Statistica Sinica **19** (2009), 1281-1296

CONSISTENT VARIABLE SELECTION IN ADDITIVE MODELS

Lan Xue

Oregon State University

Abstract: We propose a penalized polynomial spline method for simultaneous model estimation and variable selection in additive models. It approximates nonparametric functions by polynomial splines, and minimizes the sum of squared errors subject to an additive penalty on norms of spline functions. This approach sets estimators of certain function components to zero, thus performing variable selection. Under mild conditions, we show that the newly proposed method estimates the non-zero function components in the model with the same optimal mean square convergence rate as the standard polynomial spline estimators, and correctly sets the zero function components to zero with probability approaching one, as n goes to infinity. Besides being theoretically justified, the proposed method is easy to understand and straightforward to implement. Extensive Monte Carlo simulation studies show the newly proposed method compares favorably with the existing ones in finite sample performance. We also illustrate the use of the proposed method by analyzing two data sets.

Key words and phrases: Boston housing price, knot, mean square consistency, ozone data, penalized least squares, SCAD.

1. Introduction

Variable selection is of special importance in multivariate regression analysis. By effectively identifying the subset of important variables, the variable selection can not only enhance model interpretability, but also improve its prediction accuracy. Since the seminal work of Akaike (1973), many methods have been developed for variable selection.

Commonly used methods for variable selection are based on assumptions that the data are generated by linear regression models. The simple linear structure enables easy estimation of model parameters, and convenient construction of variable selection criteria such as Mallows's C_p (Mallows (1973)), the Akaike information criterion (AIC) (Akaike (1973)), and the Bayesian information criterion (BIC) (Schwarz (1978)). These criteria balance goodness of fit and model complexity with a fixed penalty on the number of non-zero coefficients. Recently, a rather modern approach emerged and gained popularity. The idea is to minimize a penalized least squares with the penalty being an irregular function of the regression coefficients. Examples include bridge regression (Frank and Friedman (1993), the nonnegative garrote (Breiman (1995)), the least absolute shrinkage and selection operator (LASSO) (Tibshirani (1996)), SCAD (Fan and Li (2001)), and least angle regression (LARS) (Efron, Hastie, Johnstone and Tibshirani (2004)). Recent works of Yuan and Lin (2006, 2007) successfully extended the nonnegative garrote, LASSO and LARS to select grouped variables, which is particularly useful in multi-level ANOVA models. What makes these methods attractive is that the penalized least squares shrink some coefficients to zero, thus simultaneously selecting the variables and estimating the coefficients without an exhaustive search over all candidate models.

Although the aforementioned methods are useful for selecting significant variables in many applications, their proper use is restricted to linear models. Many data in application, however, exhibit strong non-linearity. Various non- and semiparametric methods have been used successfully to model non-linearity, due to their ability to discover data structure that linear and parametric models fail to detect. Of special importance is the additive model (Stone (1985) and Hastie and Tibshirani (1990)) that relaxes the strict linear assumption, but retains the interpretable additive form of linear regression models. More importantly, the additive model circumvents the so-called 'curse of dimensionality' arising in multivariate nonparametric function estimation. In fact, Stone (1985) showed that the additive model can be estimated at the optimal rate of convergence for univariate functions. More recent works related to additive model include Linton and Nielsen (1995), Linton and Härdle (1996), Sperlich, Tjøstheim and Yang (2002), Yang, Sperlich and Härdle (2003) and Xue and Yang (2006a,b).

Variable selection in additive models has also been considered by several authors. Chen and Tsay (1993) suggested the use of the adaptive backfitting BRUTO algorithm; (Hastie (1989)) looked at time series data; Shively, Kohn and Wood (1999) presented a hierarchical Bayesian approach in a nonparametric manner; Lin and Zhang (2006) proposed a component selection and smoothing operator (COSSO) for model selection in more general functional ANOVA models. The work most related to ours is Huang and Yang (2004), which proposed nonparametric extensions of BIC and AIC via polynomial spline smoothing. Although shown to be consistent, the polynomial spline BIC method can be computationally inefficient for even moderates size of covariates, since it needs to search over all candidate models, and the estimation of each nonparametric additive sub-model is not easy.

In this article, we propose an efficient penalized polynomial spline estimation method for additive models. Unlike the polynomial spline BIC in Huang and Yang (2004), it selects significant variables and estimates nonparametric functions simultaneously, without exhaustive search over all candidate models. It is a nonparametric extension of the SCAD (Fan and Li (2001); and Hunter and Li (2005)), which recently gained popularity for variable selection in linear models. Like the SCAD, the proposed penalized polynomial spline method is shown to have oracle properties: when there are zero function components in the true model, they are shrunk to zero with probability approaching to 1; non-zero function components are estimated as accurately as the standard polynomial spline estimates of the correct submodel can manage.

The article proceeds as follows. Section 2 describes the additive model. Section 3 introduces the penalized polynomial spline and discusses some implementation issues, such as knot number and tuning parameter selection. Section 4 presents the asymptotic properties of the proposed estimator. Section 5 contains simulation studies and applications. The proofs are given in the Appendix.

2. The Model

Let Y be a variable of interest and $\mathbf{X} = (X_1, \ldots, X_d)^T$ be a vector of predictor variables. The additive model assumes

$$m(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x}) = \alpha_0 + \sum_{l=1}^{d} \alpha_l(x_l), \tag{2.1}$$

where $\mathbf{x} = (x_1, \ldots, x_d)^T$, α_0 is an unknown constant, and $\{\alpha_l(\cdot)\}_{l=1}^d$ are unknown nonparametric functions. For identifiability of (2.1), one assumes $E\{\alpha_l(X_l)\} = 0$, for $l = 1, \ldots, d$. Suppose now that a random sample $\{\mathbf{X}_i, Y_i\}_{i=1}^n$ has been observed. Then $\alpha_0 = E(Y)$, and can be consistently estimated by $\hat{\alpha}_0 = \sum_{i=1}^n Y_i/n$ at the rate of $1/\sqrt{n}$, which is faster than any rate of convergence for nonparametric function estimation. Thus for notational convenience, one can safely assume $\alpha_0 = 0$. As in most work on nonparametric smoothing, estimation of the nonparametric functions $\{\alpha_l(\cdot)\}_{l=1}^d$ is conducted on a compact support. Without loss of generality, let the compact set be $\mathcal{X} = [0, 1]^d$.

Following Stone (1985, p.693), for each $1 \leq l \leq d$, define the space of l-centered square integrable functions on $[0, 1]$ as $H_l^0 = \{\alpha : E\{\alpha(X_l)\} = 0, E\{\alpha^2(X_l)\} < +\infty\}$. Then we model the regression function m as a member of the model space $\mathcal{M}$, a collection of functions on $\mathcal{X}$ defined as $\mathcal{M} = \{m = \sum_{l=1}^d \alpha_l; \alpha_l \in H_l^0\}$. For any $m \in \mathcal{M}$, set $E_n(m) = \sum_{i=1}^n m(\mathbf{X}_i)/n$, and $E(m) = E\{m(\mathbf{X})\}$. For any functions $m_1, m_2 \in \mathcal{M}$, define the empirical and theoretical inner products as $\langle m_1, m_2\rangle_n = E_n(m_1 m_2)$ and $\langle m_1, m_2\rangle = E(m_1 m_2)$, respectively. The induced empirical and theoretical norms are denoted as $\|m\|_n^2 = E_n(m^2)$ and $\|m\|^2 = E(m^2)$, respectively. Then the model space $\mathcal{M}$ is a Hilbert space equipped with the theoretical inner product.

In this paper, we are particularly interested in the variable selection problem for the additive model. Suppose that only an unknown subset of covariates in (2.1) is significant with $\alpha_l(x_l) \neq 0$ a.s.. Let $\mathcal{S}_0 \subset \{1, \ldots, d\}$ be the index set of significant variables, i.e., $\alpha_l(x_l) \neq 0$ a.s., for any $l \in \mathcal{S}_0$ with $\alpha_l(x_l) = 0$ a.s., for any $l \notin \mathcal{S}_0$. Then the goal of variable selection is to correctly identify $\mathcal{S}_0$ based on $\{\mathbf{X}_i, Y_i\}_{i=1}^n$, a random sample of size n from the distribution of $(\mathbf{X}, Y)$.

3. Penalized Polynomial Spline Estimation

To select significant variables, we propose to use a penalized polynomial spline method that involves approximation of the nonparametric functions $\{\alpha_l(\cdot)\}_{l=1}^d$ by polynomial splines.

3.1. The estimators

Polynomial splines are piece-wise polynomials connected smoothly over a set of interior knots. For each $1 \leq l \leq d$, let $k_l = \{0 = x_{l,0} < x_{l,1} < \cdots < x_{l,N_l} < x_{l,N_l+1} = 1\}$ be a knot sequence with N_l interior knots on $[0, 1]$. For an integer $p \geq 1$, the polynomial spline space $\varphi_l = \varphi(p, k_l, [0, 1])$ consists of functions that are polynomials of degree p (or less) on intervals $[x_{l,i}, x_{l,i+1}), i = 0, \ldots, N_l - 1$, and $[x_{l,N_l}, x_{l,N_l+1}]$, and globally has $p - 1$ continuous derivatives. Such polynomial splines can approximate smooth functions with an error of approximation depending on a smoothing parameter $h_l = \max_{i=0,\ldots,N_l} |x_{l,i+1} - x_{l,i}|$, see de Boor (2001). A different amount of smoothing h_l can be used for each $\alpha_l(\cdot)$. Let $h = \max_{1 \leq l \leq d} h_l$.

For each l, let $\varphi_l^{0,n} = \{g_l : g_l \in \varphi_l, \sum_{i=1}^n g_l(X_{il})/n = 0\}$ be the space of empirically centered polynomial splines. Define the approximation space $\mathcal{M}_n = \{m_n(\mathbf{x}) = \sum_{l=1}^d g_l(x_l); g_l \in \varphi_l^{0,n}\}$. Then the standard polynomial spline method (Stone (1985)) estimates m by minimizing the sum of squares over the approximation space $\mathcal{M}_n$, i.e.,

$$\widetilde{m} = \operatorname*{argmin}_{m_n \in \mathcal{M}_n} \frac{1}{2} \|Y - m_n\|_n^2, \tag{3.1}$$

where $Y = Y(\cdot)$ denotes a random function which interpolates the values $Y_1, \ldots,$ Y_n at $\mathbf{X}_1, \ldots, \mathbf{X}_n$. This approach fails to reduce the model complexity when some of the explanatory variables are redundant. Here we propose to fit a penalized least squares that automatically sets small estimated functions to zero, resulting in a parsimonious model. It is defined as

$$\hat{m} = \operatorname*{argmin}_{m_n = \sum_{l=1}^d g_l \in \mathcal{M}_n} \left[\frac{1}{2} \|Y - m_n\|_n^2 + \sum_{l=1}^d p_{\lambda_n}(\|g_l\|_n) \right], \tag{3.2}$$

where $p_{\lambda_n}(\cdot) \geq 0$ is a given penalty function depending on a tuning parameter λ_n. In general, a larger λ_n results in a simpler model with fewer variables selected; the selection of λ_n will be discussed in Section 3.4. In (3.2), the empirical norm $\|\cdot\|_n$ is used instead of the theoretical one, since the definition of the theoretical norm depends on the unknown distribution of covariates $\mathbf{X}$.

The formulation (3.2) is quite general. In particular, if $\mathcal{M}_n = \{m_n(\mathbf{x}) = \sum_{l=1}^d \beta_l x_l\}$ and the covariates are normalized with $\sum_{i=1}^n x_{il}/n = 0$ and $\sum_{i=1}^n x_{il}^2/n = 1$, for $l = 1, \ldots, d$, then (3.2) reduces to a family of variable selection methods for linear models, with the penalty $p_{\lambda_n}(\|g_l\|_n) = p_{\lambda_n}(|\beta_l|)$. For example, the L_1 penalty $p_{\lambda_n}(|\beta|) = \lambda_n|\beta|$ results in the LASSO (Tibshirani (1996)), and the L_2 penalty $p_{\lambda_n}(|\beta|) = \lambda_n|\beta|^2$ results in a ridge regression. Fan and Li (2001) proposed to use the smoothly clipped absolute deviation (SCAD) penalty, whose first order derivative is given as

$$p'_{\lambda_n}(|\beta|) = \lambda_n I(|\beta| \leq \lambda_n) + \frac{(a\lambda_n - |\beta|)_+}{a-1} I(|\beta| > \lambda_n) \tag{3.3}$$

for some $a > 2$. For linear models, Fan and Li (2001) showed that the SCAD improves over other penalty functions and results in a solution with desirable properties, such as unbiasedness, sparsity, and continuity. Thus in this paper we focus on using the SCAD penalty in (3.2), and extend the asymptotic results for the SCAD penalized least squares to the additive models. In the following, we refer to the resulting method as spline SCAD.

3.2. Polynomial spline basis representation

To gain further insight, we adopt the spline basis representation of the proposed polynomial spline estimation in (3.2). For each $l = 1, \ldots, d$, write $J_l = N_l + p$, and let $\mathbf{B}_l = \{B_{l1}, \ldots, B_{lJ_l}\}$ be a basis for $\varphi_l^{0,n}$. For example, the centered truncated power basis is used in implementation, with $\{B_{lj} = b_{lj} - E_n(b_{lj})\}_{j=1}^{J_l}$, where $\mathbf{b}_l = \{b_{l1}, \ldots, b_{lJ_l}\}$ is the truncated power basis given as $\{x_l, \ldots, x_l^p, (x_l - x_{l,1})_+^p, \ldots, (x_l - x_{l,N_l})_+^p\}$, in which $(x)_+^p = (x_+)^p$.

For any fixed $\mathbf{x} = (x_1, \ldots, x_d)^T \in \mathcal{X}$, let $\mathbf{B}_l(x_l) = [B_{l1}(x_l), \ldots, B_{lJ_l}(x_l)]^T$. Then from (3.1) and (3.2), one can express $\widetilde{m}$ and $\hat{m}$ as

$$\widetilde{m}(\mathbf{x}) = \sum_{l=1}^d \widetilde{\alpha}_l(x_l),\ \widetilde{\alpha}_l(x_l) = \widetilde{\beta}_l^T \mathbf{B}_l(x_l), \quad \text{and} \quad \hat{m}(\mathbf{x}) = \sum_{l=1}^d \hat{\alpha}_l(x_l),\ \hat{\alpha}_l(x_l) = \hat{\beta}_l^T \mathbf{B}_l(x_l), \tag{3.4}$$

where the coefficients $\widetilde{\beta} = (\widetilde{\beta}_1^T \ldots, \widetilde{\beta}_d^T)^T$ and $\hat{\beta} = (\hat{\beta}_1^T \ldots, \hat{\beta}_d^T)^T$ minimize the sum of squares and the penalized sum of squares, respectively:

$$\widetilde{\beta} = \underset{\beta = (\beta_1^T \ldots, \beta_d^T)^T}{\operatorname{argmin}} \frac{1}{2} \left\| Y - \sum_{l=1}^d \beta_l^T \mathbf{B}_l \right\|_n^2, \tag{3.5}$$

$$\hat{\beta} = \operatorname*{argmin}_{\beta=(\beta_1^T \ldots, \beta_d^T)^T} \left[\frac{1}{2} \left\| Y - \sum_{l=1}^{d} \beta_l^T \mathbf{B}_l \right\|_n^2 + \sum_{l=1}^{d} p_{\lambda_n}(\|\beta_l\|_{K_l}) \right], \tag{3.6}$$

where $\|\beta_l\|_{K_l} = \sqrt{\beta_l^T \mathbf{K}_l \beta_l}$ with $\mathbf{K}_l = \sum_{i=1}^n \mathbf{B}_l(X_{il})\mathbf{B}_l^T(X_{il})/n$. Note that the solutions given in (3.4) do not depend on the particular choice of spline basis $\mathbf{B}_l$, $l = 1, \ldots, d$.

- **Remark 1**: In solving the two minimization problems, (3.6) is relatively more difficult to compute since $p_{\lambda_n}(\cdot)$ is singular at the origin and does not have second derivative at some points. But when $d = 1$ and $m(x) = \alpha(x)$,

  $$\hat{\alpha} = \{I(a\lambda_n < \|\widetilde{\alpha}\|_n) + c_1 I(2\lambda_n < \|\widetilde{\alpha}\|_n \le a\lambda_n) + c_2 I(\|\widetilde{\alpha}\|_n \le 2\lambda_n)\}\widetilde{\alpha},$$

  where $c_1 = [(a-1)\|\widetilde{\alpha}\|_n - a\lambda_n]/[(a-2)\|\widetilde{\alpha}\|_n]$, $c_2 = [(\|\widetilde{\alpha}\|_n - \lambda_n)_+]/[\|\widetilde{\alpha}\|_n]$, and $\widetilde{\alpha}$ is the standard polynomial spline estimator. This clearly shows that $\hat{\alpha}$ is a threshold rule in $\widetilde{\alpha}$. In particular, it yields zero when $\widetilde{\alpha}$ is small enough ($\|\widetilde{\alpha}\|_n \le \lambda_n$), and leaves $\widetilde{\alpha}$ unchanged when $\widetilde{\alpha}$ is large ($\|\widetilde{\alpha}\|_n \ge a\lambda_n$). In Subsection 3.3, we will discuss how to solve (3.6) under a more general framework.
- **Remark 2**: The basis representation (3.6) reveals that the spline SCAD shares a similar penalized form as the P-spline (Ruppert, Wand and Carroll (2003)). They are different in nature at least in two respects: (a) the P-spline uses a quadratic penalty on the coefficients $\{\beta_{lk}\}_{l=1,k=1}^{d,N_n}$, essentially a type of ridge regression, and is unable to produce sparse solutions; the spline SCAD uses a SCAD penalty on the function norms and is able to shrink all coefficients associated with a variable to zero simultaneously and give sparse solutions. (b) The P-spline assumes a fixed knot number, while the spline SCAD allows the knot number to increase with the sample size, which makes rigorous asymptotics possible.
- **Remark 3**: Yuan and Lin (2006, 2007) successfully extended the non-negative garrotte, LASSO, and LARS to select significant factors (groups of variables) instead of individual variables. The method proposed in this paper applies the SCAD penalty to group variables. Using the SCAD penalty, we are able to show that the proposed method enjoys oracle properties when the number of variables in a group increases with sample size. It is not clear whether the group variable selection using the non-negative garrotte, LASSO, or LARS have such an oracle property.

3.3. Local quadratic approximation algorithm

Since the SCAD penalty is singular at the origin, it is challenging to find the minimizer $\hat{\beta}$ of (3.6). However, following Fan and Li (2001), the penalty function can be locally approximated by a quadratic function, and the Newton-Raphson algorithm then can be applied to minimize the penalized sum of squares.

Let $\beta^0 = (\beta_1^{0T}, \ldots, \beta_d^{0T})^T$ be an initial value that is close to $\hat{\beta}$, and $\beta^k = (\beta_1^{kT}, \ldots, \beta_d^{kT})^T$ be the value at the kth iteration. In the implementation, we have used $\beta^0 = \widetilde{\beta}$. For $l = 1, \ldots, d$, if β_l^0 is very close to zero, i.e., $\|\beta_l^0\|_{K_l} < \epsilon$ for some small threshold value ϵ, then set $\beta_l^1 = \mathbf{0}$. In the implementation, we have used $\epsilon = 10^{-6}$. Without loss of generality, we write $\beta^1 = (\beta_1^{1T}, \ldots, \beta_{d_1}^{1T}, \beta_{d_1+1}^{1T}, \ldots \beta_d^{1T})^T = (\beta_{11}^{1T}, \beta_{22}^{1T})^T$ with the last $d - d_1$ components being zero, $\beta_{22}^1 = \mathbf{0}$. To update the first d_1 non-zero components, local quadratic approximation is used. Note that $[\partial p_{\lambda_n}(\|\beta_l\|_{K_l})]/[\partial \beta_l] = [p'_{\lambda_n}(\|\beta_l\|_{K_l})\|\beta_l\|_{K_l}^{-1}\mathbf{K}_l\beta_l]/2$, for $\beta_l \neq \mathbf{0}$. Therefore one has

$$\begin{aligned} p_{\lambda_n}\Big(\|\beta_l\|_{K_l}\Big) &\approx p_{\lambda_n}\Big(\|\beta_l^0\|_{K_l}\Big) + \frac{1}{2}p'_{\lambda_n}\Big(\|\beta_l^0\|_{K_l}\Big)\|\beta_l^0\|_{K_l}^{-1}\Big[\beta_l^{0T}\mathbf{K}_l(\beta_l - \beta_l^0)\Big] \\ &\approx p_{\lambda_n}\Big(\|\beta_l^0\|_{K_l}\Big) + \frac{1}{2}p'_{\lambda_n}\Big(\|\beta_l^0\|_{K_l}\Big)\|\beta_l^0\|_{K_l}^{-1}\Big(\beta_l^T\mathbf{K}_l\beta_l - \beta_l^{0T}\mathbf{K}_l\beta_l^0\Big) \end{aligned}$$

for $\beta_l^0 \approx \beta_l$. As a result, (3.6) can be locally approximated, up to a constant, by

$$\frac{1}{n}\sum_{i=1}^{n}\Big(Y_i - \sum_{l=1}^{d_1}\beta_l^T\mathbf{B}_l(X_{il})\Big)^2 + \sum_{l=1}^{d_1}\varphi(\beta_l^0)\beta_l^T\mathbf{K}_l\beta_l, \tag{3.7}$$

with $\varphi(\beta_l^0) = p'_{\lambda_n}(\|\beta_l^0\|_{K_l})\|\beta_l^0\|_{K_l}^{-1}$. Let $\overrightarrow{\mathbf{B}}_{11} = (\overrightarrow{\mathbf{B}}_1, \ldots, \overrightarrow{\mathbf{B}}_{d_1})$, with $\overrightarrow{\mathbf{B}}_l = \{\mathbf{B}_l(X_{1l}), \ldots, \mathbf{B}_l(X_{nl})\}^T$ and $\Sigma_{\lambda_n} = \mathrm{diag}\{\varphi(\beta_1^0)\mathbf{K}_1, \ldots, \varphi(\beta_{d_1}^0)\mathbf{K}_{d_1}\}$. Then the minimizer of (3.7) is given as $\beta_{11}^1 = \{\overrightarrow{\mathbf{B}}_{11}^T\overrightarrow{\mathbf{B}}_{11} + n\Sigma_{\lambda_n}\}^{-1}\overrightarrow{\mathbf{B}}_{11}^T\mathbf{Y}$. One repeats this procedure to convergence. In the implementation, we have used the convergence criterion that $\sqrt{(\beta^k - \beta^{k+1})^T(\beta^k - \beta^{k+1})} \leq 10^{-6}$.

Following Hunter and Li (2005), we show that our modified local quadratic approximation (LQA) algorithm is also an instance of a majorization-minimization (MM) algorithm. Then the convergence of the proposed algorithm can be studied using techniques applicable to MM algorithms in general. For $l = 1, \ldots, d$ and any vectors α, α^0 with length $N_l + p$, set

$$\Phi_{\alpha^0}^l(\alpha) = p_{\lambda_n}\Big(\|\alpha^0\|_{K_l}\Big) + \frac{1}{2}p'_{\lambda_n}\Big(\|\alpha_l^0\|_{K_l}\Big)\|\alpha_l^0\|_{K_l}^{-1}\Big(\|\alpha_l\|_{K_l}^2 - \|\alpha_l^0\|_{K_l}^2\Big).$$

Denote the penalized sum of squares by $Q(\beta) = [\|Y - \sum_{l=1}^d \beta_l^T\mathbf{B}_l\|_n^2]/2 + \sum_{l=1}^d p_{\lambda_n}(\|\beta_l\|_{K_l})$.

Theorem 1. *Suppose $\beta^k = (\beta_1^{kT}, \ldots, \beta_d^{kT})^T$, with $\beta_l^k \neq \mathbf{0}$ for all l at the kth iteration. Let $S_k(\beta) = [\|Y - \sum_{l=1}^d \beta_l^T\mathbf{B}_l\|_n^2]/2 + \sum_{l=1}^d \Phi_{\beta_l^k}^l(\beta_l)$. Then $S_k(\beta) \geq Q(\beta)$ for all β, with equality when $\beta = \beta^k$.*

Suppose β^{k+1} is the minimizer of $S_k(\beta)$. From Theorem 1, one has $S_k(\beta^k) \geq S_k(\beta^{k+1})$ implies $Q(\beta^k) \geq Q(\beta^{k+1})$. That is, a decrease in the value of a quadratic function $S_k(\beta)$ guarantees a decrease in the value of $Q(\beta)$. Thus the minimization of the penalized sum of squares $Q(\beta)$ can be replaced by minimizing a quadratic function $S_k(\beta)$ that has a closed-form solution.

3.4. Smoothing and tuning parameters

To implement the proposed spline SCAD, one needs to choose appropriate spline spaces $\{\varphi_l^n\}_{l=1}^d$ and tuning parameters a, λ_n involved in the SCAD penalty. For the choices of $\{\varphi_l^n\}_{l=1}^d$ we use splines with equally spaced knots and fixed degrees, and select only $\{N_l\}_{l=1}^d$, the number of interior knots using the data. The same strategy is also used in Huang, Wu and Zhou (2004). Following Fan and Li (2001), we take $a = 3.7$, which works well from our simulation experiences.

Denote the parameters to be selected by $\theta = (N_1, \ldots, N_d, \lambda_n)$. For fast computation, we use K-fold cross-validation to select θ, with $K = 5$ in implementation. The full data T is randomly partitioned into K groups of about the same size, denoted as T^j, for $j = 1, \ldots, K$. Then for each j, the data $T - T^j$ is used for estimation and T^j for validation. For any given θ, denote the resulting estimator of m by $\hat{m}_\theta^j$. Then the cross-validation criterion is given as $\mathrm{CV}(\theta) = \sum_{j=1}^K \sum_{i \in d_j} \{Y_i - \hat{m}_\theta^j(\mathbf{X}_i)\}^2$, where d_j denotes the indices of data points included in the j-th group. We select $\hat{\theta}$ to minimize $\mathrm{CV}(\theta)$. In practice, the minimization problem over a $d+1$ dimension can be computationally difficult. However, for smooth functions, the results are quite stable when a sufficient number of knots is used. Thus, for fast computation, we suggest taking the same number of interior knots over different directions, i.e., $N_1 = \cdots = N_d = N$. Then the minimization is a two-dimensional problem, and the tuning parameters can be estimated by a two-dimensional grid search. To be more specific, according to Theorem 2, the optimal order of N_l is $\tilde{N} = n^{1/(2p+3)}$ for each l. Thus, in the implementation, we select the optimal N from the integers in $[0.5\tilde{N}, 2\tilde{N}]$. Furthermore, to satisfy the order assumptions of λ_n in Theorem 3, we select λ_n from the interval $[0.5\log(n)\sqrt{\tilde{N}/n}, 2\log(n)\sqrt{\tilde{N}/n}]$.

4. Asymptotic Properties

In this section, we establish the asymptotic properties of the spline SCAD estimator $\hat{m}$ in (3.2). Throughout this section, the penalty function $p_{\lambda_n}(\cdot)$ is the SCAD penalty defined in (3.3). Write the true regression function as $m_0(\mathbf{x}) = \sum_{l=1}^d \alpha_{l0}(x_l) = \sum_{l=1}^s \alpha_{l0}(x_l) + \sum_{l=s+1}^d \alpha_{l0}(x_l)$ where, without loss of generality, $\alpha_{l0} = 0$ a.s. for $l = s+1, \ldots, d$, and s is the total number of nonzero function components. For any $m = \sum_{l=1}^d \alpha_l \in \mathcal{M}$, define $l(m) =$

$(\|Y - m\|_n^2)/2 + \sum_{l=1}^d p_{\lambda_n}(\|\alpha_l\|_n)$. Recall that $\hat{m}_n = \hat{m} = \text{argmin}_{g \in \mathcal{M}_n} l(g)$, where the notation $\hat{m}_n$ is used to emphasize its dependence on n.

We first show that there exists a local minimizer of $l(\cdot)$ over $\mathcal{M}_n$ whose mean square (or L_2) convergence rate is $O_p(\rho_n)$, where $\rho_n = 1/\sqrt{nh} + h^{p+1}$, the L_2 convergence rate for standard polynomial spline.

Theorem 2. *Under assumptions* (A1–A4) *in the on-line supplement, if* $\lambda_n \to 0$, *then there exists a local minimizer* $\hat{m}_n$ *of* $l(\cdot)$ *in* $\mathcal{M}_n$*, for n sufficiently large, that satisfies* $\|\hat{m}_n - m_0\| = O_p(\rho_n)$.

Theorem 2 shows that $\hat{m}_n$ enjoys the same rate of convergence as the standard polynomial spline estimators. Furthermore, write $\hat{m}_n = \sum_{l=1}^d \hat{\alpha}_l$, with $\hat{\alpha}_l \in \varphi_l^{0,n}$ (Lemma A.2 in the on-line material implies such an additive representation is essentially unique). Then $\|\hat{\alpha}_l - \alpha_{l0}\| = O_p(\rho_n)$ for each $l = 1, \ldots, d$. Now we prove that the minimizer $\hat{m}_n = \sum_{l=1}^d \hat{\alpha}_l$ in Theorem 2 possess the sparsity property, i.e., $\hat{\alpha}_l = 0$ a.s. for $l = s+1, \ldots, d$, which makes it a consistent variable selection method.

Theorem 3. *Under assumptions* (A1–A4) *in the on-line supplement, if* $\lambda_n \to 0$ *and* $\rho_n/\lambda_n arrow 0$, *then with probability approaching* 1, $\hat{\alpha}_l = 0$ *a.s. for* $l = s+1, \ldots, d$.

5. Examples

In this section, we assess the finite sample performance of the proposed spline SCAD with simulations, and illustrate its use with the analysis of the Ozone concentration data and Boston housing price data. The spline SCAD is compared with the component selection and smoothing operator (COSSO) (Lin and Zhang (2006)), multivariate additive regression splines (MARS) (Friedman (1991)), and polynomial spline BIC and AIC (Huang and Yang (2004)).

In the simulated examples, the averaged integrated squared error (AISE) is used to assess estimation accuracies. Denoting the estimator of m in the r-th ($1 \leq r \leq R$) replication as $\hat{m}_r$, and $\{\mathbf{x}_m\}_{m=1}^{n_{\text{grid}}}$ the grid points, we define $\text{AISE}(\hat{m}_r) = [\sum_{r=1}^R (1/n_{\text{grid}}) \sum_{m=1}^{n_{\text{grid}}} \{m(\mathbf{x}_m) - \hat{m}_r(\mathbf{x}_m)\}^2]/R$. Let $\mathcal{S}$ and $\mathcal{S}_0$ be the selected and true index set of significant variables respectively. Following Huang and Yang (2004), we say $\mathcal{S}$ is correct if $\mathcal{S} = \mathcal{S}_0$; $\mathcal{S}$ overfits if $\mathcal{S}_0 \subset \mathcal{S}$ and $\mathcal{S}_0 \neq \mathcal{S}$; and $\mathcal{S}$ underfits if $\mathcal{S}_0 \not\subset \mathcal{S}$.

5.1. Simulated example

In this example, we simulated 100 data sets consisting of $n = 100$, or 250 observations from the model $Y = m(\mathbf{X}) + \delta\varepsilon$, where the dimension of $\mathbf{X}$ is 10 but $m(\mathbf{x}) = 5g_1(x_1) + 3g_2(x_2) + 4g_3(x_3) + 6g_4(x_4)$, with $g_1(x) = x$, $g_2(x) = (2x -$

$1)^2$, $g_3(x) = \sin(2\pi x)/(2 - \sin(2\pi x))$, and $g_4(x) = 0.1\sin(2\pi x) + 0.2\cos(2\pi x) + 0.3\sin^2(2\pi x) + 0.4\cos^3(2\pi x) + 0.5\sin^3(2\pi x)$. The covariates $\mathbf{X}$ were generated to have a compound symmetry covariance structure: $X_l = (W_l + tU)/(1 + t)$, $l = 1, \ldots, 10$, where $(W_1, \ldots, W_{10})$ and U are i.i.d. from Uniform$[0, 1]$, and t was taken to be 0 or 2. The error ε was standard normal and the noise level σ was 1.319, which gives a signal to noise ratio $3 : 1$ when $t = 0$. The same model was also considered in Lin and Zhang (2006).

We applied spline SCAD with both linear splines (SCAD1) and cubic splines (SCAD3) to the simulated data. A random sample was simulated from the model with $t = 0$ and $n = 250$. Figure 1 plots the empirical norms of the estimated components $\|\hat{\alpha}_l(\lambda_n)\|_n$ against the tuning parameter λ_n, along with the location of $\hat{\lambda}_n$ selected from the 5-fold CV method. We fixed $a = 3.7$, and $N = 3$ for SCAD1, $N = 2$ for SCAD3. Figure 1 clearly showed the selected λ_n worked well and gave the correct model with four terms for both SCAD1 and SCAD3 in this run. Figure 2 plots the estimated component functions for the first four terms.

We then compared spline SCAD with COSSO, MARS, and spline AIC and BIC. We also considered the standard linear spline (LS) estimations of the full and oracle models. The full model is the one of all possible covariates, while the oracle model contains only non-zero components. For COSSO, the matlab code was downloaded from 'www.stat.ncsu.edu/ hzhang2'. The tuning parameters were selected using 5-fold CV with M chosen from 100 equally spaced grid points in $[0, 35]$, and λ_0 was chosen from grid points $\{2^{-j}\}_{j=1}^{100}$. For spline AIC and BIC, we used code obtained from the author, and MARS simulations were done in R, with the function "mars" in the "mda" library. For spline SCAD and LS, we always used the same N interior knots equally spaced for each function component, where N was selected from the 5-fold CV method described in Subsection 3.4.

Table 1 summarizes variable selection and estimation results from various methods. It shows that COSSO, two spline SCAD fits and spline BIC were generally comparable. Furthermore, two spline SCAD fits and spline BIC performed slightly better than COSSO when sample size was large ($n = 250$) and the covariates were uncorrelated ($t = 0$), while COSSO performed the best when the covarites were correlated ($t = 1$). Overall, spline AIC and MARS tend to overfit. Table 1 also shows SCAD1 worked as well as the linear spline estimation of the oracle model, and much better than the linear spline estimation of the full model in terms of estimation accuracy (AISEs). The simulation results support Theorems 1−2.

In this simulation study, spline SCAD, COSSO and spline BIC performed comparably. The differences among the three methods mirror the differences among SCAD, LASSO, and BIC in linear models. Similar to SCAD in linear

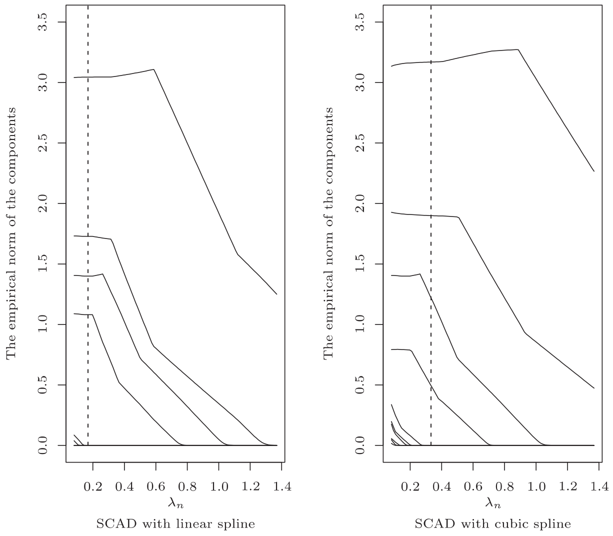

Figure 1. The empirical norm of each estimated component function is plotted against the tuning parameter λ_n in one run with $t = 0$ and $n = 250$. The dashed line indicates the location of the selected optimal λ_n from the CV method.

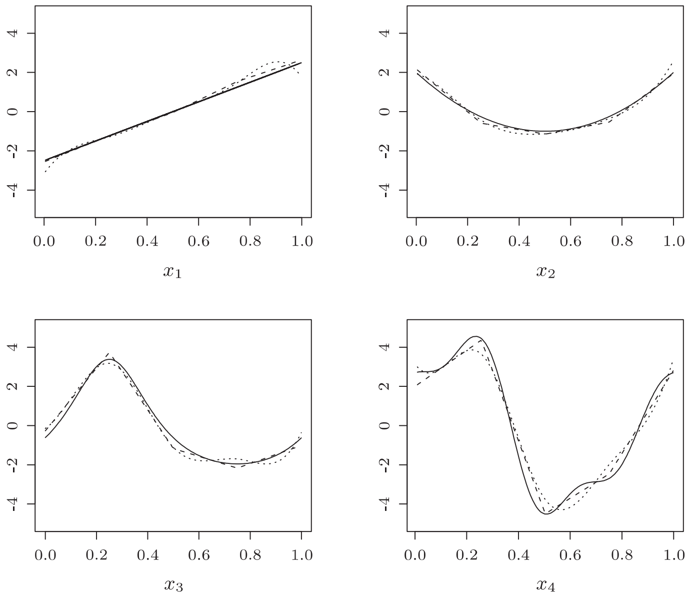

Figure 2. The estimated component functions using penalized linear spline (dashed line), using penalized cubic spline (dotted line), and the true component functions (solid line) in one run with $t = 0$ and $n = 250$. Shown are the components for the first four variables. For the other variables, both estimated and true component functions are zero.

Table 1. Simulated example, variable selection results. The columns of C, U, and O give, respectively, the numbers of correct-fitting, under-fitting, and over-fitting over 100 replications, and the columns of *Size* give the average number of selected variables.

n	Method	$t = 0$					$t = 2$				
		C	U	O	Size	AISE	C	U	O	Size	AISE
100	SCAD1	85	4	11	4.07	0.683	21	76	3	2.90	1.042
	SCAD3	81	8	11	4.01	0.709	13	84	3	2.83	1.494
	COSSO	86	2	12	4.12	0.754	37	38	25	4.2	0.898
	BIC	92	2	6	4.04	0.591	22	76	2	2.93	1.037
	AIC	33	1	66	5.39	1.009	24	26	50	5.37	1.433
	MARS	10	0	90	5.10	1.730	20	40	45	4.70	1.246
	ORACLE	100	0	0	4.00	0.539	100	0	0	4.00	0.824
	FULL	0	0	100	10.00	1.453	0	0	100	10	2.127
250	SCAD1	100	0	0	4.00	0.2897	82	18	0	3.80	0.4044
	SCAD3	98	0	2	4.02	0.2712	75	23	2	3.71	0.4614
	COSSO	91	0	9	4.17	0.2905	83	3	14	4.13	0.3714
	BIC	100	0	0	4.00	0.2832	83	17	0	3.81	0.4020
	AIC	41	0	59	4.85	0.3572	40	11	49	4.6	0.4533
	MARS	40	0	60	4.80	0.7374	38	2	60	4.91	0.5083
	ORACLE	100	0	0	4.00	0.2832	100	0	0	4.00	0.3594
	FULL	0	0	100	10.00	0.5183	0	0	100	10	0.6345

models, we showed spline SCAD also have oracle properties. However, it is conceivable from LASSO that COSSO may not have the oracle property.

5.2. Applications to ozone data and Boston housing price data

We applied the spline SCAD to two data examples: the ozone data and the Boston housing price data, available from the R library 'mlbench'. The two data sets were previously analyzed by several authors, e.g., Breiman and Friedman (1985), Buja, Hastie and Tibshirani (1989) and Breiman (1995). The ozone data consists of daily measurements of ozone concentration, and eight meteorological quantities for 330 days in 1976 in the LOS Angeles basin. To account for the seasonal effect, we also included one additional variable called date-of-the-year, as in Breiman and Friedman (1985). The Boston housing price concerns the median value of owner-occupied homes in each of the 506 census tracts in the Boston Standard Metropolitan Statistical Area in 1970. It contains 12 sociodemographic variables that are thought to affect housing price.

In our analysis, we standardized all variables (both explanatory and response variables), so that each had a zero mean and unit sample standard deviation. The methods discussed in simulation studies were applied and compared by their estimation and prediction performances. A total of m observations were randomly left out for prediction, and the remaining observations were used for modeling

Table 2. Data examples: estimation, prediction and variable selection results of several methods.

	Ozone			Housing		
Method	avg. ASEE	avg. ASPE	avg. size	avg. ASEE	avg. ASPE	avg. size
SCAD1	0.1733	0.2136	7.5	0.1413	0.1591	7.1
SCAD3	0.1792	0.2330	6.7	0.1565	0.1478	7.9
COSSO	0.1909	0.2198	7.2	0.1181	0.1750	9.4
BIC	0.1859	0.2166	6.5	0.1710	0.1623	4.3
AIC	0.1696	0.2492	8	0.1354	0.1675	8.5
MARS	0.1752	0.2542	8.2	0.1414	0.1902	9.3
FULL	0.1680	0.2675	9	0.1407	0.1969	12

fitting, where $m = 30$ and 50 for the ozone and Boston housing data, respectively. Then the averaged squared estimation errors (ASEE) and averaged squared prediction errors (ASPE) were calculated for each method. This procedure was replicated 20 times for each data. Table 2 reports the averaged ASEE, averaged ASPE and averaged number of selected variables from 20 replications. For both data sets, all variable selection methods resulted in models with smaller size and better predictions than the full model, which indicates that redundant variables exist in both data sets. Among all variable selection methods, the spline SCADs (SCAD1 and SCAD3) gave parsimonious models with the best prediction performance.

Acknowledgements

We thank an associate editor and the reviewers for their constructive comments that have greatly helped to improve the article.

Appendix

The necessary assumptions and lemmas for the following proofs are given in the on-line supplement material available at `http://www.stat.sinica.edu.tw/statistica`.

Proof of Theorem 1. The proof of the Theorem 1 is immediate. Note that the definition of $\Phi^l_{\alpha^0}(\alpha)$ depends on α, and α^0 only through $\|\alpha\|_{K_l}$ and $\|\alpha^0\|_{K_l}$. Thus it is equivalent to show that $\Phi_{\theta_0}(\theta) \geq p_{\lambda_n}(\theta)$ for any θ, $\theta_0 > 0$, where we define $\Phi_{\theta_0}(\theta) = p_{\lambda_n}(\theta_0) + [p^{'}_{\lambda_n}(\theta_0)(\theta^2 - \theta_0^2)]/(2\theta_0)$. This is proved in Proposition 3.1 of Hunter and Li (2005).

Proof of Theorem 2.

Let $\mathcal{M}_{n,0} = \{m_n(\mathbf{x}) = \sum_{l=1}^{s} g_l(x_l); g_l \in \varphi_l^{0,n}\}$, the approximation space knowing $\alpha_{l0} = 0$, for $s+1 \leq l \leq d$. Let $\hat{m}^*_{n,0} = \text{argmin}_{m_{n,0} \in \mathcal{M}_{n,0}} \|Y - m_{n,0}\|_n^2$ and $\hat{m}^*_n = \text{argmin}_{m_n \in \mathcal{M}_n} \|Y - m_n\|_n^2$, the best least square approximation of m_0

in approximation spaces $\mathcal{M}_{n,0}$, and $\mathcal{M}_n$, respectively. Write $\hat{m}^*_{n,0} = \sum_{l=1}^{s} \hat{\alpha}^*_{l0}$. Then for any $g = \sum_{l=1}^{d} g_l \in \mathcal{M}_n$, with $\|g - m_0\| = C\rho_n$, using $p_{\lambda_n}(0) = 0$ and $p_{\lambda_n}(\cdot) \geq 0$ one has

$$\begin{aligned} & l(g) - l(\hat{m}^*_{n,0}) \\ & \quad \geq \frac{(\|Y - g\|_n^2 - \|Y - \hat{m}^*_{n,0}\|_n^2)}{2} + \sum_{l=1}^{s} \Big\{ p_{\lambda_n}(\|g_l\|_n) - p_{\lambda_n}(\|\hat{\alpha}^*_{l0}\|_n) \Big\} \\ & \quad = \frac{(\|\hat{m}^*_n - g\|_n^2 - \|\hat{m}^*_n - \hat{m}^*_{n,0}\|_n^2)}{2} + \sum_{l=1}^{s} \Big\{ p_{\lambda_n}(\|g_l\|_n) - p_{\lambda_n}(\|\hat{\alpha}^*_{l0}\|_n) \}. \qquad \text{(A.1)} \end{aligned}$$

Let $A_n = \sup_{\phi \in \mathcal{M}_n} |\|\phi\|_n^2 / \|\phi\|^2 - 1|$. For the first term in (A.1),

$$\begin{aligned} & \frac{1}{2}\Big(\|\hat{m}^*_n - g\|_n^2 - \|\hat{m}^*_n - \hat{m}^*_{n,0}\|_n^2 \Big) \\ & \qquad \geq \frac{1}{2}\|\hat{m}^*_n - g\|^2 (1 - A_n) - \frac{1}{2}\|\hat{m}^*_n - \hat{m}^*_{n,0}\|^2 (1 + A_n) = I + II, \end{aligned}$$

in which, $I = (\|\hat{m}^*_n - g\|^2 - \|\hat{m}^*_n - \hat{m}^*_{n,0}\|^2)/2$ and $II = -[A_n(\|\hat{m}^*_n - g\|^2 + \|\hat{m}^*_n - \hat{m}^*_{n,0}\|^2)]/2$. By the triangle inequality,

$$\begin{aligned} 2I & \geq \|g - m_0\|^2 - 2\|g - m_0\|\|\hat{m}^*_n - m_0\| - \|\hat{m}^*_{n,0} - m_0\|^2 - 2\|\hat{m}^*_n - m_0\|\|\hat{m}^*_{n,0} - m_0\| \\ & = C^2\rho_n^2 - 2C\rho_n\|\hat{m}^*_n - m_0\| - \|\hat{m}^*_{n,0} - m_0\|^2 - 2\|\hat{m}^*_n - m_0\|\|\hat{m}^*_{n,0} - m_0\|. \qquad \text{(A.2)} \end{aligned}$$

By Lemma A.1, the last three terms in (A.2) are each of order $O_p(\rho_n^2)$. Thus by choosing a sufficient large C, the first term dominates in (A.2), uniformly for $g \in \mathcal{M}_n$ with $\|g - m_0\| = C\rho_n$. Furthermore, Lemma A.2 entails that $A_n = O_p(\sqrt{\log^2(n)/(nh)}) = O_p(1)$. Thus term II is also dominated by term I. Now for the second term in (A.1), the triangle inequality gives, for each l, $\|\alpha_{l0}\| + \|\alpha_{l0} - g_l\| \geq \|g_l\| \geq \|\alpha_{l0}\| - \|\alpha_{l0} - g_l\|$, where $\|\alpha_{l0} - g_l\| \leq \|m_0 - g\|/c_4 = C\rho_n/c_4$ from Lemma A.3. Therefore, $\|\alpha_{l0}\| + C\rho_n/c_4 \geq \|g_l\| \geq \|\alpha_{l0}\| - C\rho_n/c_4$. Furthermore Lemma A.2 gives $\|g_l\|_n \geq \|g_l\|(1 - A_n)$. Noting that $\rho_n, A_n, \lambda_n \to 0$, as $n \to \infty$, for n large enough, one has $\|g_l\| \geq a\lambda_n$, and $\|g_l\|_n \geq a\lambda_n$, for each $l = 1, \ldots, s$. Therefore, by the definition of $p_{\lambda_n}(\cdot)$, one has $p_{\lambda_n}(\|g_l\|_n) = p_{\lambda_n}(\|g_l\|) = (a+1)\lambda_n^2/2$. Similar arguments also give, for each $l = 1, \ldots, s$, $p_{\lambda_n}(\|\hat{\alpha}^*_{l0}\|_n) = p_{\lambda_n}(\|\hat{\alpha}^*_{l0}\|) = (a+1)\lambda_n^2/2$. Thus $\sum_{l=1}^{s}\{p_{\lambda_n}(\|g_l\|_n) - p_{\lambda_n}(\|\hat{\alpha}^*_{l0}\|_n)\} = 0$.

Therefore when n is sufficiently large, for any $\varepsilon > 0$ there exists a sufficiently large C such that, $P\{\inf_{g \in \mathcal{M}_n \|g - m_0\| = C\rho_n} l(g) > l(\hat{m}^*_{n,0})\} \leq 1 - \varepsilon$. Hence there exits a local maximizer $\hat{m}_n \in \mathcal{M}_n$ such that $\|\hat{m}_n - m_0\| = O_p(\rho_n)$.

Proof of Theorem 3.

Define $\mathcal{M}_{n,1}$, a subspace of $\mathcal{M}_n$, as $\mathcal{M}_{n,1} = \{m_n(\mathbf{x}) = \sum_{l=s+1}^{d} g_l(x_l);\ g_l \in \varphi_l^{0,n}\}$. It is sufficient to show that, for any $g_{n,0} = \sum_{l=1}^{s} g_l \in \mathcal{M}_{n,0}$ with $\|g_{n,0} - m_0\| = O_p(\rho_n)$, and any constant C, one has $\Lambda(g_{n,0}) = \min_{g_{n,1}\in\mathcal{M}_{n,1},\|g_{n,1}\|\le C\rho_n} \Lambda(g_{n,0} + g_{n,1})$. Consequently, one needs to show that as $narrow\infty$, uniformly for all $g_{n,0} = \sum_{l=1}^{s} g_l \in \mathcal{M}_{n,0}$ with $\|g_{n,0} - m_0\| = O_p(\rho_n)$, and for some small $\varepsilon_n = C\rho_n$ and $g_l \in \varphi_{0,n}^{l}$, $l = s+1, \dots, d$, with probability tending to 1, $\Lambda(g_{n,0}) \le \Lambda(g_{n,0} + g_l)$, $0 < \|g_l\| < \varepsilon_n$. Note that

$$\begin{aligned}
&\Lambda(g_{n,0}) - \Lambda(g_{n,0} + g_l)\\
&\quad = \frac{1}{2}\Big(\|\hat{m}_n^* - g_{n,0}\|_n^2 - \|\hat{m}_n^* - g_{n,0} - g_l\|_n^2\Big) - p_{\lambda_n}(\|g_l\|_n)\\
&\quad = \left\{\frac{1}{2}\Big(\|\hat{m}_n^* - g_{n,0}\|^2 - \|\hat{m}_n^* - g_{n,0} - g_l\|^2\Big) - p_{\lambda_n}(\|g_l\|)\right\}\{1 + o_p(1)\}\\
&\quad \le \left\{\frac{1}{2}\|g_l\|\Big(\|\hat{m}_n^* - g_{n,0}\| + \|\hat{m}_n^* - g_{n,0} - g_l\|\Big) - p_{\lambda_n}(\|g_l\|)\right\}\{1 + o_p(1)\}\\
&\quad = \left\{\lambda_n\|g_l\|\Big(\frac{R_n}{\lambda_n} - \frac{p'_{\lambda_n}(w)}{\lambda_n}\Big)\right\}\{1 + o_p(1)\},
\end{aligned}$$

in which w is a value between 0 and $\|g_l\|$, and $R_n \equiv (\|\hat{m}_n^* - g_{n,0}\| + \|\hat{m}_n^* - g_{n,0} - g_l\|)/2 = O_p(\rho_n)$, by arguments similar to those in the proof of Theorem 2. This completes the proof, observing that $\liminf_{n\to\infty} \liminf_{warrow0+} p'_{\lambda_n}(w)/\lambda_n = 1 > 0$, and $R_n/\lambda_n = o_p(1)$.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *In 2nd International Symposium on Information Theory.* (Edited by B. N. Petrov and F. Csaki), 267-281.

Buja, A., Hastie, T. and Tibshirani, R. (1989). Linear smoothers and additive models. *Ann. Statist.* **17**, 453-555.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics* **37**, 373-384.

Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Assoc.* **80**, 580-619.

Chen, R. and Tsay, R. S. (1993). Nonlinear additive ARX models. *J. Amer. Statist. Assoc.* **88**, 955-967.

de Boor, C. (2001). *A Practical Guide to Splines.* Springer, New York.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.* **32**, 407-499.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348-1360.

Frank. I. E. and Friedman. J. H. (1993). A statistical view of some chemometrics tools. *Technometrics* **35**. 109-135.

Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion). *Ann. Statist.* **19**, 1-141.

Hastie, T. J. (1989). Discussion on "Flexible parsimonious smoothing and additive modeling" by J. Friedman and B. Silverman. *Technometrics* **31**, 23-29.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models.* Chapman and Hall, London.

Huang, J. Z. (1998). Projection estimation in multiple regression with application to functional ANOVA models. *Ann. Statist.* **26**, 242-272.

Huang, J. Z. and Yang, L. (2004). Identification of non-linear additive autoregressive models. *J. Roy. Statist. Soc. Ser. B* **66**, 463-477.

Huang, J. Z., Wu, C. O. and Zhou, L. (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statist. Sinica* **14**, 763-788.

Hunter, D. R. and Li, R. (2005). Variable selection using MM algorithms. *Ann. Statist.* **33**, 1617-1642.

Lin, Y. and Zhang, H. H. (2006). Component selection and smoothing in smoothing spline analysis of variance models. *Ann. Statist.* **34**, 2272-2297.

Linton, O. B. and Härdle, W. (1996). Estimation of additive regression models with known links. *Biometrika* **83**, 529-540.

Linton, O. B. and Nielsen, J. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika* **82**, 93-100.

Mallows, C. L. (1973). Some Comments on C_P. *Technometrics* **15**, 661-675.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric Regression.* Cambridge University Press, Cambridge.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-464.

Shively, T. S., Kohn, R. and Wood, S. (1999). Variable selection and function estimation in additive nonparametric regression using a data-based prior. *J. Amer. Statist. Assoc.* **94**, 777-806.

Sperlich, S., Tjøstheim, D. and Yang, L. (2002). Nonparametric estimation and testing of interaction in additive models. *Econom. Theory* **18**, 197-251.

Stone, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13**, 689-705.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.

Xue, L. and Yang, L. (2006a). Additive coefficient modeling via polynomial spline. *Statist. Sinica* **16**, 1423-1446.

Xue, L. and Yang, L. (2006b). Estimation of semiparametric additive coefficient model. *J. Statist. Plann. Inference* **136**, 2506-2534.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. Roy. Statist. Soc. Ser. B* **68**, 49-67.

Yuan, M. and Lin, Y. (2007). On the nonnegative garrote estimator, *J. Roy. Statist. Soc. Ser. B* **69**, 143-161.

Yang, L., Sperlich, S. and Härdle, W. (2003). Derivative estimation and testing in generalized additive models. *J. Statist. Plann. Inference* **115**, 521-542.

Department of Statistics, Oregon State University, Corvallis, OR 97330, U.S.A.

E-mail: xuel@stat.orst.edu

(Received May 2007; accepted April 2008)