HHS Public Access

Author manuscript

Biometrika. Author manuscript; available in PMC 2015 December 02.

Published in final edited form as:

Biometrika. 2015 June 2; 102(2): 381–395. doi:10.1093/biomet/asu070.

Automatic structure recovery for additive models

YICHAO WU and **LEONARD A. STEFANSKI**

Department of Statistics, North Carolina State University, Raleigh, North Carolina 27519, U.S.A.

SUMMARY

We propose an automatic structure recovery method for additive models, based on a backfitting algorithm coupled with local polynomial smoothing, in conjunction with a new kernel-based variable selection strategy. Our method produces estimates of the set of noise predictors, the sets of predictors that contribute polynomially at different degrees up to a specified degree M, and the set of predictors that contribute beyond polynomially of degree M. We prove consistency of the proposed method, and describe an extension to partially linear models. Finite-sample performance of the method is illustrated via Monte Carlo studies and a real-data example.

Keywords

Backfitting; Bandwidth estimation; Kernel; Local polynomial; Measurement-error model selection likelihood; Model selection; Profiling; Smoothing; Variable selection

1. Introduction

Because of recent developments in data acquisition and storage, statisticians often encounter datasets with large numbers of observations or predictors. The demand for analysing such data has led to the current heightened interest in variable selection. For parametric models, classical methods include backward, forward and stepwise selection. More recently, many approaches to variable selection have been developed that use regularization via penalty functions. Examples include the lasso (Tibshirani, 1996), smoothly clipped absolute deviation (Fan & Li, 2001), the adaptive lasso (Zou, 2006; Zhang & Lu, 2007), and the L_0 penalty (Shen et al., 2013). Fan & Lv (2010) give a selective overview of variable selection methods.

Variable selection for nonparametric modelling has advanced at a slower pace than for parametric modelling, and has been studied primarily in the context of additive models, which are an important extension of multivariate linear regression. An additive model presupposes that each predictor contributes a possibly nonlinear effect, and that the effects of multiple predictors are additive. Such models were proposed by Friedman & Stuetzle (1981) and serve as surrogates for fully nonparametric models. Most nonparametric variable selection methods studied so far deal with additive models; see Ravikumar et al. (2009), Huang et al. (2010), Fan et al. (2011), and references therein. An exception is Stefanski et al. (2014), which considers variable selection in a fully nonparametric classification model.

wu@stat.ncsu.edu stefansk@ncsu.edu

Additionally, variable selection and structure recovery for the varying coefficient model, a popular extension of the additive model, have been studied: Xia et al. (2004) considered structure recovery towards semi-varying coefficient modelling; Fan et al. (2014) studied a new variable screening method; and for the longitudinal setting Cheng et al. (2014) investigated variable screening and selection, as well as structure recovery.

Partially linear models were proposed by Engle et al. (1986). Combining the advantages of linearity and additivity, these models assume that some covariates have nonlinear additive effects while others contribute linearly. Estimation for a partially linear model requires knowing which covariates have linear effects and which have nonlinear effects, information that is usually not available a priori. Recently, Zhang et al. (2011) and Huang et al. (2012) proposed methods to identify covariates that have linear effects and ones that have nonlinear effects.

We consider an additive model for a scalar response Y and predictors $X = (X_1, \dots, XD)^{\mathrm{T}}$,

$$Y = m(X) + \varepsilon = \alpha + \sum_{d=1}^{D} m_d(X_d) + \varepsilon, \qquad E(\varepsilon) = 0, \qquad var(\varepsilon) = \sigma^2, \quad (1)$$

under the identifiability conditions $E\{m_d(X_d)\} = 0$ ($d = 1, \dots, D$). Denote a random sample from model (1) by $\{(Y_i, Xi): i = 1, \dots, n\}$, where $X_i = (X_{i1}, \dots, X_{iD})^{\mathrm{T}}$. The goal is to estimate α and $m_d(\cdot)$ ($d = 1, \dots, D$).

We a method for estimating $m_d(\cdot)$ that first distinguishes between important predictors and predictors that are unimportant, i.e., those X_d for which $m_d(\cdot) = 0$. Next, motivated by Zhang et al. (2011) and Huang et al. (2012), the method identifies predictors that have linear effects from the estimated set of important predictors. Then, the method identifies the predictors that have quadratic effects, and so on. This process continues and results in estimates of sets of predictors that have polynomial effects at different degrees, up to some degree M, and the set of predictors for which the corresponding $m_d(\cdot)$ are not polynomial of any degree up to M.

At the core of our structure recovery method is a new nonparametric kernel-based variable selection method derived from the measurement-error model selection likelihood approach of Stefanski et al. (2014). They studied the relationship between lasso estimation and measurement error attenuation in linear models, and used that connection to develop a general approach to variable selection in nonparametric models.

2. Backfitting algorithm

Backfitting coupled with smoothing is commonly used for fitting model (1). Here we use univariate local polynomial smoothing, with $\mathscr{S}_{K,h,p}$ denoting the univariate local polynomial smoother with kernel function $K(\cdot)$, bandwidth h and degree p.

Local polynomial smoothing (Fan & Gijbels, 1996) is a well-studied nonparametric smoothing technique. To estimate the regression function $g(t) = E(Z \dots T = t)$ from an independent and identically distributed random sample $\{(T_i, Z_i) : i = 1, \dots, n\}$, local

polynomial regression uses Taylor series approximations and weighted least squares. The local polynomial smoothing estimate $\hat{g}(t_0)$ of $g(t_0)$ based on smoother $\mathscr{S}_{K,h,p}$ is given by $\hat{a}_0$, the optimizer of

$$\min_{a_0,a_1,\ldots,a_p} \sum_{i=1}^{n} \left\{ Z_i - \sum_{j=0}^{p} a_j (T_i - t_0)^j \right\}^2 K\left(\frac{T_i - t_0}{h}\right).$$

See Fan & Gijbels (1996) for a detailed account of local polynomial modelling.

Using univariate local polynomial smoothing with kernel $K(\cdot)$, bandwidth h_d and degree p_d to estimate $m_d(\cdot)$ in the additive model (1), the backfitting algorithm consists of the following steps.

***Step* 1**

Initialize by setting $\hat{\alpha} = n^{-1}\sum_{i=1}^{n} Y_i$ and $\hat{m}_d(\cdot) \equiv 0$ for $d = 1, \ldots, D$.

***Step* 2**

For $d = 1, \ldots, D$:

(a) apply the local polynomial smoother $\mathscr{S}_{K,h_d,p_d}$ to $\left\{\left\{X_{id}, Y_i - \hat{\alpha} - \sum_{k\neq d} \hat{m}_k(X_{ik})\right\} : i = 1, \ldots, n\right\}$ and set the estimated function to be the updated estimate $\hat{m}_d(\cdot)$ for $m_d(\cdot)$;

(b) if necessary, apply centring by updating $\hat{m}_d(\cdot)$ with $\hat{m}_d(\cdot) - n^{-1}\sum_{i=1}^{n} \hat{m}_d(X_{id})$.

***Step* 3**

Repeat Step 2 until the changes in all $\hat{m}_d(\cdot)$ ($d = 1, \ldots, D$) between successive iterations are less than a specified tolerance.

Denote the estimates at convergence by $\hat{m}_d^{BF}(\cdot; h, p)$ ($d = 1, \ldots, D$) and $\hat{\alpha}^{BF}$, where $h = (h_1, \ldots, h_D)^{\mathrm{T}}$ and $p = (p_1, \ldots, p_D)^{\mathrm{T}}$ are the vectors of smoothing bandwidths and local polynomial degrees, respectively. For simplicity we use the same kernel $K(\cdot)$ for each component, and thus $K(\cdot)$ is omitted from the notation $\hat{m}_d^{BF}(\cdot; h, p)$.

Backfitting works well for problems of moderate dimension D. However, even though backfitting entails only univariate smoothing, its performance deteriorates as D gets larger. Consequently, variable selection is important for the additive model.

3. Variable selection via backfitting local constants

3·1. *Variable selection*

In Step 2(a) of the backfitting algorithm, any smoothing method can be applied. In this section we consider the local constant smoother, i.e., the local polynomial smoother of degree 0, and propose a nonparametric variable selection method for the additive model (1).

With p_d =0 in Step 2(a) of the backfitting algorithm, we update the estimate $\hat{m}_d(x_d)$ of $m_d(x_d)$ by the minimizer $\hat{a}_0$ of

$$\min_{a_0} \sum_{i=1}^{n} \left[\left\{ Y_i - \hat{\alpha} - \sum_{k \neq d} \hat{m}_k(X_{ik}) \right\} - a_0 \right]^2 K\left(\frac{X_{id} - x_d}{h_d} \right). \quad (2)$$

Note that $K\{h_d^{-1}(Xi_d - x_d)\} = K(0)$ for all i when $h_d = \infty$. In this case, $m_d(\cdot)$ is globally approximated by a_0. The corresponding optimizer of (2) is given by $\hat{a}_0 = n^{-1}\sum_{i=1}^{n} \left\{ Y_i - \hat{\alpha} - \sum_{k \neq d} \hat{m}_k(X_{ik}) \right\}$, which equals zero because $\hat{a} = n^{-1}\sum_{i=1}^{n} Y_i$ and centring is applied in Step 2(b) of the backfitting algorithm.

Thus, when $p_d = 0$, the backfitting algorithm leads to the constant zero function $\hat{m}_d^{BF}(x_d; h, p)$ when $h_d = \infty$. Consequently, the jth predictor is excluded from the model when $h_d = \infty$ and is included only when $h_d < \infty$. The equivalence, $h_d < \infty$ if and only if the jth predictor is included, is key to the approach described in Stefanski et al. (2014), and we exploit it repeatedly in our method. We assume in this section that the degree of local polynomial smoothing for every function component is 0, i.e., p_d 0 for $d = 1, \ldots, D$. In this way, the smoothing bandwidth relates directly to the importance of each predictor, with small h_d corresponding to important predictors.

As in Stefanski et al. (2014), we reparameterize h_d as $\lambda_d = 1/h_d$, so that large λ_d will correspond to important predictors and $\lambda_d = 0$ to unimportant predictors. Predictor smoothing is now determined by the inverse bandwidths λ_d. Stefanski et al. (2014) show that variable selection can be obtained by minimizing a loss function with respect to the λ_d, subject to a bound on the total amount of smoothing as determined by the sum of the λ_d or, equivalently, a bound on the harmonic mean of the bandwidths.

A suitable loss function for backfitting is the sum of squared errors

$$SSE = \sum_{i=1}^{n} \left\{ Y_i - \hat{\alpha}^{BF} - \sum_{d=1}^{D} \hat{m}_d^{BF}\left(X_{id}; \lambda^{-1}, 0_D \right) \right\}^2,$$

where in the second two arguments of $m_d^{BF}(.;.,.)$, $\lambda^{-1} = (1/\lambda_1, \ldots, 1/\lambda_D)^{\mathrm{T}}$ and 0_D denotes the $D \times 1$ zero vector. We shall also write 1_s for the $s \times 1$ vector of ones.

In the absence of constraints, the minimum of SSE with respect to λ is 0, and corresponds to overfitted models. As in Stefanski et al. (2014), appropriate regularization is achieved by solving the constrained optimization problem

$$\min_{\lambda_1,\ldots,\lambda_D} \frac{1}{n}\sum_{i=1}^{n}\left\{Y_i - \hat{\alpha}^{BF} - \sum_{d=1}^{D}\hat{m}_d^{BF}\left(X_{id};\lambda^{-1},0_D\right)\right\}^2, \quad \text{subject to} \quad \lambda_d \geq 0\,(d=1,\ldots,D) \quad \text{and} \quad \sum_{d=1}^{D}\lambda_d=\tau_0. \tag{3}$$

Solving (3) distinguishes important predictors, $\hat{\lambda}_{d>0}$, from those that are unimportant, $\hat{\lambda}_{d=0}$.

3·2. *Modified coordinate descent algorithm*

The constrained optimization problem (3) is not convex because of the complicated dependence on λ through the backfitting algorithm and the univariate local polynomial smoothing. We have had success using a modified coordinate descent algorithm.

Coordinate descent for high-dimensional lasso regression (Fu, 1998; Daubechies et al., 2004) cycles through variables one at a time, solving simple marginal univariate optimization problems at each step, and thus is computationally efficient. It has been studied extensively, for example by Friedman et al. (2007) and Wu & Lange (2008), among others. However, standard coordinate descent cannot be applied directly to (3).

In our modified coordinate descent algorithm, we initialize with equal smoothing, i.e., we set $\lambda_d = \tau_0/D$ for all $d = 1, \ldots, D$. We then cycle through all components and make univariate updates. Current solutions are denoted by $\lambda_d^{\mathrm{c}}\,(d=1,\ldots,D)$, where the superscript c means current. Suppose that we are updating the d'th component. Let $\lambda_{d'}^{(1)}=(0,\ldots,0,1,0,\ldots,0)^{\mathrm{T}}$, a vector whose elements are all zero except for the d'th, which is 1, and set $\lambda_{-d'}^{\mathrm{c}}=\left(\lambda_1^{\mathrm{c}},\ldots,\lambda_{d'-1}^{\mathrm{c}},0,\lambda_{d'+1}^{\mathrm{c}},\ldots,\lambda_D^{\mathrm{c}}\right)/\sum_{d\neq d'}\lambda_d^{\mathrm{c}}$. The components of $\lambda_{d'}^{(1)}$, and $\lambda_{-d'}^{\mathrm{c}}$ sum to unity. Thus, for any $\gamma\in[0,\ \tau_0]$, the components of $\gamma\lambda_{d'}^{(1)}+(\tau_0-\gamma)\,\lambda_{-d'}^{\mathrm{c}}$, sum to τ_0 and satisfy the first constraint in (3). We then update the set of current solutions to the components of the vector $\hat{\gamma}\lambda_{d'}^{(1)}+(\tau_0-\hat{\gamma})\,\lambda_{-d'}^{\mathrm{c}}$, where $\hat{\gamma}$ is the optimizer of

$$\min_{\gamma}\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\hat{\alpha}^{BF}-\sum_{j=1}^{D}\hat{m}_j^{BF}\left[X_{ij};\left\{\gamma\lambda_{d'}^{(1)}+(\tau_0-\gamma)\,\lambda_{-d'}^{\mathrm{c}}\right\}^{-1},0_D\right]\right)^2$$

subject to $0\leq\gamma\leq\tau_0$. Cycling through $d' = 1, \ldots, D$ completes one iteration of the algorithm. Iterations continue until the change in solutions between successive iterations becomes small enough.

3·3. *Tuning*

The tuning parameter τ_0 controls the total amount of smoothing, and can be selected by standard methods such as crossvalidation, AIC or BIC. Denote the optimizer of (3) by $\hat{\lambda}\,(\tau_0)$. If an independent tuning dataset is available, then τ_0 can be selected by minimizing the sum of squared prediction errors of the estimator

$\hat{m}\left\{x;\hat{\lambda}(\tau_0),0\right\}=\hat{\alpha}^{BF}+\sum_{d=1}^{D}\hat{m}_d^{BF}\left[x_d;\left\{\hat{\lambda}(\tau_0)\right\}^{-1},0\right]$ over the tuning set. For methods such as AIC and BIC, the degrees of freedom is needed. The local constant smoothing estimator is a linear smoother, and we couple the backfitting algorithm to marginal local constant smoothing. For each function component, the trace of the corresponding smoothing matrix minus 1 can be used as the degrees of freedom. The trace is reduced by 1 to account for the centring applied in Step 2(b) of the backfitting algorithm.

4. Higher-degree local polynomial regression

The approach in § 3 extends readily to higher-degree local polynomial regression. Suppose that we use local degree p^* polynomial regression in Step 2(a) of the backfitting algorithm. With $p_d = p^*$ in Step 2(a), we update the estimate $\hat{m}_d(x_d)$ of $m_d(x_d)$ by the minimizer $\hat{a}_0$ of

$$\min_{a_0,\ldots,a_{p^*}}\sum_{i=1}^{n}\left[\left\{Y_i-\hat{\alpha}-\sum_{k\neq d}\hat{m}_k(X_{ik})\right\}-\sum_{j=0}^{p^*}a_j(X_{id}-x_d)^j\right]^2K\left(\frac{X_{id}-x_d}{h_d}\right).$$

As remarked previously, $K\{h_d^{-1}(X_{id}-x_d)\}=K(0)$ for all i when $h_d=\infty$. Consequently, $\sum_{j=0}^{p^*}a_j(X_{id}-x_d)^j$ is a global approximation of $m_d(x_d)$. For this case, in Step 2(a) of the backfitting algorithm the estimate $\hat{m}_d(x_d)$ is updated by $\sum_{j=0}^{p^*}\hat{c}_j x_d^j$, where $\hat{c}_0,\ldots\hat{c}_p$ are the optimizers of $min_{c_p,\ldots,c_p}\sum_{i=1}^{n}\left[\left\{Y_i-\hat{\alpha}-\sum_{k\neq d}\hat{m}_k(X_{ik})\right\}-\sum_{j=0}^{p^*}c_jX_{id}^j\right]^2$.

Thus, when $p_d = p^*$, the backfitting algorithm yields a polynomial of degree p^* as the estimate of $m_d(\cdot)$ when $h_d=\infty$. The interpretation is that the dth predictor makes a polynomial contribution of degree up to p^*. Based on this, we derive a method for detecting predictors that contribute polynomially up to degree p^* in the additive model.

Now we use backfitting with univariate local polynomial smoothing of degree p^* for every function component, again parameterizing via inverse bandwidths $\lambda_d = 1/h_d$. As in the previous section, the general approach of Stefanski et al. (2014) leads to solving

$$\min_{\lambda_1,\ldots,\lambda_D}\frac{1}{n}\sum_{i=1}^{n}\left\{y_i-\hat{\alpha}^{BF}-\sum_{d=1}^{D}\hat{m}_d^{BF}\left(X_{id};\lambda^{-1},p^*1_D\right)\right\}^2,\quad \text{subject to}\quad \lambda_d\geq 0\,(d=1,\ldots,D)\quad\text{and}\quad\sum_{d=1}^{D}\lambda_d=\tau_{p^*},\tag{4}$$

where τ_{p^*} plays the same role as τ_0 in (3), and in the last two arguments of $m_d^{BF}(.;.,.)$, $\lambda^{-1}=(1/\lambda_1,\ldots,1/\lambda_D)^{\mathrm{T}}$ and p^*1_D is the product of p^* and 1_D. In this case, an optimizer $\hat{\lambda}_d=0$ means that the dth predictor contributes polynomially at a degree no greater than p^*.

The modified coordinate descent algorithm and the tuning procedures discussed in § § 3·2 and 3·3 extend naturally to (4).

5. Automatic structure recovery

The procedures in § § 3 and 4 provide the building blocks for the additive model automatic structure recovery method. In combination, they enable estimation of the predictors that are not important, as well as those that contribute at the *p*th degree polynomially.

We let $\mathscr{A}_p$ denote the set of predictors that contribute polynomially at degree *p* and only at degree *p*, and we let $\mathscr{B}_p$ denote the set of predictors that contribute beyond polynomially degree *p*, which includes higher-degree polynomials as well as nonpolynomial functions. Thus $\mathscr{A}_p$ means at degree *p*, and $\mathscr{B}_p$ indicates beyond degree *p*. With these naming conventions, a predictor that contributes at degree $p = 0$ is no more informative than a constant and is therefore unimportant. Specifically, $\mathscr{A}_0 = \{d : m_d(\cdot) = 0\}$, $\mathscr{A}_p = \left\{d : m_d^{(p)}(\cdot) \neq 0, m_d^{(p+1)}(\cdot) = 0\right\}$ for $p = 1,2, \ldots$ and $\mathscr{B}_p = \left\{d : m_d^{(p+1)}(\cdot) \neq 0\right\}$ for $p = 0, 1, 2, \ldots$ Here $m_d(\cdot) = 0$ means that $m_d(t) = 0$ for any t in its domain of interest, and $m_d(\cdot) \neq 0$ means that $m_d(t) \neq 0$ for some t. Our automatic structure recovery method proceeds in the following steps.

Step 1

Identify important and unimportant predictors.

With appropriately chosen τ_0, solve (3) to obtain $\hat{\lambda}_0 = \left(\hat{\lambda}_{01}, \ldots, \hat{\lambda}_{0D}\right)^{\mathrm{T}}$, and set $\hat{\mathscr{A}}_0 = \left\{d : \hat{\lambda}_{0d} = 0\right\}$ and $\hat{\mathscr{B}}_0 = \hat{\mathscr{A}}_0^{\mathrm{C}}$, the complement of $\hat{\mathscr{A}}_0$ in the set $\{1, \ldots D\}$. Then $\hat{\mathscr{A}}_0$ and $\hat{\mathscr{B}}_0$ are estimates of the sets of unimportant and important predictors, respectively.

Step 2

Identify from $\hat{\mathscr{B}}_0$ the predictors that contribute linearly.

After identifying the set $\hat{\mathscr{B}}_0$ of important predictors, the next step is to identify a subset of $\hat{\mathscr{B}}_0$ consisting of functions that contribute linearly. We remove unimportant predictors X_d with $d \in \hat{\mathscr{A}}_0$ from consideration and apply the method of § 4, namely (4) with $p^* = 1$, to the data $\left\{\left(X_{i\hat{\mathscr{B}}_0}, Y_i\right) : i = 1, \ldots, n\right\}$, where $X_{i\hat{\mathscr{B}}_0}$ denotes the subvector of X_i with indices in $\hat{\mathscr{B}}_0$. We define $\hat{\alpha}^{BF}$ and $\hat{m}_d^{BF}\left(x_d; \lambda_{\hat{\mathscr{B}}_0}^{-1}, 1\right)\left(d \in \hat{\mathscr{B}}_0\right)$ to be the backfitting estimates obtained using the data $\left\{\left(X_{i\hat{\mathscr{B}}_0}, Y_i\right) : i = 1, \ldots, n\right\}$ with bandwidths λ_d^{-1} for $d \in \hat{\mathscr{B}}_0$. Let $\#\hat{\mathscr{B}}_0$ denote the cardinality of the set $\hat{\mathscr{B}}_0$. For an appropriately tuned τ_1, we solve the optimization problem

$$\min_{\lambda_d : d \in \hat{\mathscr{B}}_0} \frac{1}{n} \sum_{i=1}^{n} \left\{ Y_i - \hat{\alpha}^{BF} - \sum_{d \in \hat{\mathscr{B}}_0} \hat{m}_d^{BF}\left(X_{id}; \lambda_{\hat{\mathscr{B}}_0}^{-1}, 1_{\#\hat{\mathscr{B}}_0}\right) \right\}^2,$$
$$\text{subject to} \quad \lambda_d \geq 0 \left(d \in \hat{\mathscr{B}}_0\right) \quad \text{and} \quad \sum_{d \in \hat{\mathscr{B}}_0} \lambda_d = \tau_1,$$

and denote its optimizer by $\hat{\lambda}_{1d}$, for $d \in \hat{\mathscr{B}}_0$. The set $\hat{\mathscr{A}}_1 = \left\{d \in \hat{\mathscr{B}}_0 : \hat{\lambda}_{1d} = 0\right\}$ estimates the set of predictors contributing linearly, and $\hat{\mathscr{B}}_1 = \left\{d \in \hat{\mathscr{B}}_0 : \hat{\lambda}_{1d} > 0\right\}$ estimates the set of predictors that contribute beyond linearly.

Step 3

Identify from $\hat{\mathscr{B}}_1$ the predictors that contribute quadratically.

In Step 2 the set $\hat{\mathscr{A}}_1$ of linear predictors was identified. The set $\hat{\mathscr{A}}_1$ is global in principle, and this property can be ensured via profiling (Severini & Wong, 1992). For $d \in \hat{\mathscr{A}}_1$ we denote the global fit coefficient for the linear predictor X_d by β_{1d} or, in vector form, $\beta_{1\hat{\mathscr{A}}_1}$. For a given $\beta_{1\hat{\mathscr{A}}_1}$, we Apply the backfitting algorithm of degree $p_d = 2$ with bandwidth $\lambda_d^{-1}\left(d \in \hat{\mathscr{B}}_1\right)$ to the data $\left\{\left(X_{i\hat{\mathscr{B}}_1}, Y_i - \sum_{d \in \hat{\mathscr{A}}_1} X_{id}\beta_{1d}\right) : i = 1, \ldots, n\right\}$ and denote the corresponding estimates by $\hat{\alpha}^{BF}\left(\beta_{1\hat{\mathscr{A}}_1}\right)$ and $\hat{m}_d^{BF}\left(x_d; \lambda_{\hat{\mathscr{B}}_1}^{-1}, 2 \times 1_{\#\hat{\mathscr{B}}_1}, \beta_{1\hat{A}_1}\right)\left(d \in \hat{\mathscr{B}}_1\right)$; we have given $\hat{m}_d^{BF}$ a fourth argument to emphasize its dependence on $\beta_{1\hat{\mathscr{A}}_1}$. Using profiling, we solve

$$\min_{\beta_{1d}: d \in \hat{\mathscr{B}}_1} \frac{1}{n}\sum_{i=1}^{n}\left\{Y_i - \sum_{j \in \hat{\mathscr{A}}_1} X_{ij}\beta_{1j} - \hat{\alpha}^{BF}\left(\beta_{1\hat{\mathscr{A}}_1}\right) - \sum_{d \in \hat{\mathscr{B}}_1} \hat{m}_d^{BF}\left(X_{id}; \lambda_{\hat{\mathscr{B}}_1}^{-1}, 2 \times 1_{\#\hat{\mathscr{B}}_1}, \beta_{1\hat{\mathscr{A}}_1}\right)\right\}^2$$

to obtain the best estimate for $\beta_{1\hat{\mathscr{A}}_1}$, which we denote by $\hat{\beta}_{1\hat{\mathscr{A}}_1}\left(\lambda_{\hat{\mathscr{B}}_1}\right)$ to emphasize its dependence on $\lambda_{\hat{\mathscr{B}}_1}$. With the above notation and definitions, we solve the optimization problem

$$\min_{\lambda_d: d \in \hat{\mathscr{B}}_1} \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \sum_{j \in \hat{\mathscr{A}}_1} X_{ij}\hat{\beta}_{1j}\left(\lambda_{\hat{\mathscr{B}}_1}\right) - \hat{\alpha}^{BF}\left\{\hat{\beta}_{1\hat{\mathscr{A}}_1}\left(\lambda_{\hat{\mathscr{B}}_1}\right)\right\} - \sum_{d \in \hat{\mathscr{B}}_1} \hat{m}_d^{BF}\left\{X_{id}; \lambda_{\hat{\mathscr{B}}_1}^{-1}, 2 \times 1_{\hat{\mathscr{B}}_1}, \hat{\beta}_{1\hat{\mathscr{A}}_1}\left(\lambda_{\hat{\mathscr{B}}_1}\right)\right\}\right]^2,$$

$$\text{subject to} \quad \lambda_d \geq 0 \left(d \in \hat{\mathscr{B}}_1\right) \quad \text{and} \quad \sum_{d \in \hat{\mathscr{B}}_1} \lambda_d = \tau_2$$

for appropriately tuned τ_2, and denote its optimizer by $\hat{\lambda}_{2d}\left(d \in \hat{\mathscr{B}}_1\right)$. Then $\hat{\mathscr{A}}_2 = \left\{d \in \hat{\mathscr{B}}_1 : \hat{\lambda}_{2d} = 0\right\}$ estimates the set of predictors that contribute quadratically, and $\hat{\mathscr{B}}_2 = \left\{d \in \hat{\mathscr{B}}_1 : \hat{\lambda}_{2d} > 0\right\}$ estimates the set of predictors that contribute beyond quadratically.

Steps 4 to *M* + 1

Identify predictors that have *k*th-degree polynomial effects by using appropriate τ_k for *k* 3, 4, ….

After Step 3, we have estimated the set of linear predictors and the set of quadratic predictors. In a straightforward manner Step 3 can be modified to identify the sets of predictors that contribute polynomially at degrees $k = 3$, $k = 4$, and so on, up to $k = M$.

Step $M + 2$. Fit the final model.

After Step $M + 1$, we have obtained estimates $\hat{\mathscr{A}}_k$ for $k = 0, 1 \ldots, M$ and $\hat{\mathscr{B}}_M$. The final model is then estimated by combining the profiling technique and the backfitting technique as in Step 3. In this final step, we couple the backfitting algorithm with local linear smoothing, as we want to estimate only the function $m_d(\cdot)$ itself for $d \in \hat{\mathscr{B}}_M$.

6. Theoretical properties

In this section we study the consistency of the proposed structure recovery scheme. We first show that the estimated set of unimportant predictors is consistent.

Proposition 1

Suppose $\tau_0 \to \infty$ *such that* $\tau_0^3/n \to 0$ *as* $n \to \infty$, *and assume that Conditions* A1–A5 *in the Appendix hold. Then the solution to* (3) *satisfies* $\hat{\lambda}_d \to \infty$ *in probability for* $d \in \mathscr{B}_0$, *and* $\hat{\lambda}_d \to 0$ *in probability for* $d \in \mathscr{A}_0$. *Consequently,* $pr\left(\hat{\mathscr{A}}_0 = \mathscr{A}_0\right) \to 1$ *as* $n \to \infty$.

Remark 1

There is a small gap between Proposition 1 and the proposed procedure, which is confounded with the numerical analysis. In our procedure $\hat{\mathscr{A}}_0$ is defined as $\left\{d : \hat{\lambda}_d = 0\right\}$, but Proposition 1 shows that the solution to (3) satisfies $\hat{\lambda}_d \to 0$ in probability for $d \in \mathscr{A}_0$. Thus there would be closer agreement between the proposition and the algorithm if we had defined $\hat{\mathscr{A}}_0$ as $\left\{d : \hat{\lambda}_d < \delta\right\}$ for some small $\delta > 0$. However, based on our numerical experience so far, the optimizer is indeed sparse, returning exact zeros. Therefore, we shall keep the definition of $\hat{\mathscr{A}}_0$ as $\left\{d : \hat{\lambda}_d = 0\right\}$. This remark also applies to Theorem 1.

The consistency in Proposition 1 is readily extended to the estimated set of predictors that contribute polynomially at different degrees, resulting in the following consistency property for the proposed structure recovery scheme.

Theorem 1

Suppose that for $k = 0, 1, \ldots, M$, $\tau_k \to \infty$ *and* $\tau_k^{2k+3}/n \to 0$ *as* $n \to \infty$, *and assume that Conditions* A1–A5 *in the Appendix hold. Then the estimators* $\hat{\mathscr{A}}_k\ (k = 0, \ldots, M)$ *and* $\hat{\mathscr{B}}_M$ *satisfy* $pr\left(\hat{\mathscr{A}}_k = \mathscr{A}_k \text{ for } k = 0, \ldots, M, \hat{\mathscr{B}}_M = \mathscr{B}_M\right) \to 1$ *as* $n \to \infty$.

The proofs of Proposition 1 and Theorem 1 are given in the Appendix.

7. Simulation studies

We use simulation models adapted from Zhang et al. (2011) to study the finite-sample performance of the proposed method. In each model, predictors are generated as $X_j = (U_j + \eta U)/(1 + \eta)$ for $j = 1, \ldots, D$, where $U_1, \ldots, U_D, U$ are independent and identically distributed Un[0, 1] variables, with η at levels 0 and 0·5, resulting in pairwise correlations of 0 and 0·2. We compare our method with the smoothing spline analysis-of-variance method (Kimeldorf & Wahba, 1971; Gu, 2002), as well as the linear and nonlinear discovery method of Zhang et al. (2011) and its refitted version, in terms of both integrated squared error and predictor-type identification.

For an estimate $\hat{f}(x)$ of the additive model (1), define the integrated squared error $\mathrm{ISE}(\hat{\mathrm{f}}) = \mathrm{E}_{\mathrm{x}}\{\mathrm{f}(\mathrm{X}) - \hat{\mathrm{f}}(\mathrm{X})\}^2$, estimated via Monte Carlo using an independent test set of size 1000. The linear and nonlinear discovery methods, original and refitted, are designed to identify predictors having null, linear, nonlinear, or a mixture of linear and nonlinear effects, whereas the goal of the proposed method is to identify predictors that contribute polynomially at different degrees and predictors that contribute beyond polynomially of a specified degree M. In the simulation studies, we fixed M to be 2. We use the Bayesian information criterion for tuning parameter selection in all methods.

Model 1

In the first data-generation model, $Y = f_1(X_1) + f_2(X_2) + f_3(X_3) + \varepsilon$ with $f_1(t) = 3t$, $f_2(t) = 2\sin(2\pi t)$ and $f_3(t) = 2(3t - 1)^2$. Here $\varepsilon \sim N(0, \sigma^2)$ is independent of the predictors. Two values of the standard deviation, $\sigma = 1$ and $\sigma = 2$, and two dimensions, $D = 10$ and $D = 20$, are considered. For all settings, $X_4, \ldots, X_D$ are unimportant. In terms of the linear and nonlinear discovery classification scheme, X_1 has a purely linear effect, X_2 has a purely nonlinear effect, and X_3 has a mixture of linear and nonlinear effects (Zhang et al., 2011). For our method, X_1 is a linear predictor, X_3 is a quadratic predictor, and X_2 contributes beyond quadratically. Training samples of size n =100 are used. The results from 100 simulated datasets are reported in Tables 1 and 2.

Table 1 summarizes the performance of the smoothing spline method, the two linear and nonlinear discovery methods, the proposed polynomial structure classification method, and two oracle methods, which are included to assess the utility of the underlying classification schemes irrespective of estimation error. Oracle 1 is the linear and nonlinear discovery oracle that makes use of information on which predictors are noise, purely linear, purely nonlinear, or a mixture of linear and nonlinear. Oracle 2 is the oracle for our method and uses information on which predictors are noise, linear, quadratic, or beyond quadratic. The polynomial structure classification method has the smallest integrated squared error values among all methods for all generating distributions, and oracle 2 dominates oracle 1.

Table 2 summarizes the predictor classification performance of the methods under comparison. For the linear and nonlinear discovery method, the table entries are the average and standard deviation of the numbers of predictors identified as purely linear (X_1), purely nonlinear (X_2), a linear-nonlinear mixture (X_3), and noise ($X_4, \ldots, X_D$). Similarly, for the

proposed method, the table entries are the average and standard deviation of the numbers of predictors identified as linear (X_1), quadratic (X_3), beyond quadratic (X_2), and noise ($X_4, \ldots, X_D$). Also reported are the percentages of times that all variables were correctly classified according to each method's underlying classification scheme.

The proposed polynomial structure classification method performs well for all generative models, although both it and the linear and nonlinear discovery method generally perform less well for larger values of D, σ and η. Because the two methods are based on different classification schemes, predictor classification is not directly comparable, except with respect to the noise variables. The proposed method identified all noise variables for all 100 simulated datasets for all generative models, whereas the linear and nonlinear discovery method missed some noise predictors, more in cases where σ, η and especially D were large.

For one simulated dataset with $\sigma = 1$, $\eta = 0{\cdot}5$ and $D = 10$, we plot in Fig. 1 the nonzero optimizers $\hat{\lambda}_j$ of (3) in the first variable selection step as functions of τ_0. The optimizers $\hat{\lambda}$ corresponding to the important predictors, X_1, X_2 and X_3, change to nonzero values quickly as τ_0 increases and before any unimportant predictors show changes. Only four lines are visible in the graph because the optimizers corresponding to the other six predictors are zero for $0 \leq \tau_0 \leq 45$.

Model 2

In the second data-generation model, $Y = \sum_{j=1}^{7} f_j(X_j) + \varepsilon$ with $f_1(t)=3t$, $f_2(t)=-4t$, $f_3(t)=2t$, $f_4(t)=2\sin(2\pi t)$, $f_5(t)=3\sin(2\pi t)/\{2-\sin(2\pi t)\}$, $f_6(t)=5[0{\cdot}1\sin(2\pi t) + 0{\cdot}2\cos(2\pi t) + 0{\cdot}3\{\sin(2\pi t)\}^2 + 0{\cdot}4\{\cos(2\pi t)\}^3 + 0{\cdot}5\{\sin(2\pi t)\}^3] + 2t$ and $f_7(t)=2(3t-1)^2$. Here $\varepsilon \sim N(0, \sigma^2)$ is independent of the predictors, with σ controlling the noise level; we take $\sigma = 1$ or 2. The dimension is $D = 20$ and so there are 13 noise predictors, $X_8, \ldots, X_{20}$. For the linear and nonlinear discovery methods, X_1, X_2 and X_3 are linear, X_4 and X_5 are nonlinear, and X_6 and X_7 are mixed linear-nonlinear. For our method, X_1, X_2 and X_3 are linear, X_7 is quadratic, and X_4, X_5 and X_6 are beyond quadratic. Training samples of size $n = 250$ are used. Tables 3 and 4 present the results from 100 simulated datasets in the same format as for Model 1. Conclusions about the performance of the methods are similar to those for Model 1.

8. Extension to partially linear models

The partially linear model is an extension of the additive model where, in addition to predictors $X_1, \ldots, X_D$, there are predictors $Z_1, \ldots, Z_q$ known to contribute linearly, so that

$$Y = \alpha + \sum_{d=1}^{D} f_d(X_d) + \sum_{j=1}^{q} Z_j \gamma_j + \varepsilon. \quad (5)$$

The Z_j commonly include indicators for categorical variables such as gender, race and location.

We illustrate the extension of our method to the partially linear model (5) via profiling by analysing the diabetes data from Willems et al. (1997). The goal of that study was to

understand the prevalence of obesity, diabetes and other cardiovascular risk factors. The data include 18 variables on each of 403 African American subjects from central Virginia, with some missing values. The response variable in our analysis is glycosolated haemoglobin, which is of interest because a value greater than 7·0 is usually regarded as giving a positive diagnosis of diabetes.

We exclude two variables, the second systolic blood pressure and second diastolic blood pressure, as they are missingness-prone replicates of two other included variables, the first systolic and first diastolic blood pressures. Doing so leaves 15 candidate predictors. Twelve of these 15 variables are continuous and are rescaled to the unit interval: X_1 = cholesterol, X_2 = stabilized glucose, X_3 = high density lipoprotein (hdl), X_4 = cholesterol to hdl ratio, X_5 = age, X_6 = height, X_7 = weight, X_8 = first systolic blood pressure, X_9 = first diastolic blood pressure, X_{10} = waist circumference, X_{11} = hip circumference, and X_{12} = postprandial time when samples were drawn. The remaining three variables are categorical variables for location, gender and frame. Location is a factor with two levels, Buckingham and Louisa; we set $Z_1 = 0$ for Buckingham and $Z_1 = 1$ otherwise. For gender we set $Z_2 = 0$ for female and $Z_2 = 1$ for male. The frame variable has three levels; we set $Z_3 = 0$ and $Z_4 = 0$ for small frames, $Z_3 = 1$ and $Z_4 = 0$ for medium frames, and $Z_3 = 0$ and $Z_4 = 1$ for large frames.

We fit the partially linear model (5) to study the dependence of glycosolated haemoglobin on the variables $X_1, \ldots, X_{12}, Z_1, \ldots, Z_4$, using the natural extension of the automatic structure recovery algorithm to discern the effects of $X_1, \ldots, X_{12}$. We use the Bayesian information criterion to select tuning parameters. Our method identified $X_1, X_3, X_6, \ldots, X_{12}$ as unimportant predictors; X_5 was selected to have a linear effect, X_4 a quadratic effect, and X_2 a beyond-quadratic effect.

As Fig. 2(a) shows, on the original scale X_4 has one outlier, 19·3, far outside the range of the other observations, which are between 1·5 and 12·2. Upon removing the outlier and reapplying our method, $X_1, X_3, X_4, X_6, \ldots, X_{12}$ are identified as unimportant, X_5 as having a linear effect, and X_2 as having a beyond-quadratic effect. The nonlinear fit for X_2 is shown in Fig. 2(b).

9. Discussion

We have proposed and studied the properties of a new automatic structure recovery and estimation method for the additive model. The method is readily generalizable and can be extended to generalized additive models (Hastie & Tibshirani, 1990) and survival data models (Cheng & Lee, 2009).

Classifying predictors into those that are unimportant and those that contribute through polynomials of specified degree is accomplished by using a nonparametric variable selection method based on the results of Stefanski et al. (2014). Like the nonparametric classification method proposed by Stefanski et al. (2014), each step of our new method uses regularization by bounding a sum of inverse kernel bandwidths. Other choices of penalty, such as $\sum_d \lambda_d^\gamma$ for fixed $\gamma \geq 0$, are possible and worthy of study. Based on our limited experience, $\gamma = 1$ gives good overall performance. However, in light of the results in Stefanski et al. (2014), it

is expected that $\gamma < 1$ would increase sparsity whereas $\gamma > 1$ would decrease sparsity but yield regression function estimators that have better mean-squared error properties.

As noted by a referee, modifications to our method and additional development of the theory to accommodate the case of a diverging dimension D would be useful directions to pursue. Alternatively, for data with large D, one could precede the use of our method with application of one of the screening procedures recently developed by Fan et al. (2011) and Cheng et al. (2014).

Acknowledgement

We thank two reviewers, an associate editor, and the editor for their most helpful comments. This research was supported by the U.S. National Institutes of Health and National Science Foundation.

Appendix

For $d = 1, \ldots, D$, denote the density of X_d by $f_d(\cdot)$. We assume the following technical conditions from Fan & Jiang (2005), as our method is based on backfitting coupled with local polynomial smoothing.

Condition A1

The kernel function $K(\cdot)$ is bounded and Lipschitz continuous with bounded support.

Condition A2

The densities $f_d(\cdot)$ are Lipschitz continuous and bounded away from zero over their bounded supports.

Condition A3

For all pairs d and d', the joint density functions of X_d and X_d are Lipschitz continuous on their supports.

Condition A4

For all $d = 1, \ldots, D$, the $(p_d + 1)$th derivatives of $m_d(\cdot)$ exist and are bounded and continuous.

Condition A5

The error has finite fourth moment, $E(|\varepsilon|^4) < \infty$.

For the backfitting estimate coupled with local polynomial smoothing as defined in § 3, we have the following asymptotic result.

Lemma A1

Assume that Conditions A1–A5 *hold and* $h_d \to 0$ *such that* $nh_d^{2p_d+2} \to \infty$ *as n* $n \to \infty$, *for d* = 1, … , *D. Then*

$$\frac{1}{n}\sum_{i=1}^{n}\left\{Y_i - \hat{\alpha}^{BF} - \sum_{d=1}^{D}\hat{m}_d^{BF}\left(X_{id};h,p\right)\right\}^2 = \sigma^2 + O_{\mathrm{p}}\left(\sum_{d=1}^{D}h_d^{2p_d+2}\right) + O_{\mathrm{p}}\left(\sum_{d=1}^{D}\frac{1}{nh_d}\right). \quad \text{(A1)}$$

Proof

Lemma A1 is a straightforward consequence of the third step in the proof of Theorem 1 in Fan & Jiang (2005). The term on the left-hand side of (A1) is RSS_1/n in their notation, which they show converges in probability to σ^2. Fan & Jiang (2005) did not explicitly state the rate of convergence, but it is readily deduced from their proof. We use their notation in the rest of this proof.

Note that $\mathrm{RSS}_1 = \varepsilon A_{n2}\varepsilon + B^{\mathrm{T}}B + 2B^{\mathrm{T}}(W_M - I_n)\varepsilon$ and $A_{n2} = I_n + S^{\mathrm{T}}S - S - S^{\mathrm{T}} + R_{n2}$, according to Fan & Jiang (2005, p. 903). Thus it is enough to track the convergence rates of the different terms. Also, $B = O\left(\sum_{d=1}^{D}h_d^{p_d+1}\right) + \left(\sum_{d=1}^{D}\bar{m}_d\right)O(1)$ almost surely, according to (B.4) in Fan & Jiang (2005), and $\sum_{d=1}^{D}\bar{m}_d\Big) = O_p\left(n^{-1/2}\right)$ by definition. Consequently, $B^TB/n = O_{\mathrm{p}}\left(\sum_{d=1}^{D}h_d^{2p_d+2}\right)$. By line 6 from the bottom of the right-hand column of p. 901 in Fan & Jiang (2005), $B^T\left(W_M - I_n\right)\varepsilon/n = O_p\left(n^{-1} + \sum_{d=1}^{D}n^{-1/2}h_d^{p_d+1}\right)$. By (B.11)–(B.24), $\varepsilon^{\mathrm{T}}S^{\mathrm{T}}S/n - S - S^{\mathrm{T}}\varepsilon = O_p\left\{\sum_{d=1}^{D}1/\left(nh_d\right)\right\}$. As $R_{n2} = O(1^{\mathrm{T}}1/n)$ uniformly over its elements, $\varepsilon^{\mathrm{T}}R_{n2}\varepsilon/n = O_{\mathrm{p}}\left\{\sum_{d=1}^{D}1/\left(nh_d\right)\right\}$. Combining these terms and noting that $nh_d^{2p_d+2} \to \infty$ completes the proof.

For $k = 0, 1, \ldots$, let π_k be the set of all polynomial functions of degree k. Define $\mathscr{I} = \{d : m_d(\cdot) \in \pi_{p_d}, d = 1, \ldots, D\}$ to be the set of predictors with corresponding additive component functions which are polynomial of degree at most that of the corresponding local polynomial used in the backfitting algorithm. Denote its complement by $\mathscr{I}^{\mathrm{C}} = 1\{1, \ldots, D\} \setminus \mathscr{I}$.

Proposition A1

Suppose that Conditions A1–A5 *hold, that* $h_d \geq c_0 > 0$ *for* $d \in \mathscr{I}$ *and some constant* $c > 0$, *and that* $h_{d'} \to 0$ *and* $nh_{d'}^{2p_d+2} \to \infty$ *as* $n \to \infty$ *for* $d' \in \mathscr{I}^{\mathrm{C}}$. *Then*

$$\frac{1}{n}\sum_{i=1}^{n}\left\{Y_i-\hat{\alpha}^{BF}-\sum_{d=1}^{D}\hat{m}_d^{BF}\left(X_{id};h,p\right)\right\}^2=\sigma^2+O_{\mathrm{p}}\left(\sum_{d\in\mathscr{I}^{\mathrm{c}}}h_d^{2p_d+2}\right)+O_{\mathrm{p}}\left(\sum_{d\in\mathscr{I}^{\mathrm{c}}}\frac{1}{nh_d}\right). \quad \text{(A2)}$$

Proof

Opsomer & Ruppert (1999) studied a backfitting estimator for semiparametric additive models and showed that the estimator of the parametric component is $n^{1/2}$-consistent. The condition $h_d \geq c_0 > 0$ for $d \in \mathscr{I}$ and some $c_0 > 0$ implies that the smoothing bandwidth h_d is bounded away from zero for $d \in \mathscr{I}$. As $m_d(\cdot) \in \pi_{p_d}$ for $d \in \mathscr{I}$, there is no approximation bias in using local polynomial smoothing to estimate the corresponding component in the backfitting algorithm. Thus the techniques of Opsomer & Ruppert (1999) can be used to show that the estimator of $m_d(\cdot)$ for $d \in \mathscr{I}$ is $n^{1/2}$-consistent as $h_d > c_0 > 0$. The $n^{1/2}$-consistent rate is faster than the consistency rate for smoothing other components.

Proof of Proposition 1

It is easy to show by contradiction that $\hat{\lambda}_d \to \infty$ in probability for $d \in \mathscr{B}_0$. If $\hat{\lambda}_{d'}$ is bounded for some $d' \in \mathscr{B}_0$, then the objective function of (3) converges to the sum of σ^2 plus an additional positive term due to smoothing bias. The additional bias term is caused by bounded $\hat{\lambda}_{d'}$, as the corresponding smoothing bandwidth $h_{d'}=1/\hat{\lambda}_{d'}$ does not shrink to zero. This is suboptimal, as the smallest limit of the objective is σ^2. This proves that $\hat{\lambda}_d \to \infty$ in probability for $d \in \mathscr{B}_0$.

In (3), local constant smoothing is used for every component function. Thus $p_d = 0$ for d 1, … , D. The second term on the right-hand side of (A1) or (A2) is due to bias, and the third term is due to variance when using local polynomial smoothing for every component. The condition $\tau_0^3/n \to 0$ as $n \to \infty$ ensures that the variance term is dominated by the bias term. At the same time, note that a bounded and small λ_d for $d \in \mathscr{A}_0$ does not introduce any extra term.

If $\sum_{d\in\mathscr{A}_0}\hat{\lambda}_d \nrightarrow 0$ as $n \to \infty$, consider $\tilde{\lambda}_d=\hat{\lambda}_d\tau_0/\left(\tau_0-\sum_{d'\in\mathscr{A}_0}\hat{\lambda}_{d'}\right)$ for $d \in \mathscr{B}_0$ and $\tilde{\lambda}_d=0$ for $d \in \mathscr{A}_0$. In this case, $\tilde{\lambda}_d$ diverges to infinity at a faster rate than $\hat{\lambda}_d$ for $d \in \mathscr{B}_0$. Consequently, the asymptotic bias term on the right-hand side of (A2) using the $\tilde{\lambda}_d$ values is smaller than the bias term on the right-hand side of (A1) using $\hat{\lambda}_d$ values. Here, smaller is in the sense of asymptotic order if $\sum_{d\in\mathscr{A}_0}\hat{\lambda}_d \to \infty$, and in the sense of the constant multiplying the asymptotic order if $\sum_{d\in\mathscr{A}_0}\hat{\lambda}_d$ is bounded. Thus the set of $\hat{\lambda}_d$ values with $\sum_{d\in\mathscr{A}_0}\hat{\lambda}_d \nrightarrow 0$ as $n \to \infty$ is suboptimal, as we are solving a minimization problem (3). This shows that $\hat{\lambda}_d \to 0$ in probability for $d \in \mathscr{A}_0$ and therefore completes the proof.

Proof of Theorem 1

From Proposition 1, we have that $pr\left(\hat{\mathscr{A}}_0=\mathscr{A}_0\right)\to 1$ as $n\to\infty$. Conditional on $\hat{\mathscr{A}}_0=\mathscr{A}_0$, we can prove that $pr\left(\hat{\mathscr{A}}_1=\mathscr{A}_1\right)\to 1$ as $n\to\infty$ using arguments similar to those in the proof of Proposition 1. This process is iterated to complete the proof of the theorem.

References

Cheng M-Y, Lee C-Y. Local polynomial estimation of hazard rates under random censoring. J. Chin. Statist. Assoc. 2009; 47:19–38.

Cheng M-Y, Honda T, Li J, Peng H. Nonparametric independence screening and structure identification for ultra-high dimensional longitudinal data. Ann. Statist. 2014; 42:1819–49.

Daubechies IB, Defrise M, De Mol C. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. Commun. Pure Appl. Math. 2004; 57:1413–57.

Engle RF, Granger CWJ, Rice J, Weiss A. Nonparametric estimates of the relation between weather and electricity sales. J. Am. Statist. Assoc. 1986; 81:310–20.

Fan, J.; Gijbels, I. Local Polynomial Modelling and Its Applications. Chapman & Hall; London: 1996.

Fan J, Jiang J. Nonparametric inference for additive models. J. Am. Statist. Assoc. 2005; 100:890–907.

Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Statist. Assoc. 2001; 96:1348–60.

Fan J, Lv J. A selective overview of variable selection in high dimensional feature space. Statist. Sinica. 2010; 20:101–48.

Fan J, Feng Y, Song R. Nonparametric independence screening in sparse ultra-high dimensional additive models. J. Am. Statist. Assoc. 2011; 106:544–57.

Fan J, Ma Y, Dai W. Nonparametric independence screening in sparse ultra-high dimensional varying coefficient models. J. Am. Statist. Assoc. 2014; 109:1270–84.

Friedman JH, Stuetzle W. Projection pursuit regression. J. Am. Statist. Assoc. 1981; 76:817–23.

Friedman JH, Hastie TJ, Hofling H, Tibshirani RJ. Pathwise coordinate optimization. Ann. Appl. Statist. 2007; 1:302–32.

Fu WJ. Penalized regressions: The bridge versus the lasso. J. Comp. Graph. Statist. 1998; 7:397–416.

Gu, C. Smoothing Spline ANOVA Models. Springer; New York: 2002.

Hastie, TJ.; Tibshirani, RJ. Generalized Additive Models. Chapman & Hall; London: 1990.

Huang J, Horowitz JL, Wei F. Variable selection in nonparametric additive models. Ann. Statist. 2010; 38:2282–313.

Huang J, Wei F, Ma S. Semiparametric regression pursuit. Statist. Sinica. 2012; 22:1403–26.

Kimeldorf GS, Wahba G. Some results on Tchebycheffian spline functions. J. Math. Anal. Applic. 1971; 33:82–5.

Opsomer JD, Ruppert D. A root-*n* consistent backfitting estimator for semiparametric additive modeling. J. Comp. Graph. Statist. 1999; 8:715–32.

Ravikumar PK, Lafferty JD, Liu H, Wasserman L. Sparse additive models. J. R. Statist. Soc. B. 2009; 71:1009–30.

Severini TA, Wong WH. Profile likelihood and conditionally parametric models. Ann. Statist. 1992; 20:1768–802.

Shen X, Pan W, Zhu Y, Zhou H. On constrained and regularized high-dimensional regression. Ann. Inst. Statist. Math. 2013; 65:807–32.

Stefanski LA, Wu Y, White K. Variable selection in nonparametric classification via measurement error model selection likelihoods. J. Am. Statist. Assoc. 2014; 109:574–89.

Tibshirani RJ. Regression shrinkage and selection via the lasso. J. R. Statist. Soc. B. 1996; 58:267–88.

Willems JP, Saunders JT, Hunt DE, Schorling JB. Prevalence of coronary heart disease risk factors among rural blacks: A community-based study. Southern Med. J. 1997; 90:814–20. [PubMed: 9258308]

Wu TT, Lange K. Coordinate descent procedures for lasso penalized regression. Ann. Appl. Statist. 2008; 2:224–44.

Xia Y, Zhang W, Tong H. Efficient estimation for semivarying-coefficient models. Biometrika. 2004; 91:661–81.

Zhang HH, Lu W. Adaptive lasso for Cox's proportional hazards model. Biometrika. 2007; 94:691–703.

Zhang HH, Cheng G, Liu Y. Linear or nonlinear? Automatic structure discovery for partially linear models. J. Am. Statist. Assoc. 2011; 106:1099–112.

Zou H. The adaptive lasso and its oracle properties. J. Am. Statist. Assoc. 2006; 101:1418–29.

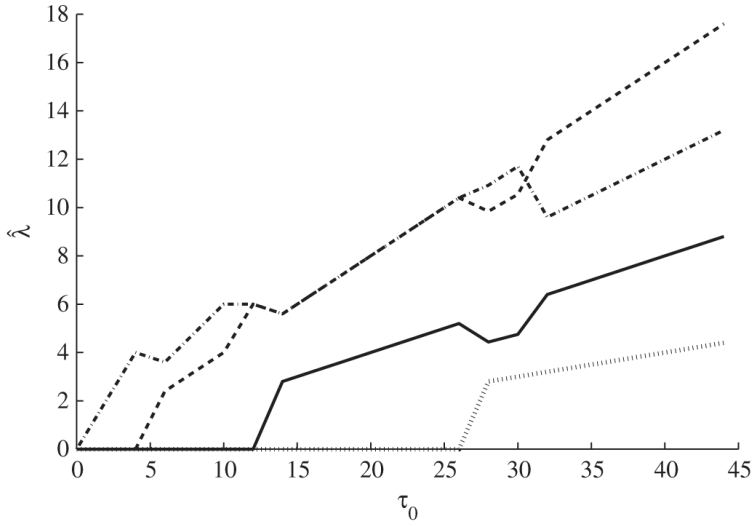

Fig. 1.

Solution paths to optimization problem (3) for one simulated dataset from Model 1, plotted over $0 < \tau_0 \leq 45$. At $\tau_0 = 40$ the paths are, from bottom to top, for variables X_9, X_1, X_3 and X_2.

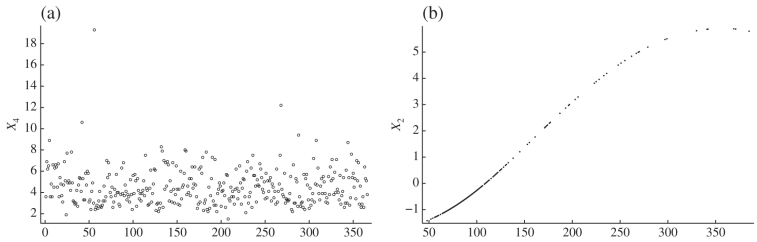

Fig. 2.

Results of applying our method to the diabetes data: (a) index plot of variable X_4, with an outlier visible in the upper left corner; (b) the beyond-quadratic fit for X_2 (stabilized glucose) after removing the X_4 outlier.

Table 1

Simulation results for Model 1: *average integrated squared errors* ($\times 10^2$) *with standard deviations* ($\times 10^2$) *in parentheses*

D	σ	η	Oracle 1	SSANOVA	LAND	LANDr	PSC	Oracle 2
10	1	0·0	12·5 (5·5)	23·9 (9·0)	11·9 (7·4)	14·9 (6·8)	11·6 (6·2)	9·8 (4·9)
		0·5	13·2 (5·4)	25·5 (8·8)	58·9 (36·8)	17·2 (7·3)	12·3 (5·9)	10·4 (4·4)
	2	0·0	44·0 (22·9)	92·7 (41·6)	82·1 (48·0)	71·2 (42·4)	52·0 (34·7)	37·3 (21·9)
		0·5	50·3 (26·3)	100·4 (35·9)	142·6 (47·9)	88·1 (38·0)	67·4 (34·9)	42·8 (21·7)
20	1	0·0	12·8 (5·4)	45·5 (14·3)	16·4 (12·6)	18·7 (9·7)	14·8 (7·8)	9·9 (4·4)
		0·5	13·4 (5·7)	50·6 (19·6)	64·4 (36·0)	24·4 (11·6)	14·3 (8·9)	11·2 (5·4)
	2	0·0	45·1 (21·0)	187·0 (65·8)	135·8 (70·9)	130·7 (60·1)	69·3 (40·7)	37·8 (18·5)
		0·5	45·7 (18·9)	192·8 (62·8)	169·5 (63·5)	155·3 (63·8)	76·2 (41·7)	40·4 (19·6)

D, dimension; σ, model error standard deviation; η, predictor correlation parameter; SSANOVA, smoothing spline analysis-of-variance method; LAND, linear and nonlinear discovery method; LANDr, linear and nonlinear discovery method, refitted version; PSC, the proposed polynomial structure classification method; Oracle 1, oracle for LAND and LANDr; Oracle 2, oracle for PSC.

Table 2

Summary of predictor classification performance of the linear and nonlinear discovery method and the proposed method for Model 1

LAND								PSC				
D	***σ***	***η***	**L**	**NL**	**L-N**	**Nil**	**CC**	**L**	**Q**	**BQ**	**Nil**	**CC**
10	1	0·0	0·9	0·8	1·0	6·5	54	0·9	1·0	1·0	7·0	91
		0·5	1·0	0·2	0·9	5·9	10	1·0	1·0	1·0	7·0	92
	2	0·0	1·0	0·4	1·0	4·8	5	0·9	1·0	0·9	7·0	74
		0·5	1·0	0·0	0·7	3·8	0	0·6	0·9	0·6	7·0	35
Average SD			0·1	0·4	0·2	1·5		0·3	0·2	0·2	0·0	
20	1	0·0	1·0	0·7	1·0	15·6	34	1·0	1·0	1·0	17·0	90
		0·5	1·0	0·2	0·9	14·8	4	1·0	1·0	1·0	17·0	93
	2	0·0	1·0	0·4	1·0	11·0	1	0·9	0·9	0·9	17·0	72
		0·5	0·9	0·1	0·7	10·1	0	0·6	0·9	0·6	17·0	33
Average SD			0·1	0·4	0·2	2·5		0·3	0·3	0·2	0·0	

LAND, linear and nonlinear discovery method; PSC, the proposed polynomial structure classification method; *D*, dimension; σ, model error standard deviation; η, predictor correlation parameter; L, linear; NL, nonlinear; L-N, linear-nonlinear mixture; Nil, noise; CC, proportion (%) of models with fully correct variable classifications; Q, quadratic; BQ, beyond quadratic; Average SD, block-averaged standard deviation.

Table 3

Simulation results for Model 2: *average integrated squared errors* ($\times 10^2$) *with standard deviations* ($\times 10^2$) *in parentheses*

σ	η	Oracle 1	SSANOVA	LAND	LANDr	PSC	Oracle 2
1	0·0	15 (4)	25 (6)	16 (5)	17 (5)	14 (4)	14 (4)
	0·5	15 (4)	25 (6)	16 (5)	17 (6)	15 (6)	13 (4)
2	0·0	52 (14)	91 (22)	58 (20)	66 (19)	56 (17)	48 (13)
	0·5	47 (16)	86 (24)	77 (34)	73 (31)	64 (26)	44 (15)

σ, model error standard deviation; η, predictor correlation parameter; SSANOVA, smoothing spline analysis-of-variance method; LAND, linear and nonlinear discovery method; LANDr, linear and nonlinear discovery method, refitted version; PSC, the proposed polynomial structure classification method; Oracle 1, oracle for LAND and LANDr; Oracle 2, oracle for PSC.

Table 4

Summary of predictor classification performance of the linear and nonlinear discovery method and the proposed method for Model 2

LAND							PSC				
σ	η	L	NL	LN	Nil	CC	L	Q	BQ	Nil	CC
1	0·0	3·0	1·6	2·0	12·5	47	3·0	1·0	3·0	13·0	99
	0·5	3·0	1·1	2·0	11·8	16	2·9	1·0	3·0	13·0	89
2	0·0	3·0	1·5	1·9	11·2	14	2·8	1·0	3·0	13·0	74
	0·5	3·0	0·8	1·8	9·7	1	2·2	1·0	2·8	13·0	23
Average SD		0·2	0·6	0·2	2·0		0·3	0·1	0·1	0·1	

LAND, linear and nonlinear discovery method; PSC, the proposed polynomial structure classification method; σ, model error standard deviation; η, predictor correlation parameter; L, linear; NL, nonlinear; L-N, linear-nonlinear mixture; Nil, noise; CC, proportion (%) of models with fully correct variable classifications; Q, quadratic; BQ, beyond quadratic; Average SD, block-averaged standard deviation.