Multivariate wavelet kernel regression method

Samir Touzani, Daniel Busby

Multivariate wavelet kernel regression method

Samir Touzani, Daniel Busby[a]

[a]*IFP Energies nouvelles, 92500 Rueil-Malmaison, France*

Abstract

The purpose of this paper is to introduce a new penalized multivariate nonparametric regression method, in the framework of wavelet decomposition. We call this method the wavelet kernel ANOVA (WK-ANOVA), which is a wavelet based reproducing kernel Hilbert space (RKHS) method with the penalty equal to the sum of blockwise RKHS norms. This method does not require design points to be equispaced or of dyadic size thus making high-dimensional wavelet estimation feasible. We also introduce a new iterative shrinkage algorithm to solve the nonegative garrote optimization problem resulting in the variable selection step. Numerical experiments on several test functions show that the WK-ANOVA provides competitive results compared to other standard methods.

Keywords: Wavelet, Reproducing kernel, Nonparametric regression, ANOVA, Landweber iterations.

1. Introduction

In this work we consider the classical multivariate nonparametric regression problem with the aim of recovering an unknown function f from noisy data

$$y_i = f(\boldsymbol{x}_i) + \epsilon_i,\ i = 1, \ldots, n$$

where $\boldsymbol{x}_i = (x_{i_1}, \ldots, x_{i_d})^T$ is d dimensional vector of inputs, the ϵ_i are i.i.d. and $N(0, \sigma^2)$ random errors.

The scope of this work is to take advantage of the multiresolution analysis provided by wavelet decomposition in multivariate function estimation. Indeed, in recent years there has been an important development in the application of wavelet methods in statistics, especially in signal processing, in image and function representation methods, with many successes in the efficient analysis and compression of noisy data.

The multiresolution analysis provides a good time frequency localization, which makes wavelet methods particularly effective to estimate functions with sharp spikes, and discontinuities. Thus wavelets are used in various nonparametric regression methods. However most of these methods are implemented only for one (signal) or two (image) dimensional problems. The reason for this is that these algorithms assumed that data is dyadic and with equally spaced points. Several algorithms have been proposed to overcome the setting of non-dyadic and non-equispaced design. Among them, Antoniadis et al. (1997) transforms the random design into equispaced data via a binning method. Kovac and Silverman (2000) apply the linear transformation to

Email addresses: samirtouzani.phd@gmail.com (Samir Touzani), daniel.busby@ifpen.fr (Daniel Busby)

the data to map it to a dyadic and equispaced set of points. Kerkyacharian and Picard (2004) project the data on an unusual non-orthonormal basis, called warped wavelet basis. Amato et al. (2006) suggested a regularization method relying on wavelet kernel reproducing Hilbert spaces, which does not require a pre-processing of data. The method also achieves optimal convergence rates in the Besov spaces when the estimation error is calculated at the design points only no matter how irregular the design is. Given that, it seems that this method is well adapted to be generalized for the multivariate regression using wavelets.
Inspired by the component selection and smoothing operator (COSSO) (Lin and Zhang, 2006), which is based on ANOVA (ANalysis Of VAriance) decomposition, and the wavelet kernel penalized estimation for non-equispaced design regression proposed by Amato et al. (2006), we introduce a new approach in the estimation of ANOVA components. Given a wavelet type expansion of f we consider a class of wavelet estimators for the nonparametric regression problem using a penalized least-squares approach. The penalties are chosen in order to control the smoothness of the resulting estimator. For this we use the same penalty as the one used for COSSO, in other words the semi-norm penalty. So we take for penalty, a weighted sum of wavelet details norms.
In this paper, we first shortly review some definitions on wavelet. Then we present a new nonparametric regression method, named Wavelet Kernel ANOVA (WK-ANOVA). A new iterative projected shrinkage algorithm based on Landweber iterations is introduced in section 4. Finally, numerical tests are presented and discussed.

2. Multiresolution analysis and wavelet kernel

Consider the univariate regression problem:

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

where $(x_i)_{i=1,\ldots,n}$ is the irregular design, the ϵ_i are i.i.d. and $N(0, \sigma^2)$ random errors and f an unknown regression function to be estimated.

2.1. Multiresolution analysis

We define a multi resolution analysis as a sequence of closed subspaces V_j, $j \in \mathbb{Z}$ in $L_2(\mathbb{R})$ and which possesses the following properties:

1. $\cap_{j\in\mathbb{Z}} V_j = 0$,
2. $\overline{\cup_{j\in\mathbb{Z}} V_j} = \boldsymbol{L}^2(\mathbb{R})$,
3. $\forall f \in \boldsymbol{L}^2(\mathbb{R})$, $\forall j \in \mathbb{Z}$, $f \in V_j$ if and only if $f(2x) \in V_{j+1}$;
4. $\forall f \in \boldsymbol{L}^2(\mathbb{R})$, $\forall k \in \mathbb{Z}$, $f \in V_0$ if and only if $f(x-k) \in V_0$,
5. there exist a scaling function $\phi \in V_0$ whose integer-translates $x \mapsto \phi(x-k)_{k\in(Z)}$ span the space V_0.

If we define $P^j f$ as the projection of a function f onto the space V_j, this is expressed by

$$P^j f = P^{j-1} f + w^{j-1}$$

where the function w^{j-1} represents the residual between the two approximations on V^j and on V^{j-1}. This function can be written in terms of dilated and translated wavelets:

$$w^{j-1} = \sum_{k\in\mathbb{Z}} \langle f, \psi_{j-1,k}\rangle \psi_{j-1,k}$$

where $\{\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k);\ k \in \mathbb{Z}\}$ is a set of functions that are orthogonal to each function of V_j and span the space W_j which is a detail space. Hence, an important property of multiresolution analysis can be defined as:

$$V_j = V_{j-1} \oplus W_{j-1}$$

By periodizing an orthonormal basis for $L_2(\mathbb{R})$ we construct an orthonormal wavelet basis for $L_2([0,1])$ generated by dilatation and translation of compactly supported scaling function ϕ^{per}, and a compactly supported wavelet ψ^{per}, where:

$$\phi^{per}_{j,k}(x) = \sum_{l\in\mathbb{Z}} \phi_{j,k}(x+l),\ \psi^{per}_{j,k}(x) = \sum_{l\in\mathbb{Z}} \psi_{j,k}(x+l)$$

and

$$V^{per}_j = \overline{span\{\phi^{per}_{j,k}, k \in \mathbb{Z}\}},\ W^{per}_j = \overline{span\{\psi^{per}_{j,k}, k \in \mathbb{Z}\}}$$

The resulting orthogonal basis provides an orthogonal decomposition

$$L_2([0,1]) = V^{per}_0 \oplus W^{per}_0 \oplus W^{per}_1 \oplus ...$$

where V^{per}_0 (spanned by $\phi_{0,0} = \psi_{-1,0}$) consists of constant function and W^{per}_j is a 2^j dimensional space. To enhance the interpretability we omit in what follows the index *per*.
For any integer $j_0 \geq 0$ any function $f \in L_2([0,1])$ is expressed in the form:

$$f(t) = \sum_{k=0}^{2^{j_0}-1} \alpha_{j_0,k}\phi_{j_0,k}(t) + \sum_{j\geq j_0}\sum_{k=0}^{2^{j_0}-1} \beta_{j,k}\psi_{j,k}(t),\ t \in [0,1]$$

where the scaling coefficients are $\alpha_{j_0,k} = \langle f, \phi_{j_0,k}\rangle$ $(k = 0, 1, ..., 2^{j_0} - 1)$ and the wavelet coefficients are $\beta_{j,k} = \langle f, \phi_{j,k}\rangle$ $(j \geq j_0, k = 0, 1, ..., 2^{j_0} - 1)$.
For more details on the mathematical aspects of wavelets and their applications in statistical settings we refer to Daubechies (1992), Vidakovic (1999), Ogden (1997) and Antoniadis et al. (2001).

2.2. Wavelet kernel

Let $G_{-1} = \{-1\} \times \{0\}$, $G_0 = \{0\} \times \{0, 1\}$ and for each integer $J \geq 1$ let $G_J = \{J\} \times \{k \in \{0, ..., 2^J\}; k/2 \notin \mathbb{Z}\}$, i.e. G_J is the index set of wavelets at resolution level J. The whole set of indexes pairs (j, k) that describes all wavelets will be denoted by $G = \bigcup_{j\geq -1} G_j$. Therefore, any function $f \in L_2([0, 1))$ admits the infinite wavelet expansion:

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

where ψ_g is the wavelet basis function indexed by $g \in G$, f_g is the corresponding expansion coefficient and $\psi_{-1,0} = \phi_{0,0}$.
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 let be $\|\cdot\|_\Gamma$ the associated norm. As G_J is a finite subset of G, we have $V_J \subset \mathcal{H}_\Gamma$ for every $J \geq 1$ defined as:

$$V_J = V_0 \oplus \bigoplus_{j=0}^{J-1} W_j$$

Moreover, for any $f \in \mathcal{H}_\Gamma$,

$$\lim_{J\to\infty} \|f - P^J(f)\|_\Gamma = 0 \tag{2}$$

where $P^J(f)$ is the projection of a function f into the space V_J. The space $\mathcal{H}_\Gamma$ is a RKHS and the corresponding reproducing kernels are given by:

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

By definition of the index set G, the kernel K can also be written as a sum of the reproducing kernels:

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

This implies that the RKHS $\mathcal{H}_\Gamma$, can be decomposed into a direct sum of wavelet RKHS's (spanned by a set of wavelets of scale j) as:

$$\mathcal{H}_\Gamma = V_0 \oplus \bigoplus_{j\geq 0} \mathcal{W}_j^\Gamma \tag{3}$$

where each space $\mathcal{W}_j^\Gamma$ is the RKHS associated to the kernel K_j^Γ. This representation involves an infinite decomposition of the detail space, in practice we truncate (3) up to a maximum resolution J, in other words, the RKHS $\mathcal{H}_{J,\Gamma} = V_0 \oplus \bigoplus_{j=0}^J \mathcal{W}_j^\Gamma$ defines a multiresolution analysis of $\mathcal{H}_\Gamma$ and the associated kernel is:

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

Furthermore, from (2)

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

We assume that Γ is only a function of j and equals to 2^{2js} on G_j and $s > 1/2$, then $\mathcal{H}_\Gamma$ equals to the Sobolev space $B_{2,2}^s([0,1])$ of index s. For more mathematical details we refer to Amato et al. (2006).

3. The wavelet kernel ANOVA

Consider now a multivariate regression problem:

$$y_i = f(\boldsymbol{x}_i) + \epsilon_i,\ i = 1, \cdots, n$$

where $\boldsymbol{x}_i = (x_{i_1}, \ldots, x_{i_d})^T$ is d dimensional vector of inputs, the ϵ_i are i.i.d. and $N(0, \sigma^2)$ random errors and f an unknown multivariate regression function to be estimated.

3.1. Definition

The idea behind the well known smoothing spline ANOVA model (Wahba, 1990) is to construct a RKHS $\mathcal{F} = \{f \in L^2([0,1]^d)\}$ corresponding to the ANOVA decomposition:

$$f(X^{(1)}, \cdots, X^{(d)}) = f_0 + \sum_{l=1}^{d} f_l(X^{(l)}) + \sum_{l<m} f_{lm}(X^{(l)}, X^{(m)}) + \ldots + f_{1,2,\ldots,d}(X^{(1)}, \cdots, X^{(d)}) \quad (4)$$

where f_0 is a constant, f_j's are univariate functions representing the main effects, f_{jl}'s are bivariate functions representing the two way interactions, and so on. Then the model space $\mathcal{F}$ is the tensor product space of $\mathcal{H}_\Gamma^l$:

$$\mathcal{F} = \bigotimes_{l=1}^{d} \mathcal{H}_\Gamma^l = \{1\} \oplus \sum_{l=1}^{d} \bar{\mathcal{H}}_\Gamma^l \oplus \sum_{l<m} [\bar{\mathcal{H}}_\Gamma^l \otimes \bar{\mathcal{H}}_\Gamma^m] \ldots \quad (5)$$

where $\mathcal{H}_\Gamma^l = \{1\} \oplus \bar{\mathcal{H}}_\Gamma^l$ and $\bar{\mathcal{H}}_\Gamma^l$ is the RKHS associated to the first-order component functions f_l of ANOVA expansion. The tensor products $[\bar{\mathcal{H}}_\Gamma^l \otimes \bar{\mathcal{H}}_\Gamma^m]$ is associated to the second-order component function f_{lm}. We denote by $\mathcal{W}_j^l$ (a detail space at scale j) the RKHS associated to wavelet kernel $K_j = \sum_{k=0}^{2^j-1} \psi_{j,k}(x)\psi_{j,k}(y)$ and the variate $X^{(l)}$, thereby for a fixed maximum resolution J the function space $\mathcal{H}_\Gamma^l$ can be written as:

$$\mathcal{H}_\Gamma^l = V_0 \oplus \bigoplus_{j=0}^{J-1} \Gamma_j^{-1} \mathcal{W}_j^l \quad (6)$$

and the tensor product $[\mathcal{H}_\Gamma^l \otimes \mathcal{H}_\Gamma^m]$ as:

$$\mathcal{H}_\Gamma^l \otimes \mathcal{H}_\Gamma^m = (V_0^l \otimes V_0^m) \bigoplus_{j=0}^{J-1} \Gamma_j^{-2} (\mathcal{W}_j^l \otimes \mathcal{W}_j^m) \quad (7)$$

It's easy to see that V_0 is also the subspace of $L_2([0,1])$ spanned by the constant function on $[0,1]$, one has $V_0 = V_0^l \otimes V_0^m = \{1\}$.
Thus, the function space $\mathcal{F}$, which is a wavelet-based RKHS, can also be written as:

$$\mathcal{F} = \{1\} \oplus \bigoplus_{\gamma=1}^{q} \mathcal{F}_\gamma \quad (8)$$

where the $\mathcal{F}_\gamma$ is an orthogonal subspaces of $\mathcal{F}$ and correspond to the subspaces $\bar{\mathcal{H}}_\Gamma^l$, $[\bar{\mathcal{H}}_\Gamma^l \otimes \bar{\mathcal{H}}_\Gamma^m]$, etc... In the additive model $q = d$ where d is the number of input parameters and in the model

with two way interactions $q = d(d+1)/2$. We assume that a second order ANOVA expansion gives a satisfactory approximation of f.
We denote by $P_\gamma^\Gamma f$ the orthogonal projection of f onto $\Gamma_j^{-1}\mathcal{W}_j$ and $\|\cdot\|$ the norm in the RKHS $\Gamma_j^{-1}\mathcal{W}_j$. Under the framework of smoothing spline ANOVA one way to estimate f is to find $f \in \mathcal{F}$ that minimizes:

$$\frac{1}{n}\sum_{i=1}^{n}\{y_i - f(\boldsymbol{x}_i)\}^2 + \lambda^2 \sum_{\gamma=1}^{q}\sum_{j=0}^{J-1}\sum_{k=0}^{2^j-1}\theta_{\gamma,j,k}^{-1}\|P_{\gamma,j,k}^{\Gamma} f\|^2 \tag{9}$$

where $\theta_{\gamma,j,k} \geq 0$. If $\theta_{\gamma,j,k} = 0$, then the minimizer is taken to satisfy $\|P_{\gamma,j,k}^{\Gamma} f\|^2 = 0$, using the convention $0/0 = 0$. The parameter λ controls the trade-off between the first term in the above expression which discourages the lack of fit of f and the second one which penalizes the roughness of f.
In analogy with COSSO (Lin and Zhang, 2006) and wavelet kernel penalized estimation (Amato et al., 2006) we propose the WK-ANOVA procedure, another way to estimate f, given by $f \in \mathcal{F}$ that minimize:

$$\frac{1}{n}\sum_{i=1}^{n}\{y_i - f(\boldsymbol{x}_i)\}^2 + \lambda^2 R_q(f) \tag{10}$$

with $R_q(f) = \sum_{\gamma=1}^{q}\sum_{j=0}^{J-1}\sum_{k=0}^{2^j-1}\|P_{\gamma,j,k}^{\Gamma} f\|$ is a sum of wavelet-based RKHS norms, instead of the squared RKHS norm employed in (9). We note that $R_q(f)$ is not a norm in $\mathcal{F}$ but a pseudo-norm in the following sense: $R_q(f) \geq 0$, $R_q(cf) = |c|R_q(f)$, $R_q(f+h) \leq R_q(f) + R_q(h)$ $\forall\ f, h \in \mathcal{F}$, and, $R_q(f) > 0$ for any non constant $f \in \mathcal{F}$. Furthermore

$$B \leq R_q(f)^2 \leq qB \tag{11}$$

with

$$B = \sum_{\gamma=1}^{q}\sum_{j=0}^{J-1}\sum_{k=0}^{2^j-1}\|P_{\gamma,j,k}^{\Gamma} f\|^2 \tag{12}$$

Note that there is only one smoothing parameter λ which should be properly chosen, instead of multiple smoothing parameters θ's in (9).

The existence of the WK-ANOVA estimate, which is due to the convexity of (10), is guaranteed by adapting Theorem 1 of Lin and Zhang (2006).

Theorem 3.1. *Let $\mathcal{F}$ be the wavelet-based RKHS of functions over $[0,1]^d$. Assume that $\mathcal{F}$ can be decomposed as (8). There exists a minimizer of (10).*

Define $\|\cdot\|_n$ as the Euclidian norm in $\mathbb{R}^n$. Under our previous assumption, $s > 1/2$ and $\Gamma_j = 2^{2sj}$. The following theorem is equivalent to theorem 2 of Lin and Zhang (2006) and shows that the WK-ANOVA estimator in the additive model has a rate of convergence $n^{-s/(2s+1)}$, where s is the order of smoothness of the components.

Theorem 3.2. *Consider the regression model $y_i = f_0(\boldsymbol{x}_i) + \varepsilon_i$, $i = 1, ..., n$, where $\boldsymbol{x}_i$'s are given deterministic points in $[0,1]^d$, and the ε_i's are i.i.d. $N(0, \sigma^2)$ noise variables. Assume f_0 lies in $\mathcal{F} = \{1\} \oplus \bigoplus_{l=1}^{d}\mathcal{H}_\Gamma^l$, with $\mathcal{H}_\Gamma^\gamma = \{1\} \oplus \bar{\mathcal{H}}_\Gamma^\gamma$ being the Sobolev space $B_{2,2}^s([0,1])$ of index s. Consider the WK-ANOVA estimator $\hat{f}$ at the design points as defined by (10).*

Then (i) if f_0 is not a constant, and $\lambda_n^{-1} = O_p(n^{s/(2s-1)})R_q^{(2s-1)/(4s+2)}(f_0)$, we have $\|\hat{f} - f_0\|_n = O_p(\lambda_n)R_q^{1/2}(f_0)$; (ii) if f_0 is a constant, we have $\|\hat{f} - f_0\|_n = O_p(max\{n^{-s/(2s-1)} \lambda_n^{-2/(2s-1)}, n^{-1/2}\})$.

The following Lemma shows that the solution of (10) is in finite dimensional space and the WK-ANOVA estimate can be computed directly from (10) by linear programming techniques.

Lemma 3.3. *Let $\hat{f} = \hat{b} + \sum_{\gamma=1}^{q} \hat{f}_\gamma$ be a minimizer of (10), with $f_\gamma \in \mathcal{F}_\gamma$. Then $\hat{f}_\gamma \in span\{K_\gamma(\mathbf{x}_i, \cdot), i = 1, ..., n\}$, where K_γ is the reproducing kernel of the space $\mathcal{F}_\gamma$*

Using the suggestion of Antoniadis and Fan (2001) for solving penalized problems with l_1 penalty, we can give an equivalent formulation of (10) for computational consideration. Consider the problem of finding $\boldsymbol{\theta} = \{\theta_{\gamma,j,k}, \gamma = 1, ..., q; j = 0, ..., J-1; k = 1, ..., 2^j - 1\}$ and $f \in \mathcal{F}$ to minimize:

$$\frac{1}{n}\sum_{i=1}^{n}\{y_i - f(\boldsymbol{x}_i)\}^2 + \lambda_0 \sum_{\gamma=1}^{q}\sum_{j=0}^{J-1}\sum_{k=0}^{2^j-1} \theta_{\gamma,j,k}^{-1}\|P_{\gamma,j,k}^{\Gamma} f\|^2 + \nu \sum_{\gamma=1}^{q}\sum_{j=0}^{J-1}\sum_{k=0}^{2^j-1}\theta_{\gamma,j,k} \tag{13}$$

subject to $\theta_{\gamma,j,k} \geq 0$, and where λ_0 is a fixed positive constant and ν is a smoothing parameter. We fix λ_0 at some value. Then

Lemma 3.4. *Set $\nu = \lambda^4/(4\lambda_0)$. (i) if $\hat{f}$ minimizes (10), set $\hat{\theta}_{\gamma,j,k} = \lambda_0^{1/2}\nu^{-1/2}\|P_{\gamma,j,k}^{\Gamma} f\|$, then the pair $(\hat{\boldsymbol{\theta}}, \hat{\mathbf{f}})$ minimizes (13). (ii) On the other hand, if a pair $(\hat{\boldsymbol{\theta}}, \hat{\mathbf{f}})$ minimizes (13), then $\hat{\mathbf{f}}$ minimizes (10).*

As already introduced by Amato et al. (2006) we can penalize the norm of coefficients by blocks, which allows reducing the number of θ's that need to be estimated and can provide a better regularization. Hence, as defined before:

$$K_{jm}^{\Gamma}(x, y) = \sum_{k \in T_{jm}} \frac{\psi_{j,k}(x)}{\Gamma_j}\psi_{j,k}(y)$$

where $m = 1, ..., M_j$ with M_j denotes the number of blocks at resolution j, and T_{jm} the blocks of length L_{jm} at resolution j. In the same way consider the decomposition:

$$\mathcal{H}_{\Gamma}^{l} = V_0 \oplus \bigoplus_{j=0}^{J-1}\sum_{m}\Gamma_j^{-1}\mathcal{W}_{j,m}^{l}$$

then replace (13) by:

$$\frac{1}{n}\sum_{i=1}^{n}\{y_i - f(\boldsymbol{x}_i)\}^2 + \lambda_0 \sum_{\gamma=1}^{q}\sum_{j=0}^{J-1}\sum_{m=1}^{M_j} \theta_{\gamma,j,m}^{-1}\|P_{\gamma,j,m}^{\Gamma} f\|^2 + \nu \sum_{\gamma=1}^{q}\sum_{j=0}^{J-1}\sum_{m=1}^{M_j}\theta_{\gamma,j,m} \tag{14}$$

We can note that the form of (14) is similar to the smoothing spline ANOVA (9) with multiple smoothing parameters and an additional penalty on the θ's. There is only one smoothing parameter ν in (14) and θ's are part of the estimate, rather than three smoothing parameters. For the WK-ANOVA procedure the sparsity on the detail components is controlled by the additional penalty on θ's in (14) makes possible to have some θ's to be zero, thus producing a sparse kernel estimate in sense of Gunn and Kandola (2002).

3.2. Algorithm

We will use an iterative optimization algorithm which is equivalent to the one used in Lin and Zhang (2006) and Amato et al. (2006). On each step of iteration, for any fixed θ we minimize (14) with respect of f, and then for this choice of f we minimize (14) with respect of θ. Note that for any fixed θ (14) is equivalent to the smoothing spline ANOVA procedure. Therefore from Wahba (1990) the solution f of (14) has the following form

$$f(\mathbf{x}) = b + \sum_{i=1}^{n} c_i \sum_{\gamma=0}^{q} \sum_{j=0}^{J-1} \sum_{m=1}^{M_j} \theta_{\gamma,j,m} K^{\Gamma}_{\gamma,j,m}(\mathbf{x}_i, \mathbf{x}) \tag{15}$$

Where $c = (c_1, ..., c_n)^T$, $b \in \mathbb{R}$, $K^{\Gamma}_{\gamma,j,m}$ is the reproducing kernel of $\Gamma_j^{-1}\mathcal{W}^l_{\gamma,j,m}$ if $\gamma \leq d$ and is the reproducing kernel of $\Gamma_j^{-2}\mathcal{W}^l_{\gamma,j,m} \otimes \mathcal{W}^p_{\gamma,j,m}$ else. In what follows, we denote by $K^{\Gamma}_{\gamma,j,m}$ the $n \times n$ matrix $\{K^{\Gamma}_{\gamma,j,m}(\mathbf{x}_i, \mathbf{x}_t)\}$, $i = 1, ..., n$, $t = 1, ..., n$, by K^{Γ}_{θ} the matrix $\sum_{\gamma=0}^{q}\sum_{j=0}^{J-1}\sum_{m=1}^{M_j} \theta_{\gamma,j,m}K^{\Gamma}_{\gamma,j,m}(\mathbf{x}_i, \mathbf{x})$ and $\mathbf{1}_n$ be the column vector consisting of n ones. Then we can write $\mathbf{f} = K^{\Gamma}_{\theta}\mathbf{c} + b\mathbf{1}_n$, it follows that (14) can be expressed as

$$\frac{1}{n}\|\mathbf{Y} - \sum_{\gamma=0}^{q} \sum_{j=0}^{J-1} \sum_{m=1}^{M_j} \theta_{\gamma,j,m} K^{\Gamma}_{\gamma,j,m}\mathbf{c} - b\mathbf{1}_n\|_n^2 + \lambda_0 \mathbf{c}^T K^{\Gamma}_{\theta}\mathbf{c} + \nu \sum_{\gamma=0}^{q} \sum_{j=0}^{J-1} \sum_{m=1}^{M_j} \theta_{\gamma,j,m} \tag{16}$$

where $\theta_{\gamma,j,m} \geq 0$, $\gamma = 1, ..., q$, $j = 0, ..., J-1$, $m = 1, ..., M_j$.
If θ's are fixed, then (16) can be written as

$$\min_{\mathbf{c},b} \|\mathbf{Y} - K^{\Gamma}_{\theta}\mathbf{c} - b\mathbf{1}_n\|_n^2 + n\lambda_0 \mathbf{c}^T K^{\Gamma}_{\theta}\mathbf{c} \tag{17}$$

which is similar to the smooting spline problem (a quadratic minimization problem) and the solution satisfy (for more details see Wahba (1990)):

$$\begin{aligned} (K^{\Gamma}_{\theta} + n\lambda_0 I)\mathbf{c} + b\mathbf{1}_n &= \mathbf{Y} \\ \mathbf{1}_n^T\mathbf{c} &= 0 \end{aligned}$$

where I is the identity matrix.
Let's fix b and c at their values from (17), denote $d_{\gamma,j,m} = K^{\Gamma}_{\gamma,j,m}\mathbf{c}$, and let D be the $n \times (\sum_\gamma \sum_j (2^j - 1))$ matrix with the (γ, j, m)th column being $d_{\gamma,j,m}$. The $\boldsymbol{\theta}$ that minimizes (16) is the same as the solution to

$$\min_{\boldsymbol{\theta}} \|\mathbf{z} - D\boldsymbol{\theta}\|_n^2 + n\nu \sum_{\gamma=0}^{q} \sum_{j=0}^{J-1} \sum_{m=1}^{M_j} \theta_{\gamma,j,m} \quad \text{subject to} \quad \theta_{\gamma,j,m} \geq 0 \tag{18}$$

where $\mathbf{z} = \mathbf{Y} - (1/2)n\lambda_0\mathbf{c} - b\mathbf{1}_n$.
By starting from a simpler estimate such as the one obtained by penalized least squares with quadratic penalties on the coefficients, a one step update procedure is sufficient to improve the WK-ANOVA estimator. Then we propose a one step update procedure:

1. Initialization: Fix $\theta_{\gamma,j,m} = 1$, $\gamma = 1, ..., q$, $j = 0, ..., J-1$, $m = 1, ..., M_j$.
2. Tune λ_0 using v-fold-cross-validation.
3. Solve for $\mathbf{c}$ and b with (17).

4. For each fixed ν, solve (18) with the $\mathbf{c}$ and b obtained in step 3. Tune ν using v-fold-cross-validation. The θ's corresponding to the best ν are the final solution at this step.
5. With the new $\boldsymbol{\theta}$ tune λ_0 using v-fold-cross-validation.
6. With the new $\boldsymbol{\theta}$ and λ_0, solve for $\mathbf{c}$ and b with (17)

A discussion of a one step procedure and fully iterated procedure can be found in Antoniadis and Fan (2001). The performance of the WK-ANOVA estimator depends on the smoothing parameter ν and the chosen resolution J. The choice of these parameters obviously involves an arbitrary decision. In our work we will fix $J = log_2 n$, but by varying the resolution level we can explore features of the data arising on different scales. We will use v fold cross validation to tune ν. It seems reasonable to take v equal to 5.
We also choose to use compactly supported wavelets, it follows that the numerical algorithm for the kernel computation is based on Daubechies cascade procedures (Daubechies, 1992). Specifically, the cascade algorithm computes the values of wavelets at dyadic points. In order to evaluate the kernel matrices $K^{\Gamma}_{\gamma,j,m}$ the values of the wavelets have been computed on a fine dyadic grid and stored in a table. Values of wavelets at arbitrary points, necessary for evaluation of $K^{\Gamma}_{\gamma,j,m}$, were then computed by considering the value at the closest point on the tabulated grid. The table construction of wavelet kernel matrices requires $O(n^2 S)$ elementary operations where S denotes the length of the wavelet filter. However, the table is constructed once and stored in memory. In addition, as the dimension of the problem grows, the number of matrices $K^{\Gamma}_{\gamma,j,m}$ also grows as well, and because of the v-fold-cross-validation these matrices must be re-computed several times. All this increases significantly the computational time, and therefore it is necessary to compute the matrices once and stored them in memory.
The formulation in (18) is a high-dimensional nonnegative garrote (NNG) optimization problem introduced by Breiman (1995) for which there exists a variety of algorithms to find the solution. However, we introduce in the next section a new iterative shrinkage algorithm. Even if the number of iterations may be high to compete with modern high-dimensional optimization algorithms, the proposed iterative algorithm is conceptually simple and easy to implement.

4. Iterative projected shrinkage algorithm

Recently, iterative shrinkage/theresholding (IST) algorithm tailored to solve the LASSO regression problem (Tibshirani, 1996), has been proposed independently by several research groups (among them Figueiredo and Nowak (2003) and Daubechies et al. (2004)). This algorithm combines the Landweber iteration (Landweber, 1951) and the soft thresholding method. The main advantages of this algorithm are its conceptual simplicity, its easiness of implementation and only involves matrix-vector multiplication.
In this section, we propose a modified version of IST algorithm that solves the NNG regression problem.

4.1. Definition

Consider the (18) regression problem:

$$\min_{\boldsymbol{\theta}} \|\mathbf{z} - D\boldsymbol{\theta}\|_n^2 + n\nu \sum_{\gamma=0}^{q} \sum_{j=0}^{J-1} \sum_{m=1}^{M_j} \theta_{\gamma,j,m} \quad \text{subject to} \quad \theta_{\gamma,j,m} \geq 0$$

The functional (18) is convex since the matrix $D^T D$ is symmetric and positive semidefinite and since the constraints $\theta_{\gamma,j,m} > 0$ define also a convex feasible set. For the convex optimization problem, the Karush-Kuhn-Tucker (KKT) conditions are necessary and sufficient for the optimal solution $\boldsymbol{\theta}^*$, where $\boldsymbol{\theta}^* = argmin_{\boldsymbol{\theta}} \|\mathbf{z} - D\boldsymbol{\theta}\|^2 + n\nu \sum_{\gamma=0}^{q} \sum_{j=0}^{J-1} \sum_{m=1}^{M_j} \theta_{\gamma,j,m}$ subject to $\theta_{\gamma,j,m} \geq 0$. This KKT conditions are defined as:

$$\begin{aligned} \{-\mathbf{d}_{\gamma,j,m}^T(\mathbf{z} - D\boldsymbol{\theta}^*) + \nu\}\theta_{\gamma,j,m}^* &= 0 \\ -\mathbf{d}_{\gamma,j,m}^T(\mathbf{z} - D\boldsymbol{\theta}^*) + \nu &\geq 0 \\ \theta_{\gamma,j,m}^* &\geq 0 \end{aligned}$$

which is equivalent to

$$-\mathbf{d}_{\gamma,j,m}^T(\mathbf{z} - D\boldsymbol{\theta}^*) + \nu = 0, \text{if } \theta_{\gamma,j,m}^* \neq 0 \tag{19}$$

$$-\mathbf{d}_{\gamma,j,m}^T(\mathbf{z} - D\boldsymbol{\theta}^*) + \nu > 0, \text{if } \theta_{\gamma,j,m}^* = 0 \tag{20}$$

where $\mathbf{d}_{\gamma,j,m}$ denotes the (γ, j, m)th column of D. Therefore, from (19) and (20) we can derive the fixed-point equation:

$$\boldsymbol{\theta}^* = \mathcal{P}_{\Omega^+}(\delta_\nu^{Soft}(\boldsymbol{\theta}^* + D^T(\mathbf{z} - D\boldsymbol{\theta}^*))) \tag{21}$$

where $\mathcal{P}_{\Omega^+}$ is the nearest point projection operator onto the nonnegative orthant (closed convex set) $\Omega^+ = \{x : x \geq 0\}$ and δ_λ^{Soft} is the soft-thresholding function defined as

$$\delta_\nu^{Soft}(x) = \begin{cases} 0 & \text{if } |x| \leq \nu \\ x - \nu & \text{if } x > \nu \\ x + \nu & \text{if } x < -\nu \end{cases}$$

Thus, we propose an iterative algorithm that we name the iterative projected shrinkage algorithm (IPS) and which is defined by

$$\boldsymbol{\theta}^{[p+1]} = \mathcal{P}_{\Omega^+}(\delta_\nu^{Soft}(\boldsymbol{\theta}^{[p]} + D^T(\mathbf{z} - D\boldsymbol{\theta}^{[p]}))) \tag{22}$$

The following theorem concerns the convergences of IPS algorithm:

Theorem 4.1. *IPS algorithm defined by (22) converge to the solution of (18), whenever such solution exists, for any starting vector* $\boldsymbol{\theta}^{[0]}$.

The proof of this theorem can be found in the Appendix. We have assumed that $\lambda_{max}(D^T D) \leq 1$ (where λ_{max} is the maximum eigenvalue). Otherwise we solve the equivalent minimization problem

$$\min_{\boldsymbol{\theta}} \|\frac{\mathbf{z}}{\alpha} - \frac{D}{\alpha}\boldsymbol{\theta}\|^2 + \frac{n\nu}{\alpha} \sum_{\gamma=0}^{q} \sum_{j=0}^{J-1} \sum_{m=1}^{M_j} \theta_{\gamma,j,m} \quad \text{subject to} \quad \theta_{\gamma,j,m} \geq 0$$

where the positive constant α ensures that $\lambda_{max}(D^T D) \leq 1$.

4.2. Stopping conditions

IPS algorithm is an iterative procedure which produces a sequence of solutions $\boldsymbol{\theta}^{[0]}, \boldsymbol{\theta}^{[0]}, \ldots, \boldsymbol{\theta}^{[p]}$ converging to the optimal solution $\boldsymbol{\theta}^*$. There is a need to stop the algorithm when the solution $\boldsymbol{\theta}^{[p]}$ is sufficiently close to the optimal solution $\boldsymbol{\theta}^*$. Several stopping conditions have been proposed in the literature (for example Defrise and De Mol (1987)). We choose to use a stopping condition based on the KKT conditions, which are easy to evaluate. This $\epsilon-$KKT conditions are defined as

$$\mathbf{d}_{\gamma,j,m}^T(\mathbf{Y} - D\boldsymbol{\theta}^*) = \nu - \epsilon, \text{if } \theta_{\gamma,j,m}^* \neq 0$$
$$\mathbf{d}_{\gamma,j,m}^T(\mathbf{Y} - D\boldsymbol{\theta}^*) \leq \nu - \epsilon, \text{if } \theta_{\gamma,j,m}^* = 0$$

where $\epsilon > 0$ is a constant which defines the precision of the solution.

5. Simulations

In this section we will study the empirical performance of WK-ANOVA, in terms of prediction accuracy. The measure of the prediction accuracy is given by Q_2 which is defined as

$$Q_2 = 1 - \frac{\sum_{i=1}^{n_{test}}(y_i - \hat{f}(\mathbf{x}_i))^2}{\sum_{i=1}^{n_{test}}(y_i - \bar{y})^2}, \text{with } n_{test} = 500 \tag{23}$$

where y_i denotes the ith test observation of the test set, $\bar{y}$ is their empirical mean and $\hat{f}(x_i)$ is the predicted value. We compare the obtained results with those obtained by COSSO and Gaussian Process (GP) (Busby, 2009). We also compare the methods for different experimental design sizes, uniformly distributed on $[0,1]^d$ and built by Latin Hypercube Design procedure (McKay et al., 1979) with maximin criterion (Santner et al., 2003) (maximinLHD). Moreover, different signal to noise ratio were applied $SNR \equiv 1:3$ (high noise) $SNR \equiv 1:7$ (medium noise) and $SNR \equiv \infty$ (without noise), with $SNR = [Var(f(X))]/\sigma^2$. For each setting of test examples 2 and 3, we perform 50 times the test.
We fixed $M_j = 2^j - 1$ for $\gamma = 1, \ldots, d$ and $M_j = 1$ for $\gamma > d$, in other words we penalize by the translation parameter k for the main effects and by resolution j for the interaction. This assumption permits us reducing significantly the computational time. The wavelets used in our tests were Daubechies wavelets with 3 vanishing moments (Daubechies, 1992).
The WK-ANOVA was implemented using R. We run the simulations on a computer operated by 32 bits-Windows OS, this latter imposes limits on the total memory allocation. Knowing that the storage of the matrices $K_{\gamma,j,m}^{\Gamma}$ is memory consuming, we limit the dimension to our examples to 8 and the sample size to estimate Q_2 to 500. To fit COSSO models we have used a modified version of the original Matlab code provided by Yi Lin and Hao Helen Zhang. We made our modification of this algorithm to make it faster. The GP code was implemented using R with a generalized power exponential Family (Busby, 2009).

5.1. Example 1

Let's consider an additive model with $\mathbf{X} \in [0,1]^6$, with the following function

$$f(\mathbf{X}) = g_1(X^{(1)}) + g_2(X^{(2)}) + g_3(X^{(3)}) + g_4(X^{(4)})$$

where

$$g_1(t) = t; \quad g_2(t) = (2t-1)^2; \quad g_3(t) = \frac{\sin(2\pi t)}{2 - \sin(2\pi t)};$$

$$g_4(t) = 0.1\sin(2\pi t) + 0.2\cos(2\pi t) + 0.3\sin^2(2\pi t) + 0.4\cos^3(2\pi t) + 0.5\sin^3(2\pi t)$$

Therefore $X^{(5)}, X^{(6)}$ are uninformative. We use an experimental design of size $n = 200$, built by maxminLHD, and $SNR \equiv \infty$. Figure 1 gives the plot of data observation with the true ANOVA component f_l and theirs WK-ANOVA estimates against inputs $X^{(l)}$, $l = 1, \ldots, 6$. The Q_2 of this WK-ANOVA estimate is equal to 0.96 which is a good performance. However, we can note that the estimation of the linear function component f_1 does suffer from using a wavelet method. Part of the reason is the boundary effects caused by using periodic wavelets.

5.2. Example 2

In this first test case, consider an additive model with $X = [0, 1]^8$, with the following function

$$f(\mathbf{X}) = g_1(X^{(1)}) + g_2(X^{(2)}) + g_3(X^{(3)}) + g_4(X^{(4)}) + \epsilon$$

where

$$\begin{aligned}
g_1(t) &= 0.1\sin(2\pi t) + 0.2\cos(2\pi t) + 0.3\sin(2\pi t) + 0.4\cos^3(2\pi t) + 0.5\sin^3(2\pi t) \\
g_2(t) &= (2t-1)^2 \\
g_3(t) &= |\sin(3\pi t)| + \frac{0.5|\sin(5\pi t)|}{2 - \sin(4\pi t)} \\
g_4(t) &= \frac{|\sin(2\pi t)|}{2 - \sin(2\pi t)}
\end{aligned}$$

Therefore $X^{(5)}, \ldots, X^{(8)}$ are uninformative. Note that all informative input of f have nonlinear response, in addition g_3 and g_4 have discontinuities on the derivative. The true ANOVA components f_l, $l = 1, \ldots, n$ with their WK-ANOVA, COSSO and GP estimates are given in figure 2. These estimates were built with an experimental design of size $n = 200$ and with noise ratio $SNR = 3$. The WK-ANOVA has more fidelity to the reality than COSSO and GP especially for the components f_3 and f_4. Indeed, WK-ANOVA captures more the discontinuities of the components f_3 and f_4. This good fit is due to the properties of wavelets analysis. In other words, our algorithm based on wavelets is well suited to this type of functions (with discontinuities on the derevatives).

We run the simulation 50 times for different sizes of experimental design ($n = 50, 100, 200$) and different signal to noise ratio $SNR \equiv 1:3$, $SNR \equiv 1:7$ and $SNR \equiv \infty$. The results are summarized in figure 3 each panel is a boxplot of the 50 estimations of Q_2. As expected, the accuracy of WK-ANOVA estimates increases when the sample size raise. We can see that WK-ANOVA procedure outperforms COSSO and GP in all the studied settings. Moreover, even though there are much more parameters to estimate with WK-ANOVA comparing to COSSO, this procedure does not seem to suffer from small sample size effect. For this example, WK-ANOVA has shown better denoising properties and better predictivity. In addition, for $n = 100$ and $n = 200$ WK-ANOVA is the most robust.

5.3. Example 3

Consider the g-Sobol function, which is strongly nonlinear and is described by a non-monotonic relationship. This function is a well-known test case in the studies of global sensitivity analysis. Figure 4 illustrates the g-Sobol function against the two most influential

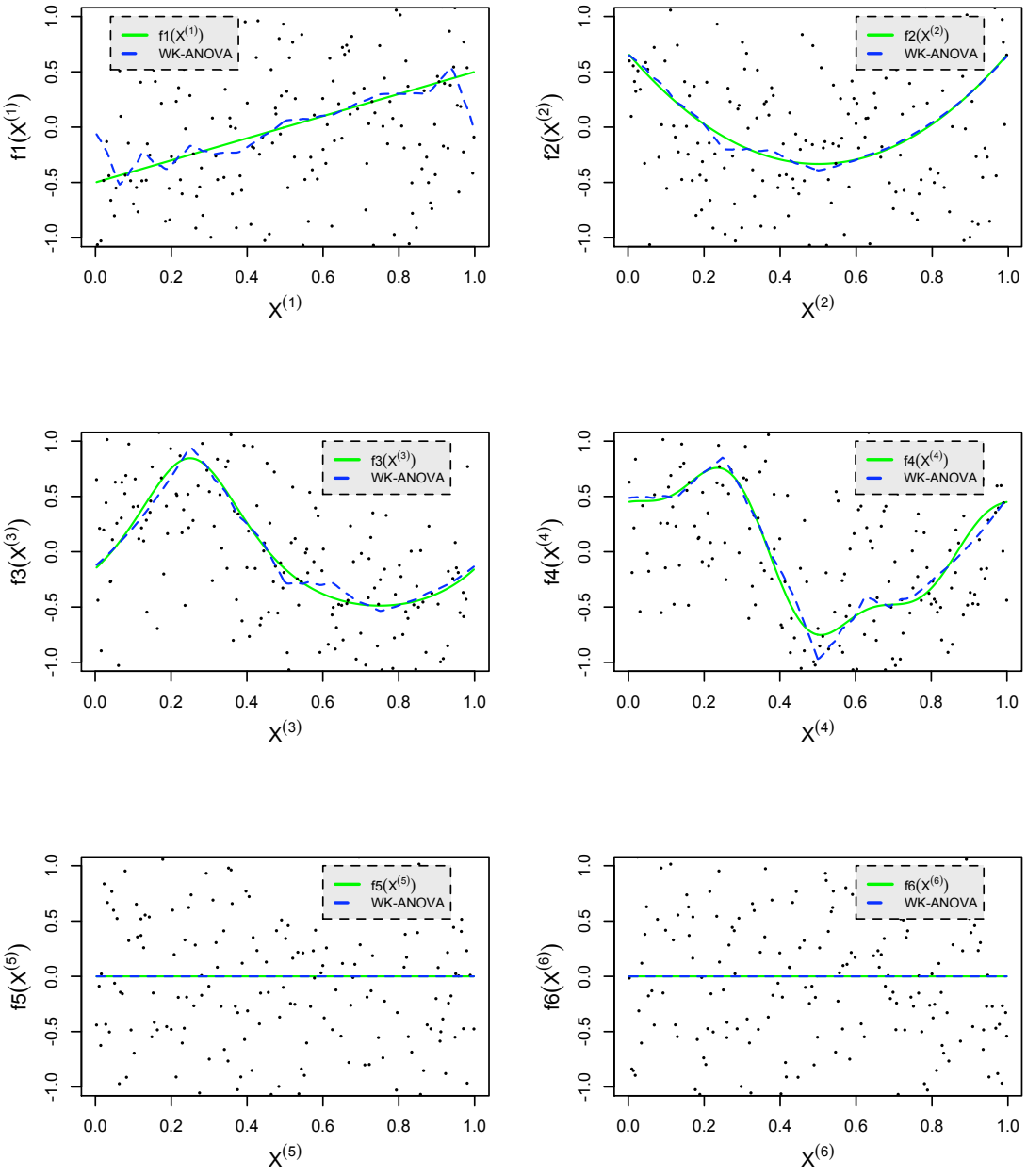

Figure 1: Plot of the six true functional components, f_l, $l = 1, \ldots, 4$ along with the data observations and theirs estimates given by WK-ANOVA for a realization from example 1

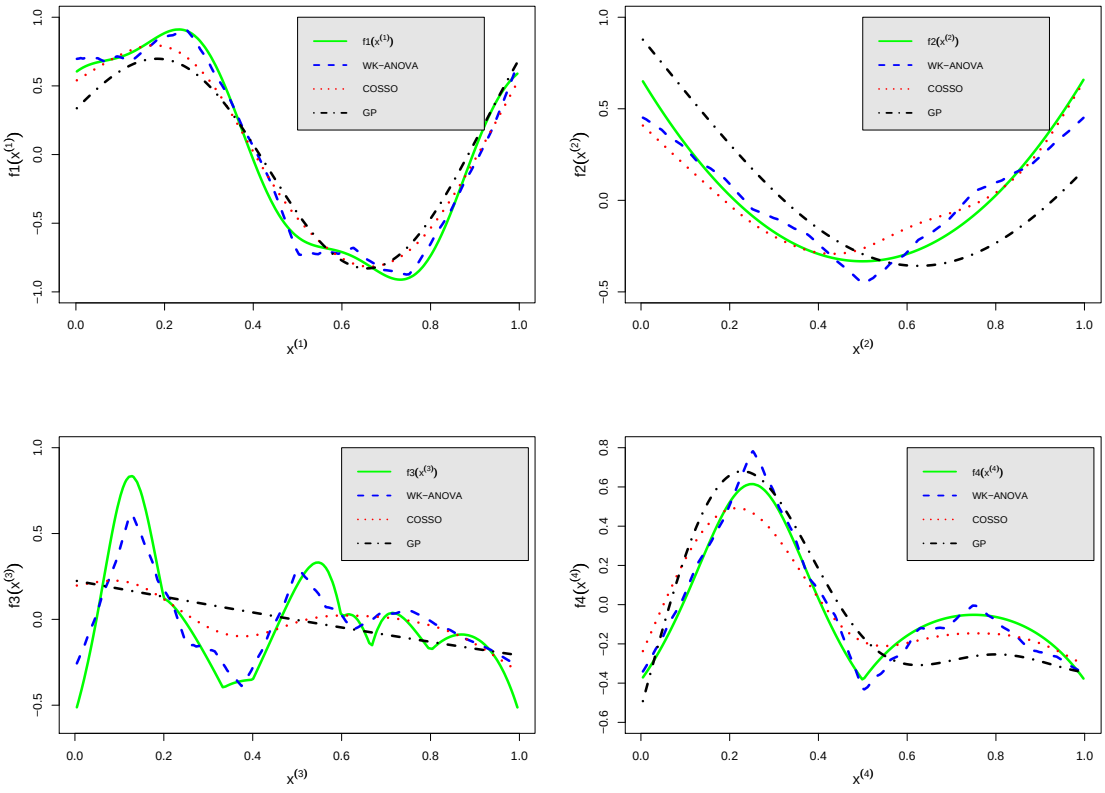

Figure 2: Plot of f_l, $l = 1, \ldots 4$ along with theirs estimates given by WK-ANOVA, COSSO and GP for a realization from example 2

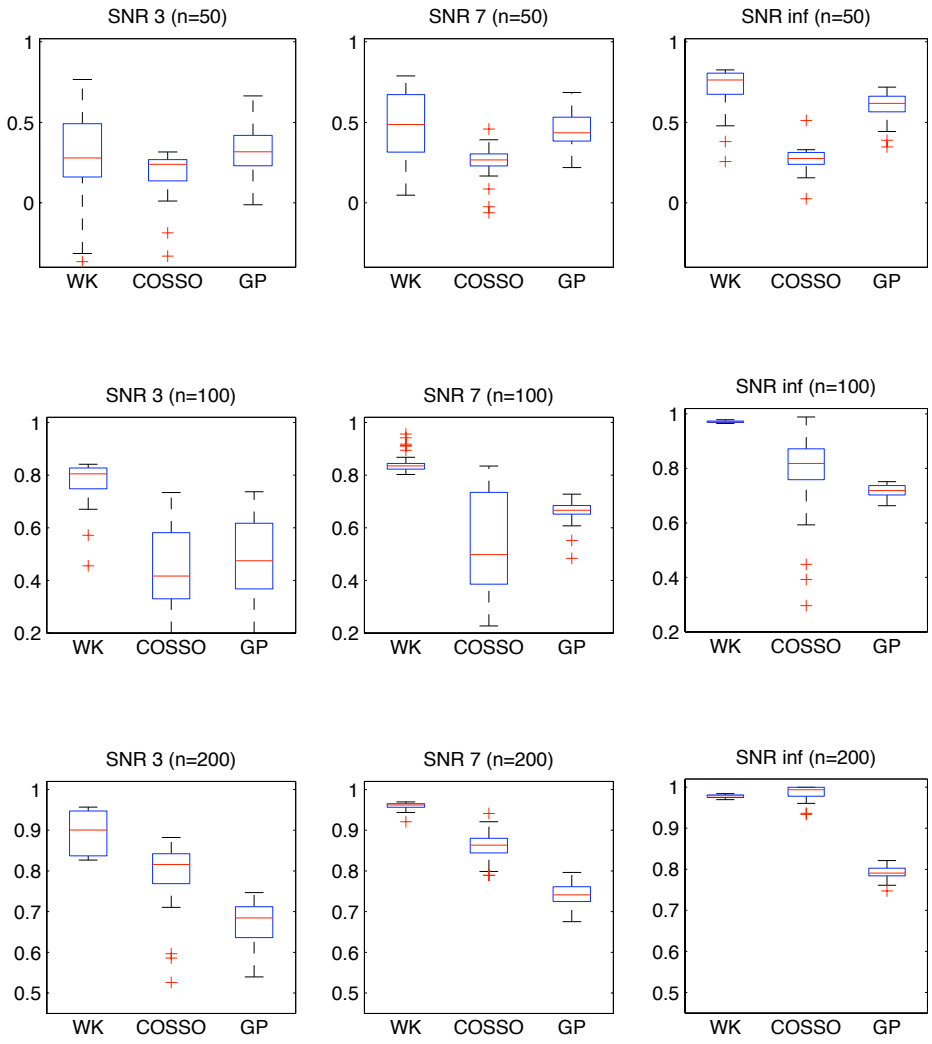

Figure 3: Q_2 results from example 2

parameters $X^{(1)}$ and $X^{(2)}$. The g-Sobol function (Saltelli et al. (2000)) is defined for 8 inputs factors as

$$g_{\text{Sobol}}(X^{(1)}, \ldots, X^{(8)}) = \prod_{k=1}^{8} g_k(X^{(k)}) \quad \text{with} \quad g_k(X^{(k)}) = \frac{\left|4X^{(k)} - 2\right| + a_k}{1 + a_k}$$

where $\{a_1, \ldots, a_8\} = \{0, 1, 4.5, 9, 99, 99, 99, 99\}$. The contribution of each input $X^{(k)}$ to the variability of the model output is represented by the weighting coefficient a_k. The lower this coefficient a_k, the more significant the variable $X^{(k)}$. For example:

$$\begin{cases} a_k = 0 \rightarrow x^{(k)} \text{ is very important,} \\ a_k = 1 \rightarrow x^{(k)} \text{ is relatively important,} \\ a_k = 4.5 \rightarrow x^{(k)} \text{ is poorly important,} \\ a_k = 9 \rightarrow x^{(k)} \text{ is non important,} \\ a_k = 99 \rightarrow x^{(k)} \text{ is non significant.} \end{cases}$$

We run the simulation 50 times for different sizes of experimental design ($n = 50, 100, 200$)

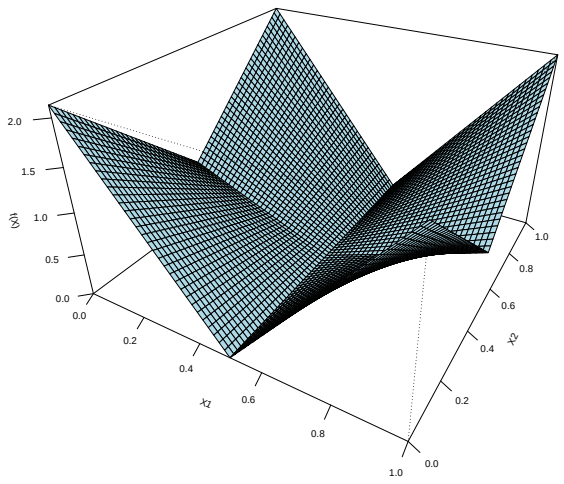

Figure 4: Plot of g-Sobol function versus inputs $X^{(1)}$ and $X^{(2)}$ with other inputs fixed at 0.5

and different signal to noise ratio $SNR \equiv 1:3$, $SNR \equiv 1:7$ and $SNR \equiv \infty$. The results are summarized in figure 5 each panel is a boxplot of the 50 estimations of Q_2. We can see that for all the tested experimental design sizes and noise ratios WK-ANOVA outperforms COSSO and GP. Moreover, the accuracy of the prediction is very good ($\bar{Q}_2 = 0.98$, where $\bar{Q}_2$ is the average of the Q_2's) even with $n = 50$ for the setting without noise, when a design with $n = 200$ design is necessary to perform a response surface with $\bar{Q}_2 = 0.94$ for GP and with $\bar{Q}_2 = 0.90$ for COSSO. Clearly, for this example WK-ANOVA has the best results in term of predictivity, denoising property and robustness.

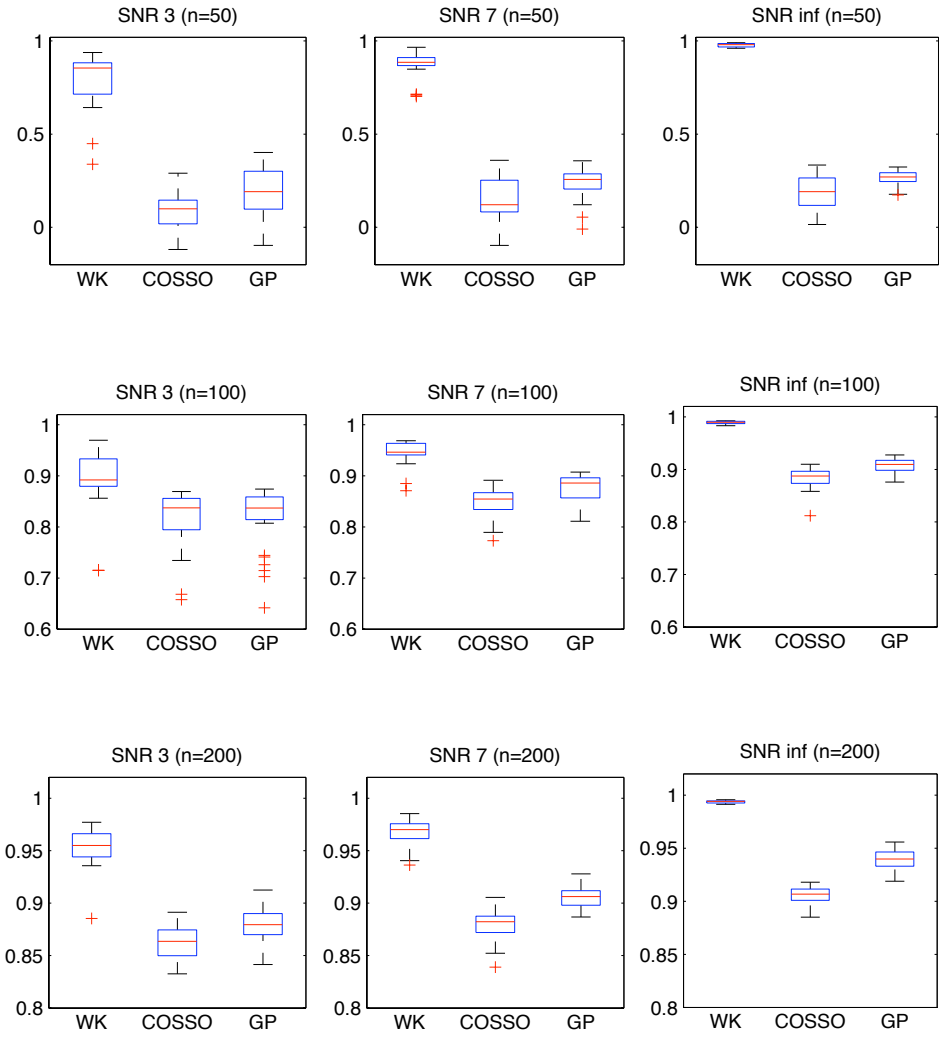

Figure 5: Q_2 results from example 3

6. Conclusion

In this article, we introduced a new regularized nonparametric regression method that we named WK-ANOVA. Differently than other wavelet methods, WK-ANOVA does require neither an equispaced experimental design points, nor a dyadic size of data.
For the tested analytical examples which contains some discontinuities, WK-ANOVA outperforms COSSO and GP. However, in example 1 the wavelet methods have undesired boundary effect in the case of the estimation of nonperiodic function, which results from the use of periodic wavelets. One future investigation topic is to use boundary adapted wavelets.

Proofs

Proof of Theorem 3.1. Consider the following decomposition of the wavelet-based RKHS $\mathcal{F}$

$$\mathcal{F} = \{1\} \oplus \mathcal{H}_q$$

where $\mathcal{H}_q = \bigoplus_{\gamma=1}^{q} \mathcal{F}_\gamma$. Denote by $A(f)$ the functional to be minimized in (10). $A(f)$ is convex and continuous. By inequality (11) we have $R_q(f) \geq \|f\|_{\mathcal{F}}$ for any $f \in \mathcal{H}_q$. Let $K_{\mathcal{H}_q}$ be the reproducing kernel on $\mathcal{H}_q$ and $\langle \cdot, \cdot \rangle_{\mathcal{H}_q}$ be the inner product of $\mathcal{H}_q$. Denote by $e_n = max_{i=1}^{n}(K_{\mathcal{H}_q})^{(1/2)}(x_i, x_i)$. By the definition of the reproducing kernel and the properties of Γ, we have for any $f \in \mathcal{H}_q$ and $i = 1, ..., n$

$$|f(x_i)| = |\langle f, K_{\mathcal{H}_q(x_i,\cdot)} \rangle_{\mathcal{H}_q}| \leq e_n \|f\|_{\mathcal{F}} \leq e_n R_q(f)$$

Let consider D a closed, convex and bounded set defined as:

$$D = \{f \in \mathcal{F}; f = b + f_1, with\ b \in \{1\}, f_1 \in \mathcal{H}_q, R_q(f) \leq v, |b| \leq v^{1/2} + (e_n + 1)v\}$$

where $v = max_i\{y_i^2 + |y_i| + 1\}$. Therefore by the theorem 4 of Tapia and Thomson (1978). There exist a minimizer $\overline{f}$ of (10) in D and $A(\overline{f}) \leq A(0) \leq v$.
On the other hand, for any $f \in \mathcal{F}$ with $R_q(f) > v$ clearly $A(f) \geq R_q(f) > v$; for any $f \in \mathcal{F}$, $f = b + f_1$ with $b \in \{1\}$, $f_1 \in \mathcal{H}_q$, $R_q(f) \leq v$ and $|b| > v^{1/2} + (e_n + 1)v$, we therefore have

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

Hence $A(f) > v$, and for any $f \notin D$, we have $A(f) > A(\overline{f})$, wich proves that $\overline{f}$ is the minimizer of (10) in $\mathcal{F}$. □

Proof of Theorem 3.2. The condition on the unknown regression function f_0 are only active for its wavelets coefficients and do not include the V_0 scaling coefficients of f_0. For any $f \in \mathcal{F}$, write $f(\mathbf{x}) = b + f_1(x^{(1)}) + ... + f_d(x^{(d)}) = b + g(\mathbf{x})$, such that $\sum_{i=1}^{n} f_l(x_i^{(l)}) = 0$, $l = 1, ..., d$ and where $b \in \{1\}$ and $g \in \bigoplus_{l=1}^{d} \mathcal{H}_\Gamma^l$. Similarly, write $f_0(\mathbf{x}) = b_0 + g_0(\mathbf{x})$, such that $g_0 \in \bigoplus_{l=1}^{d} \mathcal{H}_\Gamma^l$. By construction $\sum_{i=1}^{n}\{g_0(\mathbf{x}_i) - g(\mathbf{x}_i)\} = 0$, we can write $A(f)$ as :

$$(b - b_0)^2 + \frac{2}{n}(b - b_0)\sum_{i=1}^{n} \varepsilon_i + \frac{1}{n})\sum_{i=1}^{n}(g_0(\mathbf{x}_i) + \varepsilon_i - g(\mathbf{x}_i))^2 + \lambda_n^2 R_q(g)$$

Therefore, the minimizing $\hat{b}$ is $\hat{b} = b_0 + 1/n \sum_{i=1}^n \varepsilon_i$, which shows that $\hat{b}$ converges towards b_0 at rate $n^{-1/2}$. On the other hand, $\hat{g}$ must minimize over $\bigoplus_{l=1}^d \mathcal{H}_\Gamma^l$ the functional

$$\frac{1}{n}\sum_{i=1}^n \{g_0(\mathbf{x}_i) + \varepsilon_i - g(\mathbf{x}_i)\}^2 + \lambda_n^2 R_q(g)$$

Let $\mathcal{G} = \{g \in \mathcal{F} : g(x) = f_1(x^{(1)}) + ... + f_d(x^{(d)}), with \ \sum_{i=1}^n f_l(x_i^{(l)}) = 0, l = 1, ..., d\}$. then $g_0 \in \mathcal{G}$ and $\hat{g} \in \mathcal{G}$. The conclusion of theorem 2 follows from the following Lemma.

Lemma Appendix .1. *(Theorem* 10.2 *of Van De Geer (2000), lemma* 5.1 *of Amato et al. (2006) and lemma* 3 *of Lin and Zhang (2006))*
Let $H_\infty(\delta, \mathcal{G})$ be the $\delta-$entropy of $\mathcal{G}$ for the supremum norm. Then

$$H_\infty(\delta, \{g \in \mathcal{G} : R_q(g) \leq 1\} \leq A d^{(s+1)/s} \delta^{-1/s},$$

for all $\delta > 0$, $n \geq 1$, and some $A > 0$ and $0 < 1/s < 2$.

□

Proof of Lemma Appendix .1. Define $\mathcal{G}^l$ as the set of univariate function of $x^{(l)}$.

$$\mathcal{G}^l = \{f_l \in \mathcal{H}_\Gamma^l : R_q(f_l) \leq 1, \sum_{i=1}^n f_\gamma(x_i)^{(l)} = 0\}$$

It follows from Lemma 5.1 of (Amato et al., 2006) that

$$H_\infty(\delta, \mathcal{G}^l) \leq A\delta^{-1/s}$$

for all $\delta > 0$, and $n \geq 1$, some $A > 0$ and $0 < 1/s < 2$. By definition of $\mathcal{G}$ we see that in terms on the supreme norm, if each $\mathcal{G}^l$, $l = 1, .., d$ can be covered by N balls of radius δ, then the set $\{g \in \mathcal{G} : R_q(g) \leq 1\}$ can be covered byN^d balls with radius $d\delta$, and we get :

$$H_\infty(d\delta, \{g \in \mathcal{G} : R_q(g) \leq 1\}) \leq Ad\delta^{-1/s}$$

□

Proof of Lemma 3.3. For any $f \in \mathcal{F}$, write $f = b + \sum_{\gamma=1}^q \sum_{j=0}^{J-1} \sum_{k=0}^{2^j-1} f_{\gamma,j,k}$ with $b \in \{1\}$ and $f_{\gamma,j,k} \in \mathcal{W}_{\gamma,j,k}^\Gamma$. Let the projection of $f_{\gamma,j,k}$ onto $span\{K_{\gamma,j,k}^\Gamma(\mathbf{x}_i, \cdot), i = 1, ..., n\} \subset \mathcal{W}_{\gamma,j,k}^\Gamma$ be denoted by $\alpha_{\gamma,j,k}$ and the orthonormal complement by $\beta_{\gamma,j,k}$. Then $f_{\gamma,j,k} = \alpha_{\gamma,j,k} + \beta_{\gamma,j,k}$ and (10) can be written as

$$\frac{1}{n}\sum_{i=1}^n \{y_i - b - \sum_{\gamma=1}^q \sum_{j=0}^{J-1} \sum_{k=0}^{2^j-1} \langle K_{\gamma,j,k}^\Gamma(\mathbf{x}_i, \cdot), \alpha_{\gamma,j,k}\rangle\}^2 + \lambda^2 \sum_{\gamma=1}^q \sum_{j=0}^{J-1} \sum_{k=0}^{2^j-1} (\|\alpha_{\gamma,j,k}\|^2 + \|\beta_{\gamma,j,k}\|^2)^{1/2}$$

Therefore any minimizing f must be such that $\beta_{\gamma,j,k} = 0$, and the result follows immediately.
□

Proof of Lemma 3.4. Denote the functional in (14) by $B(\theta, f)$. For any $\gamma = 1, ..., q; j = 0, ..., J-1; k = 1, ..., 2^j - 1$, we have

$$\lambda_0 \theta_{\gamma,j,k}^{-1} \|P_{\gamma,j,k}^{\Gamma} f\|_{\mathcal{W}_{\gamma,j,k}^{\Gamma}}^2 + \nu \theta_{\gamma,j,k} \geq 2\lambda_0^{1/2} \nu^{1/2} \|P_{\gamma,j,k}^{\Gamma} f\|_{\mathcal{W}_{\gamma,j,k}^{\Gamma}} = \lambda^2 \|P_{\gamma,j,k}^{\Gamma} f\|_{\mathcal{W}_{\gamma,j,k}^{\Gamma}}.$$

for any $\theta_{\gamma,j,k} \geq 0$ and $f \in \mathcal{F}$, and the equality holds if and only if $\theta_{\gamma,j,k} = \lambda_0^{1/2} \nu^{-1/2} \|P_{\gamma,j,k}^{\Gamma} f\|_{\mathcal{W}_{\gamma,j,k}^{\Gamma}}$. Therefore $B(\theta, f) \geq A(f)$, where $A(f)$ denote the functional of (10) for any $\theta_{\gamma,j,k} \geq 0$, $\gamma = 1, ..., q$; $j = 0, ..., J-1$; $k = 1, ..., 2^j - 1$ and $f \in \mathcal{F}$, with the equality holds only if $\theta_{\gamma,j,k} = \lambda_0^{1/2} \nu^{-1/2} \|P_{\gamma,j,k}^{\Gamma} f\|_{\mathcal{W}_{\gamma,j,k}^{\Gamma}}$. The conclusion then follows.

□

Proof of Theorem 4.1. The orthogonal projection $\mathcal{P}_{\Omega} x$ of x onto Ω^+ is characterized by the following useful inequality: for all $a \in \Omega^+$ and all x we have

$$\langle a - \mathcal{P}_{\Omega^+} x, \mathcal{P}_{\Omega^+} x - x \rangle \geq 0 \tag{.1}$$

From the inequality (.1) we can say that for any Ω^+, x and z, we have

$$\langle \mathcal{P}_{\Omega^+} z - \mathcal{P}_{\Omega^+} x, \mathcal{P}_{\Omega^+} x - x \rangle \geq 0 \text{ and } \langle \mathcal{P}_{\Omega^+} z - \mathcal{P}_{\Omega^+} x, z - \mathcal{P}_{\Omega^+} z \rangle \geq 0 \tag{.2}$$

Adding, we obtain

$$\langle \mathcal{P}_{\Omega^+} z - \mathcal{P}_{\Omega^+} x, z - x \rangle \geq \|\mathcal{P}_{\Omega^+} z - \mathcal{P}_{\Omega^+} x\|^2 \tag{.3}$$

From the Cauchy inequality we conclude that

$$\|\mathcal{P}_{\Omega^+} x - \mathcal{P}_{\Omega^+} z\| \leq \|x - z\| \tag{.4}$$

Let E be nonempty set of all $\boldsymbol{\theta} \in \Omega^+$ at which the functional (18) attains its minimum value over Ω^+ and $\boldsymbol{\theta}^*$ a member of E. Then $\boldsymbol{\theta}^* = \mathcal{P}_{\Omega^+}(\delta_{\nu}^{Soft}(\boldsymbol{\theta}^*))$ and

$$\|\boldsymbol{\theta}^* - \boldsymbol{\theta}^{[p+1]}\| = \|\mathcal{P}_{\Omega^+}(\delta_{\nu}^{Soft}(\boldsymbol{\theta}^*)) - \mathcal{P}_{\Omega^+}(\delta_{\nu}^{Soft}(\boldsymbol{\theta}^{[p]}))\| \leq \|\delta_{\nu}^{Soft}(\boldsymbol{\theta}^*) - \delta_{\nu}^{Soft}(\boldsymbol{\theta}^{[p]})\| \tag{.5}$$

The convergence of the IPS algorithm follows from the following Lemma.

Lemma Appendix .2. *(Lemma 3.4 of Daubechies et al. (2004))*
$\mathcal{S}_{\nu}$ is nonexpansive, i.e., for all x and $z \in \mathbb{R}$,

$$\|\delta_{\nu}^{Soft}(x) - \delta_{\nu}^{Soft}(z)\| \leq \|x - z\| \tag{.6}$$

Since (.9) is convex the Karush-Kuhn-Tucker theorem suggests that a necessary and sufficient condition for $\boldsymbol{\theta}^*$ to be the solution of model (18) is that there is $\lambda \geq 0$ such that, for any $\gamma = 1, ..., q$, $j = 0, ..., J-1$, $m = 1, ..., M_j$

$$\{-\mathbf{d}_{\gamma,j,m}^T (\mathbf{Y} - D\boldsymbol{\theta}^*) + \nu\} \theta_{\gamma,j,m}^* = 0 \tag{.7}$$

$$-\mathbf{d}_{\gamma,j,m}^T (\mathbf{Y} - D\boldsymbol{\theta}^*) + \nu \ \geq 0 \tag{.8}$$

$$\theta_{\gamma,j,m}^* \geq 0 \tag{.9}$$

There are two possible value for $\boldsymbol{\theta}^{[p+1]}$ in (22) :

$$\theta_{\gamma,j,m}^{[p+1]} = \begin{cases} 0 \\ \theta_{\gamma,j,m}^{[p]} + \mathbf{d}_{\gamma,j,m}^{T}(\mathbf{Y} - D\boldsymbol{\theta}^{[p]}) - \nu \end{cases} \tag{.10}$$

It is not difficult to show that the KKT conditions are satisfied when $\theta_{\gamma,j,m}^{[p+1]} = 0$ and that the condition (.9) is always satisfied. Now we consider the second possibility. If $\theta_{\gamma,j,m}^{*}$ is a fixed point of the map T, with $Tx = \mathcal{P}_{\Omega^+}(\mathcal{S}_\nu(x))$, that is, $T\theta_{\gamma,j,m}^{*} = \theta_{\gamma,j,m}^{*}$. By (.10), we have

$$d_{\gamma,j,m}^{T}(Y - D\boldsymbol{\theta}^{*}) - \lambda = 0$$

The conclusion follows immediately. □

Acknowledgements

The authors are grateful to Professor Anestis Antoniadis for many useful suggestions and helpful discussions.

Amato, U., Antoniadis, A., Pensky, M., 2006. Wavelet kernel penalized estimation for non-equispaced design regression. Statistics and Computing 16, 37–56.

Antoniadis, A., Bigot, J., Sapatinas, T., 2001. Wavelet estimators in nonparametric regression: a comparative simulation study. Journal of Statistical Software 6.

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

Antoniadis, A., Gregoire, G., Vidal, P., 1997. Random design wavelet curve smoothing. Statistics and Probability Letters 35, 225–232.

Breiman, L., 1995. Better subset regression using the nonnegative garrote. Technometrics 37, 373–384.

Busby, D., 2009. Hierarchical adaptive experimental design for gaussian process emulators. Reliability Engineering and System Safety 94, 1183–1193.

Daubechies, I., 1992. Ten lectures on wavelets. SIAM.

Daubechies, I., Defrise, M., De Mol, C., 2004. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. Comm. Pure. Appl. Math 57 (11), 1413–1457.

Defrise, M., De Mol, C., 1987. A note on stopping rules for iterative regularization methods and filtered svd. P. C. Sabatier (Ed.), Inverse Problems: An interdisciplinary Study, pp. 261–268.

Figueiredo, M. A. T., Nowak, R. D., 2003. An em algorithm for wavelet-based image restoration. IEEE Transactions on Image Processing 12, 906–916.

Gunn, S., Kandola, J., 2002. Structural modeling with sparse kernels. Machine Learning 48, 115–136.

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

Kovac, A., Silverman, B. W., 2000. Extending the scope of wavelet regression methods by coefficient-dependent thresholding. Journal of American Statistical Association 95, 172–183.

Landweber, L., 1951. An iterative formula for fredholm integral equations of the first kind. American journal of mathematics 73, 615–624.

Lin, Y., Zhang, H., 2006. Component selection and smoothing in smoothing spline analysis of variance models. Annals of Statistics 34(5), 2272–2297.

McKay, M. D., Beckman, R. J., Conover, W. J., 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 21, 239–245.

Ogden, R., 1997. Essential wavelets for statistical applications and data analysis. Birkhäuser.

Saltelli, A., Chan, K., Scott, M., 2000. Sensitivity analysis. Wiley.

Santner, T. J., Williams, B. J., Notz, W. I., 2003. The design and analysis of computer experiments. Springer.

Tapia, R., Thomson, J., 1978. Nomparametric 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 Series B 58, 267–288.

Van De Geer, S., 2000. Empirical Processes in M-Estimation. Cambridge University Press.

Vidakovic, B., 1999. Statistical modeling by wavelets. Wiley-Interscience.

Wahba, G., 1990. Spline models for observational data. SIAM.