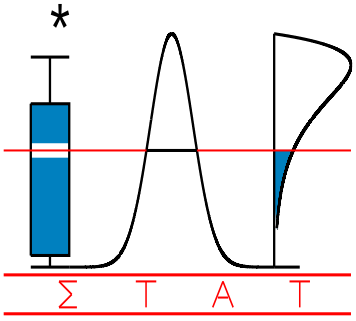
T E C H N I C A L

R E P O R T

0623

ESTIMATION OF PIECEWISE-SMOOTH FUNCTIONS BY AMALGAMATED BRIDGE REGRESSION SPLINES

ABRAMOVICH F., ANTONIADIS A., and M. PENSKY



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

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

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

Estimation of piecewise-smooth functions by amalgamated bridge regression splines

Felix Abramovich

Department of Statistics and Operations Research, Tel Aviv University

Tel Aviv 69978, Israel

Anestis Antoniadis,

Laboratoire IMAG-LMC, University Joseph Fourier,

BP 53, 38041 Grenoble Cedex 9, France.

Marianna Pensky[*]

Department of Statistics , University of Central Florida,

Orlando, FL 32816 -1364, USA.

June 6, 2006

Abstract

We consider a nonparametric estimation of an one-dimensional piecewise-smooth function observed within white Gaussian noise on the interval. We propose the two-step estimation procedure, where one first detects jump points by a wavelet-based procedure and then estimates the function on each smooth segment separately by bridge regression splines. We prove the asymptotic optimality (in the minimax sense) of the resulting amalgamated bridge regression spline estimator and demonstrate its efficiency on several simulated and a real data examples.

Key words: amalgamation; bridge regression; jumps detection; nonparametric regression; penalized regression splines; wavelets.

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

1 Introduction

In variety of nonparametric regression applications, the underlying response function is piecewise-smooth with abrupt changes between smooth segments. To cite only a few examples, we mention seismology, where the density of the sedimentary layers of the earth's crust can be locally approximated by a step function; image processing, where discontinuities present at the edges and econometric models, where structural changes due to governmental policies are not rare. "Direct" methods for estimating piecewise-smooth functions in nonparametric regression include wavelets that are known to efficiently tackle local singularities. However, in practice wavelets often produce pseudo-Gibbs phenomena and other local artifacts in reconstructing smooth regions (e.g. Coifman and Donoho, 1995; Antoniadis and Gijbels, 2002). Alternatively, following a two-step segmentation approach, one first detects the locations of change points and then applies some smooth nonparametric techniques on each segment separately (e.g. Oudshoorn, 1998; Antoniadis and Gijbels, 2002; Fink and Wells, 2004).

In this paper we consider the latter approach. We present a wavelet-based method for detecting discontinuities (jumps) of a function and then introduce amalgamated penalized regression splines for estimating the function at smooth regions. Multiresolutional nature of wavelet analysis makes it to be an excellent tool for detecting local singularities (Mallat and Hwang, 1992; Wang, 1995), while penalized regression splines is a popular statistical techniques for recovering smooth functions from noisy data due to their various optimal properties, good practical performance and computational simplicity (Eilers and Marx, 1996). For the fixed knots the traditional l_2-penalty leads to linear shrinkage (essentially ridge regression) estimator. In this paper, we consider a more general l_ρ-type penalty for $\rho > 0$. Such an approach has a direct analogy with the bridge regression of Frank and Friedman (1993) and we will call the resulting splines *bridge regression splines*. In particular, $\rho = 1$ corresponds to the LASSO estimator of Tibshirani (1996). Generally, l_ρ-penalties for $0 < \rho \leq 1$ lead to (nonlinear) spline estimators with fewer knots.

The proposed procedure shows some similarities with the recent adaptive multi-order penalized splines (AMPS) hybrid procedure of Fink and Wells (2004) who estimate the locations of jumps of a piecewise-smooth function on the basis of the first differences of the data and then fit regression splines using quadratic penalty. In terms of wavelet analysis, such jump detection corresponds to application of Haar wavelets at the finest

resolution level. As a result, AMPS procedure is not powerful enough, does not attain the optimal rates and can detect only sufficiently sharp jumps. The wavelet-based procedure proposed here is somewhat similar in spirit to that of Wang (1995) and is shown to be optimal for jumps detection. In addition, we argue that the use of the l_ρ-penalty with $0 < \rho \leq 1$ results in a spline estimator with fewer knots.

The rest of the paper is organized as follows. Section 2 contains a description of the model and a brief necessary background on wavelet transforms and amalgamated regression splines. We present the two-step amalgamated bridge regression spline estimation procedure (ABS) in Section 3 and establish its optimality in Section 4. Section 5 illustrates the performance of the ABS on several simulated and a real data examples. Some concluding remarks are made in Section 6. All the proofs are given in the Appendix.

2 Formulations and background

2.1 The model

Consider a standard nonparametric regression model

$$Y_i = f(x_i) + \sigma Z_i, \quad 0 \leq x_1 < ... < x_n \leq 1, \tag{1}$$

where f is an unknown response function and Z_i are i.i.d. standard normal $N(0,1)$.

Assume that f is a piecewise-smooth function. More precisely, assume that f belongs to the amalgam Sobolev ball $\mathcal{H}(m, R, \kappa, S)$ of radius R of functions satisfying the following conditions :

M1: $f \in L_\infty([0,1])$.

M2: f has D discontinuity (jump) points at locations $0 < \theta_1 < \cdots < \theta_D < 1$, where the integer D and the real θ_l's are unknown and $\theta_{l+1} - \theta_l > \kappa$, $l = 1, \ldots, D-1$ for some $\kappa > 0$. In particular, $D = 0$ corresponds to a continuous f.

M3: At each discontinuity point θ_l, the left and right limits $f(\theta_l-)$ and $f(\theta_l+)$ exist and $|f(\theta_l+) - f(\theta_l-)| \geq S$ for some $S > 0$.

M4:

$$\sum_{l=0}^{D} \int_{\theta_l}^{\theta_{l+1}} [f^{(m)}(x)]^2 \, dx \leq R,$$

where integer $m \geq 1$, $\theta_0 = 0$ and $\theta_{D+1} = 1$.

Note that the condition **M2** implies that the number of change points D is finite and bounded from above by $1/\kappa < \infty$.

The statistical challenges in estimating a piecewise-smooth function from $\mathcal{H}(m, R, \kappa, S)$ are therefore

1. estimating the number of jumps D and their locations $\theta_l,\ l = 1, ..., D$;

2. recovering the function at smooth regions without degrading its discontinuities.

As we have mentioned in the Introduction, we propose the following two-step procedure: first, to detect the jump points by wavelet-based procedure and then to apply amalgamated bridge regression splines for estimating f between them. We start with some brief background on wavelet transforms and amalgamated polynomial splines.

2.2 Wavelet transforms

In this Section we recall several wavelet transforms and their applications for detection function's singularities.

A wavelet ψ is a fixed function in $L_1(\mathbb{R}) \cap L_2(\mathbb{R})$ such that $\int \psi(t)dt = 0$ satisfying an admissibility condition (see Mallat, 1989). Let $\psi_{a,b}(x)$ be its dilation with a scale parameter a and translation by b: $\psi_{a,b}(x) = a^{-1/2}\psi((x-b)/a)$. The *continuous wavelet transform* (CWT) of a function $f \in L_2(\mathbb{R})$ is defined then as

$$\mathrm{CWT}_f(a, b) = \langle f, \psi_{a,b} \rangle = \frac{1}{\sqrt{a}} \int f(t)\, \psi\left(\frac{t-b}{a}\right) dt,$$

The condition $\int \psi(x)\, dx = 0$ implies that ψ oscillates. More generally, one can choose ψ with N vanishing moments satisfying $\int t^k \psi(t)\, dt = 0$, for $k = 0, \ldots, N$. An important consequence of vanishing moment conditions is that the global and the local Hölder regularities of a function f is characterized by the rate of decay of the modulus of its continuous wavelet transform $\mathrm{CWT}_f(a, b)$ across scales (see Jaffard, 1989 and Mallat, 1989 for more details).

In practical applications however one typically deals with discretely sampled, rather than continuous functions. Given a vector $\mathbf{f} = (f_1, ..., f_n)'$ of $n = 2^J$ values of the function f at equally spaced points $x_1, ..., x_n$, consider the CWT on the dyadic scales at the finest

grid, that is $\mathrm{CWT}_f(2^{-j}, k)$, $j = 0, ..., J-1$, $k = 0, ..., n-1$ (see e.g. Abry, 1994; Vidakovic, 1999). It turns out that the latter is equivalent to the non-decimated wavelet transform NWT (Shensa, 1992) also known as stationary or translation invariant discrete wavelet transform (Coifman and Donoho, 1995; Nason and Silverman, 1995). Its order of complexity is $O(n \log n)$ both for memory allocation and numerical computation (see Dutilleux, 1989; Shensa, 1992). The redundant NWT generates the equal number of n coefficients at each of J scales. The local Hölder regularity of a function at some point can be still characterized by the rate of decay of its NWT coefficients at large scales in the neighbourhood of this point provided that the mother wavelet ψ is sufficiently regular (Berkner and Wells, 1997).

The well-known discrete wavelet transform (DWT) delivers a set of n discrete wavelet coefficients $d_{jk} = \mathrm{CWT}_f(2^{-j}, 2^{-j}k) = 2^{j/2} \int f(x)\psi(2^j x - k)dx$, $k = 0, ..., n-1$. Fast algorithms of Mallat (1989) allows one to perform the DWT in $O(n)$ operations. However, unlike in the case of the NWT, a singularity point cannot be necessarily detected by the presence of "large" DWT coefficients in its neighbourhood on *each* sufficiently large scale due to the possible presence of other singularities or strong oscillations around this point (Mallat and Hwang, 1992). The discrete grid on scales for DWT is too "crude" to characterize local Hölder regularity. Nevertheless, as we shall see, for *piecewise-smooth* functions satisfying the conditions of Section 2.1, the DWT still does the job.

2.3 Amalgamated regression splines

In this subsection we present basic definitions and some background on regression splines. For general theory of spline approximations we refer the reader to de Boor (1978) and Dierckx (1993).

Polynomial splines of order p with knots $\xi_1 < \cdots < \xi_K$ are continuous piecewise $p-1$ degree polynomials with $p-2$ continuous derivatives at the knots. Any polynomial spline $s(x)$ of order p with knots $\xi_1 < \cdots < \xi_K$ can be represented as

$$s(x) = \sum_{k=0}^{p-1} \beta_k x^k + \sum_{j=1}^{K} \beta_{p-1+j}(x - \xi_j)_+^{p-1} \tag{2}$$

where $z_+ = \max(0, z)$.

Polynomial splines are useful for approximating smooth functions but evidently inappropriate for fitting functions with abrupt local changes. To model sharp local

features of a function efficiently, one can consider more general and flexible *multi-order* splines of the form

$$s(x) = \sum_{k=0}^{p-1} \beta_k x^k + \sum_{j=1}^{K} \beta_{p-1+j}(x-\xi_j)_+^{p_j}, \tag{3}$$

where the smoothness $0 \le p_j \le p-1$ at different knots ξ_j may vary (Koo, 1997; Fink and Wells, 2004). Multi-order splines allows jumps in the p_j-th derivative at ξ_j. In particular, zero-order knots ($p_j = 0$) model discontinuities of the function while first and second order knots allow one to represent sharp changes in local linear trend and local curvature respectively. Standard polynomial splines (2) of order p correspond to the particular case when $p_j = p-1$ for all $j = 1, ..., K$. A piecewise-smooth function with D jumps $\theta_1, ..., \theta_D$ can be approximated by a multi-order spline with D zero-order knots at jump points $\theta_l,\ l = 1, ..., D$ and a set of $p-1$-order knots at smooth segments (Fink and Wells, 2004). However, as it follows from (3), such a multi-order spline necessarily implies the conditions on one-sided derivatives at jump points, namely, $s^{(j)}(\theta_l-) = s^{(j)}(\theta_l+),\ j = 1, ..., p-1$. Additional flexibility can be achieved if one considers *amalgamated polynomial regression splines* of order p with zero-order knots $\theta_1, ..., \theta_D$ obtained by amalgamation of separate p-order splines at each segment. An amalgamated polynomial spline $s(x)$ of order p with D zero-order knots $\theta_1, ..., \theta_D$ and q knots $\xi_1, ..., \xi_q$ of order $p-1$ can be represented then as

$$s(x) = s_0(x)I_{\{0 \le x < \theta_1\}} + s_1(x)I_{\{\theta_1 \le x < \theta_2\}} + \cdots s_D(x)I_{\{\theta_D \le x \le 1\}}, \tag{4}$$

where each $s_l,\ l = 0, ..., D$ is a polynomial spline of order p with q_l knots located at $\xi_{1,l}, ..., \xi_{q_l,l}$ and $\sum_{l=0}^{D} q_l = q$.

3 Amalgamated bridge regression splines

We are now ready to propose a two-step adaptive amalgamated bridge regression spline (ABS) estimation procedure. We start with the estimation of a piecewise-smooth function with the *known* jump points by an amalgamated bridge regression spline and then present a wavelet-based procedure for adaptive estimation of jump points from the data.

3.1 Amalgamated bridge regression splines

Assume that $f \in \mathcal{H}(m, R, \kappa, S)$, where the jump points $0 < \theta_1 < ... < \theta_D < 1$ are *known*. Consider an amalgamated m-order spline estimator $\tilde{f}$ of f of the form (4) with q_n knots $\xi_1, ..., \xi_{q_n}$ of order $m-1$ and D knots $\theta_1, ..., \theta_D$ of order zero placed at jump points.

Re-number the observations and the $m-1$-order knots $\xi_1, ..., \xi_{q_n}$ using double indices $(x_{i,l}, Y_{i,l})$ and $\xi_{\nu,l}$, $i = 1, \cdots, n_l$, $\nu = 1, ..., q_{n_l}$, $l = 0, ..., D$, respectively, where $\theta_l \le x_{i,l} < \theta_{l+1}$ and $\theta_l \le \xi_{\nu,l} < \theta_{l+1}$. Using (2) and (4), $\tilde{f}$ can be represented by:

$$\tilde{f}(x) = \sum_{l=0}^{D} \left[\sum_{k=0}^{m-1} \beta_{k,l} x^k + \sum_{\nu=1}^{q_{n_l}} \beta_{m-1+\nu,l} (x - \xi_{\nu,l})_+^{m-1} \right] I(\theta_l \le x < \theta_{l+1}), \tag{5}$$

where $\sum_{l=0}^{D} q_{n_l} = q_n$.

By the definition of an amalgamated spline, on each interval $\theta_l \le x < \theta_{l+1}$, $\tilde{f}(x)$ is a usual m-order polynomial spline with q_{n_l} knots located at $\xi_{1,l}, ..., \xi_{q_{n_l},l}$. Unless some prior information is available, the $m-1$-order knots $\xi_{1,l}, ..., \xi_{q_{n_l},l}$ are usually placed on the sufficiently dense equidistant grid. An excessive number of $m-1$-order knots might imply too much variability in the resulting spline estimator, so one needs some regularization procedure to remove superfluous $\xi_{\nu,l}$ within each segment. We present an example of such a procedure below. The jump points are assumed meanwhile to be known, and, hence, using the representation (5), one can estimate the vector of unknown coefficients $\boldsymbol{\beta}_l$ on each l-th segment separately.

Let $\mathbf{X}^{(l)}$ be the $n \times (m + q_{n_l})$ matrix with the rows

$$(1, x_{i,l}, \cdots, x_{i,l}^{m-1}, (x_{i,l} - \xi_{1,l})_+^{m-1}, \cdots, (x_{i,l} - \xi_{q_{n_l},l})_+^{m-1})$$

and $\mathbf{Y}_j$ be the vector with components $Y_{i,l}$, $i = 1, \cdots, n_l$.

Consider the penalized maximum likelihood estimator of $\boldsymbol{\beta}_l$ with l_ρ-penalty, $\rho > 0$, derived by minimizing

$$Q_l(\boldsymbol{\beta}_l, \mathbf{Y}_l) = \|\mathbf{Y}_l - \mathbf{X}^{(l)} \boldsymbol{\beta}_l\|^2 + n_l \lambda_{n_l} \sum_{k=m}^{m-1+q_{n_l}} |\beta_{k,l}|^\rho \tag{6}$$

with respect to $\boldsymbol{\beta}_l$, where $\lambda_{n_l} > 0$ is a smoothing parameter.

The idea of l_ρ-penalty in regression was introduced in Frank and Friedman (1993) and the corresponding technique is known as the *bridge regression* estimation. The traditional l_2-penalty yields a ridge regression estimator which is based on linear shrinkage, while

$\rho = 1$ leads to the LASSO estimator of Tibshirani (1996). Any choice $0 < \rho \leq 1$ implies a thresholding estimator of $\boldsymbol{\beta}_l$ and, therefore, results in a spline with fewer $m-1$-order knots (see Antoniadis and Fan, 2001). Plugging coefficients $\tilde{\boldsymbol{\beta}}_{k,l}$ into (5) leads to the amalgamated bridge regression spline estimator $\tilde{f}(x)$ of $f(x)$.

To conclude this section, note that we have used truncated power bases for a clearer exposition of a spline-based regression. However, the truncated power bases, especially when a larger number of knots and a small penalty parameter are involved, may lead to numerical instabilities. Equivalent bases with more stable numerical properties are the B-spline bases, and it is easy to routinely transform the matrices $\mathbf{X}^{(l)}$ to the versions which are much more stable numerically by changing bases using an invertible linear transform. For this reason, we shall not further discuss numerical stability issues when we formulate the ABS estimator.

3.2 Jumps detection

The general idea behind wavelet-based detection of abrupt changes of a function is based on the characterization of its local regularity at a point by the rate of decay of its wavelet coefficients across scales around this point. Local singularities can be identified then by the presence of large wavelet coefficients at high scales in their neighbourhood.

The detection algorithm described below analyzes wavelet coefficients at an appropriately chosen scale and selects a threshold large enough to prevent the coefficients corresponding to smooth segments to penetrate by, but still small enough to allow coefficients corresponding to singularities to pass it through. The locations of jumps are then estimated by the locations of coefficients which exceed the threshold.

As we have mentioned in Section 2.2, generally, the DWT coefficients cannot be used to detect a local singularity since the discrete grid on scales for the DWT might be too "crude" due to the possible presence of other singularity points or strong oscillations around this point (Mallat and Hwang, 1992). However, we will show that for *piecewise-smooth* functions from amalgam Sobolev balls, the DWT can still be used to detect jump points.

We will assume hereafter that the variance σ^2 of the noise is known, otherwise for regression functions from amalgam Sobolev balls, it can be estimated at a parametric rate in the wavelet domain by the median of the absolute deviation of the empirical wavelet coefficients of the data at the highest resolution level divided by 0.6745.

Let ψ be a mother wavelet with a compact support $[L;U]$, $L < 0 < U$, and $\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k)$, $j = 0, ..., J-1$, $k = 0, ..., 2^j - 1$. Fix an arbitrarily small $\delta > 0$. Define j^* such that $2^{j^*} = (U-L)n/(\ln n)^{1+\delta}$ and a sequence of indices $\tau(k) = -L + (U-L)k$, $k = 0, ..., 2^{j^*}/(U-L) - 1$. Without loss of generality, we may assume that j^* and $2^{j^*}/(U-L)$ are integers; otherwise, we take the corresponding integer parts. Note that $\Omega_{j^*k} = \text{supp}\, \psi_{j^*\tau(k)} = [2^{-j^*}(U-L)k; 2^{-j^*}(U-L)(k+1)]$ and, therefore, the intersection $\Omega_{j^*k} \cap \Omega_{j^*(k+1)}$ is a zero-measure set containing a single boundary point $2^{-j^*}(U-L)(k+1)$. Hence, the unit interval is divided into a grid of $N = 2^{j^*}/(U-L)$ non-overlapping intervals of lengths $2^{-j^*}(U-L) = (\ln n)^{1+\delta}/n$. Due to **M2**, for sufficiently large n, each of these intervals can contain only a single jump point.

Let T_{j^*} be a set of indices $\tau(k)$ such that the corresponding interval Ω_{j^*k} does not contain a jump point. For an arbitrary $0 < \alpha < \delta/2$ define a threshold

$$t_n^* = \sigma \sqrt{\frac{(\log n)^{1+\delta-2\alpha}}{n}}\,. \tag{7}$$

Proposition 3.1 *Consider the model (1), where the unknown $f \in \mathcal{H}(m, R, \kappa, S)$ defined in Section 2.1. Let the wavelet ψ be differentiable with a compact support and $\hat{d}_{jk}$, $j = 0, ..., J-1$, $k = 0, ..., 2^j - 1$ be the set of the DWT coefficients of the noisy data $\boldsymbol{Y} = (Y_1, ..., Y_n)'$. Then, for the threshold t_n^* defined in (7) one has uniformly in $f \in \mathcal{H}(m, R, \kappa, S)$:*

1. $\mathbb{P}(\max_{\tau(k)\in T_{j^*}} |\hat{d}_{j^*\tau(k)}| > t_n^*) = o(n^{-\gamma})$
2. $\mathbb{P}(\min_{\tau(k)\notin T_{j^*}} |\hat{d}_{j^*\tau(k)}| < t_n^*) = o(n^{-\gamma})$

as $n \to \infty$ for an arbitrarily large $\gamma > 0$.

Proposition 3.1 shows that we can track down the jumps by the presence of large DWT coefficients $d_{j^*\tau(k)}$ with very high accuracy.

Remark 3.1 *The proposed jump detection procedure is performed at such a high resolution level j^*, that there are essentially no differences between the DWT and NWT coefficients. This explains why for a piecewise-smooth function f satisfying the conditions of Section 2.1 the DWT can be still used.*

Proposition 3.1 immediately implies that $\mathbb{P}\{\hat{D} \neq D\} = o(n^{-\gamma})$ for an arbitrarily large $\gamma > 0$. Note also that

$$\mathbb{E}(|\hat{\theta}_l - \theta_l|^2 \mathbf{I}_{\hat{D}=D}) \leq (U-L)^2\, 2^{-2j*} + \mathbb{P}\{|\hat{\theta}_l - \theta_l| > (U-L)2^{-j^*}\}, \tag{8}$$

where the first term in the right-hand side of (8) is $O((\ln n)^{2+2\delta}/n^2)$, while the second one is negligible due to the first statement of Proposition 3.1. Hence, the following Corollary holds:

Corollary 3.1 *Under the assumptions of Proposition 3.1, as* $n \to \infty$,

1. $\mathbb{P}\{\hat{D} \neq D\} = o(n^{-\gamma})$ *for any* $\gamma > 0$
2. *Uniformly in* $f \in \mathcal{H}(m, R, \kappa, S)$,

$$\mathbb{E}\left(|\hat{\theta}_l - \theta_l|^2 \mathbf{I}_{\hat{D}=D}\right) = O(2^{-2j^*}) = O\left(\frac{(\ln n)^{2(1+\delta)}}{n^2}\right),\ l = 1, \ldots, \hat{D}$$

Based on the results of this section, we suggest now the following jumps estimation procedure:

1. Consider the DWT coefficients $\hat{d}_{j^*\tau(k)}$ at the level j^* and find all $\tau(k)$ such that $|\hat{d}_{j^*\tau(k)}| > t_n^*$. If the set $\{|\hat{d}_{j^*\tau(k)}| > t_n^*\}$ is empty, set $\hat{D} = 0$. Otherwise,
2. Estimate the number of jump points D by $\hat{D} = \#\{|\hat{d}_{j^*\tau(k)}| > t_n^*\}$ and the locations θ_ℓ of the jumps by the mid-points $\hat{\theta}_\ell$ of the corresponding intervals Ω_{j^*k}, i.e., $\hat{\theta}_\ell = 2^{-j^*}(U-L)(k+1/2),\ l = 1, ..., \hat{D}$.

3.3 The ABS procedure

The resulting two-step ABS procedure naturally combines amalgamated bridge regression spline estimation with jumps detection and can be summarized as follows :

1. estimate the number of jump points $\hat{D}$ and their locations $\hat{\theta}_1, ..., \hat{\theta}_{\hat{D}}$ by the DWT-based procedure described in Subsection 3.2
2. plug in $\hat{D}$ and $\hat{\theta}_1, ..., \hat{\theta}_{\hat{D}}$ into (5) and minimize the resulting expression (6) to obtain an adaptive amalgamated bridge regression spline estimator $\hat{f}$.

4 Optimality of the ABS procedure

In this section we prove the optimality (in the minimax sense) of the proposed ABS procedure.

Consider the quadratic risk (L_2-loss) for an estimator $\hat{f}$ of f:

$$R(\hat{f}, f) = \mathbb{E}\{\|\hat{f} - f\|_2^2\}$$

The minimax quadratic risk over $\mathcal{H}(m, R, \kappa, S)$ is then defined by

$$R\left(\mathcal{H}(m, R, \kappa, S)\right) = \inf_{\hat{f}} \sup_{f \in \mathcal{H}(m,R,\kappa,S)} R(\hat{f}, f),$$

where the infimum is taken over all estimators $\hat{f}$.

Antoniadis and Gijbels (2002) derived the minimax rate over $R\left(\mathcal{H}(m, R, \kappa, S)\right)$ and showed that as n increases,

$$R\left(\mathcal{H}(m, R, \kappa, S)\right) = \mathcal{O}\left(n^{-\frac{2m}{2m+1}}\right) \tag{9}$$

An analogous result under the additional assumption $\sup_{x \in [0,1]} |f(x)| < B$ was established in Oudshoorn (1998). Note that the optimal rate (9) for estimating piecewise-smooth functions from amalgam Sobolev classes satisfying **M1**–**M4** is the same as for homogeneously smooth functions from usual Sobolev spaces.

We now show that the proposed ABS estimator attains the minimax rate (9). We first prove that for the *fixed* zero-order knots the amalgamated bridge regression spline estimator achieves the optimal rate $\mathcal{O}(n^{-2m/(2m+1)})$ and then demonstrate that the accuracy of the zero-knots estimation procedure is sufficiently high not to damage it.

Consider the amalgamated bridge regression spline estimator $\tilde{f}$ from Section 3.1 with q_n equally spaced $m-1$-order knots and *fixed* zero-order knots. Impose the following asymptotic assumptions on the design matrix $\mathbf{X}$, the number of $m-1$-order knots q_n and the smoothing parameters λ_{n_l} in (6) as $n \to \infty$ and $q_n \to \infty$:

M5: There exists $C_1 > 0$ and $C_2 > 0$ such that $0 < C_1 n < \lambda_{\min} < \lambda_{\max} < C_2 n$, where $\lambda_{\min}$ and $\lambda_{\max}$ are the minimal and the maximal eigenvalues of the matrix $\mathbf{X}^T\mathbf{X}$.

M6: $q_n = C n^{1/(2m+1)}$ for some $C > 0$.

M7: $\lambda_{n_l} n_l^{1-\rho/2} = O(1)$ as $n_l \to \infty,\ l = 1, ..., D$.

Due to the assumption **M2**, the number of observations n_l and the number of $(m-1)$-order knots q_{n_l} on each l-th segment are of the order n and q_n respectively, the same as on the entire unit interval, and, therefore, the assumptions **M5** and **M6** holds on each segment as well. Thus, when the zero-order knots are fixed, the asymptotic properties of the amalgamated estimator $\tilde{f}$ on the entire unit interval are the same as on each of its segments.

Consider then minimization of (6) under the assumptions $\mathbf{M5} - \mathbf{M7}$. The following proposition guarantees the existence of a local $\sqrt{n_l/q_{n_l}}$-consistent penalized maximum likelihood estimator $\tilde{\boldsymbol{\beta}}_l$ in (6) of $\boldsymbol{\beta}_l$:

Proposition 4.1 *Under assumptions* **M5** *and* **M7**, *there exists a local minimizer* $\tilde{\boldsymbol{\beta}}_l$ *of (6) such that* $\|\tilde{\boldsymbol{\beta}}_l - \boldsymbol{\beta}_l\| = O_p(\sqrt{q_{n_l}/n_l}) = O_p(\sqrt{q_n/n})$.

Proposition 4.1 only establishes the existence of a local $\sqrt{q_n/n}$ consistent minimizer of (6) but does not provide any tools to obtain it. A closed solution is available for $\rho = 2$. For $\rho = 1$ the minimizer of (6) is unique and can be found either by a LASSO-type algorithm (Tibshirani, 1996; Osborne *et al.*, 2000), once the matrices $\mathbf{X}^{(l)}$ are normalized to have columns of norm 1, or via surrogate functionals, a method recently introduced by Daubechies *et al.* (2004) in the context of wavelet shrinkage methods for deblurring. When $\rho < 1$, the objective function is no longer convex but one can still find a local minimizer using, for example, an approximate algorithm of Ruppert and Carrol (2000), a backfitting type algorithm of Fu and Kneight (2000) or a recently developed algorithm of Amato *et al.* (2006). We discuss these issues in Section 5 below.

The resulting ABS estimator $\tilde{f}$ is obtained by amalgamation of the corresponding estimators at each segment, that is, $\tilde{f} = X\tilde{\boldsymbol{\beta}}$. The following proposition shows that $\tilde{f}$ achieves the optimal rates (9).

Proposition 4.2 *Let assumptions* $\mathbf{M5}$ – $\mathbf{M7}$ *hold and* $\tilde{f} = X\tilde{\boldsymbol{\beta}}$. *Then,*

$$\sup_{f \in \mathcal{H}(m,R,\kappa,S)} R(\tilde{f}, f) = O\left(n^{-2m/(2m+1)}\right). \tag{10}$$

So far we considered an idealized situation where the jumps locations were assumed to be known. The following proposition shows that when zero-order knots are *estimated* by the wavelet-based jumps detection procedure from Section 3.2, the resulting ABS estimator $\hat{f}$ still attains the optimal rates (9). The high accuracy of estimating jump

points makes the additional error contribution to be negligible in the overall estimation error.

Proposition 4.3 *Let assumptions* **M1**-**M7** *hold and let* $\hat{f}$ *the ABS estimator with zero-order knots estimated by the wavelet-based procedure proposed in Section 3.2. Then, as* $n \to \infty$,

$$\sup_{f \in \mathcal{H}(m,R,\kappa,S)} R(\hat{f}, f) = O\left(n^{-2m/(2m+1)}\right). \tag{11}$$

5 Simulations and real examples

Following the proposed two-step ABS algorithm, we first estimate the number and locations of jumps of an unknown function by the wavelet-based detection procedure. We then place zero order knots on estimated jump points and a relatively large number of quadratic knots ($m = 3$) at locations fixed at "equally-spaced sample quantiles" similarly to standard penalized splines designed for estimating smooth functions (e.g. Ruppert and Carroll, 2000). As it has been noted in the previous sections, this step serves to refine the regression spline basis by allowing for additional smoothness between any zero-order features. During the second step, the ABS regression spline is fitted using either penalized least squares with a quadratic penalty or with a more general l_ρ-type penalty, penalizing coefficients of the basis functions which are the least supported by the data.

Recall that the bridge regression estimators $\tilde{\boldsymbol{\beta}}_j$ of coefficients $\boldsymbol{\beta}_j$ within each segment are obtained by minimizing Q_l in (6) separately for each $l = 0, \ldots, D$ with respect to $\boldsymbol{\beta}_j$:

$$Q_l(\boldsymbol{\beta}_l, \mathbf{Y}_l) = \|\mathbf{Y}_l - \mathbf{X}^{(l)}\boldsymbol{\beta}_l\|^2 + n_l \lambda_{n_l} \sum_{k=m}^{m-1+q_{n_l}} |\beta_{k,l}|^\rho$$

Suppressing index l in the equation above for simplicity and clarity, we consider minimization of

$$Q(\boldsymbol{\beta}, \mathbf{Y}) = \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \sum_{k=m}^{m-1+q_{n_l}} |\beta_k|^\rho.$$

As we have mentioned above, a closed solution is available for $\rho = 2$, while for $\rho = 1$ there exist numerical algorithms (e.g. LASSO). For any $0 < \rho \leq 1$, a possible numerical solution is to minimize Q iteratively, one component of $\boldsymbol{\beta}$ at a time (backfitting). Assume for simplicity that $\bar{Y} = 0$ (or replace Y_i by $Y_i - \bar{Y}$). The algorithm we have used in our numerical implementation for $\rho \leq 1$ can be described then as follows:

(0) Center the columns of $\mathbf{X}$ to have the mean 0 and scale them to have the unit variance. Using centered columns, define an initial value $\hat{\boldsymbol{\beta}}$ by using the least squares algorithm. Set $k = 1$.

(1) Define

$$Q_k(\beta_k) = \sum_{i=1}^{n}(Y_i - \sum_{j \neq k} \hat{\beta}_j x_{ij} - \beta_k x_{ik})^2 + \lambda|\beta_k|^\rho.$$

(2) Set $\hat{\beta}_k = \arg\min Q_k$. The minimization of Q_k with respect to β_k is carried out by Newton-Raphson or fixed-point iteration.

(3) If $k < m - 1$, set $k = k + 1$ else set $k = 1$.

(4) Repeat (1), (2) and (3) until convergence occurs.

The above algorithm works very well if the design is not "too collinear" (hence the interest in using B-splines), otherwise it might get stuck at a local minima. The problem is less severe when ρ is not too close to 0. For $\rho = 1$ it may be also computationally simpler than LASSO that involves linear programming techniques.

In the simulation and real data examples below we considered $\rho = 2$ (ABS2) and $\rho = 1$ (ABS1), where in the latter we used the backfitting algorithm described above. The quadratic l_2-penalty (ABS2) is equivalent to placing quadratic penalties on finite differences of adjacent B-splines coefficients and it results in shrinking all coefficients toward zero. On the contrary, the l_1-penalty on adjacent B-splines coefficients (ABS1) not only shrinks the coefficients but also thresholds them removing, therefore, the corresponding "superfluous" second order knots.

In both approaches the smoothing parameter λ_n was automatically chosen from the data by generalized cross-validation (GCV) as usual in spline smoothing (see e.g. Fan and Li, 2001).

5.1 Simulations

In this subsection we compare the estimates based on the ABS1 and ABS2 procedures with another related method, namely, the Spatially Adaptive Regression Splines (SARS) developed by Zhou and Shen (2001) which is particularly suited for functions that have jumps by themselves or in their derivatives. SARS is locally adaptive to variable

smoothness and automatically places more knots in the regions where the function is not smooth. It has been proved as an effective tool for estimating such functions. For completeness, we also compare the above estimators with a standard wavelet denoising procedure based on universal thresholding of Donoho and Johnstone (1994), since wavelet based procedures are known to efficiently denoise inhomogeneous functions.

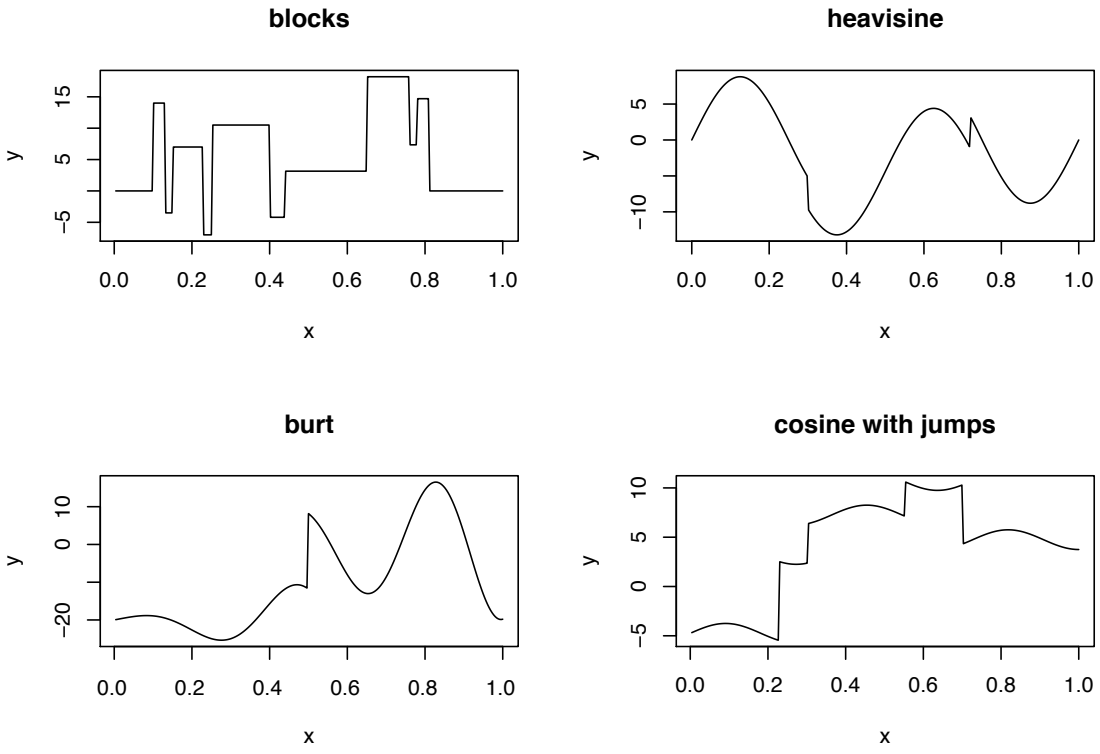


Figure 5.1: Four benchmark test functions punctuated by jump discontinuities used in the simulations.

To investigate the performance of the developed ABS estimators we conducted a simulations study based on synthetic data. We used two of the standard test functions of Donoho and Johnstone (1994) that are examples of piecewise-smooth functions and commonly used for various wavelet procedures, namely, the `blocks` and the `heavisine`. In addition, we considered two other test functions punctuated by jump discontinuities called `burt` and `cosine`, defined on $[0, 1]$, respectively, as

$$\texttt{cosine}(x) = \cos(5.5\pi x) - 4\,\mathrm{sign}(0.23 - x) - 2\,\mathrm{sign}(0.3 - x) - 1.75\,\mathrm{sign}(0.55 - x) + 3\,\mathrm{sign}(0.7 - x)$$

and

$$\texttt{burt}(x) = 20x\cos(16x^{1.2}) - 20I(x < 0.5).$$

The functions are depicted in Figure 5.1. The sample size used in the simulations was $n = 256$ and the design points were uniformly spaced within the unit interval. In each simulation, we added a normally distributed zero mean additive noise with the standard deviation σ implying a chosen signal-to-noise-ratio (SNR). SNR is measured as $sd(f)/\sigma$, where $sd(f)$ is the estimated standard deviation of the regression function over the grid. For each function and two values of SNR (4 and 6), we ran 100 simulations using 50 equally-spaced quadratic knots. The noise level σ was assumed unknown and estimated by the median of the absolute deviation of the empirical wavelet coefficients of the data at the highest resolution level divided by 0.6745. For each realization we calculated the ABS1 and ABS2 estimators using the procedures developed above, where the wavelet-based jumps detection algorithm was based on Symmlets of order 6, the SARS estimator and a wavelet denoising estimator (Wav) also used Symmlets of order 6. All simulations were performed using Matlab 7 (Mathworks 2001) and R.

The corresponding estimates from a single realization are displayed in Figure 5.2.

The accuracy of an estimator $\hat{f}$ of f was measured by the average mean square error (MSE) averaged over 100 simulation runs, where

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{f}(x_i) - f(x_i))^2$$

and $\{x_i\}$ is the set of design points. The average MSEs for each method are reported in Table 1 and the boxplots in Figure 5.3.

Figure 5.3 and Table 5.1 show that all the three spline estimators with adaptively places knots outperform a standard wavelet denoising procedure. ABS1 and ABS2 provide similar results and in most cases are better than SARS. As one can also see, application of l_1-type penalty in ABS provides a relatively small gain. This can be explained by the fact that, for each of the functions, the regions between any two jumps are relatively smooth.

5.2 A real example

Since the ABS1 has shown the better performance in our simulations we applied the ABS1 procedure to the real life data provided by Mike Battaglia from the Department of Forestry at Virginia Tech (see Battaglia, 2000). The data set contains relative light

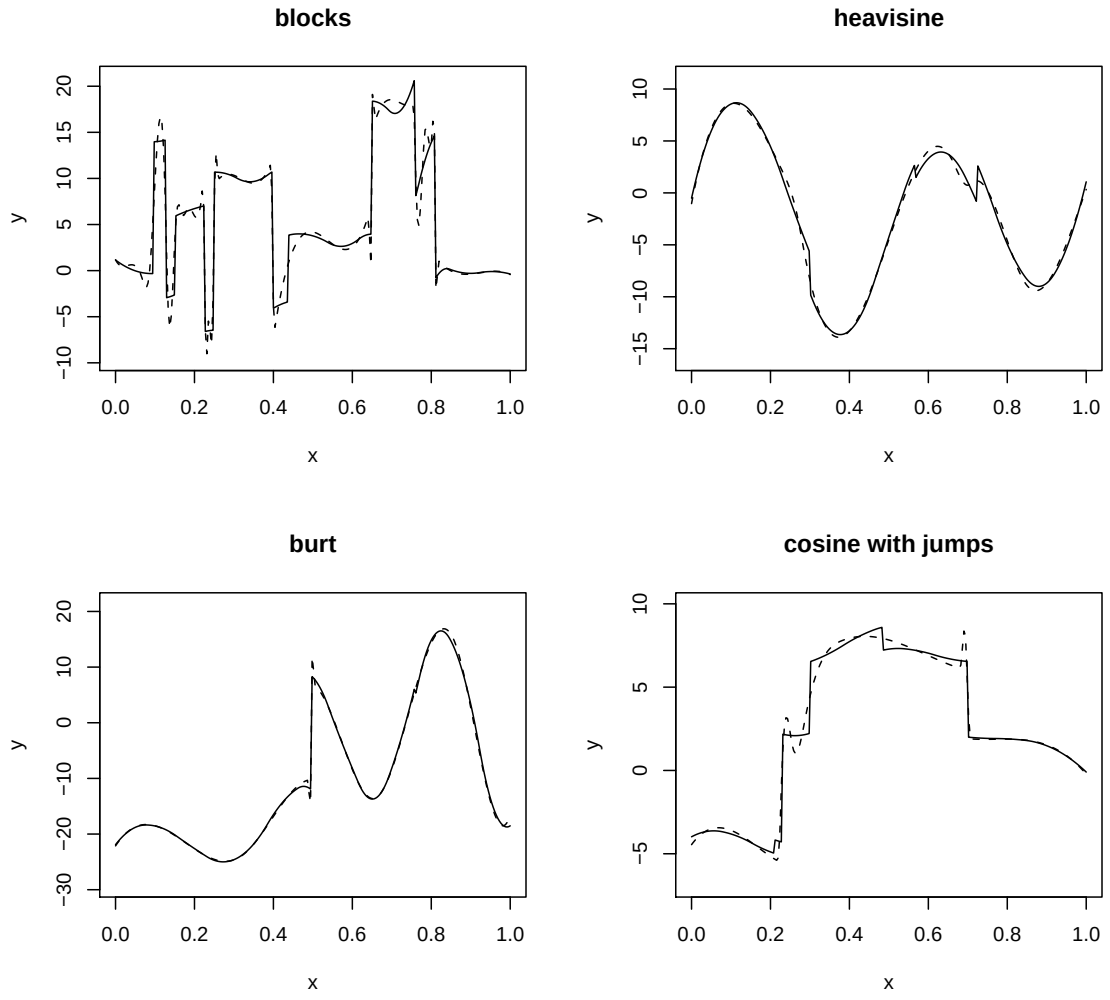


Figure 5.2: Fits from the ABS1 procedure (solid line) and SARS (dashed line) for a single realization with SNR=4.

transmittance data recorded at equal time intervals throughout the daylight hours for numerous days (see Figure 5.4 for a plot of the relative light transmittance data for one station during one day, $n = 164$). In this data set, sun light from various forest stations in plots with different cutting treatments is compared to the sun light in a nearby open plot. Cloud interference and overstory patterns (the shades produced by the trees) are the two most common phenomena that cause jump points in the relative transmittance data. Jump points that remain consistent across days may be attributed to overstory pattern while jump points that do not remain consistent across days are probably caused

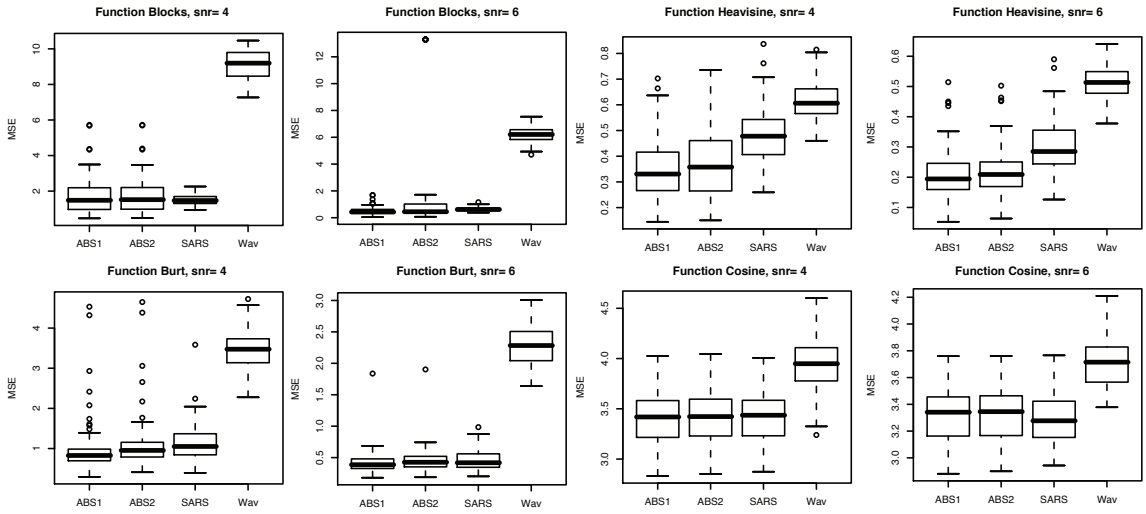


Figure 5.3: Boxplots of the mean squared errors on 100 samples obtained by four different procedures: ABS1 (l_1-penalty); ABS2 (l_2 penalty); SARS, Spatially Adaptive Regression Splines; Wav, Wavelet denoising. The hyperparameters are chosen by generalized cross-validation.

Table 1: Average mean squared errors based on 100 samples obtained using signal to noise ratios of 6 and 4 (in parenthesis) for five different procedures: ABS1 (ℓ_1-penalty); ABS2 (ℓ_2 penalty), SARS and Wav. The hyperparameters are chosen by generalized cross-validation.

estimate	blocks	heavisine	burt	cosine
ABS1	0.52 (1.85)	0.21 (0.35)	0.42 (0.96)	3.32 (3.41)
ABS2	0.58 (1.86)	0.22 (0.38)	0.45 (1.08)	3.33 (3.42)
SARS	0.63 (1.51)	0.30 (0.48)	0.46 (1.13)	3.29 (3.40)
Wav	6.21 (9.11)	0.51 (0.61)	2.27 (3.46)	3.72 (3.95)

by cloud interference. Since variation of the light availability increases as canopy gaps become larger, in order to predict the forest dynamics it is useful to estimate the relative transmittance taking into account such discontinuities.

We applied the ABS1 procedure to the data shown at Figure 5.4. Symmlets of order 6 were again used for jumps detection. The noise level σ, as usual, was estimated by the median of the absolute deviation of the empirical wavelet coefficients of the data at the highest resolution level divided by 0.6745. The procedure found three jump points

located at 0.138, 0.494 and 0.873. For construction of the ABS estimator, on each of the four resulting segments 12 quadratic knots were placed at equally spaced quantiles. The ABS1 reduced the numbers of necessary quadratic knots to 3, 4, 3 and 3 respectively. The Figure 5.4 shows the resulting ABS1 estimate that nicely fits the data. Examination of the fitted nonparametric trends of the daily relative light transmittance data and their jumps can therefore be a powerful tool for the characterization of gap openings in forest stations.

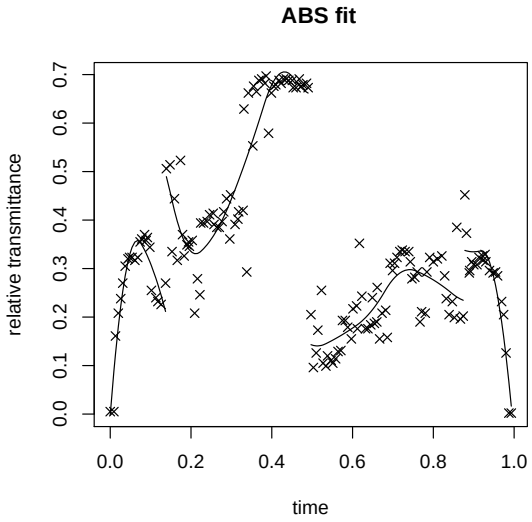


Figure 5.4: Display of the relative light transmittance data (data plotted as stars) for one station during one day from the Forestry Department of Virginia Tech. The x-axis is a daylight hours time interval rescaled to $[0, 1]$. The display also includes the ABS1 regression fit (solid line) to the transmittance data (with 3 jumps detected).

6 Concluding remarks

In this paper we developed a two-step procedure for estimating piecewise-smooth functions by amalgamated bridge regression splines. It first detects the unknown jump points by a wavelet-based method and then estimates the regression function on each smooth segment separately by bridge regression splines. We showed that the resulting amalgamated estimator achieves minimax convergence rates over amalgam Sobolev balls. From a

practical point of view, ABS is fairly accurate and computationally efficient: it requires a little more computation time than traditional penalized regression splines. The number of operations necessary for location of the zero-order knots grows linearly with n, inheriting the algorithmic complexity of the discrete wavelet transform. We demonstrated good performance of ABS on several simulated and a real data examples. Summarizing, we believe that the proposed ABS estimators are an attractive alternative to the existing estimators of piecewise-smooth functions.

7 Appendix

Throughout the proofs in the Appendix, we use C to denote a generic positive constant, not necessarily the same each time it is used, even within a single equation.

7.1 Proof of Proposition 3.1

Applying the DWT to the both sides of (1) we obtain

$$\hat{d}_{jk} = d_{jk} + \epsilon_{jk},$$

where d_{jk} and $\epsilon_{jk},\ j = 0, ..., J-1,\ k = 0, ..., 2^j - 1$ are the DWT of the unknown f and the Gaussian noise respectively. Note that ϵ_{jk} are independent Gaussian variables with $\mathrm{Var}(\epsilon_{jk}) = \sigma^2/n$.

Consider the level j^* and the corresponding sequence of indices $\tau(k)$ defined in Section 3.2. Consider first $\tau(k) \in T_{j^*}$. Since $m \geq 1$, the function f is at least of Lipschitz regularity one for all $x \in \Omega_{j^*k}$. Then, for sufficiently large n,

$$\max_{k \in T_{j^*}} |d_{j^*\tau(k)}| < C 2^{-\frac{3}{2}j^*} = O\left(\left(\frac{(\ln n)^{1+\delta}}{n}\right)^{3/2}\right) = o(t_n^*) \tag{12}$$

(e.g. Daubechies, 1992, p.299).

As n increases, for any $0 < \alpha < \delta/2$ and for all $\tau(k) \in T_{j^*}$

$$\begin{aligned} \mathbb{P}\{|\hat{d}_{j^*\tau(k)}| > t_n^*\} &\leq \mathbb{P}\{|\epsilon_{j^*\tau(k)}| > t_n^* - |d_{j^*\tau(k)}|\} \leq \mathbb{P}\{|\epsilon_{j^*\tau(k)}| > t_n^*/2\} \\ &\leq C(\log n)^{\alpha-(1+\delta)/2} \exp(-(\log n)^{1+\delta-2\alpha}/8) = o(n^{-\tilde{\gamma}}) \end{aligned} \tag{13}$$

for any $\tilde{\gamma} > 1$. Since $card\{T_{j^*}\} = O(2^{j^*}) = O(n/(\ln n)^{1+\delta})$, we have

$$\mathbb{P}\{\max_{k \in T_{j^*}} |\hat{d}_{j^*\tau(k)}| > t_n^*\} \leq \sum_{k \in T_{j^*}} \mathbb{P}\{|\hat{d}_{j^*k}| > t_n^*\} = o(n^{-\gamma}),$$

where $\gamma = \tilde{\gamma} - 1 > 0$.

Let now $\tau(k) \notin T_{j^*}$. The condition **M2** guarantees that for sufficiently large n there is a single jump point $\theta_l \in \Omega_{j^*k}$. Similarly to the arguments of Wang (1995) and Antoniadis and Gijbels (2002) but applied for the DWT, under the conditions **M1**-**M4** we derive that

$$|d_{j^*k}| = 2^{-j^*/2}\,|f(\theta_l+) - f(\theta_l-)|\left\{\left|\int \psi(\theta_l - u)sign(u)du\right| + O(1)\right\} \geq C2^{-j^*/2}. \quad (14)$$

For each $\tau(k) \notin T_{j^*}$ it then follows from (14) that

$$\mathbb{P}\{|\hat{d}_{j^*\tau(k)}| < t_n^*\} \;\leq\; \mathbb{P}\{|\epsilon_{j^*\tau(k)}| > C2^{-j^*/2}\} > C(\log n)^{-(1+\delta)}\exp(-C^2(\log n)^{1+\delta}/2) = o(n^{-\gamma})$$

for any $\gamma > 0$. Hence, for finite D,

$$\mathbb{P}\{\min_{\tau(k)\notin T_{j^*}} |\hat{d}_{j^*\tau(k)}| < t_n^*\} \leq \sum_{k=1}^{D} \mathbb{P}\{|\hat{d}_{j^*\tau(k)}| < t_n^*\} = o(n^{-\gamma})$$

that completes the proof.

□

7.2 Proof of Proposition 4.1

As we have mentioned, the asymptotic properties of $\tilde{\boldsymbol{\beta}}_l$ are same for each l and for simplicity of exposition we omit the index l throughout the proof.

Let $\boldsymbol{\beta}$ be the true set of coefficients. Let M be a relatively large number and $\alpha_n = \sqrt{q_n/n}$. In order to prove Proposition 4.1 we show that

$$\lim_{n\to\infty} P\left\{\sup_{\|\mathbf{u}\|=M} Q(\boldsymbol{\beta} + \alpha_n\mathbf{u}, \mathbf{Y}) > Q(\boldsymbol{\beta}, \mathbf{Y})\right\} = 0. \quad (15)$$

Equality (15) implies that with probability tending to one, there is a local minimum $\tilde{\boldsymbol{\beta}}$ of (6) in the ball with the center $\boldsymbol{\beta}$ and radius $\alpha_n M$ such that $\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\| = O_p(\alpha_n)$. Let $\boldsymbol{\epsilon} = \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$ and

$$\begin{aligned} D_n(\mathbf{u}) &= Q(\boldsymbol{\beta} + \alpha_n\mathbf{u}, \mathbf{Y}) - Q(\boldsymbol{\beta}, \mathbf{Y}) = \Delta_1 + \Delta_2 + \Delta_3 \quad (16)\\ &= \alpha_n^2\|\mathbf{X}\mathbf{u}\|^2 - 2\alpha_n\mathbf{u}^T\mathbf{X}^T\boldsymbol{\epsilon} + n\lambda_n\sum_{k=1}^{q_n}[|\beta_{p+k} + \alpha_n u_k|^\rho - |\beta_{p+k}|^\rho] \quad (17)\end{aligned}$$

Note that Δ_1 is a constant and, by assumption **M5**, $\Delta_1 \geq C_1\alpha_n^2 n\|\mathbf{u}\|^2 = C_1M^2\alpha_n^2 n = C_1M^2q_n$. The second term, Δ_2 is a random variable with $E\Delta_2 = 0$, so that, for any $\delta > 0$ by Markov inequality, and **M5** we have

$$P(|\Delta_2| > \delta) \;\leq\; 2\delta^{-1}\alpha_n\sqrt{E(\mathbf{u}^T\mathbf{X}^T\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\mathbf{X}\mathbf{u})} \leq 2\delta^{-1}M\sigma\alpha_n\sqrt{C_2 n}.$$

Therefore, setting $\delta = 0.5C_1M^2\alpha_n^2 n$, we obtain $P(|\Delta_2| > 0.5C_1M^2\alpha_n^2 n) \leq 4\sigma(C_1M)^{-1}(\alpha_n\sqrt{n})^{-1} \to 0$. For an upper bound for Δ_3, note that since $|x+y|^\rho - |x|^\rho \leq |y|^\rho$ as $0 < \rho \leq 1$, by condition **M7** when M is large enough we have

$$|\Delta_3| \;\leq\; n\lambda_n\alpha_n^\rho\sum_{k=1}^{q_n}|u_k|^\rho \leq n\lambda_n\alpha_n^\rho M^{\rho/2}q_n^{1-\rho/2} < 0.5\,C_1M^2q_n.$$

7.3 Proof of Proposition 4.2

The proof Proposition 4.2 is based on the following lemma.

Lemma 7.1 *Let $w(z,\alpha) = \arg\min_w[w^2 - 2wz + \alpha|w|^\rho]$, $0 < \rho \leq 1$. Then, $w(z,\alpha) = 0$ whenever $|z| < a_\rho\alpha^{1/(2-\rho)}$. If $\rho \neq 1$, then $|w(z,\alpha)| \geq b_\rho\alpha^{1/(2-\rho)}$ whenever $|z| \geq a_\rho\alpha^{1/(2-\rho)}$. Here $a_\rho = (2-\rho)(\rho/2)^{1/(2-\rho)}(1-\rho)^{(\rho-1)/(2-\rho)}$ and $b_\rho = [\rho(1-\rho)/2]^{1/(2-\rho)}$.*

Proof of Lemma 7.1: Since $w(-z,\alpha) = -w(z,\alpha)$, it is enough to consider $z > 0$. The derivative of the objective function is of the form $h(w,z) = 2w - 2z + \alpha\rho|w|^{\rho-1}\mathrm{sgn}(w)$. Note that the function $\phi(w) = 2w + \alpha\rho|w|^{\rho-1}\mathrm{sgn}(w)$ is odd and for $w > 0$ has a minimum at $w_0 = b_\rho\alpha^{1/(2-\rho)}$ equal to $a_\rho\alpha^{1/(2-\rho)}$. Hence, whenever $0 < z < a_\rho\alpha^{1/(2-\rho)}$ equation $h(w,z) = 0$ has no solutions and $h(w,z)$ is negative for any $w < 0$ and positive for any $w > 0$. Thus, in this case $w(z,\alpha) = 0$. If $z > a_\rho\alpha^{1/(2-\rho)}$, equation $h(w,z) = 0$ has two solutions $0 < w_1 < w_2$ where solution $w_1 < w_0$ corresponds to the local maximum of the objective function while $w_2 > w_0$ corresponds to its absolute minimum. □

We now complete the proof of Proposition 4.2. Let β_j be the j-th component of $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^{(-j)}$ be the vector $\boldsymbol{\beta}$ without component β_j. Similarly, let $\mathbf{X}_j$ be the j-th column of matrix $\mathbf{X}$ and $\mathbf{X}^{(-j)}$ be the matrix $\mathbf{X}$ without column j. Note that $n^{-1}Q(\boldsymbol{\beta},\mathbf{Y})$ in (6) can be rewritten as $n^{-1}Q(\beta_j,\boldsymbol{\beta}^{(-j)},\mathbf{Y}) = n^{-1}\|\mathbf{Y} - \mathbf{X}_j\beta_j - \mathbf{X}^{(-j)}\boldsymbol{\beta}^{(-j)}\|^2 + \lambda_n\sum_{k=m}^{m-1+q_n}|\beta_k|^\rho$ and $\tilde{\beta}_j = \arg\min_{\beta_j} Q(\beta_j,\boldsymbol{\beta}^{(-j)},\mathbf{Y})$. If $0 \leq j \leq m-1$, equating derivative of $n^{-1}Q(\boldsymbol{\beta},\mathbf{Y})$ over $\tilde{\beta}_j$ to zero we obtain

$$n^{-1}\mathbf{X}_j^T\mathbf{X}\tilde{\boldsymbol{\beta}} - n^{-1}\mathbf{X}_j^T\mathbf{Y} = 0. \tag{18}$$

If $m+1 \leq j \leq m+q_n$, then application of Lemma 7.1 with $\alpha = \lambda/(n^{-1}\mathbf{X}_j^T\mathbf{X}_j)$ and $z = n^{-1}\mathbf{X}_j^T(\mathbf{Y} - \mathbf{X}^{(-j)}\tilde{\boldsymbol{\beta}}^{(-j)})/(n^{-1}\mathbf{X}_j^T\mathbf{X}_j)$ yields that $\tilde{\beta}_j = 0$ whenever

$$|n^{-1}\mathbf{X}_j^T(\mathbf{Y} - \mathbf{X}^{(-j)}\tilde{\boldsymbol{\beta}}^{(-j)})| < a_\rho \lambda^{1/(2-\rho)}(n^{-1}\mathbf{X}_j^T\mathbf{X}_j)^{(1-\rho)/(2-\rho)}. \tag{19}$$

In this case, taking into account condition **M5**, $\tilde{\beta}_j = 0$ and $\mathbf{X}\tilde{\boldsymbol{\beta}} = \mathbf{X}^{(-j)}\tilde{\boldsymbol{\beta}}^{(-j)} + \mathbf{X}_j\tilde{\beta}_j$, we obtain

$$|n^{-1}\mathbf{X}_j^T(\mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\beta}})| = O\left(\lambda^{1/(2-\rho)}\right). \tag{20}$$

If inequality (19) does not hold, then $\hat{\beta}_j$ is the solution of the equation $n^{-1}\mathbf{X}_j^T(\mathbf{X}\boldsymbol{\beta} - \mathbf{Y}) + \lambda\rho|\hat{\beta}_j|^{\rho-1}\mathrm{sgn}(\tilde{\beta}_j) = 0$. For $0 < \rho < 1$, by Lemma 7.1 we obtain

$$|n^{-1}\mathbf{X}_j^T(\mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\beta}})| = \lambda\rho|\tilde{\beta}_j|^{\rho-1} = O\left(\lambda^{1+(\rho-1)/(2-\rho)}\right) = O\left(\lambda^{1/(2-\rho)}\right). \tag{21}$$

Observe that for $\rho = 1$, (21) continues to hold. Combining (18), (20) and (21), we derive that for any j

$$n^{-1}|\mathbf{X}_j^T\mathbf{X}\tilde{\boldsymbol{\beta}} - n^{-1}\mathbf{X}_j^T\mathbf{Y}| = O\left(\lambda^{1/(2-\rho)}\right). \tag{22}$$

Now, denote by $\tilde{\boldsymbol{\beta}}_*$ the global minimizer of $\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2$. It is easy to see that

$$n^{-1}\mathbf{X}_j^T\mathbf{X}\tilde{\boldsymbol{\beta}}_* - n^{-1}\mathbf{X}_j^T\mathbf{Y} = 0. \tag{23}$$

Moreover, Agarwal and Studden (1980) showed that for $\tilde{f}_* = \mathbf{X}\tilde{\boldsymbol{\beta}}_*$ under condition **M7**, one has

$$\sup_{f \in \mathcal{H}(m,R,\kappa,S)} R(\tilde{f}_*, f) = O\left(n^{-2m/(2m+1)}\right). \tag{24}$$

Subtracting equation (23) from equation (22) one derives $n^{-1}|(\mathbf{X}^T\mathbf{X}(\tilde{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_*))_j| = O\left(\lambda^{1/(2-\rho)}\right)$ for all $j = 0, \cdots, m-1+q_n$, so that (22) together with condition **M7** imply that $\|\mathbf{X}\tilde{\boldsymbol{\beta}} - \mathbf{X}\tilde{\boldsymbol{\beta}}_*\|^2 = O(q_n\lambda^{2/(2-\rho)})$. To complete the proof combine the last equality with (24) and conditions **M6** and **M7**. □

7.4 Proof of Proposition 4.3

Let $\hat{T} = \{\hat{\theta}_j, j = 1, \ldots, \hat{D}\}$ be the set of estimated jump points of f and define the event

$$F = \{\hat{D} = D\} \cap \{|\hat{\theta}_j - \theta_j| < n/(\log n)^{1+\delta}, j = 1, \ldots, \hat{D}\}.$$

We have

$$R(\hat{f}, f) = \mathbb{E}\{\|\hat{f} - f\|_2^2\} = \mathbb{E}\{\|\hat{f} - f\|_2^2\mathbf{I}_F\} + \mathbb{E}\{\|\hat{f} - f\|_2^2\mathbf{I}_{F^c}\}. \tag{25}$$

Note that

$$\mathbb{E}\{\|\hat{f}-f\|_2^2\mathbf{I}_{F^c}\} = \mathbb{E}\left(\mathbb{E}\{\|\hat{f}-f\|_2^2\mathbf{I}_{F^c}|\hat{T}\}\right) = \mathbb{E}\left(\mathbf{I}_{F^c}\mathbb{E}\{\|\hat{f}-f\|_2^2|\hat{T}\}\right).$$

Exploiting the results of Section 4 and the fact that both $\hat{f}$ and f belong to amalgam Sobolev ball $\mathcal{H}(m,R,\kappa,S)$, one easily gets $\mathbb{E}\{\|\hat{f}-f\|_2^2|\hat{T}\} = O_P(1)$. Furthermore, by Proposition 3.1, $\mathbb{P}\{F^c\} = o(n^{-\gamma})$ for an arbitrarily large $\gamma > 0$. Hence the second term in the right-hand side of (25) is $o(n^{-\gamma})$ and is negligible.

Consider now the first term in the right-hand side of (25). For simplicity of exposition, consider first the case where there is a single jump point θ and $D = 1$. The unknown f can be decomposed as

$$f(x) = \mathbf{I}_{[0,\theta]}(x)f_0(x) + \mathbf{I}_{(\theta,1]}(x)f_1(x),$$

where f_0 and f_1 both belong to the Sobolev balls $\mathcal{H}(m,R)$ of radius R. From the known properties of spline approximation, we can approximate f by an an amalgamated m-order polynomial spline s as

$$s(x) = \mathbf{I}_{[0,\theta]}(x)s_0(x) + \mathbf{I}_{(\theta,1]}(x)s_1(x),$$

where s_0 and s_1 are spline approximations of f_0 and f_1 respectively, with the approximation error

$$\|f-s\|_2^2 \le O(n^{-2m/(2m+1)})$$

Let $\hat{\theta}$ be an estimator of θ. Given the data, define amalgamated spline estimators $\tilde{f}$ and $\hat{f}$ with a zero-knot at true θ and estimated $\hat{\theta}$ respectively. Then,

$$\tilde{f}(x) = \mathbf{I}_{[0,\theta]}(x)\tilde{s}_0(x) + \mathbf{I}_{(\theta,1]}(x)\tilde{s}_1(x)$$

and

$$\hat{f}(x) = \mathbf{I}_{[0,\hat{\theta}]}(x)\hat{s}_0(x) + \mathbf{I}_{(\hat{\theta},1]}(x)\hat{s}_1(x),$$

where $\tilde{s}_j$, $j = 0,1$ and $\hat{s}_j$, $j = 0,1$ are the corresponding spline estimates.

To simplify notation, denote hereafter $\mathbb{E}\{\cdot\mathbf{I}_F\}$ by $\mathbb{E}_F\{\cdot\}$. We have $\mathbb{E}_F\{\|\hat{f}-f\|_2^2\} = \mathbb{E}_F\{\|\hat{f}-s\|_2^2\} + O(n^{-2m/(2m+1)})$.

Consider only the case $\hat{\theta} \le \theta$ since the opposite case can be treated in a similar way. Then,

$$\begin{aligned}\mathbb{E}_F\{\|\hat{f}-s\|_2^2\} &= \mathbb{E}_F\{\int_0^{\hat{\theta}}(s(x)-\hat{f}(x))^2dx\} + \mathbb{E}_F\{\int_{\hat{\theta}}^{\theta}(s(x)-\hat{f}(x))^2dx\} \\ &\quad + \mathbb{E}_F\{\int_{\theta}^{1}(s(x)-\hat{f}(x))^2dx\} = (A)+(B)+(C). \qquad (26)\end{aligned}$$

For the first term (A) in (26) we have

$$\begin{aligned}\mathbb{E}_F\{\int_0^{\hat{\theta}}(s(x)-\hat{f}(x))^2dx\} &= \mathbb{E}_F\{\int_0^{\hat{\theta}}(\hat{s}_0(x)-s_0(x))^2dx\} \\ &\leq \mathbb{E}_F\{\int_0^{\hat{\theta}}(\hat{s}_0(x)-\tilde{s}_0(x))^2dx\}+\mathbb{E}_F\{\int_0^{\hat{\theta}}(\tilde{s}_0(x)-s_0(x))^2dx\} \\ &= (A_1)+(A_2).\end{aligned}$$

By Proposition 4.2 , the term (A_2) is $O(n^{-2m/(2m+1)})$. For (A_1), note that by construction and assumption M6, the sup-norm distance between the knots of $\hat{s}$ and those of $\tilde{s}$ is bounded above by $O\left(q_n^{-1}(\theta-\hat{\theta})\right) \simeq o_P(n^{-2m/(2m+1)})$. The knots of $\hat{s}$ may be viewed as the set of knots of $\tilde{s}$ perturbed by an amount $o(n^{-2m/(2m+1)})$. Using Theorem 6.2 of Lyche and Mørken (1999) it follows that $\|\hat{s}_0(x)-\tilde{s}_0(x)\|_\infty^2 = O(n^{-2m/(2m+1)})$ and therefore (A_1) is also $O(n^{-2m/(2m+1)})$.

The third term (C) in the right-hand side of (26) can be handled exactly in the same way to verify that it is $O(n^{-2m/(2m+1)})$. Finally, it is easy to see that the remaining term (B) is $O\left(\mathbb{E}_F(|\theta-\hat{\theta}|)\right)$ which is $o(n^{-2m/(2m+1)})$ by Corollary 3.1.

So far we have proved the proposition for $D=1$. For an arbitrary (but still finite!) D, one can partition the unit interval by θ_l and $\hat{\theta}_l$, $l=1,...,D$ and then use the result established above for a single jump point similarly to the proof of Proposition 2 in Antoniadis and Gijbels (2002) in order to obtain the rates stated in Proposition 4.3. □

Acknowledgment

This research was supported by the "Projet d'Actions de Recherches Concertées", No. 93/98-164 of the Belgian Government and by NSF grant DMS 0505133. Felix Abramovich and Marianna Pensky would like to thank Anestis Antoniadis for warm hospitality while visiting Grenoble to carry out part of this work.

REFERENCES

Abry, P. (1994). *Transformées en ondelettes – Analyses multirésolutions et signaux de pression en turbulance*. Doctoral dissertation. Université Claude Bernard, Lyon, France.

Agarwal, G. and Studden, W. (1980). Asymptotic integrated mean square error using least squares and bias minimizing splines. *Ann. Statist.*, **8**, 1307–1325.

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

Antoniadis, A. and Gijbels, I. (2002). Detecting abrupt changes by wavelet methods. *Nonparametric Statistics*, **14**, 7–29.

Amato, U. , Antoniadis, A. and Pensky, M. (2006). Wavelet kernel penalized estimation for non-equispaced design regression. *Statistics and Computing*, **16**, 1, 37–56.

Battaglia, M (2000). The Influence of Overstory Structure on Understory Light Availability in a Longleaf Pine (Pinus palustris Mill) Forest. *Master's Thesis*, Department of Forestry, Virginia Tech, USA.

Berkner, K. and Wells, R.O. (1997). A fast approximation to the continuous wavelet transform with applications, *IEEE Proceedings of Asilomar' 97.*

Coifman, R.R. and Donoho, D.L. (1995). Translation invariant denoising, in *Wavelets in Statistics*, Antoniadis, A. and Oppenheim, G. (eds), *Lecture Notes in Statistics*, Springer, **103** , 125–150.

Daubechies, I. (1992). *Ten Lectures on Wavelets*, CBMS-NSF Series in Applied Mathematics, SIAM, Philadelphia.

Daubechies, I., Defrise, M. and De Mol C. (2004). An iterative thresholding algorithm for linear inverse problems with a sparsity constraint, Technical Report, Department of Mathematics, Princeton University.

De Boor, C. (1978). *A Practical Guide to Splines.* Springer-Verlag, New York.

Dierckx, P. (1993). *Curve and Surface Fitting with Splines.* Oxford University Press.

Donoho, D.L. and Johnstone, I.M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425–455.

Dutilleux, P. (1989). An implementation of the "algorithme à trou" to compute the wavelet transform. In *Wavelets Time-Frequency Methods and Phase Space*, 1–20, Combes, J.M., Grossman A. and Tchamitchian, Ph., editors, Springer-Verlag, Berlin, Heidelberg.

Eilers, P.H. and Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, **11**, 89–121.

Fan, J. and Li, R.Z. (2001). Variable selection via penalized likelihood. *J. Amer. Statist. Assoc.*, **96**, 1348–1360.

Feller, W. (1950). *An Introduction to Probabilty Theory and Its Applications*. Volume 1, Wiley.

Fink, D., and Wells, M. (2004). Adaptive multiorder penalized regression splines. Technical report, Department of Statistical Science, Cornell University.

Frank, I.E. and Friedman, J.H. (1993). A statistical view on some chemometrics regression tools (with discussion). *Technometrics*, **35**, 109–148.

Jaffard, S. (1989). Exposants de Hölder en des points donnés et coefficients d'ondelettes. *C.R.Acad.Sci. Paris*, **308**, Série I, 79–81.

Koo, J.-Y (1997). Spline estimation of discontinuous regression functions. *J. Comp. Graph. Statist.*, **6**, 3, 266-284.

Lyche, T. and Mørken, K. (1999). The sensitivity of a spline function to perturbations of the knots. *BIT*, **39**, 2, 305–322.

Mallat, S. (1989). Theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. on PAMI*, **11**, 674 – 692.

Mallat, S. and Hwang, W.L. (1992). Singularity detection and processing with wavelets. *IEEE Trans. Inform. Theory*, **2**, 617 – 643.

Nason, G.P. and Silverman, B.W. (1995). The stationary wavelet transform and some statistical applications. in *Wavelets in Statistics*, Antoniadis, A. and Oppenheim, G. (eds), *Lecture Notes in Statistics*, Springer, **103**, 281–300.

Osborne, M.R., Presnell, B. and Turlach, B.A. (2000). On the LASSO and its dual, *Journal of Computational and Graphical Statistics*, **9** , 2, 319–337.

Oudshoorn, C.G.M. (1998). Asymptotically minimax estimation of a function with jumps. *Bernoulli*, **4**, 15–33.

Ruppert, D. , and Carroll, R. J. (2000), "Spatially-adaptive penalties for spline fitting", *The Australian and New Zealand Journal of Statistics*, **42**, 205–223.

Shensa, M.J. (1992). The discrete wavelet transform: Wedding the à trous and Mallat algorithms. *IEEE Trans. Inform. Theory*, **40**, 2464–2482.

Tibshirani, R. (1996). Regression shrinkage and selection via lasso. *J. R. Statist. Soc. B*, **58**, 267–288.

Wang, Y. (1995). Jump and sharp cusp detection by wavelets. *Biometrika*, **82**, 385 – 397.

Zhou, S. and Shen, X. (2001) . Spatially adaptive regression splines and accurate knot selection schemes. *J. Am. Statist. Assoc.* **96**, 247–259.