ESTIMATION AND GROUP VARIABLE SELECTION FOR ADDITIVE PARTIAL LINEAR MODELS WITH WAVELETS AND SPLINES

Umberto Amato
Istituto per la Microelettronica e Microsistemi, Italian National Research Council, Napoli, Italy

Anestis Antoniadis [1]
University of Cape Town, Department of Statistical Sciences, Cape Town, South Africa &
University Grenoble Alpes, Laboratoire Jean Kuntzmann, Department of Statistics, France
e-mail: *Anestis.Antoniadis@univ-grenoble-alpes.fr*

Italia De Feis
Istituto per le Applicazioni del Calcolo 'M. Picone', Italian National Research Council, Napoli, Italy

Yannig Goude
EDF, OSIRIS, 7 bd Gaspard Monge, 91120 Palaiseau, France

Key words: Additive models, Backfitting, Penalisation, Proximal algorithms, Squared group-LASSO, Splines, Wavelets.

Abstract: In this paper we study sparse high dimensional additive partial linear models with nonparametric additive components of heterogeneous smoothness. We review several existing algorithms that have been developed for this problem in the recent literature, highlighting the connections between them, and present some computationally efficient algorithms for fitting such models. To achieve optimal rates in large sample situations we use hybrid P-splines and block wavelet penalisation techniques combined with adaptive (group) LASSO-like procedures for selecting the additive components in the nonparametric part of the models. Hence, the component selection and estimation in the nonparametric part may be viewed as a functional version of estimation and grouped variable selection. This allows to take advantage of several oracle results which yield asymptotic optimality of estimators in high-dimensional but sparse additive models. Numerical implementations of our procedures for proximal like algorithms are discussed. Large sample properties of the estimates and of the model selection are presented and the results are illustrated with simulated examples and a real data analysis.

1. Introduction

Nonparametric regression methods encompass a large class of flexible models which provide a means of investigating how a response variable Y depends on one or more predictor variables

[1] Corresponding author.

AMS: 62-02, 62G08, 62J07

$X^1, \dots, X^p$, without assuming a specific shape for the relationship. However, as dimension p increases, these techniques suffer from the curse of dimensionality; moreover, the ability to visually inspect estimated relationships is often lost when $p > 2$. An elegant solution to these problems is provided by additive models, an important family of structured nonparametric problems, first suggested in Friedman and Stuelze (1981) and popularised in Hastie and Tibshirani (1986). They model a random sample $\{(Y_i, \mathbf{X}_i)\}_{i=1}^n$ by

$$Y_i = \beta_0 + \sum_{j=1}^{p} f_j(X_i^j) + \varepsilon_i, \quad i = 1, \dots, n,$$

where $\mathbf{X}_i = (X_i^1, \dots, X_i^p)^T$ and the errors ε_i form a sequence of i.i.d. random variables with mean 0 and variance σ^2 independent of the predictor variables X^j. This additive combination of univariate functions—one for each covariate X^j—is less general than joint multivariate nonparametric models, but can be more interpretable and easier to fit. Moreover, since all of the unknown functions are unidimensional the difficulty associated with the curse of dimensionality is substantially reduced. In the past three decades, there has been a considerable amount of research to study additive regression models since they meet three fundamental aspects of statistical models: flexibility, dimensionality and interpretability. However, in such models, the predictors are often assumed to be continuous. Although discrete predictors can be included as indicator variables, their corresponding nonparametric effects are essentially of parametric form. Treating them as nonparametric components increases the computational cost and leads to efficiency loss in theory. This motivates us to look into high dimensional regression problems within the framework of additive partial linear models (PLAM). PLAM extends linear and additive models by modelling the effects of some predictors through a linear function and the effects of the other predictors through additively smooth functions. The additive partially linear model (PLAM) is a realistic, parsimonious candidate when one believes that the relationship between the dependent variable and some of the covariates has a parametric form, while the relationship between the dependent variable and the remaining covariates may not be linear. PLAM models are more flexible than parametric models and more efficient than nonparametric models because they combine both parametric and nonparametric components.

Hereafter we consider a random sample $\{(Y_i, \mathbf{X}_i, \mathbf{T}_i)\}_{i=1}^n$, related through the partially linear additive model (PLAM)

$$Y_i = \mathbf{X}_i^T \beta + \sum_{j=1}^{q} f_j(T_i^j) + \varepsilon_i, \quad i = 1, \dots, n, \tag{1}$$

where $\mathbf{X}_i = (X_i^1, \dots, X_i^p)^T$ is a p-dimensional covariate vector representing the linear regression component, β is the $p \times 1$ vector of corresponding regression coefficients, f_j's are unknown functions of T_i^j where $\mathbf{T}_i = (T_i^1, \dots, T_i^q)^T$ is the q-dimensional nonlinear covariate vector and the errors ε_i form a sequence of i.i.d. random variables with mean 0 and variance σ^2 independent of the predictor variables X^j and T^k. In this model, the response variable Y is linearly related to covariates X, while its relation with covariates T is not specified up to any finite number of parameters.

The additive model is a particular case of the PLAM when only a constant linear covariate is considered. Procedures that achieve simultaneous consistent variable selection and estimation in sparse additive models have been thoroughly discussed in Amato, Antoniadis and De Feis (2016) and will not be reviewed here. The interested reader is referred to the above mentioned paper. The partially linear model (PLM) is a basic (and one of the most studied) semiparametric models. See Hardle, Liang and Gao (2000) for a comprehensive review of partially linear models (PLM), a special case of model (1), which has only individual variables in the linear part and only one nonparametric component f. A lot of efforts have been devoted to estimation in this area and some of them will be reviewed in this paper. Examples include the partial spline estimator (Wahba, 1984; Engle, Granger, Rice and Weiss, 1986; Heckman, 1986) and the partial residual estimator (Robinson, 1988; Speckman, 1988; Chen, 1988). As for the estimation of the nonparametric component in PLM most of existing methods are based on smoothing splines regression techniques and have been employed in particular by Green and Yandell (1985), Engle et al. (1986), Rice (1986), Chen (1988), Chen and Shiau (1991) and Schick (1996) among others. Kernel regression (see, e.g. Speckman, 1988) and local polynomial fitting techniques (see, e.g. Hamilton and Truong, 1997) have also been used to study partially linear models. An important assumption by all these methods for the unknown nonparametric component $f(t)$ is its high smoothness. But in reality, such a strong assumption may not be satisfied. To deal with cases of a less-smooth nonparametric component, some wavelet based estimation procedures have been also proposed in the literature and will be briefly reviewed in this paper since our general methodology is inspired by those developments.

When, in PLM, p is large in the sense that $p \to \infty$ as the sample size $n \to \infty$, but $p < n$, some penalised methods have been proposed to estimate β and f, see, for example, SCAD penalised estimator (Xie and Huang, 2009). One may achieve for many of these methods, under some regularity conditions, consistency in terms of variable selection and estimation simultaneously for the linear and nonparametric component. However, most of these studies do not discuss variable selection and estimation in high-dimensional setting, in the sense that $p \gg n$ and are mainly concerned with only individual variable selection in the linear part. Liu, Wang and Liang (2011) studied variable selection in PLAM when the number of linear covariates is fixed and the nonparametric components are of similar regularity. Compared with the existing semiparametric methods, the method we propose innovates in the following aspects: (i) it can perform estimation and variable selection simultaneously on both the nonparametric and parametric components; (ii) the parametric part can have dimensions diverging with the sample size; and (iii) the nonparametric part can have a large, but fixed number of additive components of heterogeneous smoothness. Moreover, the component selection and estimation in the nonparametric part is viewed as a functional version of estimation and grouped variable selection.

The rest of the article is organised as follows. Section 2 briefly reviews existing results and procedures for fitting PLMs and PLAMs without variable selection with nonparametric components that are supposed to be smooth. Section 3 is concerned with existing procedures in PLMs with much less regular nonparametric components via wavelets, explores their connection with M-estimation and their use in variable selection in the linear part. Section 4 addresses the problem of variable selection in PLAMs with nonparametric components that are smooth. Section 5 establishes the main results addressing a general hybrid wavelet and spline regression methodology for estimating the additive components by linear combinations of basis functions selected from multiple libraries to model various features of the additive components, e.g. spline representers for smooth components and wavelets for less regular components. It also addresses variable selection results for sparse PLAMs by square-root group LASSO and two-step penalisation methods. In each Section and, whenever it is useful, we discuss some computational algorithms that are used for the numerical implementation of the procedures and simulations investigate the finite sample performance of the procedures in terms of prediction, variable selection and estimation accuracy. Some concluding remarks are given in Section 6.

2. Estimation and Inference in Partial Linear Models

Given the observations $\{(Y_i, \mathbf{X}_i, \mathbf{T}_i)\}_{i=1}^n$, where Y_i is the response, $\mathbf{X}_i = (X_i^1, \ldots, X_i^p)^T$ and $\mathbf{T}_i = (T_i^1, \ldots, T_i^q)^T$ are vectors of covariates, the partially linear model assumes that

$$Y_i = b + \mathbf{X}_i^T \beta + f(\mathbf{T}_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2}$$

where b is the intercept, β is the $p \times 1$ vector of unknown coefficients for linear terms, f is an unknown function from $\mathbb{R}^q$ to $\mathbb{R}$ and the ε_i's f are i.i.d. random variables with mean 0 and variance σ^2 independent of the covariates. In order to ensure that the model is identifiable, we have to require that the linear covariates are centred and that an identifiability condition $\int f(\mathbf{t}) d\mathbf{t} = 0$ holds. In practice, the most used model for (2) is the following special case when $q = 1$:

$$Y_i = b + \mathbf{X}_i^T \beta + f(T_i) + \varepsilon_i, \quad i = 1, \ldots, n. \tag{3}$$

For example, in longitudinal data analysis, the time covariate T is often treated as the only non-linear effect. As we already mentioned in the introduction, model estimation and inference for (3) have been actively studied under various smooth regression settings, including smoothing splines, penalised regression splines, kernel smoothing and local polynomial regression. Interesting applications include the analysis of city electricity (Engle et al., 1986), household gasoline consumption in the United States (Schmalensee and Stoker, 1999), a marketing price-volume study in the petroleum distribution industry (Green and Silverman, 1994), the logistic analysis of bioassay data (Dinse and Lagakos, 1983), the mouthwash experiment (Speckman, 1988), and so on. In this section we briefly survey some of the most important of these developments from our perspective. We will not attempt to review the many applications of PLMs since they have become too numerous to review in the limited space available here. The monograph by Hardle et al. (2000) gives an excellent overview on partially linear models, and a more comprehensive list of references can be found there.

There are many methods to estimate the components of a PLM model. Kernel regression, including local constant (Speckman, 1988) and local linear techniques (Hamilton and Truong, 1997; Opsomer and Ruppert, 1999) have been used to study the partially linear models. A remarkable characteristic of the kernel-based methods is that under-smoothing has been sometimes imposed in order to get a root-n estimator of β (Green and Yandell, 1985; Opsomer and Ruppert, 1999), but the under-smoothing restriction is unnecessary in many settings (Speckman, 1988; Severini and Staniswalis, 1994). These can be confusing, and it may not be clear for users which strategy of smoothing parameter selection is appropriate. We present here a way of clarifying the essential differences by reviewing the profile-least-squares-based estimator, briefly mentioning backfitting estimator, and analysing the reasons for applying under-smoothing and regular smoothing.

The estimation approaches for the PLM are essentially based on the idea that an estimate $\hat{\beta}$ can be found for known $f(\cdot)$, and an estimate $\hat{f}(\cdot)$ can be found for known β. This leads to a profile least squares approach which converts the PLM to a classical linear regression model which is mainly a two steps estimation. Indeed note that, given some conditional moment assumptions on the covariates, one has, for a PLM model (2):

$$\mathbb{E}(\mathbf{Y}|\mathbf{T}) = b + \mathbb{E}(\mathbf{X}|\mathbf{T})^T\beta + f(\mathbf{T}).$$

It follows then by differencing

$$\mathbf{Y} - \mathbb{E}(\mathbf{Y}|\mathbf{T}) = \{\mathbf{X} - \mathbb{E}(\mathbf{X}|\mathbf{T})\}^T\beta + \varepsilon,$$

which is a standard linear model if $\mathbb{E}(\mathbf{Y}|\mathbf{T})$ and $\mathbb{E}(\mathbf{X}|\mathbf{T})$ were known. An intuitive estimator of β may then be defined as the least-squares estimator after appropriately estimating $\mathbb{E}(\mathbf{Y}|\mathbf{T})$ and $\mathbb{E}(\mathbf{X}|\mathbf{T})$ using a nonparametric regression setup.

For the remainder of this section we will assume that $q = 1$ and that T takes its values within the interval $[0,1]$ and that the function $f \in C^2[0,1]$. Extension to the multivariate case will be discussed at the end of this Section. There is a host of nonparametric methods for estimating these two regressions, including higher degree local polynomial kernel methods, kernel methods with varying bandwidths, smoothing and regression splines, etc. Following this route, large-sample results are available for inference on β and f, when $q = 1$ and the nonparametric component is smooth (twice continuously differentiable on its compact support) and is estimated using kernel regression or regression splines. These results rely on classical smoothing techniques that are sometimes quite sensitive to the specifics of their implementation in applications. Partially motivated by the poor finite-sample performance of conventional smoothing techniques, a recent literature on penalised spline estimation has emerged and is receiving considerable attention. Proposed by O'Sullivan (1986), and later popularised by Eilers and Marx (1996), this alternative smoothing technique has generated great interest because it is perceived as a very competitive alternative to classical nonparametric estimators. Therefore, we also review this approach in this section since it is relevant to the general modelling that we are going to study later on. Without any loss of generality, we state all the results conditionally in T, which amounts in assuming that the observed values of T define a non-stochastic grid of n distinct points $0 \le t_1 < \cdots < t_n \le 1$. Equation (3) in vector-matrix form becomes

$$\mathbf{Y} = b\mathbf{1} + \mathbf{U}\beta + \mathbf{f} + \varepsilon \tag{4}$$

for $\mathbf{Y} = (Y_1,\ldots,Y_n)^T$, $\mathbf{U}^T = [\mathbf{X}_1 \ldots \mathbf{X}_n]$, $\mathbf{f} = (f(t_1),\ldots,f(t_n))^T$ and $\varepsilon = (\varepsilon_1,\ldots,\varepsilon_n)^T$. To obtain a nonparametric estimator of the smooth function f we will use a class of penalised splines based on B-spline basis functions introduced by O'Sullivan (1986). O'Sullivan penalised splines are a direct generalisation of smoothing splines, in that the latter arise when the maximal number of B-spline basis functions is included. Like smoothing splines, O'Sullivan penalised splines possess the attractive feature of natural boundary conditions (e.g. Green and Silverman, 1994). They have also become the most widely used class of penalised splines in statistical analyses, as a result of their implementation in the popular R function `smooth.spline()` and associated generalised additive model software. For a brief description of O'Sullivan splines the reader is referred to Wand and Omerod (2008). For an integer K, let $\kappa_1,\ldots,\kappa_{K+8}$ be a knot sequence such that

$$0 = \kappa_1 = \kappa_2 = \kappa_3 = \kappa_4 < \kappa_5 < \cdots < \kappa_{K+4} < \kappa_{K+5} = \kappa_{K+6} = \kappa_{K+7} = \kappa_{K+8} = 1$$

and let $B_1, \ldots, B_{K+4}$ be the cubic B-spline basis functions defined by these knots (see, for example, pp. 160–161 of Hastie, Tibshirani and Friedman, 2009). Set up the $n \times (K+4)$ design matrix $\mathbf{B}$ with (i,k)th entry $B_{ik} = B_k(t_i)$, and the $(K+4) \times (K+4)$ penalty matrix Ω with (k,k')th entry

$$\Omega_{kk'} = \int_0^1 B_k^{(2)}(x) B_{k'}^{(2)}(x)\, dx.$$

For a smoothing parameter $\lambda > 0$, we will use $S_\lambda = \mathbf{B}(\mathbf{B}^T\mathbf{B} + \lambda\Omega)^{-1}\mathbf{B}^T$ to represent the $n \times n$ smoother matrix. Note that the cubic smoothing spline arises in the special case $K = n$ and $\kappa_{k+4} = t_k$, $1 \le k \le n$, provided that the t_i's are distinct (e.g. Green and Silverman, 1994). Apart from giving a smooth (twice continuously differentiable) scatterplot $S_\lambda\mathbf{Y}$ of the data $\mathbf{Y}$ without the parametric term, the procedure leads to a smooth function $\hat{f}_\lambda(\cdot)$ estimate that has good numerical properties. The basis functions are bounded and so not prone to overflow problems. Moreover, $\mathbf{B}^T\mathbf{B}$ is four-banded, which leads to $O(n)$ algorithms when K is close to n (e.g. Hastie et al., 2009). In addition $\hat{f}_\lambda(\cdot)$ satisfies natural boundary conditions. We will therefore assume that such a linear smoother is being used to conduct the spline smoothing transformations in all that follows.

If we assume that $\tilde{\mathbf{U}} = (\mathbf{I} - S_\lambda)\mathbf{U}$ has full column rank and set $\tilde{\mathbf{Y}}_\lambda = (\mathbf{I} - S_\lambda)\mathbf{Y}$, the Speckman-like profile least-squares estimators of the parameters in (4) are given by

$$\begin{aligned} \hat{b} &= \frac{1}{n}\mathbf{1}^T\mathbf{Y} \\ \hat{\beta}_\lambda &= (\tilde{\mathbf{U}}^T\tilde{\mathbf{U}})^{-1}\tilde{\mathbf{U}}^T\tilde{\mathbf{Y}}_\lambda \\ \hat{\mathbf{f}}_\lambda &= S_\lambda(\mathbf{Y} - \hat{b}\mathbf{1} - \mathbf{U}\hat{\beta}_\lambda). \end{aligned}$$

If a value of $\hat{f}_\lambda$ is required at a non design point t it is easy to get it using the corresponding B-splines basis vector evaluated at t and applied on the spline coefficients estimates. If $\tilde{\mathbf{U}}$ is less than full rank, formulas above remain valid provided we interpret $(\tilde{\mathbf{U}}^T\tilde{\mathbf{U}})^{-1}$ as a generalised inverse.

Existing asymptotic results for the above semi-parametric partial linear model estimation procedure establish parametric rates of convergence for the linear part and minimax rates for the nonparametric part, showing in particular that the existence of a linear component does not change the rates of convergence of the nonparametric component and conversely. Assuming for example that T is continuously distributed within $[0,1]$ with Lebesgue density bounded above and below from 0, that the nonparametric function f is twice continuously differentiable on $[0,1]$, that the knots sequence is asymptotically equidistant and controlled by the tuning parameter $K_n \to \infty$ in such a way that $\ln(K_n)K_n/n \to 0$ and that $\mathbb{E}(\varepsilon_i^4|(\mathbf{X},T))$ and $\mathbb{E}(\|\mathbf{X}\|^4|T)$ are bounded, then choosing $K_n = O(n^{1/5})$ and $\lambda_n = O(n^{1/5})$ results in a square-root n consistent estimator for β and an optimal rate of convergence for the penalised B-spline estimator (see, e.g. Claeskens, Krivobokova and Opsomer, 2009; Holland, 2017).

The above asymptotic result seems to contradict the conclusion obtained by Opsomer and Ruppert (1999) for the backfitting estimator that one needs to under-smooth the nonparametric part to get a root-n consistent estimator for the linear part. Recall that the backfitting method has been suggested as an iterative algorithm to fit an additive model (see Hastie and Tibshirani, 1986; Buja, Hastie and Tibshirani, 1989). Its main idea is to regress the additive components separately on partial residuals. The PLM is again a special case, consisting of only two additive functions. To simplify notation we may assume without loss of generality that the data are centred, since the empirical mean is an efficient estimator of b. Denote by $\mathscr{P} = \mathbf{U}(\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T$ the projection matrix from a linear regression model and by S_λ the smoothing matrix as before. Then backfitting means to solve

$$\begin{aligned} \mathbf{U}\hat{\beta} &= \mathscr{P}(\mathbf{Y}-\hat{\mathbf{f}}) \\ \hat{\mathbf{f}} &= S_\lambda(\mathbf{Y}-\mathbf{U}\hat{\beta}) \end{aligned}$$

as $\mathbf{Y}-\hat{\mathbf{f}}$ are the residuals from a nonparametric fit and $\mathbf{Y}-\mathbf{U}\hat{\beta}$ the residuals from a linear regression. In this case no iteration is necessary and the explicit solution leading to the backfitting estimators $\hat{\beta}_{\text{back}}$ and $\hat{\mathbf{f}}_{\text{back}}$ is

$$\begin{aligned} \hat{\beta}_{\text{back}} &= \{\mathbf{U}^T(\mathbf{I}-S_\lambda)\mathbf{U}\}^{-1}\mathbf{U}^T(\mathbf{I}-S_\lambda)\mathbf{Y} \\ \hat{\mathbf{f}}_{\text{back}} &= S_\lambda(\mathbf{Y}-\mathbf{U}\hat{\beta}_{\text{back}}) \end{aligned}$$

These estimators differ from the Speckman-like estimators in only a subtle detail: the Speckman-like estimator for β shows $(\mathbf{I}-S_\lambda)^T(\mathbf{I}-S_\lambda)$ instead of $(\mathbf{I}-S_\lambda)$. And this is the main reason that for $\hat{\beta}_{\text{back}}$ to be root-n consistent, under-smoothing for the nonparametric part is unavoidable. This was also noticed by Rice (1986) who showed that the partial spline estimate of the parametric component in a semiparametric regression model is generally biased and it is necessary to under-smooth the nonparametric component to force the bias to be negligible with respect to the standard error. We will not pursue further the discussion here.

In practice and for finite samples assuming the sequence K_n is fixed, the smoothing parameter λ_n must be chosen in a data driven way. Let $\mathbf{H}_\lambda = S_\lambda + \tilde{\mathbf{U}}(\tilde{\mathbf{U}}^T\tilde{\mathbf{U}})^{-1}\tilde{\mathbf{U}}^T(\mathbf{I}-S_\lambda)$. To use the estimators above one must select a value for the smoothing parameter λ.Two standard data-driven choices are the minimiser of the generalised cross-validation (GCV) criterion

$$GCV(\lambda) = \frac{n\|(\mathbf{I}-\mathbf{H}_\lambda)(\mathbf{Y}-\hat{b}\mathbf{1})\|^2}{(n-\text{tr}(\mathbf{H}_\lambda))^2}$$

and the minimiser of the unbiased risk (UBR) criterion

$$R(\lambda) = n^{-1}(\|(\mathbf{I}-\mathbf{H}_\lambda)(\mathbf{Y}-\hat{b}\mathbf{1})\|^2 + 2\sigma^2\text{tr}(\mathbf{H}_\lambda)).$$

The use of generalised cross-validation is feasible because the first-order properties of the cross-validation function are basically not changed by the addition of a parametric term to the model (Speckman, 1988). However for reasons to be seen later we also use the UBR criterion. Both these criteria require computation of $\text{tr}(\mathbf{H}_\lambda)$ but one may use an $O(n)$ algorithm developed by Eubank, Kambour, Kim, Klipple, Reese and Schimek (1998) for that purpose. However, for the computation of the UBR criterion, appropriate estimates of the unknown variance σ^2 are necessary. Here, we suggest to apply an adaptation of a difference-based variance estimator due to Gasser, Sroka and Jennen-Steinmetz (1986). This estimator has been successfully applied to kernel, smoothing spline and wavelet regression (see e.g. Antoniadis and Lavergne, 1995).

To define the estimator first let

$$a_i = \frac{t_{i+1}-t_i}{t_{i+1}-t_{i-1}}, \quad b_i = \frac{t_i-t_{i-1}}{t_{i+1}-t_{i-1}}, \quad c_i = (a_i^2+b_i^2+1)^{1/2},$$

for $i = 2,\dots,n-1$. Then take $\mathbf{D}$ to be the $(n-2)\times n$ matrix whose ith row has all zero entries except for its ith, $(i+1)$th and $(i+2)$th entries which are $a_i c_i$, $-c_i$ and $b_i c_i$ respectively. Note that $\mathrm{tr}(\mathbf{D}^T\mathbf{D}) = n-2$, and hence the Gasser et al. (1986) estimator of σ^2 is

$$\tilde{\sigma}^2 = \frac{\mathbf{Y}^T\mathbf{D}^T\mathbf{D}\mathbf{Y}}{\mathrm{tr}(\mathbf{D}^T\mathbf{D})}. \tag{5}$$

When f is smooth, $\mathbf{D}\mathbf{f}$ is essentially $\mathbf{0}$ so that $\mathbf{D}\mathbf{Y} \approx \mathbf{D}\mathbf{U}\beta + \mathbf{D}\varepsilon$ and $\tilde{\sigma}^2$ in equation (5) will be efficient in estimating σ^2 when $\beta = 0$. To deal with the situation where $\beta \neq 0$ it suffices to make a simple adjustment and define the modified estimator

$$\hat{\sigma}^2 = \frac{\mathbf{Y}^T\mathbf{D}^T(\mathbf{I}-\mathbf{Q})\mathbf{D}\mathbf{Y}}{\mathrm{tr}(\mathbf{D}^T(\mathbf{I}-\mathbf{Q})\mathbf{D})},$$

with $\mathbf{Q} = \mathbf{D}\mathbf{U}(\mathbf{U}^T\mathbf{D}^T\mathbf{D}\mathbf{U})^{-1}\mathbf{U}^T\mathbf{D}^T$. Under our assumptions on the design points and the smoothness of f, Theorem 3.1 of Eubank et al. (1998) ensures that the above variance estimator has small bias and is root-n consistent.

Remark 1 The parameters in the PLM model with $q = 1$ are mainly estimated in the procedures we have described by least-squares profiling and may therefore be affected by the presence of outliers. It is possible to use a more robust approach by considering the Huber-Dutter estimators of β, scale σ for the errors and the function f respectively, using again a smoothing B-spline basis representation for f, obtained by minimising

$$\sum_{i=1}^{n} \rho\left\{\frac{\mathbf{Y}_i - b - \mathbf{X}_i^T\beta - f(t_i)}{\sigma}\right\}\sigma + v_n\sigma,$$

over b, β, σ and f approximated by its spline decomposition as before. The convex function ρ is Huber's loss function and v_n is a suitable chosen sequence of constants. Under some regularity conditions, it is shown in Tong, Cui and Zhao (2005) that the Huber-Dutter estimators of b, β and σ are asymptotically normal with convergence rate $n^{-1/2}$ and the B-spline Huber-Dutter estimator of f achieves the optimal convergence rate in nonparametric regression.

We have conducted some small simulations, just to illustrate the results given above using some standard R packages (see R, 2015). We generated data from the (2) model, with $n = 256$, $b = 0$, β a p-dimensional vector with $p = 4$, $\beta = (4,3,1,9)^T$, $\mathbf{X}$ a p-dim matrix whose columns are i.i.d. realisations of standard normal distributed variables and T a regular design of n equidistant points within $[0,1]$. We considered two nonparametric cases:

(a) $f(t) = 5\sin(5\pi t) + 100(\exp(-3.25t) - 4\exp(-6.5t) + 3\exp(-9.75t))$ (smooth function)

(b) $f(t) = 1.6603\big(10t^2 I(t \leq 0.2) - (2(t-0.65)^2 - 0.15)I(t \leq 0.7)I(t > 0.2) +$
$5(t-0.7)^2 I(t > 0.7) + 0.2572\big)$ (non regular)

The second case was chosen in order to see how the standard PLM fitting with kernel or splines behaves when the nonparametric component is not smooth. The functions are rescaled such that an added normal noise with standard deviation of 2.5 produces a preassigned signal-to-noise ratio

(SNR) of 3. Fitting was done for $M = 100$ replications. For each replication the estimation accuracy of $\hat{\beta}$ is measured by the mean squared error (MSE) defined as $||\hat{\beta} - \beta||_2^2$ and reconstruction of the theoretical nonparametric component was measured by the mean squared error (MSE), calculated as

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

where M is the number of simulation runs and $\hat{f}_r$ the estimation of nonparametric component f in the rth simulation run. Fitting of data was realised using the `gplm` R-package of Müller (2014) when using splines and the package `PMRModels` when using kernel smoothing . For each simulation run, the optimal values of the hyper-parameters are selected by means of the generalised cross-validation procedure.

Figure 1 reports the simulation results in terms of MSE.

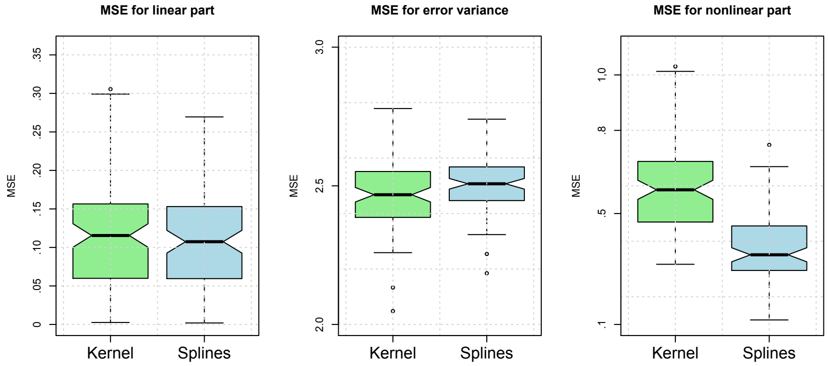

Figure 1: Simulation results for the smooth function experiment (model (2)) with linear part $X^T\beta$ and smooth nonlinear part (case (a)) $f(t) = 5\sin 5t + 100(\exp(-3.25t) - 4\exp(-6.5t) + 3\exp(-9.75t)))$: boxplot of the Mean Squared Error over 100 simulation runs for the linear part (left), error variance (middle) and nonlinear part (right) using splines (R-package `gplm`) and kernel smoothing (R-package `PMRModels`).

Figure 2 shows a typical fit of the nonparametric component by the two methods on one of the simulations (case (a)).

However, when the nonparametric component in the model is not smooth (as the one for example in case (b)), while the parameter β can still be estimated in a satisfactory way, the kernel or spline smoothing estimation method produce much more erratic results as one can see in Figure 3.

Let us now again consider the general multivariate partial linear model stated in (2) but with **T** being q-dimensional vector. As we saw earlier in this section, the case of $q = 1$ has extensively been studied and we gave several references for this. The case where $q > 1$ has received much less attention in the past. However, the general case with a high-dimensional nonlinear component makes the analysis complicated because of the "curse of dimensionality" problem. Most of the papers that discuss that model consider situations in which the nonlinear component is low-dimensional; that is, q is relatively small. Still using the idea of profile least-squares the estimation

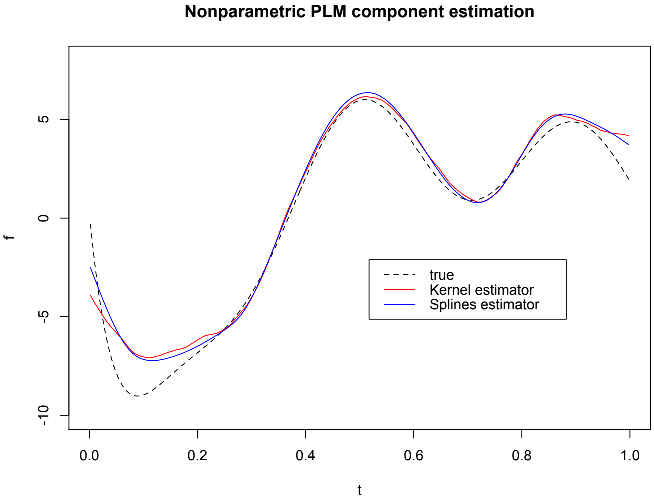

Figure 2: Typical fit of a sample realisation for the smooth function experiment (model (2)) with linear part $X^T\beta$ and smooth nonlinear part (case (a)) $f(t) = 5\sin 5t + 100(\exp(-3.25t) - 4\exp(-6.5t) + 3\exp(-9.75t)))$. True function: broken line; spline estimate (R-package `gplm`): blue solid line; kernel estimate (R-package `PMRModels`): red solid line.

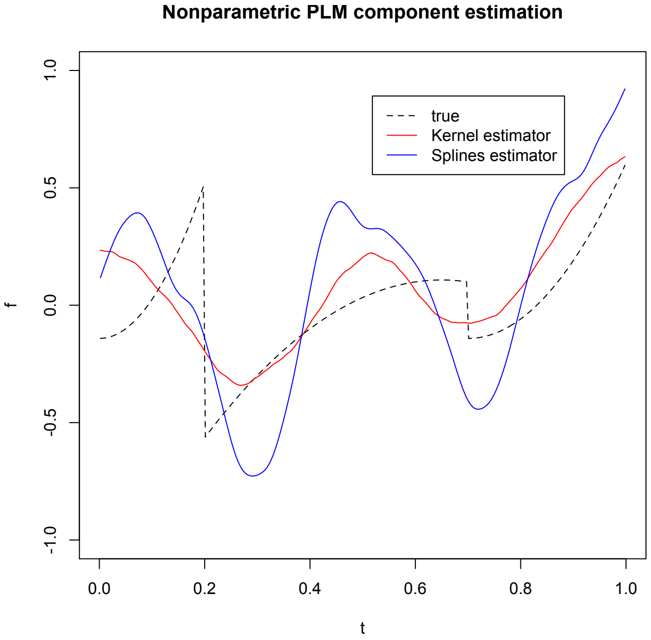

Figure 3: Typical fit of a sample realization for the nonregular function experiment (model (2)) with nonregular nonlinear part (case (b)) $f(t) = 1.6603(10t^2 I(t \le 0.2) - (2(t-0.65)^2 - 0.15)I(t \le 0.7)I(t > 0.2) + 5(t-0.7)^2 I(t > 0.7) + 0.252))$. True function: broken line; spline estimate (R-package `gplm`): blue solid line; kernel estimate (R-package `PMRModels`): red solid line.

of the nonlinear component is realised using either multivariate kernel methods or marginal integration or local scoring and local polynomial methods (see e.g. Robinson, 1988; Samarov, Spokoiny and Vial, 2005; Schick, 1996; Li, 2000). To our knowledge, for general q larger than 5, and without further structural assumptions on f, the only papers that address the problem are the recent papers by Cattaneo, Jansson and Newey (2016) relying on some earlier results by Donald and Newey (1994) on series estimators, and the already cited work of Holland (2017) on penalised spline estimation in the partially linear model. However, the "curse of dimensionality" is still mirrored in the derived asymptotic rates. To get rid of this curse, the regression modelling using PLMs can be extended to partial linear additive models (PLAM), where the nonparametric function f in the nonlinear part is now substituted by a sum of q univariate smooth functions of the underlying components of $\mathbf{T}$. Since this type of models will be studied in Section 4 together with variable selection, we postpone the derivation of the estimation procedures for such models to this later section.

3. Wavelets for PLM ($q = 1$)

In this section we consider again the regression problem stated in (3), but this time with a non-stochastic equidistant design $t_i = i/n$, $i = 1, \ldots, n$ of size $n = 2^J$ for some positive integer J, noise variables ε_i that are i.i.d. Gaussian $\mathcal{N}(0, \sigma^2)$ and with a potentially non-smooth function f that may present a wide range of irregular effects. There is not any loss of generality to assume that $b = 0$ since this parameter can always be estimated with a root-n rate by the empirical mean of the observations. We will therefore assume that the data has been centred. To deal with cases of a less-smooth nonparametric component, several wavelet based estimation procedures have been developed in the literature (see e.g. Chang and Qu, 2004; Fadili and Bullmore, 2004; Qu, 2006; Gannaz, 2007; Antoniadis, 2007; Ding, Claeskens and Jansen, 2011) and we will review here the ones that are the most relevant to our work. Using wavelets allows the nonparametric component to be parsimoniously represented by a limited number of coefficients. Models for the nonparametric component of the PLM model, that allow a wide range of irregular effects, are through the sequence space representation of Besov spaces. The (inhomogeneous) Besov spaces on the unit interval, $\mathscr{B}^s_{\pi,r}([0,1])$, consist of functions that have a specific degree of smoothness in their derivatives. The parameter π can be viewed as a degree of a function's inhomogeneity while s is a measure of its smoothness. Roughly speaking, the (not necessarily integer) parameter s indicates the number of the function's (fractional) derivatives, where their existence is required in an L^π-sense; the additional parameter r is secondary in its role, allowing for additional fine tuning of the definition of the space. For a detailed study on (inhomogeneous) Besov spaces we refer to, e.g., Donoho and Johnstone (1998). To capture key characteristics of variations in f and to exploit its sparse wavelet coefficients representation, we will assume that f belongs to $\mathscr{B}^s_{\pi,r}([0,1])$ with $s + 1/\pi - 1/2 > 0$. The last condition ensures in particular that evaluation of f at a given point makes sense. To adopt a wavelet-based model specification of the PLM model we will first present in the next subsection some relevant facts and notation about wavelet decompositions of such functions.

3.1. Wavelet series expansions and discrete wavelet transform

Throughout the paper we assume that we are working within an orthonormal basis generated by dilatations and shifts of a compactly supported scaling function, $\phi(t)$, and a compactly supported mother wavelet, $\psi(t)$, associated with an r-regular ($r \geq 0$) multi-resolution analysis of $\left(L^2[0,1], \langle\cdot,\cdot\rangle\right)$, the space of squared-integrable functions on $[0,1]$ endowed with the inner product $\langle f,g\rangle = \int_0^1 f(t)g(t)dt$. For simplicity in exposition, we work with periodic wavelet bases on $[0,1]$ (see, e.g., Mallat (1999), Section 7.5.1), letting

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

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

$$\{\phi^{\mathrm{p}}_{j_0k},\ k = 0,1,\ldots,2^{j_0}-1;\ \psi^{\mathrm{p}}_{jk},\ j \geq j_0;\ k = 0,1,\ldots,2^j-1\}$$

is then an orthonormal basis of $L^2[0,1]$. The superscript "p" will be suppressed from the notation for convenience. Despite the poor behaviour of periodic wavelets near the boundaries, where they create high amplitude wavelet coefficients, they are commonly used because the numerical implementation is particularly simple. Therefore, for any $f \in L^2[0,1]$, we denote by $c_{j_0k} = \langle f, \phi_{j_0k}\rangle$ ($k = 0,1,\ldots,2^{j_0}-1$) the scaling coefficients and by $d_{jk} = \langle f, \psi_{jk}\rangle$ ($j \geq j_0$; $k = 0,1,\ldots,2^j-1$) the wavelet coefficients of f for the orthonormal periodic wavelet basis defined above; the function f is then expressed in the form

$$f(t) = \sum_{k=0}^{2^{j_0}-1} c_{j_0k}\phi_{j_0k}(t) + \sum_{j=j_0}^{\infty}\sum_{k=0}^{2^j-1} d_{jk}\psi_{jk}(t), \quad t \in [0,1].$$

The approximation space spanned by the scaling functions $\{\phi_{j_0k},\ k = 0,1,\ldots,2^{j_0}-1\}$ is usually denoted by V_{j_0} while the details space at scale j, spanned by $\{\psi_{jk},\ k = 0,1,\ldots,2^j-1\}$ is usually denoted by W_j.

In a statistical settings, we are more usually concerned with discretely sampled, rather than continuous, functions. It is then the wavelet analogy to the discrete Fourier transform which is of primary interest and this is referred to as the discrete wavelet transform (DWT). Given a vector of real values $\mathbf{e} = (e_1,\ldots,e_n)^T$, the discrete wavelet transform of $\mathbf{e}$ is given by $\mathbf{d} = W_{n\times n}\mathbf{e}$, where $\mathbf{d}$ is an $n\times 1$ vector comprising both discrete scaling coefficients, s_{j_0k}, and discrete wavelet coefficients, w_{jk}, and $W_{n\times n}$ is an orthogonal $n\times n$ matrix associated with the orthonormal periodic wavelet basis chosen. In the following we will distinguish the blocks of $W_{n\times n}$ spanned by the scaling functions and the wavelets, respectively. The empirical coefficients s_{j_0k} and w_{jk} of $\mathbf{e}$ are given by

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

When $\mathbf{e}$ is a vector of function values $\mathbf{f} = (f(t_1),...,f(t_n))^T$ at equally spaced points t_i, the corresponding empirical coefficients s_{j_0k} and w_{jk} are related to their continuous counterparts c_{j_0k} and

d_{jk} (with an approximation error of order n^{-1}) via the relationships $s_{j_0k} \approx \sqrt{n}\, c_{j_0k}$ and $w_{jk} \approx \sqrt{n}\, d_{jk}$. Note that, because of orthogonality of $W_{n\times n}$, the inverse DWT (IDWT) is simply given by $\mathbf{f} = W_{n\times n}^{\mathrm{T}}\mathbf{d}$, where $W_{n\times n}^{\mathrm{T}}$ denotes the transpose of $W_{n\times n}$. If $n = 2^J$ for some positive integer J, the DWT and IDWT may be performed through a computationally fast algorithm (see e.g. Mallat, 1999, Section 7.3.1) that requires only order n operations. Hereafter, the coarsest wavelet decomposition level j_0 will be chosen to be the closest integer to $\log_2(\log(n)) + 1$, as suggested by Antoniadis, Bigot and Sapatinas (2001).

3.2. A wavelet-based model specification of the PLM model

We will adopt the vector-matrix form of the centred PLM model, given by (4) with $b = 0$:

$$\mathbf{Y} = \mathbf{U}\beta + \mathbf{f} + \varepsilon,$$

for $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, $\mathbf{U}^T = [\mathbf{X}_1 \ldots \mathbf{X}_n]$, $\mathbf{f} = (f(t_1), \ldots, f(t_n))^T$ and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$. After applying a linear and orthogonal wavelet transform, the discretised above model becomes

$$\mathbf{Z} = \mathbb{A}\,\beta + \gamma + \tilde{\varepsilon},$$

where $\mathbf{Z} = W_{n\times n}\mathbf{Y}$, $\mathbb{A} = W_{n\times n}\mathbf{U}$, $\gamma = W_{n\times n}\mathbf{f}$ and $\tilde{\varepsilon} = W_{n\times n}\varepsilon$. The orthogonality of the DWT matrix $W_{n\times n}$ ensures that the transformed noise vector $\tilde{\varepsilon}$ is still distributed as a Gaussian white noise with variance $\sigma^2\mathbf{I}_n$. Hence, the representation of the model in the wavelet domain not only allows one to retain the partly linear structure of the model, but also to exploit in an efficient way the sparsity of the wavelet coefficients in the representation of the nonparametric component.

We already mentioned several methods from the partially linear wavelet model literature. Both methods proposed by Chang and Qu (2004) and Fadili and Bullmore (2004) rely upon backfitting and therefore present the same asymptotic weaknesses as the backfitting procedures when smoothing splines are used. The difference between these two procedures is the choice of the thresholding parameters chosen at each iteration, with the Fadili and Bullmore algorithm significantly improving the performance. The Bayesian wavelet-based algorithm for the same problem was proposed by Qu (2006). However, we found that the implementation of that algorithm is not robust to different simulated examples and initial values of the empirical Bayes procedure, therefore, we omit it from our discussion. The procedure based on Gannaz (2007) is a wavelet thresholding based estimation procedure solved by the proposed LEGEND algorithm and the interested reader is referred to that paper for details. The formulation of the problem is based on an ℓ_1-penalised mean-shift linear model, penalising only the wavelet coefficients of the nonparametric part, but the solution is faster than backfitting. The algorithm by Ding et al. (2011) addresses variable selection in the linear part and thresholding in the wavelet part by adopting again an ℓ_1-penalised mean-shift linear model as in Gannaz (2007), but with an extra LASSO penalty on the linear parameters. No asymptotic results are discussed in their paper.

In what follows we will review the ℓ_1-penalised mean-shift linear model approach of Antoniadis (2007) and its connection with robust estimation via proximal maps since these are the ones that will be extended to the general case later. With the above notation, consider the ℓ_1-penalised mean-shift linear regression

$$\min_{\beta,\gamma} Q_n(\beta, \gamma),$$

where

$$Q_n(\beta,\gamma) = \left\{ \frac{1}{2}\|\mathbf{Z} - \mathbb{A}\beta - \gamma\|^2 + \lambda \sum_{j=1}^{n} w_j|\gamma_j| \right\},$$

for $\beta \in \mathbb{R}^p$ and $\gamma \in \mathbb{R}^n$ and w_j are given weights. When β is fixed, the objective function $\frac{1}{2}\|\mathbf{Z} - \mathbb{A}\beta - \gamma\|^2$ becomes $\frac{1}{2}\|\mathbf{r} - \gamma\|^2$ where $\mathbf{r} := \mathbf{Z} - \mathbb{A}\beta$ and the minimisation problem of $Q_n(\beta,\gamma)$ with respect to γ is separable. To solve it we need only to deal with the univariate case

$$\min_{\gamma} \left\{ \frac{(\gamma - r)^2}{2} + p_\lambda(w|\gamma|) \right\}. \tag{6}$$

When $p_\lambda(\cdot)$ is a (closed) convex function, the solution of (6) is unique, and defines the proximal mapping or proximal operator of p_λ given by (see e.g. Antoniadis, 2007):

$$\mathrm{Prox}_{p_\lambda}(r) = \arg\min_{u} \left\{ \frac{(u - r)^2}{2} + p_\lambda(w|u|) \right\}. \tag{7}$$

While proximal operators are well studied for convex functionals, the non-convex case has been of interest to researchers recently (see Hare and Sagastizabal, 2009). Very often, even when $p_\lambda(\cdot)$ is non-convex, problem (7) results in a unique solution and allows us to define an appropriate proximal map. When $p_\lambda(\cdot) = \lambda|\cdot|$ (ℓ_1 penalty) the corresponding proximal map is the soft-thresholding operator (Antoniadis, 2007)

$$\Theta_{\text{soft}}(r;\lambda) = \begin{cases} 0, & \text{if } |r| \le \lambda \\ r - \mathrm{sgn}(r)\lambda & \text{if } |r| > \lambda. \end{cases}$$

Recalling that $\mathbf{r} := \mathbf{Z} - \mathbb{A}\beta$, finding the optimal estimate $\hat{\beta}$ requires the minimisation of the profile loss function

$$Q_n(\tilde{\gamma}(\beta),\beta),$$

where $\tilde{\gamma}(\beta)$ denotes the estimate obtained by applying soft-thresholding component-wise on $w_j r_j$. One can then show (see Gannaz, 2007) that

$$Q_n(\tilde{\gamma}(\beta),\beta) = \rho_\lambda\left(\mathrm{diag}(\mathbf{w})\mathbf{r}\right),$$

where $\rho_\lambda(\cdot)$ is Huber's loss function and one sees here the connection with the Huber-Dutter estimates of β and f outlined in Remark 1. We will exploit this connection later in this section.

Using Donoho and Johnstone (1998) universal threshold $\lambda(n) = \sigma\sqrt{2\log n}$ one can show, under appropriate conditions on the design (see Theorem 1 of Gannaz, 2007) that

$$\|\hat{\beta} - \beta\|_2 = O_P\left(\sqrt{\frac{\log n}{n}}\right)$$

and that

$$\|\hat{\mathbf{g}} - \mathbf{g}\|_2 = O_P\left(\left(\frac{\log n}{n}\right)^{\frac{s}{2s+1}}\right).$$

Therefore even in the case of much less regular component f, the rates of convergence are similar to the case where f is twice continuously differentiable, but an extra logarithmic factor will appear in the rates of the parametric part and nonparametric parts, mainly due to the fact that our smoothness assumptions on the nonparametric part are weaker. A good estimate of σ is required. Gannaz (2007) makes use of an estimator that does not require estimation of β and uses a half-quadratic regularisation algorithm (LEGEND) for estimating the parameters. However the computational time of the LEGEND algorithm can be quite long.

Remark 2 Wang, Brown and Cai (2011a), inspired by Gasser-like difference-based estimators and their use in semiparametric regression by Yatchew (1997) have analysed the PLM model using a difference based approach. Their procedure estimates the linear component based on the *m*th-order differences of the observations and then estimates the nonparametric component by a wavelet thresholding method using the residuals of the linear fit. It is shown that both the estimator of the linear component and the estimator of the nonparametric component asymptotically perform as well as if the other component were known. The estimator of the linear component is asymptotically efficient (root-*n* consistent) and the estimator of the nonparametric component is asymptotically minimax rate optimal. However the asymptotic results require that the order $m \to \infty$ which is difficult to achieve in practice.

The connection between the minimisation of the profile function $Q_n(\tilde{\gamma}(\beta), \beta)$ and soft-thresholding to get the estimates of the wavelet coefficients and Huber's loss has been thoroughly discussed by She and Owen (2011) which outline the inherent difficulty of ℓ_1-penalised regression. In particular, they stress out the fact that Huber's method cannot handle even moderate leverage points well (Huber, 1981, p. 192) and is prone to masking (failing to identify outliers) and swamping (mistaking clean observations for outliers) in outlier detection. Its breakdown point is 0. They advocate using instead hard thresholding which corresponds to a mean truncation loss, which solves the masking and swamping problems noted before. One could also use SCAD thresholding (Antoniadis and Fan, 2001; Fan and Li, 2001) which is a special case of Hampel's rule (see Antoniadis, 2007). Whatever method is used, the issue of tuning the thresholding parameter λ still remains.

3.3. Tuning the parameter by square-root LASSO

A promising way to remain robust and to improve the tuning of λ is to adopt a different objective function as in Antoniadis and Fan (2001). Let us formulate the model in the wavelet domain as a high-dimensional linear regression model

$$\mathbf{Z} = [\mathbb{A}\mathbf{I}_n](\beta^T \gamma^T)^T + \tilde{\varepsilon} = \Xi\theta + \tilde{\varepsilon},$$

with $\Xi = [\mathbb{A}\mathbf{I}_n]$ and $\theta = (\beta^T \gamma^T)^T$. Note that the dimension of θ is larger than n and therefore estimation of σ is nontrivial and remains an outstanding practical and theoretical problem. To estimate θ and eventually σ one may then use square-root LASSO procedures suggested first in Antoniadis (2010) and studied in Belloni et al. (2011). Belloni et al. (2011) have shown that for a Gaussian or sub-Gaussian regression model with constant but unknown variance, the square-root LASSO with a deterministic value of the penalty parameter that depends only on known parameters achieves near-oracle performance for estimation and model selection of the mean. In our PLM case, the square-root LASSO estimation of θ eliminates the need to know or to pre-estimate σ. Using

$\lambda = 1.1n^{-1/2}\Phi^{-1}(1-0.05/(2(n+p)))$ in

$$\hat{\theta} = \text{argmin}_\theta \left(\frac{\|\mathbf{Z} - \Xi\theta\|_2}{2\sqrt{n}} + \lambda \sum_{j=1}^{n+p} \|\Xi_{\cdot,j}\|_2 |\theta_j| \right)$$

one obtains a consistent estimator of θ together with a root-n consistent estimate of σ:

$$\hat{\sigma} = \frac{1}{\sqrt{n}} \|\mathbf{Z} - \Xi\hat{\theta}\|_2.$$

The minimisation problem leading to $\hat{\theta}$ can be solved by a Second Order Cone Program (SOCP).

Just for illustration, Figure 4 illustrates a fit of the same simulated dataset fitted by splines in the previous subsection (see Figure 3), but obtained this time via some of the wavelet procedures described in this Section. Since we have used periodic Symmlets of order 7 to obtain the estimation, we display the fits within the interval $[0.07, 0.94]$ to get rid of the boundary effects since the function in case (b) is not periodic.

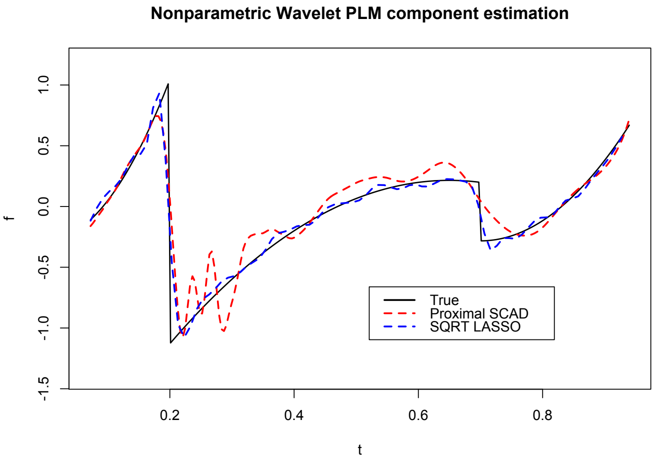

Figure 4: Typical fit of a sample realization for the nonregular function experiment (model 2) with nonregular nonlinear part (case (b)) $f(t) = 1.6603(10t^2 I(t \leq 0.2) - (2(t-0.65)^2 - 0.15)I(t \leq 0.7)I(t > 0.2) + 5(t-0.7)^2 I(t > 0.7) + 0.252))$. True function: black solid line; wavelet estimate using proximal SCAD: broken red line; wavelet estimate using SQRT LASSO: broken blue line.

A distinguishing feature of the proposed algorithm is that it can also be used for variable selection. The method proposed by Ding et al. (2011) was also developed for variable selection in the linear part of the PLM. Concerning the estimation of the nonparametric part of the PLM, their algorithm applies an iterative backfitting-like algorithm as well as soft thresholding. To do so they propose minimising the following double penalised least-squares criterion to estimate β and γ:

$$R_n(\beta, \gamma) = \left\{ \frac{1}{2} \|\mathbf{Z} - \mathbb{A}\beta - \gamma\|^2 + \lambda_1 \sum_{j=1}^{n} w_j |\gamma_j| + \lambda_2 \sum_{i=1}^{p} \tau_i |\beta_i| \right\}, \tag{8}$$

where w_j and τ_i are given weights and where the threshold value λ_1 controls the sparsity of nonparametric estimates, while for the purpose of variable selection they impose the weighted ℓ_1 penalty on the linear coefficients β and use λ_2 to control the sparsity of the parametric part of the model. For any fixed λ_1 and λ_2 they use an iterative backfitting like algorithm to solve the above estimation problem. The tuning parameters λ_1 and λ_2 are chosen in a data-driven way (GCV for λ_1 and BIC for λ_2 which again needs a consistent estimator of σ^2). While there is no asymptotic analysis and study of their procedure we believe that the same drawback of smoothing splines holds here leading again to suboptimal consistency rates for the estimators of the nonparametric component.

Remark 3 The methodology proposed in this Section, is designed for treating dyadic samples of equispaced data. The application of a wavelet analysis to irregularly spaced samples has been a subject of study for more than ten years. Most methods in the area work with a pre- and/or post-processing of the data in order to translate the problem into an equispaced one. Cai and Brown (1998) decompose the non-equispaced data into a warped wavelet basis and then project this decomposition onto a regular wavelet basis. Antoniadis and Pham (1998) implement a direct discretisation of a continuous wavelet analysis on the irregular grid to find numerical values for wavelet coefficients corresponding to regular basis functions. Kovac and Silverman (2000) interpolate the irregular observations in intermediate regular locations before starting the wavelet analysis. These and other methods require user-driven preprocessing, that might become difficult or even fail in case the data are "very" non-equidistant. As we will see in later sections it is also possible to use wavelet basis functions evaluated on irregular grids as in Antoniadis and Fan (2001) and Wand and Omerod (2011) which provide a way of handling nonequispaced predictor data.

4. Splines for Estimation and Variable Selection in Additive Partial Linear Models

We have already mentioned that partially linear additive models (PLAMs), which are a special case of additive nonparametric models, retain the parsimony and interpretability of linear models and the flexibility of nonparametric additive regression, by allowing a linear component for some predictors which are presumed to have a strictly linear effect, and an additive structure for other predictors, thus reducing the problem known as "curse of dimensionality". While PLAMs have become widely used since their introduction, their applicability has until recently been limited to problem settings where the number of covariates is modest relative to the number of observations. The last decade has seen the emergence of large data sets with big sets of variables that are more and more commonly collected in modern research studies. In such large data sets it is often fair to assume that a large number of the measured variables are irrelevant or redundant for the purpose of predicting the response and this has stimulated vast developments in efficient procedures that can perform estimation variable selection on such large data sets. Wang, Liu, Liang and Carroll (2011b) consider PLAMs, restricting to variable selection for the linear part only. Liu et al. (2011) developed a SCAD-based variable selection procedure to identify significant linear components using the smoothly clipped absolute deviation penalty (SCAD), using a spline based approximation for the nonparametric components. In the same spirit, Wei (2012) applies group variable selection for the linear parameter in high-dimensional PLAMs with a fixed number of nonparametric components using an adaptive group

LASSO and spline approximation of the nonparametric components. Du, Cheng and Liang (2012) introduced a penalisation procedure with penalty that combines the adaptive empirical L_2-norms of the nonparametric component functions and a SCAD penalty on the coefficients in the parametric part to simultaneously achieve estimation and model selection for both nonparametric and parametric parts in PLAMs with diverging dimensions of parametric components. When no linear components are present, this method is related to SpAM (Ravikumar, Liu, Lafferty and Wasserman, 2009) or additive model selection procedures studied in Amato et al. (2016). This is the method that we will review in this Section.

We suppose that the observed data $(y_i, \mathbf{x}_i, \mathbf{t}_i)$, $i = 1, \ldots, n$ is modelled by a semiparametric regression model as in (1) with $q > 1$. Hereafter, we also suppose that the responses Y_i have been centred by subtracting the sample mean $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Also, under our assumptions on the nonlinear covariates, there is no loss of generality in considering that $\mathbb{E}(f_j^2(\mathbf{T}_j))$ is proportional to $\int_0^1 f_j^2(x)dx$ or $||\mathbf{f}_j||_n^2$ (if n is large), where $\mathbb{E}$ denotes expectation with respect to the distribution of the covariates in the random design case or with respect to the empirical distribution of the covariates in the deterministic design case. We will assume that the unknown nonparametric (smooth) additive components f_j belong to the subspace of centred functions within the Sobolev space of order m. We briefly recall that the Sobolev space of order m is defined as

$$\mathscr{W}_2^m = \left\{ f : [0,1] \to \mathbb{R} | f, f^{(1)}, \ldots, f^{(m-1)} \text{ are absolutely continuous and } f^{(m)} \in L_2([0,1]) \right\}.$$

There are many possible norms that can equip $\mathscr{W}_2^m$ to make it a Hilbert space (see e.g. Adams, 1975). When dealing with mean zero functions $f \in \mathscr{W}_2^m$ one may consider the Sobolev norm $||f||_m = \sqrt{\int_0^1 \left(f^{(m)}(x)\right)^2 dx}$.

4.1. Partial splines, SpAM and SCAD

Marra and Wood (2011) have addressed some estimation procedures for PLAMs, when the additive nonparametric components f_j are represented via regression spline bases, with associated measures of function roughness which can be expressed as quadratic forms in the basis coefficients. Given such bases, model (1) can be estimated as a GAM, but to avoid overfitting it is necessary to estimate such a model by penalised least squares in which roughness measures are used to control overfit. We will therefore start the estimation procedure with an initial additive spline smoothing estimate defined as:

$$(\tilde{\beta}, \tilde{f}_1, \ldots, \tilde{f}_q) = \underset{\beta \in \mathbb{R}^p, f_j \in \mathscr{W}_2^m; j=1,\ldots,q}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \mathbf{x}_i^T \beta - \sum_{j=1}^{q} f_j(T_{ij}) \right)^2 + \lambda \sum_{j=1}^{q} J(f_j) \right\}, \quad (9)$$

where the f_j are represented via regression spline bases and the penalties $J(f_j)$ are quadratic in the spline coefficients, that are good approximations of the Sobolev norm $||f_j||_m^2$ and measure therefore the roughness of the smooth functions. The parameter λ is a smoothing parameter that controls the trade-off between fit and smoothness, and can be selected by minimisation of the generalised cross validation (GCV) score, the generalised Akaike's information criterion (AIC), and restricted maximum likelihood (REML) estimation, to name a few. The computational methods of Wood (2006) implemented in the R-package `mgcv` are available to estimate $\beta, f_1, \ldots, f_q$ minimising (9). It can be shown (see Du et al., 2012, Proposition 2.1) that under appropriate conditions on the covariate designs and the distribution of the noise and the covariates, with a dimension of the parametric component possibly diverging but remaining smaller that n while keeping the dimension q of nonparametric components fixed, if $\lambda \sim n^{-2m/(2m+1)}$ and $p_n = o(n^{1/2})$, then the initial solution of (9) has the following asymptotic rates :

$$\|\tilde{\beta} - \beta\| = O_{\mathbb{P}}(\sqrt{p_n/n}) \quad \text{and} \quad \|\tilde{f}_j - f_j\|_2 = O(\sqrt{p_n/n} \vee n^{-2m/(2m+1)} p_n^2),$$

for any $1 \le j \le q$. Thus $\tilde{f}_j$ is consistent together with $\tilde{\beta}$ if we further assume that $p_n = o(n^{m/(4m+2)})$.

4.1.1. Joint variable selection in nonparametric and parametric parts.

We now introduce the adaptive SpAM procedure to estimate the functions f_j given the coefficient vector β. Of course one could also use for this any of the procedures studied in Amato et al. (2016) for additive model selection but we prefer here to stay close to the work by Du et al. (2012). The additive component functions f_j are estimated as the minimisers of

$$\ell_\beta(f_1, \ldots, f_q)) = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \mathbf{x}_i^T \beta - \sum_{j=1}^{q} f_j(T_{ij}) \right)^2 + \lambda \sum_{j=1}^{q} w_j \|f_j\|_n, \tag{10}$$

for $f_j \in \mathscr{W}_2^m$ and where w_j's are weights chosen in a data-adaptive way. When an initial estimator $\tilde{f}_j$ is available, a choice of w_j could be $w_j = \|\tilde{f}_j\|_n^{-s}$ for some $s > 0$. Note that when $w_j = 1, \forall j$ and $\beta = 0$, (10) reduces to the SpAM model proposed in Ravikumar et al. (2009). It is easy to show, following Theorem 1 in Ravikumar et al. (2009) that the minimiser f_j in (10) is given by

$$\hat{f}_j = \left[1 - \frac{\lambda w_j}{\|P_j\|_n} \right]_+ P_j,$$

where $[\cdot]_+$ denotes the positive part and P_j is the projection of the residual $\mathbf{R}_j = \mathbf{Y} - \sum_{k \neq j} f_k(\mathbf{T}_k) - \mathbf{X}^T \beta$ on the space generated by the jth regression spline basis. A backfitting algorithm can then be used. Given estimates $\hat{f}_j$ of f_j, the parameter $\hat{\beta}$ is estimated by minimising the penalised profile least squares

$$\ell_{\hat{f}_1, \ldots, \hat{f}_q}(\beta) = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \mathbf{x}_i^T \beta - \sum_{j=1}^{q} \hat{f}_j(T_{ij}) \right)^2 + \sum_{k=1}^{p} p_{\lambda_k}(|\beta_k|), \tag{11}$$

where $p_{\lambda_k}(|\cdot|)$ is the SCAD penalty.

Thus the complete algorithm for the semiparametric variable selection and estimation procedure for PLAM is as follows.

Step 1 Start with the initial estimate $\hat{\beta}^{(0)} = \tilde{\beta}$.

Step 2 Let $\hat{\beta}^{(k-1)}$ be the estimate of β before the kth iteration. Plug $\hat{\beta}^{(k-1)}$ into (10) and solve for the nonparametric components by solving the adaptive SpAM problem of minimising (10). Let $\hat{f}_j^{(k)}$, $j = 1, \ldots, q$ be the estimates thus obtained.

Step 3 Plug $\hat{f}_j^{(k)}$, $j = 1, \ldots, q$ into (11) and solve the corresponding minimisation problem. Let $\hat{\beta}^{(k)}$ be the estimate thus obtained.

Step 4 Replace $\hat{\beta}^{(k-1)}$ in Step 2 by $\hat{\beta}^{(k)}$ and repeat Steps 2 and 3 until convergence to obtain the final estimates.

In practice to select the regularisation parameters $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_p)$ and λ in steps 3 and 2, one may use 5-fold cross validation as in Zou and Li (2008) and GCV.

Remark 4 As noted in many other settings, penalised variable selection procedures can over-shrink component estimates. To account for this, a common practice is, once only the selected components are retained, to perform a refit using a standard non-penalised estimation procedure after the variable selection step. Such refitting indeed improves the final estimation performance.

To end this section, we present some simulations used to illustrate the results given above using relevant R packages (R, 2015). We generated data from model (1), with $n = 200$, $b = 0$, $q = 8$ and β a p-dimensional vector with $p = 20$. For the parametric part we have used the parameter $\beta = (1, 0.8, 1.4, 0.6, 1.2, 0.9, 1.1, 1.2, \mathbf{0}_{12}^T)^T$, where $\mathbf{0}_{12}^T$ is the vector of zeros of length 12. The design matrix $\mathbf{X}$ is a $n \times p$ matrix whose columns are i.i.d. realisations of standard normal distributed variables, For the nonparametric part, the true functions were set

$$f_1(t) = -2\sin(2\pi t), \quad f_3(t) = 12t^2 - 11t + 1.5, \quad f_4(t) = 9\exp(-(t-0.3)^2) - 8.03,$$

all other 5 nonparametric components being equal to zero. Note that all the f_j's integrate to 0 on $[0,1]$ for identifiability reason. The $\mathbf{T}_j$'s were generated independently from the uniform distribution on $[0,1]$ and the random errors ε_i were generated from a standard normal distribution, producing a SNR of 2. For the above settings, we simulated $M = 100$ replications.

Estimation and model selection performance of the procedure was evaluated by computing the following indicators:

- Root Mean squared Error (RMSE) as

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{X}_i^T(\hat{\beta}-\beta) + \sum_{j=1}^{q}\left(\hat{f}_j(T_{ij}) - f(T_{ij})\right)^2\right]}.$$

- Number of selected nonparametric components (NNPS)

$$\text{NPNS} := |\hat{NPS}|,\ \hat{NPS} = \{j : \hat{f}_j \neq 0\}.$$

- Number of selected parametric components (NS). It is aimed at evaluating capability of the method in preserving sparsity of the linear part:

$$\text{NS} := |\hat{NS}|,\ \hat{NS} = \{k : \hat{\beta}_k \neq 0\}.$$

Table 1: Average values and standard deviations based on $m = 100$ simulations with different noise realisations.

	NS	FP	FN	NPNS	FPNP	FNNP
`Stats`	8.04 (0.2)	0.04 (0.2)	0 (0)	3.01 (0.1)	0.01 (0.1)	0 (0)

- False positives (FP) of parametric components defined as

$$\text{FP} := |\widehat{FP}|,\ \widehat{FP} := \{j : \hat{\beta}_j \neq 0 \text{ and } \beta_j = 0\}.$$

- False positives (FPNP) of nonparametric components defined as

$$\text{FPNP} := |\widehat{FPNP}|,\ \widehat{FPNP} := \{j : \hat{f}_j \neq 0 \text{ and } f_j = 0\}.$$

- False negatives (FN) of parametric components defined as

$$\text{FN} := |\widehat{FN}|,\ \widehat{FN} := \{j : \hat{\beta}_j = 0 \text{ and } \beta_j \neq 0\}.$$

- False negatives (FNNP) of nonparametric components defined as

$$\text{FNNP} := |\widehat{FNNP}|,\ \widehat{FNNP} := \{j : \hat{f}_j = 0 \text{ and } f_j \neq 0\}.$$

Figure 5 plots the retained estimated nonparametric components for a typical simulation run and also a boxplot of the RMSE over the M fits. As one can see, the estimation performances of the procedure are pretty good.

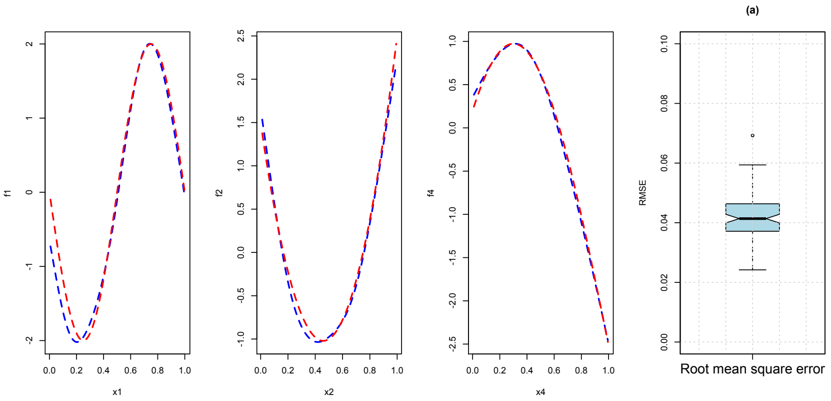

Figure 5: Estimates of selected nonparametric components for the partially linearly additive model at the end of a refitting procedure by a standard non-penalised estimation procedure after the variable selection step: from left to right functions $-2\sin(2\pi t)$, $12t^2 - 11t + 1.5$ and $9\exp(-(t-0.3)^2) - 8.03$. True functions: red lines; estimates after refitting: blue lines. Rightmost panel (a) displays a boxplot of the root mean squared estimation error for Y over the 100 simulation runs.

The other results are summarised in Table 1. From Table 1 we can see that the model selection performances of the procedure are good with most of the time correctly specified linear and nonlinear effects.

5. PLAM with Hybrid Splines and Wavelets

In the previous section the additive components were assumed to posses a similar degree of regularity: they were either belonging to Sobolev paces or to Besov spaces. The methods we reviewed could handle a large variety of shapes for the nonparametric components of PLAMs, either regular (based on penalised regression splines) or non-regular (based on wavelet decompositions) as illustrated in the examples used. A limitation of the nonparametric regression procedures that we reviewed was the use of a single class of basis functions, either splines of a given regularity for smooth PLAMs or wavelets for PLAMS with irregular components. However, a loss of efficiency occurs when the additive part is composed of both smooth functions and functions with much less regularity, a class of models that we are going to call hybrid PLAM models. Our main contribution in this section is to combine the methods used in the previous sections and obtain an estimator that can deal with and overcome the difficulties given from the non-linear part of such hybrid partial additive models. The basic idea is to process such models by exploring the advantage of using a hybrid fitting for the additive part combining regression splines and wavelets together, and use advanced model selection methods for solving the estimation and variable selection problems.

The hybrid partial linear additive models that we are going to study in this section are of the form (2). Hereafter we consider a random sample $\{Y_i, \mathbf{X}_i, \mathbf{T}_i^{(1)}, \mathbf{T}_i^{(2)}\}_{i=1,\ldots,n}$, related through the hybrid partially linear additive model (HPLAM)

$$Y_i = \mathbf{X}_i^T \beta + \sum_{j=1}^{q_s} f_j^{(1)}(T_{ij}^{(1)}) + \sum_{j=1}^{q_w} f_j^{(2)}(T_{ij}^{(2)}) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{12}$$

where $\mathbf{X}_i = (X_i^1, \ldots, X_i^p)^T$ is a p-dimensional covariate vector representing the linear regression component, β is the $p \times 1$ vector of corresponding regression coefficients, $f_j^{(1)}$s are unknown smooth functions of $T_{ij}^{(1)}$ where $\mathbf{T}_i^{(1)} = (T_{i1}^{(1)}, \ldots, T_{iq_s}^{(1)})^T$ is a q_s-dimensional nonlinear covariate vector with values in $[0,1]^{q_s}$, $f_j^{(2)}$s are unknown non-smooth functions of $T_{ij}^{(2)}$ where $\mathbf{T}_i^{(2)} = (T_{i1}^{(2)}, \ldots, T_{iq_w}^{(2)})^T$ is a q_w-dimensional nonlinear covariate vector with values in $[0,1]^{q_w}$ and the errors ε_i form a sequence of i.i.d. Gaussian random variables with mean 0 and variance σ^2 independent of the predictor variables $\mathbf{X}_i$ and $\mathbf{T}_i^{(k)}$, $k = 1, 2$. Gaussian errors is quite a strong assumption, but is not unusual. Regardless, this condition can be easily relaxed to sub-Gaussian errors. Again, for identifiability reasons we assume that $\mathbb{E}(f_j^{(k)}(\mathbf{T}_j^{(k)})) = 0$, $k = 1, 2$, for all j.

To approximate each smooth nonparametric additive component $f_j^{(1)}$, $j = 1, \ldots, q_s$, we use its expansion on O'Sullivan splines basis functions $\{B_\ell^{(j)}\}_{\ell \in \mathbb{N}}$:

$$f_j^{(1)}(t) \approx \sum_{\ell=1}^{m^{(j)}} \alpha_\ell^{(j)} B_\ell^{(j)}(t) \qquad \text{for} \quad j = 1, \ldots, q_s,$$

where $m^{(j)}$ is an appropriate truncation index that is allowed to increase to infinity with n. Throughout this section we assume that the $B_\ell^{(j)}$ are in canonical form (see e.g. Wand and Omerod, 2008, Section 4). Under reasonable smoothness assumptions, the $f_j^{(1)}(t)$ can be well approximated by the above expansions and their estimation is therefore equivalent in estimating the coefficient vector $\alpha^{(j)} = (\alpha_1^{(j)}, \ldots, \alpha_{m^{(j)}}^{(j)})^T$. When the additive components are not smooth we will approximate them using instead wavelet bases. Hence, to approximate each non smooth nonparametric additive component $f_j^{(2)}$, $j = 1,\dots,q_w$, we use its expansion on wavelet basis functions $\{W_\ell^{(j)}\}_\ell$:

$$f_j^{(2)}(t) \approx \sum_{\ell=1}^{K^{(j)}} \gamma_\ell^{(j)} W_\ell^{(j)}(t) \qquad \text{for} \quad j = 1,\dots,q_w,$$

where $K^{(j)}$ is again an appropriate truncation index that is allowed to increase to infinity with n. Under reasonable non smoothness assumptions, the $f_j^{(2)}(t)$ can be well approximated by the above expansions and their estimation is therefore equivalent in estimating the wavelet coefficient vector $\gamma^{(j)} = (\gamma_1^{(j)},\dots,\gamma_{K^{(j)}}^{(j)})^T$.

Using the O'Sullivan basis construction described in Wand and Omerod (2008) it is easy to compute, for each $j = 1,\dots,q_s$, the corresponding regression $n \times m^{(j)}$ matrices of the O'Sullivan basis functions evaluated at the observations of the corresponding predictors, i.e.

$$\mathbf{B}^{(j)} = \begin{bmatrix} B_1^{(j)}(T_{1j}^{(1)}) & \dots & B_{m^{(j)}}^{(j)}(T_{1j}^{(1)}) \\ \vdots & \ddots & \vdots \\ B_1^{(j)}(T_{nj}^{(1)}) & \dots & B_{m^{(j)}}^{(j)}(T_{nj}^{(1)}) \end{bmatrix}.$$

Similarly to the spline case, as alluded to in Antoniadis and Fan (2001) and implemented in Wand and Omerod (2011) we can also define the design matrices containing wavelet basis functions evaluated at the predictors (for the sake of completeness we give in the appendix a brief description of such a construction). Again, using this construction, for each $j = 1,\dots,q_w$, we will denote by $\mathbf{W}^{(j)}$ the corresponding wavelet regression $n \times K^{(j)}$ matrices of the wavelet basis functions evaluated at the observations of the corresponding predictors, i.e.

$$\mathbf{W}^{(j)} = \begin{bmatrix} W_1^{(j)}(T_{1j}^{(2)}) & \dots & W_{K^{(j)}}^{(j)}(T_{1j}^{(2)}) \\ \vdots & \ddots & \vdots \\ W_1^{(j)}(T_{nj}^{(2)}) & \dots & W_{K^{(j)}}^{(j)}(T_{nj}^{(2)}) \end{bmatrix}.$$

Adopt again a vector-matrix form of the HPLAM model, given by (12) to get:

$$\mathbf{Y} \approx \mathbf{U}\beta + \sum_{j=1}^{q_s} \mathbf{B}^{(j)}\alpha^{(j)} + \sum_{j=1}^{q_w} \mathbf{W}^{(j)}\gamma^{(j)} + \varepsilon, \tag{13}$$

for $\mathbf{Y} = (Y_1,\dots,Y_n)^T$, $\mathbf{U}^T = [\mathbf{X}_1 \dots \mathbf{X}_n]$ and $\varepsilon = (\varepsilon_1,\dots,\varepsilon_n)^T$. To simplify notation we will suppose that for all $j = 1,\dots,q_s$ the truncation index is the same, i.e. $m^{(1)} = m^{(2)} = \dots = m^{(q_s)} = m$ and also $K^{(1)} = K^{(2)} = \dots = K^{(q_w)} = K$ for the univariate approximations of the non-regular additive components. Each jth function in the smooth nonlinear part of the approximate hybrid partial linear model above is characterised by the m-dimensional coefficient vector $\alpha^{(j)}$ while each jth function in the non-smooth nonlinear part is characterised by K-dimensional coefficient vector $\gamma^{(j)}$. Thus the component selection and estimation in HPLAM models using the approximation (13) may be viewed as a functional version of estimation and grouped variable selection, since each nonparametric component can be expressed as a linear combination of a set of basis functions whose coefficients must be either killed or selected simultaneously and then estimated.

Let $\mathbf{B}$ be the $n \times (mq_s)$ matrix obtained by stacking block-wise the matrices $\mathbf{B}^{(j)}$, $j = 1, \ldots, q_s$:

$$\mathbf{B} = \left[\mathbf{B}^{(1)} \ldots \mathbf{B}^{(q_s)}\right]$$

and let α be the long (mq_s)-dimensional column vector $\alpha = (\alpha^{(1)^T}, \ldots, \alpha^{(q_s)^T})^T$ of spline coefficients. Use similar notation for the wavelet design matrices to define the $\mathbf{W}$ matrix of size $n \times (Kq_w)$:

$$\mathbf{W} = \left[\mathbf{W}^{(1)} \ldots \mathbf{W}^{(q_w)}\right]$$

and the corresponding (Kq_w)-dimensional wavelet coefficient vector $\gamma = (\gamma^{(1)^T}, \ldots, \gamma^{(q_w)^T})^T$. With such notation, (13) becomes

$$\mathbf{Y} = \mathbf{U}\beta + \mathbf{B}\alpha + \mathbf{W}\gamma + \varepsilon, \tag{14}$$

which is a high-dimensional linear model and the estimation task of the various components in a HPLAM model is equivalent to estimating the vectors of unknown coefficients β, α and γ, respectively. For the variable selection task, it is important to recall that the inclusion or not of a covariate affecting the nonlinear part of the mean is to be brought back to a vector of coefficients, instead of to a real-valued parameter. The nonzero additive components can therefore be selected and estimated using a group penalised method following an approach similar in spirit to those in, for example, Antoniadis, Gijbels and Verhasselt (2012) and Amato et al. (2016). To do so, let $\mathbb{G} = [\mathbf{BW}]$ be the $n \times (mq_s + Kq_w)$ matrix obtained by stacking block-wise $\mathbf{B}$ and $\mathbf{W}$ and let $\theta = (\alpha^T, \gamma^T)^T$ be the vector of unknown coefficients. The linear regression model (14) becomes

$$\mathbf{Y} = \mathbf{U}\beta + \mathbb{G}\theta + \varepsilon.$$

In a grouped linear regression setting, the $mq_s + Kq_w$ nonlinear coefficients are partitioned into $q = q_s + q_w$ groups of variables, $g_1, \ldots, g_{q_s}, g_{q_s+1}, \ldots, g_q \subseteq \{1, \ldots, mq_s + Kq_w\}$, with group sizes m for the first q_s and K for the remaining q_w. Further assuming a sparsity structure in that only a small portion of θ_{g_k}'s are nonzero, where $\theta_{g_k} \in \mathbb{R}^m$ or $\mathbb{R}^K$ is the sub-vector of θ corresponding to the kth group of features, estimation and selection that utilises the feature grouping (so that an entire group of features is selected simultaneously) can be achieved by minimising with respect to β and θ the penalised least squares criterion (group LASSO):

$$\underset{\beta,\theta}{\arg\min} \left\{ \frac{1}{2n} \|\mathbf{Y} - \mathbf{U}\beta - \mathbb{G}\theta\|_2^2 + \lambda \sum_{k=1}^{q} \sqrt{M_{g_k}} \|\theta_{g_k}\|_2 \right\}, \tag{15}$$

where the ℓ_2 norm penalty on θ_{g_k} has been rescaled relative to the size M_{g_k} of the group. Such a penalty promotes sparsity at the group level; for large λ, few groups will be selected but within any selected group, the coefficients will be dense (all nonzero). For any given θ, the $\hat{\beta}$ that minimises (15) is given by

$$\hat{\beta}(\theta) = \left(\mathbf{U}^T\mathbf{U}\right)^{-1} \mathbf{U}^T (\mathbf{Y} - \mathbb{G}\theta).$$

Let $\mathbf{H} = \mathbf{U}\left(\mathbf{U}^T\mathbf{U}\right)^{-1}\mathbf{U}^T$ be the projection matrix onto the space spanned by the columns of $\mathbf{U}$. The penalised profile criterion for θ is then

$$\frac{1}{2n}\|(\mathbf{I} - \mathbf{H})(\mathbf{Y} - \mathbb{G}\theta)\|^2 + \lambda \sum_{k=1}^{q} \sqrt{M_{g_k}} \|\theta_{g_k}\|_2.$$

More generally one may use more efficient penalties in terms of estimation and model selection consistency as the ones discussed in Antoniadis (2007) which leads to objective function of the following form:

$$\frac{1}{2n}\|(\mathbf{I}-\mathbf{H})(\mathbf{Y}-\mathbb{G}\theta)\|^2+\lambda\sum_{k=1}^{q}\vec{P}_{\lambda_k}(\theta_{g_k}), \tag{16}$$

where $\vec{P}(\cdot)$ is now a multivariate penalty function, since all coefficient vectors $\theta_{g_1},\ldots,\theta_{g_q}$ are vectors. Again the penalties are scaled with the square root of group size to penalise larger groups as in Meier, van de Geer and Bühlmann (2009).

Recall that an m-variate penalty function $\vec{P}_\lambda(\cdot)$ is obtained from a univariate penalty function $P_\lambda(\cdot)$ as those reviewed in Antoniadis and Fan (2001) and Antoniadis (2007) as follows. For any vector $\mathbf{a}$ of dimension m, we define the m-variate penalty function, based on $P_\lambda(\cdot)$, as

$$\vec{P}_\lambda(\mathbf{a})=\begin{cases}0 & \text{if} \quad \mathbf{a}=\mathbf{0}\\ \frac{\mathbf{a}}{\|\mathbf{a}\|_2}P_\lambda(\|\mathbf{a}\|_2) & \text{if} \quad \mathbf{a}\neq\mathbf{0}\,,\end{cases}$$

where $\|\mathbf{a}\|_2$ denotes the ℓ_2-norm of the vector $\mathbf{a}$.

Directly optimising (16) can be tricky for a given penalty function, especially when the penalty is non convex. To tackle the optimisation, it is more convenient to use a thresholding viewpoint with a thresholding function corresponding to the selected penalty (see Antoniadis (2007) for a survey on the one-to-one correspondence between threshold functions and penalty functions. See also She (2012)). Consider first the scalar case (each group has only one component, i.e. card(g_k)=1) and define a thresholding rule as an odd monotone unbounded shrinkage rule for t, at any λ, as follows: *A threshold function is a real valued function* $\Theta(t;\lambda)$ *defined for* $-\infty<t<\infty$ *and* $0\leq\lambda<\infty$ *such that*

1. $\Theta(-t;\lambda)=-\Theta(t;\lambda)$,
2. $\Theta(t;\lambda)\leq\Theta(t';\lambda)$ *for* $0\leq t\leq t'$
3. $\lim_{t\to\infty}\Theta(t;\lambda)=\infty$, *and*
4. $0\leq\Theta(t;\lambda)\leq t$ *for* $0\leq t<\infty$.

A *multivariate* version of Θ, denoted by $\vec{\Theta}$, is defined for any vector $\mathbf{a}\in\mathbb{R}^p$:

$$\vec{\Theta}(\mathbf{a};\lambda)=\mathbf{a}^\circ\Theta(\|\mathbf{a}\|_2;\lambda),$$

where $\mathbf{a}^\circ=\mathbf{a}/\|\mathbf{a}\|_2$, if $\mathbf{a}\neq 0$ and 0, if $\mathbf{a}=0$. Note that $\vec{\Theta}$ is still a shrinkage rule because $\|\Theta(\mathbf{a};\lambda)\|_2=\Theta(\|\mathbf{a}\|_2;\lambda)\leq\|\mathbf{a}\|_2$. The connection between thresholding rules and penalties is given now by the following result (Proposition 3.2 in Antoniadis, 2007):
Given an arbitrary thresholding rule Θ*, let* P *be any function satisfying* $P(\theta;\lambda)-P(0;\lambda)=P_\Theta(\theta;\lambda)+\nu(\theta;\lambda)$ *where* $P_\Theta(\theta;\lambda)\triangleq\int_0^{|\theta|}(\sup\{s:\Theta(s;\lambda)\leq u\}-u)\mathrm{d}u$*,* $\nu(\theta;\lambda)$ *is nonnegative and* $\nu(\Theta(t;\lambda))=0$ *for all* t*. Then, the minimisation problem*

$$\min_{\mathbf{a}\in\mathbb{R}^n}\frac{1}{2}\|\mathbf{y}-\mathbf{a}\|_2^2+P(\|\mathbf{a}\|_2;\lambda)\triangleq Q(\mathbf{a};\lambda)$$

has a unique optimal solution given by $\hat{\mathbf{a}} = \vec{\Theta}(\mathbf{y};\lambda)$ *for every* $\mathbf{y}$ *provided that* $\Theta(\cdot;\lambda)$ *is continuous at* $\|\mathbf{y}\|_2$.

Note that P (and P_Θ) may not be differentiable at 0 and may be non-convex. The above result shows that the solution of

$$\arg\min_{\mathbf{a}}(\|\mathbf{a}-\mathbf{y}\|^2/2)+P_\lambda(s\|\mathbf{a}\|_2),$$

extends the definition of the proximal operator in the scalar case to vector-valued arguments as in Bredies, Lorenz and Reiterer (2015) and can therefore also be derived as a corollary from results of Bredies et al. (2015).

Now for any given $k_0 \in \{1,\dots,q\}$, (16) can be decomposed as follows:

$$\begin{aligned}
&\frac{1}{2n}\|(\mathbf{I}-\mathbf{H})\mathbf{Y}-(\mathbf{I}-\mathbf{H})\mathbb{G}\theta\|_2^2+\sum_{j=1}^{q}\sqrt{M_{g_k}}\vec{P}_{\lambda_k}(\theta_{g_k})\\
&=\frac{1}{2n}\left\|(\mathbf{I}-\mathbf{H})\mathbf{Y}-(\mathbf{I}-\mathbf{H})\mathbb{G}_{k_0}\theta_{g_{k_0}}-\sum_{\substack{k=1\\k\neq k_0}}^{q}(\mathbf{I}-\mathbf{H})\mathbb{G}_k\theta_{g_k}\right\|_2^2+\sqrt{M_{g_{k_0}}}\vec{P}_{\lambda_{k_0}}(\theta_{g_{k_0}})\\
&\quad+\sum_{\substack{k=1\\k\neq k_0}}^{q}\sqrt{M_{g_k}}\vec{P}_{\lambda_k}(\theta_{g_k}).
\end{aligned}$$

Minimisation of (16) can therefore be done by a group-wise descent iterative procedure that cycles through each group of coefficients (see Daubechies, Defrise and Mol, 2004; She, 2012; Bredies et al., 2015). More precisely, minimising (16) with respect to $\theta_{g_{k_0}}$, leads to an iterative procedure. Indeed, let $\vec{\Theta}_\lambda(\cdot) \equiv \vec{\Theta}(\cdot;\lambda)$ be the *thresholding function* that corresponds uniquely to the penalty function $\vec{P}_\lambda(\cdot)$ (see above). To avoid the influence of the ambiguity in defining some threshold functions, we always assume that the quantity to be thresholded does not correspond to any discontinuity of $\vec{\Theta}$. This assumption is mild because a practical thresholding rule usually has at most finitely many discontinuity points and such discontinuities rarely occur in any real application. Since only one group at a time is being updated, we may restrict our attention to the subproblem of finding θ_{g_k} to minimise the objective function. Denote by $\theta_{g_k}^{(s)}$ (respectively $\theta^{(s)}$) the current value of the vector θ_{g_k} (respectively θ) at iteration step s. The updated vector $\theta_{g_k}^{(s+1)}$ is then obtained via the following updating equation

$$\theta_{g_{k_0}}^{(s+1)}=\vec{\Theta}\left(\theta_{g_{k_0}}^{(s)}+(\mathbf{I}-\mathbf{H})\mathbb{G}_{k_0}\mathbf{r}^s;n\sqrt{M_{g_{k_0}}}\lambda_{k_0}\right), \tag{17}$$

where $\mathbf{r}^s=(\mathbf{I}-\mathbf{H})(\mathbf{Y}-\mathbb{G}\theta^{(s)})$. It can then be shown, using Theorem 2.1 of She (2012), that, provided that the spectral norm of the design matrix $\mathbb{G}$ is not large, and whatever the starting value of θ is, the iterated thresholding estimates defined in (17) minimise (16). The condition on boundedness of the spectral norm of $\mathbb{G}$ is easily obtained by rescaling the vector of coefficients θ and the penalty parameter λ.

For each group g_{k_0} the thresholding functions in expression (17) involve penalty parameters $\lambda_{jg_{k_0}}$ that control the amount of regularisation. Fold cross-validation and generalised cross-validation procedures are popular methods for choosing these tuning parameters, but they are rather complicated and computationally intensive. In the scalar case we have seen in a previous section that the square-root LASSO overcomes this problem. In the same spirit we have used a grouped version of the square-root LASSO, introduced and studied by Bunea, Lederer and She (2014) and minimise

$$\frac{1}{2n}\|(\mathbf{I}-\mathbf{H})(\mathbf{Y}-\mathbb{G}\theta)\|_2+\mu\sum_{k=1}^{q}\sqrt{M_{g_k}}\|\theta_{g_k}\|_2.$$

Under appropriate conditions (compatibility condition on the design matrix $\mathbb{G}$, a finite sparsity index, and bounded norms for the smooth and irregular additive components) one may choose $\mu = 2.2K_0 n\Phi^{-1}(1-0.05/(2(n+q)))$ where Φ is the inverse cumulative function of a standard Gaussian distribution and where $K_0 = \|\mathbb{G}\|_{\text{Frob}}/\sqrt{2}$, $\|\cdot\|_{\text{Frob}}$ denoting the Frobenius norm, is the scaling constant used to make the iterations converge. Moreover, the estimator of θ may be still obtained by iterative multivariate thresholding (scale $\mathbf{r}/K_0$ and $\mathbb{G}/K_0$)

$$\theta_{g_k}^{(s+1)} = \vec{\Theta}_{soft}\left(\theta_{g_k}^{(s)}+(\mathbf{I}-\mathbf{H})\mathbb{G}_k\mathbf{r}^s;\mu\sqrt{M_{g_k}}\|((\mathbf{I}-\mathbf{H})\mathbf{r}^s\right).$$

Using the results of Bunea et al. (2014), assuming that m and K grow to infinity at appropriate rates in such a way that a compatibility condition holds for the design matrix $\mathbb{G}$ rescaled by K_0, assuming a finite sparsity index less than $n/\log q$ and a bounded entropy on the class of the smooth and irregular additive components, the variable selection procedure can effectively identify the significant non-additive components, and produce estimates that are consistent. However, using a square-root group LASSO penalisation in order to exploit Bunea's results, may sometimes over-shrink the estimated nonlinear components. In a second step, using only the selected components by the square-root group LASSO, we perform a refit using an estimation procedure developed in Chesneau, Fadili and Maillot (2015), where estimation in a d-dimensional nonparametric additive regression model with dependent observations is considered using marginal integration. We then adopt a backfitting like step to get a variable selection penalised estimation of the linear part parameter β by minimising

$$\|\mathbf{Y}-\mathbb{G}\hat{\theta}-\mathbf{U}\beta\|^2+\sum_{j=1}^{p}P_\lambda(|\beta_j|).$$

However, while the simulation results are encouraging, applying asymptotic results for such an estimation and variable selection of the parameters of the linear part is interesting, but not an easy task since the residuals $\mathbf{Y}-\mathbb{G}\hat{\theta}$ are highly correlated and we hope to address it in a future work.

Regarding the efficiency of the proposed algorithm the most computationally intensive steps for group thresholding in our algorithm are the calculation of the products $(\mathbf{I}-\mathbf{H})\mathbb{G}_k\mathbf{r}^s$, each of which requires $O(n\max m,K)$ operations. Thus, one full pass over all the groups requires $O(nq)$ operations. The fact that this approach scales linearly in q allows it to be efficiently applied to high-dimensional problems with q large, but still such that $q \ll n$. Of course, the entire time required to fit the model depends on the number of iterations, which in turn depends on the data and on the regularisation parameters.

Remark 5 We could have used other types of optimisation algorithms for non convex group penalised methods. The group coordinate descent (GCD) algorithm has been used in Yuan and Lin (2006) for the group LASSO and Wei (2012) for the group MCP, assuming that the design matrix of each group is orthogonal. One could orthonormalise the groups design matrices $\mathbf{B}$ and $\mathbf{W}$ prior to penalising the ℓ_2 norms of the group coefficients and apply such GCD algorithms. However, Simon and Tibshirani (2012) pointed out that the solution obtained with such an orthogonality transformation, while easy to compute, will not be a solution of the original problem. Moreover, in practical applications the selection of good penalty λ might be very challenging. For example it has been reported that in high dimensional settings the popular cross-validation typically leads to detection of a large number of false regressors (see e.g. Bogdan, van den Berg, Su and Candès (2014)). The iterative thresholding methods we have used allow to use no prior orthogonalisation and no matter how the predictors are grouped always guarantees the convergence to a local minimum, provided that K_0 is appropriately large.

To address the shortcomings of LASSO, a relatively new convex optimisation procedure named Sorted L-One Penalised Estimation (SLOPE) developed recently by Bogdan et al. (2014) allows for adaptive selection of regressors under sparse high dimensional designs. The idea of SLOPE using false discovery ratio (FDR) to deal with the situation when one aims at selecting whole groups of explanatory variables instead of single regressors has been recently extended by Brzyski, Su and Bogdan (2015) who formulate the respective convex optimisation problem, gSLOPE, and propose an efficient algorithm for its solution. Moreover, these authors prove that the resulting procedure adapts to unknown sparsity and is asymptotically minimax with respect to the estimation of the proportions of variance of the response variable explained by regressors from different groups. They also provide a method for the choice of the regularising sequence when variables in different groups are not orthogonal but statistically independent and illustrate its good properties with computer simulations. In the same spirit one could also use the group knockoff filter, developed recently by Dai and Foygel Barber (2016), which is a method for false discovery rate control in a linear regression setting where the features are grouped, with nice theoretical guarantees. We have used the gSLOPE implemented in the `grpSLOPE` R-package in illustrating the numerical results that follow, but an extensive study of these procedures in the PLAM context are out of the scope of the present paper and we hope to address them in a near future.

To end this section we present now a small simulation study with a data setting involving a linear part with sparse coefficients and multiple covariates influencing in a non-linear additive way the mean in a hybrid partial linear additive model with normal distributed response data. We consider again the case where many of the covariates are strictly non-informative, and therefore appropriate selection of the model's informative predictors are crucial. Performance of our algorithm was evaluated by computing the same indicators that were used at the end of Section 4.

Synthetic data was generated from model (12), with $n = 300$, centred response, $q = 8$ and β a p-dimensional vector with $p = 20$. For the parametric part we have used the parameter $\beta = (1, 0.8, 1.4, 0.6, 1.2, 0.9, 1.1, 1.2, \mathbf{0}_{12}^T)^T$. The design matrix $\mathbf{X}$ is a $n \times p$ matrix whose columns are i.i.d. realisations of standard normal distributed variables. For the nonparametric part, only 5 among

Table 2: Average values and standard deviations based on $m = 100$ simulations.

	NS	FP	FN	NPNS	FPNP	FNNP
`Stats`	4.25 (0.76)	0.02 (0.15)	3.77 (0.70)	4.69 (0.46)	0.0 (0)	0.31 (0.46)

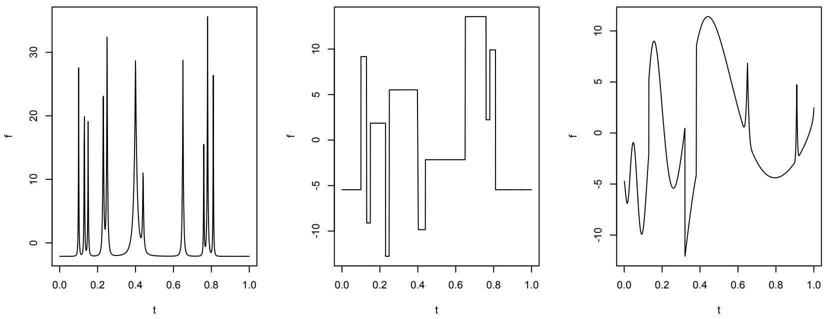

Figure 6: The true irregular functions used in the simulations (bumps, blocks and heavisine-like (see also Donoho and Johnstone, 1998)).

the q components are nonzero, numbered 1, 2, 4, 5 and 6. Among the active components the first three (bumps, blocks and heavisine-like functions (see Donoho and Johnstone (1998) and Figure 6) are non-smooth while the last two given by $f_5(x) = 3\sin(2\pi x^3)$ et $f_6(x) = 15x\exp(-x^2)$ are highly regular. The $\mathbf{T}_j$'s were generated independently from the uniform distribution on $[0,1]$ and the random errors ε_i were generated from a standard normal distribution, producing a SNR of 4. For the above setting, we simulated $M = 100$ replications. For the splines approximations of the smooth components we have used cubic P-splines with $m = 10$ equidistant knots within $[0,1]$ and for the wavelet approximations we have used Symmlets of order 6 at the finest level $K = \log_2(n) - 3$.

Figure 7 plots the retained estimated nonparametric components for a typical simulation run and also a boxplot of the RMSE over the M fits. As one can see, the estimation performances of the procedure are fairly good, but not as good as when all additive components had the same regularity. Indeed, since the nonparametric components (being of inhomogeneous smoothness) are now estimated with suboptimal rates, this accentuates their estimation bias which is reflected in the linear part and the part composed of smoother components. This explains the larger RMSE and leads to performance indicators that are less good for the parametric (linear) part. In the nonparametric part the false negatives are mainly due to the non detection of component 5 in a few runs of the simulated model.

The other results are summarised in Table 2. From Table 2 we can see that the model selection performances of the procedure are good with most of the time correctly specified linear and nonlinear effects.

As one can see from this simulation the advantage of using wavelets as opposed to splines for the non-smooth additive components is clearly visible for highly non-smooth functions where the choice of a really high resolution wavelet regression basis turns out advantageous for the estimation of the nonparametric function, it comes however at a price to pay for the mean squared error of the parametric coefficients. We noticed in simulations not reported here that when some non-smooth additive effects are misspecified as smooth the procedure in terms of estimation becomes quite poor.

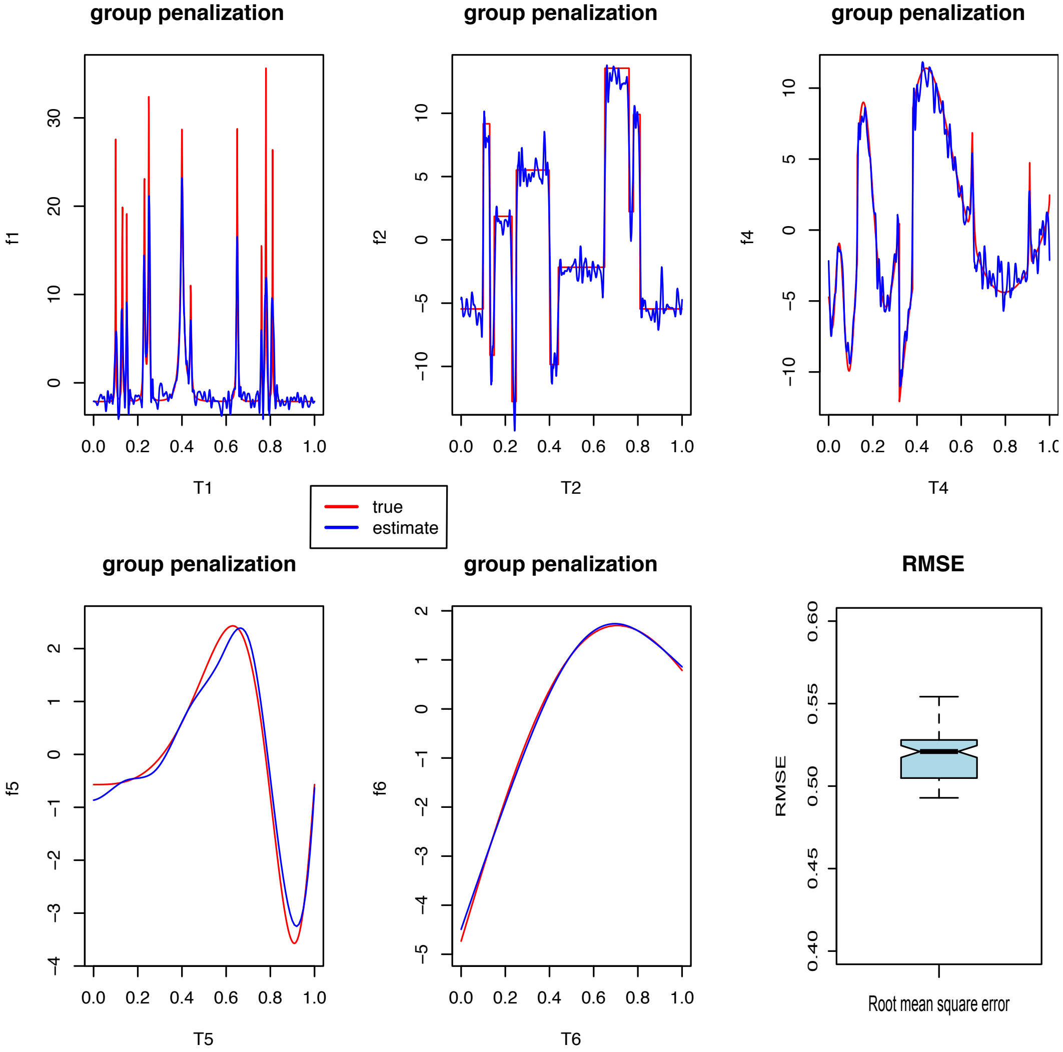

Figure 7: Estimate of some selected components from a typical sample realisation. Top line: the irregular nonparametric functions bumps, blocks and heavisine-like; bottom line from left to right: functions $3\sin(2\pi x^3)$ and $15x\exp(-x^2)$. Red lines: true functions; blue lines: estimates at the end of refitting by a standard non-penalised estimation procedure after the variable selection step. The bottom rightmost panel displays a boxplot of the root mean squared error estimation of Y over the 100 simulation runs.

6. Concluding Remarks

In this paper, taking advantage of recent developments for variable selection in parametric and nonparametric models, we reviewed several penalised least squares based estimation procedures using sparse representations for partially linear additive modelling. We have also proposed a flexible semi-parametric estimation method for partial additive linear models that have parametric components of diverging dimensions and nonparametric part with multiple additive components, both with bounded sparsity indices, composed by smooth and non-smooth functions. It retains good features for simultaneous variable selection and model estimation on both parametric and nonparametric parts. The unified approach, modelling the smooth functions by spline-based representations and the non-smooth ones by wavelet based representations, is able to estimate a broad class of functions. It is also straightforward to implement. However, for developing the results we have not allowed any dependence between the covariates and the noise, and also dependence between the responses when the data is longitudinal. All of these impose significant challenges in developing asymptotic theory and oracle properties. Moreover, in many applications, although a categorical variable is usually included in the linear part as dummies, it may be too restrictive to suppose that linear effects, smooth effects and non-smooth effects terms that affect the response are known. It is hard to determine whether the effect of a continuous predictor is linear, smooth nonlinear or non-smooth nonlinear. An idea to tackle such a point could be inspired by the linear and nonlinear discover (LAND) developed in Zhang, Cheng and Liu (2011), or the `gamsel` methodology of Chouldechova and Hastie (2015) which automatically distinguish and selects linear or nonlinear effects in an additive models with smooth components through the application of a novel penalty. Their method, extended to our hybrid setup, can surely serve as a preliminary analysis tool before the application of our method. Of course, rigorous theoretical justification of this approach requires further work.

Appendix

Wavelet basis construction

This appendix briefly reviews the wavelet basis construction for penalised wavelets mimicking the construction of O'Sullivan spline bases. The assembly of a default basis for penalised wavelets relies on classical wavelet construction over equally-spaced grids on $[0,1[$ of length N, where N is a power of 2. Choose the functions $\{b_k(\cdot); k=1,\ldots,N-1\}$, each defined on the interval $[0,1]$ such that:

$$\mathbf{W} = N^{-1/2} \left[\mathbf{1} \left\{ b_k\left(\frac{i-1}{N}\right) \right\}_{k=1,\ldots,N-1} \right]_{i=1,\ldots,N}$$

is the $N \times N$ orthogonal matrix known as a wavelet basis matrix at resolution N (see e.g. Mallat, 1999; Nason, 2010). Note that, for any fixed k, the functions $b_k(\cdot)$ do not depend on the value of N. Hence, if N is increased from 4 to 8 then the functions b_1, b_2 and b_3 remain unchanged. For an N-dimensional real vector $\mathbf{z}$, and by orthogonality of $\mathbf{W}$ we have

$$\mathbf{z} = \mathbf{W}\hat{\theta},$$

where θ is the vector of wavelet coefficients of $\mathbf{z}$ associated to the basis. A fast discrete wavelet transform algorithm ($O(N)$ operations) allows the computation of $\hat{\theta}$.

There exist a large class of compactly supported wavelet of a given regularity. They may be computed using the R package `wavethresh` (see Nason, 2010) referenced using `family="DaubExPhase"` and the regularity number which is denoted by `filter.number`. Note, however, that the Daubechies wavelet functions do not admit explicit algebraic expressions and can only be constructed via recursion over equally-spaced grids of size equal to a power of 2. Figure 8 displays four Daubechies wavelets of different smoothness.

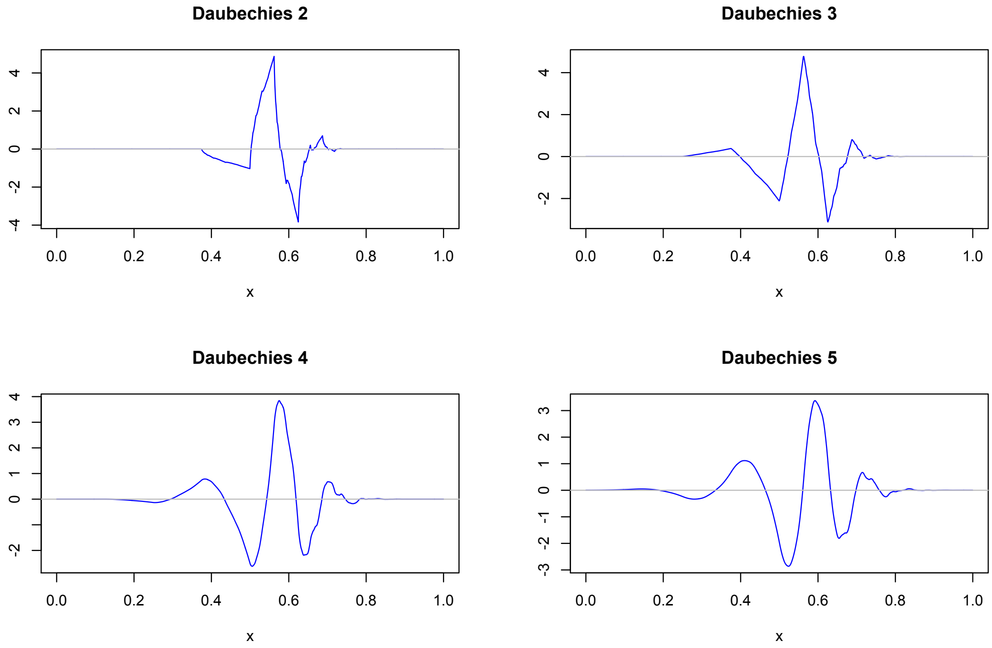

Figure 8: Some Daubechies wavelets of regularity 2, 3, 4 and 5.

The basis functions $b_k(\cdot)$ with the same amount of dilation, but different shifts, are on the same resolution level. The number of basis functions at level ℓ is $2^{\ell}-1$ for each $\ell = 1,\ldots,\log_2(N)$. The default basis definition requires that we impose the following ordering on the $b_k(\cdot)$, $1 \le k \le (N-1)$:

- $b_1(\cdot)$ is the single function on level 1.
- $b_2(\cdot)$ and $b_3(\cdot)$ are on level 2, with ordering from left to right in terms of their supports.
- and so on for levels $3,\ldots,\log_2(N)$.

Figure 9 shows the $b_k(\cdot)$ generated by a Daubechies filter of regularity 5 at resolution $N = 16$.

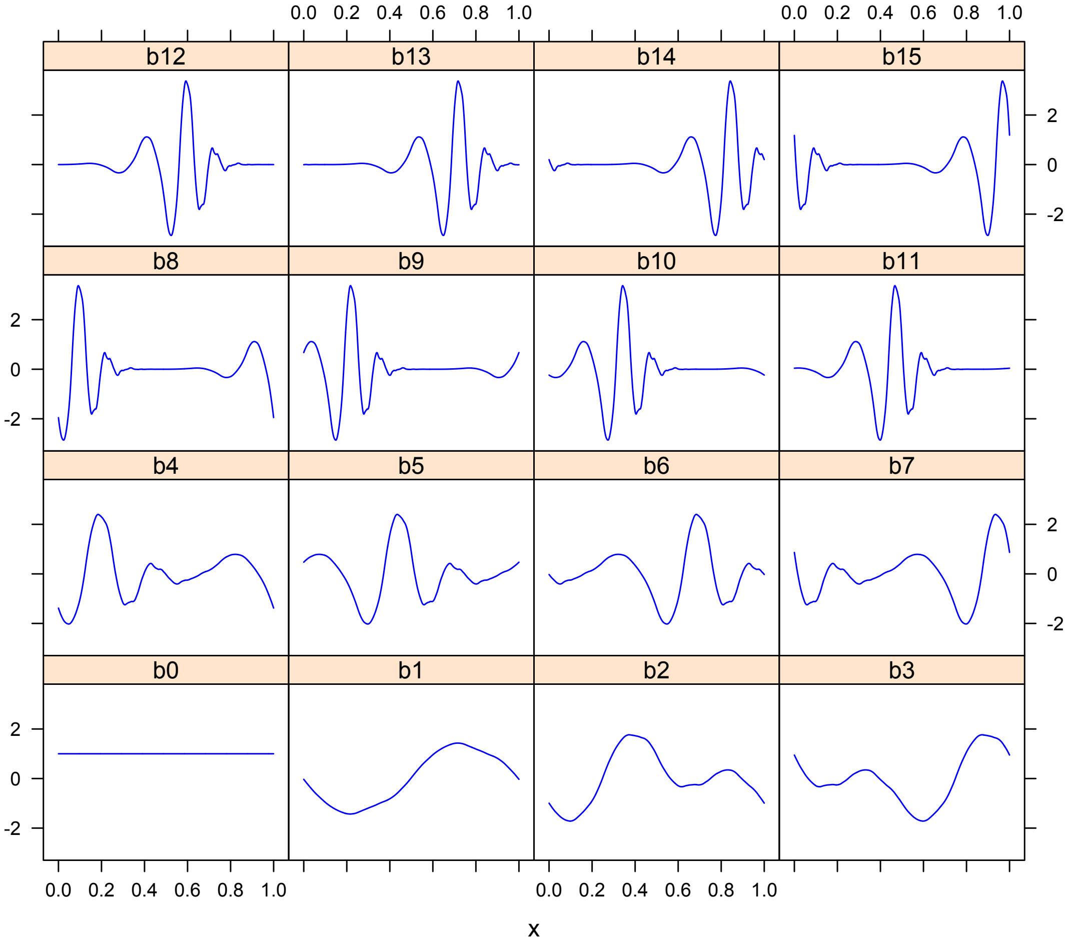

Figure 9: Daubechies wavelet basis functions of regularity 5 for $N = 16$ with the following ordering: b_0 is the mean (constant); b_1 is the single function on level 1; b_2 and b_3 are on level 2, with ordering from left to right in terms of their supports, and so on for levels 3 and 4.

To compute b_k at an arbitrary point x in $[0,1]$ we may choose as in Antoniadis and Fan (2001) a large value of N (for example $N = 16384$) and use the approximation

$$b_k(x) \simeq [1-(xN-\lfloor xN\rfloor)]b_k(\lfloor xN\rfloor/N)+(xN-\lfloor xN\rfloor)b_k((\lfloor xN\rfloor+1)/N),$$

with $b_k(1) \equiv b_k((N-1)/N)$. This allows to compute the wavelet design matrices used in this paper.

Acknowledgements

Several parts of this work were presented in an invited presentation at the 58th South African Statistical Association (SASA) Conference in Cape Town in 2016. We would like to thank Professor Francesca Little (President of SASA), Professor Sarah Radloff (Editor of SASJ) and Professor Leonard Santana (Managing Editor of SASJ) for their kind invitation to submit this paper to SASJ. We would also like to thank the reviewers for their comments. Antoniadis's work was partially supported by grants from EDF (France) and National Research Council (Italy).

References

ADAMS, R. (1975). *Sobolev Spaces*. Academic Press: New York.

AMATO, U., ANTONIADIS, A., AND DE FEIS, I. (2016). Additive model selection. *Statistical Methods & Applications*, **25** (4), 519–564.

ANTONIADIS, A. (2007). Wavelet methods in statistics: some recent developments and their applications. *Statistics Surveys*, **1**, 16–55.

ANTONIADIS, A. (2010). Comments on l_1-penalization for mixture regression models. *Test*, **19** (2), 257–258.

ANTONIADIS, A., BIGOT, J., AND SAPATINAS, T. (2001). Wavelet estimators in nonparametric regression: A comparative simulation study. *Journal of Statistical Software*, **6**.

ANTONIADIS, A. AND FAN, J. (2001). Regularization of wavelet approximations. *Journal of the American Statistical Association*, **96** (455), 939–967.

ANTONIADIS, A., GIJBELS, I., AND VERHASSELT, A. (2012). Variable selection in additive models using P-splines. *Technometrics*, **54**, 425–438.

ANTONIADIS, A. AND LAVERGNE, C. (1995). Nonparametric wavelet estimation in heteroscedastic regression models. *In* ANTONIADIS, A. AND OPPENHEIM, G. (Editors) *Wavelets and Statistics*, volume 103 of *Lecture Notes in Statistics*. Springer Verlag, Berlin, pp. 31–42.

ANTONIADIS, A. AND PHAM, D. T. (1998). Wavelet regression for random or irregular design. *Computational Statistics and Data Analysis*, **28** (4), 333–369.

BELLONI, A., CHERNOZHUKOV, V., AND WANG, L. (2011). Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, **98** (4), 791–806.

BOGDAN, M., VAN DEN BERG, E., SU, W., AND CANDÈS, E. J. (2014). Statistical estimation and testing via the ordered ℓ_1 norm. *Annals of Applied Statistics*, **9** (3), 1103–1140.

BREDIES, K., LORENZ, D. A., AND REITERER, S. B. (2015). Minimization of non-smooth, nonconvex functionals by iterative thresholding. *Journal of Optimization Theory and Applications*, **165** (1), 78–112.

BRZYSKI, D., SU, W., AND BOGDAN, M. (2015). Group SLOPE - adaptive selection of groups of predictors. arxiv:1511.09078, Faculty of Mathematics, Wroclaw University of Technology, Poland.

BUJA, A., HASTIE, T. J., AND TIBSHIRANI, R. J. (1989). Linear smoothers and additive models (with discussion). *Annals of Statistics*, **17**, 453–555.

BUNEA, F., LEDERER, J., AND SHE, Y. (2014). The group square-root lasso: Theoretical properties and fast algorithms. *IEEE Transactions on Information Theory*, **60** (2), 1313–1325.

CAI, T. AND BROWN, L. (1998). Wavelet shrinkage for non-equispaced samples. *Annals of Statistics*, **26** (5), 1783–1799.

CATTANEO, M. D., JANSSON, M., AND NEWEY, W. K. (2016). Alternative asymptotics and the partially linear model with many regressors. *Econometric Theory*, 1–25. doi:10.1017/S026646661600013X.

CHANG, X. AND QU, L. (2004). Wavelet estimation in partial linear models. *Computational Statistics and Data Analysis*, **47** (1), 31–48.

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

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

CHESNEAU, C., FADILI, J., AND MAILLOT, B. (2015). Adaptive estimation of an additive regression function from weakly dependent data. *Journal of Multivariate Analysis*, **133**, 77–94.

CHOULDECHOVA, A. AND HASTIE, T. (2015). Generalized additive model selection. arXiv preprint arXiv:1506.03850, Stanford University.

CLAESKENS, G., KRIVOBOKOVA, T., AND OPSOMER, J. (2009). Asymptotic properties of penalized spline estimators. *Biometrika*, **96** (3), 529–544.

DAI, R. AND FOYGEL BARBER, R. (2016). The knockoff filter for FDR control in group-sparse and multitask regression. *In Proceedings of the 33rd international conference on Machine Learning (ICML 2016).*

DAUBECHIES, I., DEFRISE, M., AND MOL, C. D. (2004). An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications in Pure and Applied Mathematics*, **57** (11), 1413–1457.

DING, H., CLAESKENS, G., AND JANSEN, M. (2011). Variable selection in partially linear wavelet models. *Statistical Modelling*, **11**, 409–427.

DINSE, G. E. AND LAGAKOS, S. W. (1983). Regression analysis of tumour prevalence. *Applied Statistics*, **33**, 236–248.

DONALD, S. G. AND NEWEY, W. K. (1994). Series estimation of semilinear models. *Journal of Multivariate Analysis*, **50** (1), 30–40.

DONOHO, D. L. AND JOHNSTONE, I. M. (1998). Minimax estimation via wavelet shrinkage. *Annals of Statistics*, **26** (3), 879–921.

DU, P., CHENG, G., AND LIANG, H. (2012). Semiparametric regression models with additive nonparametric components and high dimensional parametric components. *Computational Statistics and Data Analysis*, **56**, 2006–2017.

EILERS, P. H. C. AND MARX, B. D. (1996). Flexible smoothing with *b*-splines and penalties (with discussion). *Statistical Science*, **11**, 89–121.

ENGLE, R., GRANGER, C., RICE, J., AND WEISS, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association*, **81** (394), 310–320.

EUBANK, R. L., KAMBOUR, E. L., KIM, J. T., KLIPPLE, K., REESE, C. S., AND SCHIMEK, M. (1998). Estimation in partially linear models. *Computational Statistics and Data Analysis*, **29**, 27–34.

FADILI, M. AND BULLMORE, E. (2004). Penalized partially linear modelsusing sparse representation with an application to fMRI time series. *IEEE Transactions on signal processing*, **53** (9), 3436–3448.

FAN, J. AND LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**, 1348–1360.

FRIEDMAN, J. AND STUELZE, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association*, **76**, 817–823.

GANNAZ, I. (2007). Robust estimation and wavelet thresholding in partially linear models. *Statistics and Computing*, **17**, 293–310.

GASSER, T., SROKA, L., AND JENNEN-STEINMETZ, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika*, **73**, 625–633.

GREEN, P. AND SILVERMAN, B. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach.* Chapman and Hall/CRC: Boca Raton, FL.

GREEN, P. AND YANDELL, B. (1985). Semi-parametric generalized linear models. Technical Report 2847, University of Wisconsin-Madison.

HAMILTON, S. A. AND TRUONG, Y. K. (1997). Local linear estimation in partly linear models. *Journal of Multivariate Analysis*, **60**, 1–19.

HARDLE, W., LIANG, H., AND GAO, J. (2000). *Partially Linear Models.* Springer-Verlag: New York.

HARE, W. AND SAGASTIZABAL, C. (2009). Computing proximal points of non-convex functions. *Mathematical Programming, Series B*, **116** (1), 221–258.

HASTIE, T. AND TIBSHIRANI, R. (1986). Generalized additive models. *Statistical Science*, **1** (3), 297–310.

HASTIE, T., TIBSHIRANI, R., AND FRIEDMAN, J. (2009). *The Elements of Statistical Learning.* Springer: New York.

HECKMAN, N. (1986). Spline smoothing in partly linear models. *Journal of the Royal Statistical Society, Series B*, **48**, 44–248.

HOLLAND, A. D. (2017). Penalized spline estimation in the partially linear model. *Journal of Multivariate Analysis*, **153**, 211–235.

HUBER, P. (1981). *Robust Statistics.* Wiley Series in probability and Mathematical Statistics. Wiley and Sons: Hoboken, NJ.

KOVAC, A. AND SILVERMAN, B. W. (2000). Extending the scope of wavelet regression methods by coefficient-dependent thresholding. *Journal of the American Statistical Association*, **95**, 172–183.

LI, Q. (2000). Efficient estimation of additive partially linear models. *International Economic Review*, **41** (4), 1073–1092.

LIU, X., WANG, L., AND LIANG, H. (2011). Estimation and variable selection for semiparametric additive partial linear models. *Statistica Sinica*, **21**, 1225–1248.

MALLAT, S. G. (1999). *A Wavelet Tour of Signal Processing*. Second edition. Academic Press: San Diego, CA.

MARRA, G. AND WOOD, S. N. (2011). Practical variable selection for generalized additive models. *Computational Statistics and Data Analysis*, **55**, 2372–2387.

MEIER, L., VAN DE GEER, S., AND BÜHLMANN, P. (2009). High-dimensional additive modeling. *Annals of Statistics*, **37**, 3779–3821.

MÜLLER, M. (2014). An introduction to the estimation of gplms and data examples for the R `gplm` package. Package vignette. R package.
URL: `http://cran.r-project.org`

NASON, G. (2010). Wavethresh 4.5. R package.
URL: `http://cran.r-project.org`

OPSOMER, J. D. AND RUPPERT, D. (1999). A root-*n* consistent backfitting estimator for semiparametric additive modeling. *Journal of Computational and Graphical Statistics*, **8** (4), 715–732.

O'SULLIVAN, F. (1986). A statistical perspective on ill-posed inverse problems (with discussion). *Statistical Science*, **1**, 505–527.

QU, L. (2006). Bayesian wavelet estimation of partially linear models. *Journal of Statistical Computation and Simulation*, **76**, 605–617.

R (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

RAVIKUMAR, P., LIU, H., LAFFERTY, J., AND WASSERMAN, L. (2009). SpAM: Sparse Additive Models. *Journal of the Royal Statistical Society, Series B*, **71**, 1009–1030.

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

ROBINSON, P. M. (1988). Root *n*-consistent semiparametric regression. *Econometrica*, **56**, 931–954.

SAMAROV, A., SPOKOINY, V., AND VIAL, C. (2005). Component identification by structural adaptation. *Journal of the American Statistical Association*, **100** (470), 429–445.

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

SCHMALENSEE, R. AND STOKER, T. M. (1999). Household gasoline demand in the United States. *Econometrica*, **67**, 645–662.

SEVERINI, T. A. AND STANISWALIS, J. G. (1994). Quasi-likelihood estimation in semiparametric models. *Journal of the American Statistical Association*, **89**, 501–511.

SHE, Y. (2012). An iterative algorithm for fitting nonconvex penalized generalized linear models with grouped predictors. *Computational Statistics and Data Analysis*, **56**, 2976–2990.

SHE, Y. AND OWEN, A. B. (2011). Outlier detection using nonconvex penalized regression. *Journal of the American Statistical Association*, **106** (494), 626–639.

SIMON, N. AND TIBSHIRANI, R. (2012). Standardization and the group lasso penalty. *Statistica Sinica*, **22**, 983–1001.

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

TONG, X. W., CUI, H. J., AND ZHAO, H. (2005). Asymptotics of Huber-Dutter estimators for partial linear model with nonstochastic designs. *Acta Mathematicae Applicatae Sinica, English version*, **21** (2), 257–268.

WAHBA, G. (1984). Partial spline models for the semiparametric estimation of functions of several variables. *In Analyses for Time Series, Japan–US Joint Seminar*. Institue of Statistical Mathematics, Tokyo, pp. 319–329.

WAND, M. P. AND OMEROD, J. T. (2008). On semiparametric regression with O'Sullivan penalised splines. *Australian and New Zealand Journal of Statistics*, **50**, 179–198.

WAND, M. P. AND OMEROD, J. T. (2011). Penalized wavelets: Embedding wavelets into semiparametric regression. *Electronic Journal of Statistics*, **5**, 1654–1717.

WANG, L., BROWN, L. D., AND CAI, T. (2011a). A difference based approach to the semiparametric partial linear model. *Electronic Journal of Statistics*, **5**, 619–641.

WANG, L., LIU, X., LIANG, H., AND CARROLL, R. J. (2011b). Estimation and variable selection for generalized additive partial linear models. *Annals of Statistics*, **39**, 1827–185.

WEI, F. (2012). Group selection in high-dimensional partially linear additive models. *Brazilian Journal of Probability and Statistics*, **26** (3), 219–243.

WOOD, S. N. (2006). *Generalized Additive Models: An Introduction with R.* Chapman and Hall/CRC: Boca Raton.

XIE, H.-L. AND HUANG, J. (2009). SCAD-penalized regression in high-dimensional partially linear models. *Annals of Statistics*, **37**, 673–696.

YATCHEW, A. (1997). An elementary estimator of the partial linear model. *Economics Letters*, **57**, 135–143.

YUAN, M. AND LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of Royal Statistical Society, Series B*, **68**, 49–67.

ZHANG, H. H., CHENG, G., AND LIU, Y. (2011). Linear or nonlinear? Automatic structure discovery for partially linear models. *Journal of the American Statistical Association*, **106**, 1099–1112.

ZOU, H. AND LI, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models (with discussion). *Annals of Statistics*, **36**, 1509–1533.

Manuscript received, 2017-05-10, revised, 2017-06-01, accepted, 2017-06-06.