Statistica Sinica **16**(2006), 1423-1446

ADDITIVE COEFFICIENT MODELING VIA POLYNOMIAL SPLINE

Lan Xue and Lijian Yang

Oregon State University and Michigan State University

Abstract: A flexible nonparametric regression model is considered in which the response depends linearly on some covariates, with regression coefficients as additive functions of other covariates. Polynomial spline estimators are proposed for the unknown coefficient functions, with optimal univariate mean square convergence rate under geometric mixing condition. Consistent model selection method is also proposed based on a nonparametric Bayes Information Criterion (BIC). Simulations and data examples demonstrate that the polynomial spline estimators are computationally efficient and as accurate as existing local polynomial estimators.

Key words and phrases: AIC, Approximation space, BIC, German real GNP, knot, mean square convergence, spline approximation.

1. Introduction

Parametric regression models are based on the assumption that a regression function follows a pre-determined parametric form with finitely many unknown parameters. A wrong model can lead to excessive estimation biases and erroneous inferences. In contrast, nonparametric models impose less stringent assumptions on the regression function. General nonparametric models, however, need large sample sizes to obtain reasonable estimators when the predictors are high dimensional. Much effort has been made to alleviate the "curse of dimensionality" by imposing appropriate structure on the regression function.

Of special importance is the varying coefficient model (Hastie and Tibshirani (1993)), whose regression function depends linearly on some regressors, with coefficients as smooth functions of other predictor variables, called tuning variables. A special type of varying coefficient model is called the functional coefficient model by Chen and Tsay (1993b). Here all tuning variables are the same and univariate. It was studied in the time series context by Cai, Fan and Yao (2000) and Huang and Shen (2004). Xue and Yang (2006) extended the functional coefficient model to the case when the tuning variable is multivariate, with additive structure on regression coefficients to avoid the "curse of dimensionality". The regression function of the new additive coefficient model is

$$m(\mathbf{X},\mathbf{T}) = \sum_{l=1}^{d_1} \alpha_l(\mathbf{X}) T_l, \qquad \alpha_l(\mathbf{X}) = \alpha_{l0} + \sum_{s=1}^{d_2} \alpha_{ls}(X_s), \tag{1.1}$$

in which the predictor vector $(\mathbf{X},\mathbf{T}) \in R^{d_2} \times R^{d_1}$, with $\mathbf{X} = (X_1,\ldots,X_{d_2})^T$, $\mathbf{T} = (T_1,\ldots,T_{d_1})^T$. This additive coefficient model includes as special cases the varying/functional coefficient models, as well as the additive model (Chen and Tsay (1993a) and Hastie and Tibshirani (1990)) and the linear regression model, see Xue and Yang (2006) for asymptotic distributions of local polynomial marginal integration estimators of the unknown coefficients. The parameters $\{\alpha_{l0}\}_{l=1}^{d_1}$ are estimated at the parametric rate $1/\sqrt{n}$, and the nonparametric functions $\{\alpha_{ls}(x_s)\}_{l=1,s=1}^{d_1,d_2}$ are estimated at the univariate smoothing rate. Due to an integration step and its 'local' nature, the kernel method in Xue and Yang (2006) is computationally expensive. Based on a sample of size n, to estimate the coefficient functions $\{\alpha_l(\mathbf{x})\}_{l=1}^{d_1}$ in (1.1) at any fixed point $\mathbf{x}$, a total of $(d_2+1)n$ least squares estimations have to be done. So the computational burden increases dramatically as the sample size n and the dimension d_2 of the tuning variables increase.

In this paper, we propose a faster polynomial spline estimator for (1.1). In contrast to a local polynomial, a polynomial spline requires global smoothing. One solves only one least squares estimation to estimate all the components in the coefficient functions, regardless of the sample size n and the dimension of the tuning variable d_2. So the computation is substantially reduced. As an alternative to using local polynomials, polynomial splines have been used to estimate various models, for example, the additive model (Stone (1985)), the functional ANOVA model (Huang (1998a,b)), the varying coefficient model (Huang, Wu and Zhou (2002)), and the additive model for weakly dependent data (Huang and Yang (2004)). We have established the polynomial spline estimators' rate of convergence under geometric mixing conditions. A major innovation is the use of an approximation space with possibly unbounded basis. Huang, Wu and Zhou (2002), for instance, imposed the assumption that $\mathbf{T} = (T_1,\ldots,T_{d_1})^T$ in (1.1) has a compactly supported distribution to make their basis bounded. Our method, in contrast, imposes only mild moment conditions on $\mathbf{T}$.

The paper is organized as follows. Section 2 discusses the identification issue for model (1.1). Section 3 presents the polynomial spline estimators, their L_2 consistency and a model selection procedure based on Bayes Information Criterion (BIC). These estimation and model selection procedures adapt automatically to the varying coefficient model (Hastie and Tibshirani (1993)), the functional coefficient model (Chen and Tsay (1993b)), the additive model

(Hastie and Tibshirani (1990) and Chen and Tsay (1993a)), and the linear regression model, a feature not shared by any kernel type estimators. Section 4 applies the methods to simulated and empirical examples. Technical assumptions and proofs are given in the Appendix.

2. The Model

Let $\{(Y_i, \mathbf{X}_i, \mathbf{T}_i)\}_{i=1}^n$ be a sequence of strictly stationary observations, with univariate response Y_i, d_2 and d_1-variate predictors $\mathbf{X}_i$ and $\mathbf{T}_i$. With unknown conditional mean and variance functions $m(\mathbf{X}_i, \mathbf{T}_i) = E(Y_i|\mathbf{X}_i, \mathbf{T}_i), \sigma^2(\mathbf{X}_i, \mathbf{T}_i) = \mathrm{Var}\,(Y_i|\mathbf{X}_i, \mathbf{T}_i)$, the observations satisfy

$$Y_i = m\left(\mathbf{X}_i, \mathbf{T}_i\right) + \sigma\left(\mathbf{X}_i, \mathbf{T}_i\right)\varepsilon_i. \tag{2.1}$$

The errors $\{\varepsilon_i\}_{i=1}^n$ are i.i.d. with $E\left(\varepsilon_i|\mathbf{X}_i, \mathbf{T}_i\right) = 0$, $E\left(\varepsilon_i^2|\mathbf{X}_i, \mathbf{T}_i\right) = 1$, and ε_i independent of the σ-field $\mathcal{F}_i = \sigma\left\{\left(\mathbf{X}_j, \mathbf{T}_j\right), j \le i\right\}$ for $i = 1, \ldots, n$. The variables $(\mathbf{X}_i, \mathbf{T}_i)$ can consist of either exogenous variables or lagged values of Y_i. For the additive coefficient model, the regression function m takes the form in (1.1), and satisfies the identification conditions that

$$E\left\{\alpha_{ls}\left(X_{is}\right)\right\} = 0, 1 \le l \le d_1, 1 \le s \le d_2. \tag{2.2}$$

The conditional variance function $\sigma^2\left(\mathbf{x}, \mathbf{t}\right)$ is assumed to be continuous and bounded. As in most works on nonparametric smoothing, estimation of the functions $\left\{\alpha_{ls}\left(x_s\right)\right\}_{l=1,s=1}^{d_1,d_2}$ is conducted on compact sets. Without lose of generality, let the compact set be $\chi = [0,1]^{d_2}$.

Following Stone (1985, p.693), the space of s-centered square integrable functions on $[0,1]$ is

$$\mathcal{H}_s^0 = \left\{\alpha : E\left\{\alpha\left(X_s\right)\right\} = 0, E\left\{\alpha^2\left(X_s\right)\right\} < +\infty\right\}, 1 \le s \le d_2.$$

Next define the model space $\mathcal{M}$, a collection of functions on $\chi \times R^{d_1}$, as

$$\mathcal{M} = \Big\{m\left(\mathbf{x}, \mathbf{t}\right) = \sum_{l=1}^{d_1} \alpha_l\left(\mathbf{x}\right) t_l; \quad \alpha_l\left(\mathbf{x}\right) = \alpha_{l0} + \sum_{s=1}^{d_2} \alpha_{ls}(x_s); \alpha_{ls} \in \mathcal{H}_s^0\Big\},$$

in which $\{\alpha_{l0}\}_{l=1}^{d_1}$ are finite constants. The constraints that $E\left\{\alpha_{ls}\left(X_s\right)\right\} = 0, 1 \le s \le d_2$, ensure unique additive representation of α_l, but are not necessary for the definition of space $\mathcal{M}$.

In what follows, denote by E_n the empirical expectation, $E_n\varphi = \sum_{i=1}^n \varphi(\mathbf{X}_i, \mathbf{T}_i)/n$. We introduce two inner products on $\mathcal{M}$. For functions $m_1, m_2 \in \mathcal{M}$, the theoretical and empirical inner products are defined, respectively, as $\langle m_1, m_2\rangle = E\left\{m_1\left(\mathbf{X}, \mathbf{T}\right) m_2\left(\mathbf{X}, \mathbf{T}\right)\right\}$ and $\langle m_1, m_2\rangle_n = E_n\left\{m_1\left(\mathbf{X}, \mathbf{T}\right) m_2\left(\mathbf{X}, \mathbf{T}\right)\right\}$. The corresponding induced norms are $\|m_1\|_2^2 = Em_1^2(\mathbf{X}, \mathbf{T})$ and $\|m_1\|_{2,n}^2 = E_n m_1^2(\mathbf{X}, \mathbf{T})$.

The model space $\mathcal{M}$ is called *theoretically* (*empirically*) *identifiable* if, for any $m \in \mathcal{M}$, $\|m\|_2 = 0$ ($\|m\|_{2,n} = 0$) implies that $m = 0$ a.s..

Lemma 1. *Under assumptions* (C1) *and* (C2) *in the Appendix, there exists a constant* $C > 0$ *such that*

$$\|m\|_2^2 \geq C\Big\{ \sum_{l=1}^{d_1} \Big(\alpha_{l0}^2 + \sum_{s=1}^{d_2} \|\alpha_{ls}\|_2^2 \Big)\Big\}, \quad \forall\ m = \sum_{l=1}^{d_1} \Big(\alpha_{l0} + \sum_{s=1}^{d_2} \alpha_{ls}\Big) t_l \in \mathcal{M}.$$

Hence for any $m \in \mathcal{M}, \|m\|_2 = 0$ *implies* $\alpha_{l0} = 0, \alpha_{ls} = 0$ *a.s., for all* $1 \leq l \leq d_1, 1 \leq s \leq d_2$*. Consequently the model space* $\mathcal{M}$ *is theoretically identifiable.*

Proof. Let $A_l(\mathbf{X}) = \alpha_{l0} + \sum_{s=1}^{d_2} \alpha_{ls}(X_s)$ and $\mathbf{A}(\mathbf{X}) = (A_1(\mathbf{X}), \ldots, A_{d_1}(\mathbf{X}))^T$. Under (C2),

$$\begin{aligned} \|m\|_2^2 &= E\Big[\sum_{l=1}^{d_1} \Big\{\alpha_{l0} + \sum_{s=1}^{d_2} \alpha_{ls}(X_s)\Big\} T_l\Big]^2 = E\left[\mathbf{A}(\mathbf{X})^T \mathbf{T}\mathbf{T}^T \mathbf{A}(\mathbf{X})\right] \\ &\geq c_3 E\left[\mathbf{A}(\mathbf{X})^T \mathbf{A}(\mathbf{X})\right] = c_3 E\Big[\sum_{l=1}^{d_1} \Big\{\alpha_{l0} + \sum_{s=1}^{d_2} \alpha_{ls}(X_s)\Big\}^2\Big] \end{aligned}$$

which, by (2.2), is $c_3[\sum_{l=1}^{d_1} \alpha_{l0}^2 + \sum_{l=1}^{d_1} E\{\sum_{s=1}^{d_2} \alpha_{ls}(X_s)\}^2]$. Applying Lemma 1 of Stone (1985)

$$\|m\|_2^2 \geq c_3 \left[\sum_{l=1}^{d_1} \alpha_{l0}^2 + \left\{\frac{1-\delta}{2}\right\}^{d_2 - 1} \sum_{l=1}^{d_1} \sum_{s=1}^{d_2} E\alpha_{ls}^2(X_s)\right],$$

where $\delta = (1 - c_1/c_2)^{1/2}$, with $0 < c_1 \leq c_2$ as specified in (C1). By taking $C = c_3\{(1-\delta)/2\}^{d_2 - 1}$, the first part is proved. To show identifiability, notice that for any $m = \sum_{l=1}^{d_1} (\alpha_{l0} + \sum_{s=1}^{d_2} \alpha_{ls}) t_l \in \mathcal{M}$, with $\|m\|_2 = 0$, we have

$$0 = E\Big[\sum_{l=1}^{d_1} \Big\{\alpha_{l0} + \sum_{s=1}^{d_2} \alpha_{ls}(X_s)\Big\} T_l\Big]^2 \geq C\Big[\sum_{l=1}^{d_1} \alpha_{l0}^2 + \sum_{l=1}^{d_1} \sum_{s=1}^{d_2} E\left\{\alpha_{ls}^2(X_s)\right\}\Big],$$

which entails that $a_{l0} = 0$ and $\alpha_{ls}(X_s) = 0$ a.s. for all $1 \leq l \leq d_1, 1 \leq s \leq d_2$, or $m = 0$ a.s..

3. Polynomial Spline Estimation

3.1. The estimators

In this paper, denote by $C^p\left([0,1]\right)$ the space of p-times continuously differentiable functions. For each tuning variable direction, $s = 1, \ldots, d_2$, a knot sequence $k_{s,n}$ with N_n interior knots is $k_{s,n} = \{0 = x_{s,0} < x_{s,1} < \cdots < x_{s,N_n} < x_{s,N_n+1} = 1\}$.

For an integer $p \geq 1$, define $\varphi_s = \varphi^p([0,1], k_{s,n})$, a subspace of $C^{p-1}([0,1])$ consisting of functions that are polynomial of degree p (or less) on the intervals $[x_{s,i}, x_{s,i+1}), i = 0, \ldots, N_n - 1$, and $[x_{s,N_n}, x_{s,N_n+1}]$. Functions in φ_s are called polynomial splines, piecewise polynomials connected smoothly on the interior knots. A polynomial spline with degree $p = 1$ is a continuous piecewise linear function, etc. The space φ_s is determined by the polynomial degree p and the knot sequence $k_{s,n}$. Let $h_s = h_{s,n} = \max_{i=0,\ldots,N_n} |x_{s,i+1} - x_{s,i}|$, called the mesh size of $k_{s,n}$, and define $h = \max_{s=0,\ldots,d_2} h_s$, the overall smoothness measure.

Lemma 2. *For $1 \leq s \leq d_2$, let $\varphi_s^0 = \{g_s : g_s \in \varphi_s, Eg_s(X_s) = 0\}$ be the space of centered polynomial splines. There exists a constant $c > 0$ so that, for any $\alpha_s \in \mathcal{H}_s^0 \cap C^{p+1}([0,1])$, there exists a $g_s \in \varphi_s^0$ with $\|\alpha_s - g_s\|_\infty \leq c\|\alpha_s^{(p+1)}\|_\infty h_s^{p+1}$.*

Proof. According to de Boor (2001, p.149), there exists a constant $c > 0$ and a spline function $g_s^* \in \varphi_s$, such that $\|\alpha_s - g_s^*\|_\infty \leq c\|\alpha_s^{(p+1)}\|_\infty h_s^{p+1}$. Note that $|Eg_s^*| \leq |E(g_s^* - \alpha_s)| + |E\alpha_s| \leq \|g_s^* - \alpha_s\|_\infty$. Thus for $g_s = g_s^* - Eg_s^* \in \varphi_s^0$, one has

$$\|\alpha_s - g_s\|_\infty \leq \|\alpha_s - g_s^*\|_\infty + Eg_s^* \leq 2c\|\alpha_s^{(p+1)}\|_\infty h_s^{p+1}.$$

Lemma 2 entails that if the functions $\{\alpha_{ls}(x_s)\}_{l=1,s=1}^{d_1,d_2}$ in (1.1) are smooth, they are approximated well by centered splines $\{g_{ls}(x_s) \in \varphi_s^0\}_{l=1,s=1}^{d_1,d_2}$. As the definition of φ_s^0 depends on the unknown distribution of X_s, the empirically defined space $\varphi_s^{0,n} = \{g_s : g_s \in \varphi_s, E_n(g_{ls}) = 0\}$ is used. Intuitively, $m \in \mathcal{M}$ is approximated by some function from the approximate space

$$\mathcal{M}_n = \left\{ m_n(\mathbf{x}, \mathbf{t}) = \sum_{l=1}^{d_1} g_l(\mathbf{x}) t_l;\ g_l(\mathbf{x}) = \alpha_{l0} + \sum_{s=1}^{d_2} g_{ls}(x_s); g_{ls} \in \varphi_s^{0,n} \right\}.$$

Given observations $\{(Y_i, \mathbf{X}_i, \mathbf{T}_i)\}_{i=1}^n$ from (2.1), the estimator of the unknown regression function m is defined as its 'best' approximation from $\mathcal{M}_n$,

$$\hat{m} = \underset{m_n \in \mathcal{M}_n}{\operatorname{argmin}} \sum_{i=1}^{n} \{Y_i - m_n(\mathbf{X}_{i,} \mathbf{T}_i)\}^2. \tag{3.1}$$

To be precise, write $J_n = N_n + p$, and let $\{w_{s,0}, w_{s,1}, \ldots, w_{s,J_n}\}$ be a basis of the spline space φ_s, for $1 \leq s \leq d_2$. For example, the truncated power basis is used in the implementation

$$\left\{1, x_s, \ldots, x_s^p, (x_s - x_{s,1})_+^p, \ldots, (x_s - x_{s,N_n})_+^p \right\},$$

in which $(x)_+^p = (x_+)^p$. Let $\mathbf{w} = \{1, w_{1,1}, \ldots, w_{1,J_n}, \ldots, w_{d_2,1}, \ldots, w_{d_2,J_n}\}$, then $\{\mathbf{w}t_1, \ldots, \mathbf{w}t_{d_1}\}$ is a $R_n = d_1\{d_2 J_n + 1\}$ dimensional basis of $\mathcal{M}_n$, and (3.1) amounts to

$$\hat{m}\left(\mathbf{x},\mathbf{t}\right)=\sum_{l=1}^{d_1}\left\{\hat{c}_{l0}+\sum_{s=1}^{d_2}\sum_{j=1}^{J_n}\hat{c}_{ls,j}w_{s,j}\left(x_s\right)\right\}t_l, \tag{3.2}$$

in which the coefficients $\{\hat{c}_{l0},\hat{c}_{ls,j},1\le l\le d_1,1\le s\le d_2,1\le j\le J_n\}$ minimize the sum of squares

$$\sum_{i=1}^{n}\left(Y_i-\sum_{l=1}^{d_1}\left\{c_{l0}+\sum_{s=1}^{d_2}\sum_{j=1}^{J_n}c_{ls,j}w_{s,j}\left(X_{is}\right)\right\}T_{il}\right)^2 \tag{3.3}$$

with respect to $\{c_{l0},c_{ls,j},1\le l\le d_1,1\le s\le d_2,1\le j\le J_n\}$. Note that Lemma A.5 entails that, with probability approaching one, the sum of squares in (3.3) has a unique minimizer. For $1\le l\le d_1,1\le s\le d_2$, let $\alpha_{ls}^{*}(x_s)=\sum_{j=1}^{J_n}\hat{c}_{ls,j}w_{s,j}(x_s)$. Then the estimators of $\{\alpha_{l0}\}_{l=1}^{d_1}$ and $\{\alpha_{ls}\left(x_s\right)\}_{l=1,s=1}^{d_1,d_2}$ in (1.1) are

$$\begin{aligned}\hat{\alpha}_{l0}&=\hat{c}_{l0}+\sum_{s=1}^{d_2}E_n\alpha_{ls}^{*},\quad 1\le l\le d_1,\\ \hat{\alpha}_{ls}(x_s)&=\alpha_{ls}^{*}(x_s)-E_n\alpha_{ls}^{*}\quad 1\le l\le d_1,1\le s\le d_2,\end{aligned} \tag{3.4}$$

where $\{\hat{\alpha}_{ls}(x_s)\}_{l=1,s=1}^{d_1,d_2}$ are empirically centered to consistently estimate the theoretically centered function components in (1.1). These estimators are determined by the knot sequences $\{k_{s,n}\}_{s=1}^{d_2}$ and the polynomial degree p, which relates to the smoothness of the regression function. We refer to an estimator by its degree p. For example, a linear spline fit corresponds to $p=1$.

Theorem 1. *Under assumptions* (C1)−(C5) *in the Appendix, if* $\alpha_{ls}\in C^{p+1}$ $([0,1])$, *for* $1\le l\le d_1,1\le s\le d_2$, *one has*

$$\begin{aligned}&\left\|\hat{m}-m\right\|_2=O_p\left(h^{p+1}+(nh)^{-\frac{1}{2}}\right),\\ &\max_{1\le l\le d_1}\left|\hat{\alpha}_{l0}-\alpha_{l0}\right|+\max_{1\le l\le d_1,1\le s\le d_2}\left\|\hat{\alpha}_{ls}-\alpha_{ls}\right\|_2=O_p\left(h^{p+1}+(nh)^{-\frac{1}{2}}\right).\end{aligned}$$

Theorem 1 has the optimal order of h as $n^{-1/(2p+3)}$, in which case $\left\|\hat{\alpha}_{ls}-\alpha_{ls}\right\|_2$ $=O_p\left(n^{-1/(2p+3)}\right)$, which is the same rate of mean square error as achieved by the marginal integration estimators in Xue and Yang (2006).

3.2. Knot number selection

An appropriate selection of the knot sequence is important to efficiently implement the proposed polynomial spline estimation method. Stone (1985) found that the number of knots is more crucial than their location. We discuss an approach to select the number of interior knots N_n using the AIC criteria. For knot location, we use either equally spaced knots or quantile knots (sample quantiles with the same number of observations between any two adjacent knots).

According to Theorem 1, the optimal order of N_n is $n^{1/(2p+3)}$. Thus we propose selecting the 'optimal' N_n, denoted $\hat{N}_n^{\mathrm{opt}}$, from $[0.5N_r, \min(5N_r, Tb)]$, with $N_r = n^{1/(2p+3)}$ and $Tb = \{n/(4d_1) - 1\}/d_2$ to ensure that the total number of parameters in the least square estimation is less than $n/4$.

To be specific, we denote the estimator for the i-th response Y_i by $\hat{Y}_i(N_n) = \hat{m}(\mathbf{X}_i, \mathbf{T}_i)$, for $i = 1, \ldots, n$. Here $\hat{m}$ depends on the knot sequence as given in (3.2). Let $q_n = (1 + d_2 N_n) d_1$ be the total number of parameters in (3.3). Then $\hat{N}_n^{\mathrm{opt}}$ is the one minimizing the AIC value

$$\hat{N}_n^{\mathrm{opt}} = \underset{N_n \in [0.5N_r, \min(5N_r, Tb)]}{\mathrm{argmin}} \mathrm{AIC}(N_n), \tag{3.5}$$

where $\mathrm{AIC}(N_n) = \log(\mathrm{MSE}) + 2q_n/n$, with $\mathrm{MSE} = \sum_{i=1}^n \{Y_i - \hat{Y}_i(N_n)\}^2/n$.

3.3. Model selection

For the full model (1.1), a natural question to ask is whether the functions $\{\alpha_{ls}(x_s)\}_{l=1,s=1}^{d_1,d_2}$ are all significant. A simpler model, found by setting some of $\{\alpha_{ls}(x_s)\}_{l=1,s=1}^{d_1,d_2}$ to zero, may perform as well as the full model. For $1 \le l \le d_1$, let S_l denote the set of indices of the tuning variables which are significant in the coefficient function of T_l, and S the collection of indices from all the sets S_l. In particular, S for the full model is $S_f = \{S_{f1}, \ldots, S_{fd_1}\}$, where $S_{fl} \equiv \{1, \ldots, d_2\}, 1 \le l \le d_1$. For two indices $S = \{S_1, \ldots, S_{d_1}\}, S' = \{S'_1, \ldots, S'_{d_1}\}$, we say that $S \subset S'$ if and only if $S_l \subset S'_l$, for all $1 \le l \le d_1$ and $S_l \ne S'_l$, for some l. The goal is to select the smallest sub-model, with indices $S \subset S_f$, which gives the same information as the full additive coefficient model. Following Huang and Yang (2004), the Akaike Information Criterion (AIC) and the Bayes Information Criterion (BIC) are considered.

For a submodel m_S with indices $S = \{S_1, \ldots, S_{d_1}\}$, let $N_{n,S}$ be the number of interior knots used to estimate the model m_S, and $J_{n,S} = N_{n,S} + p$. As in the full model estimation, let $\{\hat{c}_{l0}, \hat{c}_{ls,j}, 1 \le l \le d_1, s \in S_l, 1 \le j \le J_{n,S}\}$ be the minimizer of the sum of squares

$$\sum_{i=1}^{n} \left(Y_i - \sum_{l=1}^{d_1} \left\{ c_{l0} + \sum_{s \in S_l} \sum_{j=1}^{J_{n,S}} c_{ls,j} w_{s,j}(X_{is}) \right\} T_{il} \right)^2. \tag{3.6}$$

Define

$$\hat{m}_S(\mathbf{x}, \mathbf{t}) = \sum_{l=1}^{d_1} \left\{ \hat{c}_{l0} + \sum_{s \in S_l} \sum_{j=1}^{J_{n,S}} \hat{c}_{ls,j} w_{s,j}(x_s) \right\} t_l. \tag{3.7}$$

Let $\hat{Y}_{i,s} = \hat{m}_S\left(\mathbf{X}_i, \mathbf{T}_i\right)$,$i = 1, \ldots, n$, $\text{MSE}_S = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_{i,s}\right)^2/n$, and $q_S = \sum_{l=1}^{d_1}\left\{1 + \#\left(S_l\right)J_{n,s}\right\}$, the total number of parameters in (3.6). Then the submodel is selected with the smallest AIC (or BIC) values, defined as $\text{AIC}_S = \log(\text{MSE}_S) + 2q_S/n$, $\text{BIC}_S = \log(\text{MSE}_S) + \log(n)q_S/n$.

Let S_0 and $\hat{S}$ be the index set of the true model and the selected model, respectively. The outcome is defined as correct fitting, if $\hat{S} = S_0$; overfitting, if $S_0 \subset \hat{S}$; and underfitting, if $S_0 \not\subset \hat{S}$, that is, $S_{0l} \not\subset \hat{S}_l$, for some l. For either overfitting or underfitting, we write $\hat{S} \neq S_0$.

Theorem 2. *Under the same conditions as in Theorem* 1, *and* $N_{n,S} \asymp N_{n,S_0} \asymp n^{1/(2p+3)}$, *the BIC is consistent: for any* $S \neq S_0$, $\lim_{n\to\infty} P\left(BIC_S > BIC_{S_0}\right) = 1$, *hence* $\lim_{n\to\infty} P(\hat{S} = S_0) = 1$.

The condition that $N_{n,S} \asymp N_{n,S_0}$ is essential for the BIC to be consistent. As a referee pointed out, the number of parameters q_S depends on the number of knots and the number of additive terms used in the model function. To ensure BIC consistency, roughly the same sufficient number of knots should be used to estimate the various models so that q_S depends only on the number of functions terms. In implementation, we have used the same number of interior knots N_n^{opt} (see (3.5), the optimal knot number for the full additive coefficient model) in the estimation of all the submodels.

4. Examples

In this section, we first analyze two simulated data sets with i.i.d. and time series set-ups, respectively. Both data sets have sample sizes $n = 100$, 250, 500 and 100 replications. Later the proposed methods are applied to an empirical example: West German real GNP.

The performance of the function estimators is assessed by the averaged integrated squared error (AISE). Denoting the estimator of α_{ls} in the i-th replication as $\hat{\alpha}_{i,ls}$, and $\{x_m\}_{m=1}^{n_{\text{grid}}}$ the grid points where the functions are evaluated, we take

$$\text{ISE}(\hat{\alpha}_{i,ls}) = \frac{1}{n_{\text{grid}}}\sum_{m=1}^{n_{\text{grid}}}\left\{\hat{\alpha}_{i,ls}(x_m) - \alpha_{ls}(x_m)\right\}^2$$

and

$$\text{AISE}(\hat{\alpha}_{ls}) = \frac{1}{100}\sum_{i=1}^{100}\text{ISE}(\hat{\alpha}_{i,ls}).$$

4.1. Simulated Example 1

The data are generated from the model

$$Y = \left\{c_1 + \alpha_{11}\left(X_1\right) + \alpha_{12}\left(X_2\right)\right\}T_1 + \left\{c_2 + \alpha_{21}\left(X_1\right) + \alpha_{22}\left(X_2\right)\right\}T_2 + \varepsilon,$$

with $c_1 = 2$, $c_2 = 1$, $\alpha_{11}(x) = \sin\{2(4x-2)\} + 2\exp\left\{-2(x-0.5)^2\right\}$, $\alpha_{12}(x) = x$, $\alpha_{21}(x) = \sin(x)$, and $\alpha_{22}(x) = 0$. The vector $\mathbf{X} = (X_1, X_2)^T$ is uniformly distributed on $[-\pi, \pi]^2$ independent of the standard bivariate normal $\mathbf{T} = (T_1, T_2)^T$. The error ε is a standard normal variable independent of $(\mathbf{X}, \mathbf{T})$.

The functions are estimated by using linear splines ($p = 1$), cubic splines ($p = 3$) and the marginal integration method of Xue and Yang (2006). For $s = 1, 2$, let $x^i_{s,\min}$, $x^i_{s,\max}$ denote the smallest and largest observation of the variable x_s in the i-th replication. Knots are placed evenly on the intervals $[x^i_{s,\min}, x^i_{s,\max}]$, with the number of interior knots N_n selected by AIC as in Subsection 3.2. The functions $\{\alpha_{ls}\}^{2,2}_{l=1,s=1}$ are estimated on a grid of equally-spaced points $x_m, m = 1, \ldots, n_{\text{grid}}$ with $x_1 = -0.975\pi, x_{n_{\text{grid}}} = 0.975\pi, n_{\text{grid}} = 62$.

Table 1 reports the means and standard errors (in the parentheses) of $\{\hat{c}_l\}_{l=1,2}$ and the AISEs of $\{\hat{\alpha}_{ls}\}^{s=1,2}_{l=1,2}$ for all the three fits. The spline fits are generally comparable, with the cubic fit better than the linear fit for larger sample sizes ($n = 250, 500$), the standard errors of the constant estimators. The AISEs of the function estimators decrease as samples size increases, confirming Theorem 1. The polynomial spline methods also perform better than the marginal integration method. Figure 1 gives the plots of the 100 cubic spline fits for all sample sizes, clearly illustrating the estimation improvements as sample size increases. Plots d1−d4 of the typical estimated curves (whose ISE is the median of the 100 ISEs from the replications) seem satisfactory for sample sizes as small as 100.

Table 1. Simulated Example 1: the means and standard errors (in parentheses) of $\hat{c}_1$, $\hat{c}_2$ and the AISEs of $\hat{\alpha}_{11}$, $\hat{\alpha}_{12}$, $\hat{\alpha}_{21}$, $\hat{\alpha}_{22}$ by two methods: marginal integration and polynomial spline.

Integration fit	$c_1 = 2$	$c_2 = 1$	α_{11}	α_{12}	α_{21}	α_{22}
$n = 100$	2.0029(0.0180)	0.9805(0.0134)	0.8339	0.4699	0.7948	0.4665
$n = 250$	2.0143(0.0081)	1.0056(0.0089)	0.4545	0.1324	0.3219	0.3741
$n = 500$	1.9994(0.0047)	0.9961(0.0042)	0.0657	0.0378	0.0629	0.0375
Spline fit $p = 1$						
$n = 100$	2.0102(0.0122)	0.9896(0.0123)	0.1279	0.0629	0.0690	0.0558
$n = 250$	1.9997(0.0046)	0.9795(0.0044)	0.0648	0.0291	0.0328	0.0261
$n = 500$	1.9992(0.0023)	0.9988(0.0022)	0.0438	0.0165	0.0164	0.0152
Spline fit $p = 3$						
$n = 100$	2.0026(0.0126)	0.9858(0.0124)	0.1699	0.0700	0.0659	0.0634
$n = 250$	1.9988(0.0046)	0.9810(0.0043)	0.0594	0.0320	0.0313	0.0285
$n = 500$	2.0003(0.0023)	0.9982(0.0022)	0.0387	0.0162	0.0156	0.0149

As mentioned earlier, the polynomial spline method enjoys great computational efficiency: it takes less than 20 seconds to run 100 simulations on a Pentium

4 PC, regardless of sample sizes. In contrast, it takes marginal integration about 2 hours to run 100 simulations with $n = 100$; and about 20 hours with $n = 500$.

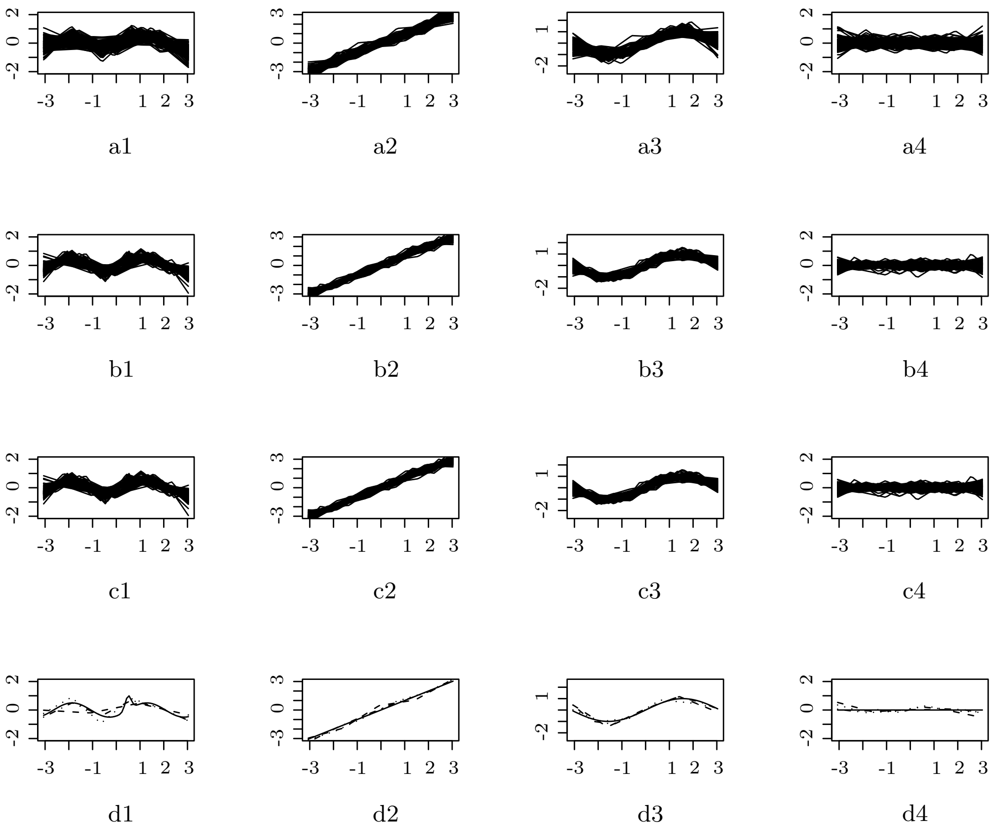


Figure 1. Plots of the estimated coefficient functions in Example 1. (a1-a4) are plots of the 100 estimated curves for $\alpha_{11}(x) = \sin(2(4x-2)) + 2\exp(-2(x-0.5)^2)$, $\alpha_{12}(x) = x$, $\alpha_{21}(x) = \sin(x)$, and $\alpha_{22}(x) = 0$, for $n = 100$. (b1−b4) and (c1−c4) are the same as (a1-a4), but for $n = 250$ and $n = 500$ respectively. (d1−d4) are plots of the typical estimators; the solid curve represents the true curve, the dotted curve is the typical estimated curve for $n = 100$, the dot-dashed and dashed curves are for $n = 250, 500$ respectively.

For each replication, model selection is also conducted according to the criteria proposed in Subsection 3.3 for polynomial splines with $p = 1, 2, 3$. The model selection results are presented in Table 2, indexed as ‘Example 1’. The BIC is rather accurate: more than 86% correct selection when the sample size is as small as 100, and absolute correct selection when sample size increases to 250 and 500, corroborating Theorem 2 on the consistency of BIC. AIC clearly tends to over-fit and never under-fits.

Table 2. Simulation results of model selection using polynomial spline fits with BIC and AIC. For each setup, the first, second and third columns give the number of underfits, correct fits and overfits in 100 simulations.

	n	BIC									AIC								
		$p=1$			$p=2$			$p=3$			$p=1$			$p=2$			$p=3$		
Example 1	100	3	94	3	0	86	14	4	95	1	0	89	11	4	95	1	0	88	12
	250	0	100	0	0	100	0	0	100	0	0	90	10	0	89	11	0	90	10
	500	0	100	0	0	100	0	0	100	0	0	91	9	0	93	7	0	92	8
Example 2	100	11	88	1	0	69	31	5	94	5	8	80	12	14	46	40	0	82	18
	250	0	100	0	0	100	0	0	100	0	0	94	6	0	96	4	0	84	16
	500	0	100	0	0	100	0	0	100	0	0	98	2	0	100	0	0	98	2

4.2. Simulated Example 2

The data is generated from a nonlinear AR model

$$Y_t=\{c_1+\alpha_{11}(Y_{t-1})+\alpha_{12}(Y_{t-2})\}Y_{t-3}+\{c_2+\alpha_{21}(Y_{t-1})+\alpha_{22}(Y_{t-2})\}Y_{t-4}+0.1\varepsilon_t,$$

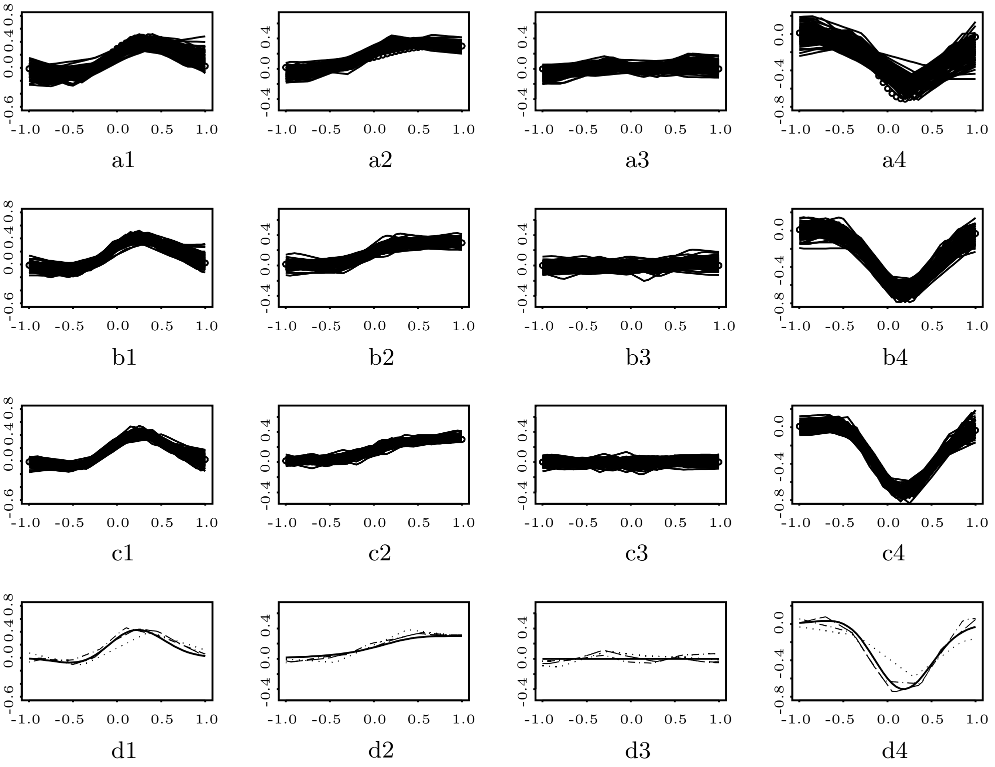
with i.i.d. standard normal noise ε_t, $c_1 = 0.2$, $c_2 = -0.3$ and

$$\alpha_{11}(u) = (0.3+u)\exp(-4u^2), \qquad \alpha_{12}(u) = \frac{0.3}{\{1+(u-1)^4\}},$$
$$a_{21}(u) = 0, \qquad \alpha_{22}(u) = -(0.6+1.2u)\exp(-4u^2).$$

In each replication, a total of $1,000+n$ observations are generated, and only the last n observations are used to ensure approximate stationarity. In this example, we have used linear splines on the quantile knot sequences. The coefficient functions $\{\alpha_{ls}\}_{l=1,s=1}^{2,2}$ are estimated on a grid of equally-spaced points on the interval $[-1,1]$, with the number of grid points $n_{\text{grid}} = 41$. Table 3 contains the means and standard errors of $\{\hat{c}_l\}_{l=1,2}$ and the AISEs of $\{\hat{\alpha}_{ls}\}_{l=1,2}^{s=1,2}$. The results are also graphically presented in Figure 2. Similar to Example 1, estimation improves as sample size increases, supporting our asymptotic result (Theorem 1). The model selection results are presented in Table 2, indexed as 'Example 2'. As in Example 1, AIC tends to overfit compared with BIC, and for $n = 250, 500$, the model selection result is satisfactory.

Table 3. Results of Example 2: the means and standard errors (in parentheses) of $\hat{c}_1$ and $\hat{c}_2$, and the AISEs of $\hat{\alpha}_{11}$, $\hat{\alpha}_{12}$, $\hat{\alpha}_{21}$, $\hat{\alpha}_{22}$ of linear splines.

Spline fit $p=1$	$c_1 = 0.2$	$c_2 = -0.3$	α_{11}	α_{12}	α_{21}	α_{22}
$n = 100$	0.2504(0.0481)	−0.2701(0.0374)	0.0113	0.0050	0.0042	0.0195
$n = 250$	0.1983(0.0271)	−0.2936(0.0279)	0.0030	0.0021	0.0025	0.0039
$n = 500$	0.1989(0.0202)	−0.2975(0.0209)	0.0015	0.0011	0.0016	0.0019



a1 a2 a3 a4

b1 b2 b3 b4

c1 c2 c3 c4

d1 d2 d3 d4

Figure 2. Plots of the estimated coefficient functions in Example 2. (a1−a4) are plots of the 100 estimated curves for $\alpha_{11}(u) = (0.3 + u)\exp(-4u^2)$, $\alpha_{12}(u) = 0.3/\{1+(u–1)^4\}$, $\alpha_{21}(u)=0$ and $\alpha_{22}(u) = -(0.6+1.2u)\exp(-4u^2)$ when $n = 100$. (b1−b4) and (c1−c4) are the same as (a1-a4), but when $n = 250, 500$ respectively. (d1−d4) are plots of the typical estimators: the solid curve represents the true curve, the dotted curve is the typical estimated curve for $n = 100$, the dot-dashed and dashed curves are for $n = 250, 500$ respectively.

4.3. West German real GNP

Now we compare the estimation and prediction performances of the polynomial spline and marginal integration methods using West German real GNP. For other empirical examples, see Xue and Yang (2005). The data consists of the quarterly West German real GNP from January 1960 to December 1990, denote as $\{G_t\}_{t=1}^{124}$, where G_t is the real GNP in the t-th quarter (the first quarter being from January 1, 1960 to April 1, 1960). From its time plot (Figure 3), $\{G_t\}_{t=1}^{124}$ appears to have both trend and seasonality. After removing the seasonal means from $\{\log (G_{t+4}/G_{t+3})\}_{t=1}^{120}$, we obtain a more stationary time series, denoted as

$\{Y_t\}_{t=1}^{120}$, whose time plot is given in Figure 4. As the nonparametric alternative to the linear autoregressive model selected by BIC,

$$Y_t = a_1 Y_{t-2} + a_2 Y_{t-4} + \sigma\varepsilon_t, \tag{4.1}$$

Xue and Yang (2006). proposed the additive coefficient model

$$Y_t = \{c_1 + \alpha_{11}(Y_{t-1}) + \alpha_{12}(Y_{t-8})\} Y_{t-2} + \{c_2 + \alpha_{21}(Y_{t-1}) + \alpha_{22}(Y_{t-8})\} Y_{t-4} + \sigma\varepsilon_t, \tag{4.2}$$

which is fitted by linear and cubic splines. Following Xue and Yang (2006), we use the first 110 observations for estimation and the last 10 observations for one-step prediction. Table 4 gives the averaged squared estimation errors (ASE) and averaged squared prediction errors (ASPE) of different fittings. Polynomial splines are better than the marginal integration method overall, while (4.2) significantly improves over (4.1) in both estimation and prediction. For visualization, plots of the function estimates are given in Figure 5.

Table 4. German real GNP: the ASE's and ASPE's of five fits.

	ASE	ASPE
Integration fit, $p = 1$	0.000201	0.000085
Integration fit, $p = 3$	0.000205	0.000077
Spline fit $p = 1$	0.000183	0.000095
Spline fit $p = 3$	0.000176	0.000082
Linear AR fit	0.000253	0.000112

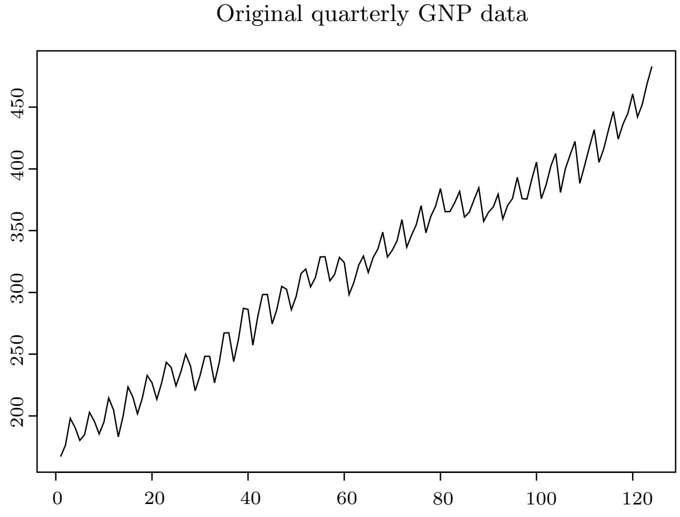


Figure 3. GNP data: time plot of the series $\{G_t\}_{t=1}^{124}$.

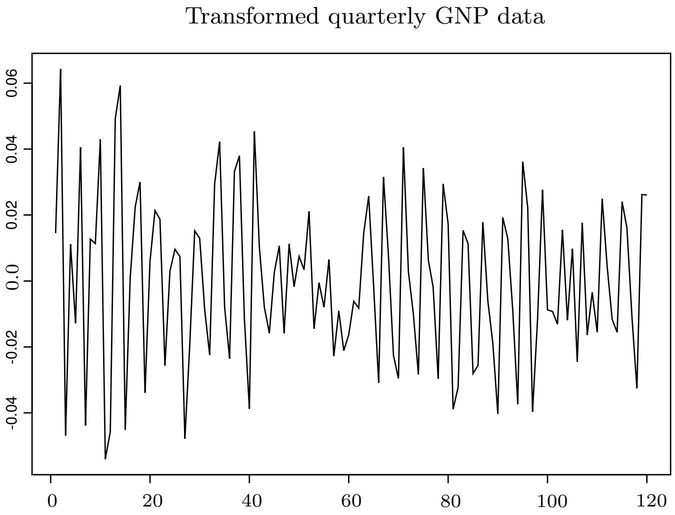


Figure 4. GNP data after transformation: time plot of the series $\{Y_t\}_{t=1}^{120}$.

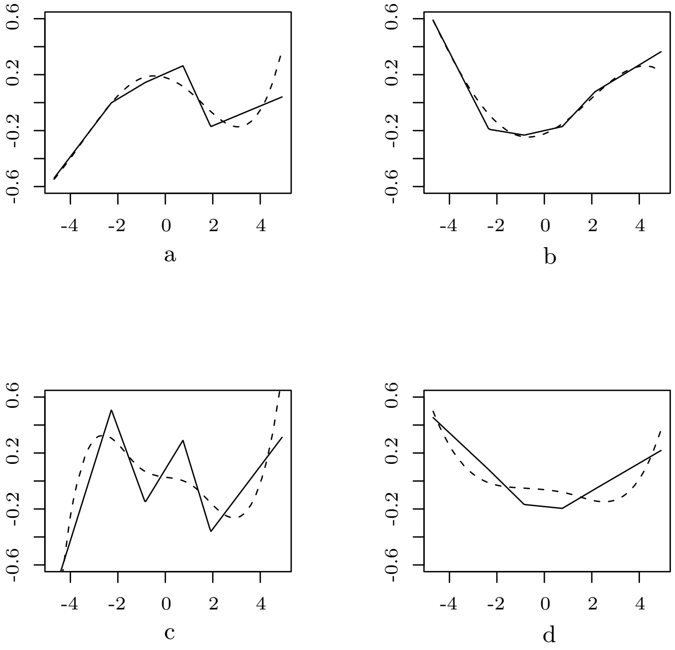


Figure 5. GNP data: spline approximation of the functions in model (4.2): (a) $\hat{\alpha}_{11}$; (b) $\hat{\alpha}_{12}$; (c) $\hat{\alpha}_{21}$; (d) $\hat{\alpha}_{22}$. Here solid lines denote the estimation results using linear splines, and dotted lines denote estimates using cubic splines.

Appendix

A.1. Assumptions and notations

The following assumptions are needed for our theoretical results.

(C1) *The tuning variables* $\mathbf{X} = (X_1, \ldots, X_{d_2})$ *are compactly supported and, without lose of generality, we assume that the support is* $\chi = [0,1]^{d_2}$. *The joint density of* $\mathbf{X}$, *denoted by* $f(\mathbf{x})$, *is absolutely continuous and* $0 < c_1 \leq \min_{\mathbf{x}\in\chi} f(\mathbf{x}) \leq \max_{\mathbf{x}\in\chi} f(\mathbf{x}) \leq c_2 < \infty$, *for positive constants* c_1 *and* c_2.

(C2) (i) *There exist positive constants* c_3 and c_4 *such that* $c_3 I_{d_1} \leq E(\mathbf{T}\mathbf{T}^T|\mathbf{X} = \mathbf{x}) \leq c_4 I_{d_1}$ *for all* $\mathbf{x} \in \chi$, *with* I_{d_1}*being the* $d_1 \times d_1$ *identity matrix. There exist positive constants* c_5 *and* c_6 *such that* $c_5 \leq E\{(T_l T_{l'})^{2+\delta_0}|\mathbf{X} = \mathbf{x}\} \leq c_6$ *a.s. for some* $\delta_0 > 0$, $l, l^{'} = 1, \ldots, d_1$.

(ii) *For some sufficient large* $m > 0$, $E\left|T_l\right|^m < +\infty$, *for* $l = 1, \ldots, d_1$.

(C3) *For the* d_2 *sets of knots* $k_{s,n} = \{0 = x_{s,0} \leq x_{s,1} \leq \cdots \leq x_{s,N_n} \leq x_{s,N_n+1} = 1\}$, $s = 1, \ldots, d_2$, *there exists* $c_7 > 0$ *such that*

$$\max_{s=1,\ldots,d_2} \frac{\max\left(x_{s,j+1} - x_{s,j}, j = 0, \ldots, N_n\right)}{\min(x_{s,j+1} - x_{s,j}, j = 0, \ldots, N_n)} \leq c_7.$$

The number of interior knots $N_n \asymp n^{(2p+3)^{-1}}$, *where* p *is the degree of the spline and '*$\asymp$*' means both sides have the same order. In particular,* $h \asymp n^{-(2p+3)^{-1}}$.

(C4) *The vector process* $\{\varsigma_t\}_{t=-\infty}^{\infty} = \{(Y_t, \mathbf{X}_t, \mathbf{T}_t)\}_{t=-\infty}^{\infty}$ *is strictly stationary and geometrically strongly mixing, that is, its* α *-mixing coefficient* $\alpha(k) \leq c\rho^k$, *for constants* $c > 0$ *and* $0 < \rho < 1$, *where* $\alpha(k) = \sup_{A\in\sigma(\varsigma_t, t\leq 0), B\in\sigma(\varsigma_t, t\geq k)} \left|P(A)P(B) - P(AB)\right|$.

(C5) *The conditional variance function* $\sigma^2(\mathbf{x}, \mathbf{t})$ *is measurable and bounded.*

Assumptions (C1)$-$(C5) are common in the nonparametric regression literature. Assumption (C1) is the same as Condition 1 on p.693 of Stone (1985), and Assumption (c), p.468 of Huang and Yang (2004). Assumption (C2) (i) is a direct extension of condition (ii), p.531 of Huang and Shen (2004). Assumption (C2) (ii) is a direct extension of condition (v), p.531 of Huang and Shen (2004), and of the moment condition A.2 (c) p.952 of Cai, Fan and Yao (2000). Assumption (C3) is the same as in (6), p.249 of Huang (1998a), and also p.59, Huang (1998b). Assumption (C4) is similar to condition (iv), p.531 of Huang and Shen (2004). Assumption (C5) is the same as one on p.242 of Huang (1998a), and p.465 of Huang and Yang (2004).

In this Appendix, whenever proofs are brief, see Xue and Yang (2005) for details.

A.2. Technical lemmas

For $1 \leq l \leq d_1, 1 \leq s \leq d_2$, let

$$\tilde{\alpha}_{l0} = \hat{c}_{l0} + \sum_{s=1}^{d_2} E\alpha_{ls}^*, \quad \tilde{\alpha}_{ls}(x_s) = \alpha_{ls}^*(x_s) - E\alpha_{ls}^*. \tag{A.1}$$

Then one can rewrite $\hat{m}$ (3.1) as $\hat{m} = \sum_{l=1}^{d_1} \left\{ \tilde{\alpha}_{l0} + \sum_{s=1}^{d_2} \tilde{\alpha}_{ls}(x_s) \right\} t_l$, with $\{\tilde{\alpha}_{ls}(x_s)\}_{1 \leq l \leq d_1} \in \varphi_s^0$. The terms in (A.1) are not directly observable and serve only as intermediaries in the proof of Theorem 1. By observing that, for $1 \leq l \leq d_1$, $1 \leq s \leq d_2$,

$$\hat{\alpha}_{l0} = \tilde{\alpha}_{l0} - \sum_{s=1}^{d_2} E_n\tilde{\alpha}_{ls}, \quad \hat{\alpha}_{ls}(x_s) = \tilde{\alpha}_{ls}(x_s) - E_n\tilde{\alpha}_{ls}, \tag{A.2}$$

the terms $\{\tilde{\alpha}_{l0}\}_{l=1}^{d_1}$, $\{\tilde{\alpha}_{ls}(x_s)\}_{l=1,s=1}^{d_1,d_2}$ and $\{\hat{\alpha}_{l0}\}_{l=1}^{d_1}$, $\{\hat{\alpha}_{ls}(x_s)\}_{l=1,s=1}^{d_1,d_2}$ differ only by a constant. In section A.3, we first prove the consistency of $\{\tilde{\alpha}_{l0}\}_{l=1}^{d_1}$, and $\{\tilde{\alpha}_{ls}(x_s)\}_{l=1,s=1}^{d_1,d_2}$ in Theorem 3. Then Theorem 1 follows by showing $\{E_n\tilde{\alpha}_{ls}\}_{s=1,l=1}^{d_1,d_2}$ are negligible.

A B-spline basis is used in the proofs. This is equivalent to the truncated power basis used in implementation, but has nice local properties (de Boor (2001)). With $J_n = N_n + p$, we denote the B-spline basis of φ_s by $\mathbf{b}_s = \{b_{s,0}, \ldots, b_{s,J_n}\}$. For $1 \leq s \leq d_2$, let $\mathbf{B}_s = \{B_{s,1}, \ldots, B_{s,J_n}\}$ with

$$B_{s,j} = \sqrt{N_n} \left(b_{s,j} - \frac{E\left(b_{s,j}\right)}{E\left(b_{s,0}\right)} b_{s,0} \right), j = 1, \ldots, J_n. \tag{A.3}$$

Note that (C1) ensures that all $B_{s,j}$'s are well defined.

Now, let $\mathbf{B} = (1, B_{1,1}, \ldots, B_{1,J_n}, \ldots, B_{d_2,1}, \ldots, B_{d_2,J_n})^T$ with $\mathbf{1}$ being the identity function defined on χ. Define $\mathbf{G} = (\mathbf{B}t_1, \ldots, \mathbf{B}t_{d_1})^T = (G_1, \ldots, G_{R_n})^T$, with $R_n = d_1(d_2J_n + 1)$. Then $\mathbf{G}$ is a set of basis of $\mathcal{M}_n$. By (3.1), one has $\hat{m}(\mathbf{x}, \mathbf{t}) = \sum_{l=1}^{d_1}\{\hat{c}_{l0}^* + \sum_{s=1}^{d_2}\sum_{j=1}^{J_n} \hat{c}_{ls,j}^* B_{s,j}(x_s)\} t_l$, in which $\{\hat{c}_{l0}^*, \hat{c}_{ls,j}^*, 1 \leq l \leq d_1, 1 \leq s \leq d_2, 1 \leq j \leq J_n\}$ minimize the sum of squares as in (3.3), with $w_{s,j}$ replaced by $B_{s,j}$. Then Lemma 1 leads to $\tilde{\alpha}_{l0} = \hat{c}_{l0}^*$, $\tilde{\alpha}_{ls} = \sum_{j=1}^{J_n} \hat{c}_{ls,j}^* B_{s,j}(x_s)$, $1 \leq l \leq d_1$, $1 \leq s \leq d_2$.

Theorem 3. *Under* (C1)–(C5), *if* $\alpha_{ls} \in C^{p+1}([0,1])$ *for* $1 \leq l \leq d_1$ *and* $1 \leq s \leq d_2$, *one has*

$$\|\hat{m} - m\|_2 = O_p\left(h^{p+1} + (nh)^{-\frac{1}{2}}\right),$$

$$\max_{1\le l\le d_1}|\tilde{\alpha}_{l0}-\alpha_{l0}|+\max_{1\le l\le d_1,1\le s\le d_2}\|\tilde{\alpha}_{ls}-\alpha_{ls}\|_2=O_p\left(h^{p+1}+(nh)^{-\frac{1}{2}}\right).$$

To prove Theorem 3, we first present the properties of the basis $\mathbf{G}$ in Lemmas A.1–A.3.

Lemma A.1. *For any $1\le s\le d_2$, and the spline basis $B_{s,j}$ of* (A.3), *one has*

(i) $E\left(B_{s,j}\right)=0$, $E\left|B_{s,j}\right|^k\asymp N_n^{k/2-1}$, for $k>1, j=1,\ldots,J_n$.

(ii) *There exists a constant $C>0$ such that for any vector $\mathbf{a}=(a_1,\ldots,a_{J_n})^T$,* as $n\to\infty$, $\|\sum_{j=1}^{J_n}a_jB_{s,j}\|_2^2\ge C\sum_{j=1}^{J_n}a_j^2$.

Proof. (i) follows from Theorem 5.4.2 of Devore and Lorentz (1993), (C1) and (C3). To prove (ii), we introduce the auxiliary knots $x_{s,-p}=\cdots=x_{s,-1}=x_{s,0}=0$, and $x_{s,N_n+p+1}=\cdots=x_{s,N_n+2}=x_{s,N_n+1}=1$ for the knots k_s. Then

$$\begin{aligned}\Big\|\sum_{j=1}^{J_n}a_jB_{s,j}\Big\|_2^2&\ge c_1\Big\|\sum_{j=1}^{J_n}a_jB_{s,j}\Big\|_2^{*2}\\&=c_1\Big\|\sum_{j=1}^{J_n}a_j\sqrt{N_n}b_{s,j}-\sum_{j=1}^{J_n}\frac{a_j\sqrt{N_n}E\left(b_{s,j}\right)}{E\left(b_{s,0}\right)}b_{s,0}\Big\|_2^{*2},\end{aligned}$$

where $\|\cdot\|_2^*$ is defined as $\|f\|_2^*=\sqrt{\int f^2(x)\,dx}$, for any square integrable function f. Let $d_{s,j}=\left(x_{s,j+1}-x_{s,j-p}\right)/(p+1)$. Then Theorem 5.4.2 of Devore and Lorentz (1993) ensures that for a $C>0$, the last term is at least

$$\begin{aligned}c_1C\Big[\sum_{j=1}^{J_n}a_j^2N_nd_{s,j}+\Big\{\sum_{j=1}^{J_n}\frac{a_j\sqrt{N_n}E\left(b_{s,j}\right)}{E\left(b_{s,0}\right)}\Big\}^2d_{s,0}\Big]&\ge c_1C\sum_{j=1}^{J_n}a_j^2N_nd_{s,j}\\&\ge\frac{c_1C\left(p+1\right)}{c_7\sum_{j=1}^{J_n}a_j^2}.\end{aligned}$$

Lemma A.2. *There exists a constant $C>0$, such that as $n\to\infty$, for any sets of coefficients $\{c_{l0},c_{ls,j},l=1,\ldots,d_1;s=1,\ldots,d_2;j=1,\ldots,J_n\}$,*

$$\Big\|\sum_{l=1}^{d_1}\Big(c_{l0}+\sum_{s=1}^{d_2}\sum_{j=1}^{J_n}c_{ls,j}B_{s,j}\Big)t_l\Big\|_2^2\ge C\sum_{l=1}^{d_1}\Big(c_{l0}^2+\sum_{s=1}^{d_2}\sum_{j=1}^{J_n}c_{ls,j}^2\Big).$$

Proof. The result follows immediately from Lemmas 1 and A.1.

Lemma A.3. *Let $\langle\mathbf{G},\mathbf{G}\rangle$ be the $R_n\times R_n$ matrix given by $\langle\mathbf{G},\mathbf{G}\rangle=(\langle G_i,G_j\rangle)_{i,j=1}^{R_n}$. Take $\langle\mathbf{G},\mathbf{G}\rangle_n$ as $\langle\mathbf{G},\mathbf{G}\rangle$, but replace the theoretical inner product with the empirical inner product. Let $\mathbf{D}=\mathrm{diag}(\langle\mathbf{G},\mathbf{G}\rangle)$. Define $Q_n=\sup|\mathbf{D}^{-1/2}(\langle\mathbf{G},\mathbf{G}\rangle_n-$*

$\langle \mathbf{G},\mathbf{G}\rangle)\mathbf{D}^{-1/2}|$, *where the* sup *is taken oven all the elements in the random matrix. Then as $n\to\infty$, $Q_n = O_p(\sqrt{n^{-1}h^{-1}\log^2(n)})$.*

Proof. For simplicity, we consider the diagonal terms. For any $1\le l\le d_2, 1\le s\le d_1$, $1\le j\le J_n$ fixed, let $\xi=(E_n-E)\left\{B_{s,j}^2\left(X_s\right)T_l^2\right\}=(1/n)\sum_{i=1}^n\xi_i$, in which $\xi_i=B_{s,j}^2\left(X_{is}\right)T_{il}^2-E\left\{B_{s,j}^2\left(X_{is}\right)T_{il}^2\right\}$. Define $\tilde{T}_{il}=T_{il}I_{\left\{|T_{il}|\le n^\delta\right\}}$, for some $0<\delta<1$, and define $\tilde{\xi}$, $\tilde{\xi}_i$ similarly as ξ and ξ_i, but replace T_l with $\tilde{T}_{l'}$. Then for any $\epsilon>0$, one has

$$P\Big(\,|\xi|\ge\epsilon\sqrt{\frac{\log^2(n)}{nh}}\Big)\le P\Big(\left|\tilde{\xi}\right|\ge\epsilon\sqrt{\frac{\log^2(n)}{nh}}\Big)+P(\xi\ne\tilde{\xi}),\tag{A.4}$$

in which $P(\xi\ne\tilde{\xi})\le P(T_{il}\ne\tilde{T}_{il}$, for some $i=1,\ldots,n)\le\sum_{i=1}^n P\left(|T_{il}|\ge n^\delta\right)$ $\le E\left|T_l\right|^m/n^{m\delta-1}$. Also note that $\sup_{0\le x_s\le 1}\left|B_{s,j}\left(x_s\right)\right|=\sup_{0\le x_s\le 1}|\sqrt{N_n}\{b_{s,j}-E(b_{s,j})b_{s,0}/E(b_{s,0})\}|\le c\sqrt{N_n}$, for some $c>0$. Then by Minkowski's inequality, for any positive integer $k\ge 3$,

$$\begin{aligned}E\left|\tilde{\xi}_i\right|^k&\le 2^{k-1}\left[E\left|B_{s,j}^2\left(X_s\right)\tilde{T}_l^2\right|^k+\left\{E\left|B_{s,j}^2\left(X_s\right)\tilde{T}_l^2\right|\right\}^k\right]\\&\le 2^{k-1}\left[n^{2\delta k}c^kN_n^k+(cN_n)^k\right]\le n^{2\delta k}c^kN_n^k.\end{aligned}$$

On the other hand

$$E\left|\tilde{\xi}_i\right|^2\ge\frac{E\left|B_{s,j}^2\left(X_s\right)T_l^2\right|^2}{2}-E^2\left\{B_{s,j}^2\left(X_s\right)T_l^2\right\}-\frac{E\left|B_{s,j}^2\left(X_s\right)T_l^2I_{\left\{|T_l|>n^\delta\right\}}\right|^2}{2},$$

in which, under (C2),

$$\begin{aligned}E\left|B_{s,j}^2\left(X_s\right)T_l^2I_{\left\{|T_l|>n^\delta\right\}}\right|^2&\le E\left|B_{s,j}^4\left(X_s\right)E\left(\frac{T_l^{4+\delta_0}}{n^{\delta\delta_0}}|\mathbf{X}\right)\right|\\&\le c_6\frac{E\left|B_{s,j}^4\left(X_s\right)\right|}{n^{\delta\delta_0}}\le\frac{cN_n}{n^{\delta\delta_0}},\end{aligned}$$

where δ_0 is as in (C2). Furthermore

$$\begin{aligned}&E^2\left\{B_{s,j}^2\left(X_s\right)T_l^2\right\}\le c_4^2E^2\left\{B_{s,j}^2\left(X_s\right)\right\}\le c,\\&E\left|B_{s,j}^2\left(X_s\right)T_l^2\right|^2\ge c_5E\left|B_{s,j}\left(X_s\right)\right|^4\ge c_1c_5\int\left|B_{s,j}\left(x_s\right)\right|^4dx_s\ge cc_1c_5N_n.\end{aligned}$$

Thus $E|\tilde{\xi}_i|^2\ge cN_n-c-(cN_n/n^{\delta\delta_0})\ge cN_n$. So there exists a constant $c>0$, such that for all $k\ge 3$,

$$E\left|\tilde{\xi}_i\right|^k\le n^{2\delta k}c^kN_n^k\le\left(cn^{6\delta}N_n^2\right)^{k-2}k!E\left|\tilde{\xi}_i\right|^2.$$

Then one can apply Theorem 1.4 of Bosq (1998) to $\sum_{i=1}^{n}\tilde{\xi}_i$, with the Cramer constant $c_r = cn^{6\delta}N_n^2$. That is, for any $\epsilon > 0$, $q \in [1, n/2]$, and $k \geq 3$, one has

$$\begin{aligned} &P\left(\frac{1}{n}\left|\sum_{i=1}^{n}\tilde{\xi}_i\right| \geq \epsilon\sqrt{\frac{\log^2(n)}{nh}}\right) \\ &\quad \leq a_1 \exp\left(-\frac{q\epsilon^2\frac{\log^2(n)}{nh}}{25m_2^2 + 5\epsilon c_r\sqrt{\frac{\log^2(n)}{nh}}}\right) + a_2(k)\alpha\left(\left[\frac{n}{q+1}\right]\right)^{\frac{2k}{2k+1}}, \end{aligned}$$

where

$$\begin{aligned} a_1 &= 2\frac{n}{q} + 2\left(1 + \frac{\epsilon^2\frac{\log^2(n)}{nh}}{25m_2^2 + 5\epsilon c_r\sqrt{\frac{\log^2(n)}{nh}}}\right), \\ a_2(k) &= 11n\left(1 + \frac{5m_p^{\frac{k}{2k+1}}}{\epsilon\sqrt{\frac{\log^2(n)}{nh}}}\right), \quad m_2^2 = E\tilde{\xi}_i^2, \quad m_p = \left\|\tilde{\xi}_i\right\|_p. \end{aligned}$$

Observe that $5\epsilon c_r\sqrt{\log^2(n)/(nh)} = 5\epsilon cn^{6\delta}N_n^2\sqrt{\log^2(n)/(nh)} = o(1)$, by taking $\delta < 2p/[12(2p+3)]$. Then, by taking $q = n/\{c_0\log(n)\}$, one has $a_1 = O(n/q) = O\{\log(n)\}$ and $a_2(k) = O(n[N_n^{k/(2k+1)}/\sqrt{\log^2(n)/(nh)}) = o\left(n^{3/2}\right)$. Thus, for n large enough,

$$\begin{aligned} &P\Big(\frac{1}{n}\Big|\sum_{i=1}^{n}\tilde{\xi}_i\Big| \geq \epsilon\sqrt{\frac{\log^2(n)}{nh}}\Big) \\ &\quad \leq c\log(n)\exp\Big\{-\frac{\epsilon^2\log(n)}{50c_0m_2^2}\Big\} + cn^{\frac{3}{2}}\exp\left\{-\log(\rho)c_0\log(n)\right\}. \end{aligned}$$

By (A.4), taking c_0, ϵ, m large enough and using (C4), one has that

$$\begin{aligned} &\sum_{n=1}^{\infty} P\left(\sup\left|\langle\mathbf{G},\mathbf{G}\rangle_n - \langle\mathbf{G},\mathbf{G}\rangle\right| \geq \epsilon\sqrt{\frac{\log^2(n)}{nh}}\right) \\ &\quad \leq \sum_{n=1}^{\infty}\left\{d_1d_2(N_n+2)\right\}^2\Big\{c\log(n)\exp\Big\{-\frac{\epsilon^2\log(n)}{25c_0}\Big\} \\ &\qquad\qquad + cn^{\frac{3}{2}}\exp\left\{-\log(\rho)c_0\log(n)\right\} + \frac{E\left|T_l\right|^m}{n^{m\delta-1}}\Big\} \\ &\quad < \sum_{n=1}^{\infty}\left\{d_1d_2(N_n+2)\right\}^2 n^{-3} < +\infty, \end{aligned}$$

in which $N_n \asymp n^{(2p+3)^{-1}}$. Then the lemma follows from Borel-Cantelli Lemma and Lemma A.1.

Lemma A.4. *As $n \to \infty$, one has*

$$\sup_{\phi_1 \in \mathcal{M}_n, \phi_2 \in \mathcal{M}_n} \left| \frac{\langle \phi_1, \phi_2 \rangle_n - \langle \phi_1, \phi_2 \rangle}{\|\phi_1\|_2 \|\phi_2\|_2} \right| = O_p\Big(\sqrt{\frac{\log^2(n)}{nh}}\Big).$$

In particular, there exist constants $0 < c < 1 < C$ such that, except on an event whose probability tends to zero as $n \to \infty$, $c\|m\|_2 \le \|m\|_{2,n} \le C\|m\|_2, \forall m \in \mathcal{M}_n$.

Proof. With vector notation, one can write $\phi_1 = \mathbf{a}_1^T \mathbf{G}$, $\phi_2 = \mathbf{a}_2^T \mathbf{G}$, for $R_n \times 1$ vectors $\mathbf{a}_1, \mathbf{a}_2$. Then

$$|\langle \phi_1, \phi_2 \rangle_n - \langle \phi_1, \phi_2 \rangle| = \sum_{i,j=1}^{R_n} |a_{1i} a_{2j}| \left| \langle G_i, G_j \rangle_n - \langle G_i, G_j \rangle \right|$$

$$\le Q_n \sum_{i,j=1}^{R_n} |a_{1i} a_{2j}| \|G_i\|_2 \|G_j\|_2 \le Q_n C \sum_{i,j=1}^{R_n} |a_{1i} a_{2j}| \le Q_n C \sqrt{\mathbf{a}_1^T \mathbf{a}_1 \mathbf{a}_2^T \mathbf{a}_2}.$$

On the other hand by Lemma A.2, $\|\phi_1\|_2^2 \|\phi_2\|_2^2 = \left(\mathbf{a}_1^T \langle \mathbf{G}, \mathbf{G} \rangle \mathbf{a}_1\right) \left(\mathbf{a}_2^T \langle \mathbf{G}, \mathbf{G} \rangle \mathbf{a}_2\right) \ge C^2 \mathbf{a}_1^T \mathbf{a}_1 \mathbf{a}_2^T \mathbf{a}_2$. Then

$$\left| \frac{\langle \phi_1, \phi_2 \rangle_n - \langle \phi_1, \phi_2 \rangle}{\|\phi_1\|_2 \|\phi_2\|_2} \right| \le \left| \frac{Q_n C \sqrt{\mathbf{a}_1^T \mathbf{a}} \sqrt{\mathbf{a}_2^T \mathbf{a}_2}}{C \sqrt{\mathbf{a}_1^T \mathbf{a}_1} \sqrt{\mathbf{a}_2^T \mathbf{a}_2}} \right| = O_p(Q_n) = O_p\Big(\sqrt{\frac{\log^2(n)}{nh}}\Big).$$

Lemma A.4 shows that the empirical and theoretical inner products are uniformly close over the approximation space $\mathcal{M}_n$. This lemma plays a role analogous to that of Lemma 10 in Huang (1998a). Our result is new, in that (i) the spline basis of Huang (1998a) must be bounded, whereas the term t in basis $\mathbf{G}$ makes it possibly bounded; (ii) Huang (1998a)'s setting is i.i.d. with uniform approximation rate of $o_p(1)$, while our setting is α-mixing, broadly applicable to time series data, with approximation rate the sharper $O_p(\sqrt{\log^2(n)/nh})$. The next lemma follows immediately from Lemmas A.2 and A.4.

Lemma A.5. *There exists constant $C > 0$ such that, except on an event whose probability tends to zero, as $n \to \infty$,*

$$\left\| \sum_{l=1}^{d_1} \Big(c_{l0} + \sum_{s=1}^{d_2} \sum_{j=1}^{J_n} c_{ls,j} B_{s,j}\Big) t_l \right\|_{2,n}^2 \ge C \sum_{l=1}^{d_1} \Big(c_{l0}^2 + \sum_{s=1}^{d_2} \sum_{j=1}^{J_n} c_{ls,j}^2\Big).$$

A.3. Proof of mean square consistency

Proof of Theorem 3. Let

$$\mathbf{Y} = (Y_1, \ldots, Y_n)^T, \qquad \mathbf{m} = \{m(\mathbf{X}_1, \mathbf{T}_1), \ldots, m(\mathbf{X}_n, \mathbf{T}_n)\}^T,$$
$$\mathbf{E} = \{\sigma(\mathbf{X}_1, \mathbf{T}_1)\varepsilon_1, \ldots, \sigma(\mathbf{X}_n, \mathbf{T}_n)\varepsilon_n\}^T.$$

Note that $\mathbf{Y} = \mathbf{m} + \mathbf{E}$ and, projecting it onto $\mathcal{M}_n$, one has $\hat{m} = \overline{m} + \overline{e}$, where $\hat{m}$ is defined in (3.1), and $\overline{m}$, $\overline{e}$ are the solution to (3.1) with Y_i replaced by $m(\mathbf{X}_i, \mathbf{T}_i)$ and $\sigma(\mathbf{X}_i, \mathbf{T}_i)\varepsilon_i$ respectively. Also one can uniquely represent $\overline{m}$ as $\overline{m} = \sum_{l=1}^{d_1} \left(\overline{\alpha}_{l0} + \sum_{s=1}^{d_2} \overline{\alpha}_{ls}\right) t_l, \overline{\alpha}_{ls} \in \varphi_s^0$. With these notations, one has the error decomposition $\hat{m} - m = \overline{m} - m + \overline{e}$, where $\overline{m} - m$ is the bias term, and $\overline{e}$ is the variance term. Since for $1 \le l \le d_1, 1 \le s \le d_2, \alpha_{ls} \in C^{p+1}([0,1])$, by Lemma 2 there exist $C > 0$ and spline functions $g_{ls} \in \varphi_s^0$, such that $\|\alpha_{ls} - g_{ls}\|_\infty \le Ch^{p+1}$. Let $m_n(\mathbf{x}, \mathbf{t}) = \sum_{l=1}^{d_1} \left\{\alpha_{l0} + \sum_{s=1}^{d_2} g_{ls}(x_s)\right\} t_l \in \mathcal{M}_n$. One has

$$\|m - m_n\|_2 \le \sum_{l=1}^{d_1} \sum_{s=1}^{d_2} \|\{\alpha_{ls} - g_{ls}\} t_l\|_2 \le c_4 \sum_{l=1}^{d_1} \sum_{s=1}^{d_2} \|\alpha_{ls} - g_{ls}\|_\infty \le c_4 C h^{p+1}. \tag{A.5}$$

Also $\|m - m_n\|_{2,n} \le Ch^{p+1}$ a.s. Then by the projection property, one has $\|m - \overline{m}\|_{2,n} \le \|m - m_n\|_{2,n} \le Ch^{p+1}$, which also implies $\|\overline{m} - m_n\|_{2,n} \le \|m - \overline{m}\|_{2,n} + \|m - m_n\|_{2,n} \le Ch^{p+1}$. By Lemma A.4 $\|\overline{m} - m_n\|_2 \le \|\overline{m} - m_n\|_{2,n} \times (1 - Q_n)^{1/2} = O_p\left(h^{p+1}\right)$. Together with (A.5), one has

$$\|m - \overline{m}\|_2 = O_p\left(h^{p+1}\right). \tag{A.6}$$

Next we consider the variance] term $\overline{e}$, which is written as $\overline{e}(\mathbf{x}, \mathbf{t}) = \sum_{j=1}^{R_n} \hat{a}_j G_j(\mathbf{x}, \mathbf{t})$, with $\widehat{\mathbf{a}} = (\hat{a}_1, \ldots, \hat{a}_{R_n})^T$. Let $\mathbf{N} = [N(G_1), \ldots, N(G_{R_n})]^T$, with $N(G_j) = (1/n) \sum_{i=1}^n G_j(\mathbf{X}_i, \mathbf{T}_i) \sigma(\mathbf{X}_i, \mathbf{T}_i) \varepsilon_i$. By the projection property, one has $\left(\langle G_j, G_j \rangle_n\right)_{j,j'=1}^{R_n} \widehat{\mathbf{a}} = \mathbf{N}$. Multiplying both sides by the same vector, one gets $\widehat{\mathbf{a}}^T \left(\langle G_j, G_j \rangle_n\right)_{j,j'=1}^{R_n} \widehat{\mathbf{a}} = \widehat{\mathbf{a}}^T \mathbf{N}$. Now, by Lemmas A.2 and A.4, the LHS is $\left\|\sum_{j=1}^{R_n} \hat{a}_j G_j\right\|_{2,n}^2 \ge C(1 - Q_n) \sum_{j=1}^{R_n} \hat{a}_j^2$, while the RHS is

$$\le \left(\sum_{j=1}^{R_n} \hat{a}_j^2\right)^{\frac{1}{2}} \left\{\sum_{j=1}^{R_n} \left(\frac{1}{n} \sum_{i=1}^n G_j(\mathbf{X}_i, \mathbf{T}_i) \sigma(\mathbf{X}_i, \mathbf{T}_i) \varepsilon_i\right)^2\right\}^{\frac{1}{2}}.$$

Hence $C(1 - Q_n) \sum_{j=1}^{R_n} \hat{a}_j^2 \le \left(\sum_{j=1}^{R_n} \hat{a}_j^2\right)^{1/2} \left\{\sum_{j=1}^{R_n} \left((1/n) \sum_{i=1}^n G_j(\mathbf{X}_i, \mathbf{T}_i) \times \sigma(\mathbf{X}_i, \mathbf{T}_i)\varepsilon_i\right)^2\right\}^{1/2}$, entailing

$$\left(\sum_{j=1}^{R_n} \hat{a}_j^2\right)^{\frac{1}{2}} \le C^{-1} (1 - Q_n)^{-1} \left\{\sum_{j=1}^{R_n} \left(\frac{1}{n} \sum_{i=1}^n G_j(\mathbf{X}_i, \mathbf{T}_i) \sigma(\mathbf{X}_i, \mathbf{T}_i) \varepsilon_i\right)^2\right\}^{\frac{1}{2}}.$$

As a result, $\|\overline{e}\|_2^2 \leq C\left(1-Q_n\right)^{-2}\left\{\sum_{j=1}^{R_n}\left((1/n)\sum_{i=1}^{n} G_j\left(\mathbf{X}_i,\mathbf{T}_i\right)\sigma\left(\mathbf{X}_i,\mathbf{T}_i\right)\varepsilon_i\right)^2\right\}$. Since the ε_i are independent of $\left\{\left(\mathbf{X}_j,\mathbf{T}_j\right), j\leq i\right\}$ for $i=1,\ldots,n$, one has

$$\begin{aligned}\sum_{j=1}^{R_n} E\Big(\frac{1}{n}\sum_{i=1}^{n} G_j\left(\mathbf{X}_i,\mathbf{T}_i\right)\sigma\left(\mathbf{X}_i,\mathbf{T}_i\right)\varepsilon_i\Big)^2 &= \sum_{j=1}^{R_n}\frac{1}{n}E\left\{G_j\left(\mathbf{X}_i,\mathbf{T}_i\right)\sigma\left(\mathbf{X}_i,\mathbf{T}_i\right)\varepsilon_i\right\}^2\\ &\leq \frac{CJ_n}{n} = O\left(\frac{1}{nh}\right),\end{aligned}$$

by (C2), (C5) and Lemma A.1(i). Therefore $\|\tilde{e}\|_2^2 = O_p\left(n^{-1}h^{-1}\right)$. This, together with (A.6), prove $\|\hat{m}-m\|_2 = O_p\left(h^{p+1}+\sqrt{1/nh}\right)$. The second part of Theorem 3 follows from Lemma 1, which entails that for $C>0$, $\|\hat{m}-m\|_2^2 \geq C\left[\sum_{l=1}^{d_1}\left\{\left(\tilde{\alpha}_{l0}-\alpha_{l0}\right)^2+\sum_{s=1}^{d_2}\|\tilde{\alpha}_{ls}-\alpha_{ls}\|_2^2\right\}\right]$.

Proof of Theorem 1. By (A.2), one only needs to show $|E_n\tilde{\alpha}_{ls}| = O_p(h^{p+1} +\sqrt{1/nh})$, for $1\leq l\leq d_1, 1\leq s\leq d_2$. Note that $|E_n\tilde{\alpha}_{ls}| \leq |E_n\left\{\tilde{\alpha}_{ls}-\alpha_{ls}\right\}| + |E_n\alpha_{ls}|$, where

$$\begin{aligned}|E_n\left\{\tilde{\alpha}_{ls}-\alpha_{ls}\right\}| &\leq \|\tilde{\alpha}_{ls}-\alpha_{ls}\|_{2,n} \leq \|\alpha_{ls}-\overline{\alpha}_{ls}\|_{2,n} + \|\tilde{\alpha}_{ls}-\overline{\alpha}_{ls}\|_{2,n}\\ &\leq \|\alpha_{ls}-g_{ls}\|_{2,n} + \|\overline{\alpha}_{ls}-g_{ls}\|_{2,n} + \|\tilde{\alpha}_{ls}-\overline{\alpha}_{ls}\|_{2,n},\end{aligned}$$

with $\|\alpha_{ls}-g_{ls}\|_{2,n} \leq \|\alpha_{ls}-g_{ls}\|_\infty \leq Ch^{p+1}$. Applying Lemmas 1 and A.3, one has

$$\|\overline{\alpha}_{ls}-g_{ls}\|_{2,n} \leq \left(1+Q_n\right)\|\overline{\alpha}_{ls}-g_{ls}\|_2 \leq \left(1+Q_n\right)\|\overline{m}-m_n\|_2 = O_p\left(h^{p+1}\right),$$

$$\|\tilde{\alpha}_{ls}-\overline{\alpha}_{ls}\|_{2,n} \leq \left(1+Q_n\right)\|\tilde{\alpha}_{ls}-\overline{\alpha}_{ls}\|_{2,n} \leq \left(1+Q_n\right)\|\tilde{e}\|_2 = O_p\left(\sqrt{\frac{1}{nh}}\right).$$

Thus $|E_n\left\{\tilde{\alpha}_{ls}-\alpha_{ls}\right\}| = O_p\left(h^{p+1}+\sqrt{1/nh}\right)$. Since $|E_n\alpha_{ls}| = O_p\left(1/\sqrt{n}\right)$, one now has $|E_n\tilde{\alpha}_{ls}| = O_p\left(h^{p+1}+(nh)^{-\frac{1}{2}}\right)$. Theorem 1 follows from the triangle inequality.

A.4. Proof of BIC consistency

We denote the model space $\mathcal{M}_S$ and the approximation space $\mathcal{M}_{n,S}$ of m_S separately as

$$\mathcal{M}_S = \left\{m\left(\mathbf{x},\mathbf{t}\right) = \sum_{l=1}^{d_1}\alpha_l\left(\mathbf{x}\right)t_l;\quad \alpha_l\left(\mathbf{x}\right) = \alpha_{l0}+\sum_{s\in S_l}\alpha_{ls}(x_s); \alpha_{ls}\in\mathcal{H}_s^0\right\},$$

$$\mathcal{M}_{n,S} = \left\{m_n\left(\mathbf{x},\mathbf{t}\right) = \sum_{l=1}^{d_1}g_l\left(\mathbf{x}\right)t_l;\quad g_l\left(\mathbf{x}\right) = \alpha_{l0}+\sum_{s\in S_l}g_{ls}(x_s); g_{ls}\in\varphi_s^0\right\}.$$

If $S \subset S_f$, $\mathcal{M}_S \subset \mathcal{M}_{S_f}$ and $\mathcal{M}_{n,S} \subset \mathcal{M}_{n,S_f}$. Let Proj_S ($\mathrm{Proj}_{n,S}$) be the orthogonal least square projector onto $\mathcal{M}_S$ ($\mathcal{M}_{n,S}$) with respect to the empirical inner product. Then $\hat{m}_S$ at (3.7) can be viewed as: $\hat{m}_S = \mathrm{Proj}_{n,S}(\mathbf{Y})$. As a special case of Theorem 1, one has the following result.

Lemma A.6. *Under the conditions of Theorem* 1, $\|\hat{m}_S - m_S\|_2 = O_p(1/N_S^{p+1} + \sqrt{N_S/n})$.

Now let $c(S, m) = \|\mathrm{Proj}_S m - m\|_2$. One has that if $m \in \mathcal{M}_{S_0}$, $\mathrm{Proj}_{S_0} m = m$, and $c(S_0, m) = 0$; if S overfits, since $m \in \mathcal{M}_{S_0} \subset \mathcal{M}_S$, $c(S, m) = 0$; if S underfits, $c(S, m) > 0$.

Proof of Theorem 2. Notice that

$$\begin{aligned}\mathrm{BIC}_S - \mathrm{BIC}_{S_0} &= \frac{\mathrm{MSE}_S - \mathrm{MSE}_{S_0}}{\mathrm{MSE}_{S_0}}\{1 + o_p(1)\} + \frac{q_S - q_{S_0}}{n}\log(n)\\ &= \frac{\mathrm{MSE}_S - \mathrm{MSE}_{S_0}}{E\{\sigma^2(\mathbf{X}, \mathbf{T})\}(1 + o_p(1))}\{1 + o_p(1)\} + n^{-\frac{2p+2}{2p+3}}\log(n),\end{aligned}$$

since $q_S - q_{S_0} \asymp n^{1/(2p+3)}$, and

$$\begin{aligned}\mathrm{MSE}_{s_0} &\le \frac{1}{n}\sum_{i=1}^n \{Y_i - m(\mathbf{X}_i, \mathbf{T}_i)\}^2 + \frac{1}{n}\sum_{i=1}^n \{\hat{m}_{S_0}(\mathbf{X}_i, \mathbf{T}_i) - m(\mathbf{X}_i, \mathbf{T}_i)\}^2\\ &= E\{\sigma^2(\mathbf{X}, \mathbf{T})\}(1 + o_p(1)).\end{aligned}$$

Case 1 (Overfitting): Suppose that $S_0 \subset S$ and $S_0 \neq S$. One has

$$\begin{aligned}&\mathrm{MSE}_s - \mathrm{MSE}_{s_0} = \|\widehat{m}_s - \widehat{m}_{s,0}\|_{2,n}^2 = \|\widehat{m}_s - \widehat{m}_{s,0}\|_2^2\{1 + o_p(1)\}\\ &\quad\le \left(\|\widehat{m}_s - m\|_2^2 + \|\widehat{m}_{s,0} - m\|_2^2\right)\{1 + o_p(1)\} = O_p\left(n^{-\frac{2p+2}{2p+3}}\right).\end{aligned}$$

Thus $\lim_{n\to+\infty}\{P(\mathrm{BIC}_s - \mathrm{BIC}_{s_0} > 0)\} = 1$. To see why the assumption $q_S - q_{S_0} \asymp n^{1/(2p+3)}$ is necessary, suppose $q_{S_0} \asymp n^r$, with $r > 1/(2p+3)$ instead. Then it can be shown that

$$\mathrm{MSE}_s - \mathrm{MSE}_{s_0} = -\frac{n^{r-1}}{E\{\sigma^2(\mathbf{X}, \mathbf{T})\}\{1 + o_p(1)\}} - n^{r-1}\log(n)\{1 + o_p(1)\},$$

which leads to $\lim_{n\to+\infty}\{P(\mathrm{BIC}_s - \mathrm{BIC}_{s_0} < 0)\} = 1$.

Case 2 (Underfitting): Similarly as in Huang and Yang (2004), we can show that if S underfits, $\mathrm{MSE}_S - \mathrm{MSE}_{S_0} \ge c^2(S, m) + o_p(1)$. Then

$$\mathrm{BIC}_S - \mathrm{BIC}_{S_0} \ge \frac{c^2(S, m) + o_p(1)}{E\{\sigma^2(\mathbf{X}, \mathbf{T})\}(1 + o_p(1))} + o_p(1),$$

which implies that $\lim_{n\to+\infty}\{P(\mathrm{BIC}_s - \mathrm{BIC}_{s_0} > 0)\} = 1$.

Acknowledgements

The paper is substantially improved thanks to the helpful comments of Editor Jane-Ling Wang, an associate editor and the referees. Research has been supported in part by National Science Foundation grants BCS 0308420, DMS 0405330, and SES 0127722.

References

Bosq, D. (1998). *Nonparametric Statistics for Stochastic Processes: Estimation and Prediction.* 2nd edition. Springer-Verlag, New York.

Cai, Z., Fan, J. and Yao, Q. W. (2000). Functional-coefficient regression models for nonlinear time series. *J. Amer. Statist. Assoc.* **95**, 941-956.

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

Chen, R. and Tsay, R. S. (1993b). Functional-coefficient autoregressive models. *J. Amer. Statist. Assoc.* **88**, 298-308.

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

Devore, R. A. and Lorentz, G. G. (1993). *Constructive Approximation.* Springer-Verlag, Berlin Heidelberg.

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

Hastie, T. J. and Tibshirani, R. J. (1993). Varying-coefficient models. *J. Roy. Statist. Soc. Ser. B* **55**, 757-796.

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

Huang, J. Z. (1998b). Functional ANOVA models for generalized regression. *J. Multivariate Anal.* **67**, 49-71.

Huang, J. Z. and Shen, H. (2004). Functional coefficient regression models for non-linear time series: a polynomial spline approach. *Scand. J. Statist.* **31**, 515-534.

Huang, J. Z., Wu, C. O. and Zhou, L. (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika* **89**, 111-128.

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

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

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

Xue, L. and Yang, L. (2005). Additive coefficient modeling via polynomial spline, downloadable at http://www.msu.edu/~yangli/addsplinefull.pdf.

Department of Statistics, Oregon State University, Corvallis, OR 97331, USA.

E-mail: xuel@stat.oregonstate.edu

Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824, USA.

E-mail: yang@stt.msu.edu

(Received September 2004; accepted August 2005)