Robust Estimation and Wavelet Thresholding in Partially Linear Models

Irène Gannaz

ROBUST ESTIMATION AND WAVELET THRESHOLDING IN PARTIAL LINEAR MODELS

Irène Gannaz
Laboratoire de Modélisation et Calcul
Université Joseph Fourier
BP 53 - 38041 Grenoble Cedex 9
France

November 2006

Abstract

This paper is concerned with a semiparametric partially linear regression model with unknown regression coefficients, an unknown nonparametric function for the non-linear component, and unobservable Gaussian distributed random errors. We present a wavelet thresholding based estimation procedure to estimate the components of the partial linear model by establishing a connection between an l_1-penalty based wavelet estimator of the nonparametric component and Huber's M-estimation of a standard linear model with outliers. Some general results on the large sample properties of the estimates of both the parametric and the nonparametric part of the model are established. Simulations and a real example are used to illustrate the general results and to compare the proposed methodology with other methods available in the recent literature.

Keywords: Semi-nonparametric models, partly linear models, wavelet thresholding, backfitting, M-estimation, penalized least-squares.

1 Introduction

Assume that responses $y_1, \ldots, y_n$ are observed at deterministic equidistant points $t_i = \frac{i}{n}$ of an univariate variable such as time and for fixed values $\mathbf{X}_i$, $i = 1, \ldots, n$, of some p-dimensional

explanatory variable and that the relation between the response and predictor values is modeled by a Partially Linear Model (PLM):

$$y_i = \mathbf{X}_i^T \boldsymbol{\beta}_0 + f(t_i) + u_i \qquad i = 1, \ldots, n, \tag{1}$$

where $\boldsymbol{\beta}_0$ is an unknown p-dimensional real parameter vector and $f(\cdot)$ is an unknown real-valued function; the u_i's are i.i.d. normal errors with mean 0 and variance σ^2 and superscript "T" denotes the transpose of a vector or matrix. Given the observed data $(y_i, \mathbf{X}_i)_{i=1,\ldots,n}$, the aim is to estimate from the data the vector $\boldsymbol{\beta}$ and the function f.

The interest in partial linear models has grown significantly within the last decade since their introduction by Engle *et al.* (1986) to analyze in a nonlinear fashion the relation between electricity usage and average daily temperature. Since then the models have been widely studied in the literature. The recent monograph by Hardle *et al.* (2000) provides an excellent survey on the theory and applications of the model in a large variety of fields, such as finance, economics, geology and biology, to name only a few. The advantages of such a model is that it allows an adequate and more flexible handling of the explanatory variables than in linear models and can be also serve as a starting point for dimension reduction by additive modeling. Although there is still lack of general theory on testing the goodness-of-fit of a partial linear model, there are some consistent specification tests such as, for example, those developed by Chen and Chen (1991).

Until now, several methods have been proposed to analyse partially linear models. One approach to estimation of the nonparametric component in these models is based on smoothing splines regression techniques and has been employed in particular by Green and Yandell (1985), Engle *et al.* (1986), Rice (1986), Chen (1987), Chen and Shiau (1991), and Schick (1996) among others. Kernel regression (see e.g. Speckman (1988)) and local polynomial fitting techniques (see e.g. Hamilton and Truong (1997)) have also been used to study partially linear models. An important assumption by all these methods for the unknown nonparametric component $f(t)$ is its high smoothness. But in reality, such a strong assumption may not be satisfied. To deal with cases of a less-smooth nonparametric component, a wavelet based estimation procedure is developed in this paper, and as such it can handle nonparametric estimation for curves lying in Besov spaces instead of the more classical Sobolev spaces.

The estimation method developed in this paper is based on a wavelet expansion of the nonparametric part of the model. The use of an appropriate thresholding strategy on the coefficients allows us to estimate in an adaptive way the nonparametric part with quasi-minimax

asymptotic rates without restrictive assumptions on its regularity. To our knowledge, only few developments in the use of nonlinear wavelet methods in the context of PLM models exist in the literature. Wavelet based estimators for the nonparametric component of a PLM have been investigated by Meyer (2003), Chang and Qu (2004) and by Fadili and Bullmore (2005), more recently. Our results will be compared to the later, since the settings adopted in their work are relatively similar to ours.

One novelty of the estimation procedure proposed in this paper is the link between wavelet thresholding and classical robust M-estimation schemes in linear models with outliers: using soft or hard thresholding or even a SCAD thresholding (see Antoniadis and Fan (2001)) amounts in estimating respectively the unknown vector β_0 of the linear part in the model by Huber's M-estimation or by a truncated mean or by Hampel's estimator. This link allows us to investigate the asymptotic minimax properties of the estimators and to derive second-order approximations for the bias and variance of the resulting estimators of β_0. This is essentially due to the fact that the nonparametric part of the model has a sparse wavelet coefficients representation, and the wavelet coefficients of a PLM in the wavelet domain appear then as outliers in the linear model composed by the linear part.

Furthermore, the above established link of our method with M-estimation theory offers the possibility to use specific M-estimation algorithms for numerically implementing the proposed method, instead of using the *backfitting* technique proposed by Fadili and Bullmore (2005). For our numerical implementation, we will adopt a class of half-quadratic optimization algorithms that have been developed recently for robust image recognition in the pattern recognition literature (see e.g. Charbonnier *et al.* (1997), Dahyot and Kokaram (2004), Vik (2004) and Nikolova and Ng (2005)).

The organization of this paper is as follows: Section 2 briefly recalls some relevant facts about the wavelet series expansion and the discrete wavelet transform that we need further and presents the wavelet decomposition used to model the observed partial linear model. In section 3, we establish the connection between wavelet thresholding estimation for the PLM and M-estimation for a linear model. Section 4 establishes the main properties of our estimators. In Section 5, we discuss the computational algorithms that are used for the numerical implementation of our procedures where we also present a small simulation study to illustrate the finite sample properties of our procedures and to compare them to the backfitting algorithm proposed by Fadili and Bullmore (2005). Proofs of our results are given in Appendix.

2 The partly linear model and its wavelet transform

2.1 THE SETUP

Suppose that y_i ($i = 1, 2, \ldots, n$) is the i-th response of the regression model at point t_i (where t is an index such as time or distance) and can be modelled as

$$y_i = \mathbf{X}_i^T \boldsymbol{\beta}_0 + f(t_i) + u_i, \tag{2}$$

where $\mathbf{X}_i^T$ are given $p \times 1$ vectors of covariate values, $t_i = \frac{i}{n}$ and $\boldsymbol{\beta}_0$ and f are respectively the parametric and nonparametric components of the partial linear model. We will assume hereafter that the noise variables u_i are i.i.d. Gaussian $\mathcal{N}(0, \sigma^2)$ and that the sample size $n = 2^J$ for some positive integer J.

In the nonparametric analysis, the nonparametric part f is modeled as a function lying in an infinite dimensional space. The underlying notion behind wavelet methods is that the unknown function has an economical wavelet expression, i.e. f is, or is well approximated by, a function with a relatively small proportion of nonzero wavelet coefficients. An approach to modelling the nonparametric component of the PLM model, that allows a wide range of irregular effects, is through the sequence space representation of Besov spaces. The (inhomogeneous) Besov spaces on the unit interval, $\mathcal{B}^s_{\pi,r}([0,1])$, consist of functions that have a specific degree of smoothness in their derivatives. The parameter π can be viewed as a degree of function's inhomogeneity while s is a measure of its smoothness. Roughly speaking, the (not necessarily integer) parameter s indicates the number of function's (fractional) derivatives, where their existence is required in an L^π-sense; the additional parameter r is secondary in its role, allowing for additional fine tuning of the definition of the space. For a detailed study on (inhomogeneous) Besov spaces we refer to, e.g., D. L. Donoho and Johnstone (1998). To capture key characteristics of variations in f and to exploit its sparse wavelet coefficients representation, we will assume that f belongs to $\mathcal{B}^s_{\pi,r}([0,1])$ with $s + 1/\pi - 1/2 > 0$. The last condition ensures in particular that evaluation of f at a given point makes sense.

We now briefly recall first some relevant facts about the wavelet series expansion and the discrete wavelet transform that we need further.

2.2 The wavelet series expansion and the discrete wavelet transform

Throughout the paper we assume that we are working within an orthonormal basis generated by dilatations and translations of a compactly supported scaling function, $\varphi(t)$, and a compactly supported mother wavelet, $\psi(t)$, associated with an r-regular ($r \geq 0$) multiresolution analysis of $\left(L^2[0,1], \langle \cdot, \cdot \rangle\right)$, the space of squared-integrable functions on $[0,1]$ endowed with the inner product $\langle f, g \rangle = \int_{[0,1]} f(t)g(t)\,dt$. For simplicity in exposition, we work with periodic wavelet bases on $[0,1]$ (see, e.g., Mallat (1999), Section 7.5.1), letting

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

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

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

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

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

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

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

chosen. In the following we will distinguish the blocs of $W_{n\times n}$ spanned respectively by the scaling functions and the wavelets. The empirical coefficients s_{j_0k} and w_{jk} of $\mathbf{e}$ are given by

$$\begin{aligned} s_{j_0,k} &\approx \frac{1}{\sqrt{n}}\sum_{i=1}^{n} e_i\varphi_{j_0,k}(t_i) \quad \text{for } k=0,\ldots,2^{j_0}-1 \\ w_{j,k} &\approx \frac{1}{\sqrt{n}}\sum_{i=1}^{n} e_i\psi_{j,k}(t_i) \quad \text{for } \begin{cases} j &= j_0,\ldots,J-1, \\ k &= 0,\ldots,2^j-1. \end{cases} \end{aligned}$$

When $\mathbf{e}$ is a vector of function values $\mathbf{F} = (f(t_1),...,f(t_n))^T$ at equally spaced points t_i, the corresponding empirical coefficients s_{j_0k} and w_{jk} are related to their continuous counterparts c_{j_0k} and d_{jk} (with an approximation error of order n^{-1}) via the relationships $s_{j_0k} \approx \sqrt{n}\, c_{j_0k}$ and $w_{jk} \approx \sqrt{n}\, d_{jk}$. Note that, because of orthogonality of $W_{n\times n}$, the inverse DWT (IDWT) is simply given by $\mathbf{F} = W_{n\times n}^{\mathrm{T}}\mathbf{d}$, where $W_{n\times n}^{\mathrm{T}}$ denotes the transpose of $W_{n\times n}$. If $n = 2^J$ for some positive integer J, the DWT and IDWT may be performed through a computationally fast algorithm (see, e.g., Mallat (1999), Section 7.3.1) that requires only order n operations.

We will further use the following notation. For a n-dimensional vector $\mathbf{e}$, its Euclidian (or l_2) norm $\left(\sum_{i=1}^{n} e_i^2\right)^{1/2}$ will be denoted by $\|\mathbf{e}\|$ and the Frobenius norm of a matrix B with general entries $b_{i,j}$ will be denoted by $\|B\| = \left(\sum_{i,j} b_{i,j}^2\right)^{1/2}$.

2.3 A wavelet-based model specification of the PLM model

In matrix notation, the PLM model specified by (2) can be written as

$$\mathbf{Y} = X\boldsymbol{\beta}_0 + \mathbf{F} + \mathbf{U}, \tag{3}$$

where $\mathbf{Y} = \left(y_1,\ldots,y_n\right)^T$, $X^T = \left(\mathbf{X}_1,\ldots,\mathbf{X}_n\right)$ is the $p\times n$ design matrix, and $\mathbf{F} = \left(f(t_1),\ldots f(t_n)\right)^T$. The noise vector $\mathbf{U} = \left(u_1,\ldots,u_n\right)^T$ is a Gaussian vector with mean 0 and variance matrix $\sigma^2 I_n$.

For the model to be asymptotically identifiable, we will assume:

(A1) The vector $\frac{1}{n}X^T\mathbf{F}$ tends to 0 as n goes to infinity.

(A2) The matrix X is full rank, i.e. $\frac{1}{n}X^TX$ converges towards an invertible matrix.

Expressing the vector of coefficients of the linear part as

$$\boldsymbol{\beta}_0 = \left(\frac{1}{n}X^TX\right)^{-1}X^T(\mathbf{Y}-\mathbf{F}-\mathbf{U}),$$

clearly shows that conditions (A1) and (A2) are sufficient to asymptotically ensure the identifiability of the PLM model. As it will be seen in the Appendix, none of these assumptions is restrictive.
Let now $\mathbf{Z} = W_{n\times n}\mathbf{Y}$, $A = W_{n\times n}X$, $\boldsymbol{\theta}_0 = W_{n\times n}\mathbf{F}$ and $\boldsymbol{\varepsilon} = W_{n\times n}\mathbf{U}$. Then premultiplying (1) by W, we obtain the transformed model

$$\mathbf{Z} = A\boldsymbol{\beta}_0 + \boldsymbol{\theta}_0 + \boldsymbol{\varepsilon}. \tag{4}$$

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

3 Soft Thresholding and Huber's M-estimation

The wavelet shrinkage estimators that are classically obtained by hard or soft thresholding can be regarded as an extension of the penalized least squares (PLS) estimator (see Antoniadis and Fan (2001)). We therefore propose estimating the parameters $\boldsymbol{\beta}_0$ and $\boldsymbol{\theta}_0$ in model (4) by penalized least squares. To be specific, our wavelet based estimators will be defined as follows:

$$(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\theta}}_n) = \underset{(\boldsymbol{\beta},\boldsymbol{\theta})}{argmin} \left\{ J_n(\boldsymbol{\beta},\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{1}{2}(z_i - \mathbf{A}_i^T\boldsymbol{\beta} - \theta_i)^2 + \lambda \sum_{i=i_0}^{n} |\theta_i| \right\}, \tag{5}$$

for a given penalty parameter λ, where $i_0 = 2^{j_0} + 1$. The penalty term in the above expression penalizes only the empirical wavelet coefficients of the nonparametric part of the model and not its scaling coefficients. The choice l^1 of the penalty function produces the soft thresholding rule.
The regularization method proposed above is closely related to the method proposed recently by Chang and Qu (2004), but these authors essentially concentrate on the backfitting algorithms involved in the optimization, without any theoretical study of the resulting estimates. The method also relates to the recent one developed by Fadili and Bullmore (2005) where a variety of penalties is discussed. Note, however, that their study is limited to quadratic penalties which amounts essentially in assuming that the underlying function f belongs to some Sobolev space and does not exploit the sparse representation of f.
In order to establish the link with Huber's estimation we will have a closer look at the minimization

of the criterion J_n stated in (5). For a fixed value of $\boldsymbol{\beta}$, the criterion $J_n(\boldsymbol{\beta}, \cdot)$ is minimum at

$$\tilde{\theta}_i(\boldsymbol{\beta}) = \begin{cases} z_i - \mathbf{A}_i^T \boldsymbol{\beta} & \text{if } i < i_0, \\ \text{sign}(z_i - \mathbf{A}_i^T \boldsymbol{\beta}) \left(|z_i - \mathbf{A}_i^T \boldsymbol{\beta}| - \lambda \right)_+ & \text{if } i \geq i_0. \end{cases} \tag{6}$$

Therefore, finding $\hat{\boldsymbol{\beta}}_n$, a solution to problem (5), amounts in finding $\hat{\boldsymbol{\beta}}_n$ minimizing the criterion $J_n(\tilde{\boldsymbol{\theta}}(\boldsymbol{\beta}), \boldsymbol{\beta})$. However, note that

$$J_n(\tilde{\boldsymbol{\theta}}(\boldsymbol{\beta}), \boldsymbol{\beta}) = \sum_{i=i_0}^{n} \rho_\lambda(z_i - \mathbf{A}_i^T \boldsymbol{\beta}) \tag{7}$$

where ρ_λ is Huber's cost functional with threshold λ, defined by:

$$\rho_\lambda(u) = \begin{cases} u^2/2 & \text{if } |u| \leq \lambda, \\ \lambda |u| - \lambda^2/2 & \text{if } |u| > \lambda. \end{cases} \tag{8}$$

The above facts can be derived as follows. Let $i \geq i_0$. Minimizing expression (5) with respect to θ_i is equivalent in minimizing $j(\theta_i) := \frac{1}{2}(z_i - \mathbf{A}_i^T \boldsymbol{\beta} - \theta_i)^2 + \lambda |\theta_i|$. The first order condition for this is: $j'(\theta_i) = \theta_i - (z_i - \mathbf{A}_i^T \boldsymbol{\beta}) + \text{sign}(\theta_i)\lambda = 0$ where j' denotes the derivative of j. Now,

- if $\theta_i \geq 0$, then $j'(\theta_i) = 0$ if and only if $\theta_i = z_i - \mathbf{A}_i^T \boldsymbol{\beta} - \lambda$. Hence, if $z_i - \mathbf{A}_i^T \boldsymbol{\beta} \leq \lambda$, $\theta_i = 0$ and otherwise $\theta_i = z_i - \mathbf{A}_i^T \boldsymbol{\beta} - \lambda$.
- if $\theta_i \leq 0$, $j'(\theta_i)$ is zero if and only if $\theta_i = z_i - \mathbf{A}_i^T \boldsymbol{\beta} + \lambda$; therefore, if $z_i - \mathbf{A}_i^T \boldsymbol{\beta} \geq -\lambda$, $\theta_i = 0$ and otherwise $\theta_i = z_i - \mathbf{A}_i^T \boldsymbol{\beta} + \lambda$.

This proves that for a fixed value of $\boldsymbol{\beta}$, the criterion (5) is minimal for $\tilde{\boldsymbol{\theta}}(\boldsymbol{\beta})$ given by expression (6). If we now replace $\boldsymbol{\theta}$ in the objective function J_n we obtain $J_n(\boldsymbol{\beta}, \tilde{\boldsymbol{\theta}}(\boldsymbol{\beta})) = \frac{1}{2} \sum_{i=i_0}^{n} \left((z_i - \mathbf{A}_i^T \boldsymbol{\beta} - \tilde{\theta}_i)^2 + \lambda |\tilde{\theta}_i| \right)$ since $\tilde{\theta}_i = z_i - \mathbf{A}_i^T \boldsymbol{\beta}$ for $i < i_0$. Now denoting by I the set $I := \left\{ j = i_0, \ldots, n, \quad |z_j - \mathbf{A}_j \boldsymbol{\beta}| < \lambda \right\}$, we find that $J_n(\boldsymbol{\beta}, \tilde{\boldsymbol{\theta}}(\boldsymbol{\beta})) = \frac{1}{2} \sum_{I} (z_i - \mathbf{A}_i^T \boldsymbol{\beta})^2 + \frac{1}{2} \sum_{I^C} \lambda^2 + \lambda \sum_{I^C} \left(|z_i - \mathbf{A}_i^T \boldsymbol{\beta}| - \lambda \right)$ by replacing $\tilde{\theta}_i$ with (6), which is exactly Huber's functional.

The mathematical equivalence of the solution of the two classes of estimation can be stated in the following proposition.

Proposition 1. *If $\hat{\boldsymbol{\beta}}_n$ and $\hat{\boldsymbol{\theta}}_n$ are solutions of the optimization problem (5), then they satisfy*

$$\hat{\boldsymbol{\beta}}_n = \underset{\boldsymbol{\beta}}{argmin} \sum_{i=i_0}^{n} \rho_\lambda(z_i - \mathbf{A}_i^T \boldsymbol{\beta}), \tag{9}$$

$$\hat{\theta}_{i,n} = \begin{cases} z_i - \mathbf{A}_i^T \hat{\boldsymbol{\beta}}_n & \text{if } i < i_0 \\ \gamma_{soft,\lambda}(z_i - \mathbf{A}_i^T \hat{\boldsymbol{\beta}}_n) & \text{if } i \geq i_0, \end{cases}, \quad i = 1, \ldots, n, \tag{10}$$

with ρ_λ being Huber's cost functional defined in (8) and $\gamma_{soft,\lambda}$ the soft-thresholding function with threshold λ defined by $\gamma_{soft,\lambda}(u) = sign(u)\,(|u| - \lambda)_+$.

This result allows the computation of the estimators $\hat{\boldsymbol{\beta}}_n$ et $\hat{\boldsymbol{\theta}}_n$ in a non-iterative fashion. We can estimate the parameter $\boldsymbol{\beta}_0$ directly from the observed data without caring about the nonparametric part of the model by means of eq.(9), and then determine $\hat{\boldsymbol{\theta}}_n$, thence $\hat{\mathbf{F}}_n$ using eq.(10).
The resulting form of the estimators allows us to study their asymptotic properties. Moreover, as we shall see in the simulation section of this paper, another benefit is that we can design estimation algorithms that are much faster than those based on backfitting. Lastly, Propostion 1 leads to a nice interpretation of the estimators.
We may summarize the estimation procedure as follows. Using the observed data $(\mathbf{Y}, X)$:

1. Apply the DWT of order $J = \log_2(n)$ on X and $\mathbf{Y}$ to get their corresponding representation A and $\mathbf{Z}$ in the wavelet domain.

2. The parameter $\boldsymbol{\beta}_0$ is then Huber 's robust estimator which is obtained without taking care of the nonparametric component in the PLM model, given by the optimization problem (9). In other words this amounts in considering the linear model $z_i = \mathbf{A}_i^T \boldsymbol{\beta}_0 + e_i$ with noise $e_i = \theta_{0i} + \varepsilon_i$.

3. The vector $\boldsymbol{\theta}$ of wavelet coefficients of the function f is estimated by soft thresholding of $\mathbf{Z} - A\hat{\boldsymbol{\beta}}_n$, i.e. by equation (10). The estimation of f is then obtained by applying the inverse discrete wavelet transform. Note that this last step corresponds to a standard soft-thresholding nonparametric estimation of f in the model:

$$y_i - \mathbf{X}_i^T \hat{\boldsymbol{\beta}}_n = f(t_i) + v_i, \quad i = 1, \ldots, n,$$

where $v_i = \mathbf{X}_i^T(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}_n) + u_i$.

Remark 1. *The above estimation procedure is in phase with the one advocated by Speckman (1988) who suggests that it is usually preferable to estimate first the linear component in a PLM and to then proceed to the estimation of the nonparametric one. Indeed, we propose to estimate $\boldsymbol{\beta}_0$ and $\mathbf{F}$ by: $\hat{\boldsymbol{\beta}} = (X^T S^T S X)^{-1} S\mathbf{Y}$ and $\hat{\mathbf{F}} = (I - S)(\mathbf{Y} - X\hat{\boldsymbol{\beta}})$, with $S = (I - T)W$, T being the threshold operator. We recognize the exact same form as those of Speckman (1988), differing only on the fact that the smoothing operator S is not anymore linear.*

The wavelet soft-thresholding procedure proposed in this section was derived by establishing the connection between an l_1 based penalization of the wavelet coefficients of f and Huber's M-estimators in a linear model. Other penalties, leading to different thresholding procedures can also be seeing as M-estimation procedures. For example, if γ_λ denotes the resulting thresholding function, we can show in a similar way that the estimators verify

$$\begin{aligned}\hat{\boldsymbol{\beta}}_n &= \underset{\boldsymbol{\beta}}{argmin} \sum_{i=i_0}^{n} \rho_\lambda(z_i - \mathbf{A}_i^T\boldsymbol{\beta}),\\ \hat{\theta}_{i,n} &= \begin{cases} z_i - \mathbf{A}_i^T\boldsymbol{\beta} & \text{if } i < i_0,\\ \gamma_\lambda(z_i - \mathbf{A}_i^T\boldsymbol{\beta}) & \text{if } i \geq i_0, \end{cases}, \quad i = 1,\dots,n,\end{aligned}$$

with ρ_λ being the primitive of $u \mapsto u - \gamma_\lambda(u)$. From what precedes, one sees that hard thresholding corresponds to mean truncation, while SCAD thresholding is associated to Hampel's M-estimation. The above thresholding procedures and the corresponding criteria are illustrated in Figure 1.

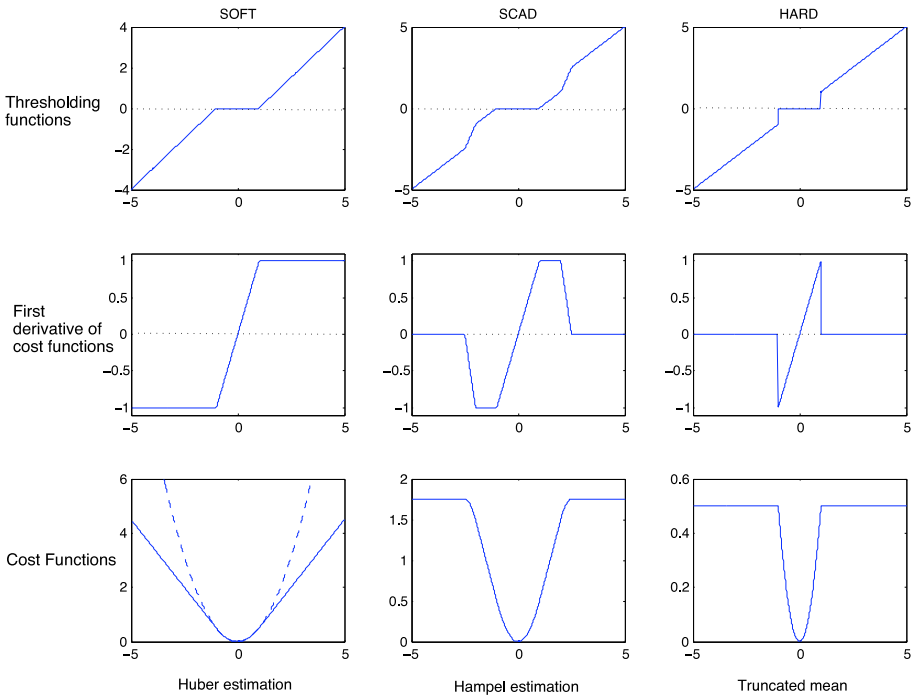


Figure 1: Link between different thresholdings and M-estimation. The dashed line displays the least squares criterion.

However, in this paper, we only concentrate on the properties of estimators obtained by soft thresholding, those corresponding to other rules presenting avenues for further research that hope

will be addressed in the future.

4 Asymptotic properties

Huber's M-estimation was introduced as an alternative to least squares in order to limit the sensitivity of the least-squares estimates to each individual observation. While Huber's M-estimators do not have finite breakdown points, one can show they are quite robust to outliers (see e.g. Hampel *et al.* (1986)). Huber's M-estimation appears therefore a natural approach for robustly fitting the linear part of a PLM, interpreting the wavelet coefficients of the nonparametric part as outliers. In what follows, relying upon this analogy, we study the asymptotic properties of our estimator. To establish our asymptotic results we will require several assumptions to hold.

First a condition which ensures the unicity of $\hat{\boldsymbol{\beta}}_n$ defined in (9):

(A3) The series (K_n) defined by $K_n := \frac{1}{n}\sum_{i=1}^{n} \mathbf{A}_i \mathbf{A}_i^T \rho''_\lambda(\theta_{0i} + \varepsilon_i)$, converges in the L^2-norm towards a non-singular matrix K_0.

The next assumption deals with the structure of the regression design matrix. Since the dicrete wavelet transform W is orthogonal it follows that $A^T A = X^T X$ and, therefore when (A2) holds the matrix $A^T A$ is non-singular for n sufficiently large. Consequently, the projection matrix on the space spanned by the columns of A, say $H = A(A^T A)^{-1} A^T$, has a rank p. In such a case, if $(h_1, \ldots, h_n)$ denotes the diagonal of H, the equality $\sum h_i = p$ holds. With regards to the design, we will also use the assumption:

(A4) The quantity $h := \max_{i=1,\ldots,n} \mathbf{A}_i^T (A^T A)^{-1} \mathbf{A}_i$ tends to 0 when n goes to infinity.

Assumption (A4) is common in a robust regression framework, validating among other things the use of the Lindeberg-Feller criterion. The only difference in our case is that the regression matrix that we consider is the wavelet transformed A rather than X, but the relevant discussion in the Appendix shows that such an assumption is reasonable.

Existing results for semi-parametric partial linear models establish parametric rates of convergence for the linear part and minimax rates for the nonparametric part, showing in particular that the existence of a linear component does not changes the rates of convergence of the nonparametric

component. Within the framework adopted in this paper, the rates of convergence are similar, but an extra logarithmic term will appear in the rates of the parametric part, mainly due to the fact that our smoothness assumptions on the nonparametrric part are weaker. We are now in position to give our asymptotic results.

Theorem 1. *Let $\hat{\boldsymbol{\beta}}_n$ and $\hat{\boldsymbol{\theta}}_n$ be the estimators defined by (9,10) in the model (1). Consider that the penalty parameter λ is the universal threshold: $\lambda = \sigma\sqrt{2\log(n)}$. Under assumptions (A1)– (A4), we have*

$$\begin{aligned} \hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0 &= \bigcirc_{\mathbb{P}}\left(\sqrt{\frac{\log(n)}{n}}\right), \\ \text{and} \qquad \sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) &= K_0^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\rho'_\lambda(\theta_{0i}+\varepsilon_i)\mathbf{A}_i\right) + o_{\mathbb{P}}(\sqrt{\log(n)}). \end{aligned}$$

If in addition we assume that the scaling function φ and the mother wavelet ψ belong to $\mathcal{C}^R$ and that ψ has N vanishing moments, then, for f belonging to the Besov space $\mathcal{B}^s_{\pi,r}$ with $0 < s - 1/2 + 1/\pi$ and $1/\pi < s < \min(R, N)$, we have

$$\|\hat{f}_n - f\|_2 = \bigcirc_{\mathbb{P}}\left(\left(\frac{\log(n)}{n}\right)^{\frac{s}{1+2s}}\right),$$

where $\|\hat{f}_n - f\|_2^2 = \int_0^1(\hat{f}_n - f)^2$.

The Theorem is proved in the Appendix. As noted previously, we lose a factor $\sqrt{\log(n)}$ in the estimation of the vector of parameters $\boldsymbol{\beta}$. The presence of a logarithmic loss lies on the choice of the threshold λ: taking λ which tends to 0, as suggested by Fadili and Bullmore (2005), would lead to a minimax rate in the estimation of $\boldsymbol{\beta}$. The drawback is that the quality of the estimation for the nonparametric part of the PLM would not be anymore quasi-minimax. This phenomenon was put in evidence by Rice (1986): a compromise must be done between the optimality of the linear part estimation with an oversmoothing of the functional estimation and a loss in the linear regression parameter convergence rate but a correct smoothing of the functional part.

The method of estimation that we propose leads to quasi-minimax convergence rates and is applicable for a large class of functions f. An important remark is also that our procedure is adaptative relatively to the regularity of f, thanks to the use of threshold techniques in the wavelet decomposition. Note also that Theorem 1 give a Bahadur's representation of $\hat{\boldsymbol{\beta}}_n$, allowing to elaborate appropriate testing procedures; such inferential problems are out of the scope of the present paper, but interesting for future work.

4.1 Estimation of the variance

Our estimation procedure relies upon knowledge of the variance σ^2 of the noise, appearing in the expression of the threshold λ (recall that we have adopted the universal threshold: $\lambda = \sigma\sqrt{2\log(n)}$). In practice, this variance is unknown and needs to be estimated. One could estimate σ^2 in an iterative way, i.e. with a backfitting algorithm. We propose instead a direct method of estimation based on a QR decomposition of the linear part.

In wavelet approaches for standard nonparametric regression, a popular and well behaved estimator for the unknown standard deviation of the noise is the median absolute deviation (MAD) of the finest detail coefficients of the response divided by 0.6745 (see D. Donoho *et al.* (1995)). The use of the MAD makes sense provided that the wavelet representation of the signal to be denoised is sparse. However, such an estimation procedure cannot be applied without some pretreatment of the data in a partially linear model because the wavelet representation of the linear part of a PLM may be not sparse. Indeed, in practice we have observed that for many partly linear models such a procedure leads to biased estimations.

A QR decomposition on the regression matrix of the PLM allows to eliminate this bias. Since often the function wavelet coefficients at weak resolutions are not sparse, we only consider the wavelet representation at level $J = \log_2(n)$. Let A_J be the wavelet representation of the design matrix X at level J. The QR decomposition ensures that there exist an orthogonal matrix Q and an upper triangular matrix R such that

$$A_J = Q \begin{pmatrix} R \\ 0 \end{pmatrix}.$$

If $\mathbf{Z}_J$, $\boldsymbol{\theta}_{0,J}$ and $\boldsymbol{\varepsilon}_J$ denote respectively the vector of the wavelets coefficients at resolution J of Y, $\mathbf{F}$ and U, model (4) gives

$$Q^T\mathbf{z}_J = \begin{pmatrix} R \\ 0 \end{pmatrix} \boldsymbol{\beta}_0 + Q^T\boldsymbol{\theta}_{0,J} + Q^T\boldsymbol{\varepsilon}_J.$$

It is easy to see that applying the MAD estimation on the last components of $Q^T\mathbf{z}_J$ rather than on $\mathbf{z}_J$ will lead to a satisfactory estimation of σ. Indeed thanks to the QR decomposition the linear part does not appear anymore in the estimation and thus the framework is similar to the one used in nonparametric regression. Following D. Donoho *et al.* (1995), the sparsity of the functional part representation ensures good properties of the resulting estimator.

5 Simulation study

The purpose of this section is to study through simulations several algorithms for estimating the linear part of a PLM model but also to evaluate the performance of the proposed estimators. Our wavelet estimation method for PLM will be also compared with a wavelet backfitting algorithm proposed by Fadili and Bullmore (2005). As already noted, our estimation method allows us to first estimate the linear regression parameter vector $\boldsymbol{\beta}_0$ independently of the nonparametric part, and to then proceed to the estimation of the functional part of the PLM model. The M-estimation $\hat{\boldsymbol{\beta}}_n$ of $\boldsymbol{\beta}_0$ is obtained by means of iterative optimization procedures that are more or less efficient, but usually much faster than backfitting procedures, as we shall see. Before proceeding to the analysis of our simulation results, we briefly recall two particular optimization algorithms that may be used for estimating the linear part.

5.1 Half-quadratic algorithms

The minimization problem we have to solve is of the form:

$$\hat{\boldsymbol{\beta}}_n = \underset{\boldsymbol{\beta}}{argmin}\, J(\boldsymbol{\beta}) \;\text{ with }\; J(\boldsymbol{\beta}) = \sum_{i=1}^{n} \rho_\lambda(z_i - \mathbf{A}_i^T \boldsymbol{\beta}). \tag{11}$$

Minimizers of $J(\boldsymbol{\beta})$ can be obtained using standard optimization tools such as relaxation, gradient, conjugated gradient and so on, but even if the loss function ρ_λ is convex, its second derivative is large near to zero, so the optimization may be slow. For this reason, specialized optimization schemes have been conceived. A very successful approach is *half-quadratic optimization*, proposed in Geman and Reynolds (1992) and Geman and Yang (1995) for cost functions of the above form. The idea is to associate with every $\boldsymbol{\beta}$ in (11) an auxiliary variable $\mathbf{c}$ and to construct an augmented criterion K, such that for every $\mathbf{c}$ fixed, the function $\boldsymbol{\beta} \rightarrow K(\boldsymbol{\beta}, \mathbf{c})$ is quadratic (hence quadratic programming can be used) whereas for every $\boldsymbol{\beta}$ fixed, each $\mathbf{c}$ can be computed independently using an explicit formula. The augmented criterion K is chosen to have the same minimum as J, attained for the same value of $\boldsymbol{\beta}$. The optimization problem of the augmented energy can be solved iteratively. At each iteration one realizes an optimization with respect to $\boldsymbol{\beta}$ for $\mathbf{c}$ fixed and a second with respect to $\mathbf{c}$ for $\boldsymbol{\beta}$ fixed. More precisely, if $\boldsymbol{\beta}^{(m)}$ and $\mathbf{c}^{(m)}$ are the values given after m

iterations, the $(m+1)^{th}$ step of the algorithm actualizes these values through:

$$\begin{aligned} \boldsymbol{\beta}^{(m+1)} &= \underset{\boldsymbol{\beta}}{argmin}\, K(\boldsymbol{\beta}, \mathbf{c}^{(m)}) \\ \mathbf{c}^{(m+1)} &= \underset{\mathbf{c}}{argmin}\, K(\boldsymbol{\beta}^{(m+1)}, \mathbf{c}) \end{aligned} \tag{12}$$

This procedure leads to two algorithms, namely ARTUR and LEGEND, that are also referenced in the literature as IRLS and IMR. We refer to Nikolova and Ng (2005) for some theory on the their use with Huber M-estimation. These algorithms are used for example in robust recognition (see e.g. Charbonnier *et al.* (1997), Dahyot *et al.* (2004) or Vik (2004)). Vik (2004) in particular stresses the link between ARTUR and LEGEND and Huber's approach.

ARTUR

The algorithm described hereafter is referenced as the ARTUR algorithm in the optimization literature or as *Iterative Reweighted Least Squares* (IRLS) in the robustness literature. Geman and Reynolds's theorem leads to an augmented criterion of the form

$$K(\boldsymbol{\beta}, \mathbf{c}) = \sum_{i=1}^{n} c_i (z_i - \mathbf{A}_i^T \boldsymbol{\beta})^2 + \Psi(\mathbf{c}).$$

The auxiliary variable $\mathbf{c}$ corresponds to a weight on the residuals of the least squares fit, thus explaining the IRLS terminology. Intuitively, weights on large residuals have a tendency to eliminate the corresponding responses from the fit. For $\boldsymbol{\beta}$ fixed, the minimum is reached for $c_i = \frac{\rho'_\lambda(r_i)}{r_i}$ where r_i is the ith residual $r_i = z_i - \mathbf{A}_i^T \boldsymbol{\beta}$. At this point the value of Ψ is $\rho_\lambda(r_i) - \rho'_\lambda(r_i) r_i / 2$. The $m+1$ step of the ARTUR algorithm can therefore be described as follows:

$$\begin{cases} r_i^{(m)} &= z_i - \mathbf{A}_i^T \boldsymbol{\beta}^{(m)} \\ c_i^{(m+1)} &= \frac{\rho'_\lambda(2r_i^{(m)})}{2r_i^{(m)}}, \qquad \forall i \in \{1, \ldots, n\} \\ \boldsymbol{\beta}^{(m+1)} &= (A^T \mathbf{c}^{(m+1)} A)^{-1} A^T \mathbf{c}^{(m+1)} \mathbf{Z} \end{cases}$$

LEGEND

LEGEND, or *Iterative Modified Residuals* (IMR), is a slightly different algorithm. The auxiliary variable doesn't weight the residuals anymore but subtracts the larger values of the residuals instead. The existence of the corresponding augmented energy functional follows from the second theorem of Geman and Reynolds (1992). The criterion to be minimized can be written as

$$K(\boldsymbol{\beta}, \mathbf{c}) = \sum_{i=1}^{n} (z_i - \mathbf{A}_i^T \boldsymbol{\beta} - c_i)^2 + \xi(\mathbf{c}).$$

For $\boldsymbol{\beta}$ fixed, the minimum is reached for $c_i = r_i \left(1 - \frac{\rho'_\lambda(r_i)}{2r_i}\right)$ where r_i ith residual $r_i = z_i - \mathbf{A}_i^T \boldsymbol{\beta}$ and at this point the function ξ takes the value $\rho_\lambda(r_i) - \rho'_\lambda(r_i)^2/4$.
With similar notation as for the ARTUR algorithm, the $m+1$ step of the LEGEND algorithm can be described as follows:

$$\begin{cases} r^{(m)} &= \mathbf{Z} - A\boldsymbol{\beta}^{(m)} \\ c_i^{(m+1)} &= r_i^{(m)} \left(1 - \frac{\rho'_\lambda(2r_i^{(m)})}{2r_i^{(m)}}\right) \qquad \forall i \in \{1, \ldots, n\} \\ \boldsymbol{\beta}^{(m+1)} &= (A^T A)^{-1} A^T (\mathbf{Z} - \mathbf{c}^{(m+1)}) \end{cases}$$

Both ARTUR and LEGEND are very easy to program. Nikolova and Ng (2005) show that the risk obtained via the multiplicative form ARTUR is always smaller than the one obtained via the additive form, but the later one is numerically faster. The main reason for this is that under the multiplicative form a matrix inversion is performed within each iteration.

5.2 Numerical simulations

In this subsection, we give some simulation results. All the calculations were carried out in MATLAB 7.0 on a unix environment. For the DWT, we used the WaveLab toolbox developed by Donoho and his collaborators at the Statistics Department of Stanford University (http://www-stat.stanford.edu/~wavelab). For each of the simulated examples in the sequel, we may summarize the various ingredients of our fitting procedure as follows:

1. Application on the observed data of the discrete wavelet transform (DWT) using the pyramidal algorithm of Mallat (1989);
2. Estimation of the variance σ^2 by means of a QR decomposition on the matrix of wavelet coefficients at maximal resolution followed by a MAD estimation;
3. Estimation of $\boldsymbol{\beta}_0$ with ARTUR or LEGEND, solving (9);
4. Estimation of $\boldsymbol{\theta}_0$ by soft thresholding of $\mathbf{Z} - A\hat{\boldsymbol{\beta}}_n$, given by (10);
5. Finally, estimation of $\hat{f}_n$ by applying the inverse DWT on $\hat{\boldsymbol{\theta}}_n$.

We will compare with Fadili and Bullmore's procedure that estimates conjointly $\boldsymbol{\beta}_0$ and $\boldsymbol{\theta}_0$ using a backfitting algorithm.
In order to reduce the number of iterations, we have used a stopping criterion in both ARTUR

et LEGEND: while fixing a larger upper bound for the total number of iterations allowed, we also consider that the algorithm has converged as soon as the difference between two successive iterations is smaller than some given threshold δ. More precisely, the iterations are stopped as soon as $\frac{\|\boldsymbol{\beta}^{(m+1)}-\boldsymbol{\beta}^{(m)}\|_2}{\|\boldsymbol{\beta}^{(m)}\|_2} < \delta$ or whenever we attain their upper limit.

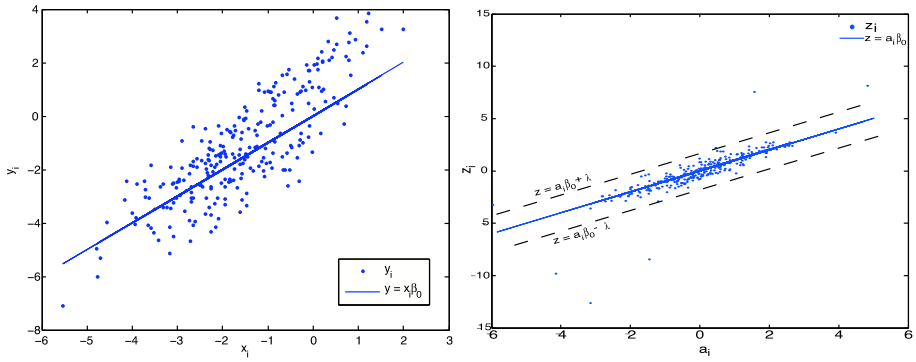
For illustration, we generated three test problems as follows. The nonparametric component f_0 was selected among two different functions, one sinusoidal function and one piecewise constant function. The covariate is chosen as $\mathbf{X}_i = g(i/n) + \eta_i$ with polynomial functions g and with the $(\eta_i)_{i=1,\ldots,n}$ generated independently from a centered distribution with finite variance, as explained in Section 6. For DWT, the filter we used is the Daubechies Symmlet filter with 8 vanishing moments. The sample size we took was $n = 2^8$. For each setting, 500 replicates of data with different X and u were generated. The variance of the noise was chosen such as the signal-to-noise ratios of the nonparametric and parametric component respectively were equal to 2.2 and 4.38. Such choices seem reasonable. With the simulated data, we then used the proposed algorithms to estimate the unknown parameters. For wavelet thresholding the universal threshold was used, while the termination tolerance δ was set to 10^{-5} for ARTUR and 10^{-10} for LEGEND. For *Backfitting*, we have used the algorithm of Fadili and Bullmore (2005) with a tolerance level δ equal to 10^{-20}. To save computational time we have also specified an upper limit of 2000 for the maximum number of iterations allowed.

Example 1: Sinusoidal test function

In examples 1 and 2, the covariate was generated using the polynomial function $g(t) = t^5 + 2t$ and with the $(\eta_i)_{i=1,\ldots,n}$ generated independently from $N(0,1)$. We have also run some numerical simulations with different design functions g such as $g(t) = 2^t$, $g(t) = e^{-t^2}$ or $g(t) = cos(t)$ with similar results, not reported here by the lack of space. It seems that assumption (A4) is not really necessary for asymptotic consistency.

We first consider the case of a sinusoidal function for the nonparametric part. In such a case one could obviously use smoothing splines based semiparametric estimation but it is interesting to see how our wavelet based procedure behaves. Figure 2 displays the wavelet transform of the data and of the design matrix. Note that the sparse representation of the nonparametric part allows an efficient reduction of the bias between the observations and a linear model. The dashed lines in the plot displayed in Figure 2, represent the lines $X_i\boldsymbol{\beta}_0 \pm \lambda$. Observations lying far out from these lines

do not affect the estimation of $\boldsymbol{\beta}_0$.



(a) (b)

Figure 2: Wavelet transform of the data. Figure (a) represents the scatter plot of the observations y_i versus the values of the covariates X_i. The line is the linear part of the model, of equation $y_i = X_i\boldsymbol{\beta}_0$. Figure (b) is the scatter plot in (a) after the Discrete Wavelet Transform: it represents the coefficients z_i versus $\mathbf{A}_i$. The solid line is the linear part of the model (equation $z_i = A_i\boldsymbol{\beta}_0$) and the dashed lines are the lines of equations $z_i = A_i\boldsymbol{\beta}_0 \pm \lambda$.

We now evaluate the effect of the QR decomposition on the estimation of the noise, and we compare the computational time required by each of the algorithms, namely ARTUR, LEGEND and *Backfitting* over the 500 replications of the experiment.

Estimation of σ by MAD		
True value	without QR	with QR
0.5	1.2222(0.0955)	0.5023(0.0511)

Table 1: The mean values of the estimates and their standard deviation over the 500 simulations in Example 1 with $n = 2^8$ (the standard deviation appears in brackets).

From Table 1, we get a fairly good impression on the effect of the QR decomposition on the estimation of the noise variance: the presence of the linear part introduces a strong bias in the MAD estimator, bias which is strongly diminished when using the QR decomposition. This also explains why in the comparison of their various thresholded estimators, Fadili and Bullmore (2005) often obtain estimators that are over-smoothed, since the variance that is used in their thresholds is over estimated. To be fair, we therefore have adopted for all methods the universal threshold

$\lambda = \sigma\sqrt{2\log(n)}$ with σ estimated by MAD after a QR decomposition.

Estimation of $\boldsymbol{\beta}_0$			
True value	*Backfitting*	ARTUR	LEGEND
1	0.9000(0.0273)	0.9417(0.0327)	0.9417(0.0327)
Average computing time	0.0936	0.0232	0.0151

Table 2: The mean values of the estimates and their standard deviation over the 500 simulations in Example 1 (standard deviation appears in brackets) with $n = 2^8$. The average MISE for the nonparametric part for these simulations is 0.1029 for ARTUR and LEGEND and 0.1098 for *Backfitting*.

From the last row of Table 2 one can see that both half-quadratic procedures (ARTUR and LEGEND) are faster than *Backfitting* and the quality of estimation of both the parametric and nonparametric parts in terms on mean squared error is also better. The differences observed in estimating $\boldsymbol{\beta}_0$ between the various procedures is mainly due to the different tolerance levels δ used by each. Note also that *Backfitting* always stops because the maximal number of iterations is reached. The estimation given by *Backfitting* could be improved but at the cost of a much larger computational time.

Recall that for both half-quadratic based algorithms, once the unknown parameter $\boldsymbol{\beta}_0$ is estimated, a nonparametric wavelet based estimation procedure is applied to the resulting residuals $y_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}_n$ for estimation of the nonparametric part. Figure 3 displays a typical example of these residuals and of the corresponding nonparametric estimation using ARTUR on one replication.

For the value of the signal-to-noise ratio ($SNR_f = 2.2$) adopted in our simulations for the nonparametric part, the estimator does not detect the discontinuity. However it produces results very similar to those by standard wavelet denoising of an identical nonparametric signal (without a linear part) with the same SNR, supporting our claim that the presence of the linear part in a PLM doesn't affect the estimation of the nonparametric part.

In their numerical implementation of ARTUR et LEGEND, both Vik (2004) and Dahyot and Kokaram (2004) conclude that LEGEND converges faster, supporting the theoretical results of Nikolova and Ng (2005). To share some light on this fact we have run some simulations with a larger number sample size. With $n = 2^{10}$ observations and the same signal-to-noise ratio as before one can see a clear difference in computational time among the two algorithm for estimators with

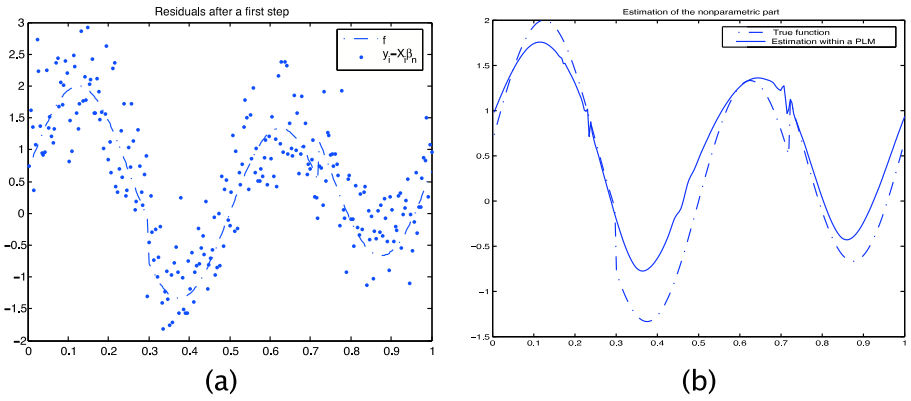


Figure 3: Estimation of the nonparametric part in Example 1. Figure (a) represents the residuals obtained after estimation of the linear part of the models, meaning $z_i - \mathbf{A}_i\hat{\boldsymbol{\beta}}_n$, and the true functionnal part (dash). in Figure (b) we have the resulting estimation of the function (solid) and the true function (dash).

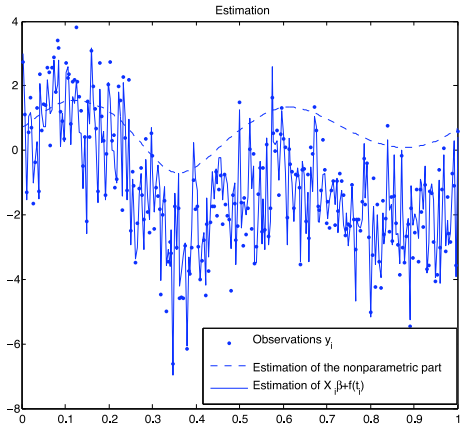


Figure 4: A typical partial linear fit from Example 1. The figure represents the scatter plot of the observations, the estimated functionnal part (dash) and the parial linear fit (solid) for one of the simulation.

equivalent qualities, as reported in Table 3.

Example 2: piecewise linear function

We would like now to illustrate our estimation procedure when the nonparametric part is highly non regular. We thus consider a function f_0 which is piecewise constant. It is obvious that for such a function, our wavelet based procedure is better suited than a spline based procedure. All other setting adopted for these simulations are the same as those for example 1.

Estimation of β_0		
True value	ARTUR	LEGEND
1	0.9762(0.0127)	0.9762(0.0127)
Average computing time	0.2331	0.0166
Average number of iterations	7	59

Table 3: The mean values of the estimates and their standard deviation over the 500 simulations in Example 1 (the standard deviation appears in brackets) with $n = 2^{10}$. LEGEND is much faster than ARTUR.

Estimation of σ by MAD	
True value	with QR
0.5	0.49961(0.052741)

Table 4: The mean values of the estimates and their standard deviation over the 500 simulations in Example 2 (the standard deviation appears in brackets) for $n = 2^8$.

The results given in Table 5 reinforce our claim from example 1 that half-quadratic algorithms are more efficient than *Backfitting*. Note moreover that the non regularity of the nonparametric part does not seem to affect the quality of the estimation of the vector of regression parameters.

Estimation of β_0 for $n = 2^8$			
True value	*Backfitting*	ARTUR	LEGEND
1	0.8999(0.0273)	0.9548(0.0309)	0.9548(0.0309)
Average computing time	0.0744	0.0209	0.0139

Table 5: The mean values of the estimates and their standard deviation over the 500 simulations in Example 2 (standard deviation appears in brackets). The average MISE for the nonparametric part for these simulations is 0.1012 for ARTUR and LEGEND and 0.1078 for *Backfitting*.

As in example 1, one can see from Table 5 and Table 6 that LEGEND outperforms ARTUR, and that the difference of computing time increases with the number of observations n.
The estimation of the nonparametric part does not detect the discontinuities of the function. Yet

Estimation of β_0 for $n = 2^{10}$		
True value	ARTUR	LEGEND
1	0.9554(0.0149)	0.9554(0.0149)
Average computing time	0.3036	0.0209

Table 6: The mean values of the estimates and their standard deviation over the 500 simulations in Example 2 (standard deviation appears in brackets). The average MISE for the nonparametric part for these simulations is 0.0584 for ARTUR and LEGEND.

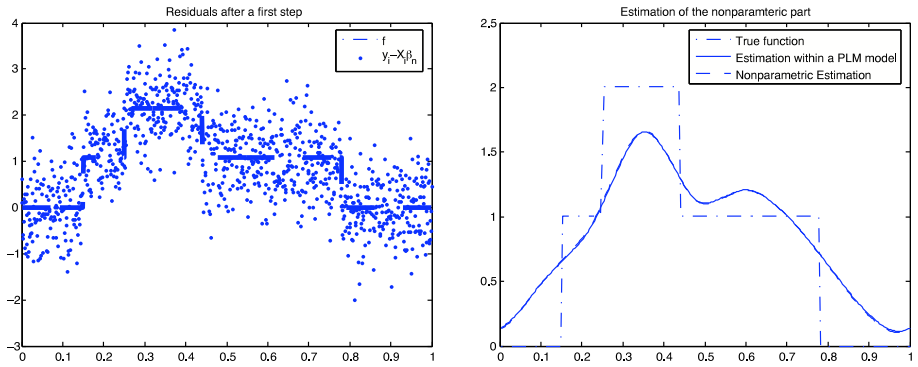


Figure 5: Estimation of the nonparametric part in Example 2. Figure (a) represents the residuals obtained after estimation of the linear part of the models, meaning $z_i - \mathbf{A}_i\hat{\boldsymbol{\beta}}_n$, and the true functionnal part (dash). In Figure (b) we have the resulting estimation of the function (solid) and the true function (dash).

compared to standard wavelet denoising in a nonparametric regression model with the same SNR, the estimation obtained in the PLM is very similar. The bad visual quality of the estimation results from the choice of the signal-to-noise ratio ($SNR_f = 2.2$) adopted in our simulations rather than the presence of the linear part.

Example 3: dimension 4

We now consider a case where the vector of parameter $\boldsymbol{\beta}$ belongs to $\mathbb{R}^4$ (the dimension of the design regression matrix X is then $n \times 4$). The nonparametric part f_0 is the same as in example 2, meaning that the function is highly irregular. The SNR for the global model was chosen equal to 5.99, with a SNR equal to 4.38 for the nonlinear part. One may summarize the results for this example in the above tables.

Estimation of σ by MAD with QR	
True value	with QR
0.5	0.52261(0.053808)

Table 7: The mean values of the estimates and their standard deviation over the 500 simulations in Example 3 (the standard deviation appears in brackets).

Estimation of β_0			
True value	*Backfitting*	ARTUR	LEGEND
-1	-1.4969(0.45822)	-0.7203(0.461)	-0.7203(0.461)
3	2.8563(0.09770)	2.9168(0.09941)	2.9168(0.09941)
0	-0.1201(0.33685)	0.0125(0.34415)	0.0125(0.34415)
8	7.5601(0.16772)	7.7112(0.18525)	7.7112(0.18525)
Mean squared error	0.8434	0.5438	0.5438
Average computing time	0.1602	0.0305	0.0234

Table 8: The mean values of the estimates and their standard deviation over the 500 simulations in Example 3 (the standard deviation appears in brackets) for a given value of the true β_0. The average MISE for the nonparametric part for these simulations is 0.2140 for ARTUR and LEGEND and 0.2164 for *Backfitting*.

As one can see with computational times that are similar for all procedures, both half-quadratic algorithms outperform *Backfitting* in terms of the MSE.
As for examples 1 and 2, when the sample size increases, among the half-quadratic algorithms the LEGEND one is much faster.

Conclusion

This paper develops a powerful penalized least squares estimation in partially linear models, based on a wavelet expansion of the nonparametric part. Choosing an appropriate penalty on the wavelet coefficients of the function, the procedure leads to an estimation of the linear part of partly linear models independent from the nonparametric part, while the estimation of the nonparametric part is adaptative relatively to the smoothness of the function. Since the functionnal part of the model has

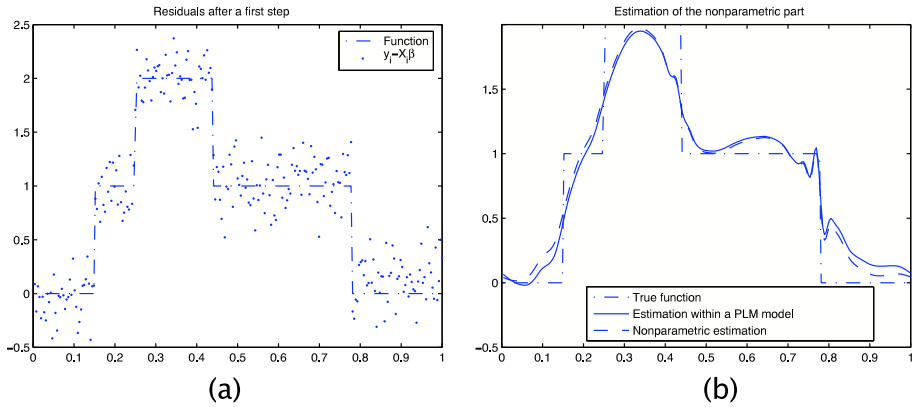


Figure 6: Estimation of the nonparametric part in Example 3. Figure (a) represents the residuals obtained after estimation of the linear part of the models, meaning $z_i - \mathbf{A}_i\hat{\boldsymbol{\beta}}_n$, and the true functionnal part (dash). In Figure (b) we have the resulting estimation of the function (solid) and the true function (dash).

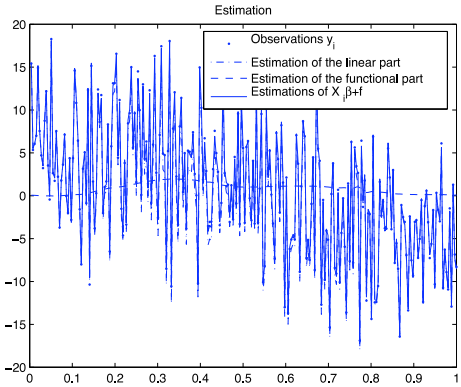


Figure 7: A typical partial linear fit from Example 3. The figure represents the scatter plot of the observations, the estimated functionnal part (dash) and the partial linear fit (solid) for one of the simulation.

a sparse representation, the estimation of the regression parameters vector is moreover interpreted as a common M-estimation. In the particular case of an l^1-penalty (leading to soft thresholding and Huber's estimator) the near-minimaxity of the estimation of both parametric and nonparametric parts of a partially linear model is established, and the result is avalaible for a large class of functions, including nonsmooth irregular functions. From an implementation point of view, half-quadratic algorithms are proposed that appear to give good results on simulation studies.

Our ongoing research is focusing on exploring the asymptotic properties of the procedure for other thresholding schemes and in more general frameworks such as nonequidistant designs for the nonparametric part.

Acknowledgements

Part of this work was supported by the 'IAP Research Network P5/24'. The author would like to thank Dr. Fadili and Dr. Bullmore for kindly providing the Matlab codes implementing the backfitting procedures used in the paper.

6 Appendix

Appendix A. Discussion of the assumptions.

In this Section, we study wether the assumptions made in Theorem 1 are reasonable in practice. Following Rice (1986) or Speckman (1988) we suppose that the design matrix X can be written as a sum of a deterministic function and a noise term. The (i,j)-component of X can be written as $x_{i,j} = g_i(t_j) + \xi_{i,j}$ with functions g_i such that $\int f g_i = 0$ and where $\xi_{i,j}$ denotes a realization of a random variable ξ_i. The variables $(\xi_i)_{i=1,\dots,n}$ are supposed to be independent and identically distributed, centered and with finite variance, independent from the u_i. With these notation, assumptions (A1), (A2) and (A4) become:

(A1) The norm of $\frac{1}{n}X^T\mathbf{F}$ can be decomposed as follows:
$\|\frac{1}{n}X^T\mathbf{F}_0\|^2 = \sum_{j=1}^{p}\left(\frac{1}{n}\sum_{j=1}^{n} g_i(t_j)f(t_j) + \frac{1}{n}\sum_{j=1}^{n}\xi_{i,j}f(t_j)\right)^2$.
The convergence towards 0 of the first term is ensured by the assumption that $\int f g_i = 0$ for all $i = 1,\dots,p$. We can prove that the second term tends to 0 almost surely.

Remark 2. *When we suppose that $\forall i,\ \int f g_i = 0$, this impose that either the integral of f is equal to zero or the vector $\mathbb{1}_{n\times 1}$ is not in the space spanned by the columns of X. This is the usual assumption for identifiability in PLM (e.g. Chen (1988) or Donald and Newey (1994)).*

(A2) Let $V(g)$ be the matrix with entries $\int g_i g_j$ and V denotes the covariance matrix of the variables $(\xi_i)_{i=1,\dots,n}$. One can prove that $\frac{1}{n}X^TX$ converges almost surely to $V(g) + V$. It is sufficient to assume that the family $(g_i)_{i=1,\dots,n}$ is $\mathbb{L}^2$-orthogonal in order that the matrix $V(g) + V$ is non singular.

(A4) Actually, it is equivalent to prove that $\frac{1}{n}\sup\|\mathbf{A}_i\|^2 \to 0$ to get (A4). For $i \in \{1,\dots,n\}$ given, $\frac{1}{n}\|\mathbf{A}_i\|^2$ is equal to $n^{-1}\mathbf{A}_i^T\mathbf{A}_i = \sum_{l=1}^{p}\left[\frac{1}{n}\sum_{j=1}^{n}\psi_i(t_j)X_{j,l}\right]^2$. With the previous notation,

$x_{j,l} = g_l(t_j) + \xi_{j,l}$ and we can establish that $n^{-1}\mathbf{A}_i^T\mathbf{A}_i$ tends almost surely to $\sum_{l=1}^{p}\left(\int \psi_i g_l\right)^2$. This can also be written as $n^{-1}\mathbf{A}_i^T\mathbf{A}_i \sim \sum_{l=1}^{p}(w_i^l)^2$ with $(w_i^l)_{i=1,\dots,n}$ wavelets coefficients of the functions g_l.

If, for all $l = 1,\dots,p$, g_l is a polynomial function whose degree is less than or equal to the number of vanishing moments N of the wavelet mother, then this assumption holds.

Hypothesis (A3) is not detailled here because even if it does not seem very constraining, it is difficult to study its feasibility.

To conclude, when the design X_i, $i = 1,\dots,n$ can be written as $X_i = g_i + \xi_i$ with g_i orthogonal polynomial functions with a degree less than or equal to N, and with ξ_i centered independent random variables with finite variance, whenever $\int f g_i = 0$ for all i, assumptions (A1), (A2) and (A4) hold.

Appendix B. Proofs of the main results

B.1. Preliminary result

Proposition 2.

When assumptions (A2) and (A3) hold,

$$\frac{1}{\sqrt{n}}\sum_{i=i_0}^{n}\rho'_\lambda(\theta_{0i}+\varepsilon_i)\mathbf{A}_i = \bigcirc_{\mathbb{P}}(\lambda)$$

This result comes from Bernstein's inequality applied to the random variables $Y_{i,j} = \frac{A_{i,j}}{\sqrt{n}}\frac{\rho'_\lambda(\theta_{0i}+\varepsilon_i)}{\lambda}$, $i = 1,\dots,n$, for any fixed j in $\{1,\dots,p\}$. Indeed, these variables are almost surely uniformly bounded and $\sum_{i=1}^{n}\mathbb{E}[Y_{i,j}^2]$ is bounded, due to the following lemma:

Lemma 3. *If (A2) and (A3) hold,*

(i) $n^{-1/2}\sup_{i=1,\dots,n}\|\mathbf{A}_i\| \to 0$

(ii) $n^{-1}\sum_{i=1,\dots,n}\|\mathbf{A}_i\|^2 = \bigcirc(1)$

This result lies on the observation that $\|\mathbf{A}_i\|^2 = \mathbf{A}_i^T(A^TA)^{1/2}(A^TA)^{-1/2}\mathbf{A}_i$, and consequently $\|\mathbf{A}_i\| \le n^{1/2}\|(\frac{1}{n}A^TA)^{1/2}\|h_i^{1/2}$.

B.2. Variables transform

Let us recall that we are studying the model

$$z_i = \mathbf{A}_i^T \boldsymbol{\beta}_0 + \theta_{0i} + \varepsilon_i \quad \text{under (A2)-(A4)} \tag{13}$$

(Assumption (A1) is an identifiability assumption and does not intervene in the proofs). Following Huber (1981) or Bai *et al.* (1992), we build an equivalent model by a change of variables. Let us define the following transforms:

$$\begin{aligned} R &= A(A^TA)^{-1/2}, \\ \boldsymbol{\alpha} &= \frac{1}{\lambda}(A^TA)^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) \\ d_i &= \frac{1}{\lambda}(\theta_i + \varepsilon_i). \end{aligned}$$

The results may be established equivalently for the following model:

$$z_i = \mathbf{R}_i^T \boldsymbol{\alpha}_0 + d_i \quad \text{under (A2'')-(A4'');} \tag{14}$$

(A2'') $R^TR = I_p$.

(A3'') $h = \max_{i=i_0,\ldots,n} \mathbf{R}_i^T \mathbf{R}_i$ tends to 0.

(A4'') $K_n'' := \sum_{i=i_0}^n \mathbf{R}_i \mathbf{R}_i^T \mathbb{E}\left[\rho_1''(d_i)\right]$ tends to K_0'', non singular matrix.

As the Huber cost function has scale transform properties:

$$\text{for any } v > 0, \;\; \rho_\lambda(u) = v^2 \rho_{\lambda/v}(u/v), \tag{15}$$

we then can prove that in the model (14), the estimator $\hat{\boldsymbol{\alpha}}_n$ is solution of the minimization problem

$$\hat{\boldsymbol{\alpha}}_n = \underset{\boldsymbol{\alpha}}{argmin} \sum_{i=1}^n \rho_1(d_i - \mathbf{R}_i^T \boldsymbol{\alpha}).$$

As $\rho_\lambda'(u) = \lambda \rho_1'(u/\lambda)$ and $\rho_\lambda''(u) = \rho_1''(u/\lambda)$, we have $K_0'' \sim \Sigma^{-1} K_0$ and Proposition 2 becomes in (14):

$$\sum_{i=i_0}^n \rho_1'(d_i) \mathbf{R}_i = \bigcirc_{\mathbb{P}}(1).$$

In all the proofs, we will consider the model (14) and obtain the consistency results thanks to the mentionned transforms.

B.3. Convergence of the criterion

Proposition 4.

Let c be a strictly positive constant. Suppose (A1) to (A4) hold. Then,

$$\sup_{\{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\leq c\lambda n^{-1/2}\}} \frac{1}{\lambda^2}\Big|\sum_{i=i_0}^{n}\left(\rho_\lambda(\theta_{0i}+\varepsilon_i-\mathbf{A}_i^T(\boldsymbol{\beta}-\boldsymbol{\beta}_0))-\rho_\lambda(\theta_{0i}+\varepsilon_i)\right) + \sum_{i=i_0}^{n}\rho'_\lambda(\theta_{0i}+\varepsilon_i)\mathbf{A}_i^T(\boldsymbol{\beta}-\boldsymbol{\beta}_0) - n\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T K_0(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\Big| \xrightarrow{\mathbb{P}} 0.$$

The proof is built on two phases: we first approximate the Huber cost function ρ with a smoother function, keeping a control on the third derivative; secondly, we develop a scheme of proof very similar to Bai *et al.* (1992) in the transformed model (14). The main argument is the convexity of ρ, which allows in particular the use of Rockafellar's theorems.

B.3.1. Approximation of Huber cost function

The approximation is built by three successive integrations. Let $0<\delta<1$. We define r_δ^3 on $\mathbb{R}$:

$$r_\delta^3 : u \mapsto \begin{cases} \frac{6}{\delta^3}(u-(1-\delta/2))(u-(1+\delta/2)) & \text{if } 1-\delta/2<|u|<1+\delta/2 \\ 0 & \text{otherwise} \end{cases}.$$

We introduce next, r_δ^2 primitive of r_δ^3 equal to zero at $1+\delta/2$, r_δ^1 primitive of r_δ^2 equal to zero at 0 and r_δ, primitive of r_δ^1 equal to zero at 0.

The function series $\tilde{\rho}_1 = r^3_{1/n^2}$ is a series of convex functions $\mathcal{C}^3$, which converges uniformly towards ρ_1 when n goes to infinity. We can furthermore prove that $\int|\tilde{\rho}_1^{(3)}|\leq 12$, and that

$$n\|\tilde{\rho}_1-\rho_1\|_\infty \underset{n\to\infty}{\longrightarrow} 0, \tag{16}$$

$$n\|\tilde{\rho}'_1-\rho'_1\|_\infty \underset{n\to\infty}{\longrightarrow} 0, \tag{17}$$

$$\|\tilde{\rho}''_1-\rho''_1\|_\infty \leq 1. \tag{18}$$

Moreover, $\tilde{\rho}''_1$ and ρ''_1 only differ from each others on two intervals of length $1/n^2$.

B.3.2. Preliminary tools

Proposition 5. *Let C be an open compact set of $\mathbb{R}^m$. We consider $(f_n)_{n\in\mathbb{N}}$ and f a family of convex functions defined on C and taking their values in a given probability space (Ω, P, μ). Suppose for all $u\in C$,*

$f_n(u) - f(u)$ converges in probability to 0. Then the convergence in probability of $\sup_{\{u\in C\}} f_n(u) - f(u)$ towards 0 is acquired.

Proof. We recall a theorem given in Rockafellar (1970) (Theorem 10.8, page 90):

Proposition 6. *Let $\mathcal{C}$ be an open compact set of $\mathbb{R}^m$. We consider $(f_n)_{n\in\mathbb{N}}$ and f a family of finite convex functions defined on C. Suppose the series $(f_n)_{n\in\mathbb{N}}$ converges simply to f on $\mathcal{C}$. Then the convergence is uniform on $\mathcal{C}$.*

In order to obtain a similar result for the convergence in probability, we may use the following characterization of such a convergence:

Lemma 7. *Let $(X_n)_n$ be a series of random variables and X a random variable. The series (X_n) converges in probability towards X if and only if from all subsequence of X_n we can extract a series which tends almost surely to X.*

Consider $f_{\nu(n)}$ a subsequence of f_n. We would like to find $\boldsymbol{\beta}(n)$, subsequence of $\nu(n)$, such that for all $u \in \mathcal{C}$, $f_{\boldsymbol{\beta}(n)}(u) - f(u) \xrightarrow{a.s.} 0$. The Lemma 7 tells us that for all $u \in \mathcal{C}$ there exists $\eta_u(n)$ extraction of $\nu(n)$ such that $f_{\eta_u(n)}(u) - f(u) \xrightarrow{a.s.} 0$. Let us consider $\mathcal{D} = \{u_0, u_1, u_2 \dots\}$ dense and countable subset of $\mathcal{C}$. Using a diagonal procedure, we can exhibit $(\boldsymbol{\beta}(n))$ such that for all $u \in \mathcal{D}$, we have $f_{\boldsymbol{\beta}(n)}(u) - f(u) \xrightarrow{a.s.} 0$. Afterwards, the convergence of $f_{\boldsymbol{\beta}(n)} - f$ on $\mathcal{C}$ holds by density of $\mathcal{D}$ and continuity of $f_{\boldsymbol{\beta}(n)} - f$. Applying Rockafellar's Theorem, we obtain that $\sup_{u\in C} f_{\boldsymbol{\beta}(n)}(u) - f(u)$ tends almost surely to 0.

To conclude, we have proved that from all subsequence $\sup_{u\in C} f_{\nu(n)}(u) - f(u)$ of $\sup_{u\in C} f_n(u) - f(u)$ we could extract a series which converges almost surely to 0. This finishes the proof using Lemma 7. □

B.3.3. Convergence criterion

Let $c > 0$. We are going to prove that in model (14) we have:

$$\sup_{\{\|\boldsymbol{\alpha}\|\leq c\}} \left| \sum_{i=i_0}^{n} \Big(\rho_1(d_i - \mathbf{R}_i^T\boldsymbol{\alpha}) - \rho_1(e_i) \Big) + \sum_{i=1}^{n} \rho_1'(d_i)\mathbf{R}_i^T\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}^T K_0''\boldsymbol{\alpha} \right| \xrightarrow{\mathbb{P}} 0. \tag{19}$$

Note that in the initial model (13), this is equivalent to

$$\sup_{\{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\leq c\lambda n^{-1/2}\}} \frac{1}{\lambda^2} \Big| \sum_{i=i_0}^{n} \big(\rho_\lambda(\theta_{0i} + \varepsilon_i - \mathbf{A}_i^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0)) - \rho_\lambda(\theta_{0i} + \varepsilon_i) \big) + \sum_{i=i_0}^{n} \rho_\lambda'(\theta_{0i} + \varepsilon_i)\mathbf{A}_i^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0) - n\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T K_0(\boldsymbol{\beta} - \boldsymbol{\beta}_0) \Big| \xrightarrow{\mathbb{P}} 0.$$

• We introduce:

$$\Delta(\boldsymbol{\alpha}) := \sum_{i=1}^{n} \left(\tilde{\rho}_1(d_i - \mathbf{R}_i^T\boldsymbol{\alpha}) - \tilde{\rho}_1(d_i) + \tilde{\rho}_1'(d_i)\mathbf{R}_i^T\boldsymbol{\alpha} \right).$$

The cost function $\tilde{\rho}_1$ is convex. For every i, it gives the upper bound:

$$\left| \tilde{\rho}_1(d_i - \mathbf{R}_i^T\boldsymbol{\alpha}) - \tilde{\rho}_1(d_i) + \tilde{\rho}_1'(d_i)\mathbf{R}_i^T\boldsymbol{\alpha} \right| \leq |\tilde{\rho}_1'(d_i - \mathbf{R}_i^T\boldsymbol{\alpha}) - \tilde{\rho}_1'(d_i)||\mathbf{R}_i^T\boldsymbol{\alpha}|. \tag{20}$$

This inequality gives a bound of the variance of $\Delta(\boldsymbol{\alpha})$:

$$Var(\Delta(\boldsymbol{\alpha})) \leq \sum_{i=1}^{n} \mathbb{E}\left[\left(\tilde{\rho}_1'(d_i - \mathbf{R}_i^T\boldsymbol{\alpha}) - \tilde{\rho}_1'(d_i) \right)^2 \right] |\mathbf{R}_i^T\boldsymbol{\alpha}|^2.$$

The function $\tilde{\rho}_1'$ being 1-Lipschitz,

$$\forall n \in \mathbb{N},\ \forall i = 1, \ldots, n,\ \forall u \in \mathbb{R}^+,\ \mathbb{E}\left(\tilde{\rho}_1'(d_i + u) - \tilde{\rho}_1'(d_i) \right)^2 \leq u^2.$$

Consequently,

$$Var(\Delta(\boldsymbol{\alpha})) \leq \sum_{i=1}^{n} |\mathbf{R}_i^T\boldsymbol{\alpha}|^4 \leq \|\boldsymbol{\alpha}\|^4 \sum_{i=1}^{n} \|\mathbf{R}_i\|^4.$$

As $\boldsymbol{\alpha}$ is supposed to be bounded and $\sum_{i=1}^{n} |\mathbf{R}_i|^4 \leq h \sum h_i = hp$ tends to 0, we obtain that $Var(\Delta(\boldsymbol{\alpha}))$ tends to 0. Bienaymé-Tchebychev inequality ensures then that $|\Delta(\boldsymbol{\alpha}) - \mathbb{E}\Delta(\boldsymbol{\alpha})|$ converges towards 0 in probability.

• **The term $\mathbb{E}\Delta(\boldsymbol{\alpha})$.**

As the function $\tilde{\rho}$ is $\mathcal{C}^3$, the Taylor expansion of degree 2 with a rest of an integral form of $\tilde{\rho}_1$ on a neighborhood of d_i exists. It gives:

$$\begin{aligned} &\tilde{\rho}_1(d_i - \mathbf{R}_i^T\boldsymbol{\alpha}) - \tilde{\rho}_1(d_i) + \tilde{\rho}_1'(d_i)\mathbf{R}_i^T\boldsymbol{\alpha} - \frac{1}{2}\tilde{\rho}_1''(d_i)\boldsymbol{\alpha}^T\mathbf{R}_i\mathbf{R}_i^T\boldsymbol{\alpha} \\ &\qquad = -\int \tilde{\rho}_1^{(3)}(t)(d_i - t)^3 \mathbb{1}_{d_i - \mathbf{R}_i^T\boldsymbol{\alpha} \leq t \leq d_i} dt/6. \end{aligned}$$

Using the bound $\int \left| \tilde{\rho}_1^{(3)}(t) \right| dt \leq 12$, obtained when constructing $\tilde{\rho}$, we obtain:

$$\mathbb{E}\left| \sum_{i=1}^{n} \left(\tilde{\rho}_1(d_i - \mathbf{R}_i^T\boldsymbol{\alpha}) - \tilde{\rho}_1(d_i) + \tilde{\rho}_1'(d_i)\mathbf{R}_i^T\boldsymbol{\alpha} - \frac{1}{2}\tilde{\rho}_1''(d_i)\boldsymbol{\alpha}^T\mathbf{R}_i\mathbf{R}_i^T\boldsymbol{\alpha} \right) \right| \leq 2\|\boldsymbol{\alpha}\|^3 \sum_{i=1}^{n} \|\mathbf{R}_i\|^3.$$

Note that $\sum_{i=1}^{n} \|\mathbf{R}_i\|^3 \leq h^{1/2} \sum h_i = h^{1/2} p \to 0$. Therefore, when $\|\boldsymbol{\alpha}\| \leq c$,

$$\mathbb{E}\Delta(\boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{\alpha}^T \tilde{K}_n'' \boldsymbol{\alpha} + o(1), \text{ with } \tilde{K}_n'' = \sum_{i=i_0}^{n} \mathbf{R}_i\mathbf{R}_i^T \mathbb{E}\left[\tilde{\rho}_{i,1}''(d_i) \right].$$

Actually, $\tilde{K}_n''$ converges towards K_0''. Let us decompose $\|\tilde{K}_n'' - K_0''\|$ in

$$\|\tilde{K}_n'' - K_0''\| \leq \|\tilde{K}_n'' - K_n''\| + \|K_n'' - K_0''\|.$$

The convergence to 0 of the second term is ensured by hypothesis (A3''). The first term is:

$$\tilde{K}_n'' - K_n'' = \sum \mathbf{R}_i^T \mathbb{E}(\tilde{\rho}_1''(d_i) - \rho_1''(d_i)).$$

The functions $\tilde{\rho}_1''$ and ρ_1'' only differ on intervals whose total length is $2/(n^2)$. Consequently, $\mathbb{E}(\tilde{\rho}_\lambda''(d_i) - \rho_\lambda''(d_i)) \leq 2/(n^2)\|\tilde{\rho}_\lambda'' - \rho_\lambda''\|_\infty \|f_\varepsilon\|_\infty$ where f_ε denotes the density function of ε_i. We obtain the inequality: $\|\tilde{K}_n'' - K_n''\| \leq \frac{1}{n}h^{1/2}C$, with C a constant. As h tends to 0 under (A4''), we deduce that $\tilde{K}_n''$ converges towards K_0'' and thus $\mathbb{E}\Delta(\boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{\alpha}^T K_0'' \boldsymbol{\alpha} + o_{\mathbb{P}}(1)$.

When $\|\boldsymbol{\alpha}\| \leq c$, the convergence in probability of $|\Delta(\boldsymbol{\alpha}) - \mathbb{E}\Delta(\boldsymbol{\alpha})|$ to 0 implies:

$$\left| \sum_{i=1}^n \left(\tilde{\rho}_1(d_i - \mathbf{R}_i^T\boldsymbol{\alpha}) - \tilde{\rho}_1(d_i) + \tilde{\rho}_1'(d_i)\mathbf{R}_i^T\boldsymbol{\alpha} \right) - \frac{1}{2}\boldsymbol{\alpha}^T K_0''\boldsymbol{\alpha} \right| \xrightarrow{\mathbb{P}} 0.$$

If $\tilde{D}$ and D respectively denote $\tilde{D} := \sum_{i=1}^n \tilde{\rho}_1(d_i - \mathbf{R}_i^T\boldsymbol{\alpha}) - \tilde{\rho}_1(d_i)$ and $D := \sum_{i=1}^n \rho_1(d_i - \mathbf{R}_i^T\boldsymbol{\alpha}) - \rho_1(d_i)$, then $|D - \tilde{D}| \leq n\|\tilde{\rho}_1 - \rho_1\|_\infty$. Using (16), we obtain the almost sure convergence of $D - \tilde{D}$ to 0. In the same way, if $\tilde{B} := \sum_{i=1}^n \tilde{\rho}_1'(d_i)\mathbf{R}_i^T\boldsymbol{\alpha}$ and $B := \sum_{i=1}^n \rho_1'(d_i)\mathbf{R}_i^T\boldsymbol{\alpha}$, we then have $|B - \tilde{B}| \leq n\|\tilde{\rho}_1' - \rho_1'\|_\infty \|\boldsymbol{\alpha}\| h^{1/2}$. When $\|\boldsymbol{\alpha}\| \leq c$, properties (17) implie that $B - \tilde{B}$ tends almost surely to 0. All together, we have:

$$\left| \sum_{i=1}^n \left(\rho_1(d_i - \mathbf{R}_i^T\boldsymbol{\alpha}) - \rho_1(d_i) + \rho_1'(d_i)\mathbf{R}_i^T\boldsymbol{\alpha} \right) - \frac{1}{2}\boldsymbol{\alpha}^T K_0''\boldsymbol{\alpha} \right| \xrightarrow{\mathbb{P}} 0.$$

- We may prove now that the convergence is uniform on the set $\{\|\boldsymbol{\alpha}\| \leq c\}$.

  The functions in $\boldsymbol{\alpha}$:

$$\sum_{i=1}^n \left(\rho_1(d_i - \mathbf{R}_i^T\boldsymbol{\alpha}) - \rho_1(d_i) + \rho_1'(d_i)\mathbf{R}_i^T\boldsymbol{\alpha} \right) \text{ and } \frac{1}{2}\boldsymbol{\alpha}^T K_0''\boldsymbol{\alpha}$$

  are convex and the set $\{\|\boldsymbol{\alpha}\| \leq c\}$ is convex, compact and independent from n. Proposition 5 completes the proof.

B.4. Proof of Theorem 1

B.4.1. Consistency

In the model (14), we are willing to prove that $\hat{\boldsymbol{\alpha}}_n = \bigcirc_{\mathbb{P}}(1)$. Let $c_n \to \infty$. We may prove that $\mathbb{P}(\|\hat{\boldsymbol{\alpha}}_n\| > c_n) \to 0$. We can deduce from (19) that there exists a series c_n' such that $c_n' \to \infty$, $c_n' \leq c_n$ and

$$\sup_{\{\|\boldsymbol{\alpha}\| \leq c_n'\}} \left| \sum_{i=1}^n \left(\rho_1(d_i - \mathbf{R}_i^T\boldsymbol{\alpha}) - \rho_1(d_i) + \rho_1'(d_i)\mathbf{R}_i^T\boldsymbol{\alpha} \right) - \frac{1}{2}\boldsymbol{\alpha}^T K_0''\boldsymbol{\alpha} \right| \xrightarrow{\mathbb{P}} 0.$$

It is sufficient then to prove that $\mathbb{P}\left(\|\hat{\boldsymbol{\alpha}}_n\| > c'_n\right) \to 0$.

- Suppose $\|\boldsymbol{\alpha}\| = c'_n$.

  We have

$$\sum_{i=1}^{n}\Big(\rho_1(d_i - \mathbf{R}_i^T\boldsymbol{\alpha}) - \rho_1(d_i)\Big) = -\sum_{i=1}^{n}\rho'_1(d_i)\mathbf{R}_i^T\boldsymbol{\alpha} + \frac{1}{2}\boldsymbol{\alpha}^T K''_0\boldsymbol{\alpha} + o_{\mathbb{P}}(1).$$

  First,

$$\|\frac{1}{2}\boldsymbol{\alpha}^T K''_0\boldsymbol{\alpha}\| \geq \frac{1}{2}\underline{s}(K''_0)(c'_n)^2,$$

  with $\underline{s}(K''_0)$ smallest eigenvalue of K''_0. As the matrix K''_0 is nonsingular, $\underline{s}(K''_0) > 0$. Next, Proposition 2 implies that

$$\|\sum_{i=1}^{n}\rho'_1(d_i)\mathbf{R}_i^T\boldsymbol{\alpha}\| = \bigcirc_{\mathbb{P}}(c'_n).$$

  As a consequence, the probability that the quantity

$$\sum_{i=1}^{n}\rho_1(d_i - \mathbf{R}_i^T\boldsymbol{\alpha}) - \rho_1(d_i) = -\sum\rho'_1(d_i)\mathbf{R}_i^T\boldsymbol{\alpha} + \frac{1}{2}\boldsymbol{\alpha}^T K''_0\boldsymbol{\alpha} + o_{\mathbb{P}}(1)$$

  is negative tends to 0. This result is true uniformly for $\boldsymbol{\alpha}$ verifying $\|\boldsymbol{\alpha}\| = c'_n$. We obtain:

$$\mathbb{P}\left(\inf_{\{\boldsymbol{\alpha},\ \|\boldsymbol{\alpha}\|=c'_n\}}\sum_{i=1}^{n}\rho_1(d_i - \mathbf{R}_i^T\boldsymbol{\alpha}) - \rho_1(d_i) \leq 0\right) \to 0. \tag{21}$$

- Let $\boldsymbol{\alpha}$ be such that $\|\boldsymbol{\alpha}\| \geq c'_n$.

  We define $t = \frac{c'_n}{\|\boldsymbol{\alpha}\|} \in]0;1]$ and $\boldsymbol{\alpha}' = t\boldsymbol{\alpha}$. With the equality $d_i - \mathbf{R}_i^T\boldsymbol{\alpha}' = (1-t)d_i + t(d_i - \mathbf{R}_i^T\boldsymbol{\alpha})$, together with the convexity of ρ, we have:

$$\rho_1(d_i - \mathbf{R}_i^T\boldsymbol{\alpha}') - \rho_1(d_i) \leq t\left(\rho_1(d_i - \mathbf{R}_i^T\boldsymbol{\alpha}) - \rho_1(d_i)\right).$$

  As $\|\boldsymbol{\alpha}'\| = c'_n$, it comes that:

$$\mathbb{P}\left(\inf_{\{\boldsymbol{\alpha},\ \|\boldsymbol{\alpha}\|\geq c'_n\}}\sum_{i=1}^{n}\rho_1(d_i - \mathbf{R}_i^T\boldsymbol{\alpha}) - \rho_1(d_i) \leq 0\right) \to 0, \tag{22}$$

  or equivalently:

$$\mathbb{P}\left(\inf_{\{\boldsymbol{\alpha},\ \|\boldsymbol{\alpha}\|\geq c'_n\}} J_n(\boldsymbol{\alpha}) \leq J_n(0)\right) \to 0.$$

  The estimator $\hat{\boldsymbol{\alpha}}_n$ has been defined as the argument realizing the minimum of J_n, and so, $\mathbb{P}\left(\|\hat{\boldsymbol{\alpha}}_n\| \geq c'_n\right)$ tends towards zero, which achieves the proof.

B.4.2. Bahadur's representation

We want to prove that in model (14), we have

$$\hat{\boldsymbol{\alpha}}_n = K_0''^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\rho_1'(d_i)\mathbf{R}_i\right) + o_{\mathbb{P}}(1).$$

Let us first recall this result given in Rockafellar (1970):

Proposition 8. *Let $\mathcal{C}$ be an open convex set. Let f_n be a family of differentiable convex functions and f be a differentiable convex function. If f_n converges simply towards f on $\mathcal{C}$, then ∇f_n converges simply towards ∇f on $\mathcal{C}$ and the convergence is uniform on every compact set of $\mathcal{C}$.*

Similarly to Proposition 5, this Proposition can be generalized to a convergence in probability (using Lemma 7).

Applying this Proposition to the result (19) gives us that, for all $c > 0$,

$$\sup_{\|\boldsymbol{\alpha}\|\leq c}\left|\sum_{i=1}^{n}\left(\rho_1'(d_i - \mathbf{R}_i^T\boldsymbol{\alpha})\mathbf{R}_i - \rho_1'(d_i)\mathbf{R}_i\right) + K_0''\boldsymbol{\alpha}\right| \xrightarrow{\mathbb{P}} 0.$$

We have proved precedently that $\hat{\boldsymbol{\alpha}}_n = \bigcirc_{\mathbb{P}}(1)$. Then,

$$\left|\sum_{i=1}^{n}\left(\rho_1'(d_i - \mathbf{R}_i^T\hat{\boldsymbol{\alpha}}_n)\mathbf{R}_i - \rho_1'(d_i)\mathbf{R}_i\right) + K_0''\hat{\boldsymbol{\alpha}}_n\right| \xrightarrow{\mathbb{P}} 0. \tag{23}$$

By definition of $\hat{\boldsymbol{\alpha}}_n$, $\sum_{i=1}^{n}\rho_1'(d_i - \mathbf{R}_i^T\hat{\boldsymbol{\alpha}}_n)\mathbf{R}_i = 0$. The convergence of (23) becomes:

$$\hat{\boldsymbol{\alpha}}_n = K_0''^{-1}\left(\sum_{i=1}^{n}\rho_1'(d_i)\mathbf{R}_i\right) + o_{\mathbb{P}}(1),$$

which is the announced result.

B.4.3. Asymptotic behavior of the functionnal part

The model considered for this part of the proof is the model (13) contrarily to what precedes.

Parseval equality gives: $\|\hat{f}_n - f\|_2 \sim \frac{1}{n}\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|$. We decompose this bound into: $\frac{1}{n}\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| \leq \frac{1}{n}\|\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n\| + \frac{1}{n}\|\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|$ where

$$\tilde{\boldsymbol{\theta}}_{i,n} = \begin{cases} z_i - \mathbf{A}_i^T\boldsymbol{\beta}_0 & \text{if } i < i_0 \\ \operatorname{sign}(z_i - \mathbf{A}_i^T\boldsymbol{\beta}_0)\left(|z_i - \mathbf{A}_i^T\boldsymbol{\beta}_0| - \lambda\right)_+ & \text{if } i \geq i_0 \end{cases}.$$

D. Donoho (1992) proved that there exists a constant C such that $\mathbb{E}\frac{1}{n}\|\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| \leq C\left(\frac{\log(n)}{n}\right)^{\frac{s}{1+2s}}$. The convergence in L^2 implies the convergence in probability.

The term $\frac{1}{n}\|\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n\|$ verifies the inequality $\frac{1}{n}\|\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n\| \leq \frac{1}{n}\|A\|\|\hat{\boldsymbol{\beta}}_n - \hat{\boldsymbol{\beta}}_0\| + 2\frac{\lambda}{n}$. Assumptions (A2) and (A3) ensure that $\frac{1}{\sqrt{n}}\|A\| = \left(\frac{1}{n}\sum\|\mathbf{A}_i\|^2\right)^{1/2}$ is bounded and that $\|\hat{\boldsymbol{\beta}}_n - \hat{\boldsymbol{\beta}}_0\| = \bigcirc_{\mathbb{P}}(\frac{\lambda}{\sqrt{n}})$ through the first part of the Theorem. Then, $\frac{1}{n}\|\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n\| = \bigcirc_{\mathbb{P}}(\frac{\lambda}{n}) = \bigcirc_{\mathbb{P}}(\frac{\log(n)^{1/2}}{n})$.

References

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

Bai, Z., Rao, C. and Wu, Y. (1992). M-estimation of multivariate linear regression parameters under a convex discrepancy function. *Statistica Sinica*, **2**, 237–254.

Chang, X. and Qu, L. (2004). Wavelet estimation of partially linear models. *Computationnal statistics and data analysis*, **47**(1), 31-48.

Charbonnier, P., Blanc-Feraud, G. and Barlaud, M. (1997). Deterministic edge-preserving regularization in computed imaging. *Transactions on Image Processing*, **6**(2), 298–311.

Chen, H. (1987). Estimation of semiparametric generalized linear models. Technical Report. State University of New York.

Chen, H. (1988). Convergence rates for parametric components in a partly linear model. *The Annals of Statistics*, **16**(1), 136–146.

Chen, H. and Chen, K.-W. (1991). Selection of the splined variables and convergence rates in a partial spline model. *The Canadian Journal of Statistics*, **19**(3), 323–339.

Chen, H. and Shiau, J.-J. H. (1991). A two-stage spline smoothing method for partially linear models. *Journal of Statistical Planning and Inference*, **27**, 187–201.

Dahyot, R., Charbonnier, P. and Heitz, F. (2004). A bayesian approach to object detection using probabilistic appearance-based models. *Pattern Analysis and Applications*, **7**, 317–332.

Dahyot, R. and Kokaram, A. (2004). Comparison of two algorithms for robust M-estimation of global motion parameters. *http://citeseer.ist.psu.edu/709403.html*.

Donald, S. and Newey, W. (1994). Series estimation of semilinear models. *Journal of Multivariate Analysis*, **50**, 30–40.

Donoho, D. (1992). De-noising by soft-thresholding. Technical Report. Department of statistics, Stanford University.

Donoho, D., Johnstone, I., Kerkyacharian, G. and Picard, D. (1995). Wavelet shrinkage: asymptotia? *Journal of Royal Statistics Society*, **57**(2), 301–369.

Donoho, D. L. and Johnstone, I. M. (1998). Minimax estimation via wavelet shrinkage. *Annals of Statistics*, **26**(3), 879–921.

Engle, R., Granger, C., Rice, J. and Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association*, **81**(394), 310–320.

Fadili, J. and Bullmore, E. (2005). Penalized partially linear modelsusing sparse representation with an application to fMRI time series. *IEEE Transactions on signal processing*, **53**(9), 3436–3448.

Geman, D. and Reynolds, G. (1992). Constrained restoration and the recovery of discontinuities. *IEEE Transactions of pattern Analysis of machine intelligence*, **14**, 367–383.

Geman, D. and Yang, C. (1995). Nonlinear image recovery with half-quadratic regularization. *IEEE Transaction on Image Processes*, **4**, 932–946.

Green, P. and Yandell, B. (1985). Semi-parametric generalized linear models. Technical Report No. 2847. University of Wisconsin-Madison.

Hamilton, S. and Truong, Y. (1997). Local estimation in partly linear models. *Journal of Multivariate Analysis*, **60**, 1–19.

Hampel, F. R., Rousseeuw, P. J., Ronchetti, E. and Stahel, W. A. (1986). *Robust statistics: The approach based on influence functions.* Wiley Series in probability and Mathematical Statistics.

Hardle, W., Liang, H. and Gao, J. (2000). *Partially linear models.* New-York: Springer-Verlag.

Huber, P. (1981). *Robust statistics.* Wiley Series in probability and Mathematical Statistics.

Mallat, S. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE transactions on pattern analysis and machine intelligence*, **11**(7), 674–693.

Mallat, S. (1999). *A wavelet tour on signal processing. (2 ed.).* Academic press.

Meyer, F. (2003). Wavelet-based estimation of a semiparametric generalized linear model of fMRI time-series. *IEEE transactions on medical imaging*, **22**, 315–324.

Nikolova, M. and Ng, M. (2005). Analysis of half-quadratic minimization methods for signal and image recovery. *SCIAM Journal of Scientific Computing*, **27**(3), 937–966.

Rice, J. (1986). Convergence rates for partially splined models. *Statistics ans Probability Letters*, **4**, 203–208.

Rockafellar, R. (1970). *Convex analysis.* Princeton University Press.

Schick, A. (1996). Root-n-consistent and efficient estimation in semiparametric additive regression models. *Statistics ans Probability Letters*, **30**, 45–51.

Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of Royal Statistical Society*, **50**(3), 413–436.

Vik, T. (2004). Modèles statistiques d'apparence non gaussiens. Application à la création d'un atlas probabiliste de perfusion cérebrale en imagerie médicale. Ph. D. dissertation, Université Strasbourg 1.