Parsimonious Additive Models

Marta Avalos, Yves Grandvalet, Christophe Ambroise

Marta Avalos[1,2], Yves Grandvalet[1], Christophe Ambroise[1]

[1] *HeuDiaSyC, UMR CNRS 6599, Université de Technologie de Compiègne 60205 Compiègne Cedex, FRANCE*

[2] *Equipe de Biostatistique INSERM E0338, ISPED, Université de Bordeaux 2 33076 Bordeaux Cedex, FRANCE*

Abstract

We present a new method for function estimation and variable selection, specifically designed for additive models fitted by cubic splines. Our method involves regularizing additive models using the l_1–norm, which generalizes Tibshirani's lasso to the nonparametric setting. As in the linear case, it shrinks coefficients, some of them reducing exactly to zero. It gives parsimonious models, select significant variables, and reveal nonlinearities in the effects of predictors. Two strategies for finding a parsimonious additive model solutions are proposed. Both algorithms are based on a fixed point algorithm, combined with a singular value decomposition that considerably reduces computation. The empirical behavior of parsimonious additive models is compared to the adaptive backfitting BRUTO algorithm. The results allow us to characterise the domains in which our approach is effective: it performs significantly better than BRUTO when model estimation is challenging. An implementation of this method is illustrated using real data from the Cophar 1 ANRS 102 trial. Parsimonious additive models are applied to predict the indinavir plasma concentration in HIV patients. Results suggest that our method is a promising technique for the research and application areas.

Key words: model selection, supervised learning, nonparametric regression, function estimation, splines, smoothing, variable selection, lasso, penalization, interpretable models.

* Corresponding author: Marta Avalos, Institut de Santé Publique, Epidémiologie et Développement, Université Victor Segalen Bordeaux 2, 146 rue Léo Saignat 33076 Bordeaux, FRANCE.
E-mail: marta.avalos@isped.u-bordeaux2.fr, Tel: 33(0)5 57 57 15 34, Fax: 33 (0)5 56 24 00 81.

Preprint submitted to Elsevier Science

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 $X_1, ..., X_p$, 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*, popularized by Hastie and Tibshirani [23]. An additive model is defined by

$$Y = \alpha_0 + \sum_{j=1}^{p} f_j(X_j) + \varepsilon, \tag{1}$$

where the error ε is independent of the predictor variables X_j, $\mathbb{E}(\varepsilon) = 0$ and $\text{var}(\varepsilon) = \sigma^2$. f_j are univariate smooth functions, defined such that $\mathbb{E}_{X_j}(f_j) = 0$ in order to ensure unicity, and α_0 is a constant.

The additive structure does not assume a rigid form for the dependence of Y on $X_1, ..., X_p$ so nonparametric flexibility is preserved. Also, the additive model retains an important interpretive feature of the linear model: we can represent the functions f_j to analyze the effects of the predictors on the response. Moreover, it overcomes problems of high–dimensionality: since the response variable is modeled as the sum of univariate functions of predictor variables, the number of observations required to get variance–stable estimates grows only linearly in p. The price to pay for such interesting properties is the possible misspecification of the model.

As in any statistical learning task, model selection is an important issue in the estimation of additive models. The problem of determining the model structure that best fits the data can be decomposed in two subproblems: complexity tuning and variable selection. In nonparametric regression there is a fundamental trade–off between the bias and variance of the estimate, which is typically governed by a regularization or smoothing parameter. Complexity tuning addresses the question "what is the right amount of smoothing" [23,24]. Variable selection consists in selecting input variables that are most predictive of a given outcome. Appropriate variable selection aims at improving prediction performance, enhancing understanding of the underlying concept that generated the data and reducing training time [21].

Subset selection strategies have been applied to additive models. These proposals exploit the fact that additive regression generalizes linear regression. Thus, we find hypothesis tests [12,10,22,40,16], techniques based on a prediction error estimator [9,7,32], as well as Bayesian approaches [34,33].

A different approach to variable selection consists of regularizing additive models using the l_1–norm. We present a new method for variable selection and complexity tuning specifically designed for additive models fitted by cubic splines. We use a three–part objective function that includes goodness–of–fit and a double penalty: on the l_1–norm of linear components of cubic splines coefficients and on the (generalized) l_1–norm of nonlinear components of cubic spline coefficients. Because of their nature, these penalties shrink linear and nonlinear compounds, some of them reducing exactly to zero. Hence they give parsimonious models, select significant variables, and reveal nonlinearities in the effects of predictors.

Two strategies for finding a parsimonious additive model solutions are proposed. In both of them, curve fitting is based on a fixed–point algorithm solving the penalized optimization problem, combined with a singular value decomposition that considerably reduces computation. The empirical behavior of parsimonious additive models is compared to the adaptive backfitting BRUTO algorithm for additive models [23]. The results allow us to deduce conditions of application for each method. Our method performs significantly better than BRUTO when model estimation is challenging.

This new approach is applied to predict indinavir (an antiretroviral from the protease inhibitor class) plasma concentration in HIV patients, from the Cophar 1 ANRS 102 trial.

This article is organized as follows. In section 2 we review penalization techniques. Additive models fitted by cubic splines are introduced in section 3. We present our approach in section 4. In section 5 we discuss estimation of the regularization parameters. Simulation studies are described in section 6. A real data example is given in section 7. Finally, section 8 contains concluding remarks and perspectives.

2 Penalization Techniques for Linear Models

In this section we provide a brief review of penalization techniques for linear models by way of an introduction to our approach for additive models.

Consider the usual linear regression setting. The ordinary least–squares estimate is obtained by minimizing the residual squared error. However, if the number of covariates p is large (with respect to the number of examples n) or if the regressor variables are highly correlated, then the variance of the least–squares estimate may be high, leading to prediction inaccuracy.

Penalization is extensively used to address these difficulties. It decreases the

predictor variability to improve the accuracy of prediction. It also tends to produce models with few non–zero coefficients if interpretation is planned. Ridge regression (l_2 penalization) and subset selection (l_0 penalization) are the two main penalization procedures. The former is stable, but does not shrink parameters to zero, whereas the latter gives simple models, but is unstable [5].

l_1 penalization, termed the *lasso* (least absolute shrinkage and selection operator) [36,24] provides an alternative to these techniques. The lasso estimates the vector of linear regression coefficients by minimizing the residual sum of squares, subject to a constraint on the l_1–norm of the coefficient vector. An attractive feature of the l_1–norm constraint is that it shrinks coefficients and sets some of them to zero. The smooth form of the constraint leads to a convex optimization problem which provides a stable model.

Suppose we have data $(x_{i1}, \ldots, x_{ip}, y_i)$, $i = 1, \ldots, n$, where x_{ij} are the standardized predictor variables and $\mathbf{y} = (y_1, \ldots, y_n)^t$ are the centered responses. The observations are assumed to be independent and identically distributed. We denote by $\mathbf{X}$ the design matrix $\{x_{ij}\}$, and by $\mathbf{x}_j$ the vector $(x_{1j}, \ldots, x_{nj})^t$, $i = 1, \ldots, n$, $j = 1, \ldots, p$. To simplify the notation, we suppose there is no intercept in the model.

The lasso estimator solves the optimization problem

$$\min_{\alpha_1,\ldots,\alpha_p} \left\| \mathbf{y} - \sum_{j=1}^{p} \mathbf{x}_j \alpha_j \right\|_2^2 \quad \text{subject to} \quad \sum_{j=1}^{p} |\alpha_j| \leq \tau, \tag{2}$$

where the predefined value τ controls model complexity.

A convenient formulation of the lasso is given by the adaptive ridge [18], which was proposed as a means of automatically balancing penalization on each variable. The two procedures are equivalent, in the sense that they produce the same estimate [17].

The adaptive ridge estimate is the minimizer of the problem

$$\begin{aligned} \min_{\substack{\alpha_1,\ldots,\alpha_p \\ \mu_1,\ldots,\mu_p}} \quad & \left\| \mathbf{y} - \sum_{j=1}^{p} \mathbf{x}_j \alpha_j \right\|_2^2 + \sum_{j=1}^{p} \mu_j \alpha_j^2 \\ \text{subject to } & \sum_{j=1}^{p} \frac{1}{\mu_j} = \frac{p}{\mu} \\ & \mu_j > 0, \end{aligned} \tag{3}$$

where the predefined value μ controls global model complexity, and the values of μ_j are automatically induced from the sample.

Sketch of proof of equivalence (a rigorous detailed proof is given in appendix A)

The Lagrangian $\mathcal{L}$ corresponding to problem (3) is

$$\left\| \mathbf{y} - \sum_{j=1}^{p} \mathbf{x}_j \alpha_j \right\|_2^2 + \sum_{j=1}^{p} \mu_j \alpha_j^2 + \nu \left(\sum_{j=1}^{p} \frac{1}{\mu_j} - \frac{p}{\mu} \right) + \sum_{j=1}^{p} \eta_j \mu_j \ ,$$

where ν and η_j's are the Lagrange multipliers pertaining to the constraints in problem (3). A necessary condition for optimality is obtained by deriving the Lagrangian with respect to μ_j, which reads

$$\mu_j = \frac{\sqrt{\nu}}{|\alpha_j|} \ ,$$

and pluging this expression in the first constraint of problem (3) yields

$$\sqrt{\nu} = \frac{\mu}{p} \sum_{j=1}^{p} |\alpha_j| \ .$$

The optimal μ_j are then obtained from μ and α_k's, so that problem (3) is rewritten

$$\min_{\alpha_1,\ldots,\alpha_p} \left\| \mathbf{y} - \sum_{j=1}^{p} \mathbf{x}_j \alpha_j \right\|_2^2 + \frac{\mu}{p} \left(\sum_{j=1}^{p} |\alpha_j| \right)^2 \ ,$$

which is equivalent to minimizing the squared error loss subject to $\sum_{j=1}^{p} |\alpha_j| \leq t$ for some t, which is exaclty the lasso problem.

□

The adaptive ridge formulation of lasso does not explicitely entail sparse solutions, but it suggests means to generalize the lasso idea to the sums of quadratic penalties. It inspired the parsimonious additive models described in section 4.

3 Additive Models

The cubic smoothing spline estimator is defined as the minimizer of a penalized least–squares criterion over functions belonging to a reproducing kernel Hilbert space (RKHS), $\mathcal{H}$,

$$\min_{f \in \mathcal{H}} \|\mathbf{y} - f(\mathbf{x})\|_2^2 + \lambda \left\| D^2 f \right\|_{L_2}^2 \ , \tag{4}$$

where $f(\mathbf{x})$ denotes the vector $(f(x_1), \ldots, f(x_n))^t$, the differential operator D^2 maps functions to their second derivative, $\|\cdot\|_{L_2}$ denotes the L_2 norm and $\mathcal{H}$ is defined as the space of twice–continuously–differentiable functions f, with all points evaluation functional and finite $\|D^2 f\|_{L_2}$.

Cubic splines are extended to additive models in a straightforward manner [38,24]:

$$\min_{(f_1,\ldots,f_p)\in\mathcal{H}_1\times\ldots\times\mathcal{H}_p} \left\| \mathbf{y} - \sum_{j=1}^{p} f_j(\mathbf{x}_j) \right\|_2^2 + \sum_{j=1}^{p} \lambda_j \left\| D^2 f_j \right\|_{L_2}^2 \quad , \tag{5}$$

where each space $\mathcal{H}_j$ is defined as the space of twice–continuously–differentiable functions f_j of X_j, with all points evaluation functional and finite $\|D^2 f_j\|_{L_2}$.

Each function in (5) is penalized by a smoothing parameter λ_j. Large values of λ_j produce smoother curves for the jth component, while smaller values produce more wiggly curves [23]. At the one extreme, as $\lambda_j \to \infty$, the penalty term dominates, forcing $D^2 f_j \equiv 0$, and thus the solution for the jth component is the least–squares line. At the other extreme, $\lambda_j \to 0$, the penalty term becomes unimportant and the solution for the jth component tends to an interpolating twice–differentiable function.

Before solving (5), we should have already determined p smoothing parameters λ_j. Several methods have been proposed to estimate smoothing parameters. These methods are based on generalizing univariate techniques such as generalized cross validation (used at each step of the backfitting procedure in BRUTO [23], optimised by a Newton method [19,40] or combined with a diagonalization technique for penalized splines [31]), Akaike information criteria [23,25,27], Bayesian information criteria [27], plug in [28], or hypothesis testing [8].

A different approach to this problem was proposed by Grandvalet *et al.* [17,18]. It involves the extension of the lasso to additive models fitted by cubic splines, using the adaptive ridge formulation

$$\begin{aligned} \min_{\substack{(f_1,\ldots,f_p)\in\mathcal{H}_1\times\ldots\times\mathcal{H}_p \\ \lambda_1,\ldots,\lambda_p}} & \left\| \mathbf{y} - \sum_{j=1}^{p} f_j(\mathbf{x}_j) \right\|_2^2 + \sum_{j=1}^{p} \lambda_j \left\| D^2 f_j \right\|_{L_2}^2 \\ \text{subject to} \quad & \sum_{j=1}^{p} \frac{1}{\lambda_j} = \frac{p}{\lambda} \\ & \lambda_j > 0, \end{aligned} \tag{6}$$

where the penalization on each variable is optimized to minimize residuals, and consequently only λ has to be defined prior to the estimation procedure.

Transposing the linear adaptive ridge (3) to additive cubic spline fitting (6) amounts essentially to transforming a l_2–norm in a L_2–norm.

Expression (6) shows that we address the standard additive spline model (5), except that the penalization terms λ_j applied to each additive component are optimized subject to constraints. This writing can be motivated from a hierarchical Bayesian viewpoint in the maximum a posteriori framework.

Again, if some λ_j goes to infinity, the solution for the jth component is the least–squares line. Hence no predictor is likely to be eliminated by solving (6). The following section presents two proposals dedicated at removing irrelevant variables from the model.

4 Parsimonious Additive Models

Following [23], additive models form the subspace $\mathcal{H}^{\mathtt{add}}$, which can be decomposed as $\mathcal{H}^{\mathtt{add}} = \mathcal{H}_1 + \ldots + \mathcal{H}_p \subset \mathcal{H}$. Furthermore, the RKHS of cubic spline functions $\mathcal{H}$ can be decomposed in the direct sum of two components $\mathcal{H}^{\mathtt{L}} \oplus \widetilde{\mathcal{H}}$. The space of linear functions $\mathcal{H}^{\mathtt{L}}$ corresponds to the null space of the semi-norm $\|D^2 f\|^2_{L_2}$, that is, ultimately smooth functions according to the roughness penalty. The orthogonal complement $\widetilde{\mathcal{H}}$ is the space where the roughness penalty defines a norm.

Our approach to Parsimonious Additive Models (PAM) consists in defining new roughness penalties that are norms on $\mathcal{H}^{\mathtt{add}}$, for which the solution can still be expressed as an additive cubic spline model. The abovementioned decompositions of $\mathcal{H}^{\mathtt{add}}$ and $\mathcal{H}$ suggest the two penalization schemes detailed below.

4.1 Modified Roughness Penalties (PAM1)

A first penalization scheme stems from the usual decomposition $\mathcal{H}^{\mathtt{add}} = \mathcal{H}_1 + \ldots + \mathcal{H}_p \subset \mathcal{H}$, followed by $\mathcal{H}_j = \mathcal{H}_j{}^{\mathtt{L}} \oplus \widetilde{\mathcal{H}_j}$. The idea is to use a standard additive model, where we modify the roughness penalty on each component. We first define the norm on the subspaces $\mathcal{H}_j{}^{\mathtt{L}}$ as the l_1–norm on the expansion on the basis $(1, x_j)$.

The optimization problem is now:

$$\begin{aligned}
&\min_{\substack{\boldsymbol{\alpha},(f_1,\ldots,f_p)\in\widetilde{\mathcal{H}_1}\times\ldots\times\widetilde{\mathcal{H}_p}\\ \lambda_1,\ldots,\lambda_p}} \left\|\mathbf{y}-\sum_{j=1}^{p}\mathbf{x}_j\alpha_j-\sum_{j=1}^{p}f_j(\mathbf{x}_j)\right\|_2^2+\mu\sum_{j=1}^{p}|\alpha_j|+\sum_{j=1}^{p}\lambda_j\left\|D^2f_j\right\|_{L_2}^2\\
&\text{subject to}\qquad \sum_{j=1}^{p}\frac{1}{\lambda_j}=\frac{p}{\lambda}\\
&\qquad\qquad\quad \lambda_j>0\ .
\end{aligned}\tag{7}$$

where the two predefined values μ and λ tune the global model complexity, while the induced values λ_j control the individual complexities of f_j.

In (7), each function f_j is restricted to lie in $\widetilde{\mathcal{H}_j}$, the orthogonal of linear functions in $(1, x_j)$. This optimisation problem can be reformulated using functions of the usual cubic spline spaces $\mathcal{H}_1 \ldots \mathcal{H}_p$ as follows

$$\begin{aligned}
&\min_{\substack{\boldsymbol{\alpha},(f_1,\ldots,f_p)\in\mathcal{H}_1\times\ldots\times\mathcal{H}_p\\ \lambda_1,\ldots,\lambda_p}} \left\|\mathbf{y}-\sum_{j=1}^{p}\mathbf{x}_j\alpha_j-\sum_{j=1}^{p}f_j(\mathbf{x}_j)\right\|_2^2+\mu\sum_{j=1}^{p}|\alpha_j|+\sum_{j=1}^{p}\lambda_j\left\|D^2f_j\right\|_{L_2}^2\\
&\text{subject to}\qquad \sum_{j=1}^{p}\frac{1}{\lambda_j}=\frac{p}{\lambda}\\
&\qquad\qquad\quad \lambda_j>0\\
&\qquad\qquad\quad \langle\mathbf{f}_j,\mathbf{1}\rangle=0\quad j=1,\ldots,p\\
&\qquad\qquad\quad \langle\mathbf{f}_j,\mathbf{x}_j\rangle=0\quad j=1,\ldots,p\ ,
\end{aligned}\tag{8}$$

where $\mathbf{1}$ is a n–dimensional vector of ones and $\mathbf{f}_j$ is the vector of the jth additive component evaluated at $\mathbf{x}_j$.

The parsimony of (8) follows from the equivalence between adaptive ridge and l_1 penalization. If, after convergence, $\frac{1}{\lambda_j}=0$, then $\|D^2f_j\|_{L_2}^2$ is shrunk to zero and the effect of the jth variable is linearized. If α_j is null, the effect of the jth variable is estimated to be strictly nonlinear (since $\langle\mathbf{f}_j,\mathbf{x}_j\rangle=0$). Finally, if $\alpha_j=0$ and $\|D^2f_j\|_{L_2}^2=0$, the corresponding variable is removed from the model.

We can represent (8) in terms of spline bases. Let $\mathbf{N}_j$ denote the $n\times(n+2)$ matrix of the unconstrained natural B–spline basis, evaluated at x_{ij}. Let $\boldsymbol{\Omega}_j$ be the $(n+2)\times(n+2)$ matrix corresponding to the penalization of the second derivative of f_j. The coefficients of f_j in the unconstrained B–spline basis are noted $\boldsymbol{\beta}_j$. We thus have $\mathbf{f}_j=\mathbf{N}_j\boldsymbol{\beta}_j$, and $\|D^2f_j\|_{L_2}^2=\boldsymbol{\beta}_j^t\boldsymbol{\Omega}_j\boldsymbol{\beta}_j$. Problem (8) can

be rewritten as

$$\begin{aligned}
\min_{\substack{\boldsymbol{\alpha},\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_p \\ \lambda_1,\ldots,\lambda_p}} \quad & \left\| \mathbf{y} - \sum_{j=1}^{p} \mathbf{x}_j \alpha_j - \sum_{j=1}^{p} \mathbf{N}_j \boldsymbol{\beta}_j \right\|_2^2 + \mu \sum_{j=1}^{p} |\alpha_j| + \sum_{j=1}^{p} \lambda_j \boldsymbol{\beta}_j^t \boldsymbol{\Omega}_j \boldsymbol{\beta}_j \\
\text{subject to } & \sum_{j=1}^{p} \frac{1}{\lambda_j} = \frac{p}{\lambda} \\
& \lambda_j > 0 \\
& \mathbf{1}^t \mathbf{N}_j \boldsymbol{\beta}_j = 0 \quad j = 1, \ldots, p \\
& \mathbf{x}_j^t \mathbf{N}_j \boldsymbol{\beta}_j = 0 \quad j = 1, \ldots, p \ .
\end{aligned} \tag{9}$$

4.2 Additive Nonlinear Effects (PAM2)

The second penalization scheme originates from the opposite processing, where the decomposition $\mathcal{H} = \mathcal{H}^{\mathsf{L}} \oplus \widetilde{\mathcal{H}}$, is followed by $\widetilde{\mathcal{H}} \supset \widetilde{\mathcal{H}}_1 + \ldots + \widetilde{\mathcal{H}}_p$. The idea now is to process linear and nonlinear components separately, the nonlinear component being handled by the additive model. Such a decomposition favors interpretability, since the nonlinear components are restricted to explain what cannot be described by linear effects [1] . We now define the norm on the subspace $\mathcal{H}^{\mathsf{L}}$ as the l_1–norm on the expansion on the basis $(1, x_1, \ldots, x_p)$ and the optimization problem is

$$\begin{aligned}
\min_{\substack{\boldsymbol{\alpha},(f_1,\ldots,f_p)\in\widetilde{\mathcal{H}}_1\times\ldots\times\widetilde{\mathcal{H}}_p \\ \lambda_1,\ldots,\lambda_p}} \quad & \left\| \mathbf{y} - \sum_{j=1}^{p} \mathbf{x}_j \alpha_j - \sum_{j=1}^{p} f_j(\mathbf{x}_j) \right\|_2^2 + \mu \sum_{j=1}^{p} |\alpha_j| + \sum_{j=1}^{p} \lambda_j \left\| D^2 f_j \right\|_{L_2}^2 \\
\text{subject to } & \sum_{j=1}^{p} \frac{1}{\lambda_j} = \frac{p}{\lambda} \\
& \lambda_j > 0 \ .
\end{aligned} \tag{10}$$

In (10) the functions f_j are restricted to lie in $\widetilde{\mathcal{H}}_j$, which are subsets of $\widetilde{\mathcal{H}}$, the orthogonal of linear functions in $(1, x_1, \ldots, x_p)$. This optimisation problem can be reformulated using functions of the usual cubic spline spaces $\mathcal{H}_1 \ldots \mathcal{H}_p$

[1] This restraint occurs in the usual additive spline model, as a consequence of the shrinking of all nonlinear effects coupled with the projection on linear effects.

as follows

$$\begin{aligned}
\min_{\substack{\boldsymbol{\alpha},(f_1,\ldots,f_p)\in\mathcal{H}_1\times\ldots\times\mathcal{H}_p \\ \lambda_1,\ldots,\lambda_p}} & \left\|\mathbf{y}-\sum_{j=1}^{p}\mathbf{x}_j\alpha_j-\sum_{j=1}^{p}f_j(\mathbf{x}_j)\right\|_2^2+\mu\sum_{j=1}^{p}|\alpha_j|+\sum_{j=1}^{p}\lambda_j\left\|D^2f_j\right\|_{L_2}^2 \\
\text{subject to} \quad & \sum_{j=1}^{p}\frac{1}{\lambda_j}=\frac{p}{\lambda} \\
& \lambda_j>0 \\
& \langle\mathbf{f}_j,\mathbf{1}\rangle=0 \quad j=1,\ldots,p \\
& \langle\mathbf{f}_j,\mathbf{x}_k\rangle=0 \quad j=1,\ldots,p \quad k=1,\ldots,p \ .
\end{aligned} \tag{11}$$

Note that this approach differs from the previous one only in the last constraints, where $\mathbf{f}_j$ are required to be orthogonal to $\mathbf{x}_k$ for $k\neq j$.

In terms of a spline bases, (11) can be rewritten as

$$\begin{aligned}
\min_{\substack{\boldsymbol{\alpha},\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_p \\ \lambda_1,\ldots,\lambda_p}} & \left\|\mathbf{y}-\sum_{j=1}^{p}\mathbf{x}_j\alpha_j-\sum_{j=1}^{p}\mathbf{N}_j\boldsymbol{\beta}_j\right\|_2^2+\mu\sum_{j=1}^{p}|\alpha_j|+\sum_{j=1}^{p}\lambda_j\boldsymbol{\beta}_j^t\boldsymbol{\Omega}_j\boldsymbol{\beta}_j \\
\text{subject to} \quad & \sum_{j=1}^{p}\frac{1}{\lambda_j}=\frac{p}{\lambda} \\
& \lambda_j>0 \\
& \mathbf{1}^t\mathbf{N}_j\boldsymbol{\beta}_j=0 \quad j=1,\ldots,p \\
& \mathbf{x}_k^t\mathbf{N}_j\boldsymbol{\beta}_j=0 \quad j=1,\ldots,p \quad k=1,\ldots,p \ .
\end{aligned} \tag{12}$$

4.3 Actual decomposition into linear and nonlinear subspaces

Splines are linear smoothers: that is, the univariate fits can be written as $\widehat{\mathbf{f}}_j=\mathbf{S}_j\mathbf{y}$, where $\mathbf{S}_j$ is a $n\times n$ matrix, called the *smoother matrix*, free of $\mathbf{y}$.

For the jth covariate, the smoother matrix is computed as

$$\mathbf{S}_j=\mathbf{N}_j(\mathbf{N}_j^t\mathbf{N}_j+\lambda_j\boldsymbol{\Omega}_j)^{-1}\mathbf{N}_j^t \ . \tag{13}$$

It has two unitary eigenvalues corresponding to the constant and linear functions (its projection part), and $n-2$ non–negative eigenvalues strictly smaller than 1 corresponding to higher–order functions (its *shrinking* part). For the purpose of minimizing (9) and (12), we decompose the smoother matrix: $\mathbf{S}_j=\mathbf{H}_j+\widetilde{\mathbf{S}}_j$, where $\mathbf{H}_j$ is the matrix that projects onto the space of eigenvalue 1 for the jth smoother (the *hat* matrix corresponding to least–squares

Algorithm PAM1:

(1) Fix μ and λ

(2) Initialize $\boldsymbol{\beta}_j$ ($\mathbf{0}$ by default) and λ_j (λ by default), $j = 1, \ldots, p$.

(3) Singular value decomposition and hat matrices:

(a) Compute eigenvalue decomposition: $\boldsymbol{\Omega}_j = \mathbf{P}_j \mathbf{D}_j \mathbf{P}_j^t$.

(b) Replace the eigenvalues corresponding to the null space (linear and constant functions) by a positive value (1 by default).

(c) Denote: $\mathbf{Q}_j = \mathbf{N}_j \mathbf{P}_j \mathbf{D}_j^{-1/2}$.

(d) Compute singular value decomposition: $\mathbf{Q}_j = \mathbf{U}_j \mathbf{Z}_j \mathbf{V}_j^t$.

(e) Compute hat matrices: $\mathbf{H}_j = \mathbf{x}_j (\mathbf{x}_j^t \mathbf{x}_j)^{-1} \mathbf{x}_j^t$.

(4) Linear components:

(a) Compute residual: $\mathbf{r} = \mathbf{y} - \sum_{j=1}^p \mathbf{N}_j \boldsymbol{\beta}_j$.

(b) Compute coefficients:

$$\boldsymbol{\alpha} = \text{argmin}_{\boldsymbol{\alpha}} \left\| \mathbf{r} - \sum_{j=1}^p \mathbf{x}_j \alpha_j \right\|_2^2 + \mu \sum_{j=1}^p |\alpha_j| \ .$$

(5) Nonlinear components:

(a) Estimate nonlinear coefficients via backfitting:

(i) Compute smoother and shrinking matrices:

$$\mathbf{S}_j = \mathbf{U}_j \mathbf{Z}_j \left(\mathbf{Z}_j^t \mathbf{Z}_j + \lambda_j \mathbf{I} \right)^{-1} \mathbf{Z}_j^t \mathbf{U}_j^t \text{ and } \widetilde{\mathbf{S}}_j = \mathbf{S}_j - \mathbf{H}_j \ .$$

(ii) Compute partial residual: $\mathbf{r}_j = \mathbf{y} - \sum_{k \neq j} \mathbf{N}_j \boldsymbol{\beta}_j$.

(iii) Compute coefficients: $\mathbf{N}_j \boldsymbol{\beta}_j = \widetilde{\mathbf{S}}_j \mathbf{r}_j$.

(b) Re-estimate penalizers: $\lambda_j = \lambda \dfrac{\sum_{j=1}^p \sqrt{\boldsymbol{\beta}_j^t \boldsymbol{\Omega}_j \boldsymbol{\beta}_j}}{p \sqrt{\boldsymbol{\beta}_j^t \boldsymbol{\Omega}_j \boldsymbol{\beta}_j}}$.

(6) Repeat (4) and (5) until convergence.

Fig. 1. Iterative algorithm PAM1 (modified roughness penalties) for carrying out both function estimation and variable selection in additive models.

regression on $\mathbf{x}_j$), and $\widetilde{\mathbf{S}}_j$ is the *shrinking* matrix [23]. This decomposition is practical for PAM1 as for PAM2: using $\widetilde{\mathbf{S}}_j$ (instead of $\mathbf{S}_j$) in the backfitting loop improves the numerical stability.

4.4 Algorithms

Problems (9) and (12) can be solved by a fixed point algorithm including backfitting.

The general outline of the algorithm (figures 1 and 2) is the following. Firstly, the two regularization parameters are fixed (step 1) and penalization terms are initialized (step 2). Secondly, the matrices $\mathbf{S}_j$ are decomposed to improve efficiency in the backfitting loop (step 3). Thirdly, in step 4, the linear components are estimated by solving a lasso problem, using the algorithm proposed by Osborne *et al.* [29]. Recently, Efron *et al.* [14] developed a least–angle regression which can readily provide all lasso solutions in a highly efficient fashion. Using this algorithm could marginally increase the computational efficiency for

Algorithm PAM2:

(1) Fix μ and λ

(2) Initialize $\boldsymbol{\beta}_j$ ($\mathbf{0}$ by default) and λ_j (λ by default), $j = 1, \ldots, p$.

(3) Singular value decomposition and hat matrices:

(a) Compute eigenvalue decomposition: $\boldsymbol{\Omega}_j = \mathbf{P}_j \mathbf{D}_j \mathbf{P}_j^t$.

(b) Replace the eigenvalues corresponding to the null space (linear and constant functions) by a positive value (1 by default).

(c) Denote: $\mathbf{Q}_j = \mathbf{N}_j \mathbf{P}_j \mathbf{D}_j^{-1/2}$.

(d) Compute singular value decomposition: $\mathbf{Q}_j = \mathbf{U}_j \mathbf{Z}_j \mathbf{V}_j^t$.

(e) Compute hat matrices: $\mathbf{H}_j = \mathbf{x}_j (\mathbf{x}_j^t \mathbf{x}_j)^{-1} \mathbf{x}_j^t$.

(4) Linear components:

(a) Compute the ordinary least–squares estimate:

$$\boldsymbol{\alpha}_{\text{ols}} = \text{argmin}_{\boldsymbol{\alpha}} \left\| \mathbf{y} - \sum_{j=1}^{p} \mathbf{x}_j \alpha_j \right\|_2^2 .$$

(b) Compute coefficients

$$\boldsymbol{\alpha} = \text{argmin}_{\boldsymbol{\alpha}} \left\| \mathbf{y} - \sum_{j=1}^{p} \mathbf{x}_j \alpha_j \right\|_2^2 + \mu \sum_{j=1}^{p} |\alpha_j| .$$

(5) Nonlinear components:

(a) Estimate nonlinear coefficients via backfitting

(i) Compute smoother and shrinking matrices:

$$\mathbf{S}_j = \mathbf{U}_j \mathbf{Z}_j \left(\mathbf{Z}_j^t \mathbf{Z}_j + \lambda_j \mathbf{I} \right)^{-1} \mathbf{Z}_j^t \mathbf{U}_j^t \text{ and } \widetilde{\mathbf{S}}_j = \mathbf{S}_j - \mathbf{H}_j .$$

(ii) Compute partial residual: $\mathbf{r}_j = \mathbf{y} - \mathbf{X}\boldsymbol{\alpha}_{\text{ols}} - \sum_{k \neq j} \mathbf{N}_j \boldsymbol{\beta}_j$.

(iii) Compute coefficients: $\mathbf{N}_j \boldsymbol{\beta}_j = \widetilde{\mathbf{S}}_j \mathbf{r}_j$.

(b) Re-estimate penalizers: $\lambda_j = \lambda \dfrac{\sum_{j=1}^{p} \sqrt{\boldsymbol{\beta}_j^t \boldsymbol{\Omega}_j \boldsymbol{\beta}_j}}{p \sqrt{\boldsymbol{\beta}_j^t \boldsymbol{\Omega}_j \boldsymbol{\beta}_j}}$.

(c) Repeat 5.(b) and 5.(c) until convergence.

Fig. 2. Iterative algorithm PAM2 (additive nonlinear effects) for carrying out both function estimation and variable selection in additive models.

PAM2, where linear and nonlinear components are computed independently. Note that step 4(b) differs for the two algorithms: the predicted variable is the residual of the nonlinear fit in PAM1, wheras it is the response variable for PAM2. Finally, given a current estimate for penalization terms, the spline coefficients are calculated (step 5(a)). These coefficient values are then used to get a new estimate for the penalization terms (step 5(b)). In PAM1, this is followed by the re-estimation of the linear component (step 4), and in PAM2, the two steps are iterated until convergence is achieved.

Step 5(a) is an iteration of the backfitting algorithm, which is used to estimate the nonlinear components. First, given the current estimate for penalization terms, smoother and shrinking matrices are calculated (step 5(a)i). Smoother matrices are computed efficiently, using the singular value decomposition of step 3 detailed below. Shrinking matrices are then obtained by an orthogonal projection of smoother matrices onto the space spanned by nonlinear components. Secondly, partial residuals are computed (step 5(a)ii).

Here, PAM1 and PAM2 differ. In PAM1, partial residuals are obtained by subtracting from $\mathbf{y}$ the nonlinear fits from the other covariates. in PAM2, the ordinary least–squares fit is also subtracted from $\mathbf{y}$. This subtraction ensures the orthogonality of linear and nonlinear fits. Thirdly, in step 5(a)iii, using $\widetilde{\mathbf{S}}_j$ to fit residuals ensures the orthogonality constraints $\langle \mathbf{f}_j, \mathbf{x}_j \rangle = 0$.[2] The coefficients $\boldsymbol{\beta}_j$ themselves are in fact not computed explicitely, since we are only interested in $\mathbf{N}_j\boldsymbol{\beta}_j$ in step 5(a)iii and $\boldsymbol{\beta}_j^t\boldsymbol{\Omega}_j\boldsymbol{\beta}_j$ in step 5(b). The latter can be directly computed using the singular value decomposition results of step 3 as $\mathbf{r}_j^t\mathbf{U}_j\mathbf{Z}_j(\mathbf{Z}_j^t\mathbf{Z}_j + \lambda_j\mathbf{I})^{-2}\mathbf{Z}_j^t\mathbf{U}_j^t\mathbf{r}_j$.

Singular Value Decomposition We propose a singular value decomposition (step 3 in figures 1 and 2) that allows us to speed up the computation of the nonlinear components by avoiding matrix inversions in the backfitting inner loop.

The first part of the procedure is to transform $\boldsymbol{\Omega}_j$ in a full rank matrix. The penalization matrix has two null eigenvalues corresponding to the constant and linear functions (as the second derivative of these functions is the null function). However, if a penalization on the linear functions is added, then $\boldsymbol{\Omega}_j$ becomes a full rank matrix. This is performed by replacing the null eigenvalues in the penalization matrix by positive values (step 3(a)). This substitution does not change coefficient estimates, since linear and nonlinear components are treated independently.

Let $\boldsymbol{\Omega}'_j$ be the full rank matrix such that $\mathbf{S}'_j = \mathbf{N}_j(\mathbf{N}_j^t\mathbf{N}_j + \lambda_j\boldsymbol{\Omega}'_j)^{-1}\mathbf{N}_j^t$ has the same eigenvalues and eigenvectors as $\mathbf{S}_j$, except for the two unit eigenvalues (and corresponding eigenvectors). This matrix is obtained as

$$\boldsymbol{\Omega}'_j = \boldsymbol{\Omega}_j + \frac{1}{n}\mathbf{N}_j^t\mathbf{x}\mathbf{x}^t\mathbf{N} + \frac{1}{n}\mathbf{N}_j^t\mathbf{1}\mathbf{1}^t\mathbf{N}. \tag{14}$$

Let $\mathbf{P}_j$ be a unitary matrix and let $\mathbf{D}_j$ be a diagonal matrix such that $\boldsymbol{\Omega}'_j = \mathbf{P}_j\mathbf{D}_j\mathbf{P}_j^t$ (step 3(b)). Define $\mathbf{Q}_j = \mathbf{N}_j\mathbf{P}_j\mathbf{D}_j^{-1/2}$ (step 3(c)) and let $\mathbf{Q}_j = \mathbf{U}_j\mathbf{Z}_j\mathbf{V}_j^t$ be its singular value decomposition (step 3(d)), where $\mathbf{Z}_j$ is a diagonal matrix of the same dimension as $\mathbf{Q}_j$ and with nonnegative diagonal elements, and $\mathbf{U}_j$ and $\mathbf{V}_j$ are unitary matrices. Then we can write

$$\mathbf{S}'_j = \mathbf{U}_j\mathbf{Z}_j(\mathbf{Z}_j^t\mathbf{Z}_j + \lambda_j\mathbf{I})^{-1}\mathbf{Z}_j^t\mathbf{U}_j^t. \tag{15}$$

[2] For PAM2, this constraint is ensured by the subtraction of the ordinary least–squares fit in step 5(a)ii. However, using $\widetilde{\mathbf{S}}_j$ improves stability. Note also that the subtraction of the ordinary least–squares fit could be made only once, after step 4(b), but including it in the loop also improves stability.

The projections of $\mathbf{S}'_j$ and $\mathbf{S}_j$ onto the nonlinear space ($\widetilde{\mathbf{S}}'_j$ and $\widetilde{\mathbf{S}}_j$, respectively) coincide.

Using the singular value decomposition we obtain a simple calculation of the smoother matrices. Matrices $\mathbf{U}_j$ and $\mathbf{Z}_j$ do not depend on either coefficients $\boldsymbol{\beta}_j$ or penalizers λ_j, and so factors $\mathbf{U}_j\mathbf{Z}_j$ and $\mathbf{Z}_j^t\mathbf{Z}_j$ only need to be calculated once for given data. On the other hand, we avoid matrix inversions in the iterative step, since matrices $\left(\mathbf{Z}_j^t\mathbf{Z}_j + \lambda_j\mathbf{I}\right)$ are diagonal.

Notice that the computations needed for the singular value decomposition do not depend on λ. Model selection methods based on evaluation over a grid of (μ, λ) values (section 5) will make use of this fact. Indeed, step 3 only needs to be computed once for the entire set of (μ, λ) values.

Algorithm improvements The practical difference between PAM1 and PAM2 is that, in the latter, linear and nonlinear components span orthogonal spaces, they are computed independently, whereas in PAM1, steps 4 and 5 have to be iterated until convergence (step 6 of figure 1).

Independence of linear and nonlinear steps in PAM2 has an important implication for the model selection problem (section 5). As the model complexity is tuned by only two parameters, μ and λ, (no matter how many input variables there are in the data set), it is possible to perform regularization parameter selection by direct grid search optimization of a given criterion. Direct search methods solve unconstrained optimization problems without forming or estimating derivatives. They are based on evaluating a criterion over an admissible set of values, which may be a grid or random, and selecting the optimization parameter values that minimize this criterion. In the present case, independence of linear and nonlinear components implies that the grid of values is not quadratic (the number of μ–values $\times$ the number of λ–values) but linear (the number of μ–values + the number of λ–values). Model selection for the PAM1 algorithm is also performed by direct grid search optimization, but it involves much more calculations.

Adapting the algorithms for finding the lasso solution [29,14] to the estimation of the nonlinear components (step 5) would improve the computational efficiency. However, lasso–type problems are only a special case of the more general problems (9) or (12), for which no such algorithm is currently available.

An alternative to backfitting is to fit all the smooth components simultaneously, achievable using penalized regression splines [27,40,31]. This approach, which has been shown to be computationally efficient, can be integrated into our algorithm in a straightforward manner.

4.5 More General Models

Parsimonious additive models can be extended to generalized parsimonious additive models using an iteratively reweighted least–squares (IRLS) procedure to compute coefficients [23,24]. Thus we can solve the penalized problem by iterative application of the weighted version of algorithm in figures 1 and 2 within an IRLS loop.

This algorithm nevertheless presents new difficulties. First, the hat and smoother matrices depend on the weights, which change at each IRLS iteration. The singular value decomposition step (step 3) has to be incorporated within the iterated nonlinear coefficients estimation step (step $5(a)$). Secondly, the estimation of linear and nonlinear coefficients is not independent anymore for PAM2, since the two procedures interact via the weight matrices. Thus, complexity parameter selection based on direct grid search optimization (section 5) implies the evaluation of a quadratic, instead of a linear number of values. Consequently, computation becomes more intensive than in the Gaussian–type responses case.

4.6 Related Methods

The l_1–based penalizer is used in the context of linear [36,24], wavelet [11] and kernel [30,20] regressions. It has also been adapted to additive models fitted by cubic smoothing splines [18,3]. Nevertheless, as pointed out previously, selected variables are not eliminated, but linearized.

Our solution is close to the COSSO (COmponent Selection and Smoothing Operator), a general regularization scheme for smoothing spline ANOVA models, where the penalty functional is defined as the sum of component norms [26].

In the context of additive cubic smoothing splines, problem (6) is a special form of the COSSO, as it can rewritten as

$$\min_{(f_1,\ldots,f_p)\in\mathcal{H}_1\times\ldots\times\mathcal{H}_p} \left\| \mathbf{y} - \sum_{j=1}^{p} f_j(\mathbf{x}_j) \right\|_2^2 + \frac{\lambda}{p} \left(\sum_{j=1}^{p} \left\| D^2 f_j \right\|_{L_2} \right)^2 . \qquad (16)$$

In the present paper, we depart from this formulation, with the aim to encourage variable selection. This goal was also pursued by Lin and Zhang, who considered the space of univariate functions defined on the second order

Sobolev Hilbert space $W_2[0,1]$ endowed with the norm

$$\|f\|^2 = \left(\int_0^1 f(t)dt\right)^2 + \left(\int_0^1 f'(t)dt\right)^2 + \int_0^1 \left(f''(t)\right)^2 dt \ .$$

In this approach, only one penalization term is needed to penalize linear and nonlinear components [26].

PAM differs from the COSSO in the respect that the amount of regularization on the linear and nonlinear component is not tied by the definition of the regularization functional. PAM thus requires a second tuning parameter at the selection step, but this burden is compensated by the additional flexibility, which results in invariance with respect to global scale changes (where all explicative variables are multiplied by a constant). In this regard, PAM is similar to Likelihood Basis Pursuit (LBP) [41]. The main effect model in [41] is an additive model expressed as a linear combination of kernels, where two regularization parameters are provided: one for the parametric component and the other one for the nonparametric component.

PAM departs from LBP in the estimation of the nonparametric component. The estimate returned by LBP is a sparse expansion of kernel coefficients, but all variables are likely to contribute to the model, each one being represented by a small number of kernels. The regularizers of PAM aim at providing sparsity with respect to the number of variables entering the nonparametric expansion. When a significant nonlinear effect is detected, all $\boldsymbol{\beta}_j$ are non-zero: LBP favors compact representations and PAM favors interpretability.

5 Complexity Tuning

Model selection refers to the problem of selecting, among a class of models, the one that minimizes the prediction error. This task is difficult to implement for additive models in the form (5) since there are p complexity parameters. Our proposal requires only two complexity parameters to be tuned, whatever the number of variables in the model. The optimal (μ, λ) values are those that minimize prediction error. This error is unknown and has to be estimated.

A popular criterion for choosing complexity parameters is K–fold cross–validation (CV), which is an unbiased estimate of the expected prediction error [35]. This computer intensive technique uses all available examples as training and test examples. It mimics the use of training and test sets by repeatedly training the algorithm K times with (approximately) a fraction $1/K$ of training examples left out for testing purposes. If K equals the sample size, this is called "leave–one–out" cross–validation. Leave–one–out CV for

linear smoother operators admits an analytic formulation, then no resampling is required and computing time is accelerated [23].

6 Experiments

We evaluate the performance of our method by comparing it to BRUTO, an adaptive method for estimating an additive model using smoothing splines that combines backfitting and model selection, allowing a continuous regimen of fits for each term [23]. Model selection is based on an approximation to the GCV criterion, which is used to determine the λ_j parameters, one parameter at a time, at each step of the backfitting procedure. Once the selection process stops, the model is backfit using the chosen amount of smoothing. Simulations with BRUTO were carried out using the mda package of R 2.1.1. Results obtained by the constant model (CM), estimated by the mean response, are also provided as a reference.

Parsimonious additive models are computed by the two proposed algorithms: PAM1 and PAM2. Model selection for parsimonious additive models is carried out using CV criterias, which are evaluated over a 8×8 grid of (μ, λ) values regularly spaced on a logarithmic scale. For PAM1, model selection is performed by 5–fold CV. The analytical approximation of leave-one-out CV for linear smoothers turned out to perform well enough for PAM2, with the benefit of avoiding a great deal of calculations. The performance of CV is compared to the optimal performance, achieved by choosing the model that minimizes the test error (the *crystal ball model*[3], using Breiman's terminology [5]), and calculated over the same grid of values. Simulations with parsimonious additive models were carried out using Matlab 6.5.

6.1 Protocol

The synthetic data sets were randomly generated as follows. There are $p = 18$ explanatory variables identically distributed from the standard normal distribution, and 1 response variable. Explanatory variables are grouped in 6 clusters of 3 variables: $\boldsymbol{X}^k = (X_1^k, X_2^k, X_3^k)$, $k = 1, \ldots, 6$. The variables belonging to different clusters are independent, and the variables within each group are correlated: $\boldsymbol{X}^k \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Lambda})$, $\Lambda_{ij} = \rho^{|i-j|}$, where ρ is the parameter controlling correlation. Dealing with small clusters allows us, first, to control

[3] The *crystal ball method* consists in picking up the predictor having minimum prediction error. Without prior information, one uses for example, test sets.

correlation in a precise way and, secondly, to localize redundant information easily.

The underlying functions in each group are: $f_1(x_1^k) = x_1^k$, $f_2(x_2^k) = \cos(\frac{\pi}{2}x_2^k)$, $f_3(x_3^k) = \frac{1}{2}x_3^k + \frac{1}{2}\sin(\pi x_3^k)$, $k = 1, \ldots, 6$. Hence we take into account a wide range of functions with respect to their curvature. The response is calculated as

$$y = \sum_{k=1}^{6} \delta^k [f_1(x_1^k) + f_2(x_2^k) + f_3(x_3^k)] + \varepsilon, \tag{17}$$

where $\delta^k \in \{0, 1\}$ controls the relevance of cluster k and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. The noise level defines R^2.

We consider the following scenarios:

- Low ($\rho = 0.1$) and severe ($\rho = 0.9$) intra–clusters correlation,
- Few ($\delta^1 = \delta^2 = 1$, $\delta^3 = \delta^4 = \delta^5 = \delta^6 = 0$) and several ($\delta^1 = \delta^2 = \delta^3 = \delta^4 = \delta^5 = 1$, $\delta^6 = 0$) relevant variables ($d = 6$ among 18 and $d = 15$ among 18, respectively).
- Low ($R^2 = 0.95$) and moderate noise ($R^2 = 0.75$),
- Small ($n = 50$) and moderate ($n = 200$) sample size,

Table 1
Summary of analyzed scenarios as functions of control parameters.

Case	Correlation	Relevant variables	Noise	Observations
1	low ($\rho = 0.1$)	low ($d = 6$)	low ($R^2 = 0.95$)	low ($n = 50$)
2	low ($\rho = 0.1$)	low ($d = 6$)	low ($R^2 = 0.95$)	moderate ($n = 200$)
3	low ($\rho = 0.1$)	low ($d = 6$)	moderate ($R^2 = 0.75$)	low ($n = 50$)
4	low ($\rho = 0.1$)	low ($d = 6$)	moderate ($R^2 = 0.75$)	moderate ($n = 200$)
5	low ($\rho = 0.1$)	high ($d = 15$)	low ($R^2 = 0.95$)	low ($n = 50$)
6	low ($\rho = 0.1$)	high ($d = 15$)	low ($R^2 = 0.95$)	moderate ($n = 200$)
7	low ($\rho = 0.1$)	high ($d = 15$)	moderate ($R^2 = 0.75$)	low ($n = 50$)
8	low ($\rho = 0.1$)	high ($d = 15$)	moderate ($R^2 = 0.75$)	moderate ($n = 200$)
9	severe ($\rho = 0.9$)	low ($d = 6$)	low ($R^2 = 0.95$)	low ($n = 50$)
10	severe ($\rho = 0.9$)	low ($d = 6$)	low ($R^2 = 0.95$)	moderate ($n = 200$)
11	severe ($\rho = 0.9$)	low ($d = 6$)	moderate ($R^2 = 0.75$)	low ($n = 50$)
12	severe ($\rho = 0.9$)	low ($d = 6$)	moderate ($R^2 = 0.75$)	moderate ($n = 200$)
13	severe ($\rho = 0.9$)	high ($d = 15$)	low ($R^2 = 0.95$)	low ($n = 50$)
14	severe ($\rho = 0.9$)	high ($d = 15$)	low ($R^2 = 0.95$)	moderate ($n = 200$)
15	severe ($\rho = 0.9$)	high ($d = 15$)	moderate ($R^2 = 0.75$)	low ($n = 50$)
16	severe ($\rho = 0.9$)	high ($d = 15$)	moderate ($R^2 = 0.75$)	moderate ($n = 200$)

Table 1 shows a summary of the different scenarios studied here as functions of control parameters. For each one of the 16 scenarios, 50 experiments were conducted.

Table 2
Mean test error of BRUTO and parsimonious additive models (PAM1 and PAM2). The mean test error of the constant model (CM) and optimal parsimonious additive model (CB1 and CB2) are also given for reference. Values are means and standard deviations of prediction error over 50 simulations. For each scenario, when the PAM1 error is smaller than the BRUTO error, the PAM1 value is shown in bold type (and similar for PAM2). A significant difference (T–test, $p < 0.05$) between BRUTO and PAM1 or BRUTO and PAM2 is indicated by $*$.

Case	CM	CB1	PAM1	CB2	PAM2	BRUTO
1	3.581 (1.043)	0.315 (0.075)	**0.363** (0.108)*	0.488 (0.137)	0.722 (0.198)*	0.566 (0.237)
2	3.297 (0.436)	0.137 (0.012)	0.140 (0.013)*	0.175 (0.020)	0.209 (0.044)*	0.128 (0.012)
3	3.877 (1.032)	1.287 (0.188)	**1.531** (0.343)*	1.433 (0.227)	**1.559** (0.268)	1.629 (0.338)
4	3.864 (0.577)	0.767 (0.043)	0.787 (0.050)	0.814 (0.058)	0.826 (0.065)*	0.739 (0.048)
5	13.210 (4.221)	1.405 (0.396)	**1.783** (0.665)*	1.833 (0.328)	**2.473** (0.496)*	3.344 (0.714)
6	12.828 (2.172)	0.382 (0.038)	**0.382** (0.037)*	0.553 (0.073)	0.753 (0.175)*	0.416 (0.045)
7	13.871 (3.499)	3.897 (0.469)	**4.611** (0.891)*	4.130 (0.510)	**4.429** (0.589)*	5.818 (0.823)
8	14.249 (1.927)	2.112 (0.128)	**2.165** (0.161)*	2.308 (0.172)	**2.434** (0.218)	2.441 (0.241)
9	4.626 (1.148)	0.321 (0.050)	**0.367** (0.083)*	0.439 (0.088)	**0.526** (0.119)*	0.791 (0.293)
10	4.374 (0.627)	0.190 (0.017)	**0.193** (0.017)*	0.213 (0.023)	0.215 (0.023)*	0.194 (0.022)
11	5.867 (1.379)	1.537 (0.161)	**1.722** (0.258)*	1.591 (0.200)	**1.717** (0.254)*	2.240 (0.582)
12	5.185 (0.588)	1.160 (0.057)	**1.182** (0.067)*	1.180 (0.065)	**1.192** (0.071)*	1.240 (0.091)
13	15.297 (4.054)	1.306 (0.216)	**1.552** (0.401)*	1.509 (0.196)	**1.894** (0.286)*	2.695 (0.353)
14	15.450 (2.294)	0.544 (0.053)	**0.555** (0.059)*	0.656 (0.069)	0.723 (0.076)	0.704 (0.133)
15	18.236 (4.356)	4.394 (0.411)	**5.001** (0.688)*	4.500 (0.463)	**4.726** (0.524)*	6.274 (1.069)
16	17.550 (2.342)	3.112 (0.152)	**3.170** (0.184)*	3.187 (0.163)	**3.269** (0.168)*	3.564 (0.253)

6.2 Results

Table 2 shows the performances achieved by BRUTO (for which model selection is carried out by GCV), and parsimonious additive models (PAM1 and PAM2, for which model selection is carried out by 5-fold CV and leave–one–out CV, respectively). For reference, we also provide the mean test errors of the constant model (CM) and of the *crystal ball* models CB1 and CB2, that is the optimaly tuned models obtained respectively by PAM1 and PAM2. Prediction errors are estimated on a test set of size 10000 and performances are reported by the means and standard deviations of prediction error over 50 experiments.

PAM1 performs significatively better than BRUTO, except in cases 2 and 4. For the latter, the difference is not significative, and the former is the simpliest estimation problem, with low correlation, few relevant variables, low noise and the highest number of observations.

PAM2 achieves slightly lower results. BRUTO performs significantly better in five cases (1, 2, 4, 6, 10) out of 13 significant differences. The five situations where BRUTO wins are characterized by relatively easy setups: low noise, larger sample sizes and low correlation between covariates, or low noise, larger sample sizes and few relevant variables.

The results obtained by BRUTO are more variable than those obtained by PAM, especially when either the sample size is small, the number of relevant variables is high, variables are correlated or the noise level is high. We suspect that BRUTO unstability may be due to the local search of the model selection technique, where one λ_j parameter is optimized at a time.

The 5–fold CV model selection technique used with PAM1 is, in general, close to the optimal performance (reported as CB1). This is an indicator of the stability of the algorithm. The analytical leave–one–out CV used for PAM2 performs also well (the optimal performance is CB2), except for low sample size and little noise.

Finally, BRUTO executes faster than PAM. However, we note that the computing time of PAM does not depend on the number of relevant variables.

Concerning differences between the two proposed algorithms, the optimaly tuned PAM1 (CB1) is always better than CB2. This is explained by the fact that PAM1 explores a larger space. PAM2 is only better than PAM1 in very difficult situations with high noise and small sample sizes, but avoiding the exact CV procedure reduces drastically the computing time.

Table 3
Average number of total and irrelevant eliminated variables obtained by BRUTO and parsimonious additive models (PAM1 and PAM2).

Case	Total			Irrelevant		
	PAM1	PAM2	BRUTO	PAM1	PAM2	BRUTO
1	2.5	4.5	11.7	2.5	4.4	11.1
2	3.0	3.0	11.4	3.0	3.0	11.4
3	5.0	4.9	12.8	4.8	4.7	10.9
4	2.6	2.2	11.3	2.6	2.2	11.3
5	1.9	2.3	10.6	1.0	1.1	2.8
6	1.0	0.3	2.8	1.0	0.3	2.8
7	3.7	3.0	12.4	1.4	1.0	2.7
8	1.1	0.6	3.7	1.1	0.6	3.0
9	3.9	6.6	9.1	3.9	6.5	7.4
10	4.1	5.4	8.5	4.1	5.4	8.5
11	6.9	6.7	9.3	6.4	6.0	6.8
12	3.8	4.6	8.0	3.8	4.6	7.5
13	2.0	4.3	8.0	1.4	1.9	1.8
14	1.0	1.0	2.7	1.0	1.0	2.0
15	4.4	4.6	9.5	1.7	1.3	1.8
16	1.5	1.4	5.7	1.3	1.1	2.2

The average number of eliminated variables and the average number of eliminated irrelevant variables are presented in table 3. As a general rule, BRUTO discarded most of the irrelevant variables but it also eliminated desired variables. Conversely, PAM selected most relevant variables, but few irrelevant variables were eliminated. Analyzing PAM results in detail, we see that even if few irrelevant or redundant variables are eliminated, these variables are

severely penalized.

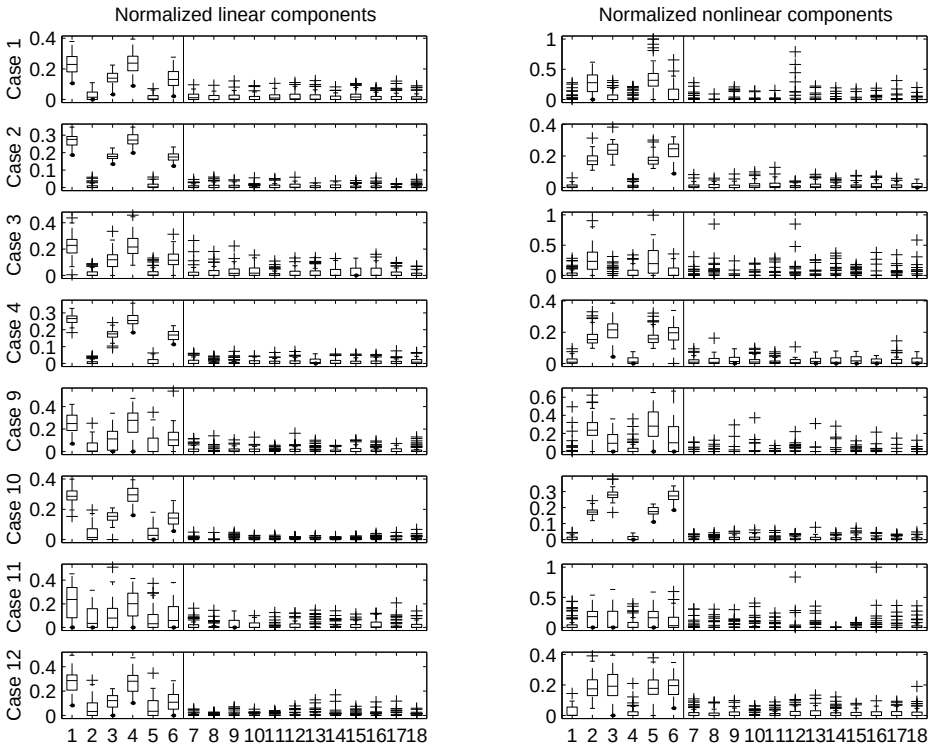

Fig. 3. Box plot summarizing the distribution (over 50 simulations) of normalized linear and nonlinear components: $\frac{|\alpha_j|}{\sum_{k=1}^{p}|\alpha_k|}$ and $\frac{\|D^2 f_j\|_{L_2}}{\sum_{k=1}^{p}\|D^2 f_k\|_{L_2}}$, respectively, obtained by PAM2, for all 18 input variables and for the eight scenarios corresponding to 6 relevant variables. Vertical lines separate relevant from irrelevant variables.

Figures 3 and 4 show box plots of normalized relevance index for linear components, $\frac{|\alpha_j|}{\sum_{k=1}^{p}|\alpha_k|}$ and normalized nonlinear components, $\frac{\|D^2 f_j\|_{L_2}}{\sum_{k=1}^{p}\|D^2 f_k\|_{L_2}}$, obtained by PAM2, for the 18 input variables. Figure 3 gathers all scenarios in which there are 6 relevant variables and figure 4 corresponds to the ones where there are 15 relevant variables. As pointed out above, many indexes corresponding to irrelevant variables ($j = 7, \ldots, 18$, in figure 3 and $j = 16, \ldots, 18$, in figure 4) are near–zeroes, but most of them are not exactly null.

We also observe that variables that are only linearly related to the response variable ($j = 1, 4$, in figure 3 and $j = 1, 4, 7, 10, 13$, in figure 4) present a severely penalized nonlinear components. Similarly, underlying functions without a linear effect ($j = 2, 5$, in figure 3, and $j = 2, 5, 8, 11, 14$, in figure 4) present a severely penalized linear component although their nonlinear component is important. Thus, linear and nonlinear trends are well identified.

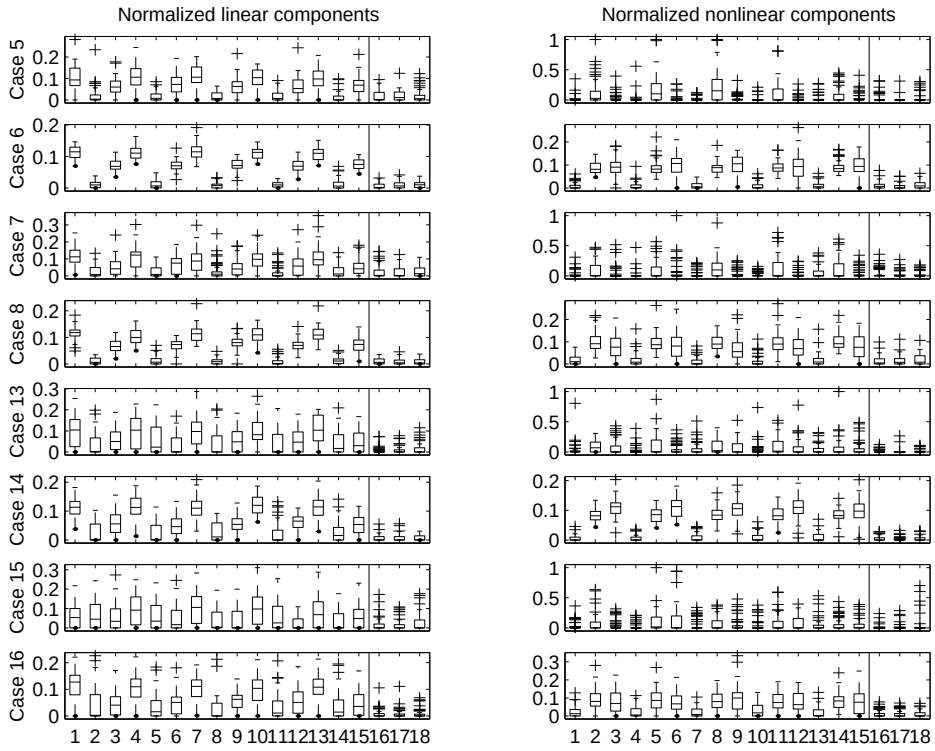

Fig. 4. Box plot summarizing the distribution (over 50 simulations) of normalized linear and nonlinear components: $\frac{|\alpha_j|}{\sum_{k=1}^{p}|\alpha_k|}$ and $\frac{\|D^2 f_j\|_{L_2}}{\sum_{k=1}^{p}\|D^2 f_k\|_{L_2}}$, respectively, obtained by PAM2, for all 18 input variables and for the eight scenarios corresponding to 15 relevant variables. Vertical lines separate relevant from irrelevant variables.

Alternative model selection criteria Inferring an estimator of the effective number of parameters, would allow us to adapt normal model selection criteria (as generalized cross–validation or Akaike information criteria) to tune the complexity of our problem. Orthogonality constraints of PAM2 justify calculation of the effective number of parameters as the effective number of parameters associated to linear components plus the effective number of parameters associated to nonlinear components. The estimator proposed by Fu [15] could be used to approximate the effective number of parameters associated to linear components. Another possibility consists in considering the problem under the adaptive ridge formulation and exploiting analogies between ridge and adaptive ridge regressions. The effective number of parameters associated to nonlinear components can be calculated as the addition of individual effective number of parameters [23]. This proposal must to be tested thoroughly, yet our experience has been very favourable.

7 Prediction of Indinavir Plasma Concentration

Pharmacokinetic characteristics (absorption, distribution and elimination) of certain antiretrovirals, and especially of protease inhibitors, are very variable. The concentration/effect (therapeutic or toxic) relationship is therefore a better indicator than the dose/effect relationship. The Cophar 1 ANRS 102 trial aims at establishing a window of efficacy and safe plasma concentrations for protease inhibitor treatments. We apply our approach to data concerning the protease inhibitor indinavir [6].

The data-set corresponds to 42 HIV–infected patients (one of them having missing values is excluded), described by several demographic and clinical characteristics: 1. gender (female/male); 2. age at examination in years; 3. weight in kg; 4. body mass index (BMI) in kg/m^2; 5. body surface area (BSA) in m^2; 6. number of lymphocytes CD4 cells/mm^3; 7. disease stage according to the CDC classification (1=no immunosuppression, 2=moderate immunosuppression and 3=severe immunosuppression); 8. duration of treatment (T) in months; 9. duration of indinavir treatment (IT) in months; 10. and 11. number of different molecules included in the antiretroviral multitherapy treatment (M1, ranged from 0 to 2, and M2, ranged from 0 to 1); 12. daily indinavir dose (dosage) in mg; 13. indinavir dose per unit intake in mg; and 14. absence/presence (0/1) of ritonavir treatment, that allows less frequent dosing of indinavir by slowing elimination. The Gaussian type response variable is the log plasma through concentration of indinavir (LPTCI), in ng/ml. We fitted LPTCI on standardized predictors using a parsimonious additive model (using PAM2). Categorical unordered covariates (gender and ritonavir) appear in the form of linear parametric functions and are penalized by only one constraint. Model selection was achieved by leave–one–out CV and 10–fold CV. Both criteria selected the same (μ, λ) values.

Figure 5 shows important effects of age, weight, dosage, BMI, IT and M1 (although the poor representativeness of one of the values for the later covariate may lead to overestimation). Weight and dosage have negative linear effects on the plasma concentration. The former is unsurprising, the latter can be explained by the fact that the smallest doses correspond to ritonavir treatments, and so they are accompanied by a slower elimination. The BMI covariates and the duration of indinavir treatment seem to have a quadratic effect on the LPTCI. The highest concentrations are found in the window of the standard BMI values, and the lowest concentrations are found in the overweight and underweight regions. Inversely, the LPTCI decreases slowly during the two first months of indinavir treatment, and increases afterwards. Concentration is initially a linear function of age, having a quadratic behavior that reaches its maxima between 45 and 55 years old.

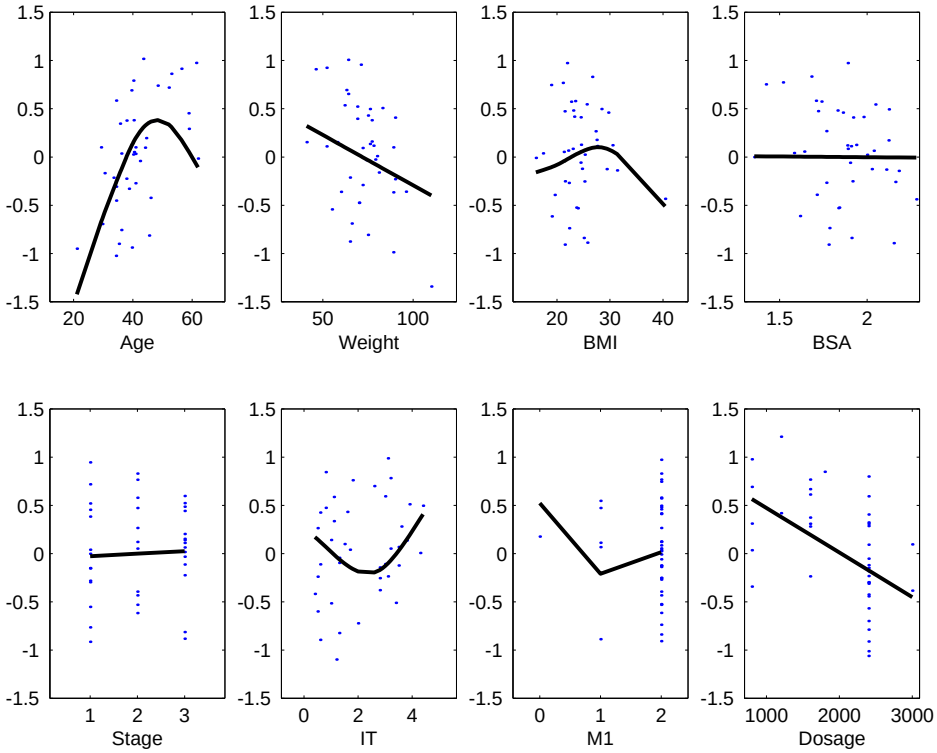

Fig. 5. Estimated univariate components for selected variables (solid curves) and partial residuals (dots).

The disease stage and BSA effects are very low. Indeed, the CDC classification is not always a good indicator of the severity of disease, since it does not take into account the regression of disease when the patient is undergoing treatment. As regards the effect of BSA, similar explanations can be given as for weight and BMI.

The dose, CD4, duration of treatment, M2, gender and ritonavir covariates are eliminated. Duration of treatment is, in part, explained by IT. Similarly, dosage is an important factor that depends on dose and ritonavir.

BRUTO was also applied to fit an additive model to the 14 predictors and to perform variable selection. Variables retained in the final model are age, which shows a quadratic effect, weight, BMI, BSA and dosage. Weight and dosage have negative linear effects on the plasma concentration, similar to the ones obtained with parsimonious additive models. However, the negative influence of weight is much more important according to BRUTO, while BMI and BSA have strong positive linear effects. In this respect, BRUTO differs significantly from PAM. One may be suspicious about BRUTO's results, since weight, BMI and BSA are highly positively correlated. The high negative influence of weight compensated by the strong positive effect of BSA is a further indication of the unstable behavior of BRUTO, which may result from the local search of complexity parameters, which are tuned one at a time.

Parsimonious additive models provided an adequate analysis of the concentrations of the Cophar 1 data set. On the one hand, they uncovered highly nonlinear effects on some of the covariates (BMI and IT), which would have been hard to detect by a parametric strategy. On the other hand, they could handle a small data set, where the relatively high number of covariates (with respect to the small sample size) makes difficult the use of standard nonparametric procedures, and where the high correlation between covariates may cause problems to local search strategies like BRUTO.

8 Conclusions

Additive and generalized additive models provide a flexible alternative to the standard linear and generalized linear models, preserving the ability to summarize relationships in an intuitive way. These models are thus applied in several domains including economics [34,4], engineering [39] and public health [2,13]. Most of these applications deal with few predictor variables. Additive models are seldom applied to variable selection problems, owing to the limitations of current methods.

In this paper we have proposed an extension of the lasso technique to additive models. We examined the relative merits of the adaptive backfitting procedure called BRUTO and parsimonious additive models PAM1 and PAM2 in 16 different scenarios.

BRUTO eliminates most irrelevant variables, but it may also discard significant variables. Conversely, parsimonious additive models select most relevant variables, but few irrelevant or redundant variables are eliminated. These variables are however severely penalized. In nonparametric additive regression, variable elimination may be more difficult, since it demands a zero coefficient for the linear component and a zero coefficient for its nonlinear component.

Globally, PAM1 performs (generally) better than PAM2, which is itself superior to BRUTO. The latter is however competitive in the easiest situation characterized by large sample sizes, low noise, low correlation between variables and/or few relevant variables. Parsimonious additive models are well adapted when model estimation is challenging. Hence, the results reported here suggest that our approach leads to promising techniques for the parsimonious additive modelling of the relationship between a response and several continuous covariates.

9 Acknowledgements

The authors wish to thank the Cophar 1–ANRS 102 study group for giving us access to data used in section 7, especially Dr C. Goujard, main investigator, and Dr A. M. Taburet, pharmacologic coordinator from the Bicêtre University Hospital. The authors are also indebted to the INSERM U738 research team (especially, F. Mentré et X. Panhard), for assistance in obtaining the data and interpreting results.

A Proof of equivalence

This appendix details the derivation between adaptive ridge and lasso [17].

Let L be any differentiable loss function (throughout this paper, L is the quadratic loss function). To simplify, suppose that the responses are centered. The adaptive ridge solution $\widehat{\boldsymbol{\alpha}} = (\widehat{\alpha}_1, \ldots, \widehat{\alpha}_p)$ is the minimizer of

$$\begin{cases} (\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\mu}}) = \arg\min_{\boldsymbol{\alpha},\boldsymbol{\mu}} L(\boldsymbol{\alpha}) + \sum_{j=1}^{p} \mu_j \alpha_j^2, \\ \text{subject to} \qquad \sum_{j=1}^{p} \frac{1}{\mu_j} = \frac{p}{\mu}, \qquad \mu_j > 0, \end{cases} \tag{A.1}$$

where $\mu \in]0, +\infty[$. The parameterization avoiding divergent solution is

$$\gamma_j = \sqrt{\frac{\mu_j}{\mu}} \alpha_j \qquad \text{and} \qquad c_j = \sqrt{\frac{\mu}{\mu_j}} \qquad \text{for } j = 1, \ldots, p \tag{A.2}$$

The optimization problem of adaptive ridge is then stated as

$$\begin{cases} (\widehat{\mathbf{c}}, \widehat{\boldsymbol{\gamma}}) = \arg\min_{\mathbf{c},\boldsymbol{\gamma}} L(\mathbf{c}, \boldsymbol{\gamma}) + \mu \sum_{j=1}^{p} \gamma_j^2, \\ \text{subject to} \qquad \sum_{j=1}^{p} c_j^2 = p, \qquad c_j \geq 0. \end{cases} \tag{A.3}$$

The corresponding Lagrangian $\mathcal{L}$ is

$$\mathcal{L}(\mathbf{c}, \boldsymbol{\gamma}) = L(\mathbf{c}, \boldsymbol{\gamma}) + \mu \sum_{j=1}^{p} \gamma_j^2 + \nu \left(\sum_{j=1}^{p} c_j^2 - p \right) - \boldsymbol{\xi}^t \mathbf{c}, \tag{A.4}$$

where ν and $\boldsymbol{\xi}$ are the Lagrange multipliers corresponding, respectively, to the equality and the positivity constraints on $\{c_j\}$. The normal equations for

(A.4) are thus

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial \boldsymbol{\gamma}} = \dfrac{\partial L(\mathbf{c}, \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} + 2\mu\boldsymbol{\gamma} \\ \dfrac{\partial \mathcal{L}}{\partial \mathbf{c}} = \dfrac{\partial L(\mathbf{c}, \boldsymbol{\gamma})}{\partial \mathbf{c}} + 2\nu\mathbf{c} - \boldsymbol{\xi}. \end{cases} \tag{A.5}$$

First we state a relation between the partial derivatives of L with respect to $\mathbf{c}$ and $\boldsymbol{\gamma}$. This relation stems from the relation $\alpha_j = c_j\gamma_j$:

$$\begin{cases} \dfrac{\partial L}{\partial \gamma_j} = c_j \dfrac{\partial L}{\partial \alpha_j} \\ \dfrac{\partial L}{\partial c_j} = \gamma_j \dfrac{\partial L}{\partial \alpha_j}. \end{cases} \tag{A.6}$$

For this system, we have

$$\gamma_j \frac{\partial L}{\partial \gamma_j} = c_j \frac{\partial L}{\partial c_j}. \tag{A.7}$$

This equation is used to derive a relationship between $\widehat{c}_j$ and $\widehat{\gamma}_j$, independently of L and the Lagrange multipliers:

$$\begin{cases} \mathrm{diag}(\widehat{\boldsymbol{\gamma}}) \dfrac{\partial \mathcal{L}}{\partial \boldsymbol{\gamma}} = \mathrm{diag}(\widehat{\boldsymbol{\gamma}}) \dfrac{\partial L(\mathbf{c}, \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}\Big|_{(\widehat{\mathbf{c}}, \widehat{\boldsymbol{\gamma}})} + 2\mu \mathrm{diag}(\widehat{\boldsymbol{\gamma}})\widehat{\boldsymbol{\gamma}} \\ \mathrm{diag}(\widehat{\mathbf{c}}) \dfrac{\partial \mathcal{L}}{\partial \mathbf{c}} = \mathrm{diag}(\widehat{\mathbf{c}}) \dfrac{\partial L(\mathbf{c}, \boldsymbol{\gamma})}{\partial \mathbf{c}}\Big|_{(\widehat{\mathbf{c}}, \widehat{\boldsymbol{\gamma}})} + 2\nu \mathrm{diag}(\widehat{\mathbf{c}})\widehat{\mathbf{c}} - \mathrm{diag}(\widehat{\mathbf{c}})\boldsymbol{\xi}. \end{cases} \tag{A.8}$$

Since a Lagrange multiplier is zero for inactive constraints, we have $\mathrm{diag}(\widehat{\mathbf{c}})\boldsymbol{\xi} = 0$. As (A.7) holds for $(\widehat{\mathbf{c}}, \widehat{\boldsymbol{\gamma}})$, and that optimality of $(\widehat{\mathbf{c}}, \widehat{\boldsymbol{\gamma}})$ implies $\dfrac{\partial \mathcal{L}}{\partial \boldsymbol{\gamma}} = \dfrac{\partial \mathcal{L}}{\partial \mathbf{c}} = \mathbf{0}$, then, from (A.8), we have

$$\widehat{c}_j^2 = \frac{\mu}{\nu}\widehat{\gamma}_j^2, \quad \forall j. \tag{A.9}$$

The equality constraint (A.3) on $\{c_j\}$ implies:

$$\widehat{c}_j = \frac{\sqrt{p}|\widehat{\gamma}_j|}{\sqrt{\sum_{k=1}^p \widehat{\gamma}_k^2}}, \quad \forall j. \tag{A.10}$$

We finally use this equation to give the optimality conditions as a function of the original variables $\widehat{\alpha}_j$. As $|\widehat{\alpha}_j| = \widehat{c}_j|\widehat{\gamma}_j|$, we have

$$|\widehat{\alpha}_j| = \frac{\sqrt{p}\widehat{\gamma}_j^2}{\sqrt{\sum_{k=1}^p \widehat{\gamma}_k^2}} \Rightarrow \frac{|\widehat{\alpha}_j|}{\sum_{k=1}^p |\widehat{\alpha}_k|} = \frac{\widehat{\gamma}_j^2}{\sum_{k=1}^p \widehat{\gamma}_k^2} \Leftrightarrow \widehat{c}_j^2 = \frac{p|\widehat{\alpha}_j|}{\sum_{k=1}^p |\widehat{\alpha}_k|}. \tag{A.11}$$

This value of $\widehat{c}_j$ is now plugged into the first equation of system (A.5) evaluated at $(\widehat{\mathbf{c}}, \widehat{\boldsymbol{\gamma}})$, using the first equation of system (A.6):

$$\widehat{c}_j = \frac{\partial L}{\partial \alpha_j}\Big|_{\widehat{\alpha}_j} + 2\mu\widehat{\gamma}_j = 0, \quad \forall j. \tag{A.12}$$

Therefore, either $\widehat{c}_j = \widehat{\gamma}_j = \widehat{\alpha}_j = 0$, either $\left.\frac{\partial L}{\partial \alpha_j}\right|_{\widehat{\alpha}_j} + 2\mu\frac{\widehat{\gamma}_j}{\widehat{c}_j} = 0$. From (A.11), $\widehat{\gamma}_j/\widehat{c}_j$ can be rewritten using $\boldsymbol{\alpha}$ as follows:

$$\begin{aligned}\frac{\widehat{\gamma}_j}{\widehat{c}_j} &= \widehat{\gamma}_j\widehat{c}_j\frac{1}{\widehat{c}_j^2} \\ &= \widehat{\alpha}_j\frac{\sum_{k=1}^{p}|\widehat{\alpha}_k|}{p|\widehat{\alpha}_j|} \\ &= \frac{1}{p}\text{sign}(\widehat{\alpha}_j)\sum_{k=1}^{p}|\widehat{\alpha}_k|.\end{aligned} \tag{A.13}$$

The optimality conditions are thus

$$\begin{cases}\left.\frac{\partial L}{\partial \alpha_j}\right|_{\widehat{\alpha}_j} + 2\frac{\mu}{p}\text{sign}(\widehat{\alpha}_j)\sum_{k=1}^{p}|\widehat{\alpha}_k| = 0, & \forall j, \\ \text{or} \quad \widehat{\alpha}_j = 0,\end{cases} \tag{A.14}$$

wich are recognized as the normal equations of

$$L(\boldsymbol{\alpha}) + \frac{\mu}{p}\left(\sum_{k=1}^{p}|\alpha_k|\right)^2. \tag{A.15}$$

This concludes the proof.

References

[1] Avalos, M., Grandvalet Y. and Ambroise C., 2003. Regularization methods for additive models. In: M. R. Berthold, H. J. Lenz, E. Bradley, R. Kruse, and C. Borgelt, (Ed.), *5th International Symposium on Intelligent Data Analysis.*, Springer, LNCS, 509–520.

[2] Bacchetti, P. and Quale, C., 2002. Generalized additive models with interval–censored data and time–varying covariates: application to human immunodeficiency virus infection in hemophiliacs. *Biometrics*, 58(2) 443–447.

[3] Bakin, S., 1999. *Adaptive Regression and Model Selection in Data Mining Problems*. PhD thesis, School of Mathematical Sciences, The Australian National University, Canberra.

[4] Beck, N. and Jackman S., 1998. Beyond linearity by default: Generalized additive models. *American Journal of Political Science*, 42, 596–627.

[5] Breiman, L., 1996. Heuristics of instability and stabilization in model selection. *Annals of Statistics*, 24(6) 2350–2383.

[6] Brendel, K., Legrand, M., Taburet, A. M., Baron, G., Goujard, C., Mentre, F. and Cophar 1–ANRS 102 Trial Group, 2005. Population pharmacokinetic analysis of indinavir in HIV–infected patient treated with a stable antiretroviral therapy. *Fundamental and Clinical Pharmacology*, 19(3) 373–383.

[7] Brumback, B. A., Ruppert, D. and Wand, M.P., 1999. Comment on "Variable selection and function estimation in additive nonparametric regression using a data–based prior" by Shively, T.S. and Khon, R. and Wood, S. *Journal of the American Statistical Association*, 94(447) 794–797.

[8] Cantoni, E. and Hastie, T.J., 2002. Degrees of freedom tests for smoothing splines. *Biometrika*, 89, 251–263.

[9] Chambers, J.M. and Hastie, T.J., 1993. *Statistical Models in S*. Computer Science Series. Chapman & Hall, London.

[10] Chen, R., Härdle, W., Linton, O.B. and Severance–Lossin, E., 1996. Nonparametric estimation of additive separable regression models. In: W. Härdle and M.G. Schimek (Ed.) *Statistical Theory and Computational Aspects of Smoothing: Proceedings of the COMPSTAT'94 Satellite Meeting*, Heidelberg, Physica–Verlag, 247–265.

[11] Chen, S., Donoho, D. and Saunders, M, 1995. Atomic decomposition by basis pursuit. Technical Report 479, Department of Statistics, Stanford University.

[12] Chen, Z., 1993. Fitting multivariate regression functions by interaction spline models. *J. R. Statist. Soc. B*, 55(2) 473–491.

[13] Dominici, F., McDermott, A., Zeger, S.L. and Samet, J.M., 2002. On the use of generalized additive models in time–series studies of air pollution and health. *American Journal of Epidemiology*, 156(3) 193–203.

[14] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R., 2004. Least angle regression. *Annals of Statistics*, 32(2) 407–499.

[15] Fu, W.J., 1998. Penalized regression: the bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7(3) 397–416.

[16] González–Manteiga, W., Quintela–del Río, A. and Vieu, P., 2002. A note on variable selection in nonparametric regression with dependent data. *Stat. Probab. Lett.*, 57(3) 259–268.

[17] Grandvalet, Y., 1998. Least absolute shrinkage is equivalent to quadratic penalization. In: L. Niklasson, M. Bodén and T. Ziemske (Ed.), *ICANN'98*, volume 1 of *Perspectives in Neural Computing*, Springer, 201–206.

[18] Grandvalet, Y. and Canu, S., 1998. Outcomes of the equivalence of adaptive ridge with least absolute shrinkage. In: M.S. Kearns, S.A. Solla, and D.A. Cohn (Ed.) *Advances in Neural Information Processing Systems 11*, MIT Press, 445–451.

[19] Gu, C. and Wahba, G., 1991. Minimizing GCV/GML scores with multiple smoothing parameters via the newton method. *SIAM J. Sci. Statist. Comput.*, 12, 383–398.

[20] Gunn, S.R. and Kandola, J.S., 2002. Structural modeling with sparse kernels. *Mach. Learning*, 10, 581–591.

[21] Guyon, I. and Elisseeff, A., 2003. An introduction to variable and feature selection. *Journal of Machine Learning Research, Special Issue on Variable/Feature Selection*, 3, 1157–1182.

[22] Härdle, W. and Korostelev, A., 1996. Search for significant variables in nonparametric additive regression. *Biometrika*, 83(3) 541–549.

[23] Hastie, T.J. and Tibshirani, R.J., 1990. *Generalized Additive Models*, volume 43 of *Monographs on Statistics and Applied Probability*. Chapman & Hall.

[24] Hastie, T.J., Tibshirani, R.J. and Friedman, J., 2001. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Springer Series in Statistics, Springer, New York.

[25] Hurvich, C.M., Simonoff, J.S. and Tsai, C.L., 1998. Smoothing parameter selection in non parametric regression using an improved akaike information criteria. *Journal of the Royal Statistical Society, B*, 60(2) 271–293.

[26] Lin, Y. and Zhang, H.H., 2002. Component selection and smoothing in smoothing spline analysis of variance models. Technical Report 1072r, University of Winconsin – Madison and North Carolina State University. Rev janvier 2003.

[27] Marx, B.D. and Eilers, P.H.C., 1998. Direct generalized additive modeling with penalized likelihood. *Computational Statistics & Data Analysis*, 28, 193–209.

[28] Opsomer, J.D. and Ruppert, D., 1998. A fully automated bandwidth selection method for fitting additive models. *J. Multivariate Analysis*, 73, 166–179.

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

[30] Roth, V., 2001. Sparse kernel regressors. In: G. Dorfner, H. Bischof and K. Hornik, (Ed.) *Artificial Neural Networks–ICANN 2001*, Springer, LNCS 2130, 339–346.

[31] Ruppert, D., Wand, M.P. and Carroll, R. J., 2003. *Semiparametric regression*, volume 12 of *Cambridge Series on Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

[32] Shi, P. and Tsai, C.-L., 1999. Semiparametric regression model selections. *Journal of Statist. Plann. Inference*, 77(1) 119–139.

[33] Shively, T.S., Khon, R. and Wood, S., 1999. Variable selection and function estimation in additive nonparametric regression using a data–based prior. *Journal of the American Statistical Association*, 94(447) 777–806.

[34] Smith, M. and Kohn, R., 1996. Nonparametric regression using bayesian variable selection. *Journal of Econometrics*, 75(2) 317–343.

[35] Stone, M., 1974. Cross–validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, B*, 36(1) 111–147.

[36] Tibshirani, R.J., 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, B*, 58(1) 267–288.

[37] Tibshirani, R.J. and Knight, K., 1999. The covariance inflation criterion for adaptive model selection. *Journal of the Royal Statistical Society, B*, 61(3) 529–546.

[38] Wahba, G., 1990. *Spline Models for Observational Data*. Number 59 in Regional Conference Series in Applied Mathematics, SIAM, Philadelphia, PA.

[39] Walker, E. and Wright, S.P., 2002. Comparing curves using additive models. *Journal of Quality Technology*, 34(1) 118–129.

[40] Wood, S.N., 2000. Modelling and smoothing parameter estimation with multiple quadratic penalties. *J. R. Statist. Soc. B*, 62(2) 413–428.

[41] Zhang, H.H., Wahba, G., Lin, Y., Voelker, M., Ferris, M. Klein, R. and Klein, B., 2004. *Variable Selection and Model Building via Likelihood Basis Pursuit*. *Journal of the American Statistical Association*, 99, 659–672.