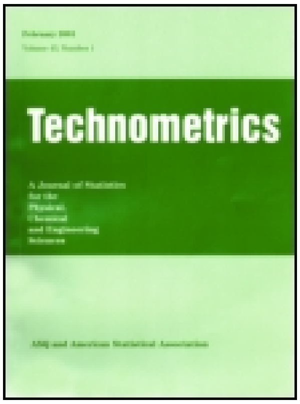

Technometrics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/utch20

Variable Selection in Additive Models Using P-Splines

Anestis Antoniadis [a] , Irène Gijbels [c] & Anneleen Verhasselt [b]

[a] Laboratoire Jean Kuntzmann, Université Joseph Fourier, Grenoble, France

[b] Department of Mathematics and Computer Science, University of Antwerp, Antwerpen, Belgium

[c] Department of Mathematics and Leuven Statistics Research Center (LStat), Katholieke Universiteit Leuven, Leuven, Belgium

Version of record first published: 28 Nov 2012.

To cite this article: Anestis Antoniadis , Irène Gijbels & Anneleen Verhasselt (2012): Variable Selection in Additive Models Using P-Splines, Technometrics, 54:4, 425-438

To link to this article: http://dx.doi.org/10.1080/00401706.2012.726000

Variable Selection in Additive Models Using P-Splines

Anestis ANTONIADIS

Laboratoire Jean Kuntzmann
Université Joseph Fourier
Grenoble, France
(Anestis.Antoniadis@imag.fr)

Irène GIJBELS

Department of Mathematics and Leuven
Statistics Research Center (LStat)
Katholieke Universiteit Leuven
Leuven, Belgium
(Irene.Gijbels@wis.kuleuven.be)

Anneleen VERHASSELT

Department of Mathematics and Computer Science
University of Antwerp
Antwerpen, Belgium
(Anneleen.Verhasselt@ua.ac.be)

This article extends the nonnegative garrote method to a component selection method in a nonparametric additive model in which each univariate function is estimated with P-splines. We also establish the consistency of the procedure. An advantage of P-splines is that the fitted function is represented in a rather small basis of B-splines. A numerical study illustrates the finite-sample performance of the method and includes a comparison with other methods. The nonnegative garrote method with P-splines has the advantage of being computationally fast and performs, with an appropriate parameter selection procedure implemented, overall very well. Real data analysis leads to interesting findings. Supplementary materials for this article (technical proofs, additional numerical results, R code) are available online.

KEY WORDS: Additive modeling; Nonnegative garrote; Nonparametric smoothing; Penalized spline regression; Selection of variables

1. INTRODUCTION

Regularization and variable selection techniques have become very popular for analyzing data. Consider, for example, data from a concrete slump flow test (Yeh 2007; Frank and Asuncion 2010) for measuring workability of fresh concrete. Nowadays high-performance concrete (HPC) can be made using the four basic ingredients in conventional concrete, that is, Portland cement, fine and coarse aggregates, and water; supplementary cementitious materials such as fly ash and blast furnace slag; and chemical admixtures such as superplasticizer. The quality of HPC is characterized by, among other things, high workability with good consistency. In the slump flow test, the workability is measured with a simple measure of plasticity of fresh concrete (see the online supplementary materials for a description). Modeling the influence of the ingredients on the slump flow of HPC is a difficult task. Yeh (2007) modeled the slump flow using parametric second-order regression modeling as well as modeling through artificial neural networks. Using the same data, Chien et al. (2010) employed a back-propagation network to estimate the slump flow. An advantage of artificial neural network type of modeling is the accuracy of the model fitting. Disadvantages are the overparameterization in networks and the instability of final parameter estimates.

Substantial improvements in adequate modeling can be achieved by using additive nonparametric models. These models often provide very flexible and effective descriptions of regression data where assumptions of parametric relationships are too restrictive. In an additive model, we have

$$Y_i = f_0 + \sum_{j=1}^{d} f_j(X_{ij}) + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{1}$$

with $(Y_i, X_{i1}, \ldots, X_{id})$ n independent and identically distributed (iid) observations from $(Y, X_1, \ldots, X_d)$, where Y is the response; $X_1, \ldots, X_d$ are the d explanatory variables; and ε is the random noise term with mean 0 and variance σ^2. Often only a few components f_j are different from 0 and thus important in explaining the response. Therefore we want to *select* and *estimate* the nonzero components and we want to do this in one single action.

In this article, we discuss a selection and estimation technique based on P-splines. For the HPC data example, the method reveals that the ingredients with the highest influence are water and slag. The influence of the ingredients fly ash and superplasticizer is far less pronounced. The most interesting finding concerns the influence of slag, which is clearly of a nonlinear form. See Figure 3 in Section 5.1.

Variable selection is an important issue in regression. When the explanatory variables in a regression model are assumed to

TECHNOMETRICS, NOVEMBER 2012, VOL. 54, NO. 4
DOI: 10.1080/00401706.2012.726000

have a linear effect, several penalized multiple linear regression methods that perform variable selection in a continuous fashion have been proposed in the literature. See, for example, Hastie, Tibshirani, and Friedman (2009). A nonparametric multiple regression model with a general smooth multivariate regression function relaxes the strong assumptions made by a linear model, but the statistical challenge is then how to reduce the dimension of the predictors in what would otherwise be a severely ill-posed problem. The additive combination of univariate functions proposed in model (1), while being less general than joint multivariate nonparametric models, is more interpretable and easier to fit. When performing variable selection in such additive models, one avoids testing and therefore this is a very useful technique especially when the number of covariates is large. Penalized multiple linear regression methods that encourage sparse models have been extended to nonparametric models (see, e.g., Antoniadis, Gijbels, and Nikolova 2011). Lin and Zhang (2006) introduced an extension of the least absolute shrinkage and selection operator (LASSO) for the additive model (1), resulting into the component selection and smoothing operator (COSSO) in the context of smoothing spline analysis of variance (ANOVA), for the case where the component functions f_j in (1) belong to a reproducing kernel Hilbert space (RKHS). They penalized the sum of the RKHS norms of the functional components. The ACOSSO (adaptive COSSO, Storlie et al., 2011) method adds a weight to the penalty for each component and therefore allows more flexible estimation of each component, in the same way, the adaptive LASSO (Zou 2006) is a weighted version of the LASSO for multiple linear regression models.

Breiman's (1995) nonnegative garrote method for multiple linear regression has been extended to perform variable selection in nonparametric additive models by Cantoni, Flemming, and Ronchetti (2011) and Yuan (2007). Cantoni, Flemming, and Ronchetti (2011) formulated the nonnegative garrote for additive models and proposed some methods to choose the smoothing parameters. The nonnegative garrote techniques accomplish such a selection by shrinking some of the additive components toward zero in addition to setting some of the components exactly equal to zero. For such models, an idea of the proof of the estimation and component selection consistency is given in the article by Yuan (2007). The nonnegative garrote uses an initial estimator and searches for shrinkage factors, which will lead to a sparser representation. Cantoni, Flemming, and Ronchetti (2011) and Yuan (2007) used smoothing splines as an initial estimator. We, on the other hand, will use P-splines [introduced by Eilers and Marx (1996)] as an initial estimator, extended to additive models by backfitting, to initially estimate the components f_j. We prove that the P-spline estimator combined with backfitting in additive models is consistent. The estimation and variable selection consistency of the nonnegative garrote with P-splines then follows from an application of the consistency result of Yuan (2007). We are thus able to provide full theoretical support for the selection procedure based on P-splines. In addition, several nonnegative garrote data-driven choices of the smoothing and regularization parameters were implemented. For brevity of presentation, we only report on an implemented L-curve selection procedure.

The article is organized as follows. In Section 2, we introduce the nonnegative garrote and the functional nonnegative garrote (i.e., extended to the additive model context). Section 3 is devoted to the consistency of the nonnegative garrote with P-splines (and backfitting). The details of the proofs are given in Section 1 in the online supplementary materials, which also discuss an additional Additive P-spline Selection Operator. We investigate the finite-sample performance of the method in a simulation study, which also includes comparisons with other methods in Section 4. In Section 5, the methods are applied to some real datasets. Some further discussions and conclusions are given in Section 6.

2. NONNEGATIVE GARROTE

2.1 Original Nonnegative Garrote

Breiman (1995) proposed the nonnegative garrote for subset regression in a multiple linear regression model. It starts from the ordinary least squares (OLS) estimator as an initial estimator and it shrinks or puts some coefficients of the OLS equal to zero. The data $(Y_i, X_{i1}, \ldots, X_{id})$ for $i = 1, \ldots, n$, come from a multiple linear regression model

$$Y_i = \beta_0 + \sum_{j=1}^{d} \beta_j X_{ij} + \varepsilon_i, \quad i = 1, \ldots, n.$$

The nonnegative garrote shrinkage factors $\hat{c}_j$ are found by solving

$$\begin{cases} (\hat{c}_1, \ldots, \hat{c}_d) = \arg\min_{c_1,\ldots,c_d} \\ \qquad \dfrac{1}{2}\displaystyle\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0^{\mathrm{OLS}} - \sum_{j=1}^{d} c_j \hat{\beta}_j^{\mathrm{OLS}} X_{ij}\right)^2 \\ \text{where } 0 \le c_j \ (j = 1, \ldots, d), \quad \displaystyle\sum_{j=1}^{d} c_j \le s, \end{cases}$$

for given s, or equivalently

$$\begin{cases} (\hat{c}_1, \ldots, \hat{c}_d) = \arg\min_{c_1,\ldots,c_d} \\ \left\{ \dfrac{1}{2}\displaystyle\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0^{\mathrm{OLS}} - \sum_{j=1}^{d} c_j \hat{\beta}_j^{\mathrm{OLS}} X_{ij}\right)^2 + \theta \sum_{j=1}^{d} c_j \right\} \\ \text{where } 0 \le c_j \ (j = 1, \ldots, d), \end{cases}$$

for given θ, where $\hat{\beta}_j^{\mathrm{OLS}}$ is the OLS estimator of the regression coefficient of the jth component and $s > 0$ and $\theta > 0$ are regularization parameters. See Xiong (2010) for recent work on the selection of θ. The nonnegative garrote estimator of the regression coefficient is then

$$\hat{\beta}_j^{\mathrm{NNG}} = \hat{c}_j \hat{\beta}_j^{\mathrm{OLS}}.$$

In the special case of an orthogonal design, that is, $\mathbf{X}'\mathbf{X} = \mathbf{I}_n$ (with $\mathbf{I}_n$ the $n \times n$ identity matrix), the nonnegative garrote estimates are

$$\hat{c}_j = \left(1 - \frac{\theta}{(\hat{\beta}_j^{\mathrm{OLS}})^2}\right)_+ ,$$

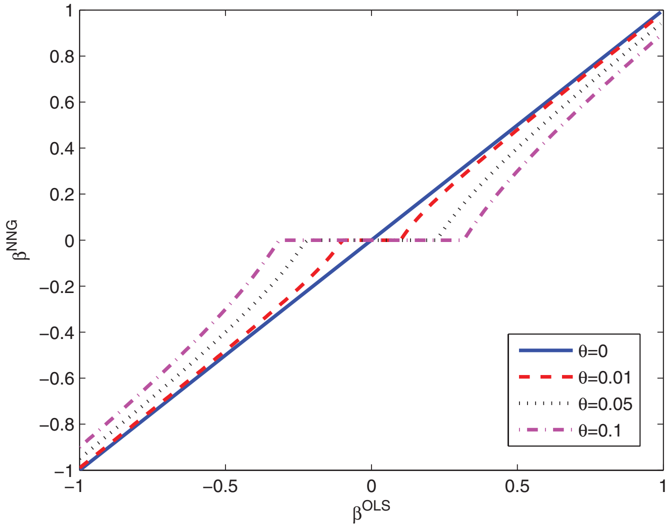

Figure 1. Shrinkage effect of the nonnegative garrote for different θ's. The online version of this figure is in color.

where $z_+ = \max(z, 0)$. The effect of the nonnegative garrote method is thus to "shrink" or to "put to zero" an initial regression coefficient, a property that the estimator shares with a soft-thresholded estimator of the initial coefficients. The nonnegative garrote estimate is presented in Figure 1 for different values of θ. The larger the θ, the stronger the shrinkage effect.

2.2 Functional Nonnegative Garrote

The nonnegative garrote of Breiman (1995) was extended to additive model (1) by Cantoni, Flemming, and Ronchetti (2011) and Yuan (2007). In this functional nonnegative garrote, one starts with an initial estimator $\hat{f}_j^{\text{init}}(X_j)$ for the function describing the jth component, which replaces $\hat{\beta}_j^{\text{OLS}} X_j$ in the original nonnegative garrote. The nonnegative garrote shrinkage factors are then found by solving

$$\begin{cases} \min_{c_1,\ldots,c_d} \left\{ \dfrac{1}{2} \displaystyle\sum_{i=1}^{n} \left(Y_i - \hat{f}_0^{\text{init}} - \sum_{j=1}^{d} c_j \hat{f}_j^{\text{init}}(X_{ij}) \right)^2 \right. \\ \qquad \left. + \, \theta \displaystyle\sum_{j=1}^{d} c_j \right\} \\ \text{where } 0 \le c_j \ (j = 1, \ldots, d). \end{cases} \tag{2}$$

The resulting nonnegative garrote estimate of the jth component is then given by

$$\hat{f}_j^{\text{NNG}}(\cdot) = \hat{c}_j \hat{f}_j^{\text{init}}(\cdot).$$

Cantoni, Flemming, and Ronchetti (2011) compared the nonnegative garrote with smoothing splines with COSSO on different simulated and real datasets. They also compared different algorithms for the initial smoothing spline fit. We will compare the effect of changing the basis functions, that is, using P-splines or smoothing splines. Moreover, in the above-mentioned article, there are no theoretical results. On the other hand, using P-splines, we are able to obtain theoretical results. In addition, the advantage of using P-splines is that they are understandable for any readership and are an efficient penalization method that is widely used.

3. CONSISTENCY OF NONNEGATIVE GARROTE WITH P-SPLINES

P-splines were first introduced by Eilers and Marx (1996) in the univariate nonparametric smoothing context

$$Y_i = f(X_i) + \varepsilon_i \quad \text{for } i = 1, \ldots, n,$$

where the ε_i's are iid zero mean random variables with finite variance σ^2. P-splines are an extension of regression splines with a penalty on the coefficients of adjacent B-splines. Suppose that we have data (X_i, Y_i), for $i = 1, \ldots, n$, with $X_i \in [a, b] \subset I\!R$. Without loss of generality, we will assume that $[a, b] = [0, 1]$. To estimate $f(\cdot)$, we use a regression spline model $f(x) = \sum_{j=1}^{m} \alpha_j B_j(x; q)$, where $\{B_j(\cdot; q) : j = 1, \ldots, K + q = m\}$ is the qth degree B-spline basis, using normalized B-splines such that $\sum_j B_j(x; q) = 1$, with $K + 1$ equidistant knot points $t_0 = 0, t_1 = \frac{1}{K}, \ldots, t_K = 1$ in $[0, 1]$ and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)'$ is the unknown column vector of regression coefficients. The penalized least squares estimator $\hat{\alpha}$ is the minimizer of

$$S(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{m} \alpha_j B_j(x_i; q) \right)^2 + \lambda \sum_{j=k+1}^{m} (\Delta^k \alpha_j)^2, \tag{3}$$

where $\lambda > 0$ is the smoothing parameter and Δ the differencing operator, that is, $\Delta^k \alpha_j = \sum_{t=0}^{k} (-1)^t \binom{k}{t} \alpha_{j-t}$, with $k \in I\!N$. In particular, for $k = 1$ and $k = 2$, this is $\Delta^1 \alpha_j = \alpha_j - \alpha_{j-1}$ and $\Delta^2 \alpha_j = \alpha_j - 2\alpha_{j-1} + \alpha_{j-2}$, respectively. Rewriting (3) in matrix notation, we obtain

$$S(\boldsymbol{\alpha}) = (\mathbf{Y} - \mathbf{B}\boldsymbol{\alpha})'(\mathbf{Y} - \mathbf{B}\boldsymbol{\alpha}) + \lambda \boldsymbol{\alpha}' \mathbf{D}_k' \mathbf{D}_k \boldsymbol{\alpha},$$

where the elements B_{ij} of $\mathbf{B}$ ($\in I\!R^{n \times m}$) are $B_j(x_i; q)$ and $\mathbf{D}_k$ ($\in I\!R^{(m-k) \times m}$) is the matrix representation of the kth-order differencing operator Δ^k.

The consistency of the nonnegative garrote with P-splines (the proposed method) is based on three key ingredients: (1) on a consistency result for (univariate) P-splines, (2) on an extension of a univariate smoothing estimator to additive models via backfitting, and (3) on a consistency result for the functional nonnegative garrote.

Our theoretical contributions are in Theorem 2 (consistency of each univariate P-spline estimator in additive models) and Theorem 3, which establishes the consistency result for the nonnegative garrote with P-splines.

Claeskens, Krivobokova, and Opsomer (2009) showed the consistency of the univariate P-spline for deterministic design (see Theorem 1). The fact that the rate of convergence for a univariate P-spline estimator combined with backfitting in an additive model is the same as the rate for the univariate estimator is established in Theorem 2, the proof of this relies on a result of Horowitz, Klemelä, and Mammen (2006). Finally Yuan (2007) proved that, given an initial estimator that is consistent, the nonnegative garrote is variable selection and estimation consistent. This result combined with the established consistency results in the additive modeling context allows us to prove the variable selection and estimation consistency of the proposed procedure in Theorem 3. The various essential parts are discussed in the next subsections.

3.1 Consistency of P-Spline Estimator

We combine the univariate P-spline estimator with a backfitting algorithm to use the P-splines in the multivariate setting and to find an initial P-spline estimator for each component f_j.

In the sequel, we denote by $\mathbf{f}_j$ the column vector of dimension $n \times 1$ containing the function f_j evaluated in the observed points $X_{1j}, \ldots, X_{nj}$, that is, $\mathbf{f}_j = (f_j(X_{1j}), \ldots, f_j(X_{nj}))'$. Similarly $\hat{\mathbf{f}}_j$ denotes the column vector of the estimate $\hat{f}_j$ calculated in the n X_j-observations. The L_2-norm of $\mathbf{f}_j$ equals $\|\mathbf{f}_j\|_2 = \sqrt{\sum_{i=1}^n f_j^2(X_{ij})}$ and similarly for the L_2-norm of $\hat{\mathbf{f}}_j$. Furthermore, a vector of zeros is denoted by $\mathbf{0}$ and the vector of Y-observations is $\mathbf{Y} = (Y_1, \ldots, Y_n)'$. With these notations, the backfitting algorithm is as follows:

1. Set $\hat{f}_0 = \bar{Y} = n^{-1} \sum_{i=1}^n Y_i$ and $\hat{\mathbf{f}}_j = \mathbf{0}$ for $j = 1, \ldots, d$.
2. Repeat Steps 3–5 until the estimates $\hat{\mathbf{f}}_j$ stop changing, that is, if the relative error $\frac{\|\hat{\mathbf{f}}_j^{(\iota+1)} - \hat{\mathbf{f}}_j^{(\iota)}\|_2}{\|\hat{\mathbf{f}}_j^{(\iota)}\|_2}$ is smaller than a certain precision ϵ (where $\hat{\mathbf{f}}_j^{(\iota)}$ is the estimate at the ιth iteration).
3. For $j = 1, \ldots, d$, repeat Steps 4 and 5.
4. Calculate the partial residuals $\mathbf{e}_j = \mathbf{Y} - \sum_{h \neq j} \hat{\mathbf{f}}_h$.
5. Set $\hat{\mathbf{f}}_j$ equal to the result of smoothing $\mathbf{e}_j$ with respect to X_j (using univariate P-splines).

Suppose that $f(\cdot) \in C^{q+1}([0, 1])$ is a $q + 1$ times continuously differentiable function on [0, 1]. The asymptotic properties of the P-spline estimator are considered under the following assumptions.

Assumption 1.

1. For deterministic points $x_i \in [0, 1]$, $i = 1, \ldots, n$, assume that there exists a distribution function Q with corresponding positive continuous design density p such that, with Q_n the empirical distribution of $x_1, \ldots, x_n$, $\sup_{x \in [a,b]} |Q_n(x) - Q(x)| = o(1/K)$.
2. $K = o(n)$.
3. $k \leq 2q - 1$.

Assumption 1.1 is classical in nonparametric regression (see, e.g., Eubank and Speckman 1990). The other two assumptions are natural when using P-splines modeling (see Eilers and Marx 1996).

Since the following theorem holds for deterministic design, we will condition on $X_1, \ldots, X_d$ in the sequel. However, another possibility is to consider a random design which is often the case in observational studies. Under appropriate extra conditions on the distribution of the X_i's (see Assumption 2.2 in the next subsection) using arguments analogous to those in the proof of Theorem 1 and the boundedness of the X_i's, a similar result holds in the random design case.

The consistency of the (univariate) P-spline is a result of Claeskens, Krivobokova, and Opsomer (2009), stated in the following theorem. Herein the column vector $(f(x_1), \ldots, f(x_n))'$ is denoted by $\mathbf{f}$ and the column vector of the P-spline estimate $\hat{f}(\cdot) = \sum_{j=1}^m \hat{\alpha}_j B_j(\cdot; q)$ evaluated at $(x_1, \ldots, x_n)$ as $\hat{\mathbf{f}}$.

Theorem 1. (Claeskens, Krivobokova, and Opsomer 2009) Under Assumptions 1.1–1.3 and with $K = O(n^{\frac{1}{2q+3}})$ and $\lambda = O(n^{\chi})$ with $\chi \leq \frac{q+2+k}{2q+3}$ and AMSE$(\hat{\mathbf{f}}|X_1 = x_1, \ldots, X_n = x_n)$ defined by $\frac{1}{n}\mathrm{E}((\hat{\mathbf{f}} - \mathbf{f})'(\hat{\mathbf{f}} - \mathbf{f})|X_1 = x_1, \ldots, X_n = x_n)$, we have that

$$\begin{aligned}\mathrm{AMSE}(\hat{\mathbf{f}}|X_1 = x_1, \ldots, X_n = x_n)\\ = \frac{1}{n}\sum_{i=1}^n \mathrm{E}((\hat{f}(x_i) - f(x_i))^2 = O\big(n^{\frac{-2(q+1)}{2q+3}}\big).\end{aligned}$$

Forthwith, we adopt these rates for λ and K.

3.2 Consistency of Univariate Smoother in Additive Model

The consistency of the P-spline with backfitting relies on a result of Horowitz, Klemelä, and Mammen (2006) for linear smoothers. They proved that each additive component can be estimated as well as it could be if the other components were known (i.e., the asymptotic rates for a linear smoother in additive models are the same as in a nonparametric regression with one component).

We consider model (1) with covariates $\mathbf{X}_i = (X_{i1}, \ldots, X_{id})$ random and iid (with density p) in a bounded interval, assume without loss of generality in [0, 1], and error variable ε_i iid and independent of $X_{i1}, \ldots, X_{id}$, with $\mathrm{E}(\varepsilon_i) = 0$ and $\mathrm{var}(\varepsilon_i) = \sigma^2$. The consistency result of Horowitz, Klemelä, and Mammen (2006) holds also for nonrandom design that approximately follows a smooth design density p (see Assumption 1.1).

For identifiability, we assume that

$$\mathrm{E}(f_j(X_j)) = 0 \tag{4}$$

for $j = 1, \ldots, d$. We consider estimation of f_1 (without loss of generality) and prove that asymptotically f_1 can be estimated as well for $d > 1$ as for $d = 1$. Therefore we consider a model with $d > 1$ and suppose that $f_2, \ldots, f_d$ are known. Then, following the strategy of Horowitz, Klemelä, and Mammen (2006), an estimator $\hat{f}_1$ of f_1 can be constructed using the pseudo data (X_{i1}, Z_i) with

$$Z_i = Y_i - f_2(X_{i2}) - \cdots - f_d(X_{id}) = f_0 + f_1(X_{i1}) + \varepsilon_i. \tag{5}$$

The Z_i are replaced by local averages $\widehat{Z}_i$ (a presmoother), which are found by using a regular regression with B-splines of degree 0 and with $L_n + 1$ equidistant knots ($L_n + 1$ will be larger than the number of knots $K + 1$ of the univariate P-spline smoother). The smoothing of Z_i and $\widehat{Z}_i$ leads to asymptotically equivalent estimators.

The presmoother $\widehat{Y}_i$ is found by a regular regression with B-splines of degree 0 and $L_n + 1$ knots, where the regression coefficients $\widehat{\alpha}_{jh}$ are found by minimizing

$$\sum_{i=1}^n \left(Y_i - \bar{Y} - \sum_{j=1}^d \sum_{h=1}^{L_n} \alpha_{jh} B_h(X_{ij}; 0) \right)^2$$

with respect to α_{jh}. Moreover the presmoother $\widehat{Y}_i$ and $\widehat{Z}_i$ differ only in higher-order terms.

Horowitz, Klemelä, and Mammen (2006) considered a linear smoother. Note that our univariate P-spline estimator is also a linear smoother with

$$\begin{aligned}\hat{f}_1(x) &= \sum_{i=1}^{n} (\mathbf{B}(x)(\mathbf{B}'\mathbf{B} + \lambda \mathbf{D}_k'\mathbf{D}_k)^{-1}\mathbf{B}')_i (Z_i - \bar{Z}) \\ &= \frac{1}{n}\sum_{i=1}^{n} w_i(x)(Z_i - \bar{Z}), \qquad (6)\end{aligned}$$

where $\mathbf{B}(x)$ is a vector with the B-spline basis functions evaluated at x. This estimator is compared with the following P-splines estimator (which smooths the presmoothed data $\hat{Y}_i$)

$$\tilde{f}_1(x) = \frac{1}{n}\sum_{i=1}^{n} w_i(x)\hat{Y}_i. \qquad (7)$$

From Theorem 2, we will have that the difference between $\hat{f}_1(x)$ and $\tilde{f}_1(x)$ is asymptotically negligible relative to the difference between $\hat{f}_1$ and f_1 under the following assumptions:

Assumption 2.

1. Let $c_1\ n^{\frac{q+1}{2q+3}} \le L_n \le c_2\ n^{\frac{q+1}{2q+3}}$, for some positive c_1 and c_2 (with $c_1 \le c_2$).
2. All covariates take values in a bounded interval, [0, 1], say. The one- and two-dimensional marginal densities p_j and $p_{j,l}$ of X_{ij} and (X_{ij}, X_{il}), respectively, are bounded away from zero and infinity.
3. For all additive components f_j, the first derivative exists almost surely and satisfies $\int f_j'(x)^2\ dx < \infty$ for $j = 1, \dots, d$.
4. There is a finite constant $C > 0$ such that $|f_j(x) - f_j(y)| \le C|x - y|$ for each $j = 1, \dots, d$. Moreover $\mathrm{E}(|\varepsilon_i|^{2+\delta}) < \infty$ with $\delta = (q + 1)^{-1}$.
5. $k \le q - 1$.
6. $\lambda \frac{K}{n} \to 0$ for $n \to \infty$.
7. $\frac{K}{n} \to$ constant.
8. For $j = 1, \dots, d$, $\sup_{i,i'} |X_{ij} - X_{i'j}| < \xi$ (with $\xi \le \frac{1}{q+1} < 1$) holds with probability tending to 1 (i.e., $\lim_{n\to\infty} P(\sup_{i,i'} |X_{ij} - X_{i'j}| < \xi) = 1$).

Note that Assumption 2.1 is an assumption on the construction of the presmoothers $\hat{Z}_i$ and $\hat{Y}_i$. The boundedness of the covariates is a natural condition in practice. Assumptions 2.3 and 2.4 are conditions on the smoothness of the underlying functions. The other assumptions concern the degree, knots, smoothing parameter, and differencing order of the P-splines and are needed to satisfy the conditions on $w_i(x)$ in the theorem of Horowitz, Klemelä, and Mammen (2006). Note that Assumptions 2.6 and 2.7 are satisfied with the rates for λ and K of Theorem 1. It is easy to check that, for example, the models for Examples 1 and 3 in Section 4 satisfy the above assumptions.

Theorem 2 of Horowitz, Klemelä, and Mammen (2006) is applied to the P-spline estimator results in the following theorem. The proof of this theorem is provided in Section 1.1 in Part I of the online supplementary materials.

Theorem 2. Suppose that (1) holds with (4) and let Z_i be as defined in (5), $\hat{f}_1$ as in (6), and $\tilde{f}_1$ as in (7).

1. Under Assumptions 2.1–2.3 and 2.5–2.8

$$\int_0^1 (\hat{f}_1(x) - \tilde{f}_1(x))^2\ dx = O_P\big(n^{-\frac{2q+2}{2q+3}}\big).$$

2. Under Assumptions 2.1, 2.2, and 2.4–2.8

$$\sup_{0\le x\le 1} |\hat{f}_1(x) - \tilde{f}_1(x)| = O_P\big(n^{-\frac{q+1}{2q+3}}\big).$$

Therefore we have that

$$\sup_{0\le x_{i1}\le 1} (\widehat{f}_1(x_{i1}) - \widetilde{f}_1(x_{i1}))^2 = O_P\big(n^{\frac{-2(q+1)}{2q+3}}\big)$$

$$\frac{1}{n}\sum_{i=1}^{n}(\widehat{f}_1(x_{i1}) - \widetilde{f}_1(x_{i1}))^2 = \frac{1}{n}||\widehat{\mathbf{f}}_1 - \widetilde{\mathbf{f}}_1||_2^2 = O_P\big(n^{\frac{-2(q+1)}{2q+3}}\big).$$

Then (by the triangle inequality)

$$\begin{aligned}\frac{1}{\sqrt{n}}||\widetilde{\mathbf{f}}_1 - \mathbf{f}_1||_2 &\le \frac{1}{\sqrt{n}}||\widehat{\mathbf{f}}_1 - \widetilde{\mathbf{f}}_1||_2 + \frac{1}{\sqrt{n}}||\widehat{\mathbf{f}}_1 - \mathbf{f}_1||_2 \\ &= O_P(n^{\frac{-(q+1)}{2q+3}})\end{aligned}$$

and thus $\frac{1}{n}||\widetilde{\mathbf{f}}_1 - \mathbf{f}_1||_2^2 = O_P(n^{\frac{-2(q+1)}{2q+3}})$.

This $\widetilde{f}_j$, which is a P-spline estimator combined with backfitting, will be the initial estimator $\widehat{f}_j^{\mathrm{init}}$ in our nonnegative garrote.

3.3 Consistency of Nonnegative Garrote With P-Splines

In this section, we consider optimization problem (2). In the following, we condition on $X_{ij} = x_{ij}$. From Theorem 2 and the above calculation, we have that the initial P-spline estimator is consistent in L_2-norm, that is,

$$\frac{1}{n}||\widetilde{\mathbf{f}}_1 - \mathbf{f}_1||_2^2 = O_P\big(n^{\frac{-2(q+1)}{2q+3}}\big) = O_P\big(\kappa_n^2\big),$$

with $\kappa_n = n^{\frac{-(q+1)}{2q+3}}$.

The following theorem (for a proof, see Section 1.2 in Part I of the online supplementary materials), which is a combination of theorem 2 of Yuan (2007) and the consistency of the backfitted P-spline, establishes the consistency of the nonnegative garrote estimator with P-splines under Assumption 3.

Assumption 3.

1. $||\mathbf{f}_j||_2 < \infty$.
2. $||(\frac{1}{n}\mathbf{f}_{\cdot 1}'\mathbf{f}_{\cdot 1})^{-1}||_2 < \infty$, where $\mathbf{f} = (\mathbf{f}_1, \dots, \mathbf{f}_d)$ is a matrix and $\mathbf{f}_{\cdot 1}$ are the columns of $\mathbf{f}$ containing the nonzero components.
3. $\varepsilon_i = O_P(1)$.

Theorem 3. Assume that Assumptions 1–3 hold. If $\frac{\theta}{n}$ tends to 0 such that $\kappa_n = o(\frac{\theta}{n})$, then (given $X_{ij} = x_{ij}$)

1. $P(\hat{\mathbf{f}}_j^{\mathrm{NNG}} = \mathbf{0}) \to 1$ for any j such that $f_j = 0$,
2. $\sup_j \frac{1}{n}||\mathbf{f}_j - \hat{\mathbf{f}}_j^{\mathrm{NNG}}||_2^2 = O_P((\frac{\theta}{n})^2)$.

This theorem states that the nonnegative garrote with P-splines is variable selection consistent (1) and estimation consistent (2). Thus it has the desired property that it tends to estimate a true-zero coefficient as a zero function.

Table 1. Methods, codes, and implementations

Method	Codes	Parameter selection procedures
NGP	Custom MATLAB code, using lsqlin.m (constrained linear least squares)	θ in NG with L-curve
NGSM	Custom MATLAB code for the nonnegative garrote part, and custom R-code, using, for example, the gam function of the mgcv R-package	θ in NG with L-curve
COSSO	ACOSSO-code of Storlie et al. (2011)	BIC and 5-CV
ACOSSO	Idem (see *www.math.unm.edu/∼storlie/acosso/acosso.R*)	BIC and 5-CV

4. NUMERICAL STUDY

In this section, we compare the performances of the following methods:

NGP: nonnegative garrote with P-splines with $q = 3$, $k = 2$, and $K = 9$;
NGSM: nonnegative garrote with cubic smoothing splines (e.g., Cantoni, Flemming, and Ronchetti 2011);
COSSO (Lin and Zhang 2006);
ACOSSO (Storlie et al., 2011)

on three simulated examples with different covariates. The results for the proposed NGP method are presented first, followed by the results for the available methods (NGSM, COSSO, and ACOSSO).

In Table 1, we list the different software implementations that were used. The MATLAB-code for the NGP method is provided in the online supplementary materials.

All methods involve the choice of smoothing/regularization parameters, and the performance of a method often (highly) depends on good data-driven choices for these parameters. When performing our numerical studies, we obtained the simulation results using several data-driven criteria for selecting the procedures parameters, being limited of course by these that are implemented in existing codes for available methods. For calculating the initial P-splines or smoothing splines estimators, we used the Akaike's information criterion (AIC) to choose the smoothing parameter. For our own codes developed for the NGP method, we implemented the following parameter selection criteria for the nonnegative garrotte (NG) selection part: AIC, Bayesian information criterion (BIC), five-fold cross-validation (5-CV), generalized cross-validation (GCV), and an L-curve criterion (L-curve). For each of the methods involved in our comparative study, we tried to find a possible "best way" (in general) to choose the parameters of the method. For COSSO and ACOSSO, we report on BIC and 5-CV for the choice of the λ parameter (see also Section 3 in Part I of the online supplementary materials). These are the implemented choices in the available software codes. All together this resulted in the choices indicated in Table 1.

The BIC criterion means that we choose the parameters such that

$$\mathrm{BIC} = \ln\left(\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2\right) + \frac{\ln(n)}{n}\mathrm{df},$$

with df the number of nonzero shrinkage factors, is minimized. The L-curve criterion is based on the so-called L-curve, which plots (for a set of values of a regularization parameter, λ say) the fitted (standardized) quantity $\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2)/(\max_{\lambda}(\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2)) \quad - \min_{\lambda}(\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2)))$ against the penalty term (standardized in a similar manner) and then searches for the point where this curve has the largest curvature. See, for example, Belge, Kilmer, and Miller (2002) and Antoniadis, Gijbels, and Nikolova (2011). This criterion is called an L-curve criterion because the resulting plot has the shape of the letter L, and the task is to find the cornerpoint of the L. One possible algorithm for finding this point is to search for the tangent line that has slope -1. We do this by calculating the running gradients based on all subsequent pairs of points and selecting the one closest to -1. In our tables, we report on the results using this search algorithm. Other implementations of the L-curve criterion (e.g., searching for the point with the smallest distance to the origin) led to comparable results.

Finally, it should be noted that inherent to additive modeling are the constraints (4). These are dealt with rather easily in the implementation as follows. Before starting the analysis, we center the Y-observations, and further each fitted component is centered by subtracting its mean value. When providing plots of fitted (univariate) components, we again add this mean value and as such can present the data points and the fitted univariate components in the original range of Y-values.

To investigate the performances of the methods, we report on several evaluation criteria, described in Table 2, based on 100 simulations. The first two criteria (MS and ER) are also used in Huang, Horowitz, and Wei (2010). Here the estimation error is defined as $\mathrm{ER} = \frac{1}{n}\sum_{i=1}^{n}\{\sum_j (f_j(X_{ij}) - \hat{f}_j(X_{ij}))\}^2$. The third and fourth criteria were used in Fan and Li (2001) and are complemented with the information on the true and false positives. The criteria PercT, AverR, and AverI yet provide another useful summary of the quality of the selection procedure and were used in Wang, Li, and Huang (2008). It is important to mention that to have an overall idea of the performance of the methods, it is important to look at all criteria together. For example, in case of a true model with four nonzero components, it is possible that the median number of selected components MS equals 4, but this might involve also selected components that are true zero components. Information on the occurrence of such events is to be found in MFZ and MTZ. The numbers in brackets in the tables are the standard errors for real-valued criteria (to be found as $(\cdot)$), or the first and third quartiles for discrete-valued criteria (to be found as $(\cdot, \cdot)$ in the tables).

An important criterion for practical use of a statistical method is also its computational complexity. We therefore also record for each method the average computation time for one simulation, listed under the "time" criterion. Since the various methods are implemented using different software environments

Table 2. Evaluation criteria

MS	Median number of selected components
ER	Estimation error
MTZ	Median of zero components restricted to the true zero components (true zeros)
MFZ	Median of zero components restricted to the true nonzero components (false zeros)
MTP	Median of the nonzero components restricted to the true nonzero components (true positives)
MFP	Median of the nonzero components restricted to the true zero components (false positives)
PercT	Percentage of replications in which the exact true model was selected
AverR	Average of the number of relevant variables selected in the model
AverI	Average of the number of irrelevant variables selected in the model
time	Average of the computing time (in seconds)

(MATLAB, R), the differences in computing time may be partially due to these different environments.

4.1 Example 1

We consider a simulated data example of Lin and Zhang (2006), which has also been considered by Cantoni, Flemming, and Ronchetti (2011), with 10 explanatory variables of which 4 are informative. The components are

$$f_1(x) = 5x \qquad f_3(x) = 4\frac{\sin(2\pi x)}{2-\sin(2\pi x)}$$

$$f_2(x) = 3(2x-1)^2 \qquad f_4(x) = 6(0.1\sin(2\pi x) + 0.2\cos(2\pi x) + 0.3\sin^2(2\pi x) + 0.4\cos^3(2\pi x) + 0.5\sin^3(2\pi x))$$

and the model is $Y_i = f_1(X_{i1}) + f_2(X_{i2}) + f_3(X_{i3}) + f_4(X_{i4}) + \varepsilon_i$. The sample size is 100.

The covariates are generated as follows. We first generate $W_{i1}, \ldots, W_{i10}, U_i$ independently from a Uniform[0, 1] for $i = 1, \ldots, n$. Then we take $X_{ij} = \frac{W_{ij}+tU_i}{1+t}$ for $j = 1, \ldots, 10$. The correlation among the predictors is controlled by t via $\mathrm{cor}(X_{ij}, X_{il}) = \frac{t^2}{1+t^2}$ for $j \neq l$. The uniform design corresponds to the case where $t = 0$. The error term ε_i is normally distributed with mean 0 and we consider two variances, namely, 1.74 [resulting in a signal to noise ratio (SNR) of 3 in the uniform case] and 3.9 (with SNR 2 in the uniform case).

For brevity of presentation, we only discuss here the results for $t = 0$ (uncorrelated covariates) and SNR = 2. The other simulation results are provided in Section 2 of the online supplementary materials.

The simulation results are summarized in Tables 3 and 4. In the latter table, we also indicate in one of the top rows, the "ideal" values for the criteria. From Table 3 and Table 14 (in the online supplementary materials), it can be seen that in case the covariates are correlated (case $t = 1$), all methods tend to first remove the nonzero component X_2. Some general conclusions from Tables 3 and 4 and Tables 13–17 (in the online supplementary materials) are as follows:

Table 3. Simulation example 1. Appearance frequency of the variables ($t = 0$ and SNR = 2)

Method	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
NGP	100	78	100	100	2	1	1	2	1	1
NGSM	100	83	100	100	7	13	5	9	6	6
COSSO BIC	100	61	100	100	1	1	0	2	1	1
COSSO 5-CV	100	76	100	100	6	10	8	10	7	4
ACOSSO BIC	100	73	100	100	4	3	2	3	3	2
ACOSSO 5-CV	100	89	100	100	25	28	26	31	26	25

1. The ACOSSO procedures often have the smallest estimation error.
2. According to the median number of false positives criterion (MFP), the ACOSSO (with 5-CV) method seems to be most sensitive to start including true zero components into the model.
3. In case of correlated covariates and low SNR, the performance of all methods degrades, but, for example, COSSO (with BIC) has the lowest AverR value.
4. It was noticed that the results for the COSSO and ACOSSO methods can be quite different when using either the 5-CV or BIC criterion for selecting the procedures parameters. In contrast, the *L*-curve criterion for the NGP method was observed to be a quite stable selection method.
5. The NGP method is quite fast. For the ACOSSO and COSSO methods, the computation time is always the highest.

In Figure 2, we present the "median" fitted curves, for the four true nonzero components, for the case of uncorrelated covariates (i.e., $t = 0$), and SNR = 3. This "median" curve for a given method is the estimated curve that resulted from the simulation leading to the median ER value across all simulations for that method.

4.2 Example 2

This second simulation example is an example of Huang, Horowitz, and Wei (2010) with 21 explanatory variables of which 4 are informative:

$$f_1(x) = 4x - 4 \qquad f_3(x) = -8x^3$$

$$f_2(x) = \cos(2\pi x) \qquad f_4(x) = \sqrt{x(1-x)}\sin\left(\frac{2\pi(1+2^{-0.6})}{x+2^{-0.6}}\right),$$

where the model is $Y_i = f_1(X_{i1}) + f_2(X_{i2}) + f_3(X_{i3}) + f_4(X_{i4}) + \varepsilon_i, \quad i = 1, \ldots, n$. The sample size $n = 100$. The covariates are generated as follows. We first generate $W_{i1}, \ldots, W_{i21}, U_i, V_i$ independently from a Uniform[0, 1] for $i = 1, \ldots, n$. Then we take $X_{ij} = \frac{W_{ij}+tU_i}{1+t}$ for $j = 1, \ldots, 4$ and $X_{ij} = \frac{W_{ij}+tV_i}{1+t}$ for $j = 5, \ldots, 21$. The correlation among the predictors is again controlled by t via $\mathrm{cor}(X_{ij}, X_{il}) = \frac{t^2}{1+t^2}$ for $(1 \leq j \leq 4,\ 1 \leq l \leq 4)$ and $(5 \leq j \leq 21,\ 5 \leq l \leq 21)$, and the

Table 4. Simulation example 1. Evaluation criteria for $t = 0$, SNR= 2

Method	ER	MS	MTZ	MFZ	MTP	MFP	PercT	AverR	AverI	Time
Ideal value	0	4	6	0	4	0	1	4	0	0
NGP	1.1278	4	6	0	4	0	0.7500	3.7800	0.0800	0.5035
	(0.3522)	(3,4)	(6,6)	(0,0)	(4,4)	(0,0)		(0.4163)	(0.2727)	(0.0439)
NGSM	1.2049	4	6	0	4	0	0.5500	3.8300	0.4600	2.1248
	(0.3622)	(4,5)	(5,6)	(0,0)	(4,4)	(0,1)		(0.3775)	(0.6878)	(0.8659)
COSSO BIC	1.3043	4	6	0	4	0	0.5600	3.6100	0.0600	0.9566
	(0.4940)	(3,4)	(6,6)	(0,1)	(3,4)	(0,0)		(0.4902)	(0.2778)	(0.0756)
COSSO 5-CV	1.2306	4	6	0	4	0	0.5400	3.7600	0.4500	4.1852
	(0.4080)	(4,4)	(6,6)	(0,0)	(4,4)	(0,0)		(0.4292)	(1.0088)	(0.8601)
ACOSSO BIC	1.0336	4	6	0	4	0	0.6200	3.7300	0.1700	0.9299
	(0.3801)	(3,4)	(6,6)	(0,1)	(3,4)	(0,0)		(0.4462)	(0.4935)	(0.0764)
ACOSSO 5-CV	1.0925	5	5	0	4	1	0.3600	3.8900	1.6100	4.0361
	(0.3465)	(4,7)	(3,6)	(0,0)	(4,4)	(0,3)		(0.3145)	(1.8635)	(0.8221)

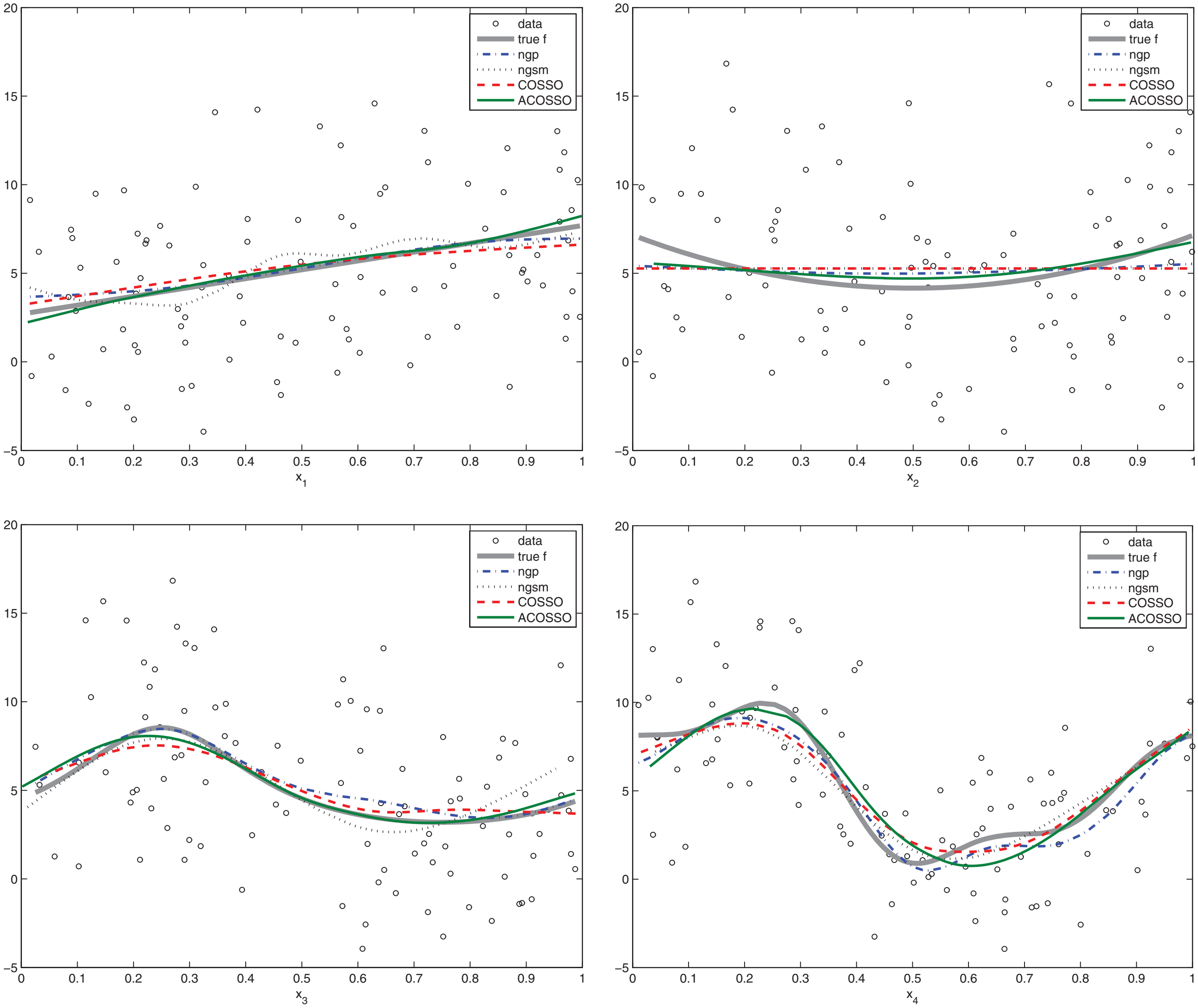

Figure 2. Simulated Example 1 with true component functions (bold solid curves). Fitted components for $t = 0$ and SNR= 3 for the four methods. The online version of this figure is in color.

Table 5. Simulation example 2. Evaluation criteria for $t = 0$

Method Ideal value	ER 0	MS 4	MTZ 17	MFZ 0	MTP 4	MFP 0	PercT 1	AverR 4	AverI 0	Time 0
NGP	0.2023	4	17	0	4	0	0.5100	3.7500	0.3700	1.9660
	(0.0838)	(4,4)	(16,17)	(0,0)	(3,4)	(0,1)		(0.4352)	(0.5972)	(0.1025)
NGSM	0.2057	5	15	1	3	2	0.0800	3.4900	1.9300	4.2162
	(0.0837)	(4,6)	(14,16)	(0,1)	(3,4)	(1,3)		(0.5024)	(1.3872)	(1.8305)
COSSO BIC	0.2280	4	17	1	3	0	0.1200	3.2000	0.6300	1.8096
	(0.0810)	(3,4)	(16,17)	(1,1)	(3,3)	(0,1)		(0.402)	(0.9063)	(0.2369)
COSSO 5-CV	0.2441	3	17	1	3	0	0.0700	3.0900	0.4900	5.0471
	(0.1220)	(3,4)	(17,17)	(1,1)	(3,3)	(0,0)		(0.3786)	(1.2268)	(0.7439)
ACOSSO BIC	0.2525	5	16	1	3	1	0.0700	3.2800	1.7100	1.7311
	(0.1173)	(4,6)	(15,16)	(0,1)	(3,4)	(1,2)		(0.514)	(1.5195)	(0.1781)
ACOSSO 5-CV	0.2525	4	16	1	3	1	0.0500	3.3800	3.2400	4.8641
	(0.0954)	(3,10)	(11,17)	(0,1)	(3,4)	(0,6)		(0.4878)	(4.3114)	(0.7468)

covariates of the nonzero components and the zero components are independent. The error term ε_i is normally distributed with mean 0 and variance 1.

Table 5 and Table 18 (provided in the online supplementary materials) summarize the simulation results for the various evaluation criteria. Also from this example, we can see that when the covariates are correlated ($t = 1$), COSSO and ACOSSO start to eliminate nonzero components and including more irrelevant components (see the columns "MFZ" and "MFP" in Table 18). In this simulation example, the NGP method seems to be a clear winner for most evaluation criteria.

4.3 Example 3

We now consider a model with 10 explanatory variables of which 6 are informative, but for which, 3 of the 6 informative variables have a far smaller influence on the response than the 3 other covariates. This simulation model is thus well suited for testing whether the methods can catch up the difference between variables that have no influence (true zero components) and variables that have a weak influence (i.e., true nonzero but weak components). The true nonzero components are

$$f_1(x) = 40x \qquad f_4(x) = 240/9(0.1\sin(2\pi x) + 0.2\cos(2\pi x) + 0.3\sin^2(2\pi x) + 0.4\cos^3(2\pi x) + 0.5\sin^3(2\pi x))$$

$$f_2(x) = 8(2x-1)^2 \qquad f_5(x) = -8x^3$$

$$f_3(x) = 32\frac{\sin(2\pi x)}{2-\sin(2\pi x)} \qquad f_6(x) = 4\cos(2\pi x)$$

and the model is $Y_i = f_1(X_{i1}) + f_2(X_{i2}) + f_3(X_{i3}) + f_4(X_{i4}) + f_5(X_{i5}) + f_6(X_{i6}) + \varepsilon_i$. The second, fifth, and sixth components are weak components, with a range around 8, whereas the other components have a range around 40. The sample size is 100.

The covariates $X_{i1}, \ldots, X_{i10}$ are generated independently from a Uniform [0, 1] for $i = 1, \ldots, n$. The error term ε_i is normally distributed with mean 0 and variance 0.6.

Table 6 reports on the frequencies (out of 100) that a covariate was included in the selected model. Note that all methods perform in fact quite well, with a somewhat worse performance of the COSSO and ACOSSO methods, by including nonrelevant variables (for ACOSSO) and leaving out the relevant variable X_5 for the COSSO method.

Table 7 provides the summary of the simulation results for the various evaluation criteria from Table 2. From this table, we have similar conclusions as for Table 6. Also the findings from the previous examples remain valid.

5. APPLICATION TO REAL DATA

For the real data analysis, we mention the following:

1. Categorical variables have been treated in the following way. For each categorical variable, we introduced an appropriate number of dummies, entering the model in a parametric way. Except in Section 5.2, we do not provide plots of fitted components for categorical variables since visual interpretation is reduced.
2. The theoretical results on the proposed methods rely on some assumptions. For all data examples, we analyzed the residuals after the selection and estimation, checking for possible problems with underlying assumptions.

In all examples, we use B-splines of degree 3 and difference order penalty $k = 2$ for the NGP method.

5.1 Concrete Data

We consider the concrete data (with 103 data points) of Section 1. The response in our analysis is the slump flow (in cm) of the fresh concrete. The covariates are cement, fine aggregate,

Table 6. Simulation example 3: Appearance frequency of the variables

Method	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
NGP	100	100	100	100	100	100	0	0	0	0
NGSM	100	100	100	100	100	100	0	0	0	0
COSSO BIC	100	100	100	100	99	100	0	0	0	0
COSSO 5-CV	100	100	100	100	98	100	0	1	0	0
ACOSSO BIC	100	100	100	100	100	100	5	4	3	3
ACOSSO 5-CV	100	100	100	100	100	100	4	5	5	7

Table 7. Simulation example 3: evaluation criteria

Method	ER	MS	MTZ	MFZ	MTP	MFP	PercT	AverR	AverI	Time
ideal value	0	6	4	0	6	0	1	6	0	0
NGP	3.6487	6	4	0	6	0	1	6	0	0.4397
	(0.8580)	(6,6)	(4,4)	(0,0)	(6,6)	(0,0)		(0)	(0)	(0.0509)
NGSM	1.6986	6	4	0	6	0	1	6	0	1.8047
	(0.3427)	(6,6)	(4,4)	(0,0)	(6,6)	(0,0)		(0)	(0)	(1.2873)
COSSO BIC	4.6560	6	4	0	6	0	0.9900	5.9900	0	0.8851
	(0.3894)	(6,6)	(4,4)	(0,0)	(6,6)	(0,0)		(0.1000)	(0)	(0.0805)
COSSO 5-CV	4.4859	6	4	0	6	0	0.9700	5.9800	0.0100	2.6541
	(0.3837)	(6,6)	(4,4)	(0,0)	(6,6)	(0,0)		(0.1407)	(0.1000)	(0.1506)
ACOSSO BIC	0.4396	6	4	0	6	0	0.8700	6	0.1500	0.8699
	(0.0835)	(6,6)	(4,4)	(0,0)	(6,6)	(0,0)		(0)	(0.4113)	(0.0712)
ACOSSO 5-CV	0.4383	6	4	0	6	0	0.8900	6	0.2100	2.6264
	(0.0818)	(6,6)	(4,4)	(0,0)	(6,6)	(0,0)		(0)	(0.7426)	(0.1556)

coarse aggregate, water, fly ash, slag, and superplasticizer. The units of all covariates are kg/m^3.

Table 8 indicates with a $\surd$ which variables were selected for each of the methods, and in Figure 3, we plot the fitted components (restricting to the selected components). For COSSO and ACOSSO, we present the plots for the results based on the BIC criterion.

It is remarkable that all four methods select slag and water as ingredients having the highest influence on the slump flow. Three of the four methods select only these two variables. The NGSM method (nonnegative garrote method with smoothing splines) also selects the variable fly ash, but with an estimated coefficient very close to zero (i.e., very small effect of that variable). From Figure 3, it is seen that the influence of water has a linear shape with more water in the HPC leading to an increased slump flow; hence, water has a decreasing effect on workability, since adding water makes the mixture more liquid. The influence of slag on slump flow, however, appears as nonlinear, revealing some local minima in slump flow. In Yeh (2007), it was reported that also fly ash and superplasticizer are ingredients having a strong effect (see page 478 of Yeh 2007). This finding is in contrast with our findings, in which all four methods do not select the variable superplasticizer and only one selects the ingredient fly ash but showing a weak influence. In addition, in Yeh (2007), the estimated effect of water showed a decreasing effect of water on the slump flow for water contents larger than 195 kg/m^3. Since all four state-of-the-art variable selection methods, using flexible modeling with data-driven choices of hyperparameters, point into the same direction, we believe that the impact of fly ash and superplasticizer has been overestimated in the literature, whereas the influence of slag has been largely underestimated.

Finally, in Table 9, we report on the quality of the fits with the selected variables, via

$$\mathrm{RSS} = ||\mathbf{Y} - \hat{\mathbf{Y}}||_2^2,$$

where $\hat{Y}_i = \sum_{j=1}^{d} \hat{f}_{ij}$. The quantity RSS$/n$ is thus an average residual sum of squares. The smaller the RSS$/n$, the better the fit. Note that the lowest RSS$/n$ value is obtained for the NGP method.

As good practice, we investigated the residual plots. They look quite similar for all four methods. For all residual plots, normality was not rejected when tested for. However, we should emphasize that our methods do not require normality of the errors. The residual plot for the NGP method is provided in Figure 7(a) in the online supplementary materials. One remarkable observation from the residual plot is the presence of a kind of line structure in the lower left part of the cloud of points. Investigation revealed that this structure is coming from the set of observations on slump flow that equal the minimum of

Table 8. Concrete data. Selected components when using all observations (indicated with $\surd$) and when only using observations with slump flow > 20 cm (indicated by $*$)

Variable	NGP	NGSM	COSSO BIC	COSSO 5-CV	ACOSSO BIC	ACOSSO 5-CV
Cement				$*$		$*$
Coarse aggregation						$*$
Fine aggregation				$*$		
Water	$\surd *$	$\surd *$	$\surd *$	$\surd *$	$\surd *$	$\surd *$
Slag	$\surd *$	$\surd *$	$\surd$	$\surd *$	$\surd$	$\surd *$
Fly ash		$\surd *$		$*$		$*$
Superplasticizer				$*$		$*$

Table 9. Concrete data: NS = number of selected components and RSS

	All observations ($n = 103$)		Without "20 cm" observations ($n = 86$)	
Method	NS	RSS/n	NS	RSS/n
NGP	2	128.2504	2	90.5216
NGSM	3	149.5926	3	95.1743
COSSO BIC	2	137.0876	1	109.3269
COSSO 5-CV	2	138.9469	6	78.2811
ACOSSO BIC	2	136.0557	1	109.3312
ACOSSO 5-CV	2	136.0557	6	66.4930

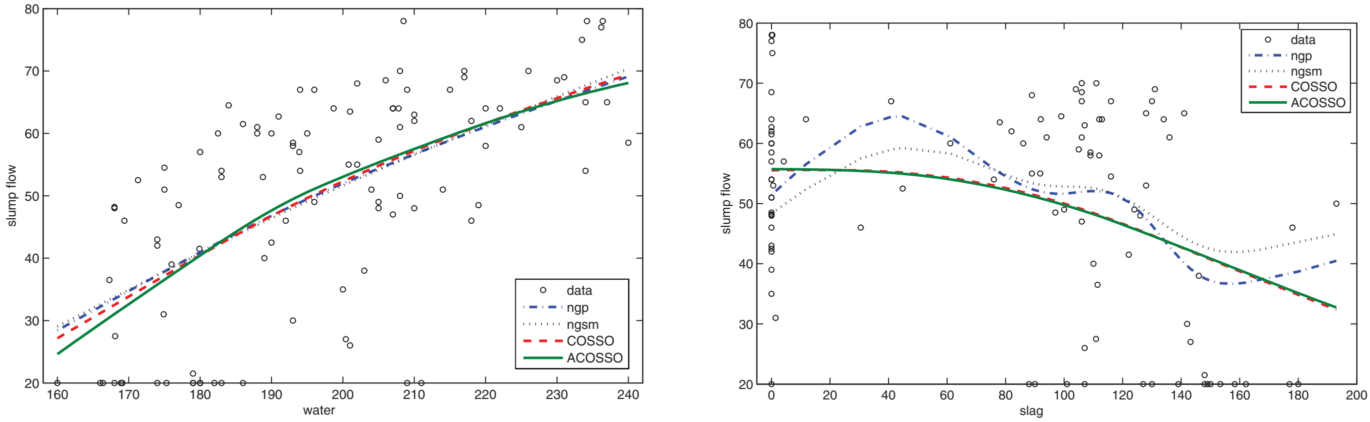

Figure 3. Concrete data. Fitted (nonzero) components for all methods. The online version of this figure is in color.

20 cm. Such an observation in fact means that the slump under investigation is a so-called true slump meaning that the concrete simply subsides and keeps more or less its shape. Given the observed line structure in the residuals, we redid the analysis using only the 86 out of 103 observations for which the slump flow is strictly larger than the minimum of 20 cm. The results of this analysis are also summarized in Tables 8 and 9. Note that for all, but the (A)COSSO BIC methods, this resulted in selecting again the variables water and slag, together with some other variables (except for the NGP method). The estimated effects for the water and slag variables look very similar as for the analysis based on all observations, and the estimated effects of the additionally selected variables are (very) small. From Table 9, it can be seen that there is overall a drop in the RSS/n value.

In Figure 7(b) (in the online supplementary materials), we depict the residual plot for this second analysis for the NGP method. Now, the main cloud of the residual points looks very much alike that in Figure 7(a) except that there is no longer the visible line structure.

5.2 Prostate Cancer Data

The Prostate cancer data come from a study by Stamey et al. (1989). They examined the correlation between the level of prostate-specific antigen and a number of clinical measures in 97 men who were about to receive a radical prostatectomy. The response variable is the level of prostate-specific antigen and the candidate predictor variables are log cancer volume (lcavol), log of prostate weight (lweight), age, log of the amount of benign prostatic hyperplasia (lbph), seminal vesicle invasion (svi), log of capsular penetration (lcp), Gleason score (gleason), and percent of Gleason scores 4 or 5 (pgg45).

The selected components by each of the four methods are listed in Table 10. In the nonnegative garrote with P-splines, we took $K = 9$. In our analysis, we removed observation number 32, which corresponds to a man with log prostate weight equal to 6.1076 and 2.008 as a level of prostate-specific antigen. This observation seemed to be an outlier.

Note that all four methods select the variables cancer volume (lcavol), prostate weight (lweight), and seminal vesicle invasion (svi), with a strong increasing effect of the first two variables on the level of prostate-specific antigen (see Figure 4; presenting only the selected components). The effect of the variable svi is less strong. For the two nonnegative garrote methods (NGP and NGSM), there is no difference in the level of prostate-specific antigen for the different Gleason scores (which is a measure for the malignancy of the tissue). Note again the very similar behavior of the fitted components for all four methods. Yuan (2007) analyzed these data using nonnegative garrote component selection, resulting into the selection of lcavol, lweight, lbph, and svi, with a reported prediction error of more than 50%. His analysis revealed a nonlinear effect of the variable “log of the amount of benign prostatic hyperplasia” (lbph) that was not confirmed by any of the four methods.

Table 11 (left) gives the RSS and the number of selected components. The lowest RSS value is for COSSO with 5-CV, including all variables. The second lowest RSS value is noted for ACOSSO 5-CV with 6 selected variables.

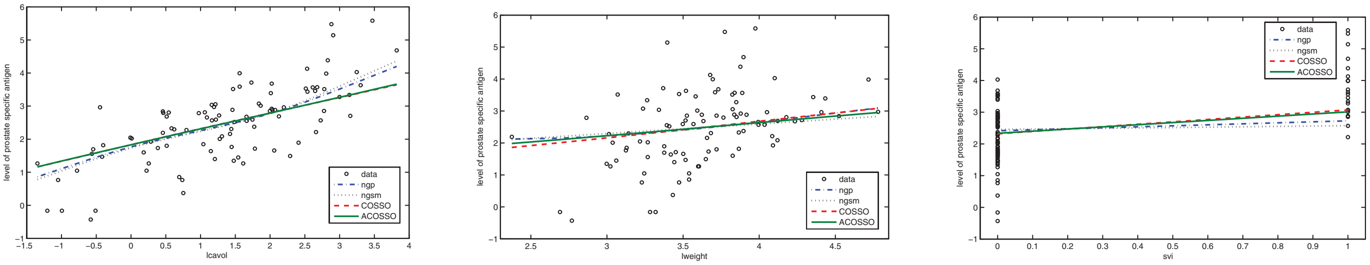

Figure 4. Prostate cancer data. Fitted (nonzero) components for all methods. The online version of this figure is in color.

Table 10. Prostate cancer data: selected components

Variable	NGP	NGSM	COSSO BIC	COSSO 5-CV	ACOSSO BIC	ACOSSO 5-CV
lcavol	✓	✓	✓	✓	✓	✓
lweight	✓	✓	✓	✓	✓	✓
age				✓		✓
lbph				✓		✓
svi	✓	✓	✓	✓	✓	✓
lcp				✓		
gleason				✓	✓	✓
pgg45				✓		

5.3 Waste Isolation Pilot Plant Data

We consider data (with 300 observations) from a sensitivity analysis of a model for two-phase fluid flow, carried out as part of the 1996 compliance certification application for a waste isolation pilot plant (WIPP) in New Mexico by Sandia National Laboratory. The simulated model corresponds to a drilling intrusion at 1000 years that penetrates both the repository and an underlying region of pressurized brine and simulates the waste panel's condition as a deterministic function of model inputs. The simulated model contains several variables describing the environmental conditions. The aim of our study is to determine which variables have an important effect on the response, namely, cumulative brine flow (in m^3), into a waste repository at 10,000 years for a drilling intrusion at 1000 years.

The brine enters the repository mainly by flow through the anhydrite ($CaSO_4$) marker beds, drainage from the disturbed rock zone, brine flow up the borehole from a pressurized brine pocket, and flow down the intruding borehole from overlying formations. We consider 11 explanatory variables that relate to the various environmental conditions (see Table 19 in the online supplementary materials and Vaughn et al. 2000).

We use B-splines with 16 knots for the nonnegative garrote with P-splines method. Note first of all from Table 12 that the ACOSSO method using BIC selects none of the variables (so the fitted model is just a constant model), whereas the ACOSSO method with 5-CV selects all the variables (but several fitted components are close to zero). Also the COSSO method selects all variables. The two nonnegative garrote-based methods are more selective, and each select the variables ANHPRM (the anhydrite permeability), BHPRM (the borehole permeability), BPCOMP (the bulk compressibility of the brine pocket), HALPOR (the halite porosity), and WMICDFLG (microbial degradation of cellulose) as relevant variables in explaining the brine flow. See also Figure 5 in which we plot the fitted effect of the first four variables. These have an increasing effect on the brine flow. This finding corresponds to the conclusions from the analysis of these data in Reich, Storlie, and Bondell (2009). Of interest is, for example, the flat part in the effect of ANHPRM: the anhydrite permeability needs to exceed a certain threshold before it becomes sufficient to counteract the pressure in the repository and to allow for brine to flow from the marker beds. In Reich, Storlie, and Bondell (2009), a Bayesian nonparametric variable selection method was used, but no clear guideline for how to choose the involved hyperparameter was provided, and the reported inclusion or not of, for example, the variable HALPRM depends on the choice of that parameter.

Table 11. Data examples: NS = number of selected components and RSS

Prostate data			WIPP data		
Method	NS	RSS/n	Method	NS	RSS/n
NGP	3	0.4960	NGP	5	$2.1662 \cdot 10^8$
NGSM	3	0.5321	NGSM	5	$2.3217 \cdot 10^8$
COSSO BIC	3	0.4906	COSSO BIC	11	$1.8075 \cdot 10^8$
COSSO 5-CV	8	0.4361	COSSO 5-CV	11	$1.7411 \cdot 10^8$
ACOSSO BIC	4	0.4805	ACOSSO BIC	0	$7.5941 \cdot 10^8$
ACOSSO 5-CV	6	0.4596	ACOSSO 5-CV	11	$1.1274 \cdot 10^8$

Table 12. WIPP data: selected components

Variable	NGP	NGSM	COSSO BIC	COSSO 5-CV	ACOSSO BIC	ACOSSO 5-CV
ANHPRM	✓	✓	✓	✓		✓
BHPRM	✓	✓	✓	✓		✓
BPCOMP	✓	✓	✓	✓		✓
BPPRM			✓	✓		✓
HALPOR	✓	✓	✓	✓		✓
HALPRM			✓	✓		✓
SHPRMASP			✓	✓		✓
SHPRMCLY			✓	✓		✓
SHPRMHAL			✓	✓		✓
SHRBRSAT			✓	✓		✓
WMICDFLG	✓	✓	✓	✓		✓

In the right panel of Table 11, the RSS values of the fitted models are given. The smallest values are to be noted for the models that include all variables. Among the two nonnegative garrote methods, the smallest RSS value is recorded for the NGP method.

6. FURTHER DISCUSSION AND CONCLUSIONS

In this article, we proved the estimation and variable selection consistency of the nonnegative garrote with P-splines in additive models. An advantage of P-splines is that they are easy to understand, since they are based on a development in a B-spline basis and use a rather simple (discrete) penalty based on kth-order differences. In addition, methods based on P-splines are very economical in computing time and are also very well suited for generalizations to mixed effects models (see, e.g., Ruppert, Wand, and Carroll 2003) that occur in longitudinal data analysis. Hence the work presented here is useful for further dealing with flexible modeling for complex data settings.

We compared nonnegative garrote with P-splines and smoothing splines, with (A)COSSO on simulated and real data. When the covariates are correlated, the methods start removing more easily nonzero components. This is particularly the case for the (A)COSSO methods. It was noted that the performance of the COSSO method can be quite different when using BIC or 5-CV for selecting the regularization parameters. The NGP and NGSM methods with an L-curve criterion for selecting the (smoothing/regularization) parameters turned out to be

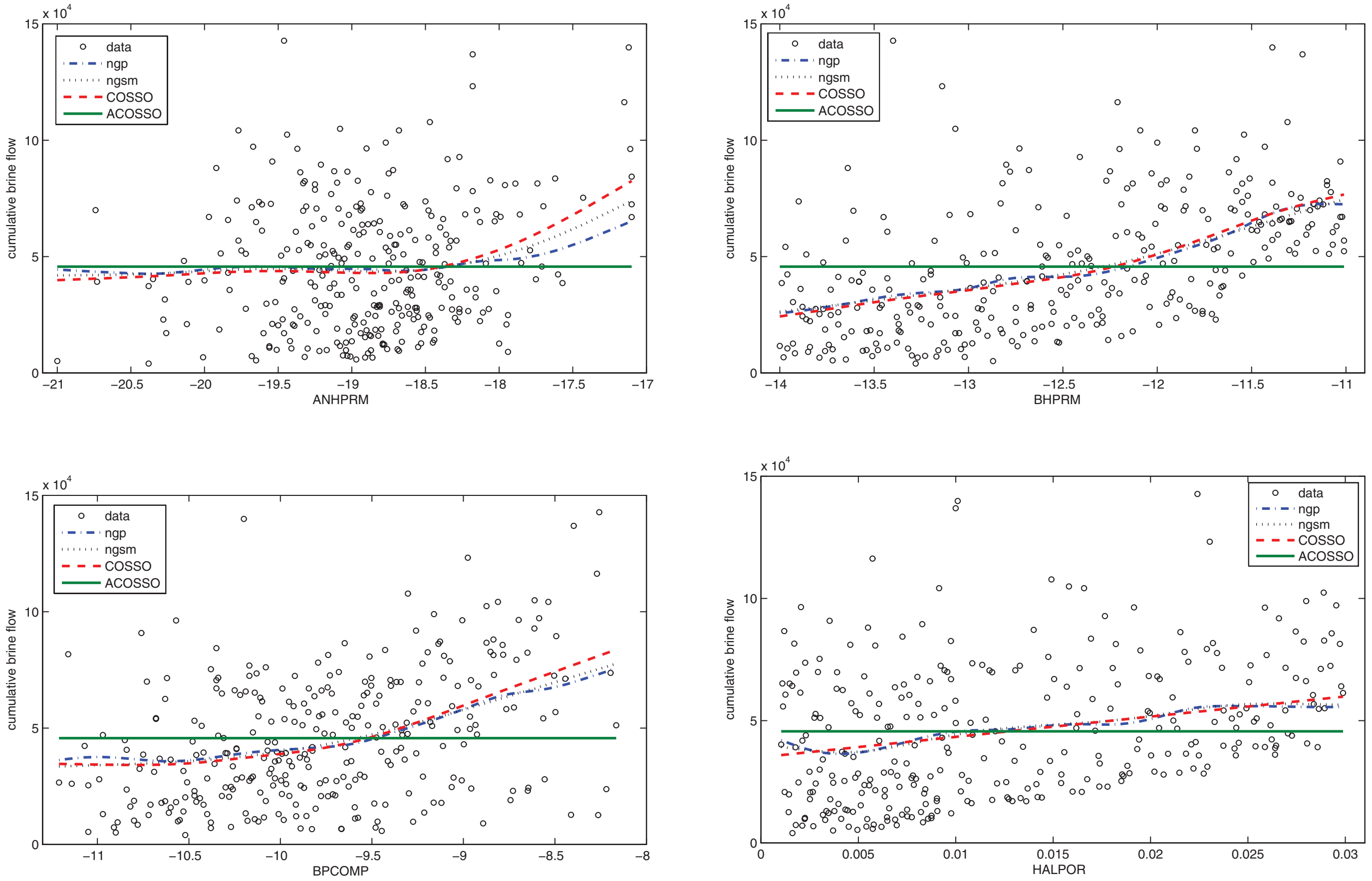

Figure 5. WIPP data. Fitted components for the variables ANHPRM, BHPRM, BPCOMP, and HALPOR, for all methods. The online version of this figure is in color.

reasonable stable methods. Among these two, a preference is given for the NGP method (using P-splines) over the NGSM method (using smoothing splines) since they use less parameters. Indeed the NGSM method builds on n knots, whereas the P-splines based methods use K knots with K far less than n. The computational advantage of the NGP method is clear from the time criterion that is included in the simulation study.

Working with P-splines as an ingredient involves certain choices to be made beforehand, for example, there is the degree of the spline q, the difference order of the penalty k, the number of knot points K, and the smoothing parameter λ. There is an interplay between these parameters as was theoretically investigated in Gijbels and Verhasselt (2010a,b), among others. In particular, also the choices of K and λ are related. This can be seen, for example, from the assumptions imposed on these parameters in the theoretical results. In this article, we assume that the univariate functions in the additive model are smooth and taking $q = 3$ and $k = 2$ are common (and recommended) choices. Giving the interplay between K and λ, it is a good practice to fix K and have a good data-driven choice for λ. This is the angle taken in this article. Given the well-understood (theoretical) relationship between K and λ, choosing K approximately $\sqrt{n}$ seems a good choice that we applied here.

In this article, we only consider basic additive models that allow functions of individual variables. There are situations though in which, for example, the combined effect of particular components, say components i and j, is not additive and an interaction term is needed. This requires the replacement of the additive form $f_i(x_i) + f_j(x_j)$ for variables x_i and x_j with the bivariate component $f_{ij}(x_i, x_j)$. Assuming again that the bivariate interaction component is smooth along both variables, it is rather straightforward to extend the nonnegative garrote method using penalized B-spline tensor products, where appropriate difference penalties are placed on the rows and columns of the tensor product coefficients as in Marx and Eilers (2005).

SUPPLEMENTARY MATERIALS

Part I. Proofs, alternative algorithm and additional results: A pdf file containing the following:

1. **Proofs of Theorems 2 and 3:** Proofs of the theorems are provided.
2. **Additive P-Spline Selection Operator:** An alternative approach using a different type of penalty is investigated.
3. **Additional simulation results for Examples 1 and 2:** This part contains additional simulation results for Examples 1 and 2, as well as for the method presented in Section 2 of the supplementary materials.
4. **Application to real data: Additional description and analysis:** This section contains more information on

the concrete data and WIPP data example, as well as residual analysis for the concrete data example.

Part II. Codes: Information and codes are provided for (1) the data generation and the nonnegative garrotte method in MATLAB, (2) the smoothing spline method using R-software, and (3) running the COSSO and ACOSSO software in R (see file codes.pdf).

ACKNOWLEDGMENTS

The authors thank the Editor, an Associate Editor, and three reviewers for very valuable comments on the manuscript that led to a considerable improvement. Part of the work was done while the third author was affiliated at the Katholieke Universiteit Leuven. This author also thanks the Université Joseph Fourier for its kind hospitality during the visits that initiated the research. This research is supported by the IAP (Interuniversity Attraction Poles) Research Network P6/03 of the Belgian State (Belgian Science Policy), the FWO (Fonds Wetenschappelijk Onderzoek) project G.0328.08N of the Flemish Science Foundation, and the Research Fund of the KU Leuven (GOA/07/04-project, GOA/12/014-project, and PDMk/10/176).

[Received December 2009. Revised June 2012.]

REFERENCES

Antoniadis, A., Gijbels, I., and Nikolova, M. (2011), "Penalized Likelihood Regression for Generalized Linear Models With Nonquadratic Penalties," *The Annals of the Institute of Statistical Mathematics*, 63, 585–615. [426,430]

Belge, M., Kilmer, M. E., and Miller, E. L. (2002), "Efficient Determination of Multiple Regularization Parameters in a Generalized *L*-Curve Approach," *Inverse Problems*, 18, 1161–1183. [430]

Breiman, L. (1995), "Better Subset Regression Using the Nonnegative Garrote," *Technometrics*, 51, 373–384. [426,427]

Cantoni, E., Flemming, J., and Ronchetti, E. (2011), "Variable Selection in Additive Models by Nonnegative Garrote," *Statistical Modelling*, 11, 237–252. [426,427,430,431]

Chien, W.-H., Chen, L., Wei, C.-C., Hsu, H.-H., and Wang, T.-S. (2010), "Modeling Slump Flow of High-Performance Concrete Using a Back-Propagation Network," *Applied Mechanics and Materials*, 20–23, 838–842. [425]

Claeskens, G., Krivobokova, T., and Opsomer, J. D. (2009), "Asymptotic Properties of Penalized Spline Estimators," *Biometrika*, 96, 529–544. [427,428]

Eilers, P., and Marx, B. (1996), "Flexible Smoothing With B-Splines and Penalties," *Statistical Science*, 11, 89–102. [426,427,428]

Eubank, R. L., and Speckman, P. (1990), "Curve Fitting by Polynomial-Trigonometric Regression," *Biometrika*, 77, 1–9. [428]

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. [430]

Frank, A., and Asuncion, A. (2010), UCI Machine Learning Repository. "Concrete Slump Test Data Set" [online]. Available at *http://archive.ics.uci.edu/ml*. [425]

Gijbels, I., and Verhasselt, A. (2010a), "P-Splines Regression Smoothing and Difference Type of Penalty," *Statistics and Computing*, 20, 499–511. [437]

——— (2010b), "Regularisation and P-Splines in Generalised Linear Models," *Journal of Nonparametric Statistics*, 22, 271–295. [437]

Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference and Prediction* (2nd ed.), New York: Springer. [426]

Horowitz, J., Klemelä, J., and Mammen, E. (2006), "Optimal Estimation in Ddditive Regression Models," *Bernoulli*, 12, 271–298. [427,428,429]

Huang, J., Horowitz, J., and Wei, F. (2010), "Variable Selection in Nonparametric Additive Models," *The Annals of Statistics*, 38, 2282–2313. [430,431]

Lin, B. Y., and Zhang, H. H. (2006), "Component Selection and Smoothing in Multivariate Nonparametric Regression," *The Annals of Statistics*, 32, 2272–2297. [426,430,431]

Marx, B., and Eilers, P. (2005), "Multidimensional Penalized Signal Regression," *Technometrics*, 47, 13–22. [438]

Reich, B., Storlie, C. B., and Bondell, H. D. (2009), "Variable Selection in Bayesian Smoothing Spline ANOVA Models: Application to Deterministic Computer Codes," *Technometrics*, 51, 110–120. [436]

Ruppert, D., Wand, M., and Carroll, R. J. (2003), *Semiparametric Regression*, Cambridge: Cambridge University Press. [437]

Stamey, T., Kabalin, J., McNeal, J., Johnston, I., Freiha, F., Redwine, E., and Yang, N. (1989), "Prostate Specific Antigen in the Diagnosis and Treatment of Adenocarcinoma of the Prostate, ii: Radical Prostatectomy Treated Patients," *Journal of Urology*, 16, 1076–1083. [435]

Storlie, C., Bondell, H., Reich, B., and Zhang, H. (2011), "Surface Estimation, Variable Selection, and the Nonparametric Oracle Property," *Statistica Sinica*, 21, 697–705. [426,430]

Vaughn, P., Bean, J. E., Helton, J. C., Lord, M. E., MacKinnon, R. J., and Schreiber, J. D. (2000), "Representation of Two-Phase Flow in the Vicinity of the Repository in the 1996 Performance Assessment for the Waste Isolation Pilot Plant," *Reliability Engineering and System Safety*, 69, 205–226. [436]

Wang, L., Li, H., and Huang, J. Z. (2008), "Variable Selection in Nonparametric Varying-Coefficient Models for Analysis of Repeated Measurements," *Journal of the American Statistical Association*, 103, 1556–1569. [430]

Xiong, S. (2010), "Some Notes on the Nonnegative Garrote," *Technometrics*, 52, 349–361. [426]

Yeh, I. (2007), "Modeling Slump Flow of Concrete Using Second-Order Regressions and Artificial Neural Networks," *Cement and Concrete Composites*, 29, 474–480. [425,434]

Yuan, M. (2007), "Nonnegative Garrote Component Selection in Functional ANOVA models," *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics March 21–24, 2007, San Juan, Puerto Rico*, 2, 660–666. [426,427,429,436]

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Pproperties," *Journal of the American Statistical Association*, 101, 1418–1429. [426]