
Volume 141
Issue 1
January 2011
ISSN 0378-3758
ELSEVIER
journal of statistical planning and inference
ScienceDirect

Journal of Statistical Planning and Inference 141 (2011) 204–219

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

# Spline-backfitted kernel smoothing of partially linear additive model

Shujie Ma, Lijian Yang *

*Department of Statistics and Probability, Michigan State University, USA*

ARTICLE INFO

*Article history:*
Received 10 August 2009
Received in revised form
31 May 2010
Accepted 31 May 2010
Available online 8 June 2010

*Keywords:*
Bandwidths
B spline
Knots
Local linear estimator
Mixing
Nadaraya–Watson estimator
Nonparametric regression

ABSTRACT

A spline-backfitted kernel smoothing method is proposed for partially linear additive model. Under assumptions of stationarity and geometric mixing, the proposed function and parameter estimators are oracally efficient and fast to compute. Such superior properties are achieved by applying to the data spline smoothing and kernel smoothing consecutively. Simulation experiments with both moderate and large number of variables confirm the asymptotic results. Application to the Boston housing data serves as a practical illustration of the method.



## 1. Introduction

Since the 1980s, non- and semiparametric analysis of time series has been vigorously pursued, see, for example, Tjøstheim and Auestad (1994) and Huang and Yang (2004). There are few satisfactory smoothing tools for multi-dimensional time series data, however, due to the poor convergence rate of nonparametric estimation of multivariate functions, known as the "curse of dimensionality". One solution is the partially linear additive model (PLAM) studied in Li (2000), Fan and Li (2003) and Liang et al. (2008)

$$Y_i = m(\mathbf{X}_i,\mathbf{T}_i)+\sigma(\mathbf{X}_i,\mathbf{T}_i)\varepsilon_i, \quad m(\mathbf{x},\mathbf{t}) = c_{00}+\sum_{l=1}^{d_1} c_{0l}t_l+\sum_{\alpha=1}^{d_2} m_\alpha(x_\alpha) \tag{1}$$

in which the sequence $\{Y_i,\mathbf{X}_i^T,\mathbf{T}_i^T\}_{i=1}^n = \{Y_i,X_{i1},\ldots,X_{id_2},T_{i1},\ldots T_{id_1}\}_{i=1}^n$. The functions $m$ and $\sigma$ are the mean and standard deviation of the response $Y_i$ conditional on the predictor vector $\{\mathbf{X}_i,\mathbf{T}_i\}$, and $\varepsilon_i$ is a white noise conditional on $\{\mathbf{X}_i,\mathbf{T}_i\}$. For identifiability, both additive and linear components must be centered, i.e., $Em_\alpha(X_{i\alpha})\equiv 0,\ 1\le\alpha\le d_2,\ ET_{il}=0,\ 1\le l\le d_1$.

If parameters $c_{0l}\equiv 0,\ 1\le l\le d_1,\ (T_{i1},\ldots T_{id_1})$ are redundant, and $\{Y_i,X_{i1},\ldots,X_{id_2}\}_{i=1}^n$ follow an additive model. For applications of additive model, see, Nácher et al. (2006), Roca-Pardiñas et al. (2006) and González-Manteiga et al. (2008). Additive model, however, is only appropriate to model nonparametric effects of continuous predictors $(X_{i1},\ldots,X_{id_2})$ supported on compact intervals. The effects of possibly discrete and/or unbounded predictors can be neatly modeled as some of the variables $(T_{i1},\ldots T_{id_1})$ in the PLAM (1), see the simulation example in Section 3 where $T_{i1}$, $T_{i2}$ are normal conditional on $\mathbf{X}_i$, $T_{i3}$ is discrete and $T_{i4}$ has positive density over a compact interval, and Section 4 which shows that the simpler PLAM fits the Boston housing data much better than an additive model. For general references on partially linear

* Corresponding author.
*E-mail address:* yang@stt.msu.edu (L. Yang).

0378-3758/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2010.05.028

model, see Schimek (2000) and Liang (2006). For applications of partially linear model to panel data, see Su and Ullah (2006), while for data with measurement errors, see Liang et al. (2007, 2008).

Satisfactory estimators of functions $\{m_\alpha(x_\alpha)\}_{\alpha=1}^{d_2}$ and constants $\{c_{0l}\}_{l=0}^{d_1}$ in model (1) based on $\{Y_i,\mathbf{X}_i^T,\mathbf{T}_i^T\}_{i=1}^n$ should be (i) computationally expedient; (ii) theoretically reliable and (iii) intuitively appealing. Kernel procedures for PLAM, such as Fan and Li (2003) and Liang et al. (2008) satisfy criterion (iii) and partly (ii) but not (i) since they are computationally intensive when sample size $n$ is large, as illustrated in the Monte-Carlo results of Xue and Yang (2006) and Wang and Yang (2007). It is mentioned in Li (2000) that the computation time of estimating a PLAM is about $n$ times of estimating a partially linear model with $d_2$=1 by using the kernel marginal integration method. For discussion of computation burden issues by kernel methods, see Li (2000). Spline approaches of Li (2000), Schimek (2000) to PLAM, do not satisfy criterion (ii) as they lack limiting distribution, but are fast to compute, thus satisfying (i). The SBK estimator we propose combines the best features of both kernel and spline methods, and is essentially as fast and accurate as an univariate kernel smoothing, satisfying all three criteria (i)–(iii).

We propose to extend the "spline-backfitted kernel smoothing" (SBK) of Wang and Yang (2007) to PLAM (1). If the regression coefficients $\{c_{0l}\}_{l=0}^{d_1}$ and the component functions $\{m_\beta(x_\beta)\}_{\beta=1,\beta\neq\alpha}^{d_2}$ were known by "oracle", one could create $\{Y_{i\alpha},X_{i\alpha}\}_{i=1}^n$ with $Y_{i\alpha}=Y_i-c_{00}-\sum_{l=1}^{d_1}c_{0l}T_{il}-\sum_{\beta=1,\beta\neq\alpha}^{d_2}m_\beta(X_{i\beta})=m_\alpha(X_{i\alpha})+\sigma(\mathbf{X}_i,\mathbf{T}_i)\varepsilon_i$, from which one could compute an "oracle smoother" to estimate the only unknown function $m_\alpha(x_\alpha)$, bypassing the "curse of dimensionality". The idea was to obtain approximations to the unobservable variables $Y_{i\alpha}$ by substituting $m_\beta(X_{i\beta})$, $1\le i\le n$, $1\le\beta\le d_2$, $\beta\neq\alpha$, with spline estimates and argue that the error incurred by this "cheating" is of smaller magnitude than the rate $O(n^{-2/5})$ for estimating $m_\alpha(x_\alpha)$ from the unobservable data. Lemmas A.9, A.14, A.17 and A.18 establish the estimators' uniform oracle efficiency by "reducing bias via undersmoothing (step one) and averaging out the variance (step two)", via the joint asymptotics of kernel and spline functions. A major theoretical innovation is to resolve the dependence between $\mathbf{T}$ and $\mathbf{X}$, making use of Assumption (A5), which is not needed in Wang and Yang (2007). Another significant innovation is the $\sqrt{n}$-consistency and asymptotic distribution of estimators for parameters $\{c_{0l}\}_{l=0}^{d_1}$, which is trivial for the additive model of Wang and Yang (2007).

The paper is organized as follows. The SBK estimators are introduced in Section 2 with theoretical properties. Section 3 contains Monte Carlo results to demonstrate the asymptotic properties of SBK estimators. The SBK estimator is applied to the Boston housing data in Section 4. Proofs of technical lemmas are in Appendix A.

## 2. The SBK estimators

For convenience, we denote vectors as $\mathbf{x}=(x_1,\ldots,x_d)^T$ and take $\|\cdot\|$ as the usual Euclidean norm on $R^d$, i.e., $\|\mathbf{x}\|=\sqrt{\sum_{\alpha=1}^d x_\alpha^2}$, and $\|\cdot\|_\infty$ the sup norm, i.e., $\|\mathbf{x}\|_\infty=\sup_{1\le\alpha\le d}|x_\alpha|$. We denote by $\mathbf{I}_r$ the $r\times r$ identity matrix, $\mathbf{0}_{r\times s}$ the zero matrix of dimension $r\times s$, and diag$(a,b)$ the $2\times 2$ diagonal matrix with diagonal entries $a,b$. Let $\{Y_i,\mathbf{X}_i^T,\mathbf{T}_i^T\}_{i=1}^n$ be a sequence of strictly stationary observations from a geometrically $\alpha$-mixing process following model (1), where $Y_i$ and $(\mathbf{X}_i,\mathbf{T}_i)=\{(X_{i1},\ldots,X_{id_2})^T,(T_{i1},\ldots T_{id_1})^T\}$ are the $i$-th response and predictor vector. Denote $\mathbf{Y}=(Y_1,\ldots,Y_n)^T$ the response vector. Without loss of generality, we assume $X_\alpha$ is distributed on [0,1], $1\le\alpha\le d_2$. An integer $N=N_n\sim n^{1/4}\log n$ is pre-selected. Denote the space of the $k$-th order smooth functions as $C^{(k)}[0,1]=\{m|m^{(k)}\in C[0,1]\}$, the class of Lipschitz continuous functions for a constant $C>0$ as $\mathrm{Lip}([0,1],C)=\{m||m(x)-m(x')|\le C|x-x'|,\ \forall x,x'\in[0,1]\}$.

Define the second order B spline (or linear B spline) basis (de Boor, 2001, p. 89) as $b_J(x)=(1-|x-\xi_J|/H)_+$, $0\le J\le N+1$, where $0=\xi_0<\xi_1<\cdots<\xi_N<\xi_{N+1}=1$ are equally spaced points called knots, knots $\xi_1,\ldots,\xi_N$ are called interior knots, $\xi_0=0$, $\xi_{N+1}=1$ the boundary knots, and $\xi_{-1}=0$, $\xi_{N+2}=1$ the degenerate knots. The distance between neighboring interior or boundary knots is $H=H_n=(N+1)^{-1}$.

Define next the space $G$ of partially linear additive spline functions as the linear space spanned by $\{1,t_l,b_J(x_\alpha),\ 1\le l\le d_1,\ 1\le\alpha\le d_2,\ 1\le J\le N+1\}$, and let $\{1,\{T_l,b_J(X_{i\alpha})\}_{i=1}^n,1\le l\le d_1,1\le\alpha\le d_2,1\le J\le N+1\}$ span the space $G_n\subset R^n$. As $n\to\infty$, with probability approaching 1, the dimension of $G_n$ becomes $\{1+d_1+d_2(N+1)\}$. The spline estimator of $m(\mathbf{x},\mathbf{t})$ is the unique element $\hat{m}(\mathbf{x},\mathbf{t})=\hat{m}_n(\mathbf{x},\mathbf{t})$ from $G$ so that $\{\hat{m}(\mathbf{X}_i,\mathbf{T}_i)\}_{1\le i\le n}^T$ best approximates the response vector $\mathbf{Y}$. To be precise, we define

$$\hat{m}(\mathbf{x},\mathbf{t})=\hat{c}_{00}+\sum_{l=1}^{d_1}\hat{c}_{0l}t_l+\sum_{\alpha=1}^{d_2}\sum_{J=1}^{N+1}\hat{c}_{J,\alpha}b_J(x_\alpha), \tag{2}$$

where the coefficients $(\hat{c}_{00},\hat{c}_{0l},\hat{c}_{J,\alpha})_{1\le l\le d_1,1\le J\le N+1,1\le\alpha\le d_2}$ minimize

$$\sum_{i=1}^n\left\{Y_i-c_0-\sum_{l=1}^{d_1}c_lT_{il}-\sum_{\alpha=1}^{d_2}\sum_{J=1}^{N+1}c_{J,\alpha}b_J(X_{i\alpha})\right\}^2.$$

Pilot estimators of $\mathbf{c}^T=\{c_{0l}\}_{l=0}^{d_1}$ and $m_\alpha(x_\alpha)$ are $\hat{\mathbf{c}}^T=\{\hat{c}_{0l}\}_{l=0}^{d_1}$ and $\hat{m}_\alpha(x_\alpha)=\sum_{J=1}^{N+1}\hat{c}_{J,\alpha}b_J(x_\alpha)-n^{-1}\sum_{i=1}^n\sum_{J=1}^{N+1}\hat{c}_{J,\alpha}b_J(X_{i\alpha})$, which are used to define pseudo-responses $\hat{Y}_{i\alpha}$, estimates of the unobservable "oracle" responses $Y_{i\alpha}$:

$$\hat{Y}_{i\alpha}=Y_i-\hat{c}_{00}-\sum_{l=1}^{d_1}\hat{c}_{0l}T_{il}-\sum_{\beta=1,\beta\neq\alpha}^{d_2}\hat{m}_\beta(X_{i\beta}),$$

$$Y_{i\alpha} = Y_i - c_{00} - \sum_{l=1}^{d_1} c_{0l} T_{il} - \sum_{\beta=1, \beta \neq \alpha}^{d_2} m_\beta(X_{i\beta}). \tag{3}$$

Based on $\{\hat{Y}_{i\alpha}, X_{i\alpha}\}_{i=1}^n$, the SBK estimator $\hat{m}_{\mathrm{SBK},\alpha}(x_\alpha)$ of $m_\alpha(x_\alpha)$ mimics the would-be Nadaraya–Watson estimator $\tilde{m}_{\mathrm{K},\alpha}(x_\alpha)$ of $m_\alpha(x_\alpha)$ based on $\{Y_{i\alpha}, X_{i\alpha}\}_{i=1}^n$, if the unobservable responses $\{Y_{i\alpha}\}_{i=1}^n$ were available

$$\hat{m}_{\mathrm{SBK},\alpha}(x_\alpha) = \left\{ n^{-1} \sum_{i=1}^n K_h(X_{i\alpha} - x_\alpha) \hat{Y}_{i\alpha} \right\} \Big/ \hat{f}_\alpha(x_\alpha),$$

$$\tilde{m}_{\mathrm{K},\alpha}(x_\alpha) = \left\{ n^{-1} \sum_{i=1}^n K_h(X_{i\alpha} - x_\alpha) Y_{i\alpha} \right\} \Big/ \hat{f}_\alpha(x_\alpha), \tag{4}$$

with $\hat{Y}_{i\alpha}, Y_{i\alpha}$ in (3), $\hat{f}_\alpha(x_\alpha) = n^{-1} \sum_{i=1}^n K_h(X_{i\alpha} - x_\alpha)$ an estimator of $f_\alpha(x_\alpha)$.

Without loss of generality, let $\alpha = 1$. Under Assumptions (A1)–(A5) and (A7), it is straightforward to verify (as in Bosq, 1998) that as $n \to \infty$,

$$\sup_{x_1 \in [h, 1-h]} |\tilde{m}_{\mathrm{K},1}(x_1) - m_1(x_1)| = o_p(n^{-2/5} \log n),$$

$$\sqrt{nh}\{\tilde{m}_{\mathrm{K},1}(x_1) - m_1(x_1) - b_1(x_1)h^2\} \xrightarrow{D} \mathrm{N}\{0, v_1^2(x_1)\},$$

where $b_1(x_1) = \int u^2 K(u)\, du \{m_1''(x_1) f_1(x_1)/2 + m_1'(x_1) f_1'(x_1)\} f_1^{-1}(x_1)$,

$$v_1^2(x_1) = \int K^2(u)\, du\, E[\sigma^2(\mathbf{X}, \mathbf{T}) | X_1 = x_1] f_1^{-1}(x_1). \tag{5}$$

It is shown in Li (2000) and Schimek (2000) that the direct spline estimator $\hat{m}_1(x_1)$ in step one converges uniformly to $m_1(x_1)$ with certain rate, but without asymptotic distribution. In contrast, Theorem 1 below states that the difference between $\hat{m}_{\mathrm{SBK},1}(x_1)$ and $\tilde{m}_{\mathrm{K},1}(x_1)$ is $o_p(n^{-2/5})$ uniformly, dominated by the size of the error $\tilde{m}_{\mathrm{K},1}(x_1) - m_1(x_1)$. So $\hat{m}_{\mathrm{SBK},1}(x_1)$ has identical asymptotic distribution as $\tilde{m}_{\mathrm{K},1}(x_1)$.

**Theorem 1.** *Under Assumptions* (A1)–(A7), *as* $n \to \infty$, *the SBK estimator* $\hat{m}_{\mathrm{SBK},1}(x_1)$ *given in* (4) *satisfies* $\sup_{x_1 \in [0,1]} |\hat{m}_{\mathrm{SBK},1}(x_1) - \tilde{m}_{\mathrm{K},1}(x_1)| = o_p(n^{-2/5})$. *Hence with* $b_1(x_1)$ *and* $v_1^2(x_1)$ *as defined in* (5), *for any* $x_1 \in [h, 1-h]$, $\sqrt{nh}\{\hat{m}_{\mathrm{SBK},1}(x_1) - m_1(x_1) - b_1(x_1)h^2\} \xrightarrow{D} \mathrm{N}\{0, v_1^2(x_1)\}$.

Instead of Nadaraya–Watson estimator, one can use local polynomial estimator, see Fan and Gijbels (1996). Under Assumptions (A1)–(A7), for any $\alpha \in (0,1)$, an asymptotic $100(1-\alpha)\%$ confidence intervals for $m_1(x_1)$ is

$$\hat{m}_{\mathrm{SBK},1}(x_1) - \hat{b}_1(x_1)h^2 \pm z_{\alpha/2} \hat{v}_1(x_1)(nh)^{-1/2}, \tag{6}$$

where $\hat{b}_1(x_1)$ and $\hat{v}_1^2(x_1)$ are estimators of $b_1(x_1)$ and $v_1^2(x_1)$, respectively.

The following corollary provides the asymptotic distribution of $\hat{m}_{\mathrm{SBK}}(\mathbf{x})$. The proof of this corollary is straightforward and therefore omitted.

**Corollary 1.** *Under Assumptions* (A1)–(A7) *and the additional assumption* $m_\alpha \in C^{(2)}[0,1]$, $2 \le \alpha \le d_2$. *Let* $\hat{m}_{\mathrm{SBK}}(\mathbf{x}) = \sum_{\alpha=1}^{d_2} \hat{m}_{\mathrm{SBK},\alpha}(x_\alpha)$, $b(\mathbf{x}) = \sum_{\alpha=1}^{d_2} b_\alpha(x_\alpha)$, $v^2(\mathbf{x}) = \sum_{\alpha=1}^{d_2} v_\alpha^2(x_\alpha)$, *for any* $\mathbf{x} \in [0,1]^{d_2}$, *with SBK estimators* $\hat{m}_{\mathrm{SBK},\alpha}(x_\alpha)$, $1 \le \alpha \le d_2$, *defined in* (4), *and* $b_\alpha(x_\alpha)$, $v_\alpha^2(x_\alpha)$ *similarly defined as in* (5), *as* $n \to \infty$, $\sqrt{nh}\{\hat{m}_{\mathrm{SBK}}(\mathbf{x}) - \sum_{\alpha=1}^{d_2} m_\alpha(x_\alpha) - b(\mathbf{x})h^2\} \xrightarrow{D} \mathrm{N}\{0, v^2(\mathbf{x})\}$

Next theorem describes the asymptotic behavior of estimator $\hat{\mathbf{c}}$ for $\mathbf{c}$.

**Theorem 2.** *Under Assumptions* (A1)–(A6), *as* $n \to \infty$, $\|\hat{\mathbf{c}} - \mathbf{c}\| = O_p(n^{-1/2})$. *With the additional Assumption* (A8),

$$\sqrt{n}(\hat{\mathbf{c}} - \mathbf{c}) \to_d \mathrm{N}\left(0, \sigma_0^2 \begin{Bmatrix} 1 & 0_{d_1}^T \\ 0_{d_1} & \Sigma^{-1} \end{Bmatrix}\right),$$

*for* $\Sigma = \mathrm{cov}(\tilde{\mathbf{T}})$ *with random vector* $\tilde{\mathbf{T}}$ *defined in* (A.3).

To construct confidence sets for $\mathbf{c}$, $\Sigma$ is consistently estimated by $n^{-1} \sum_{i=1}^n \hat{T}_{i,l,n} \hat{T}_{i,l',n}$ in which $\hat{T}_{l,n} = T_l - \sum_{\alpha=1}^{d_2} \sum_{J=1}^{N+1} \hat{a}_{J,\alpha} b_{J,\alpha}^*(x_\alpha)$, where $b_{J,\alpha}^*(x_\alpha) \equiv b_J(x_\alpha) - n^{-1} \sum_{i=1}^n b_J(X_{i\alpha})$ is the empirical centering of $b_J(x)$ for the $\alpha$-th variable $X_\alpha$, defined in Appendix A and $(\hat{a}_{J,\alpha})_{1 \le J \le N+1, 1 \le \alpha \le d_2}$ minimize $\|T_l - \sum_{\alpha=1}^{d_2} \sum_{J=1}^{N+1} a_{J,\alpha} b_{J,\alpha}^*(X_\alpha)\|_n^2$.

## 3. Simulation

In this section, we analyze some simulated data examples to illustrate the finite-sample behavior of SBK estimators. The number of interior knots $N$ in (2) is given by $N=\min([c_1 n^{1/4}\log n]+c_2,[(n/2-1-d_1)d_2^{-1}-1])$, in which $[a]$ denotes the integer part of $a$. In our implementation, we have used $c_1=c_2=1$. The additional constraint that $N \le (n/2-1-d_1)d_2^{-1}-1$ ensures that the number of terms in the linear least squares problem (2), $1+d_1+d_2(N+1)$, is no greater than $n/2$, which is necessary when the sample size $n$ is moderate.

The i.i.d. data $\{Y_i,\mathbf{X}_i,\mathbf{T}_i\}_{i=1}^n$ are generated according to the partially linear additive model (1), which satisfies Assumptions (A1)–(A5), and (A8)

$$Y_i = 2+\sum_{l=1}^{d_1} T_{il}+\sum_{\alpha=1}^{d_2} m_\alpha(X_{i\alpha})+\sigma_0\varepsilon_i, \quad m_\alpha(x)\equiv \sin(2\pi x), \quad 1\le\alpha\le d_2,$$

where $\sigma_0=2$, $\varepsilon_i \sim \mathrm{N}(0,1)$ is independent of $(\mathbf{X}_i,\mathbf{T}_i)$,$\mathbf{T}_i=(T_{i1},T_{i2},T_{i3},T_{i4})$ such that $T_{i3},T_{i4}$, $(T_{i1},T_{i2})$ are independent, $T_{i3}=\pm 1$ with probability 1/2, $T_{i4}\sim \mathrm{U}(-0.5,0.5)$, $(T_{i1},T_{i2})' \sim \mathrm{N}((0,0)',\mathrm{diag}(a(X_{i1}),a(X_{i2})))$, $a(x)=(5-\sin(2\pi x))/(5+\sin(2\pi x))$. $\mathbf{X}_i=\{(X_{i\alpha})_{\alpha=1}^{d_2}\}^T$ is generated from the vector autoregression (VAR) equation $X_{i\alpha}=\Phi\{(1-a^2)^{1/2}Z_{i\alpha}\}-1/2$, $1\le\alpha\le d_2$ with stationary distribution $\mathbf{Z}_i=(Z_{i1},\ldots,Z_{id_2})^T \sim \mathrm{N}(0_{d_2},(1-a^2)^{-1}\boldsymbol{\Sigma})$

$$\mathbf{Z}_1 \sim \mathrm{N}(0_{d_2},(1-a^2)^{-1}\boldsymbol{\Sigma}), \quad \mathbf{Z}_i = a\mathbf{Z}_{i-1}+\boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_i\sim \mathrm{N}(0,\boldsymbol{\Sigma}), \ 2\le i\le n,$$

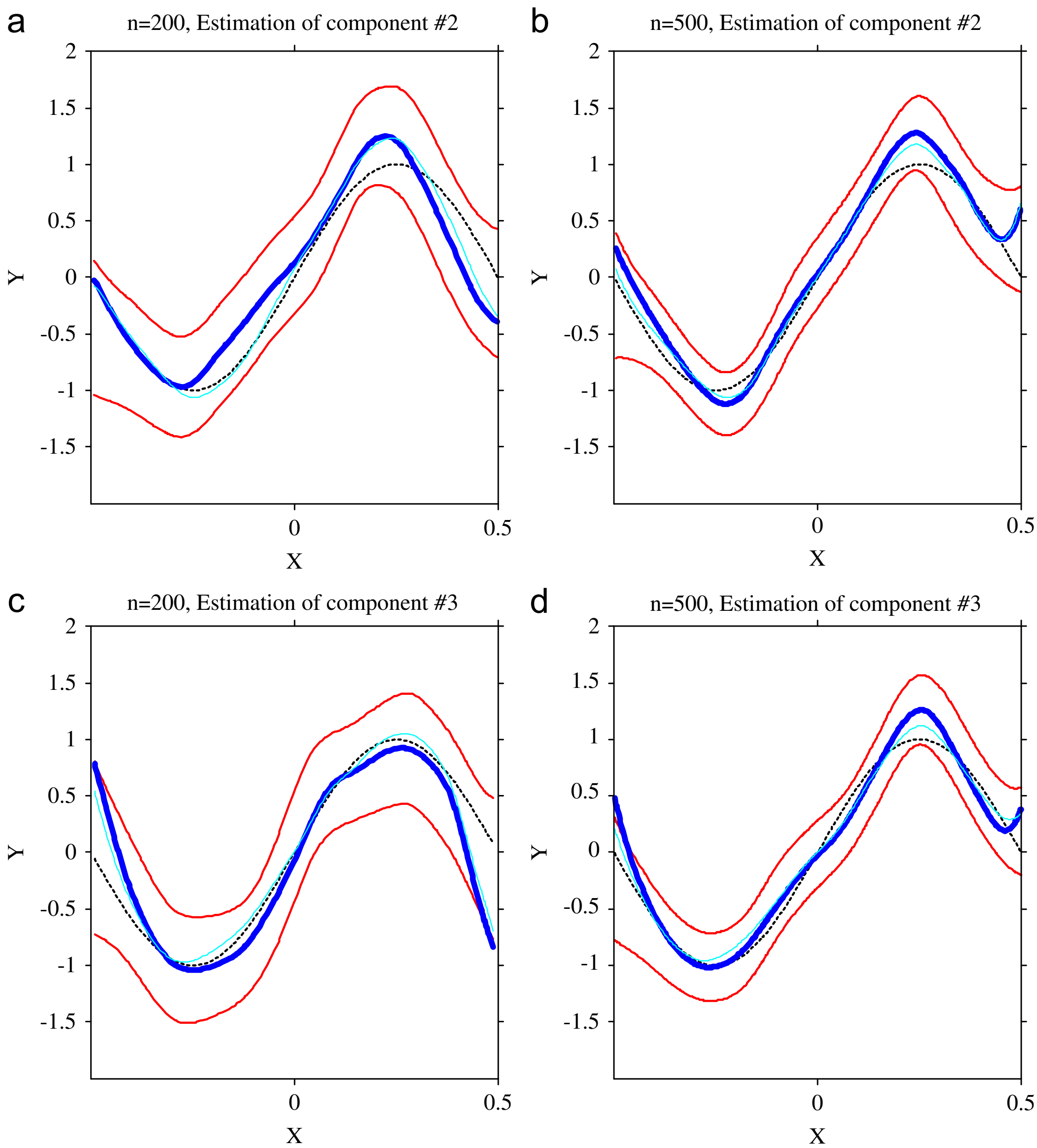


**Fig. 1.** Plots of the oracle smoother $\tilde{m}_{K,\alpha}(x_\alpha)$ (thin curve), the SBK estimator $\hat{m}_{SBK,\alpha}(x_\alpha)$ (thick curve) defined in (4), and the 95% pointwise confidence intervals constructed by (6) (upper and lower medium curves) of the function components $m_\alpha(x_\alpha)$, $\alpha=2,3$ (dashed curve), $d_1=4$, $d_2=3$.

$$\mathbf{\Sigma} = (1-r)\mathbf{I}_{d_2\times d_2} + r\mathbf{1}_{d_2}\mathbf{1}_{d_2}^T, \quad 0<a<1, \ 0<r<1.$$

So $\{\mathbf{X}_i\}_{i=1}^n$ is geometrically $\alpha$-mixing with marginal distribution $U[-0.5,0.5]$.

We obtained for comparison the SBK estimator $\hat{m}_{\text{SBK},\alpha}(x_\alpha)$ and the "oracle" smoother $\tilde{m}_{\text{K},\alpha}(x_\alpha)$ by Nadaraya–Watson regression using quartic kernel and the rule-of-thumb bandwidth. To see that $\hat{m}_{\text{SBK},\alpha}(x_\alpha)$ is as efficient as $\tilde{m}_{\text{K},\alpha}(x_\alpha)$ for numerical performance, we define the empirical relative efficiency of $\hat{m}_{\text{SBK},\alpha}(x_\alpha)$ with respect to $\tilde{m}_{\text{K},\alpha}(x_\alpha)$ as

$$\text{eff}_\alpha = \left[\frac{\sum_{i=1}^n \{\tilde{m}_{\text{K},\alpha}(x_\alpha) - m_\alpha(X_{i\alpha})\}^2}{\sum_{i=1}^n \{\hat{m}_{\text{SBK},\alpha}(x_\alpha) - m_\alpha(X_{i\alpha})\}^2}\right]^{1/2}. \tag{7}$$

Theorem 1 indicates $\mathit{eff}_\alpha$ should be close to 1 for $1 \le \alpha \le d_2$. Fig. 2 provides the kernel density estimates of 100 empirical efficiencies $\alpha = 2,3$, sample sizes $n$=100 (solid lines), 200 (dashed lines), 500 (thin lines) and 1000 (thick lines) at $\sigma_0 = 2$, $d_2$=3 for (a), (b) and $d_2$=30 for (c), (d). The vertical line at efficiency =1 is the standard line for the comparison of $\hat{m}_{\text{SBK},\alpha}(x_\alpha)$ and $\tilde{m}_{\text{K},\alpha}(x_\alpha)$. One clearly sees that the center of the density plots is going toward the standard line at 1 with narrower spread when sample size $n$ is increasing, confirmative to Theorem 1.

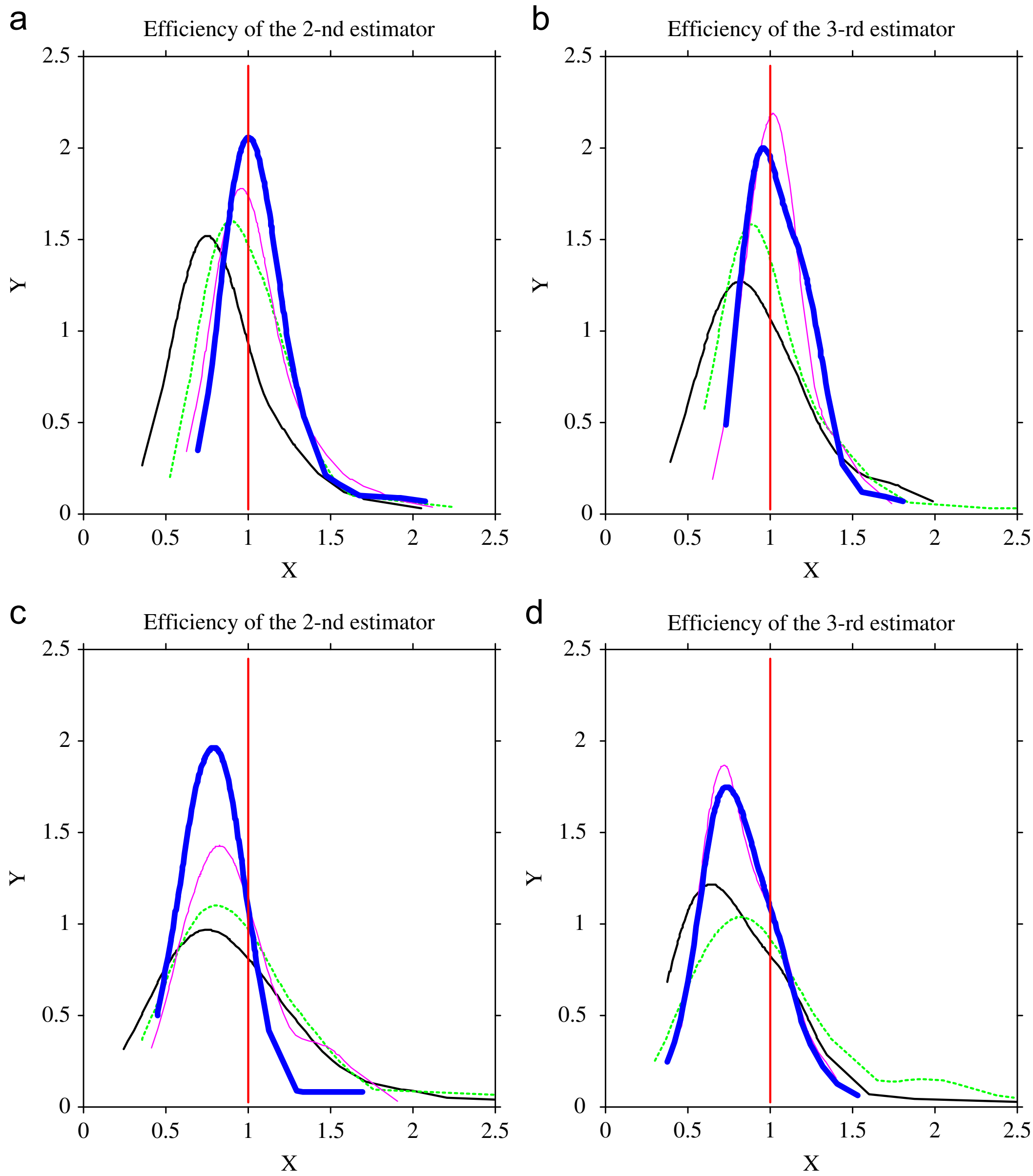


**Fig. 2.** Kernel density plots of the 100 empirical efficiencies of $\hat{m}_{\text{SBK},\alpha}(x_\alpha)$ to $\tilde{m}_{\text{K},\alpha}(x_\alpha)$, computed according to (7): (a) $\alpha=2$, $d_1=4$, $d_2=3$; (b) $\alpha=3$, $d_1=4$, $d_2=3$; (c) $\alpha=2$, $d_1=4$, $d_2=30$; (d) $\alpha=3$, $d_1=4$, $d_2=30$.

**Table 1**
Estimation of $c=(c_{00},c_{01},c_{02},c_{03},c_{04})'$ with $d_2$=3, $n$=200 (outside parentheses), $n$=500 (inside parentheses).

| | $r$ | $a$ | 95% CI coverage frequency | MSE | Asymptotic efficiency |
|---|---|---|---|---|---|
| $c_{00}$ | 0 | 0 | 0.92(0.92) | 0.0241 (0.010) | 0.8806 (0.8406) |
| | 0.3 | 0 | 0.92 (0.91) | 0.0264 (0.0095) | 0.8403 (0.8588) |
| | 0 | 0.3 | 0.89 (0.92) | 0.0263 (0.0096) | 0.8446 (0.8536) |
| | 0.3 | 0.3 | 0.89 (0.92) | 0.0282 (0.0103) | 0.8146 (0.8270) |
| $c_{01}$ | 0 | 0 | 0.95 (0.90) | 0.0330 (0.0152) | 0.8795 (0.8892) |
| | 0.3 | 0 | 0.99 (0.91) | 0.0297 (0.0143) | 0.9217 (0.9069) |
| | 0 | 0.3 | 0.98 (0.95) | 0.0296 (0.0121) | 0.9157 (0.9949) |
| | 0.3 | 0.3 | 0.96 (0.94) | 0.0336 (0.0134) | 0.8635 (0.9491) |
| $c_{02}$ | 0 | 0 | 0.96 (0.95) | 0.0306 (0.0115) | 0.8809 (0.8659) |
| | 0.3 | 0 | 0.97 (0.97) | 0.0378 (0.0118) | 0.7914 (0.8553) |
| | 0 | 0.3 | 0.95 (0.95) | 0.0329 (0.0112) | 0.8523 (0.8757) |
| | 0.3 | 0.3 | 0.97 (0.97) | 0.0336 (0.0104) | 0.8397 (0.9039) |
| $c_{03}$ | 0 | 0 | 0.96 (0.97) | 0.0259 (0.0087) | 0.8892 (0.8983) |
| | 0.3 | 0 | 0.92 (0.98) | 0.0301 (0.0074) | 0.7527 (0.9327) |
| | 0 | 0.3 | 0.93 (0.96) | 0.0362 (0.0078) | 0.8264 (0.9178) |
| | 0.3 | 0.3 | 0.96 (0.97) | 0.0258 (0.0078) | 0.8919 (0.9101) |
| $c_{04}$ | 0 | 0 | 0.95 (0.96) | 0.4006 (0.1229) | 0.7873 (0.9181) |
| | 0.3 | 0 | 0.94 (0.95) | 0.3771 (0.1111) | 0.8117 (0.9661) |
| | 0 | 0.3 | 0.92 (0.95) | 0.3867 (0.1154) | 0.8019 (0.9470) |
| | 0.3 | 0.3 | 0.93 (0.96) | 0.3533 (0.1138) | 0.8388 (0.9537) |

To see that $\hat{c}_{0l}$ is as efficient as $\tilde{c}_{0l}$, we define the asymptotic efficiency of $\hat{c}_{0l}$ with respect to $\tilde{c}_{0l}$ as $\mathrm{eff}_l=[\sum_{t=1}^{100}\{\tilde{c}_{0l,t}-c_{0l}\}^2/100/\sum_{t=1}^{100}\{\hat{c}_{0l,t}-c_{0l}\}^2/100]^{1/2}$, where $\tilde{c}_{0l,t}$,$\hat{c}_{0l,t}$ are values of $\tilde{c}_{0l}$,$\hat{c}_{0l}$ for the $t$-th replication in the simulation. For $n$=200,500, $d_2$=3, Table 1 lists the frequencies of 95% confidence interval coverage of the SBK estimators for the regression coefficients $\{c_{0l}\}_{l=0}^{d_1}$, the sample mean squared error (MSE) and the asymptotic efficiency. The coverage frequencies are all close to the nominal level of 95%. As expected, increase in sample size reduces the sample MSE and increases the asymptotic efficiency.

For visualization of the actual function estimates, at noise level $\sigma_0=2$ with sample size $n$=200,500, we plot $\tilde{m}_{\mathrm{K},\alpha}(x_\alpha)$ (thin curves), $\hat{m}_{\mathrm{SBK},\alpha}(x_\alpha)$ (thick curves) and their 95% pointwise confidence intervals (upper and lower medium curves) for $m_\alpha$ (dashed curves) in Fig. 1. The SBK estimators seem rather satisfactory and their performance improves with increasing $n$.

## 4. Application

In this section we apply our method to the well-known Boston housing data, which contains 506 different houses from a variety of locations in Boston Standard Metropolitan Statistical Area in 1970. The median value and 13 sociodemographic statistics values of the Boston houses were first studied by Harrison and Rubinfeld (1978) to estimate the housing price index model. Breiman and Friedman (1985) did further analysis to deal with the multi-collinearity for overfitting by using a stepwise method. The response and explanatory variables of interest are:

MEDV: median value of owner-occupied homes in $1000s;
RM: average number of rooms per dwelling;
TAX: full-value property-tax rate per $10,000;
PTRATIO: pupil-teacher ratio by town school district;
LSTAT: proportion of population that is of "lower status" in %.

Wang and Yang (2009) fitted an additive model using RM, log (TAX), PTRSATIO and log(LSTAT) as predictors to test the linearity of the components and found that only PTRATIO is accepted at the significance level 0.05 for the linearity hypothesis test. Based on the conclusion drawn from Wang and Yang (2009), we fitted a partial linear additive model as follows:

$$\mathrm{MEDV} = c_{00} + c_{01} \times \mathrm{PTRATIO} + m_1(\mathrm{RM}) + m_2(\log(\mathrm{TAX})) + m_3(\log(\mathrm{LSTAT})) + \varepsilon.$$

As in Wang and Yang (2009), the number of interior knots is $N=5$.

In Fig. 3, the univariate nonlinear function estimates (dashed lines) and corresponding simultaneous confidence bands (thin lines) are displayed together with the "pseudo-data points" (dots) with pseudo response as the backfitted response after subtracting the sum function of the remaining covariates as in (3). The confidence bands are used to test the linearity of the nonparametric components. In Fig. 3 the straight solid lines are the least squares regression lines through the pseudo-data points. The first figure confidence band with 0.999999 confidence level does not totally cover the straight regression line, i.e., the $p$-value is less than 0.000001. Similarly the linearity of the component functions for log(TAX) and log(LSTAT) are rejected at the significance levels 0.017 and 0.007, respectively. The estimators $\hat{c}_{00}$ and $\hat{c}_{01}$ of $c_{00}$ and $c_{01}$ are 33.393 and $-0.58845$ and both are significant with $p$-values close to 0. The correlation between the estimated and observed values of MEDV is 0.89944, much higher than 0.80112 obtained by Wang and Yang (2009). This improvement is due to fitting the variable PTRATIO directly as linear with the higher accuracy of parametric model instead of treating it unnecessarily as a nonparametric variable. In other words, our simpler partially linear additive model (PLAM) fits the housing data much better than the additive model of Wang and Yang (2009).

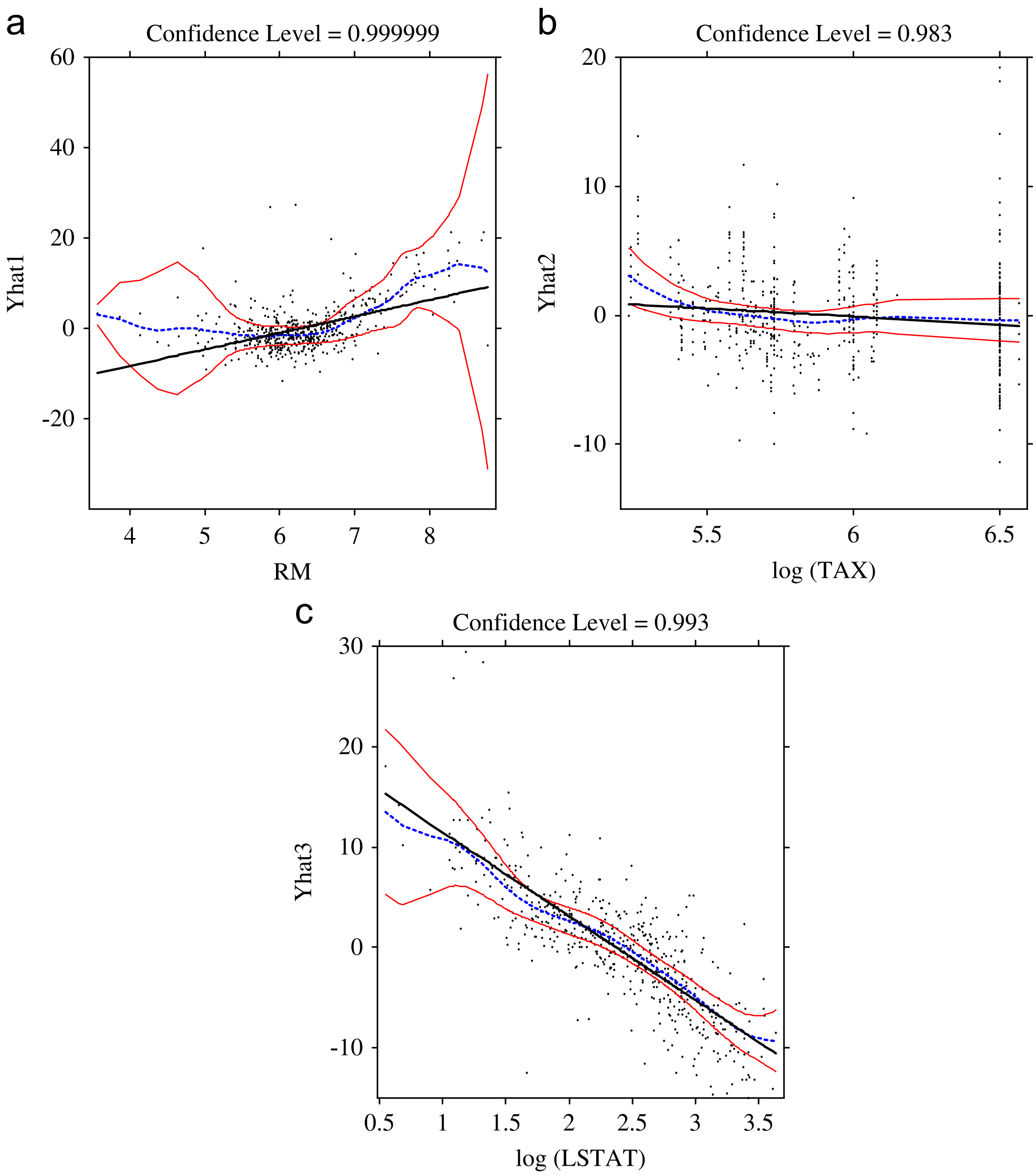


**Fig. 3.** Plots of the least squares regression estimator (solid line), confidence bands (upper and lower thin lines), the spline estimator (dashed line) and the data (dot).

**Acknowledgement**

This work is part of the first author's dissertation under the supervision of the second author and has been partially supported by NSF grant DMS 0706518. The helpful comments from a referee are gratefully acknowledged.

## Appendix A

Throughout this section, $a_n \gg b_n$ means $\lim_{n\to\infty} b_n/a_n = 0$, and $a_n \sim b_n$ means $\lim_{n\to\infty} b_n/a_n = c$, where $c$ is some nonzero constant.

We state the following assumptions.

(A1) Given $1 \le \alpha \le d_2$, $m_\alpha \in C^{(2)}[0,1]$, while there is a constant $0 < C_\infty < \infty$, such that $m'_\beta \in \mathrm{Lip}([0,1],C_\infty)$, $\forall 1 \le \beta \le d_2$ and $\beta \ne \alpha$.

(A2) Vector process $\{\mathbf{Z}_t\}_{t=-\infty}^{\infty} = \{(\mathbf{X}_t^T,\mathbf{T}_t^T,\varepsilon_t)\}_{t=-\infty}^{\infty}$ is strictly stationary and geometrically strongly mixing, that is, its $\alpha$-mixing coefficient $\alpha(k) \le K_0 e^{-\lambda_0 k}$, for $K_0,\lambda_0 > 0$, where

$$\alpha(k) = \sup_{B\in\sigma\{\mathbf{Z}_t, t\le 0\}, C\in\sigma\{\mathbf{Z}_t, t\ge k\}} |P(B\cap C) - P(B)P(C)|. \tag{A.1}$$

(A3) The noise $\varepsilon_i$ satisfies $E(\varepsilon_i|\mathcal{F}_i)=0$, $E(\varepsilon_i^2|\mathcal{F}_i)=1$, $E(|\varepsilon_i|^{2+\delta}|\mathcal{F}_i) < M_\delta < +\infty$ for some $\delta > 2/3$, $M_\delta > 0$, and $\sigma$-fields $\mathcal{F}_i = \sigma\{(\mathbf{X}_j,\mathbf{T}_j), j\le i; \varepsilon_j, j \le i-1, 1\le i\le n\}$. Function $\sigma(\mathbf{x},\mathbf{t})$ is continuous with

$$0 < c_\sigma \le \inf_{\mathbf{x}\in[0,1]^{d_2},\mathbf{t}\in\mathbf{R}^{d_1}} \sigma(\mathbf{x},\mathbf{t}) \le \sup_{\mathbf{x}\in[0,1]^{d_2},\mathbf{t}\in\mathbf{R}^{d_1}} \sigma(\mathbf{x},\mathbf{t}) \le C_\sigma < \infty.$$

(A4) The density function $f(\mathbf{x})$ of $X$ and the marginal densities $f_\alpha(x_\alpha)$ of $X_\alpha$ satisfy $f \in C[0,1]^{d_2}$, $0 < c_f \le \inf_{\mathbf{x}\in[0,1]^{d_2}} f(\mathbf{x}) \le \sup_{\mathbf{x}\in[0,1]^{d_2}} f(\mathbf{x}) \le C_f < \infty$, $f_\alpha \in C^{(1)}[0,1]$.

(A5) There exist constants $0 < c_\delta \le C_\delta < +\infty, 0 < c_\mathbf{Q} \le C_\mathbf{Q} < +\infty$ such that $c_\delta \le E(|T_l|^{2+\delta}|\mathbf{X}=\mathbf{x}) \le C_\delta$, $\forall x \in [0,1]^{d_2}$, $1 \le l \le d_1$, and $c_\mathbf{Q} I_{(d_1+1)\times(d_1+1)} \le Q(\mathbf{x}) \le C_\mathbf{Q} I_{(d_1+1)\times(d_1+1)}$, where $\mathbf{Q}(\mathbf{x}) = E\{(1\ \mathbf{T}^T)^T(1\ \mathbf{T}^T)|\mathbf{X}=\mathbf{x}\}$.

(A6) The number of interior knots $N = N_n \sim n^{1/4}\log n$, i.e., $c_N n^{1/4}\log n \le N \le C_N n^{1/4}\log n$ for some positive constants $c_N$, $C_N$.

(A7) The kernel function $K \in \mathrm{Lip}([-1,1],C_\infty)$ for $C_\infty > 0$ is bounded, nonnegative, symmetric, and supported on $[-1,1]$. The bandwidth $h \sim n^{-1/5}$, i.e., $c_h n^{-1/5} \le h \le C_h n^{-1/5}$ for positive constants $C_h$, $c_h$.

Assumption (A1) on the smoothness of the component functions is greatly relaxed and is close to being the minimal. Assumption (A2) is typical in time series literature while Assumptions (A3)–(A5) are typical in nonparametric smoothing literature, see for instance, Fan and Gijbels (1996).

For $\phi,\varphi$ on $[0,1]^{d_2}\times\mathbf{R}^{d_1}$, define the empirical inner product and empirical norm as $\langle\phi,\varphi\rangle_n = n^{-1}\sum_{i=1}^n \phi(\mathbf{X}_i,\mathbf{T}_i)\varphi(\mathbf{X}_i,\mathbf{T}_i)$, $\|\phi\|_n^2 = n^{-1}\sum_{i=1}^n \phi^2(\mathbf{X}_i,\mathbf{T}_i)$. If $\phi,\varphi$ are $L^2$-integrable, we define the theoretical inner product and theoretical $L^2$ norm as $\langle\phi,\varphi\rangle = E\{\phi(\mathbf{X}_i,\mathbf{T}_i)\varphi(\mathbf{X}_i,\mathbf{T}_i)\}$, $\|\phi\|^2 = E\{\phi^2(\mathbf{X}_i,\mathbf{T}_i)\}$ and denote $E_n\phi = \langle\phi,1\rangle_n$. $\phi$ is empirically(theoretically) centered if $E_n\phi = 0 (E\phi = 0)$. For theoretical analysis, define the centered spline basis as $b_{J,\alpha}(x_\alpha) = b_J(x_\alpha) - (c_{J,\alpha}/c_{J-1,\alpha} b_{J-1})(x_\alpha)$, $\forall 1\le\alpha\le d_2$, $1\le J\le N+1$, where $c_{J,\alpha} = Eb_J(X_\alpha) = \int b_J(x_\alpha)f_\alpha(x_\alpha)\,dx_\alpha$. The standardized basis is

$$B_{J,\alpha}(x_\alpha) = b_{J,\alpha}(x_\alpha)/\|b_{J,\alpha}\|, \quad \forall 1\le\alpha\le d_2,\ 1\le J\le N+1. \tag{A.2}$$

For the proof of Theorem 2, define the Hilbert space

$$\mathcal{H} = \left\{ p(\mathbf{x}) = \sum_{\alpha=1}^{d_2} p_\alpha(x_\alpha), Ep_\alpha(X_\alpha) = 0, E^2 p_\alpha(X_\alpha) < \infty \right\}$$

of theoretically centered $L_2$ additive functions on $[0,1]^{d_2}$, while denote by $\mathcal{H}_n$ its subspace spanned by $\{B_{J,\alpha}(x_\alpha), 1\le\alpha\le d_2, 1\le J\le N+1\}$. Denote

$$\mathrm{Proj}_\mathcal{H} T_l = p_l(\mathbf{X}) = \operatorname*{argmin}_{p\in\mathcal{H}} E\{T_l - p(\mathbf{X})\}^2, \quad \tilde{T}_l = T_l - \mathrm{Proj}_\mathcal{H} T_l,$$

$$\mathrm{Proj}_{\mathcal{H}_n} T_l = \operatorname*{argmin}_{p\in\mathcal{H}_n} E\{T_l - p(\mathbf{X})\}^2, \quad \tilde{T}_{l,n} = T_l - \mathrm{Proj}_{\mathcal{H}_n} T_l,$$

for $1\le l\le d_1$, where $\mathrm{Proj}_\mathcal{H} T_l$ and $\mathrm{Proj}_{\mathcal{H}_n} T_l$ are orthogonal projections of $T_l$ unto subspaces $\mathcal{H}$ and $\mathcal{H}_n$ respectively. Denote next in vector form

$$\tilde{\mathbf{T}}_n = \{\tilde{T}_{l,n}\}_{1\le l\le d_1}, \quad \tilde{\mathbf{T}} = \{\tilde{T}_l\}_{1\le l\le d_1}. \tag{A.3}$$

The next assumption is needed for the second part of Theorem 2.

(A8) Functions $p_l \in C[0,1]^{d_2}$, $1\le l\le d_1$ while $\sigma(\mathbf{x},\mathbf{t}) \equiv \sigma_0, (\mathbf{x},\mathbf{t}) \in [0,1]^{d_2}\times R^{d_1}$.

$\hat{m}(\mathbf{x},\mathbf{t})$ can be expressed in terms of the standardized basis

$$\hat{m}(\mathbf{x},\mathbf{t}) = \hat{c}_{00} + \sum_{l=1}^{d_1} \hat{c}_{0l} t_l + \sum_{\alpha=1}^{d_2} \sum_{J=1}^{N+1} \hat{c}_{J,\alpha} B_{J,\alpha}(x_\alpha), \tag{A.4}$$

where $(\hat{c}_{00}, \hat{c}_{0l}, \hat{c}_{J,\alpha})_{1 \le l \le d_1, 1 \le J \le N+1, 1 \le \alpha \le d_2}$ minimize

$$\sum_{i=1}^{n} \left\{ Y_i - c_0 - \sum_{l=1}^{d_1} c_l T_{il} - \sum_{\alpha=1}^{d_2} \sum_{J=1}^{N+1} c_{J,\alpha} B_{J,\alpha}(X_{i\alpha}) \right\}^2. \tag{A.5}$$

While (2) is used for statistical implementation, algebraically equivalent (A.4) is for mathematical analysis. Pilot estimators of $m_\alpha(x_\alpha)$ and $\mathbf{c}^T$ are

$$\hat{m}_\alpha(x_\alpha) = \sum_{J=1}^{N+1} \hat{c}_{J,\alpha} B_{J,\alpha}^*(x_\alpha), \quad \hat{\mathbf{c}}^T = \{\hat{c}_{00}, \hat{c}_{0l}\}_{l=1}^{d_1}, \tag{A.6}$$

where $B_{J,\alpha}^*(x_\alpha) \equiv B_{J,\alpha}(x_\alpha) - E_n B_{J,\alpha} = B_{J,\alpha}(x_\alpha) - n^{-1} \sum_{i=1}^{n} B_{J,\alpha}(X_{i\alpha})$ is the empirical centering of $B_{J,\alpha}(x_\alpha)$. The evaluation of $\hat{m}(\mathbf{x},\mathbf{t})$ at the $n$ observations results in an $n$-dimensional vector $\{\hat{m}(\mathbf{X}_i,\mathbf{T}_i)\}_{1 \le i \le n}^T$, the projection of $\mathbf{Y}$ onto $G_n$ with respect to the empirical inner product $\langle \cdot, \cdot \rangle_n$. In general, for any $n$-dimensional vector $\mathbf{\Lambda} = \{\Lambda_i\}_{1 \le i \le n}^T$, we define $\mathbf{P}_n \mathbf{\Lambda}(\mathbf{x},\mathbf{t})$ as the projection of $\mathbf{\Lambda}$ onto $(G_n, \langle \cdot, \cdot \rangle_n)$, i.e., $\mathbf{P}_n \mathbf{\Lambda}(\mathbf{x},\mathbf{t}) = \hat{\lambda}_0 + \sum_{l=1}^{d_1} \hat{\lambda}_l t_l + \sum_{\alpha=1}^{d_2} \sum_{J=1}^{N+1} \hat{\lambda}_{J,\alpha} B_{J,\alpha}(x_\alpha)$, with $Y_i$ replaced by $\Lambda_i$ and coefficients $\{\hat{\lambda}_0, \hat{\lambda}_l, \hat{\lambda}_{J,\alpha}\}$ given in (A.5). Define the empirically centered additive components as $\mathbf{P}_{n,\alpha} \mathbf{\Lambda}(x_\alpha) = \sum_{J=1}^{N+1} \hat{\lambda}_{J,\alpha} B_{J,\alpha}^*(x_\alpha), 1 \le \alpha \le d_2$, and the linear component as $(\mathbf{P}_{n,c} \mathbf{\Lambda})^T = \{\hat{\lambda}_0, \hat{\lambda}_l\}_{1 \le l \le d_1}$. Rewrite estimators $\hat{m}(\mathbf{x})$, $\hat{m}_\alpha(x_\alpha)$, $\hat{\mathbf{c}}$ defined in (A.4) and (A.6) as $\hat{m}(\mathbf{x},\mathbf{t}) = \mathbf{P}_n \mathbf{Y}(\mathbf{x},\mathbf{t}), \hat{m}_\alpha(x_\alpha) = \mathbf{P}_{n,\alpha} \mathbf{Y}(x_\alpha), \hat{\mathbf{c}} = \mathbf{P}_{n,c} \mathbf{Y}$. The noiseless and noisy components are

$$\tilde{m}(\mathbf{x},\mathbf{t}) = \mathbf{P}_n \mathbf{m}(\mathbf{x},\mathbf{t}), \quad \tilde{m}_\alpha(x_\alpha) = \mathbf{P}_{n,\alpha} \mathbf{m}(x_\alpha), \quad \tilde{\mathbf{c}}_m = \mathbf{P}_{n,c} \mathbf{m},$$

$$\tilde{\varepsilon}(\mathbf{x},\mathbf{t}) = \mathbf{P}_n \mathbf{E}, \quad \tilde{\varepsilon}_\alpha(x_\alpha) = \mathbf{P}_{n,\alpha} \mathbf{E}(x_\alpha), \quad \tilde{\mathbf{c}}_\varepsilon = \mathbf{P}_{n,c} \mathbf{E}, \tag{A.7}$$

for $\mathbf{m} = \{m(\mathbf{X}_i,\mathbf{T}_i)\}_{i=1}^n$, $\mathbf{E} = \{\sigma(\mathbf{X}_i,\mathbf{T}_i)\varepsilon_i\}_{i=1}^n$. Linearity of $\mathbf{P}_n$, $\mathbf{P}_{n,c}$, $\mathbf{P}_{n,\alpha}$, $1 \le \alpha \le d_2$, and the relation $\mathbf{Y}=\mathbf{m}+\mathbf{E}$ imply a crucial decomposition

$$\hat{m}(\mathbf{x},\mathrm{t}) = \tilde{m}(\mathbf{x},\mathbf{t}) + \tilde{\varepsilon}(\mathbf{x},\mathbf{t}), \quad \hat{m}_\alpha(x_\alpha) = \tilde{m}_\alpha(x_\alpha) + \tilde{\varepsilon}_\alpha(x_\alpha), \quad \hat{\mathbf{c}} = \tilde{\mathbf{c}}_m + \tilde{\mathbf{c}}_\varepsilon. \tag{A.8}$$

Let $\tilde{\mathbf{a}} = (\tilde{a}_0, \tilde{a}_l, \tilde{a}_{J,\alpha})_{1 \le l \le d_1, 1 \le J \le N+1, 1 \le \alpha \le d_2}^T$ be the minimizer of

$$\sum_{i=1}^{n} \left\{ \sigma(\mathbf{X}_i,\mathbf{T}_i)\varepsilon_i - a_0 - \sum_{l=1}^{d_1} a_l T_{il} - \sum_{\alpha=1}^{d_2} \sum_{J=1}^{N+1} a_{J,\alpha} B_{J,\alpha}(X_{i\alpha}) \right\}^2. \tag{A.9}$$

Similarly, $\tilde{\mathbf{c}} = (\tilde{c}_{00}, \tilde{c}_{0l}, \tilde{c}_{J,\alpha})_{1 \le l \le d_1, 1 \le J \le N+1, 1 \le \alpha \le d_2}^T$ minimizes

$$\sum_{i=1}^{n} \left\{ m(\mathbf{X}_i,\mathbf{T}_i) - c_0 - \sum_{l=1}^{d_1} c_l T_{il} - \sum_{\alpha=1}^{d_2} \sum_{J=1}^{N+1} c_{J,\alpha} B_{J,\alpha}(X_{i\alpha}) \right\}^2. \tag{A.10}$$

Then $\tilde{\varepsilon}(\mathbf{x},\mathbf{t}) = \tilde{\mathbf{a}}^T \mathbf{B}(\mathbf{x},\mathbf{t})$, $\tilde{\mathbf{a}} = (\mathbf{B}^T \mathbf{B})^{-1} \mathbf{B}^T \mathbf{E}$, with matrices

$$\mathbf{B}(\mathbf{x},\mathbf{t}) = \{1, t_l, B_{J,\alpha}(x_\alpha)\}_{1 \le l \le d_1, 1 \le J \le N+1, 1 \le \alpha \le d_2}^T, \quad \mathbf{B} = \{\mathbf{B}(\mathbf{X}_i,\mathbf{T}_i)\}_{1 \le i \le n}^T, \tag{A.11}$$

and $\tilde{\mathbf{a}}$, the solution of (A.9), equals to

$$\left\{ \begin{matrix} 1 & E_n T_l & E_n B_{J,\alpha} \\ (E_n T_{l'})^T & \langle T_{l'}, T_l \rangle_n & \langle T_{l'}, B_{J,\alpha} \rangle_n \\ (E_n B_{J',\alpha'})^T & \langle B_{J',\alpha'}, T_l \rangle_n & \langle B_{J',\alpha'}, B_{J,\alpha} \rangle_n \end{matrix} \right\}_{1 \le l,l' \le d_1, 1 \le \alpha,\alpha' \le d_2, 1 \le J,J' \le N+1}^{-1} \left\{ \begin{matrix} n^{-1} \sum_{i=1}^{n} \sigma(\mathbf{X}_i,\mathbf{T}_i)\varepsilon_i \\ n^{-1} \sum_{i=1}^{n} T_{il} \sigma(\mathbf{X}_i,\mathbf{T}_i)\varepsilon_i \\ n^{-1} \sum_{i=1}^{n} B_{J,\alpha}(X_{i\alpha}) \sigma(\mathbf{X}_i,\mathbf{T}_i)\varepsilon_i \end{matrix} \right\}_{1 \le l \le d_1, 1 \le J \le N+1, 1 \le \alpha \le d_2}. \tag{A.12}$$

Bernstein inequality below under geometric $\alpha$-mixing is used in many proofs.

**Lemma A.1** (Bosq, 1998, Theorem 1.4, p. 31). *Let $\{\xi_t, t \in \mathbf{Z}\}$ be a zero mean real valued $\alpha$-mixing process, $S_n = \sum_{t=1}^{n} \xi_t$. Suppose there exists $c > 0$ such that for $t = 1, \ldots, n$, $k = 3, 4, \ldots, E|\xi_t|^k \le c^{k-2} k! E\xi_t^2 < +\infty$ (Cramér's condition) then for $n > 1$, integer*

*$q \in [1,n/2]$, $\varepsilon > 0$ and $k \ge 3$*

$$P(|S_n| \ge n\varepsilon) \le a_1 \exp\left(-\frac{q\varepsilon^2}{25m_2^2+5c\varepsilon}\right)+a_2(k)\alpha\left(\left[\frac{n}{q+1}\right]\right)^{2k/(2k+1)},$$

*where $\alpha(\cdot)$ is the $\alpha$-mixing coefficient defined in* (A.1) *and*

$$a_1 = 2\frac{n}{q}+2\left(1+\frac{\varepsilon^2}{25m_2^2+5c\varepsilon}\right), \quad a_2(k) = 11n\left(1+\frac{5m_k^{2k/(2k+1)}}{\varepsilon}\right),$$

*with $m_r = \max_{1 \le i \le n}\|\xi_t\|_r$, $r \ge 2$.*

**Lemma A.2.** *Under Assumptions* (A2), (A4), (A6), *$c_f H/2 \le c_{J,\alpha} \le C_f H$ and* (i) *$\exists$ constants $c_0(f)$, $C_0(f)$ depending on $f_\alpha(x_\alpha)$, $1 \le \alpha \le d_2$, such that $c_0(f)H \le \|b_{J,\alpha}\|^2 \le C_0(f)H$.* (ii) *Uniformly for $1 \le J,J' \le N+1, 1 \le \alpha,\alpha' \le d_2, E|B_{J,\alpha}(X_\alpha)| \le CH^{1/2}$, $E|B_{J,\alpha}(X_{i\alpha})B_{J',\alpha'}(X_{i\alpha'})|^2 \ge c_f C_f^{-2} > 0$, $E|B_{J,\alpha}(X_{i\alpha})B_{J',\alpha'}(X_{i\alpha'})|^k \le C^k H^{2-k}$, $k \ge 1$.* (iii) *Uniformly for $1 \le J, J' \le N+1, 1 \le \alpha \le d_2$,*

$$E\{B_{J,\alpha}(X_{i\alpha})B_{J',\alpha}(X_{i\alpha})\} \sim \begin{cases} 1, & J'=J, \\ -1/3, & |J'-J|=1, \\ 1/6, & |J'-J|=2, \end{cases}$$

$$E|B_{J,\alpha}(X_{i\alpha})B_{J',\alpha}(X_{i\alpha})|^k \begin{cases} \le C^k H^{1-k}, & |J'-J| \le 2, \\ 0, & |J'-J| > 2, \end{cases} \quad k \ge 1.$$

**Lemma A.3.** *Under Assumptions* (A2), (A4), (A6) *and* (A7), *denote*

$$\omega_{J,\alpha}(\mathbf{X}_l,x_1) = K_h(X_{l1}-x_1)B_{J,\alpha}(X_{l\alpha}), \quad \mu_{J,\alpha}(x_1) = E\omega_{J,\alpha}(\mathbf{X}_l,x_1), \tag{A.13}$$

*as $n \to \infty$, $\sup_{x_1 \in [0,1]}\sup_{1 \le \alpha \le d_2, 1 \le J \le N+1}|\mu_{J,\alpha}(x_1)| = O(\sqrt{H})$.*

**Lemma A.4.** *Under Assumptions* (A2), (A4), (A6) *and* (A7), *as $n \to \infty$,*

$$\sup_{x_1 \in [0,1]}\sup_{1 \le \alpha \le d_2, 1 \le J \le N+1}\left|n^{-1}\sum_{l=1}^{n}\{\omega_{J,\alpha}(\mathbf{X}_l,x_1)-\mu_{J,\alpha}(x_1)\}\right| = O_p\{\log n/\sqrt{nh}\},$$

*where $\omega_{J,\alpha}(\mathbf{X}_l,x_1)$ and $\mu_{J,\alpha}(x_1)$ are given in* (A.13).

**Lemma A.5.** *Under Assumptions* (A2), (A4)–(A6), *as $n \to \infty$,*

$$A_{n,1} = \sup_{J,\alpha}|E_n B_{J,\alpha}| = O_p(n^{-1/2}\log n), \tag{A.14}$$

$$A_{n,2} = \sup_{J,J',\alpha}|\langle B_{J,\alpha},B_{J',\alpha}\rangle_n - \langle B_{J,\alpha},B_{J',\alpha}\rangle| = O_p\{(nH)^{-1/2}\log n\}, \tag{A.15}$$

$$A_{n,3} = \sup_{\alpha \ne \alpha'}|\langle B_{J,\alpha},B_{J',\alpha'}\rangle_n - \langle B_{J,\alpha},B_{J',\alpha'}\rangle| = O_p(n^{-1/2}\log n), \tag{A.16}$$

$$A_{n,4} = \sup_{l,J,\alpha}|\langle T_l,B_{J,\alpha}\rangle_n - \langle T_l,B_{J,\alpha}\rangle| = O_p(n^{-1/2}\log n). \tag{A.17}$$

**Lemma A.6.** *Under Assumptions* (A2), (A4)–(A7), *denote*

$$\zeta_l(X_{i1},T_{il},x_1) = K_h(X_{i1}-x_1)T_{il}, \quad \mu_l(x_1) = E\zeta_l(X_{i1},T_{il},x_1), \tag{A.18}$$

*as $n \to \infty$, $\sup_{1 \le l \le d_1, x_1 \in [0,1]}|\mu_l(x_1)| = O(1)$, while*

$$\sup_{1 \le l \le d_1, x_1 \in [0,1]}\left|n^{-1}\sum_{i=1}^{n}\{\zeta_l(X_{i1},T_{il},x_1)-\mu_l(x_1)\}\right| = O_p\{n^{-1/2}h^{-1/2}\log n\}.$$

For proofs of Lemmas A.2–A.6, see Ma and Yang (2009). Let $(\nu_{(J',\alpha'),(J,\alpha)}) = \langle B_{J',\alpha'},B_{J,\alpha}\rangle, (\nu_{l',l}) = \langle T_{l'},T_l\rangle, (\nu_{l',(J,\alpha)}) = \langle T_{l'},B_{J,\alpha}\rangle$, $(\nu_{(J',\alpha'),l}) = \langle B_{J',\alpha'},T_l\rangle$. Denote by $\mathbf{V}$ the theoretical inner product matrix of the standardized basis $\{1,t_l,B_{J,\alpha}(x_\alpha), 1 \le l \le d_1, 1 \le J \le N+1, 1 \le \alpha \le d_2\}$, i.e.

$$\mathbf{V} = \begin{pmatrix} 1 & \mathbf{0}_{d_1}^T & \mathbf{0}_{d_2(N+1)}^T \\ \mathbf{0}_{d_1} & (\nu_{l',l}) & (\nu_{l',(J,\alpha)}) \\ \mathbf{0}_{d_2(N+1)} & (\nu_{(J',\alpha'),l}) & (\nu_{(J',\alpha'),(J,\alpha)}) \end{pmatrix}_{1 \le l,l' \le d_1, 1 \le \alpha \le \alpha' \le d_2, 1 \le J,J' \le N+1}. \tag{A.19}$$

Denote by $\mathbf{S}$ the inverse matrix of $\mathbf{V}$

$$\mathbf{S}=\begin{pmatrix} 1 & \mathbf{0}_{d_1}^T & \mathbf{0}_{d_2(N+1)}^T \\ \mathbf{0}_{d_1} & (s_{l',l}) & (s_{l',(J,\alpha)}) \\ \mathbf{0}_{d_2(N+1)} & (s_{(J',\alpha'),l}) & (s_{(J',\alpha'),(J,\alpha)}) \end{pmatrix}_{1\le l,l'\le d_1, 1\le\alpha\le\alpha'\le d_2, 1\le J,J'\le N+1}. \tag{A.20}$$

Next, we denote by $\hat{\mathbf{V}}$ the empirical version of $\mathbf{V}$, i.e.

$$\hat{\mathbf{V}}=\begin{pmatrix} 0 & E_nT_l & E_nB_{J,\alpha} \\ (E_nT_{l'})^T & \langle T_{l'},T_l\rangle_n & \langle T_{l'},B_{J,\alpha}\rangle_n \\ (E_nB_{J',\alpha'})^T & \langle B_{J',\alpha'},T_l\rangle_n & \langle B_{J',\alpha'},B_{J,\alpha}\rangle_n \end{pmatrix}_{1\le l,l'\le d_1, 1\le\alpha\le\alpha'\le d_2, 1\le J,J'\le N+1}. \tag{A.21}$$

**Lemma A.7.** *Under Assumptions* (A2), (A4)–(A7), *for matrices* $\mathbf{V}$, $\mathbf{S}$ and $\hat{\mathbf{V}}$ *defined in* (A.19), (A.20) *and* (A.21) (i) *there exist constants* $C_V>c_V>0$, $C_S=c_V^{-1}$, $c_S=C_V^{-1}$ *such that*

$$c_V\mathbf{I}_{1+d_1+d_2(N+1)}\le\mathbf{V}\le C_V\mathbf{I}_{1+d_1+d_2(N+1)},$$

$$c_S\mathbf{I}_{1+d_1+d_2(N+1)}\le\mathbf{S}\le C_S\mathbf{I}_{1+d_1+d_2(N+1)}. \tag{A.22}$$

(ii) *Define* $A_n=\sup_{g_1,g_2\in G}|\langle g_1,g_2\rangle_n-\langle g_1,g_2\rangle|\|g_1\|^{-1}\|g_2\|^{-1}$, *then* $A_n=O_p(n^{-1/2}H^{-1}\log n)$. (iii) *With probability approaching* 1 *as* $n\to\infty$,

$$c_V\mathbf{I}_{1+d_1+d_2(N+1)}\le\hat{\mathbf{V}}\le C_V\mathbf{I}_{1+d_1+d_2(N+1)},$$

$$c_S\mathbf{I}_{1+d_1+d_2(N+1)}\le\hat{\mathbf{V}}^{-1}\le C_S\mathbf{I}_{1+d_1+d_2(N+1)}. \tag{A.23}$$

**Lemma A.8.** *Under Assumptions* (A2)–(A7), *as* $n\to\infty$,

$$\left|n^{-1}\sum_{i=1}^n\sigma(\mathbf{X}_i,\mathbf{T}_i)\varepsilon_i\right|+\max_{J,\alpha}\left|n^{-1}\sum_{i=1}^nB_{J,\alpha}(X_{i\alpha})\sigma(\mathbf{X}_i,\mathbf{T}_i)\varepsilon_i\right|+\max_l\left|n^{-1}\sum_{i=1}^nT_{il}\sigma(\mathbf{X}_i,\mathbf{T}_i)\varepsilon_i\right|=O_p(n^{-1/2}\log n).$$

For proofs of Lemmas A.7 and A.8, see Ma and Yang (2009).

**Corollary A.1.** *Under Assumptions* (A2)–(A7), *as* $n\to\infty$, $\|n^{-1}\mathbf{B}^T\mathbf{E}\|=O_p(n^{-1/2}N^{1/2}\log n)$, $\|n^{-1}\mathbf{B}^T\mathbf{E}\|_\infty=O_p(n^{-1/2}\log n)$.

Corollary A.1 follows from Lemma A.8 directly.

Next we study the difference between $\hat{m}_{\mathrm{SBK}}(x_1)$ and $\tilde{m}_{\mathrm{K},1}(x_1)$, both given in (4).

Denote $\mathbf{c}=\{c_{0l}\}_{l=0}^{d_1}$, the decomposition (A.8) implies that $\tilde{m}_{\mathrm{K},1}(x_1)-\hat{m}_{\mathrm{SBK}}(x_1)=\{\Psi_{Tb}(x_1)+\Psi_{Tv}(x_1)+\Psi_b(x_1)+\Psi_v(x_1)\}/\hat{f}_1(x_1)$, where

$$\Psi_{Tb}(x_1)=n^{-1}\sum_{i=1}^nK_h(X_{i1}-x_1)(1,\mathbf{T}_i^T)(\tilde{\mathbf{c}}_m-\mathbf{c}),$$

$$\Psi_{Tv}(x_1)=n^{-1}\sum_{i=1}^nK_h(X_{i1}-x_1)(1,\mathbf{T}_i^T)\tilde{\mathbf{c}}_\varepsilon, \tag{A.24}$$

$$\Psi_b(x_1)=n^{-1}\sum_{i=1}^nK_h(X_{i1}-x_1)\sum_{\alpha=2}^{d_2}\{\tilde{m}_\alpha(X_{i\alpha})-m_\alpha(X_{i\alpha})\},$$

$$\Psi_v(x_1)=n^{-1}\sum_{i=1}^nK_h(X_{i1}-x_1)\sum_{\alpha=2}^{d_2}\tilde{\varepsilon}_\alpha(X_{i\alpha}). \tag{A.25}$$

First we show $\Psi_b(x_1)$ is uniformly of order $O_p(n^{-1/2})$ for $x_1\in[0,1]$.

**Lemma A.9.** *Under Assumptions* (A1), (A2), (A4)–(A7), *as* $n\to\infty$,

$$\sup_{x_1\in[0,1]}|\Psi_b(x_1)|=O_p(n^{-1/2}+H^2)=O_p(n^{-1/2}).$$

**Lemma A.10.** *Under Assumptions* (A1), (A2) *and* (A6), *there exist functions* $g_\alpha\in G$, $1\le\alpha\le d_2$, *such that as* $n\to\infty$, $\|\tilde{m}-g+\sum_{\alpha=1}^{d_2}E_ng_\alpha(X_\alpha)\|_n=O_p(n^{-1/2}+H^2)$, *where* $g(\mathbf{x},\mathbf{t})=c_{00}+\sum_{l=1}^{d_1}c_{0l}t_l+\sum_{\alpha=1}^{d_2}g_\alpha(x_\alpha)$ *and* $\tilde{m}$ *is defined in* (A.7).

For proofs of Lemmas A.10 and A.9, see Ma and Yang (2009). Next we show $\Psi_v(x_1)$ in (A.25) is uniformly of order $o_p(n^{-2/5})$. For $\tilde{a}_{J,\alpha}$ given in (A.12), define an auxiliary entity

$$\tilde{\varepsilon}_\alpha^* = \sum_{J=1}^{N+1} \tilde{a}_{J,\alpha} B_{J,\alpha}(x_\alpha). \tag{A.26}$$

The $\tilde{\varepsilon}_\alpha(x_\alpha)$ in (A.7) is the empirical centering of $\tilde{\varepsilon}_2^*(x_2)$, i.e.

$$\tilde{\varepsilon}_\alpha(x_\alpha) \equiv \tilde{\varepsilon}_\alpha^*(x_\alpha) - n^{-1} \sum_{i=1}^{n} \tilde{\varepsilon}_\alpha^*(X_{i\alpha}). \tag{A.27}$$

According to (A.27), we can write $\Psi_v(x_1) = \Psi_v^{(2)}(x_1) - \Psi_v^{(1)}(x_1)$, where

$$\Psi_v^{(1)}(x_1) = n^{-1} \sum_{l=1}^{n} K_h(X_{l1} - x_1) \sum_{\alpha=2}^{d_2} n^{-1} \sum_{i=1}^{n} \tilde{\varepsilon}_\alpha^*(X_{i\alpha}), \tag{A.28}$$

$$\Psi_v^{(2)}(x_1) = n^{-1} \sum_{l=1}^{n} K_h(X_{l1} - x_1) \sum_{\alpha=2}^{d_2} \tilde{\varepsilon}_\alpha^*(X_{l\alpha}), \tag{A.29}$$

for $\tilde{\varepsilon}_\alpha^*(X_{i\alpha})$ in (A.26). By (A.12) and (A.26), $\Psi_v^{(2)}(x_1)$ can be rewritten as

$$\Psi_v^{(2)}(x_1) = n^{-1} \sum_{\alpha=2}^{d_2} \sum_{l=1}^{n} \sum_{J=1}^{N+1} \tilde{a}_{J,\alpha} \omega_J(\mathbf{X}_l, x_1), \tag{A.30}$$

for $\omega_{J,\alpha}(\mathbf{X}_l, x_1)$ given in (A.13). $\Psi_v^{(1)}(x_1)$ and $\Psi_v^{(2)}(x_1)$ are of order $O_p\{Nn^{-1}(\log n)^2\}$ and $O_p(n^{-1/2}\log n)$ uniformly, given in Lemmas A.12 and A.13. The next lemma provides the size of $\tilde{\mathbf{a}}^T\tilde{\mathbf{a}}$, where $\tilde{\mathbf{a}}$ is the least square solution defined by (A.9).

**Lemma A.11.** *Under Assumptions* (A2)–(A6), *as* $n \to \infty$,

$$\tilde{\mathbf{a}}^T\tilde{\mathbf{a}} = \tilde{a}_0^2 + \sum_{l=1}^{d_1} \tilde{a}_l^2 + \sum_{\alpha=1}^{d_2} \sum_{J=1}^{N+1} \tilde{a}_{J,\alpha}^2 = O_p\{Nn^{-1}(\log n)^2\}. \tag{A.31}$$

**Proof.** By (A.11), (A.12), (A.23), with probability approaching 1 as $n \to \infty$, $\|\tilde{\mathbf{a}}\| \|n^{-1}\mathbf{B}^T\mathbf{E}\| \ge \tilde{\mathbf{a}}^T(n^{-1}\mathbf{B}^T\mathbf{E}) = \tilde{\mathbf{a}}^T\hat{\mathbf{V}}\tilde{\mathbf{a}} \ge c_V \|\tilde{\mathbf{a}}\|^2$ with Corollary A.1 imply $\|\tilde{\mathbf{a}}\|^2 \le c_V^{-1} \|\tilde{\mathbf{a}}\| \|n^{-1}\mathbf{B}^T\mathbf{E}\| = c_V^{-1}\|\tilde{\mathbf{a}}\| \times O_p\{N^{1/2}n^{-1/2}\log n\}$. Thus $\|\tilde{\mathbf{a}}\| = O_p\{N^{1/2}n^{-1/2}\log n\}$, $n \to \infty$. □

**Lemma A.12.** *Under Assumptions* (A2)–(A7), *as* $n \to \infty$, $\Psi_v^{(1)}(x_1)$ *in* (A.28) *satisfies* $\sup_{x_1 \in [0,1]} |\Psi_v^{(1)}(x_1)| = O_p\{Nn^{-1}(\log n)^2\}$.

For proof, see Ma and Yang (2009). The vector $\tilde{\mathbf{a}}$ in (A.12) is

$$\tilde{\mathbf{a}} = (\hat{\mathbf{V}})^{-1}(n^{-1}\mathbf{B}^T\mathbf{E}). \tag{A.32}$$

Next define theoretical versions $\hat{\mathbf{a}}$ of $\tilde{\mathbf{a}}$ and $\hat{\Psi}_v^{(2)}(x_1)$ of $\Psi_v^{(2)}(x_1)$ in (A.30) as

$$\hat{\mathbf{a}} = \mathbf{V}^{-1}(n^{-1}\mathbf{B}^T\mathbf{E}) = \mathbf{S}(n^{-1}\mathbf{B}^T\mathbf{E}), \tag{A.33}$$

$$\hat{\Psi}_v^{(2)}(x_1) = n^{-1} \sum_{i=1}^{n} \sum_{\alpha=1}^{d_2} \sum_{J=1}^{N+1} \hat{a}_{J,\alpha} \omega_{J,\alpha}(\mathbf{X}_i, x_1). \tag{A.34}$$

**Lemma A.13.** *Under Assumptions* (A2)–(A7), *as* $n \to \infty$, $\Psi_v^{(2)}(x_1)$ *in* (A.29) *satisfies* $\sup_{x_1 \in [0,1]} |\Psi_v^{(2)}(x_1)| = O_p(n^{-1/2}\log n) = o_p(n^{-2/5})$.

**Proof.** According to (A.32) and (A.33), one has $\mathbf{V}\hat{\mathbf{a}} = (\hat{\mathbf{V}})\tilde{\mathbf{a}}$, which implies that $(\hat{\mathbf{V}} - \mathbf{V})\tilde{\mathbf{a}} = \mathbf{V}(\hat{\mathbf{a}} - \tilde{\mathbf{a}})$. Using (A.14)–(A.17) one obtains that $\|\mathbf{V}(\hat{\mathbf{a}} - \tilde{\mathbf{a}})\| = \|(\hat{\mathbf{V}} - \mathbf{V})\tilde{\mathbf{a}}\| \le O_p(n^{-1/2}H^{-1}\log n)\|\tilde{\mathbf{a}}\|$. According to Lemma A.11, $\|\tilde{\mathbf{a}}\| = O_p(n^{-1/2}N^{1/2}\log n)$, so as $n \to \infty$, $\|\mathbf{V}(\hat{\mathbf{a}} - \tilde{\mathbf{a}})\| \le O_p\{n^{-1}N^{3/2}(\log n)^2\}$. By Lemmas A.7 and A.11, as $n \to \infty$,

$$\|\hat{\mathbf{a}} - \tilde{\mathbf{a}}\| = O_p\{n^{-1}N^{3/2}(\log n)^2\}, \tag{A.35}$$

$$\|\hat{\mathbf{a}}\| \le \|\hat{\mathbf{a}} - \tilde{\mathbf{a}}\| + \|\tilde{\mathbf{a}}\| = O_p(n^{-1/2}N^{1/2}\log n), \tag{A.36}$$

$$\begin{aligned} \sup_{x_1 \in [0,1]} |\Psi_v^{(2)}(x_1) - \hat{\Psi}_v^{(2)}(x_1)| &\le \sup_{x_1 \in [0,1]} \left| \sum_{\alpha=2}^{d_2} \sum_{J=1}^{N+1} (\tilde{a}_{J,\alpha} - \hat{a}_{J,\alpha}) n^{-1} \sum_{l=1}^{n} \omega_J(\mathbf{X}_l, x_1) \right| \\ &= \sqrt{d_2(N+1)}\, O_p(n^{-1}H^{-3/2}\log^2 n) O_p(H^{1/2}) = O_p(n^{-1}H^{-3/2}\log^2 n). \end{aligned} \tag{A.37}$$

Note that $|\hat{\Psi}_{v}^{(2)}(x_1)| \le Q_1(x_1) + Q_2(x_1)$, where $Q_1(x_1) = |\sum_{\alpha=2}^{d_2}\sum_{J=1}^{N+1}\hat{a}_{J,\alpha}\mu_{J,\alpha}(x_1)|$,

$$Q_2(x_1) = \left|\sum_{\alpha=2}^{d_2}\sum_{J=1}^{N+1}\hat{a}_{J,\alpha}n^{-1}\sum_{i=1}^{n}\{\omega_{J,\alpha}(\mathbf{X}_i,x_1)-\mu_{J,\alpha}(x_1)\}\right|.$$

Using the discretization idea, we divide interval [0,1] into $M_n \sim n$ equally spaced intervals with disjoint endpoints $0 = x_{1,0} < \cdots < x_{1,M_n} = 1$, then $\sup_{x_1 \in [0,1]} Q_1(x_1) \le T_1 + T_2$, where $T_1 = \max_{1 \le k \le M_n}|\sum_{\alpha=2}^{d_2}\sum_{J=1}^{N+1}\hat{a}_{J,\alpha}\mu_{J,\alpha}(x_{1,k})|$,

$$T_2 = \sup_{x_1 \in [x_{1,k-1},x_{1,k}], 1 \le k \le M_n}\left|\sum_{\alpha=2}^{d_2}\sum_{J=1}^{N+1}\{\hat{a}_{J,\alpha}\mu_{J,\alpha}(x_1)-\hat{a}_{J,\alpha}\mu_{J,\alpha}(x_{1,k})\}\right|.$$

$$\hat{a}_{J,\alpha} = \frac{1}{n}\sum_{i=1}^{n}\sigma(\mathbf{X}_i,\mathbf{T}_i)\varepsilon_i\left\{\sum_{l=1}^{d_1}T_{il}s_{(J,\alpha),l} + \sum_{\alpha'=1}^{d_2}\sum_{J=1}^{N+1}B_{J',\alpha'}(X_{i\alpha'})s_{(J,\alpha),(J',\alpha')}\right\},$$

$$\sum_{J=1}^{N+1}\hat{a}_{J,\alpha}\mu_{J,\alpha}(x_{1,k}) = n^{-1}\sum_{i,l,J}\mu_{J,\alpha}(x_{1,k})s_{(J,\alpha),l}T_{il}\sigma(\mathbf{X}_i,\mathbf{T}_i)\varepsilon_i + n^{-1}\sum_{i,\alpha',J,J'}\mu_{J,\alpha}(x_{1,k})s_{(J,\alpha),(J',\alpha')}B_{J',\alpha'}(X_{i\alpha'})\sigma(\mathbf{X}_i,\mathbf{T}_i)\varepsilon_i.$$

Define next $W_{\alpha,l} = \max_{1 \le k \le M_n}|n^{-1}\sum_{i=1}^{n}\sum_{J=1}^{N+1}\mu_{J,\alpha}(x_{1,k})s_{(J,\alpha),l}T_{il}\sigma(\mathbf{X}_i,\mathbf{T}_i)\varepsilon_i|$, $W_{\alpha,\alpha'} = \max_{1 \le k \le M_n}|n^{-1}\sum_i\sum_{J,J'}\mu_{J,\alpha}(x_{1,k})\ s_{(J,\alpha),(J',\alpha')}$ $B_{J',\alpha'}\ (X_{i\alpha'})\sigma(\mathbf{X}_i,\mathbf{T}_i)\varepsilon_i|$. So $T_1 \le \sum_{\alpha=1}^{d_2}(\sum_{l=1}^{d_1}W_{\alpha,l} + \sum_{\alpha'=1}^{d_2}W_{\alpha,\alpha'})$. $\sup_{\alpha,l}W_{\alpha,l}$ is bounded by

$$\sup_l\left|n^{-1}\sum_{i=1}^{n}T_{il}\sigma(\mathbf{X}_i,\mathbf{T}_i)\varepsilon_i\right|\max_{1 \le k \le M_n}\sup_{\alpha,l}\left|\sum_{J=1}^{N+1}\mu_{J,\alpha}(x_{1,k})s_{(J,\alpha),l}\right| = O_p(n^{-1/2}\log n)O_p(1) = O_p(n^{-1/2}\log n),$$

which follows from Lemmas A.8 and A.3. Let $D_n = n^{\theta_0}$, $(2+\delta)^{-1} < \theta_0 < 3/8$, where $\delta$ is the same as in Assumption (A3). Define

$$\varepsilon_{i,D}^{-} = \varepsilon_i I(|\varepsilon_i| \le D_n), \quad \varepsilon_{i,D}^{+} = \varepsilon_i I(|\varepsilon_i| > D_n), \quad \varepsilon_{i,D}^{*} = \varepsilon_{i,D}^{-} - E(\varepsilon_{i,D}^{-}|\mathbf{X}_i,\mathbf{T}_i),$$

$$U_{i,k} = \boldsymbol{\mu}_\alpha(x_{1,k})^T\{s_{(J,\alpha),(J',\alpha')}\}_{1 \le J,J' \le N+1}\{B_{J',\alpha'}(X_{i\alpha'})\}_{J'}^T\sigma(\mathbf{X}_i,\mathbf{T}_i)\varepsilon_{i,D}^{*}.$$

Denote $W_{\alpha,\alpha'}^D = \max_{1 \le k \le M_n}|n^{-1}\sum_{i=1}^{n}U_{i,k}|$ as truncated centered version of $W_{\alpha,\alpha'}$. To show $|W_{\alpha,\alpha'} - W_{\alpha,\alpha'}^D| = U_p(n^{-1/2}\log n)$, note $|W_{\alpha,\alpha'} - W_{\alpha,\alpha'}^D| \le \Lambda_1 + \Lambda_2$,

$$\Lambda_1 = \max_{1 \le k \le M_n}\left|\frac{1}{n}\sum_{i,J,J'}\mu_{J,\alpha}(x_{1,k})s_{(J,\alpha),(J',\alpha')}B_{J',\alpha'}(X_{i\alpha'})\sigma(\mathbf{X}_i,\mathbf{T}_i)E(\varepsilon_{i,D}^{-}|\mathbf{X}_i,\mathbf{T}_i)\right|$$

$$\Lambda_2 = \max_{1 \le k \le M_n}\left|\frac{1}{n}\sum_i\sum_{J,J'}\mu_{J,\alpha}(x_{1,k})s_{(J,\alpha),(J',\alpha')}B_{J',\alpha'}(X_{i\alpha'})\sigma(\mathbf{X}_i,\mathbf{T}_i)\varepsilon_{i,D}^{+}\right|.$$

Let $\boldsymbol{\mu}_\alpha(x_{1,k}) = \{\mu_{1,\alpha}(x_{1,k}),\ldots,\mu_{N+1,\alpha}(x_{1,k})\}^T$, then $\Lambda_1$ equals to

$$\max_{1 \le k \le M_n}\left[\boldsymbol{\mu}_\alpha(x_{1,k})^T \times \{s_{(J,\alpha),(J',\alpha')}\}_{J,J'}\left\{\frac{1}{n}\sum_{i=1}^{n}B_{J',\alpha'}(X_{i\alpha'})\sigma(\mathbf{X}_i,\mathbf{T}_i)E(\varepsilon_{i,D}^{-}|\mathbf{X}_i,\mathbf{T}_i)\right\}_{J'}\right] \le \max_{1 \le k \le M_n}\left[\boldsymbol{\mu}_\alpha(x_{1,k})^T\boldsymbol{\mu}_\alpha(x_{1,k})\left[\{s_{(J,\alpha),(J',\alpha')}\}_{1 \le J,J' \le N+1}\right.\right.$$

$$\left.\left.\times\left\{n^{-1}\sum_{i=1}^{n}B_{J',\alpha'}(X_{i\alpha'})\sigma(\mathbf{X}_i,\mathbf{T}_i)E(\varepsilon_{i,D}^{-}|\mathbf{X}_i,\mathbf{T}_i)\right\}_{J'}\right]^T\left[\{s_{(J,\alpha),(J',\alpha')}\}_{J,J'}\left\{n^{-1}\sum_{i=1}^{n}B_{J',\alpha'}(X_{i\alpha'})\sigma(\mathbf{X}_i,\mathbf{T}_i)E(\varepsilon_{i,D}^{-}|\mathbf{X}_i,\mathbf{T}_i)\right\}_{J'=1}^{N+1}\right]\right]^{1/2}$$

$$\le C_S'\max_k\left\{\sum_{J,J'}\mu_{J,\alpha}^2(x_{1,k})\left\{\frac{1}{n}\sum_i B_{J',\alpha'}(X_{i\alpha'})\sigma(\mathbf{X}_i,\mathbf{T}_i)E(\varepsilon_{i,D}^{-}|\mathbf{X}_i,\mathbf{T}_i)\right\}^2\right\}^{1/2}.$$

By Assumption (A3), $|E(\varepsilon_{i,D}^{-}|\mathbf{X}_i,\mathbf{T}_i)| = |E(\varepsilon_{i,D}^{+}|\mathbf{X}_i,\mathbf{T}_i)| \le M_\delta D_n^{-(1+\delta)}$. By Lemma A.1, $\sup_{J',\alpha'}|n^{-1}\sum_{i=1}^{n}B_{J',\alpha'}(X_{i\alpha'})\sigma(\mathbf{X}_i,\mathbf{T}_i)| = O_p(n^{-1/2}\log n)$. So

$$\sup_{\alpha,\alpha'}\Lambda_1 \le C_S M_\delta D_n^{-(1+\delta)}N\max_k\sup_{J,\alpha}|\mu_{J,\alpha}(x_{1,k})|\sup_{J',\alpha'}\left|\frac{1}{n}\sum_{i=1}^{n}B_{J',\alpha'}(X_{i\alpha'})\sigma(\mathbf{X}_i,\mathbf{T}_i)\right| = O_p\{Nn^{-1}D_n^{-(1+\delta)}\log^2 n\} = O_p\{(\log n)^2Nn^{-3/2}\},$$

where the last step follows from the choice of $D_n$. Meanwhile

$$\sum_{n=1}^{\infty}P(|\varepsilon_n| \ge D_n) \le \sum_{n=1}^{\infty}\frac{E|\varepsilon_n|^{2+\delta}}{D_n^{2+\delta}} = \sum_{n=1}^{\infty}\frac{E(E|\varepsilon_n|^{2+\delta}|\mathbf{X}_n,\mathbf{T}_n)}{D_n^{2+\delta}} \le \sum_{n=1}^{\infty}\frac{M_\delta}{D_n^{2+\delta}} \le \infty,$$

since $\delta > 2/3$. By Borel–Cantelli lemma, for large $n$, with probability 1,

$$n^{-1}\sum_{i=1}^{n}\sum_{J',J=1}^{N+1}\mu_{J,\alpha}(x_{1,k})s_{(J,\alpha),(J',\alpha')}B_{J',\alpha'}(X_{i\alpha'})\sigma(\mathbf{X}_i,\mathbf{T}_i)\varepsilon^{+}_{i,D}=0,$$

$\sup_{\alpha,\alpha'}|W_{\alpha,\alpha'}-W^D_{\alpha,\alpha'}|\le \sup_{\alpha,\alpha'}\Lambda_1+\sup_{\alpha,\alpha'}\Lambda_2=\mathrm{O_p}((\log n)^2 N n^{-3/2})$. Next we show that $W^D_{\alpha,\alpha'}=U_p(n^{-1/2}\log n)$. The variance of $U_{i,k}$ is

$$\boldsymbol{\mu}_\alpha(x_{1,k})^T\{s_{(J,\alpha),(J',\alpha')}\}_{J,J'}\mathrm{var}(\{B_{J,\alpha'}(X_{i\alpha'})\}^T_{1\le J\le N+1}\sigma(\mathbf{X}_i,\mathbf{T}_i)\varepsilon^*_{i,D})\{s_{(J,\alpha),(J',\alpha')}\}^T_{1\le J,J'\le N+1}\boldsymbol{\mu}_\alpha(x_{1,k}).$$

By Assumption (A3), $c^2_\sigma(\nu_{(J,\alpha'),(J',\alpha')})_{1\le J,J'\le N+1}$ is bounded by

$$\mathrm{var}(\{B_{J,\alpha'}(X_{i\alpha'})\}^T_{1\le J\le N+1}\sigma(\mathbf{X}_i,\mathbf{T}_i))\le C^2_\sigma(\nu_{(J,\alpha'),(J',\alpha')})_{1\le J,J'\le N+1},$$

$$\begin{aligned}\mathrm{var}(U_{i,k})&\sim \boldsymbol{\mu}_\alpha(x_{1,k})^T\{s_{(J,\alpha),(J',\alpha')}\}_{1\le J,J'\le N+1}(\nu_{(J,\alpha'),(J',\alpha')})_{1\le J,J'\le N+1}\{s_{(J,\alpha),(J',\alpha')}\}^T_{1\le J,J'\le N+1}\boldsymbol{\mu}_\alpha(x_{1,k})V_{\varepsilon,D}\\&\sim \boldsymbol{\mu}_\alpha(x_{1,k})^T\{s_{(J,\alpha),(J',\alpha')}\}_{1\le J,J'\le N+1}\{s_{(J,\alpha),(J',\alpha')}\}^T_{1\le J,J'\le N+1}\boldsymbol{\mu}_\alpha(x_{1,k})V_{\varepsilon,D},\end{aligned}$$

where $V_{\varepsilon,D}=\mathrm{var}\{\varepsilon^*_{i,D}|\mathbf{X}_i,\mathbf{T}_i\}$. Let $\kappa(x_{1,k})=\{\boldsymbol{\mu}_\alpha(x_{1,k})^T\boldsymbol{\mu}\mu_\alpha(x_{1,k})\}^{1/2}$,

$$c'^2_S c^2_\sigma\{\kappa(x_{1,k})\}^2V_{\varepsilon,D}\le \mathrm{var}(U_{i,k})\le C'^2_S C^2_\sigma\{\kappa(x_{1,k})\}^2V_{\varepsilon,D}.$$

Since $E|U_{i,k}|^r\le\{c_0\kappa(x_{1,k})D_nH^{-1/2}\}^{r-2}r!E|U_{i,k}|^2<+\infty$, for $r\ge 3,\{U_{i,k}\}^n_{i=1}$ satisfies the Cramér's condition with Cramér's constant $c_*=c_0\kappa(x_{1,k})D_nH^{-1/2}$, hence by Lemma A.1

$$P\left\{\left|n^{-1}\sum_{l=1}^{n}U_{i,k}\right|\ge\rho_n\right\}\le a_1\exp\left(-\frac{q\rho_n^2}{25m_2^2+5c_*\rho_n}\right)+a_2(3)\alpha\left(\left[\frac{n}{q+1}\right]\right)^{6/7},$$

where $m_2^2\sim\kappa^2(x_{1,k})V_{\varepsilon,D}$, $m_3\le\{c\kappa^3(x_{1,k})H^{-1/2}D_nV_{\varepsilon,D}\}^{1/3}$, $\rho_n=\rho\log n/\sqrt{n}, a_1=2n/q+2(1+\rho_n^2/(25m_2^2+5c_*\rho_n))$, $a_2(3)=11n(1+5m_3^{6/7}/\rho_n)$. Similar as in Lemma A.4,

$$\frac{q\rho_n^2}{25m_2^2+5c_*\rho_n}\ge\frac{cn\log n^{-1}(\rho n^{-1/2}\log n)^2}{25c^*+5c_0\kappa(x_{1,k})D_nH^{-1/2}\rho n^{-1/2}\log n}\sim\log n\quad\text{as } n\to\infty.$$

Taking $c_0$, $\rho$ large enough, one has

$$P\left\{\frac{1}{n}\left|\sum_{i=1}^{n}U_{i,k}\right|>\rho n^{-1/2}\log n\right\}\le c\log n\exp\{-c_2\rho^2\log n\}+Cn^{2-6\lambda_0c_0/7}\le n^{-3},$$

for $n$ large enough. Hence $\sum_{n=1}^{\infty}P(|W^D_{\alpha,\alpha'}|\ge\rho n^{-1/2}\log n)$ is bounded by

$$\sum_{n=1}^{\infty}\sum_{k=1}^{M_n}P\left(\left|n^{-1}\sum_{i=1}^{n}U_{i,k}\right|\ge\rho n^{-1/2}N^{1/2}\log n\right)\le\sum_{n=1}^{\infty}M_nn^{-3}<\infty.$$

Thus, Borel–Cantelli lemma entails $W^D_{\alpha,\alpha'}=U_p(n^{-1/2}\log n)$, as $n\to\infty$. Note $|W_{\alpha,\alpha'}-W^D_{\alpha,\alpha'}|=U_p((\log n)^2Nn^{-3/2})$, then $W_{\alpha,\alpha'}=U_p(\log n/\sqrt{n})$. Thus $T_1\le\sum_{\alpha=1}^{d_2}(\sum_{l=1}^{d_1}W_{\alpha,l}+\sum_{\alpha'=1}^{d_2}W_{\alpha,\alpha'})$.

So as $n\to\infty$, $T_1\le d_1\mathrm{O_p}(n^{-1/2}\log n)+d_2^2\mathrm{O_p}(n^{-1/2}\log n)=\mathrm{O_p}(n^{-1/2}\log n)$.

Employing Cauchy–Schwartz inequality and Lipschitz continuity of kernel $K$, Assumption (A5), Lemma A.2(ii) and (A.36) lead to

$$T_2\le d_2\mathrm{O_p}(n^{-1/2}N^{1/2}\log n)\left\{\sum_{J=1}^{N+1}EB^2_{J,2}(X_{12})\right\}^{1/2}(h^2M_n)^{-1}=\mathrm{o_p}(n^{-1/2}).$$

Therefore, $\sup_{x_1\in[0,1]}Q_1(x_1)\le T_1+T_2=\mathrm{O_p}(n^{-1/2}\log n)$. Noting that

$$\sup_{x_1\in[0,1]}Q_2(x_1)=\mathrm{O_p}(n^{-1/2}N^{1/2}(\log n)d_2^{1/2}N^{1/2}n^{-1/2}h^{-1/2}\log n)=\mathrm{o_p}(n^{-1/2}),$$

by Cauchy–Schwartz inequality, (A.36), Lemma A.4, Assumptions (A6) and (A7). Thus $\sup_{x_1\in[0,1]}|\hat{\Psi}^{(2)}_{v}(x_1)|\le\sup_{x_1\in[0,1]}Q_1(x_1)+\sup_{x_1\in[0,1]}Q_2(x_1)=\mathrm{O_p}(n^{-1/2}\log n)=\mathrm{o_p}(n^{-2/5})$. The desired result follows from the above result and (A.37). □

**Lemma A.14.** *Under Assumptions* (A2)–(A4), (A6) *and* (A7), *as* $n\to\infty$,

$$\sup_{x_1\in[0,1]}|\Psi_v(x_1)| = O_p(n^{-1/2}\log n) = o_p(n^{-2/5}).$$

Lemma A.14 follows from Lemmas A.12 and A.13. Next we bound $\tilde{\mathbf{c}}_m^T-\mathbf{c}^T$, $\tilde{\mathbf{c}}_\varepsilon^T$ defined in (A.10) and (A.9). Denote by $\mathbf{I}_{r\times d}$ the matrix $(\mathbf{I}_r\ \mathbf{0}_{r\times d})$.

**Lemma A.15.** *Under Assumptions* (A1), (A2), (A4)–(A7), *as* $n\to\infty$, $\|\tilde{\mathbf{c}}_m-\mathbf{c}\| = o_p(n^{-1/2})$.

**Proof.** By the result on p. 149 of de Boor (2001), $\exists\, C_\infty>0$ such that for $m_\alpha\in C^1[0,1]$ with $m_\alpha'\in \mathrm{Lip}([0,1],C_\infty)$, there is a function $g_\alpha\in G$ such that $Eg_\alpha(X_\alpha)=0, \|g_\alpha-m_\alpha\|_\infty\le C_\infty\|m_\alpha'\|H^2$, $1\le\alpha\le d_2$. Then

$$\begin{aligned}\tilde{\mathbf{c}}_m-\mathbf{c} &= \mathbf{I}_{(1+d_1)\times\{d_2(N+1)\}}\times(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{m}-\mathbf{c} = \mathbf{I}_{(1+d_1)\times\{d_2(N+1)\}}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\left\{c_0+\sum_{l=1}^{d_1}c_lT_{il}+\sum_{\alpha=1}^{d_2}g_\alpha(X_{i\alpha})\right\}_{1\le i\le n}-\mathbf{c}\\ &\quad+\mathbf{I}_{(1+d_1)\times\{d_2(N+1)\}}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\left\{\sum_{\alpha=1}^{d_2}m_\alpha(X_{i\alpha})-\sum_{\alpha=1}^{d_2}g_\alpha(X_{i\alpha})\right\}_{1\le i\le n}\\ &= \mathbf{I}_{(1+d_1)\times\{d_2(N+1)\}}\hat{\mathbf{V}}^{-1}n^{-1}\mathbf{B}^T\left[\sum_{\alpha=1}^{d_2}\{m_\alpha(X_{i\alpha})-g_\alpha(X_{i\alpha})\}\right]_{1\le i\le n}\end{aligned}$$

with $\hat{\mathbf{V}}$ defined in (A.21). So by Lemma A.7, $\|\tilde{\mathbf{c}}_m-\mathbf{c}\|^2$ is bounded with probability approaching 1 by

$$\begin{aligned}&\left\|\hat{\mathbf{V}}^{-1}n^{-1}\mathbf{B}^T\left[\sum_{\alpha=1}^{d_2}\{m_\alpha(X_{i\alpha})-g_\alpha(X_{i\alpha})\}\right]_{1\le i\le n}\right\|^2\le C_S^2\left\|n^{-1}\mathbf{B}^T\left[\sum_{\alpha=1}^{d_2}\{m_\alpha(X_{i\alpha})-g_\alpha(X_{i\alpha})\}\right]_{1\le i\le n}\right\|^2\\ &\le C_S^2\left(\sum_{\alpha=1}^{d_2}\|g_\alpha-m_\alpha\|_\infty\right)^2+C_S^2\sum_{l=1}^{d_1}\left(\sum_{\alpha=1}^{d_2}\|g_\alpha-m_\alpha\|_\infty n^{-1}\sum_{i=1}^n|T_{il}|\right)^2+C_S^2\sum_{\alpha'=1}^{d_2}\sum_{J=1}^{N+1}\left(\sum_{\alpha=1}^{d_2}\|g_\alpha-m_\alpha\|_\infty n^{-1}\sum_{i=1}^n|B_{J,\alpha'}(X_{i\alpha'})|\right)\\ &\le C_S^2\left(\sum_{\alpha=1}^{d_2}\|g_\alpha-m_\alpha\|_\infty\right)^2\left\{1+d_1\left(\sup_{1\le l\le d_1}n^{-1}\sum_{i=1}^n|T_{il}|\right)^2+(N+1)d_2\left(\sup_{1\le\alpha'\le d_2,1\le J\le N+1}n^{-1}\sum_{i=1}^n|B_{J,\alpha'}(X_{i\alpha'})|\right)^2\right\}.\end{aligned}$$

$\sup_{1\le\alpha'\le d_2,1\le J\le N+1}n^{-1}\sum_{i=1}^n|B_{J,\alpha'}(X_{i\alpha'})| = O_{a.s.}(H^{1/2})$, $\sup_{1\le l\le d_1}n^{-1}\sum_{i=1}^n|T_{il}| = O_{a.s.}(1)$, so $\|\tilde{\mathbf{c}}_m-\mathbf{c}\|^2\sim[\sum_{\alpha=1}^{d_2}\|g_\alpha-m_\alpha\|_\infty]^2 = O_p(H^4)$, as $n\to\infty$. □

**Lemma A.16.** *Under Assumptions* (A1)–(A7), *as* $n\to\infty$, $\|\tilde{\mathbf{c}}_\varepsilon\| = O_p(n^{-1/2})$.

**Proof.** Let $(\mathbf{Q}_1,\mathbf{Q}_2)=\mathbf{I}_{(1+d_1)\times\{d_2(N+1)\}}(\hat{\mathbf{a}},\tilde{\mathbf{a}}-\hat{\mathbf{a}})$. By (A.33), (A.35), (A.20)

$$\tilde{\mathbf{c}}_\varepsilon^T=\mathbf{I}_{(1+d_1)\times\{d_2(N+1)\}}\tilde{\mathbf{a}}=\mathbf{Q}_1+\mathbf{Q}_2, \tag{A.38}$$

so $\|\mathbf{Q}_2\| = O_p\{n^{-1}N^{3/2}(\log n)^2\}$, while

$$\mathbf{Q}_1=\left(n^{-1}\sum_{i=1}^n\sigma(\mathbf{X}_i,\mathbf{T}_i)\varepsilon_i, n^{-1}\sum_{i=1}^n\xi_{il'}\right)^T_{1\le l'\le d_1} \tag{A.39}$$

in which $\xi_{il'}=\{\sum_l s_{l'l}T_{il}+\sum_{\alpha,J}s_{l',(J,\alpha)}B_{J,\alpha}(X_{i\alpha})\}\sigma(\mathbf{X}_i,\mathbf{T}_i)\varepsilon_i$. Clearly $E\xi_{il'}=0$ while $E\xi_{il'}^2\le C_\sigma^2E[\{\sum_l s_{l'l}T_l+\sum_{\alpha,J}s_{l',(J,\alpha)}B_{J,\alpha}(X_\alpha)\}]^2\le C_\sigma^2(0,s_{l'l},s_{l',(J,\alpha)})\mathbf{V}(0,s_{l'l},s_{l',(J,\alpha)})^T\le C_\sigma^2C_V\|(0,s_{l'l},s_{l',(J,\alpha)})\|^2\le C_\sigma^2C_VC_S^2$. It is easily checked $E(\xi_{il'}\xi_{jl'})=0$ for $i\ne j$ thus by Markov inequality, $\sup_{1\le l'\le d_1}|n^{-1}\sum_{i=1}^n\xi_{il'}| = O_p(n^{-1/2})$. Likewise $n^{-1}\sum_{i=1}^n\sigma(\mathbf{X}_i,\mathbf{T}_i)\varepsilon_i = O_p(n^{-1/2})$. Then, Lemma A.16 follows from the above results. □

**Lemma A.17.** *Under Assumptions* (A1), (A2), (A4)–(A7), *as* $n\to\infty$, $\sup_{x_1\in[0,1]}|\Psi_{Tb}(x_1)| = o_p(1/\sqrt{n})$.

For proof, see Ma and Yang (2009). Define a theoretical version of $\Psi_{Tv}(x_1)=n^{-1}\sum_{i=1}^nK_h(X_{i1}-x_1)(\tilde{a}_0+\sum_{l=1}^{d_1}\tilde{a}_lT_{il})$ as

$$\hat{\Psi}_{Tv}(x_1)=\hat{a}_0n^{-1}\sum_{i=1}^nK_h(X_{i1}-x_1)+\sum_{l=1}^{d_1}\hat{a}_ln^{-1}\sum_{i=1}^n\zeta_l(X_{i1},T_{il},x_1).$$

**Lemma A.18.** *Under Assumptions* (A1)–(A7), *as* $n\to\infty$, $\sup_{x_1\in[0,1]}|\Psi_{Tv}(x_1)| = O_p\{n^{-1/2}(\log n)^4\} = o_p(n^{-2/5})$.

For proof, see Ma and Yang (2009).

**Lemma A.19.** *Under Assumptions* (A1)–(A6) *and* (A8), $(s_{l',l}) = \mathrm{cov}(\tilde{\mathbf{T}}_n)^{-1}$ *and as* $n \to \infty$, $\mathrm{cov}(\tilde{\mathbf{T}}_n)^{-1} \to \mathrm{cov}(\tilde{\mathbf{T}})^{-1}$, *where* $(s_{l',l})$ *defined in* (A.20), $\tilde{\mathbf{T}}_n$ *and* $\tilde{\mathbf{T}}$ *defined in* (A.3).

**Proof.** $(s_{l',l}) = \mathrm{cov}(\tilde{\mathbf{T}}_n)^{-1}$ is induced by basic linear algebra. According to the result on p. 149 of de Boor (2001), there is a constant $C_\infty > 0$ and functions $g_l(\mathbf{x}) \in \mathcal{H}_n, 1 \le \alpha \le d_2$ such that $\sup_{1 \le l \le d_1} \|g_l - p_l\|_\infty = o(1)$. Since $\mathrm{Proj}_{\mathcal{H}_n} T_l = \mathrm{Proj}_{\mathcal{H}_n}(\mathrm{Proj}_{\mathcal{H}} T_l)$, for $\forall 1 \le l \le d_1$ according to Hilbert space theory, $E(\mathrm{Proj}_{\mathcal{H}_n} T_l - \mathrm{Proj}_{\mathcal{H}} T_l)^2 \le E\{g_l(\mathbf{X}) - \mathrm{Proj}_{\mathcal{H}} T_l\}^2 = E\{g_l(\mathbf{X}) - p_l(\mathbf{X})\}^2 = o(1)$, as $n \to \infty$. Thus $\mathrm{cov}(\tilde{\mathbf{T}}_n)^{-1} \to \mathrm{cov}(\tilde{\mathbf{T}})^{-1}$, as $n \to \infty$. □

**Proof of Theorem 1.** The term $\Psi_{Tb}(x_1)$ in (A.24) has order $o_p(n^{-1/2})$ and other terms have order $o_p(n^{-2/5})$ by Lemmas A.17, A.18, A.9 and A.14. Standard theory ensures that $\hat{f}_1(x_1) = n^{-1} \sum_{i=1}^n K_h(X_{i1} - x_1)$ has a positive lower bound. Theorem 1 then follows. □

**Proof of Theorem 2.** The first part of Theorem 2 follows from Lemmas A.15 and A.16. By (A.8) and (A.38) and Lemma A.15, $\sqrt{n}(\hat{\mathbf{c}} - \mathbf{c}) = \sqrt{n}\tilde{\mathbf{c}}_\varepsilon + \sqrt{n}(\tilde{\mathbf{c}}_m - \mathbf{c}) = \sqrt{n}(\mathbf{Q}_1 + \mathbf{Q}_2) + \sqrt{n}(\tilde{\mathbf{c}}_m - \mathbf{c}) = \sqrt{n}\mathbf{Q}_1 + o_p\{1\}$. It is easily verified $E(\sqrt{n}\mathbf{Q}_1) = 0$. By (A.39) and Lemma A.19,

$$\mathrm{var}(\sqrt{n}\mathbf{Q}_1) = n \times \sigma_0^2 \mathbf{I}_{(1+d_1)\times\{d_2(N+1)\}} \mathrm{var}(\hat{a}\hat{a}^T) \mathbf{I}^T_{(1+d_1)\times\{d_2(N+1)\}} = n \times n^{-1} \sigma_0^2 \mathbf{I}_{(1+d_1)\times\{d_2(N+1)\}} \mathbf{S}\mathbf{I}^T_{(1+d_1)\times\{d_2(N+1)\}}$$
$$= \sigma_0^2 \begin{Bmatrix} 1 & 0_{d_1}^T \\ 0_{d_1} & (s_{l',l}) \end{Bmatrix} \to \sigma_0^2 \begin{Bmatrix} 1 & 0_{d_1}^T \\ 0_{d_1} & \mathrm{cov}(\tilde{\mathbf{T}})^{-1} \end{Bmatrix} \quad \text{as } n \to \infty$$

Theorem 2 then follows by applying Theorem 1 of Sunklodas (1984). □

## References


Bosq, D., 1998. Nonparametric Statistics for Stochastic Processes. Springer-Verlag, New York.
Breiman, L., Friedman, J.H., 1985. Estimating optimal transformations for multiple regression and correlation. J. Amer. Statist. Assoc. 80, 580–619.
de Boor, C., 2001. A Practical Guide to Splines. Springer-Verlag, New York.
Fan, J., Gijbels, I., 1996. Local Polynomial Modelling and its Applications. Chapman and Hall, London.
Fan, Y., Li, Q., 2003. A kernel-based method for estimating additive partially linear models. Statist. Sinica 13, 739–762.
González-Manteiga, W., Martínez-Miranda, M.D., Raya-Miranda, R., 2008. SiZer Map for inference with additive models. Statist. Comput. 18, 297–312.
Harrison, D., Rubinfeld, D.L., 1978. Hedonic housing prices and the demand for cleaning air. J. Econom. Manage. 5, 81–102.
Huang, J.Z., Yang, L., 2004. Identification of nonlinear additive autoregression models. J. Roy. Statist. Soc. Ser. B Statist. Methodol. 66, 463–477.
Li, Q., 2000. Efficient estimation of additive partially linear models. Internat. Econom. Rev. 41, 1073–1092.
Liang, H., 2006. Estimation in partially linear models and numerical comparisons. Comput. Statist. Data Anal. 50, 675–687.
Liang, H., Wang, S., Carroll, R., 2007. Partially linear models with missing response variables and error-prone covariates. Biometrika 94, 185–198.
Liang, H., Thurston, S., Ruppert, D., Apanasovich, T., Hauser, R., 2008. Additive partial linear models with measurement errors. Biometrika 95, 667–678.
Ma, S., Yang, L., 2009. Spline-backfitted kernel smoothing of partially linear additive model. Manuscript.
Nácher, V., Ojeda, S., Cadarso-Suárez, C., Roca-Pardiñas, J., Acuña, C., 2006. Neural correlates of memory retrieval in the prefrontal cortex. Eur. J. Neurosci. 24, 925–936.
Roca-Pardiñas, J., Cadarso-Suárez, C., Nácher, V., Acuña, C., 2006. Bootstrap-based methods for testing factor-by-curve interactions in generalized additive models: assessing prefrontal cortex neural activity related to decision-making. Statist. Med. 25, 2483–2501.
Schimek, M., 2000. Estimation and inference in partially linear models with smoothing splines. J. Statist. Plann. Inference 91, 525–540.
Su, L., Ullah, A., 2006. Profile likelihood estimation of partially linear panel data models with fixed effects. Econom. Lett. 92, 75–81.
Sunklodas, J., 1984. On the rate of convergence in the central limit theorem for strongly mixing random variables. Lithuanian Math. J. 24, 182–190.
Tjøstheim, D., Auestad, B., 1994. Nonparametric identification of nonlinear time series: projections. J. Amer. Statist. Assoc. 89, 1398–1409.
Wang, J., Yang, L., 2009. Efficient and fast spline-backfitted kernel smoothing of additive models. Ann. Inst. Statist. Math. 61, 663–690.
Wang, L., Yang, L., 2007. Spline-backfitted kernel smoothing of nonlinear additive autoregression model. Ann. Statist. 35, 2474–2503.
Xue, L., Yang, L., 2006. Additive coefficient modeling via polynomial spline. Statist. Sinica 16, 1423–1446.