



ELSEVIER

Available at

www.ElsevierMathematics.com

POWERED BY SCIENCE @ DIRECT®

Computational Statistics & Data Analysis III (IIII) III-III

**COMPUTATIONAL
STATISTICS
& DATA ANALYSIS**

www.elsevier.com/locate/csda

Wavelet estimation of partially linear models

Xiao-Wen Chang^{a,*}, Leming Qu^b

^a*School of Computer Science, McGill University, 3480 University Street, Montreal, Quebec, Canada H3A 2A7*

^b*Department of Mathematics, Boise State University, Boise, ID 83725, USA*

Received 1 August 2003; accepted 23 October 2003

Abstract

A wavelet approach is presented for estimating a partially linear model (PLM). We find an estimator of the PLM by minimizing the square of the l_2 norm of the residual vector while penalizing the l_1 norm of the wavelet coefficients of the nonparametric component. This approach, an extension of the wavelet approach for nonparametric regression problems, avoids the restrictive smoothness requirements for the nonparametric function of the traditional smoothing approaches for PLM, such as smoothing spline, kernel and piecewise polynomial methods. To solve the optimization problem, an efficient descent algorithm with an exact line search is presented. Simulation results are given to demonstrate effectiveness of our method.

© 2003 Elsevier B.V. All rights reserved.

Keywords: Partially linear models; Wavelet estimation; Discrete wavelet transform (DWT); Penalized least squares; Descent algorithms

1. Introduction

Considering the following regression model:

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + f(t_i) + \varepsilon_i, \quad i = 1, 2, \dots, n, \quad (1)$$

where $\{y_i\}$ are observations, $\{\mathbf{x}_i\}$ are known design points and each is a fixed p -dimensional vector with $p \leq n$, $\{t_i\}$ are values of an extra univariate variable such as the time at which the observation is made, $\boldsymbol{\beta}$ is an unknown p -dimensional parameter vector, f is an unknown function, and $\{\varepsilon_i\}$ are random errors assumed to be iid $N(0, \sigma^2)$ distributed. Without loss of generality, we assume $t \in [0, 1]$. The goal is to

* Corresponding author. Tel.: +1-514-398-8259; fax: +1-514-398-3883.

E-mail addresses: chang@cs.mcgill.ca (X.-W. Chang), qu@math.boisestate.edu (L. Qu).

estimate the unknown parameter vector $\boldsymbol{\beta}$ and nonparametric function $f(t)$ from the data $\{(y_i, x_i, t_i)\}$.

In matrix–vector notation, the model is written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{f} + \boldsymbol{\varepsilon}, \quad (2)$$

where

$$\begin{aligned} \mathbf{y} &= (y_1, \dots, y_n)^\top, & \mathbf{X}^\top &= [\mathbf{x}_1, \dots, \mathbf{x}_n], \\ \mathbf{f} &= (f(t_1), \dots, f(t_n))^\top, & \boldsymbol{\varepsilon} &= (\varepsilon_1, \dots, \varepsilon_n)^\top. \end{aligned} \quad (3)$$

As in most literature, we assume that \mathbf{X} has full column rank.

This model has received a considerable amount of research in the past two decades. One reason is that it is much more flexible than the standard linear model since it combines both parametric and nonparametric components. Another reason is that it allows easier interpretation of the effect of each variable compared to a completely nonparametric regression. Because of its relation to the classical linear regression model, this model is called “partially linear model” or “partly linear model” (PLM) in the literature. Engle et al. (1986) were among the first to apply this model in analyzing the relationship between weather and electricity sales. Recently, the monograph by Hardle et al. (2000) studies this model exclusively.

All the existing approaches for the PLM are based on different nonparametric regression procedures. For example, the partial spline solution for model (1) is based on the fact that the cubic spline is a linear estimator for the nonparametric regression problem. So, the nonparametric procedure can be naturally extended to handle the PLM. Among the most important approaches are the spline methods by Eubank et al. (1998), Green et al. (1985), Green and Silverman (1994, Chapter 4), Heckman (1986), Schimek (2000) and Wahba (1990, Chapter 6); the kernel methods by Robinson (1988) and Speckman (1988); the piecewise polynomials method by Chen (1988); and the local linear smoothing method by Hamilton and Truong (1997).

In this paper, we extend the wavelet nonparametric regression procedure to the PLM, and develop an iterative algorithm for the l_1 -penalized least-squares criterion for fitting the PLM. The main reason for adopting the wavelet approach for the PLM is that an important assumption by all the existing approaches for $f(t)$ is its high smoothness. But in reality, the assumption may not be satisfied. In some practical areas, such as signal and image processing, objects are frequently inhomogeneous. For the wavelet approach, it is well known that the hypotheses of degrees of smoothness of the underlying function $f(t)$ are less restrictive.

The rest of the paper is organized as follows. In Section 2 we give a brief review about the wavelet nonparametric regression and the discrete wavelet transform (DWT). We then extend this approach to the PLM and discuss the necessary and sufficient conditions for the optimal wavelet estimator in Section 3. Based on these conditions, we propose a descent algorithm in Section 4. In order to demonstrate effectiveness of our method, we give some simulation results in Section 5. Finally, some remarks and suggestions for future research conclude this article in Section 6.

2. The background: wavelet nonparametric regression

The classical nonparametric regression problem is to recover $f(t)$ after observing data $\{(y_i, t_i)\}$ from the standard “signal-plus-noise” model

$$y_i = f(t_i) + \varepsilon_i, \quad i = 1, \dots, n,$$

where $f(t)$ is an unknown function, and $\{\varepsilon_i\}$ are noises and are usually assumed to be iid $N(0, \sigma^2)$ distributed. Note that this model is a special case of model (1) where $\beta = 0$. In order to apply DWT, we assume that $\{t_i\}$ are equally spaced and n is power of 2.

In vector notation (see (3)), the above model is written

$$\mathbf{y} = \mathbf{f} + \boldsymbol{\varepsilon}. \quad (4)$$

The DWT can be represented by an orthogonal matrix \mathbf{W} . Applying \mathbf{W} to the noisy observation \mathbf{y} , we obtain the wavelet transform of \mathbf{y} : $\mathbf{w} = \mathbf{W}\mathbf{y}$. Let $\boldsymbol{\theta} = \mathbf{W}\mathbf{f}$ be the wavelet transform of \mathbf{f} , then $\mathbf{f} = \mathbf{W}^T\boldsymbol{\theta}$ with \mathbf{W}^T the inverse discrete wavelet transform (IDWT). Thus from (4) the observed data can be expressed as a linear model on the wavelet domain

$$\mathbf{y} = \mathbf{W}^T\boldsymbol{\theta} + \boldsymbol{\varepsilon}. \quad (5)$$

The ordinary least-squares estimator of $\boldsymbol{\theta}$ is simply $\hat{\boldsymbol{\theta}}_{\text{LS}} = \mathbf{W}\mathbf{y}$, i.e., the empirical wavelet coefficients. The $\hat{\boldsymbol{\theta}}_{\text{LS}}$ is an unbiased estimator of $\boldsymbol{\theta}$ and its covariance matrix is $\sigma^2\mathbf{I}$, where \mathbf{I} is the $n \times n$ identity matrix. The estimator $\hat{\mathbf{f}} = \mathbf{W}^T\hat{\boldsymbol{\theta}}_{\text{LS}}$ simply interpolates the observed data \mathbf{y} . So it does not denoise at all.

To denoise the data, one usually takes the penalized least-squares approach. By penalizing some measure of $\boldsymbol{\theta}$, such as its norm, one loses unbiasedness of $\hat{\boldsymbol{\theta}}$, but may get smaller variance and covariance matrix $\text{Var}(\hat{\boldsymbol{\theta}})$, thus obtain a smaller mean squared error (MSE) overall. Since wavelet coefficients of $f(t)$ in a wide range of function space are usually sparse, one usually penalizes the l_1 norm of $\boldsymbol{\theta}$ to get a sparse solution. For a given $\lambda > 0$, the solution to the penalized least-squares problem

$$\min_{\boldsymbol{\theta}} \left\{ \frac{1}{2} \|\mathbf{y} - \mathbf{W}^T\boldsymbol{\theta}\|_2^2 + \lambda \|\boldsymbol{\theta}\|_1 = \frac{1}{2} \|\mathbf{w} - \boldsymbol{\theta}\|_2^2 + \lambda \|\boldsymbol{\theta}\|_1 \right\}.$$

is the soft thresholding of \mathbf{w} :

$$\hat{\boldsymbol{\theta}}_S = \text{sgn}(\mathbf{w}) \circ (|\mathbf{w}| - \lambda \mathbf{e})_+,$$

where $\mathbf{e} = (1, \dots, 1)^T$, $\mathbf{x}_+ = (\max(x_1, 0), \dots, \max(x_n, 0))^T$ for any vector $\mathbf{x} = (x_1, \dots, x_n)^T$, and \circ denotes the componentwise product of two vectors. Here we have used the fact that

$$\text{sgn}(w_i)(|w_i| - \lambda)_+ = \arg \min_{\theta_i} \frac{1}{2}(w_i - \theta_i)^2 + \lambda|\theta_i|, \quad i = 1, \dots, n.$$

The choice of the smoothing parameter or the threshold λ is crucial. There are several approaches in the literature. It can be chosen to be the universal threshold $\lambda_{\text{UV}} = \sigma\sqrt{2 \log(n)}$ (Donoho and Johnstone, 1994), or determined by minimizing Stein

unbiased risk estimate (SURE) (Donoho and Johnstone, 1995), or by the method of cross validation (Nason, 1996).

For DWT and IDWT, a fast algorithm developed by Mallat (1989) can be used to perform the transform $\mathbf{w} = \mathbf{W}\mathbf{y}$ in $O(n)$ operations and a matrix–vector multiplication is avoided. However, the use of the fast DWT and IDWT requires equally spaced t_i 's and n to be power of 2. This requirement is not a real restriction. Methods exist to overcome these limitations, allowing the DWT to be applied on unequally spaced data with any length (Kovac and Silverman, 2000).

3. Wavelet estimation for partially linear model

We now extend the idea of wavelet nonparametric regression in the previous section to the estimation of the partial linear model (1). We assume equally spaced time points $t_i = i/n$ and n is power of 2 in model (1). In the wavelet domain, the observed data can be expressed as a linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}^T\boldsymbol{\theta} + \boldsymbol{\varepsilon}.$$

If $\boldsymbol{\beta}$ is known, the model is the same as (5) in the previous section. So our focus here is to estimate $\boldsymbol{\beta}$. By penalizing the l_1 norm of $\boldsymbol{\theta}$, for a given λ , one finds $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ which minimize the quantity

$$l(\boldsymbol{\beta}, \boldsymbol{\theta}) = \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{W}^T\boldsymbol{\theta}\|_2^2 + \lambda \|\boldsymbol{\theta}\|_1.$$

Then, by the orthogonality of the matrix \mathbf{W} ,

$$\begin{aligned} l(\boldsymbol{\beta}, \boldsymbol{\theta}) &= \frac{1}{2} \|\mathbf{W}\mathbf{y} - \mathbf{W}\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\theta}\|_2^2 + \lambda \|\boldsymbol{\theta}\|_1 \\ &= \frac{1}{2} \|\mathbf{w} - \mathbf{U}\boldsymbol{\beta} - \boldsymbol{\theta}\|_2^2 + \lambda \|\boldsymbol{\theta}\|_1, \end{aligned} \quad (6)$$

where $\mathbf{U} = [\mathbf{u}_1, \dots, \mathbf{u}_n]^T = \mathbf{W}\mathbf{X}$ is the DWT of the matrix \mathbf{X} .

Notice that $l(\boldsymbol{\beta}, \boldsymbol{\theta})$ is a convex, continuous function of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ and is bounded below. Also notice that \mathbf{U} is of full column rank. Thus $l(\boldsymbol{\beta}, \boldsymbol{\theta})$ it has finite minimizers. Since $l(\boldsymbol{\beta}, \boldsymbol{\theta})$ is not differentiable with respect to $\boldsymbol{\theta}$ when $\theta_i = 0$ for some i , we need to use directional derivatives to study the characterization of the minimizers.

Definition (Rockafellar, 1970, p. 213). Let L denote a function on \mathcal{R}^m , and let \mathbf{x} be a point where $L(\mathbf{x})$ is finite. The “one-sided directional derivative” L' of L at \mathbf{x} with respect to a direction \mathbf{h} is the limit (if it exists):

$$L'(\mathbf{x}; \mathbf{h}) = \lim_{\alpha \rightarrow 0^+} \frac{L(\mathbf{x} + \alpha\mathbf{h}) - L(\mathbf{x})}{\alpha}.$$

Note that

$$-L'(\mathbf{x}; -\mathbf{h}) = \lim_{\alpha \rightarrow 0^-} \frac{L(\mathbf{x} + \alpha\mathbf{h}) - L(\mathbf{x})}{\alpha},$$

so the one-sided directional derivative $L'(\mathbf{x}; \mathbf{h})$ is two-sided if and only if $L'(\mathbf{x}; -\mathbf{h})$ exists and

$$L'(\mathbf{x}; -\mathbf{h}) = -L'(\mathbf{x}; \mathbf{h}).$$

In our case, $l(\boldsymbol{\beta}, \boldsymbol{\theta})$ may not have two-sided directional derivative at some points, for example, at $\boldsymbol{\theta} = \mathbf{0}$. The minimizers of $L(\mathbf{x})$ can be characterized according to the following lemma, which will be used later.

Lemma 1. *Let $L(\mathbf{x})$ be a convex function on \mathcal{R}^m , having one-sided directional derivative at any point. Then, $\hat{\mathbf{x}}$ is a minimizer of $L(\mathbf{x})$ if and only if*

$$L'(\hat{\mathbf{x}}; \mathbf{h}) \geq 0 \quad \text{for all } \mathbf{h} \in \mathcal{R}^m. \quad (7)$$

Proof. See p.264 of Rockafellar (1970). \square

With this lemma, we can establish the following theorem to characterize the estimator of our model.

Theorem 1.

$$\{\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}\} = \arg \min_{\{\boldsymbol{\beta}, \boldsymbol{\theta}\}} l(\boldsymbol{\beta}, \boldsymbol{\theta}) = \arg \min_{\{\boldsymbol{\beta}, \boldsymbol{\theta}\}} \frac{1}{2} \|\mathbf{w} - \mathbf{U}\boldsymbol{\beta} - \boldsymbol{\theta}\|_2^2 + \lambda \|\boldsymbol{\theta}\|_1 \quad (8)$$

if and only if the following conditions hold:

$$\mathbf{U}^T(\mathbf{w} - \mathbf{U}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\theta}}) = \mathbf{0}, \quad (9)$$

$$\hat{\boldsymbol{\theta}} = \text{sgn}(\mathbf{w} - \mathbf{U}\hat{\boldsymbol{\beta}}) \circ (|\mathbf{w} - \mathbf{U}\hat{\boldsymbol{\beta}}| - \lambda \mathbf{e})_+. \quad (10)$$

Proof. For convenience, we define the following index sets for a given $\boldsymbol{\theta} \in \mathcal{R}^n$

$$\mathcal{I}(\boldsymbol{\theta}) = \{i : \theta_i = 0\}, \quad \bar{\mathcal{I}}(\boldsymbol{\theta}) = \{i : \theta_i \neq 0\}. \quad (11)$$

By Lemma 1, it is sufficient to show that conditions (9) and (10) are equivalent to

$$l'(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}; \mathbf{q}, \mathbf{t}) \geq 0, \quad \text{for all } \mathbf{q} \in \mathcal{R}^p \quad \text{and} \quad \mathbf{t} \in \mathcal{R}^n,$$

where

$$l'(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}; \mathbf{q}, \mathbf{t}) = \lim_{\alpha \rightarrow 0^+} \frac{l(\hat{\boldsymbol{\beta}} + \alpha \mathbf{q}, \hat{\boldsymbol{\theta}} + \alpha \mathbf{t}) - l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})}{\alpha}.$$

For $\alpha > 0$, we define the incremental ratio

$$\Delta l(\alpha) = \frac{1}{\alpha} [l(\hat{\boldsymbol{\beta}} + \alpha \mathbf{q}, \hat{\boldsymbol{\theta}} + \alpha \mathbf{t}) - l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})].$$

Simple algebraic operations with (6) and (11) give

$$\begin{aligned} \Delta l(\alpha) &= \frac{1}{2} \alpha (\mathbf{U}\mathbf{q} + \mathbf{t})^T (\mathbf{U}\mathbf{q} + \mathbf{t}) + (\mathbf{U}\mathbf{q} + \mathbf{t})^T (-\mathbf{w} + \mathbf{U}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\theta}}) \\ &\quad + \lambda \sum_{i \in \mathcal{I}(\hat{\boldsymbol{\theta}})} |t_i| + \frac{1}{\alpha} \sum_{i \in \bar{\mathcal{I}}(\hat{\boldsymbol{\theta}})} (|\hat{\theta}_i + \alpha t_i| - |\hat{\theta}_i|). \end{aligned} \quad (12)$$

Notice that for $\alpha > 0$ sufficiently small

$$\text{sgn}(\hat{\theta}_i + \alpha t_i) = \text{sgn}(\hat{\theta}_i), \quad \forall i \in \bar{\mathcal{J}}(\hat{\boldsymbol{\theta}}),$$

then we have from (12) that

$$\begin{aligned} \Delta l(\alpha) &= \frac{1}{2} \alpha (\mathbf{U}\mathbf{q} + \mathbf{t})^\top (\mathbf{U}\mathbf{q} + \mathbf{t}) + (\mathbf{U}\mathbf{q} + \mathbf{t})^\top (-\mathbf{w} + \mathbf{U}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\theta}}) \\ &\quad + \lambda \sum_{i \in \mathcal{J}(\hat{\boldsymbol{\theta}})} |t_i| + \lambda \sum_{i \in \bar{\mathcal{J}}(\hat{\boldsymbol{\theta}})} \text{sgn}(\hat{\theta}_i) t_i. \end{aligned} \tag{13}$$

Taking the limit as $\alpha \rightarrow 0^+$, we obtain

$$\begin{aligned} l'(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}; \mathbf{q}, \mathbf{t}) &= (\mathbf{U}\mathbf{q} + \mathbf{t})^\top (-\mathbf{w} + \mathbf{U}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\theta}}) + \lambda \sum_{i \in \mathcal{J}(\hat{\boldsymbol{\theta}})} |t_i| + \lambda \sum_{i \in \bar{\mathcal{J}}(\hat{\boldsymbol{\theta}})} \text{sgn}(\hat{\theta}_i) t_i \\ &= \mathbf{q}^\top \mathbf{U}^\top (-\mathbf{w} + \mathbf{U}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\theta}}) + \sum_{i \in \mathcal{J}(\hat{\boldsymbol{\theta}})} \left[-w_i + \mathbf{u}_i^\top \hat{\boldsymbol{\beta}} + \lambda \text{sgn}(t_i) \right] t_i \\ &\quad + \sum_{i \in \bar{\mathcal{J}}(\hat{\boldsymbol{\theta}})} \left[-w_i + \mathbf{u}_i^\top \hat{\boldsymbol{\beta}} + \hat{\theta}_i + \lambda \text{sgn}(\hat{\theta}_i) \right] t_i. \end{aligned} \tag{14}$$

Then it is easy to verify that $l'(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}; \mathbf{q}, \mathbf{t}) \geq 0$ for all $\mathbf{q} \in \mathcal{R}^p$ and $\mathbf{t} \in \mathcal{R}^n$ if and only if the equalities (9) and (10) hold. \square

Our proof is similar to that for Theorem 2 in Alliney and Ruzinsky (1994), which gives necessary and sufficient conditions for the solution of $\min_{\boldsymbol{\beta}} \|\mathbf{w} - \mathbf{U}\boldsymbol{\beta}\|_2^2 + \|\boldsymbol{\beta}\|_1$. A slightly different result from Theorem 1 was given by Fu (1998) in the context of penalized regressions comparing bridge versus Lasso, where the result was proved by mathematical induction on dimension p . Another slightly different result with fixed large number of linear predictors in the context of high-dimensional generalized linear models appeared in Klinger (2001), where the detailed proof of the theorem is not given, although it can be proved by formulating the target function as an equivalent constraint maximum likelihood problem and characterization of corresponding conditions. The key difference between our result and the ones in Fu (1998) and Klinger (2001) is that the dimension of parameters in our case (Theorem 1) is $p + n$, which is greater than sample size n by a constant p , whereas the dimension of parameters in Fu (1998) and Klinger (2001) is a fixed constant p .

Now we give another remark. We can easily argue that (9) and (10) are necessary conditions for $\{\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}\}$ to be a minimizer of $l(\boldsymbol{\beta}, \boldsymbol{\theta})$ without using Lemma 1. In fact, when $\boldsymbol{\theta}$ is known, by the normal equations for linear least-squares estimation, $\hat{\boldsymbol{\beta}}$ is a minimizer of $l(\boldsymbol{\beta}, \boldsymbol{\theta})$ if and only if $\mathbf{U}^\top (\mathbf{w} - \mathbf{U}\hat{\boldsymbol{\beta}} - \boldsymbol{\theta}) = \mathbf{0}$. On the other hand, when $\boldsymbol{\beta}$ is known, according to Section 2, $\hat{\boldsymbol{\theta}}$ is a minimizer of $l(\boldsymbol{\beta}, \boldsymbol{\theta})$ if and only if $\hat{\boldsymbol{\theta}} = \text{sgn}(\mathbf{w} - \mathbf{U}\boldsymbol{\beta}) \circ (|\mathbf{w} - \mathbf{U}\boldsymbol{\beta}| - \lambda \mathbf{e})_+$.

4. Algorithms for computing the wavelet estimator

From the structure of Eqs. of (9) and (10), naturally we see that we can use the following naive iterative scheme to find the solution:

Backfitting algorithm:

Start with $\boldsymbol{\theta}^0 = \mathbf{0}$.

For $k = 1, 2, \dots$ until convergence

$$\boldsymbol{\beta}^k = (\mathbf{U}^T \mathbf{U})^{-1} \mathbf{U}^T (\mathbf{w} - \boldsymbol{\theta}^{k-1}) \quad (15)$$

$$\boldsymbol{\theta}^k = \text{sgn}(\mathbf{w} - \mathbf{U} \boldsymbol{\beta}^k) \circ (|\mathbf{w} - \mathbf{U} \boldsymbol{\beta}^k| - \lambda \mathbf{e})_+. \quad (16)$$

The above scheme is in spirit similar to the iterative *backfitting algorithm* (Hastie and Tibshirani, 1990), a general algorithm that enables one to fit an additive model using any regression type fitting mechanisms, hence we call this scheme the backfitting algorithm.

The backfitting algorithm is simple, but not very efficient. Based on it, we will derive a more sophisticated one, a descent iterative algorithm with an exact line search. In order to do this, we will first analyze the backfitting algorithm to find the descent direction implicitly used by the algorithm at each iterative step. Then we will propose to modify the algorithm with an exact line search, resulting in the line search algorithm. We will show how to efficiently compute the steplength used in the line search.

First let us look at the backfitting algorithm more closely. Define

$$\mathbf{r} = \mathbf{w} - \mathbf{U} \boldsymbol{\beta}.$$

This is the (transformed) residual at $\boldsymbol{\beta}$ if we do not have the nonparametric component in our model. Naturally, the residual at $\boldsymbol{\beta}^k$ is defined by

$$\mathbf{r}^k = \mathbf{w} - \mathbf{U} \boldsymbol{\beta}^k. \quad (17)$$

For later use, we define the following two index sets:

$$\mathcal{L}(\mathbf{r}) = \{i : |r_i| \leq \lambda\}, \quad \bar{\mathcal{L}}(\mathbf{r}) = \{i : |r_i| > \lambda\}.$$

From (16) the i th element of $\boldsymbol{\theta}^k$ satisfies

$$\theta_i^k = \text{sgn}(r_i^k) (|r_i^k| - \lambda)_+ = \begin{cases} 0, & i \in \mathcal{L}(\mathbf{r}^k), \\ r_i^k - \lambda \text{sgn}(r_i^k), & i \in \bar{\mathcal{L}}(\mathbf{r}^k), \end{cases} \quad i = 1, \dots, n. \quad (18)$$

Thus

$$\boldsymbol{\theta}^k = \mathbf{r}^k - \mathbf{z}^k, \quad \text{where} \quad \mathbf{z}^k = \begin{cases} r_i^k, & i \in \mathcal{L}(\mathbf{r}^k), \\ \lambda \text{sgn}(r_i^k), & i \in \bar{\mathcal{L}}(\mathbf{r}^k), \end{cases} \quad i = 1, \dots, n. \quad (19)$$

Then from the iteration formula (15) with (19) and (17),

$$\boldsymbol{\beta}^{k+1} = (\mathbf{U}^T \mathbf{U})^{-1} \mathbf{U}^T (\mathbf{w} - \mathbf{r}^k + \mathbf{z}^k) = \boldsymbol{\beta}^k + (\mathbf{U}^T \mathbf{U})^{-1} \mathbf{U}^T \mathbf{z}^k. \quad (20)$$

Thus \mathbf{q}^k defined by

$$\mathbf{q}^k = (\mathbf{U}^T \mathbf{U})^{-1} \mathbf{U}^T \mathbf{z}^k \quad (21)$$

is a search direction with respect to $\boldsymbol{\beta}$ at step k . Later we will show \mathbf{q}^k is a descent direction if $\mathbf{q}^k \neq \mathbf{0}$. But the unit steplength in (20) may not be optimal or even possibly $l(\boldsymbol{\beta}^{k+1}, \boldsymbol{\theta}^{k+1}) > l(\boldsymbol{\beta}^k, \boldsymbol{\theta}^k)$. So we would like to incorporate a line search, taking (cf. (20), (21) and (16))

$$\boldsymbol{\beta}^{k+1} \equiv \boldsymbol{\beta}^{k+1}(\alpha) = \boldsymbol{\beta}^k + \alpha \mathbf{q}^k, \quad (22)$$

$$\boldsymbol{\theta}^{k+1} \equiv \boldsymbol{\theta}^{k+1}(\alpha) = \text{sgn}(\mathbf{w} - \mathbf{U} \boldsymbol{\beta}^{k+1}(\alpha)) \circ (|\mathbf{w} - \mathbf{U} \boldsymbol{\beta}^{k+1}(\alpha)| - \lambda \mathbf{e})_+. \quad (23)$$

The optimal steplength α will be determined later. Note that the new definition of $\boldsymbol{\theta}^{k+1}$ is consistent with (16). Since $\boldsymbol{\theta}^{k+1}$ is determined by $\boldsymbol{\beta}^{k+1}$, we do not need to know what the corresponding search direction with respect to $\boldsymbol{\theta}$ is.

If the search direction $\mathbf{q}^k = \mathbf{0}$, then $\{\boldsymbol{\beta}^k, \boldsymbol{\theta}^k\}$ is a minimizer. In fact, with a line search in each iteration step, we still have (17), (19) and (21), thus

$$\begin{aligned} \mathbf{q}^k &= (\mathbf{U}^T \mathbf{U})^{-1} \mathbf{U}^T \mathbf{z}^k = (\mathbf{U}^T \mathbf{U})^{-1} \mathbf{U}^T (\mathbf{r}^k - \boldsymbol{\theta}^k) \\ &= (\mathbf{U}^T \mathbf{U})^{-1} \mathbf{U}^T (\mathbf{w} - \mathbf{U} \boldsymbol{\beta}^k - \boldsymbol{\theta}^k) = \mathbf{0}. \end{aligned}$$

So $\{\boldsymbol{\beta}^k, \boldsymbol{\theta}^k\}$ satisfies condition (9). Since it also satisfies condition (10) (cf. (23)), it must be a minimizer of $l(\boldsymbol{\beta}, \boldsymbol{\theta})$.

In the following, we will show that \mathbf{q}^k is a descent direction under the assumption that $\mathbf{q}^k \neq \mathbf{0}$. Define

$$\mathbf{h}^k = -\mathbf{U} \mathbf{q}^k. \quad (24)$$

Then with (22),

$$\mathbf{r}^{k+1}(\alpha) = \mathbf{w} - \mathbf{U} \boldsymbol{\beta}^{k+1}(\alpha) = \mathbf{w} - \mathbf{U}(\boldsymbol{\beta}^k + \alpha \mathbf{q}^k) = \mathbf{r}^k + \alpha \mathbf{h}^k,$$

which shows how the residual \mathbf{r} is updated when $\boldsymbol{\beta}$ is updated according to (22), and from (23)

$$\boldsymbol{\theta}^{k+1}(\alpha) = \text{sgn}(\mathbf{r}^k + \alpha \mathbf{h}^k) \circ (|\mathbf{r}^k + \alpha \mathbf{h}^k| - \lambda \mathbf{e})_+.$$

With $\mathbf{z}^{k+1}(\alpha) = \mathbf{r}^{k+1}(\alpha) - \boldsymbol{\theta}^{k+1}(\alpha)$ (cf. (19)), we have

$$\begin{aligned} J(\alpha) &\equiv l(\boldsymbol{\beta}^{k+1}(\alpha), \boldsymbol{\theta}^{k+1}(\alpha)) = \frac{1}{2} \|\mathbf{z}^{k+1}(\alpha)\|_2^2 + \lambda \|\boldsymbol{\theta}^{k+1}(\alpha)\|_1 \\ &= \sum_{i=1}^n \frac{1}{2} (z_i^{k+1}(\alpha))^2 + \lambda |\theta_i^{k+1}(\alpha)|. \end{aligned}$$

Simple calculations give

$$J_i(\alpha) \equiv \frac{1}{2}(\mathbf{z}_i^{k+1})^2(\alpha) + \lambda|\theta_i^{k+1}(\alpha)| = \begin{cases} \frac{1}{2}(r_i^k + \alpha h_i^k)^2, & i \in \mathcal{Z}(\mathbf{r}^k + \alpha \mathbf{h}^k), \\ \frac{1}{2}(2\lambda|r_i^k + \alpha h_i^k| - \lambda^2), & i \in \bar{\mathcal{Z}}(\mathbf{r}^k + \alpha \mathbf{h}^k). \end{cases}$$

We observe that $J_i(\alpha)$ is a piecewise quadratic, convex, continuously differentiable function of α and

$$J'_i(\alpha) = \begin{cases} h_i^k(r_i^k + \alpha h_i^k), & i \in \mathcal{Z}(\mathbf{r}^k + \alpha \mathbf{h}^k), \\ h_i^k \lambda \operatorname{sgn}(r_i^k + \alpha h_i^k), & i \in \bar{\mathcal{Z}}(\mathbf{r}^k + \alpha \mathbf{h}^k). \end{cases} \quad (25)$$

Thus by (19), (24) and (21), we have

$$\begin{aligned} J'(0) &= \sum_{i=1}^n J'_i(0) = \sum_{i \in \mathcal{Z}(\mathbf{r}^k)} h_i^k r_i^k + \sum_{i \in \bar{\mathcal{Z}}(\mathbf{r}^k)} h_i^k \lambda \operatorname{sgn}(r_i^k) \\ &= (\mathbf{h}^k)^\top \mathbf{z}^k = -(\mathbf{q}^k)^\top \mathbf{U}^\top \mathbf{z}^k = -(\mathbf{z}^k)^\top \mathbf{U}(\mathbf{U}^\top \mathbf{U})^{-1} \mathbf{U}^\top \mathbf{z}^k. \end{aligned}$$

Since $\mathbf{q}^k = (\mathbf{U}^\top \mathbf{U})^{-1} \mathbf{U}^\top \mathbf{z}^k \neq \mathbf{0}$, we must have $J'(0) < 0$. Thus indeed \mathbf{q}^k is a descent direction.

Now we would like to determine the optimal steplength, denoted by $\hat{\alpha}$. Since $J(\alpha)$ is a piecewise quadratic, convex, and continuously differentiable function of α , $\hat{\alpha}$ is a minimizer of $J(\alpha)$ if and only if $J'(\hat{\alpha}) = 0$. Since $\mathbf{q}^k \neq \mathbf{0}$ and \mathbf{U} is of full column rank, $\mathbf{h}^k = -\mathbf{U}\mathbf{q}^k \neq \mathbf{0}$. Then from (25) we observe that $J'(\alpha)$ is a nondecreasing function of α and $\lim_{\alpha \rightarrow \infty} J(\alpha) = \infty$. Since $J'(0) < 0$, $\hat{\alpha}$ must be finite and positive. In order to find the optimal $\hat{\alpha}$, we define the set

$$\mathcal{A} = \{\alpha : |r_i^k + \alpha h_i^k| = \lambda, \alpha > 0, h_i^k \neq 0, i = 1, \dots, n\}. \quad (26)$$

Note that the number of elements in \mathcal{A} is at most $2n$. Suppose \mathcal{A} has m distinct elements in the following order

$$0 < \alpha_1 < \alpha_2 < \dots < \alpha_m.$$

For convenience, define $\alpha_0 = 0$. If j is the smallest index such that $J'(\alpha_{j+1}) \geq 0$, then $\hat{\alpha}$ must lie in $(\alpha_j, \alpha_{j+1}]$; or if $J'(\alpha_m) < 0$, then $\hat{\alpha}$ must lie in (α_m, ∞) . We would like to compute $J'(\alpha_j)$ from $j = 1$ until the first j such that $J'(\alpha_j) \geq 0$ in an efficient way.

From (25) we have

$$\begin{aligned} J'(\alpha) &= \left(\sum_{i \in \mathcal{Z}(\mathbf{r}^k + \alpha \mathbf{h}^k)} h_i^k r_i^k + \lambda \sum_{i \in \bar{\mathcal{Z}}(\mathbf{r}^k + \alpha \mathbf{h}^k)} h_i^k \operatorname{sgn}(r_i^k + \alpha h_i^k) \right) \\ &\quad + \alpha \sum_{i \in \mathcal{Z}(\mathbf{r}^k + \alpha \mathbf{h}^k)} (h_i^k)^2. \end{aligned} \quad (27)$$

When $\alpha \in (\alpha_j, \alpha_{j+1})$, the index sets $\mathcal{Z}(r^k + \alpha h^k)$ and $\bar{\mathcal{Z}}(r^k + \alpha h^k)$ remain unchanged, so we use more sensible notation \mathcal{Z}_j and $\bar{\mathcal{Z}}_j$ to replace them, respectively. From (27) we can write

$$J'(\alpha) = c_j + \alpha d_j, \quad \alpha \in (\alpha_j, \alpha_{j+1}),$$

where

$$c_j = \sum_{i \in \mathcal{Z}_j} h_i^k r_i^k + \lambda \sum_{i \in \bar{\mathcal{Z}}_j} h_i^k \operatorname{sgn}(r_i^k + \alpha_j h_i^k), \quad d_j = \sum_{i \in \mathcal{Z}_j} (h_i^k)^2. \quad (28)$$

Here we have used the fact that $\operatorname{sgn}(r_i^k + \alpha h_i^k) = \operatorname{sgn}(r_i^k + \alpha_j h_i^k)$ for $i \in \bar{\mathcal{Z}}_j$. Since c_j and d_j are constants, $J'(\alpha)$ is a linear function in (α_j, α_{j+1}) . But $J'(\alpha)$ is continuous at any α , so we can compute $J'(\alpha_{j+1})$ by $J'(\alpha_{j+1}) = c_j + \alpha_{j+1} d_j$.

In order to compute $J'(\alpha_{j+1})$ quickly, we need an efficient way to compute c_j and d_j , which in fact can be obtained by updating c_{j-1} and d_{j-1} . Notice that when α moves from the interval (α_{j-1}, α_j) to the interval (α_j, α_{j+1}) , each index i_j corresponding to α_j (i.e., $|r_{i_j}^k + \alpha_j h_{i_j}^k| = \lambda$) moves from one type of index sets to the other (i.e., if $i_j \in \mathcal{Z}_{j-1}$, then $i_j \in \bar{\mathcal{Z}}_j$, and if $i_j \in \bar{\mathcal{Z}}_{j-1}$, then $i_j \in \mathcal{Z}_j$), but the other indices do not. Thus from (28) we then have

$$c_j = c_{j-1} + h_{i_j}^k r_{i_j}^k - \lambda h_{i_j}^k \operatorname{sgn}(r_{i_j}^k + \alpha_{j-1} h_{i_j}^k), \quad d_j = d_{j-1} + (h_{i_j}^k)^2, \quad \text{if } i_j \in \bar{\mathcal{Z}}_{j-1};$$

$$c_j = c_{j-1} - h_{i_j}^k r_{i_j}^k + \lambda h_{i_j}^k \operatorname{sgn}(r_{i_j}^k + \alpha_{j-1} h_{i_j}^k), \quad d_j = d_{j-1} - (h_{i_j}^k)^2, \quad \text{if } i_j \in \mathcal{Z}_{j-1}.$$

If there are more than one i_j corresponding to α_j , we simply use a for loop to repeat the above updating process to get the final c_j and d_j . We see that updating c_j and d_j is very simple. But we need the initial c_0 and d_0 (see (28)):

$$c_0 = \sum_{i \in \mathcal{Z}_0} h_i^k r_i^k + \lambda \sum_{i \in \bar{\mathcal{Z}}_0} h_i^k \operatorname{sgn}(r_i^k), \quad d_0 = \sum_{i \in \mathcal{Z}_0} (h_i^k)^2,$$

where it is easy to show that

$$\mathcal{Z}_0 = \{i: |r_i^k| < \lambda\} \cup \{i: |r_i^k| = \lambda, r_i^k h_i^k \leq 0\},$$

$$\bar{\mathcal{Z}}_0 = \{i: |r_i^k| > \lambda\} \cup \{i: |r_i^k| = \lambda, r_i^k h_i^k > 0\}.$$

Since $J'(\alpha)$ is continuous at $\alpha = 0$, we can also compute c_0 using

$$c_0 = J'(0) = \sum_{i \in \mathcal{Z}(r^k)} h_i^k r_i^k + \lambda \sum_{i \in \bar{\mathcal{Z}}(r^k)} h_i^k \operatorname{sgn}(r_i^k).$$

If j is the smallest index such that $J'(\alpha_{j+1}) \geq 0$, then the optimal steplength $\hat{\alpha}$ satisfies

$$\hat{\alpha} = -c_j/d_j.$$

Otherwise $J'(\alpha_m) < 0$ and then

$$\hat{\alpha} = -c_m/d_m.$$

Note that updating c_j and d_j once needs only $O(1)$ operations. If the updating process is repeated l times (note $l \leq m \leq 2n$), then we need $O(l)$ operations. If we

sort the set \mathcal{A} by the binary insertion sorting algorithm at the beginning of the line search process, we need $O(m \log m)$ operations. Thus the whole line search will also need $O(m \log m)$ operations. There is another approach instead of sorting. For each $j \geq 1$, before computing c_j and d_j by the updating process, we find α_j , the smallest element among the remaining elements in \mathcal{A} after $\alpha_1, \dots, \alpha_{j-1}$ have been found in the previous updating processes. This requires $O(m - j)$ operations. Thus the whole line search process requires $O(ml)$ operations. Note that the second approach is more efficient than the first one when $l \leq \log m$.

We are now ready to summarize the whole algorithm.

Line search algorithm:

Given the data \mathbf{y} and \mathbf{X} and a tolerance δ .

Compute the DWT of \mathbf{y} and \mathbf{X} : $\mathbf{w} = \mathbf{W}\mathbf{y}$ and $\mathbf{U} = \mathbf{W}\mathbf{X}$

Compute $\boldsymbol{\beta}^1 = (\mathbf{U}^T \mathbf{U})^{-1} \mathbf{U}^T \mathbf{w}$

For $k = 1, 2, \dots$

 Compute the search direction $\mathbf{q}^k = (\mathbf{U}^T \mathbf{U})^{-1} \mathbf{U}^T \mathbf{z}^k$, where \mathbf{z}^k is computed by (19)

 Compute the optimal steplength $\hat{\alpha} = \arg \min_{\alpha} J(\alpha)$

 Set $\boldsymbol{\beta}^{k+1} = \boldsymbol{\beta}^k + \hat{\alpha} \mathbf{q}^k$

 If $\frac{\|\boldsymbol{\beta}^{k+1} - \boldsymbol{\beta}^k\|_2}{\|\boldsymbol{\beta}^k\|_2} \leq \delta$, set $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{k+1}$

 compute $\hat{\boldsymbol{\theta}} = \text{sgn}(\mathbf{w} - \mathbf{U}\hat{\boldsymbol{\beta}}) \circ (|\mathbf{w} - \mathbf{U}\hat{\boldsymbol{\beta}}| - \lambda \mathbf{e})_+$ and its IDWT $\hat{\mathbf{f}} = \mathbf{W}^T \hat{\boldsymbol{\theta}}$

 stop

Now we give some remarks about this algorithm. In computing $\boldsymbol{\beta}^1$ and \mathbf{q}^k , we can use the QR factorization of \mathbf{U} :

$$\mathbf{U} = \mathbf{Q}_1 \mathbf{R},$$

where $\mathbf{Q}_1 \in \mathcal{R}^{n \times p}$ has orthonormal columns, i.e., $\mathbf{Q}_1^T \mathbf{Q}_1 = \mathbf{I}$, and $\mathbf{R} \in \mathcal{R}^{p \times p}$ is a non-singular upper triangular matrix. The QR factorization of \mathbf{U} can be computed by the Householder transformations in $O(mp^2)$ operations (see for example Golub and van Loan (1996, Chapter 50)). Then we can obtain $\boldsymbol{\beta}^1$ and \mathbf{q}^k by solving the upper triangular systems

$$\mathbf{R}\boldsymbol{\beta}^1 = \mathbf{Q}_1^T \mathbf{w}, \quad \mathbf{R}\mathbf{q}^k = \mathbf{Q}_1^T \mathbf{z}^k.$$

With the QR factorization of \mathbf{U} , computing $\boldsymbol{\beta}^1$ or \mathbf{q}^k needs $O(mp)$ operations. Notice that our algorithm needs to compute the QR factorization only once. This makes the algorithm attractive.

Note that $\boldsymbol{\beta}^1$ is simply the ordinary least-squares estimator. If $|\mathbf{r}^1| = |\mathbf{w} - \mathbf{U}\boldsymbol{\beta}^1| < \lambda \mathbf{e}$, then $\boldsymbol{\theta}^1 = \mathbf{0}$ (see (18)), $\mathbf{z}^1 = \mathbf{r}^1$ (see (19)), and $\mathbf{q}^1 = \mathbf{0}$ (see (21)). Thus $\boldsymbol{\beta}^1$ is the optimal estimator. This means the $\boldsymbol{\theta}$ part should not be included into the model or λ is so big that the $\boldsymbol{\theta}$ is penalized too much.

5. Numerical simulations

In this section, we give some simulation results. All the calculations were carried out in MATLAB 6.5 on a Pentium III running Windows 2000. For the DWT, we used the *WaveLab* developed by the team from the Statistics Department of Stanford University (<http://www-stat.stanford.edu/~wavelab>).

We generated the test problems as follows. For the nonparametric component we select different functions

$$f(t) = cf_i(t), \quad \max_{t \in [0,1]} f_i(t) = 1, \quad i = 1, 2, 3, 4$$

$$f_1(t) = 4.26(\exp(-3.25t) - 4\exp(-6.5t) + 3\exp(-9.75t)),$$

$$f_2(t) = \begin{cases} 4t^2(3 - 4t) & \text{if } 0 \leq t \leq 0.5, \\ \frac{4}{3}t(4t^2 - 10t + 7) - 1.5 & \text{if } 0.5 < t \leq 0.75, \\ \frac{16}{3}t(t - 1)^2 & \text{if } 0.75 < t \leq 1. \end{cases}$$

$$f_3(t) = \text{'Bumps'},$$

$$f_4(t) = \text{'Doppler'}. \quad (29)$$

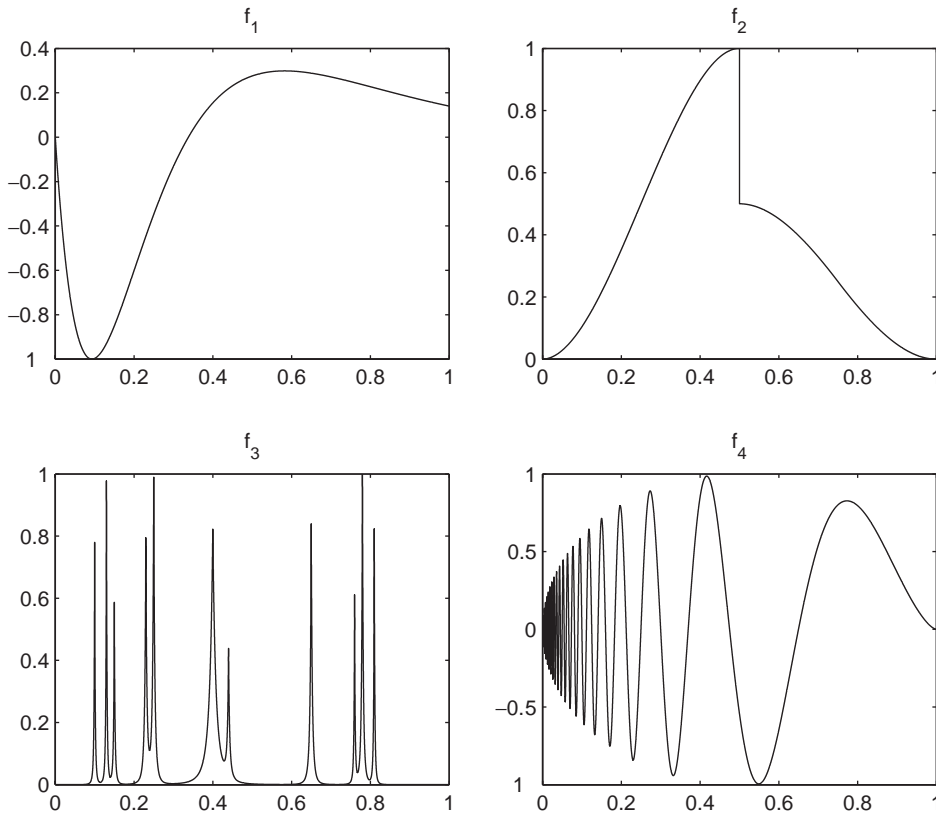
Here f_1 , given in Schimek (2000), is a smoothing function; f_2 , given in Nason (1996), is a piecewise polynomial with discontinuity; f_3 , given in Donoho (1994), is a function with many bumps; and f_4 , given in Donoho (1994), is a function with changing frequencies. Fig. 1 displays the plots of these functions.

With $p = 2$, we took $\beta_1 = 0.5$ and $\beta_2 = 1$ for the regression coefficients and generated x_{i1} and x_{i2} independently from $N(0, 1)$ following Heckman (1986). We simulated ε as a white noise vector following $N(\mathbf{0}, \mathbf{I}_n)$. For DWT, the filter we used is the Daubechies filter with 8 vanishing moments. We chose $c = 9$ in (29) to have a reasonable signal-to-noise ratios of the nonparametric and parametric component. The sample sizes we took were $n = 32, 64, 128$ and 256 . For each sample size, 100 replicates of data with different X and ε were generated. With the simulated data, we then used the proposed algorithm to estimate the true β_1 and β_2 . In our algorithm, the universal threshold $\lambda_{UV} = \sigma\sqrt{2\log(n)}$ was used, and the termination tolerance δ was set to be 10^{-6} .

The estimates of β by the proposed wavelet method were compared with those by partial spline approach. Discussions on the partial spline approach can be found in Wahba (1990, Chapter 6) and Green and Silverman (1994, Chapter 4). Gu (2002, Section 4.1) gave one example for fitting the partial spline models using Gu's *ssanova* facilities in the R package GSS. We used Gu's implementation of partial spline approach in our simulation. We called *ssanova* running in the R for Windows from MATLAB program. The R command for fitting the partial spline model was

```
fit <- ssanova(y~t, partial=cbind(x1,x2,)method="u", varht=1).
```

The smoothing parameter was chosen by the unbiased risk estimator criterion, and the variance estimate was the specified value. Note that the computational complexity

Fig. 1. Plots of the nonparametric function $f(t)$.

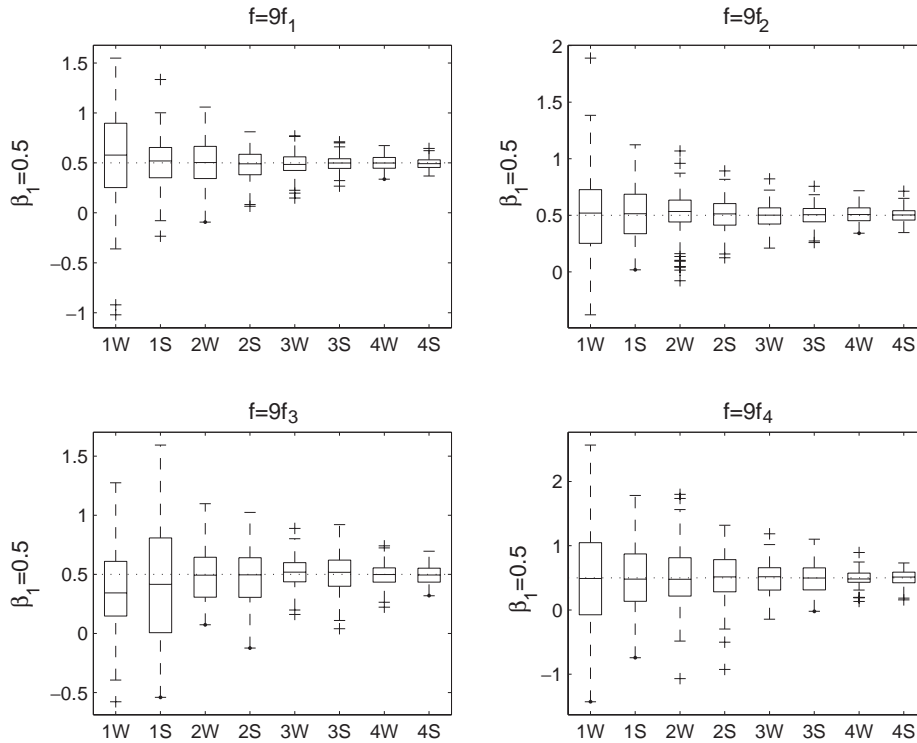
of *ssanova* was $O(n^3)$; this was the main reason that we chose relatively small sample sizes here.

For each setting, we obtained 100 estimates $\hat{\beta}_1$ and $\hat{\beta}_2$. Fig. 2 gives the box plots of the estimated regression coefficients $\hat{\beta}_1$. The x -axis labels in the box plot read as follows: ‘1W’ denotes the case for $n = 32$ using the wavelet approach; ‘1S’ denotes the case for $n = 32$ using the partial spline approach; and so on. Since the results for $\hat{\beta}_2$ are similar to those for $\hat{\beta}_1$, we omit them for brevity.

We also report the observed mean squared errors (MSE) here:

$$\text{MSE}_i = \frac{1}{100} \sum_{j=1}^{100} (\hat{\beta}_{ij} - \beta_i)^2, \quad i = 1, 2.$$

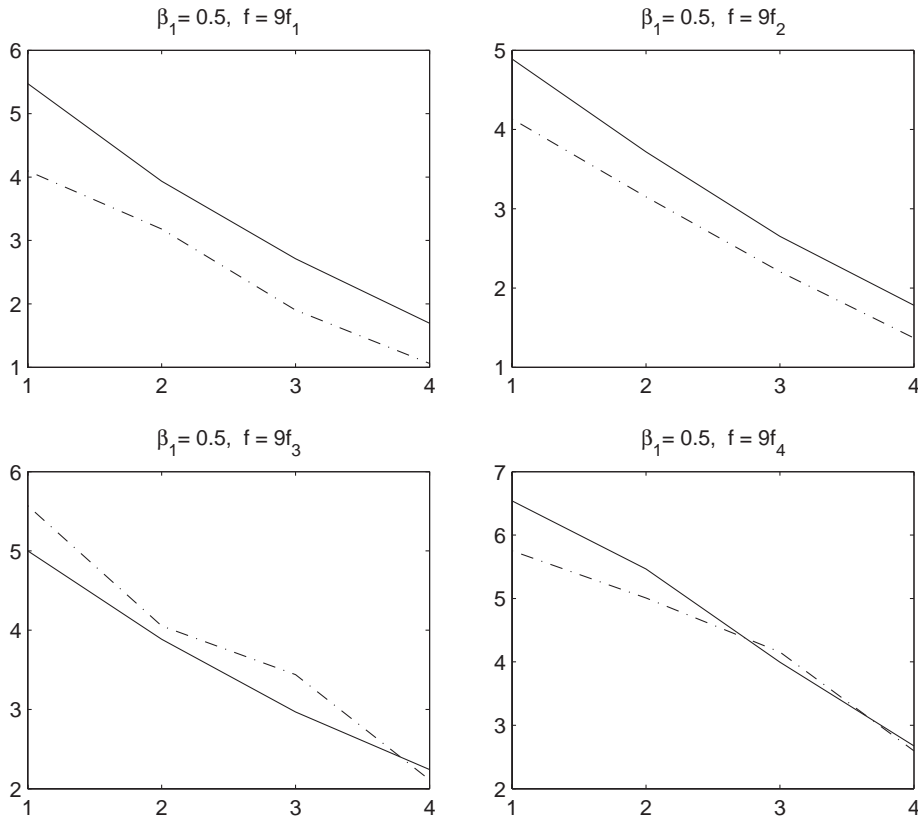
Figs. 3 and 4 display $\log(1000 \times \text{MSE}_1)$ and $\log(1000 \times \text{MSE}_2)$, respectively, where the solid line and the dotted line denote the rescaled MSE by the wavelet method and the partial spline method, respectively.

Fig. 2. Box plots of the estimates $\hat{\beta}_1$.

From these plots, we see that on average, both the wavelet method and the partial spline method gave fairly good estimates of β_1 and β_2 . Reduction of the range of estimates with growing sample size is clearly identified in these figures. For the irregular function like ‘Bumps’, the wavelet method slightly outperforms the partial spline method in terms of the observed MSE. But for the smooth function like f_1 , the partial spline method slightly outperforms the wavelet method. It should be pointed out that the smoothing parameter for the partial spline method was chosen optimally in the expected MSE sense by minimizing the unbiased risk estimator, while the threshold for the wavelet method here was not. A data driven method for choosing a good threshold in the wavelet estimation of the PLM will be a subject of future research.

In order to get some ideas about the computational efficiency of the line search algorithm, we counted the number of iterations of the algorithm for various cases. For brevity, we considered only f_1 and f_2 . The data y and X were constructed the same way as before for $n = 128, 256, 512, 1024$; $p = 2, 20, 40$; $\beta_i = 0.5i$, $i = 1, \dots, p$, and $c = 9$ and 90. For comparison, we also counted the number of iterations for the backfitting algorithm. As before, for each case, we generated 100 replicates of the data. The mean number of iterations for each case was given in Table 1, where ‘LS’ stands for ‘Line Search’ and ‘BF’ stands for ‘BackFitting’.

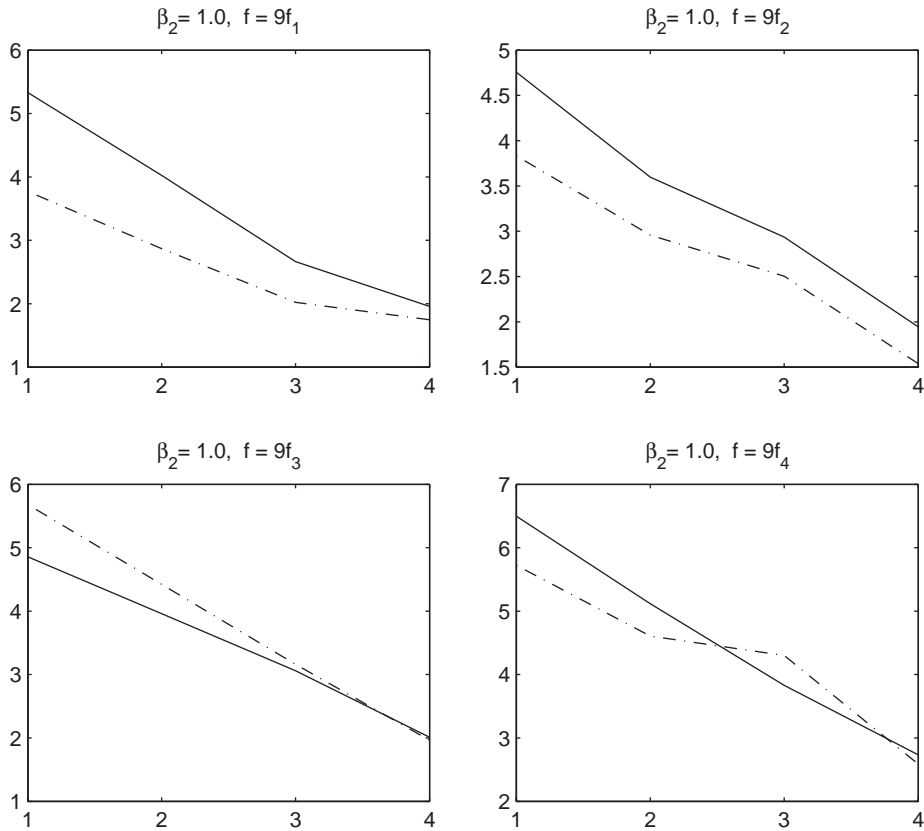
From Table 1, we see that the proposed linear search algorithm converges quickly. When both c and p are small, the backfitting algorithm also converges quickly, costing

Fig. 3. Rescaled MSE of $\hat{\beta}_1$.

only a few more iterations. But when c and p are large, the line search can bring a significant reduction in the number of iterations. For example, with $n = 128$, $p = 40$ and $f = 90f_2$, on average, the backfitting algorithm took 47.86 iterations, while the linear search algorithm took only 23.65 iterations. Note that our line search is quite simple and does not cost much. When n increases, the number of iterations for both algorithms decreases. The reason is that with n increasing the initial estimate of β , which is the ordinary least squares estimate, is closer to the optimal estimate of β . When c increases from 9 to 90, the nonparametric function is more dominant, so that the ordinary least squares estimate is moving further away from the optimal estimate, leading in the increase in the number of iterations.

6. Concluding remarks

To formalize the notion of sparsity of the discrete wavelet transform of functions in a wide range of function spaces (such as Besov spaces), wavelet nonparametric regression usually penalizes the l_1 norm of the population wavelet coefficients. For the

Fig. 4. Rescaled MSE of $\hat{\beta}_2$.

penalized wavelet regression estimates of a partially linear model, we derived the necessary and sufficient conditions for the minimum solution. Based on these conditions, we developed a descent iterative algorithm. The threshold can be taken to be the universal threshold or determined by a data-driven method. The Monte Carlo simulation results confirmed that this wavelet approach (with the universal threshold) has good performance and can give slightly better results than the partial spline approach for some irregular functions. Complete comparisons between this wavelet approach with different thresholds and other approaches need further study.

Future developments include methods for non-equally spaced designs. In the wavelet nonparametric regression context, many approaches have been developed to handle this situation. These methods are mostly based on interpolation or approximation, either in the original function domain or in the wavelet domain. Most recently, [Antoniadis and Fan \(2001\)](#) introduced the nonlinear regularized wavelet estimators by using a large class of penalty functions. It will be of interest to extend this regularized Sobolev interpolator to the partially linear models.

Table 1
The mean number of iterations

n	p	$f = 9f_1$		$f = 90f_1$		$f = 9f_2$		$f = 90f_2$	
		LS	BF	LS	BF	LS	BF	LS	BF
128	2	4.50	7.09	5.94	12.25	4.50	6.59	5.86	13.59
256	2	3.99	5.72	5.06	8.95	3.93	5.51	5.12	9.70
512	2	3.62	4.90	4.49	7.16	3.75	4.98	4.46	7.50
1024	2	3.38	4.21	3.95	5.76	3.27	4.26	4.09	6.34
128	20	7.68	10.45	13.33	25.60	7.29	9.77	13.46	26.76
256	20	5.69	7.17	8.19	13.96	5.57	7.02	8.79	15.74
512	20	4.71	5.57	6.22	9.73	4.68	5.55	6.69	10.53
1024	20	4.00	4.66	5.06	7.25	4.00	4.72	5.32	7.87
128	40	9.91	14.27	23.17	45.85	9.33	13.54	23.65	47.86
256	40	6.43	8.57	10.26	18.85	6.28	8.27	10.92	21.07
512	40	5.01	6.10	7.04	11.59	4.97	6.16	7.44	12.63
1024	40	4.00	4.94	5.69	8.25	4.02	4.97	6.00	9.09

In real world applications, we are most likely to encounter correlated data. [Johnstone and Silverman \(1997\)](#) proposed the level-dependent thresholding approach for data with correlated noise in the wavelet nonparametric regression. This can be naturally extended to the partially linear regression settings. But it will be more computationally intensive.

In order to make inference about the linear coefficients, it is important to derive the asymptotic distribution of the penalized estimators. It is also of interest to study the asymptotic behavior of the estimator for the nonparametric part. The rate of convergence of these estimators is another important problem we would like to investigate in the future.

Acknowledgements

This research was supported by NSERC of Canada Grant RGPIN217191-03 and FCAR of Quebec, Canada Grant 2001-NC-66487 for Xiao-Wen Chang, and by the Faculty Research Associates Program (2003–2004) of the Boise State University for Leming Qu.

We would like to thank Professor Chong Gu for his help on using the *GSS* package, Professor Mary Ellen Bock for her useful suggestions which improved the presentation. The associate editor and the referee are greatly acknowledged for their helpful comments and suggestions.

References

- Alliney, S., Ruzinsky, S.A., 1994. An algorithm for the minimization of mixed l_1 and l_2 norms with application to Bayesian Estimation. *IEEE Trans. Signal Process.* 42 (3), 618–627.

- Antoniadis, A., Fan, J., 2001. Regularization of wavelets approximations (Disc: p956–967). *J. Amer. Statist. Assoc.* 96, 939–956.
- Chen, H., 1988. Convergence rates for parametric components in a partly linear model. *Ann. Statist.* 16, 136–146.
- Donoho, D.L., Johnstone, I.M., 1994. Ideal spatial adaptation by wavelet shrinkage. *Biometrika* 81, 425–455.
- Donoho, D.L., Johnstone, I.M., 1995. Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.* 90, 1200–1224.
- Engle, R.F., Granger, C.W.J., Rice, J., Weiss, A., 1986. Semiparametric estimates of the relation between weather and electricity sales. *J. Amer. Statist. Assoc.* 81, 310–320.
- Eubank, R.L., Kambour, E.L., Kim, J.T., Klipple, K., Reese, C.S., Schimek, M., 1998. Estimation in partially linear models. *Comput. Statist. Data Anal.* 29, 27–34.
- Fu, W., 1998. Penalized regressions: the bridge versus the lasso. *J. Comput. Graph. Statist.* 7 (3), 397–416.
- Golub, G.H., van Loan, C.F., 1996. *Matrix Computations*, 3rd Edition. The Johns Hopkins University Press, Baltimore, MD.
- Green, P., Silverman, B.W., 1994. *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman & Hall, London.
- Green, P., Jennison, C., Scheult, A., 1985. Analysis of field experiments by least squares smoothing. *J. Roy. Statist. Soc. Ser. B* 47, 299–315.
- Gu, C., 2002. *Smoothing Spline ANOVA Models*. Springer, Berlin.
- Hamilton, S.A., Truong, Y.K., 1997. Local linear estimation in partly linear models. *J. Multivariate Anal.* 60 (1), 1–19.
- Hardle, W., Liang, H., Gao, J., 2000. *Partially Linear Models*. Physica-Verlag, Heidelberg.
- Hastie, T.J., Tibshirani, R.J., 1990. *Generalized Additive Models*. Chapman & Hall, London.
- Heckman, N., 1986. Spline smoothing in a partly linear model. *J. Roy. Statist. Soc. Ser. B* 48, 244–248.
- Johnstone, I.M., Silverman, B.W., 1997. Wavelet threshold estimators for data with correlated noise. *J. Roy. Statist. Soc. Ser. B* 59, 319–351.
- Klinger, A., 2001. Inference in high dimensional generalized linear models based on soft thresholding. *J. Roy. Statist. Soc. Ser. B* 63, 377–392.
- Kovac, A., Silverman, B.W., 2000. Extending the scope of wavelet regression methods by coefficient-dependent thresholding. *J. Amer. Statist. Assoc.* 95, 172–183.
- Mallat, S.G., 1989. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pattern Anal. Mach. Intell.* 11, 674–693.
- Nason, G.P., 1996. Wavelet shrinkage using cross-validation. *J. Roy. Statist. Soc. Ser. B* 58, 463–479.
- Robinson, P.M., 1988. Root-n-consistent semiparametric regression. *Econometrica* 56, 931–954.
- Rockafellar, R.T., 1970. *Convex Analysis*. Princeton University Press, Princeton, NJ.
- Schimek, M.G., 2000. Estimation and inference in partially linear models with smoothing splines. *J. Statist. Plann. Inference* 91, 525–540.
- Speckman, P., 1988. Kernel smoothing in partial linear models. *J. Roy. Statist. Soc. Ser. B* 50, 413–436.
- Wahba, G., 1990. *Spline Models for Observational Data*. SIAM, Philadelphia, PA.