq v$ and $|b| > v^{1/2} + (e_n + 1)v\}$, we therefore have

$$|b + f_1(t_i) - y_i| > (v^{1/2} + (e_n + 1)v) - e_n v - v = v^{1/2}.$$

Thence $A(f) > v$, and for any $f \notin D$, we have $A(f) > A(\bar{f})$ which proves that $\bar{f}$ is the minimizer of (17) in $\mathcal{H}_{J,\Gamma}$.

PROOF OF THEOREM 3.2. The conditions on the unknown regression function f_0 in Theorem 3.2 are only active for its wavelet coefficients and do not include the V_0 scaling coefficient of f_0. This is what essentially makes the difference between the set $\mathcal{B}_{p,q}^s$ and the unit Besov ball of $B_{p,q}^s([0,1))$. To deal with this we will follow the following arguments. For any $f \in \mathcal{H}_{J,\Gamma}$, write $f = b + f_1$ where $b \in V_0$ and $f_1 \in \mathcal{V}_{J,\Gamma}$. The conditions of Theorem 3.2 are equivalent to the fact that the function f_0 is such that $f_{01} \in \mathcal{V}_{J,\Gamma}$. One can also write $A(f)$ as

$$(b - b_0)^2 + \frac{2}{n}(b - b_0)\sum_{i=1}^n \epsilon_i + \frac{1}{n}\sum_{i=1}^n (f_{01}(t_i) + \epsilon_i - f_1(t_i))^2 + \lambda_n R_J(f_1).$$

Therefore, the minimizing $\hat{b}$ is $\hat{b} = b_0 + \frac{1}{n}\sum_{i=1}^n \epsilon_i$, which showns that $\hat{b}$ converges towards b_0 at rate $n^{-1/2}$. On the other hand, $\hat{f}_1$ must minimize over $\mathcal{V}_{J,\Gamma}$, the functional

$$\frac{1}{n}\sum_{i=1}^{n}(f_{01}(t_i) + \epsilon_i - f_1(t_i))^2 + \lambda_n R_J(f_1).$$

We can now apply Lemma 5.1 with $R = R_J$ and $\eta = 1/s$. That the δ-entropy of the corresponding set in Lemma 5.1 is bounded by $A\delta^{-1/s}$ follows the fact that $R_J(f_1 - f_{01}) \leq R_J(f_1) + R_J(f_{01})$ and from Kerkyacharian and Picard's inequality (30), given the fact that the wavelet coefficients of any f_1 in the set specified by Lemma 5.1 satisfy the inequality (19) for $\rho = 2/(2s+1)$. The conclusion of Theorem 3.2 follows.

Proof of Theorem 3.3. Denote by $\|\cdot\|_2^2$ the integrated squared norm and by $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\beta}}$ the vectors of wavelet coefficients of f and $\hat{f}$ respectively (here, the first coefficient is the coefficient for the unit scaling function). Observe that $\|\hat{f} - f\|_2^2 = \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_n^2$ and $\|\hat{f} - f\|_n^2 = \|\Psi(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_n^2$. Hence,

$$\|\hat{f} - f\|_2^2 \leq \|(\Psi^T\Psi)^{-1}\|\|\hat{f} - f\|_n^2$$

where $\|B\|$ is the L^2-norm of the matrix B, and the result follows immediately.

Proof of Lemma 3.1. For any $f \in \mathcal{H}_{J,\Gamma}$, write again $f = b + \sum_{j=0}^J \sum_m f_{jm}$ where $b \in V_0$ and $f_{jm} \in W_{j,m,\Gamma}$. Let the projection of $f_{j,m}$ onto $\text{span}\{K_{jm}(t_i,\cdot), i = 1,\ldots,n\}$ be denoted by $\ell_{j,m}$ and the orthogonal complement by $h_{j,m}$. Then $f_{jm} = \ell_{jm} + h_{jm}$ and (17) can be written as

$$\frac{1}{n}\sum_{i=1}^{n}\{y_i - b - \sum_{j=0}^{J}\sum_m \langle K_{jm}(t_i,\cdot), \ell_{jm}\rangle\}^2 + \lambda^2 \sum_{j=0}^{J}\sum_m (\|\ell_{jm}\|^2 + \|h_{jm}\|^2)^{1/2}.$$

Therefore any minimizing f must be such that $h_{jm} = 0$, $j = 0,\ldots,J; m = 1,\ldots,M_j$ and the conclusion of the lemma follows.

Proof of Lemma 3.2. Denote the functional in (20) by $B(\boldsymbol{\theta}, f)$. For any $j = 0,\ldots,J; m = 1,\ldots,M_j$, we have

$$\lambda_0\theta_{jm}^{-1}\|P_{jm}f\|_{\mathcal{H}_{J,\Gamma}}^2 + \nu\theta_{jm} \geq 2\lambda_0^{1/2}\nu^{1/2}\|P_{jm}f\|_{\mathcal{H}_{J,\Gamma}} = \lambda^2\|P_{jm}f\|_{\mathcal{H}_{J,\Gamma}}$$

for any $\theta_{jm} \geq 0$ and $f \in \mathcal{H}_{J,\Gamma}$, and the equality holds if and only if $\theta_{jm} = \lambda_0^{1/2}\nu^{1/2}\|P_{jm}f\|_{\mathcal{H}_{J,\Gamma}}$. Therefore $B(\boldsymbol{\theta}, f) \geq A(f)$ for any $\theta_{jm} \geq 0$, $j = 0,\ldots,J; m = 1,\ldots,M_j$ and $f \in \mathcal{H}_{J,\Gamma}$, and the equality holds if and only if $\theta_{jm} = \lambda_0^{1/2}\nu^{1/2}\|P_{jm}f\|_{\mathcal{H}_{J,\Gamma}}$. The conclusion then follows.

References

Abramovich F., Bailey T. & Sapatinas T. (2000). Wavelet analysis and its statistical applications. *The Statistician - Journal of the Royal Statistical Society*, Ser. D , **49**, 1–29.

Amato, U. and Vuza, D.T. (1997). Wavelet approximation of a function from samples affected by noise. *Rev. Roumaine Math. Pures Appl.*, **42**, 481–493.

Antoniadis, A. (1996). Smoothing noisy data with tapered coiflets series. *Scandinavian Journal of Statistics*, **23**, 313–330.

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

Antoniadis, A. and Fan, J. (2001). Regularization by Wavelet Approximations, *J. Amer. Statist. Assoc.*, **96**, 939–967.

Aronszajn, N. (1950). Theory of reproducing kernels. *Trans. Am. Math. Soc.*, **68**, 337–404.

Birgé, L. and Massart, P. (2000). An adaptive compression algorithm in Besov spaces, *Journal of Constructive Approximation*, **16**, 1–36.

Birman, M.S. and Solomjak, M.Z. (1967). Piecewise-polynomial approximation of functions of the classes W^p. *Mat. Sbornik.*, **73**, 295–317.

Brinkman, N. (1981). Ethanol fuel - a single-cylinder engine study of efficiency and exhaust emissions. SAE Transactions, **90**, 1414-1424.

Cai, T. (1999). Adaptive wavelet estimation: a block thresholding and oracle inequality approach. *Ann. Statist.*, **27**, 898–924.

Cai, T. (2001). Discussion of "Regularization of Wavelets Approximations" by A. Antoniadis and J. Fan. *J. American Statistical Association*, **96**, 960–962.

Cai, T. and Silverman, B.W. (2001). Incorporating information on neighboring coefficients into wavelet estimation. *Sankhya*, **63**, 127–148.

Canu,S., Mary, X., and Rakotomamonjy, A. (2003). Functional learning through kernel, in *Advances in Learning Theory: Methods, Models and Applications*, NATO Science Series III: Computer and Systems Sciences, Eds Suykens, J *et al.* , IOS Press, Amsterdam , **90**, 89–110.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions. *Numer. Math.*, **31**, 377–403.

Daubechies, I. (1992). *Ten Lectures on Wavelets*. Philadelphia: SIAM.

DeVore, R.A. and Popov, V. (1988). Interpolation of Besov Spaces. *Transactions of the American Mathematical Society*, **305**, 397–414.

Donoho D.L., Elad M and Temlyakov, V. (2004). Stable Recovery of Sparse Overcomplete Representations in the Presence of Noise. Technical report, Stanford University.

Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1995). Wavelet shrinkage: asymptopia? (with discussion). *Journal of the Royal Statistical Society*, Series B, **57**, 301–337.

Eubank, R.L. (1988) *Spline Smoothing and Nonparametric Regression*, New York: Marcel Dekker, Inc.

van de Geer, S. (2000). *Empirical Processes in M-Estimation*. Cambridge University Press.

Gradshtein, I.S., and Ryzhik, I.M. (1980) *Tables of Integrals, Series, and Products.* Academic Press, New York.

Green, P.J. and Silverman, B.W. (1994). *Nonparametric Regression and Generalised Linear Models*. London: Chapman and Hall.

Gunn, S. R. and Kandola, J. S. (2002). Structural modeling with sparse kernels. *Mach. Learning*, **48**, 115–136.

Hall, P., Kerkyacharian, G. and Picard, D. (1999). On the minimax optimality of block thresholded wavelet estimators. *Statist. Sinica*, **9**, 33–50.

Härdle,W., Kerkyacharian, G., Picard, D and Tsybakov, A. (1998). *Wavelets, Approximation, and Statistical Applications*, Lecture Notes in Statistics, **129**, Springer-Verlag, New-York.

Karlovitz, L.A. (1970). Construction of nearest points in the l_p, p even and l_1 norms, *Journal of Approximation Theory*, **3**, 123–127.

Kerkyacharian, G. and Picard, D. (2003). Replicant compression coding in Besov spaces, *ESAIM: P & S*, **7**, 239–250.

Kimeldorf G., and Wahba, G (1971). Some results on Tchebycheffian spline functions. *J. Math. Anal. Applic.*, **33**, 82–95.

Kovac, A. and Silverman, B. (2000). Extending the scope of wavelet regression methods by coefficient-dependent thresholding. *J. Am. Stat. Assoc.*, **95**, 172–183.

Lin, Y. and Zhang, H. H. (2003). Component Selection and Smoothing in Smoothing Spline Analysis of Variance Models, Technical report, University of Wisconsin - Madison.

Lin, X., Wahba, G., Xiang, D., Gao, F., Klein, R. and Klein, B. (2000). Smoothing spline ANOVA models for large data sets with Bernoulli observations and the randomized GACV. *Ann. Statist.*, **28**, 1570–1600.

Loubes, M. and van de Geer, S. (2002), Adaptive estimation with soft thresholding penalties, *Statistica Neerlandica*, **56**, 454–479.

Mallat, S.G. (1999). *A Wavelet Tour of Signal Processing*. 2nd ed. San Diego: Academic Press.

Meyer, Y. (1992). *Wavelets and Operators*. Cambridge: Cambridge University Press.

Nason, G. (1998). *WaveThresh3 Software*. Department of Mathematics, University of Bristol, Bristol, UK.

Silverman, B. W. (1985) Some aspects of the spline smoothing approach to non-parametric curve fitting. *Journal of the Royal Statistical Society series B.*, **47**, 1–52.

Tapia, R. and Thompson, J. (1978). *Nonparametric Probability Density Estimation*. Baltimore, MD, Johns Hopkins University Press.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society, B*, **58**, 267–288.

Triebel, H. (1983). *Theory of Function Spaces*. Birkhäuser Verlag, Basel.

Vidakovic, B. (1999). *Statistical Modeling by Wavelets*. New York: John Wiley & Sons.

Wahba, G. (1990). *Spline Models for Observational Data*, SIAM. CBMS-NSF Regional Conference Series in Applied Mathematics, **59**.

Wahba, G., Wang, Y.,Gu, C., Klein, R. and Klein, B. (1995) Smoothing spline ANOVA for exponential families, with application to the Wisconsin Epidemiological Study of Diabetic Retinopathy. *Ann. Statist.*, **23**, 1865–1895.

Zhang, H., Wahba, G., Lin, Y., Voelker, M., Ferris, M., Klein, R. and Klein, B. (2002). Variable selection and model building via likelihood basis pursuit. Technical report, University of Wisconsin - Madison.