# THEORY OF SPLINE REGRESSION WITH APPLICATIONS TO TIME SERIES, LONGITUDINAL, AND CATEGORICAL DATA, AND DATA WITH JUMPS

By

Shujie Ma

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Statistics

2011

# ABSTRACT

## THEORY OF SPLINE REGRESSION WITH APPLICATIONS TO TIME SERIES, LONGITUDINAL, AND CATEGORICAL DATA, AND DATA WITH JUMPS

**By**

**Shujie Ma**

Modern technological advances have led to the explosion in the collection of complex data such as functional/longitudinal, nonlinear time series, and mixed data, and data with jumps. In recent years, there has been a growing interest in developing statistical tools to analyze these data primarily due to the fact that traditional parametric methods are unrealistic in applications. Non- and semi- parametric methods as alternatives have been widely recognized as powerful tools for complex data analysis, which relax the usual assumptions of parametric methods and enable us to explore the data more flexibly so as to uncover data structure that might otherwise be missed.

This dissertation develops statistical theories and methods in spline regression for those complex data mentioned before, with applications to medical science, finance and economics.

In Chapter 2, procedures to detect jumps in the regression function via constant and linear splines are proposed based on the maximal differences of the spline estimators among neighboring knots. Simulation experiments corroborate with the asymptotic theory, while the computing is extremely fast. The detecting-procedure is illustrated in analyzing the thickness of pennies data set.

In Chapter 3, asymptotically simultaneous confidence bands are obtained for the mean function of the functional regression model, using piecewise constant spline estimation. Simulation experiments corroborate the asymptotic theory. The confidence band procedure is

illustrated by analyzing CD4 cell counts of HIV infected patients.

A spline-backfitted kernel smoothing method is proposed in Chapter 4 for partially linear additive autoregression model. Under assumptions of stationarity and geometric mixing, the proposed function and parameter estimators are oracally efficient and fast to compute. Simulation experiments confirm the asymptotic results. Application to the Boston housing data serves as a practical illustration of the method.

Chapter 5 considers the problem of estimating a relationship nonparametrically using regression splines when there exist both continuous and categorical predictors. The resulting estimator possesses substantially better finite-sample performance than either its frequency-based peer or cross-validated local linear kernel regression or even additive regression splines (when additivity does not hold). Theoretical underpinnings are provided and Monte Carlo simulations are undertaken to assess finite-sample behavior.

I dedicate this thesis to my husband, Fulin Wang, my parents, Sanbao Ma and Xiaoli Xue, and my grandmother, Peilan Sun.

## ACKNOWLEDGMENT

This thesis would not have been possible without the support of many people. First of all, I would like to express my deepest gratitude to my supervisor Professor Lijian Yang for his patient guidance, invaluable assistance, encouragement and excellent advice throughout this study. His continued support led me to the right way.

I would like to thank Professor Yuehua Cui, Professor Lyudmila Sakhanenko, and Professor Jeffrey Wooldridge for serving as members of my doctoral committee and for their invaluable suggestions.

Thanks also to my collaborator Professor Jeff Racine for his help and for many interesting discussions and insights.

I am grateful to the entire faculty and staff in the Department of Statistics and Probability who have taught me and assisted me during my study at MSU. My special thanks go to Professor James Stapleton and Professor Dennis Gilliland for their help, support, and encouragement.

I am thankful to Dr. Steven J. Pierce for accepting me as one of the statistical consultants at CSTAT, which provided me with plenty of opportunities to work with students and faculties from different disciplines.

I am thankful to the Graduate School and the Department of Statistics and Probability for providing me with the Dissertation Completion Fellowship (2011), Summer Support Fellowship (2010) and Stapleton Fellowship for working on this dissertation. This dissertation is also supported in part by NSF award 0706518.

I also thank my academic sisters and brothers: Dr. Lan Xue, Dr. Jing Wang, Dr. Li

Wang, Dr. Rong Liu, Dr. Qiongxia Song, Guanqun Cao, and Shuzhuan Zheng for their generous help.

Finally, I take this opportunity to express my profound gratitude to my beloved husband, my parents and grandmother for their love, endless support and encouragement over all these years.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# Chapter 1

# Introduction

## 1.1 Nonparametric Regression

Regression analysis is one of the most widely used tools in data analysis. Regression analysis models and analyzes the relationship between a dependent variable $Y$ and a vector of independent variables $\mathbf{X}$. Most commonly, regression analysis estimates the conditional expectation of the dependent variable given the independent variables. Classic parametric models, although their properties are very well established, need restrictive assumptions, which have encountered various limitations in applications. Mis-specification in parametric models could lead to large bias. Nonparametric modelling as an alternative reduces modeling bias by imposing no specific model structure and enables people to explore the data more flexibly.

We begin by considering the nonparametric regression model

$$Y = m(\mathbf{X}) + \epsilon, \tag{1.1}$$

where the error term $\epsilon$ contributes a roughness to the raw data, and functional form $m(\cdot)$ is unknown and satisfies $m(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$.

In many cases, the unknown function $m(\cdot)$ is a smooth function, then we can estimate it nonparametrically by such methods as kernel and spline smoothing. For example, Figure (1.1) shows a nonparametric smoothing curve fitted by quadratic polynomial spline as well as the data points for a fossil data. The data contains ratios of strontium isotopes found in fossil shells millions of years ago, which can reflect global climate.

In other cases, however, the unknown function $m(\cdot)$ is not smooth. Ignoring possible jumps in the regression function $m(\cdot)$ may result in a serious error in drawing inference about the process under study. In Chapter 2, we propose procedures to detect jumps in the regression function $m(\cdot)$ via constant and linear spline estimation methods in a random-design nonparametric regression model for i.i.d. case. The detecting procedure is illustrated by simulation experiments and by analyzing a thickness of pennies data set.

In Chapter 3, we consider a sparse functional data case which has the form $\left\{X_{ij}, Y_{ij}\right\}$, $1 \leq i \leq n, 1 \leq j \leq N_i$, in which $N_i$ observations are taken for the $i^{th}$ subject, with $X_{ij}$ and $Y_{ij}$ the $j^{th}$ predictor and response variables, respectively, for the $i^{th}$ subject, and $N_i$'s are i.i.d. copies of an integer valued positive random variable. Asymptotically simultaneous confidence bands are obtained for the mean function $m(\cdot)$ of the functional regression model, using piecewise constant spline estimation. For illustration, Figure (1.2) shows a confidence band and confidence interval at 95% confidence level, and the estimated function by piecewise constant spline for a spinal bone mineral density data set.

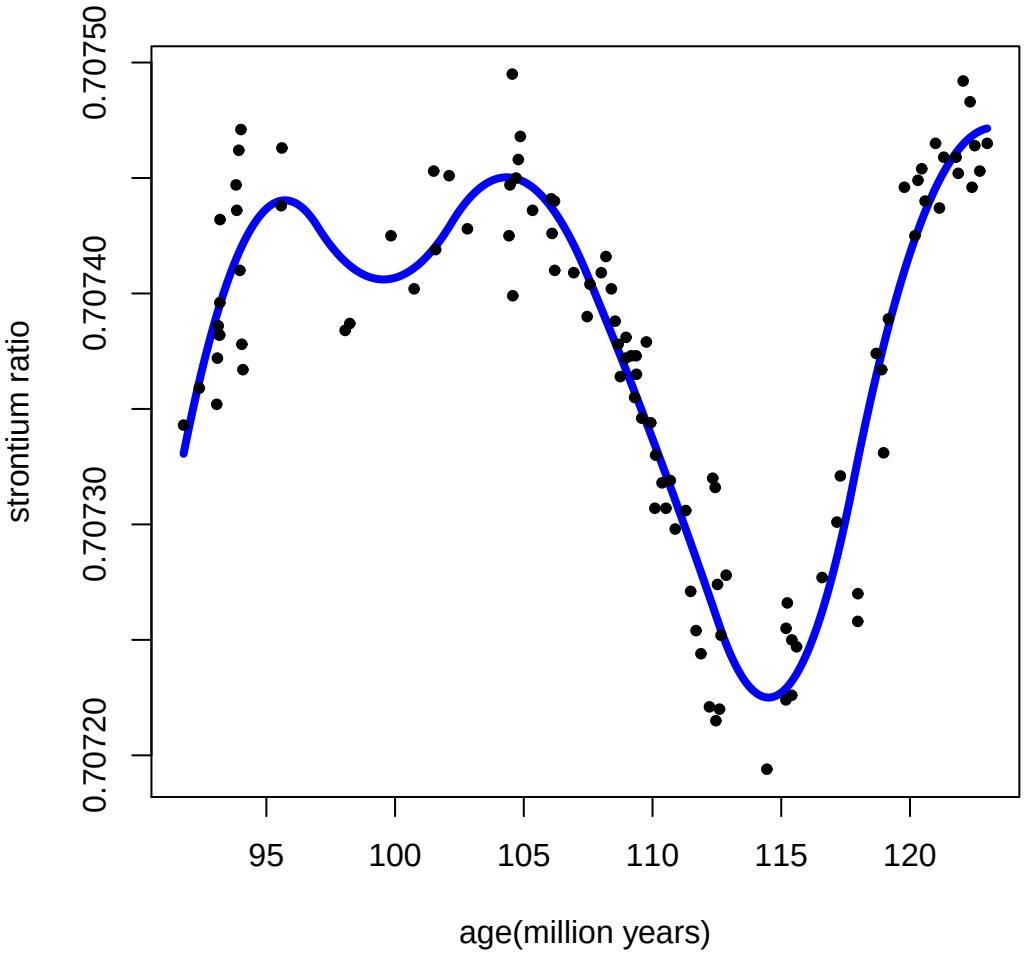


Figure 1.1: Spline estimate plot for the regression function of the fossil data

Note: the quadratic spline estimator (solid line) and the data points (circle) for ratios of strontium isotopes over time. For interpretation of the references to color in this and all other figures, the reader is referred to the electronic version of this dissertation.

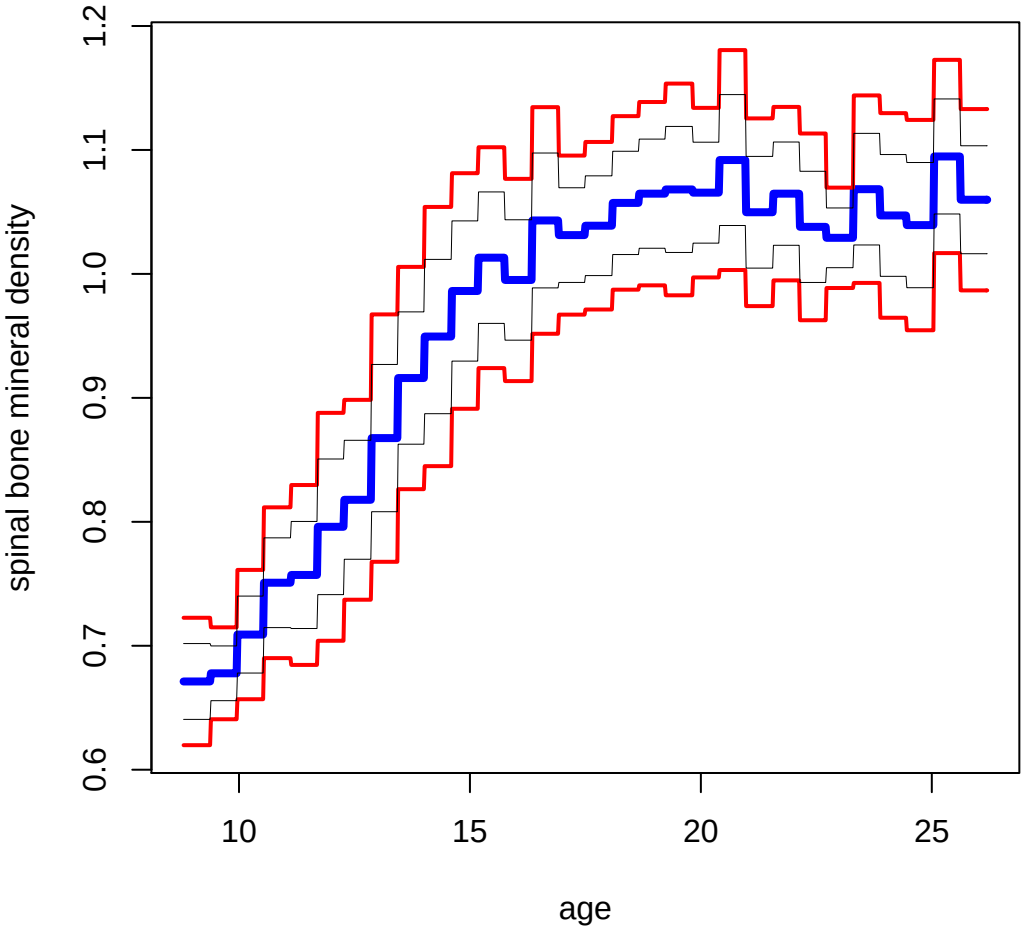


Figure 1.2: Confidence band plot for spinal bone mineral density

Note: the constant spline estimator (middle curve), the confidence band (solid line) and the confidence interval (thin line) at confidence level 95% for spinal bone mineral density over age.

## 1.2 Partially Linear Additive Models

Nonparametric modeling imposes no specific model structure and enables one to explore the data more flexibly, but it does not perform well when the dimension of the covariates is high, and the variances of the resulting estimates tend to be unacceptably large due to the sparseness of data, which is the so-called "curse of dimensionality". To overcome these difficulties, many different semi-parametric models have been proposed and developed, among which partially linear additive model is becoming very popular in data analysis, which are important multivariate semiparametric models by allowing linearity in some variables.

A partially linear additive model (PLAM) is given as

$$Y_i = m\left(\mathbf{X}_i, \mathbf{T}_i\right) + \sigma\left(\mathbf{X}_i, \mathbf{T}_i\right)\varepsilon_i, m(\mathbf{x}, \mathbf{t}) = c_{00} + \sum_{l=1}^{d_1} c_{0l} t_l + \sum_{\alpha=1}^{d_2} m_\alpha\left(x_\alpha\right) \quad (1.2)$$

in which the sequence $\left\{Y_i, \mathbf{X}_i^T, \mathbf{T}_i^T\right\}_{i=1}^n = \left\{Y_i, X_{i1}, ..., X_{id_2}, T_{i1}, \ldots T_{id_1}\right\}_{i=1}^n$. 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 $\{\mathbf{X}_i, \mathbf{T}_i\}$.

From equation (1.2) we observe that a PLAM consists of a linear part and a nonparametric additive part, which retains the merits of both additive and linear models, see Stone (1985) and Hastie and Tibshirani (1990) for additive models. It is much more flexible than parametric models, since it eschews many assumptions, such as finitely many unknown parameters imposed on the model and presumed distribution structure set on the data, and meanwhile it is more interpretable than nonparametric regression surfaces. It avoids the so-called "curse of dimensionality" for high dimensional models, since the unknown functions are one-dimensional. Moreover, PLAMs are more parsimonious and flexible and easier

interpretable than purely additive model by allowing a subset of the independent variables to be discrete, while additive models only admit continuous predictors, as well as allowing the intersection terms among the elements of the additive part to enter as the linear part of the model.

In Chapter 4, we extend the "spline-backfitted kernel smoothing" (SBK) of Wang and Yang (2007) to partially linear additive autoregression models. It establishes the uniform oracle efficiency of the estimators by "reducing bias via undersmoothing (step one) and averaging out the variance (step two)" via the joint asymptotic properties of kernel and spline functions. The proposed SBK estimators satisfy (i) computationally expedient; (ii) theoretically reliable; and (iii) intuitively appealing.

## 1.3 Polynomial Splines

Polynomial splines have been widely accepted as an appealing and effective tool for non-parametric regression estimation because of its great flexibility, fast computation, simple implementation and explicit expression, and they have been applied to a wide range of statistical problems such as hazard regressions, derivative estimation, semi-parametric models, longitudinal, functional and high dimensional data analysis, jump detection, etc. In this section, I will introduce polynomial spline functions.

Without loss of generality, we take the range of $X$ a univariate predictor to be $[0, 1]$. To introduce the spline functions, divide the finite interval $[0, 1]$ into $(N + 1)$ equal subintervals $\chi_J = \big[t_J, t_{J+1}\big)$, $J = 0, ...., N - 1$, $\chi_N = \big[t_N, 1\big]$. A sequence of equally-spaced points

$\{t_J\}_{J=1}^N$, called interior knots, are given as

$$t_0 = 0 < t_1 < \cdots < t_N < 1 = t_{N+1},\ t_J = Jh,\ J = 0, ..., N+1,$$

in which $h = 1/(N+1)$ is the distance between neighboring knots. We denote by $G^{(p-2)} = G^{(p-2)}[0,1]$ the space of functions that are polynomials of degree $p-1$ on each $\chi_J$ and has continuous $(p-2)$-th derivative. For example, $G^{(-1)}$ denotes the space of functions that are constant on each $\chi_J$, and $G^{(0)}$ denotes the space of functions that are linear on each $\chi_J$ and continuous on $[0,1]$.

Denote by $\|\phi\|_2$ the theoretical $L^2$ norm of a function $\phi$ on $[0,1]$, $\|\phi\|_2^2 = E\left\{\phi^2(X)\right\} = \int_0^1 \phi^2(x) f(x)\, dx = \int_0^1 \phi^2(x)\, dx$, and the empirical $L^2$ norm as $\|\phi\|_{2,n}^2 = n^{-1}\sum_{i=1}^n \phi^2(X_i)$. Corresponding inner products are defined by

$$\langle \phi, \varphi \rangle = \int_0^1 \phi(x)\varphi(x) f(x)\, dx = \int_0^1 \phi(x)\varphi(x)\, dx = E\{\phi(X)\varphi(X)\},$$

$\langle \phi, \varphi \rangle_n = n^{-1}\sum_{i=1}^n \phi(X_i)\varphi(X_i)$, for any $L^2$-integrable functions $\phi, \varphi$ on $[0,1]$. Clearly $E\langle \phi, \varphi \rangle_n = \langle \phi, \varphi \rangle$. We now introduce the B-spline basis $\left\{b_{J,1}(x), 1-p \le J \le N\right\}$ of $G^{(p-2)}$, the space of splines of order $p$, for theoretical analysis. The B-spline basis of $G^{(-1)}$, the space of piecewise constant splines, are indicator functions of intervals $\chi_J$, $b_{J,1}(x) = I_J(x) = I_{\chi_J}(x)$, $0 \le J \le N$. The B-spline basis of $G^{(0)}$, the space of piecewise linear splines, are $\left\{b_{J,2}(x)\right\}_{J=-1}^N$

$$b_{J,2}(x) = K\left(\frac{x - t_{J+1}}{h}\right),\ -1 \le J \le N,\ \text{for } K(u) = (1-|u|)_+.$$

In Chapters 2, 3, and 4, polynomial spline techniques are applied in different circumstances.

## 1.4 Tensor Product Splines

In section 1.3, we introduced spline smoothing and B-spline functions for univariate variables. In this section, we generalize B-spline to functions of multivariate variables by the tensor product construction.

Let $\mathbf{X} = \left(X_1, \ldots, X_q\right)^{\mathrm{T}}$ be a $q$-dimensional vector of continuous predictors. Assume for $1 \leq l \leq q$, each $X_l$ is distributed on a compact interval $\left[a_l, b_l\right]$, and without loss of generality, we take all intervals $\left[a_l, b_l\right] = [0, 1]$. Let $G_l = G_l^{\left(m_l - 2\right)}$ be the space of polynomial splines of order $m_l$ and pre-select an integer $N_l = N_{n,l}$. Divide $[0, 1]$ into $\left(N_l + 1\right)$ subintervals $I_{J_l,l} = \left[t_{J_l,l}, t_{J_l+1,l}\right)$, $J_l = 0, \ldots, N_l - 1$, $I_{N_l,l} = \left[t_{N_l,l}, 1\right]$, where $\left\{t_{J_l,l}\right\}_{J_l=1}^{N_l}$ is a sequence of equally-spaced points, called interior knots, given as

$$t_{-\left(m_l-1\right),l} = \cdots = t_{0,l} = 0 < t_{1,l} < \cdots < t_{N_l,l} < 1 = t_{N_l+1,l} = \cdots = t_{N_l+m_l,l},$$

in which $t_{J_l,l} = J_l h_l$, $J_l = 0, 1 \ldots, N_l + 1$, $h_l = 1/\left(N_l + 1\right)$ is the distance between neighboring knots. Then $G_l$ consists of functions $\varpi$ satisfying (i) $\varpi$ is a polynomial of degree $m_l - 1$ on each of the subintervals $I_{J_l,l}, J_l = 0, \ldots, N_l$; (ii) for $m_l \geq 2$, $\varpi$ is $m_l - 2$ times continuously differentiable on $[0, 1]$. Let $K_l = K_{n,l} = N_l + m_l$, where $N_l$ is the number of interior knots and $m_l$ is the spline order, $B_l\left(x_l\right) = \left\{B_{J_l,l}\left(x_l\right) : 1 - m_l \leq J_l \leq N_l\right\}^{\mathrm{T}}$ be a basis system of the space $G_l$. We define the space of tensor-product polynomial splines by

$\mathcal{G} = \otimes_{l=1}^{q} G_l$. It is clear that $\mathcal{G}$ is a linear space of dimension $\mathbf{K}_n = \prod_{l=1}^{q} K_l$. Then

$$\mathcal{B}(\mathbf{x}) = \left[ \left\{ \mathcal{B}_{J_1,\ldots,J_q}(\mathbf{x}) \right\}_{J_1=1-m_1,\ldots,J_q=1-m_q}^{N_1,\ldots,N_q} \right]_{\mathbf{K}_n \times 1} = B_1(x_1) \otimes \cdots \otimes B_q\left(x_q\right) \quad (1.3)$$

is a basis system of the space $\mathcal{G}$, where $\mathbf{x} = \left(x_l\right)_{l=1}^{q}$.

In Chapter 5, tensor product splines and categorical kernel functions which will be introduced in this chapter are applied to study nonparametric regression with multivariate continuous and categorical variables.

# Chapter 2

# A Jump-detecting Procedure based on Polynomial Spline Estimation

## 2.1 Introduction

This chapter is based on Ma and Yang (2011a). In application of regression methods, ignoring possible jump points may result in a serious error in drawing inference about the process under study. Whenever there is no appropriate parametric method available, one may start from nonparametric regression. Two popular nonparametric techniques are kernel and spline smoothing. For properties of kernel estimators in the absence of jump points, see Mack and Silverman (1982), Fan and Gijbels (1996), Xia (1998) and Claeskens and Van Keilegom (2003), and for spline estimators, see Zhou, Shen and Wolfe (1998), Huang (2003) and Wang and Yang (2009a).

One is often interested in detecting jump points and estimating regression function with jumps. We assume that observations $\{(X_i, Y_i)\}_{i=1}^n$ and unobserved errors $\{\varepsilon_i\}_{i=1}^n$ are

i.i.d. copies of $(X, Y, \varepsilon)$ satisfying the regression model

$$Y = m(X) + \sigma\varepsilon, \tag{2.1}$$

where the joint distribution of $(X, \varepsilon)$ satisfies Assumptions (A3) and (A4) in Section 2.2. The unknown mean function $m(x)$, defined on interval $[a, b]$, may have a finite number of jump points.

Jump regression analysis started in the early 1990s and has become an important research topic in statistics. See for instance, Qiu, Asano and Li (1991), Müller (1992), Wu and Chu (1993), Qiu (1994) and Qiu and Yandell (1998) for procedures that detect the jumps explicitly before estimating the regression curve, Kang, Koo and Park (2000) for comparing two estimators of the regression curve after the jump points are detected, Qiu (2003) and Gijbels Lambert and Qiu (2007) for jump-preserving curve estimators, Joo and Qiu (2009) for jump detection in not only the regression curve but also its derivatives. For a comprehensive view on jump regression, see Qiu (2005).

Jump detection has been tackled with many techniques, including local polynomial smoothing [Qiu (2003) and Gijbels et al. (2007)], smoothing spline [Shiau (1987)], wavelet methods [Hall and Patil (1995), Wang (1995) and Park and Kim (2004, 2006)], and for two-dimensional cases, see Qiu (2007). We propose a spline smoothing method to detect jumps by solving one optimization problem over the range of $x$ instead of each point, which is computationally more expedient than kernel-type method in Müller (1992). Spline method was also discussed in Koo (1997), which proposed estimating discontinuous regression function without providing theoretical justifications. In contrast, asymptotic distributions in Theorem 2.1 are established by using the strong approximation results in Wang and Yang (2009a),

Normal Comparison Lemma in Leadbetter, Lindgren and Rootzén, H. (1983), and a convenient formula from Kılıç (2008) for inverting tridiagonal matrix. The automatic procedures proposed for detecting jumps are based on implementing the asymptotics of Theorem 2.1.

This chapter is organized as follows. Section 2.2 states main theoretical results based on (piecewise) constant and linear splines. Section 2.3 provides steps to implement the procedure based on the asymptotic result. Section 2.4 reports findings in both simulation and real data studies. All technical proofs are contained in Appendices.

## 2.2 Main Results

We denote the space of the $p$-th order smooth functions as $C^{(p)}[a,b] = \left\{\varphi \left| \varphi^{(p)} \in C[a,b] \right.\right\}$, for $p = 1, 2$. Without loss of generality, we take the range of $X$ to be $[0,1]$. We use the spline functions introduced in Section 1.3 of Chapter 1. Define the spline estimator based on data $\left\{\left(X_i, Y_i\right)\right\}_{i=1}^n$ drawn from model (3.3) as

$$\hat{m}p\left(x\right) = \text{argmin}_{g \in G^{(p-2)}[0,1]} \sum_{i=1}^n \left\{Y_i - g\left(X_i\right)\right\}^2, \ p = 1, 2. \tag{2.2}$$

The unknown function $m\left(x\right)$ in (3.3) may be smooth, or have jump points $\left\{\tau_i\right\}_{i=1}^k$, for $0 = \tau_0 < \tau_1 < \cdots < \tau_k < \tau_{k+1} = 1$. Technical assumptions are listed as follows:

(A1) *There exists a function* $m_0\left(x\right) \in C^{(p)}[0,1]$ *and a vector* $\mathbf{c} = \left(c_1, ..., c_k\right)$ *of jump magnitudes such that the regression function* $m\left(x\right) = c_l + m_0\left(x\right), x \in \left[\tau_l, \tau_{l+1}\right)$, *for* $l = 1, \cdots k-1, m\left(x\right) = m_0\left(x\right), x \in \left[\tau_0, \tau_1\right), m\left(x\right) = c_k + m_0\left(x\right), x \in \left[\tau_k, \tau_{k+1}\right]$.

(A2) *The number of interior knots* $N = o\left(n^{1/(2p+1)+\vartheta}\right)$ *for any* $\vartheta > 0$ *while* $N^{-1} = o\left(n^{-1/(2p+1)}/\log n\right)$.

(A3) *$X$ is uniformly distributed on interval $[0,1]$, i.e. the density function of $X$ is $f(x) = I(0 \leq x \leq 1)$.*

(A4) *The joint distribution $F(x,\varepsilon)$ of random variables $(X,\varepsilon)$ satisfies the following:*

(a) *The error is a white noise: $E(\varepsilon\,|X=x) = 0$, $E\left(\varepsilon^2\,|X=x\right) = 1$.*

(b) *There exists a positive value $\eta > 1$ and finite positive $M_\eta$ such that $E|\varepsilon|^{2+\eta} < M_\eta$ and $\sup_{x\in[0,1]} E\left(|\varepsilon|^{2+\eta}\,|X=x\right) < M_\eta$.*

Assumption (A1) is similar to Müller and Song (1997). Assumption (A2) is similar to the undersmoothing condition in Claeskens and Van Keilegom (2003), thus the subinterval length $h = o\left(n^{-1/(2p+1)}/\log n\right)$ while $h^{-1} = o\left(n^{1/(2p+1)+\vartheta}\right)$ for any $\vartheta > 0$. A uniform distribution of $X$ in Assumption (A3) is for the simplicity of proofs, which can be relaxed to any distribution with continuous and positive density function on $[0,1]$. Assumption (A4) is identical with (C2) (a) of Mack and Silverman (1982). All are typical assumptions for nonparametric regression, with Assumption (A4) weaker than the corresponding assumption in Härdle (1989).

We use the B-spline basis introduced in Section 1.3 of Chapter 1 for theoretical analysis. Define next the theoretical norms of spline functions

$$\begin{aligned}
c_{J,n} &= \left\|b_{J,1}\right\|_2^2 = \int_0^1 I_J^2(x)\,dx = \int_0^1 I_J(x)\,dx = h,\ 0 \leq J \leq N, \\
d_{J,n} &= \left\|b_{J,2}\right\|_2^2 = \int_0^1 K^2\left(\frac{x-t_{J+1}}{h}\right)dx = \begin{cases} 2h/3, 0 \leq J \leq N-1 \\ h/3, J = -1, N \end{cases}, \\
\left\langle b_{J,2}, b_{J',2}\right\rangle &= \int_0^1 K\left(\frac{x-t_{J+1}}{h}\right)K\left(\frac{x-t_{J'+1}}{h}\right)dx = \begin{cases} h/6, \left|J-J'\right| = 1 \\ 0, \left|J-J'\right| > 1 \end{cases} \qquad (2.3)
\end{aligned}$$

We introduce the rescaled B-spline basis $\left\{B_{J,p}(x)\right\}_{J=1-p}^{N}$, for $G^{(p-2)}$,

$$B_{J,p}(x) \equiv b_{J,p}(x) \left\|b_{J,p}\right\|_2^{-1}, J = 1-p, ..., N. \tag{2.4}$$

The inner product matrix $V$ of the B-spline basis $\left\{B_{J,2}(x)\right\}_{J=-1}^{N}$ is denoted as

$$\mathbf{V} = \left(v_{J'J}\right)_{J,J'=-1}^{N} = \left(\left\langle B_{J',2}, B_{J,2}\right\rangle\right)_{J,J'=-1}^{N}$$

$$= \begin{pmatrix} 1 & \sqrt{2}/4 & & & & 0 \\ \sqrt{2}/4 & 1 & 1/4 & & & \\ & 1/4 & 1 & \ddots & & \\ & & \ddots & \ddots & 1/4 & \\ & & & 1/4 & 1 & \sqrt{2}/4 \\ 0 & & & & \sqrt{2}/4 & 1 \end{pmatrix}_{(N+2)\times(N+2)} = \left(l_{ik}\right)_{(N+2)\times(N+2)}^{-1}, \tag{2.5}$$

which computed via (2.3). Denote the inverse of $\mathbf{V}$ by $\mathbf{S}$ and for $J = 1, ..., N$, $3 \times 3$ diagonal submatrices of $\mathbf{S}$ are expressed as

$$\mathbf{S} = \left(s_{J'J}\right)_{J,J'=-1}^{N} = V^{-1},\ S_J = \begin{pmatrix} {}^{s}(J-2),(J-2) & {}^{s}(J-2),(J-1) & {}^{s}(J-2),J \\ {}^{s}(J-1),(J-2) & {}^{s}(J-1),(J-1) & {}^{s}(J-1),J \\ {}^{s}J,(J-2) & {}^{s}J,(J-1) & {}^{s}JJ \end{pmatrix}. \tag{2.6}$$

To detect jumps in $m$, one tests the hypothesis $\mathcal{H}_0$: $m \in C^{(p)}[0,1]$ vs $\mathcal{H}_1$: $m \notin C[0,1]$. Denote by $\|\mathbf{c}\|_2 = \left(c_1^2 + \cdots + c_k^2\right)^{1/2}$, the Euclidean norm of the vector $\mathbf{c}$ of all the $k$ jump

magnitudes, then under Assumption (A1), one can write alternatively $\mathcal{H}_0 : \|\mathbf{c}\|_2 = 0$ vs $\mathcal{H}_1 : \|\mathbf{c}\|_2 > 0$. For $\hat{m}_p(x)$ given in (2.2), $p = 1, 2$, define the test statistics

$$T_{1n} = \max_{0 \le J \le N-1} \hat{\delta}_{1J},\ \hat{\delta}_{1J} = \left|\hat{m}_1\left(t_{J+1}\right) - \hat{m}_1\left(t_J\right)\right| / \sigma_{n,1},$$
$$T_{2n} = \max_{1 \le J \le N} \hat{\delta}_{2J},\ \hat{\delta}_{2J} = \left|\left\{\hat{m}_2\left(t_{J+1}\right) + \hat{m}_2\left(t_{J-1}\right)\right\}/2 - \hat{m}_2\left(t_J\right)\right| / \sigma_{n,2,J}, \tag{2.7}$$

$$\text{where } \sigma^2_{n,1} = 2\sigma^2/(nh),\ \sigma^2_{n,2,J} = \sigma^2 (8nh/3)^{-1} \zeta^T S_J \zeta,\ \zeta = \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix} \tag{2.8}$$

with $S_J$ defined in (2.6). To state our main results, denote

$$d_N(\alpha) = 1 - \{2\log N\}^{-1} \left[\log\left\{-\frac{1}{2}\log(1-\alpha)\right\} + \frac{1}{2}\{\log\log(N) + \log 4\pi\}\right]. \tag{2.9}$$

THEOREM 2.1. *Under Assumptions (A1)-(A4) and* $\mathcal{H}_0$,

$$\lim_{n\to\infty} P\left[T_{pn} > \{2\log(N-2p+2)\}^{1/2} d_{N-2p+2}(\alpha)\right] = \alpha,\ p = 1, 2.$$

A similar result by kernel smoothing with fixed-design regression model exists in Theorem 3 of Wu and Chu (1993). The proof of that result, however, does not contain sufficient details for us to further comment. It is feasible to derive similar asymptotic result for $T_{pn}$ under $\mathcal{H}_1$ but that is beyond the scope of this chapter so we leave it to future work.

## 2.3 Implementation

In this section, we describe how to implement in XploRe (Härdle, Hlávka and Klinke (2000)) the jump points detection procedures by using the results in Theorem 2.1.

Given any sample $\left\{\left(X_i, Y_i\right)\right\}_{i=1}^{n}$ from model (3.3), denote $X_{\min} = \min\left(X_1, ..., X_n\right)$ and $X_{\max} = \max\left(X_1, ..., X_n\right)$. Then we transform $\left\{X_i\right\}_{i=1}^{n}$ onto interval $[0, 1]$ by subtracting each $X_i$ from $X_{\min}$, then dividing by $X_{\max} - X_{\min}$. The definition of $\hat{m}_p\left(x\right)$ in (2.2) entails

$$\hat{m}_p\left(x\right) \equiv \sum\nolimits_{J=1-p}^{N} \hat{\lambda}'_{J,p} b_{J,p}\left(x\right), \ p = 1, 2, \tag{2.10}$$

where coefficients $\left\{\hat{\lambda}'_{1-p,p}, ..., \hat{\lambda}'_{N,p}\right\}^T$ are solutions of the following least squares problem

$$\left\{\hat{\lambda}'_{1-p,p}, ..., \hat{\lambda}'_{N,p}\right\}^T = \underset{\left\{\lambda_{1-p,p}, ..., \lambda_{N,p}\right\} \in R^{N+p}}{\operatorname{argmin}} \sum_{i=1}^{n} \left\{Y_i - \sum\nolimits_{J=1-p}^{N} \lambda_{J,p} b_{J,p}\left(X_i\right)\right\}^2.$$

By Assumption (A2), the number of interior knots $N$ is taken to be $N = \left[n^{1/3} \left(\log n\right)^2 / 5\right]$ for constant spline ($p = 1$), and $N = \left[n^{1/5} \left(\log n\right)^2 / 5\right]$ for linear spline ($p = 2$), in which $[a]$ denotes the integer part of $a$. Denote the estimator for $Y_i$ by $\hat{Y}_{i,p} = \hat{m}_p\left(X_i\right)$, for $i = 1, \cdots, n$, with $\hat{m}_p$ given in (2.10), and define the estimator of $\sigma$ as

$$\hat{\sigma} p = \left\{\sum\nolimits_{i=1}^{n} \left(Y_i - \hat{Y}_{i,p}\right)^2 / \left(n - N - p\right)\right\}^{1/2}.$$

Basic spline smoothing theory as in Wang and Yang (2009a) ensures that $\hat{\sigma}_p^2 \to_{\mathrm{p}} \sigma^2$, as $n \to \infty$, hence Theorem 2.1 holds if one replaces $\sigma$ by $\hat{\sigma} p$. The asymptotic p-value $p_{\text{value},p}$ are obtained from solving the equation $T_{pn} = \left\{2 \log\left(N - 2p + 2\right)\right\}^{1/2} d_{N-2p+2}\left(p_{\text{value},p}\right)$,

$p = 1, 2$ with $T_{pn}$ defined in (2.7) and estimated by replacing $\sigma^2$ with $\hat{\sigma}_p^2$, then

$$p_{\text{value},p} = 1 - \exp\left[-2\exp\left[2\log\left(N-2p+2\right)\left\{1 - \left\{2\log\left(N-2p+2\right)\right\}^{-1/2} T_{pn}\right\} - 2^{-1}\left\{\log\log\left(N-2p+2\right) + \log 4\pi\right\}\right]\right]. \tag{2.11}$$

When the p-value is below a pre-determined $\alpha$, one concludes that there exist jump points in $m$. The jump locations and magnitudes are estimated as follows. We use the BIC criteria proposed in Xue and Yang (2006) to select the "optimal" $N$, denoted $\hat{N}^{\text{opt}}$, from $\left[\left[4n^{1/3}\right] + 4, \min\left(\left[10n^{1/3}\right], [n/2] - 1\right)\right]$, which minimizes the BIC value $\text{BIC}(N) = \log\left(\hat{\sigma}_1^2\right) + (N+1) \times \log(n)/n$. By letting $p = 1$ and replacing $T_{1n}$ with $\hat{\delta}_{1J}$, for $0 \leq J \leq N-1$ in (2.11), we obtain the p-value $p_{\text{value},1,J}$ for each $\hat{\delta}_{1J}$. The jump locations $\tau_i, 1 \leq i \leq k$ are estimated by $\hat{\tau}_i = \left(t_{l_i} + t_{l_i+1}\right)/2$, for $p_{\text{value},1,l_i} < \alpha$, with $\hat{c}_i = \hat{m}_1\left(t_{l_i+1}\right) - \hat{m}_1\left(t_{l_i}\right)$ as the estimated jump magnitudes, for $0 \leq l_1, \ldots, l_k \leq N-1$. It is apparent that for $\tau_i \in \left[t_{l_i}, t_{l_i+1}\right]$, $\hat{\tau}_i \to \tau_i, 1 \leq i \leq k$, as $n \to \infty$.

## 2.4 Examples

### 2.4.1 Simulation example

Here, we examine the finite-sample performance of the procedure described in Section 2.3 where $m(x)$ has at most one jump. The data set is generated from model (3.3) with $X \sim U[-1/2, 1/2], \varepsilon \sim N(0,1)$, and with $m(x) = \sin(2\pi x) + c \times I(\tau_1 \leq x \leq 1/2)$, for $\tau_1 = \sqrt{2}/4$. The noise level $\sigma = 0.2, 0.5$, sample size $n = 200, 600, 1000$ and significant level $\alpha = 0.05, 0.01$. Let $n_{\text{S}}$ be the number of replications. Denote the asymptotic powers based on constant and linear splines by $\hat{\beta}_p(c)$, $p = 1, 2$, calculated from

| $c$ | $\sigma$ | sample size $n$ | $\hat{\beta}_2(c)$ $\alpha = 0.05$ | $\hat{\beta}_2(c)$ $\alpha = 0.01$ | $\hat{\beta}_1(c)$ $\alpha = 0.05$ | $\hat{\beta}_1(c)$ $\alpha = 0.01$ |
|---|---|---|---|---|---|---|
| 0 | 0.2 | 200 | 0.100 | 0.032 | 0.640 | 0.280 |
| | | 600 | 0.062 | 0.014 | 0.390 | 0.140 |
| | | 1000 | 0.046 | 0.010 | 0.220 | 0.050 |
| | 0.5 | 200 | 0.058 | 0.012 | 0.220 | 0.070 |
| | | 600 | 0.054 | 0.006 | 0.180 | 0.040 |
| | | 1000 | 0.050 | 0.010 | 0.120 | 0.030 |
| 2 | 0.2 | 200 | 1.000 | 0.998 | 1.000 | 1.000 |
| | | 600 | 1.000 | 1.000 | 1.000 | 1.000 |
| | | 1000 | 1.000 | 1.000 | 1.000 | 1.000 |
| | 0.5 | 200 | 0.942 | 0.776 | 0.890 | 0.680 |
| | | 600 | 1.000 | 0.980 | 1.000 | 0.970 |
| | | 1000 | 1.000 | 1.000 | 1.000 | 1.000 |

Table 2.1: Power calculation for the simulated example in Chapter 2

Note: powers calculated from the test statistic $T_{pn}$ defined in (2.7) by constant and linear splines, respectively, over $n_{\mathrm{S}} = 500$ replications.

$\hat{\beta}_p(c) = \sum_{q=1}^{n_{\mathrm{S}}} I\left[T_{n,p,q} > \{2\log(N-2p+2)\}^{1/2} d_{N-2p+2}(\alpha)\right]/n_{\mathrm{S}}$, where $T_{n,p,q}$ is the $q$-th replication of $T_{pn}$, with $T_{pn}$ given in (2.7), and $d_N(\alpha)$ given in (2.9), for $p = 1, 2$. Table 2.1 shows values of $\hat{\beta}_p(c)$ for $c = 0$ and $c = 2$.

In Table 2.1, $\hat{\beta}_p(2)$, $p = 1, 2$ approach to 1 rapidly. Meanwhile $\hat{\beta}_2(0)$ approaches $\alpha$ as the sample size increase, which shows very positive confirmation of Theorem 2.1, in contrast to $\hat{\beta}_1(0)$, the convergent rate of which is much slower, indicating that the linear spline method outperforms the constant spline method. Table 2.1 also shows the noise level has more influence on the constant spline method than the linear spline method. Table 2.2 shows the average computing time (in seconds) of generating data and detecting jump by constant and linear spline methods, which are comparable.

There are 500 replications for $n = 200, 600$ satisfying $p_{\mathrm{value},2} < \alpha = 0.05$, with $p_{\mathrm{value},2}$

| sample size $n$ | constant | linear |
|---|---|---|
| 200 | 0.04 | 0.06 |
| 600 | 0.21 | 0.30 |
| 1000 | 0.55 | 0.60 |

Table 2.2: Computing time for simulated example in Chapter 2

Note: computing time (in seconds) per replication over $n_S = 500$ replications of generating data and detecting jump by constant and linear spline methods.

given in (2.11), when $c = 2$, $n_S = 500$. Figures 2.1 and 2.2 show the kernel estimators of the densities of $\hat{\tau}_1$ and $\hat{c}_1$ given in Section 2.3 with sample size $n = 200$ (thick lines) and $n = 600$ (median lines) at $\sigma = 0.2$.

The vertical lines at $\sqrt{2}/4$ and 2 are the standard lines for comparing $\hat{\tau}_1$ to $\tau_1$ and $\hat{c}_1$ to $c_1$ respectively. One clearly sees that both of the centers of the density plots are going toward the standard lines with much narrower spread when the sample size $n$ is increasing. The frequencies over 500 replications for $\tau_1$ falling between $t_{l_1}$ and $t_{l_1+1}$ described in Section 2.3 are 0.994 and 1 for $n = 200$ and 600 respectively.

For visual impression of the actual function estimates, at noise level $\sigma = 0.2$ with sample size $n = 600$, we plot the spline estimator $\hat{m}_2(x)$ (solid curves) for the true functions $m(x)$ (thick solid curves) in Figures 2.3 and 2.4. The spline estimators seem rather satisfactory.

### 2.4.2 Real data analysis

We apply the jump detection procedures in Section 2.3 to the thickness of pennies data set given in Scott (1992), which consists of measurements in mils of the thickness of 90 US Lincoln pennies. There are two measurements taken as the response variable $Y$ each year,

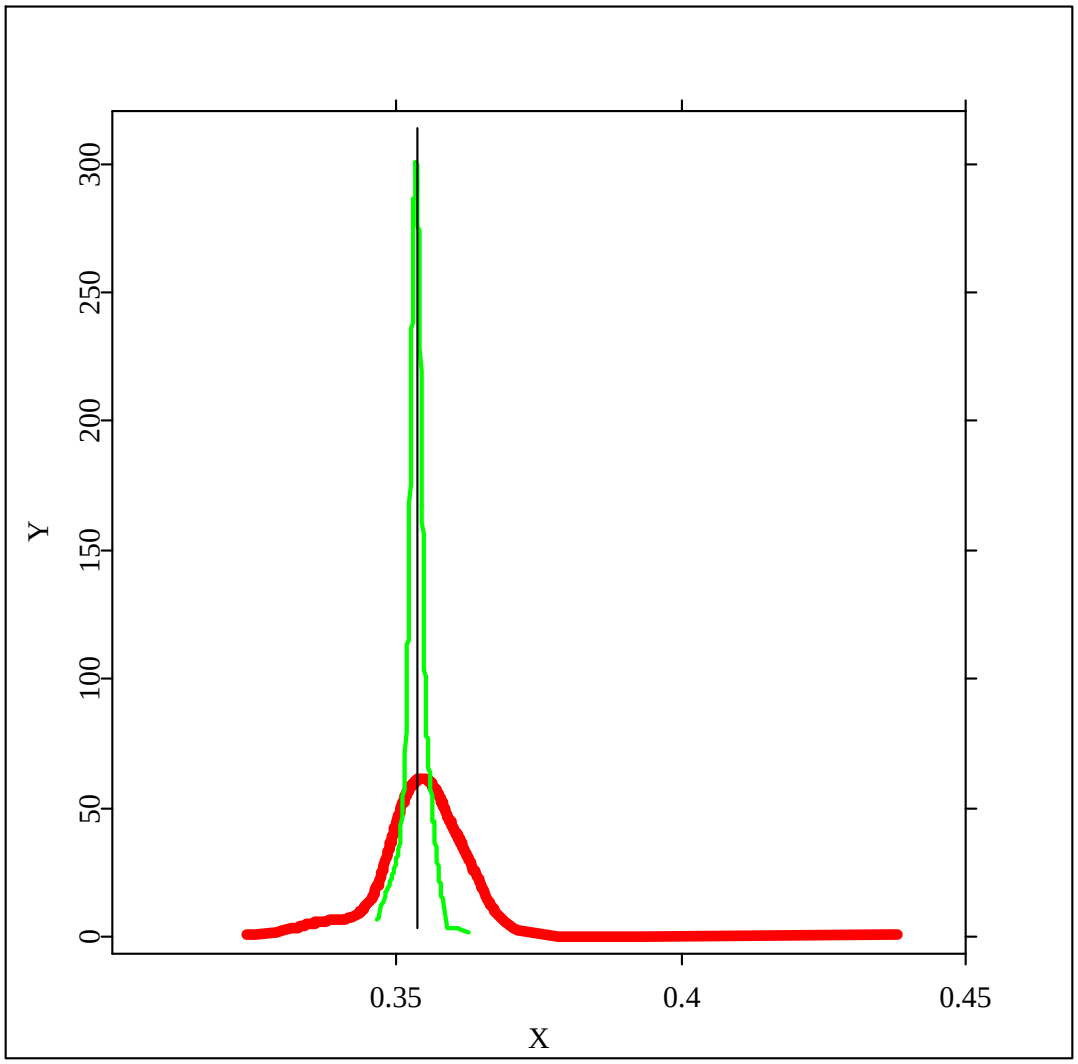


Figure 2.1: Kernel density plots of $\hat{\tau}_1$ in Chapter 2

Note: kernel density plots of $\hat{\tau}_1$ over 500 replications of sample size $n = 200$ (thick solid) and $n = 600$ (solid) for which $\mathcal{H}_0$ is rejected.

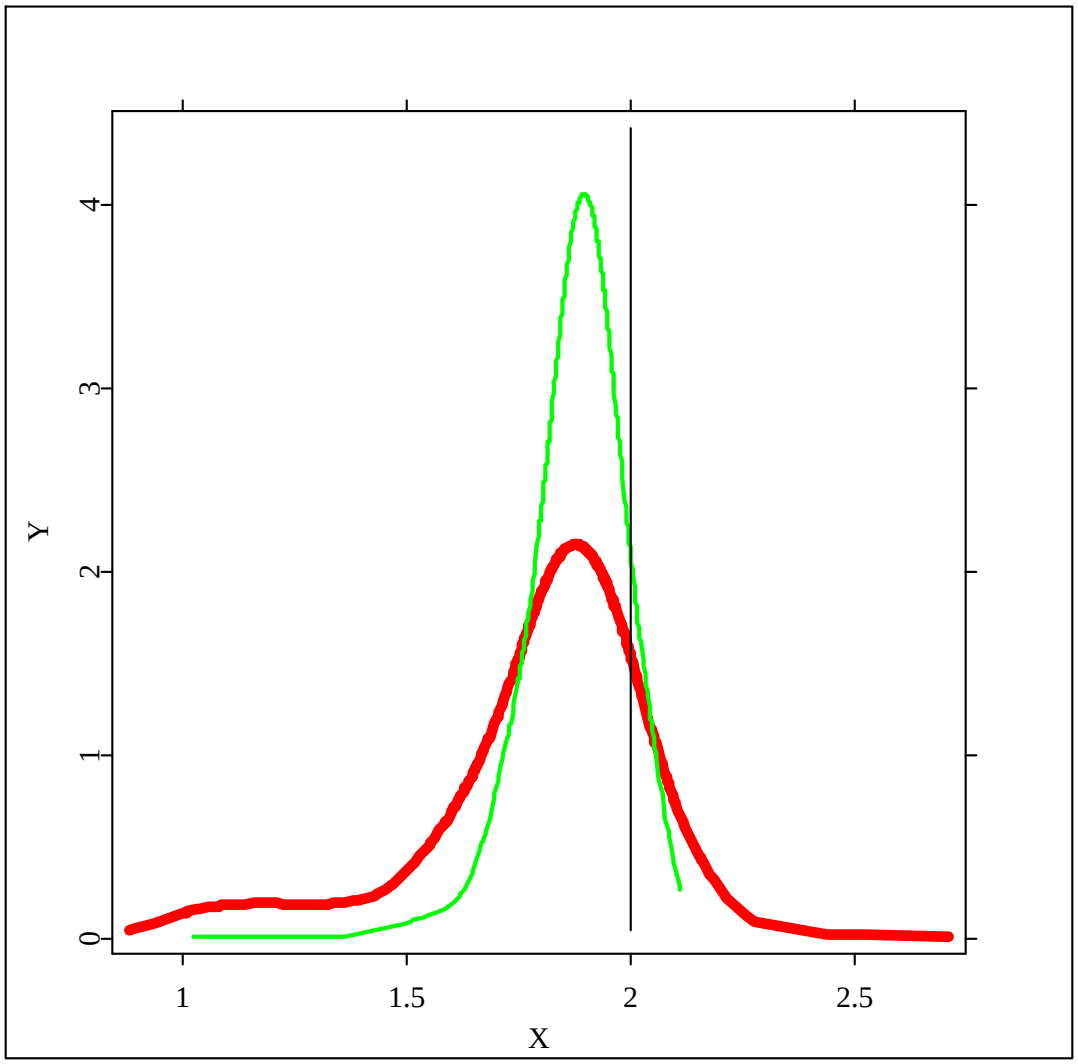


Figure 2.2: Kernel density plots of $\hat{c}_1$ in Chapter 2

Note: kernel density plots of $\hat{c}_1$ over 500 replications of sample size $n = 200$ (thick solid) and $n = 600$ (solid) for which $\mathcal{H}_0$ is rejected.

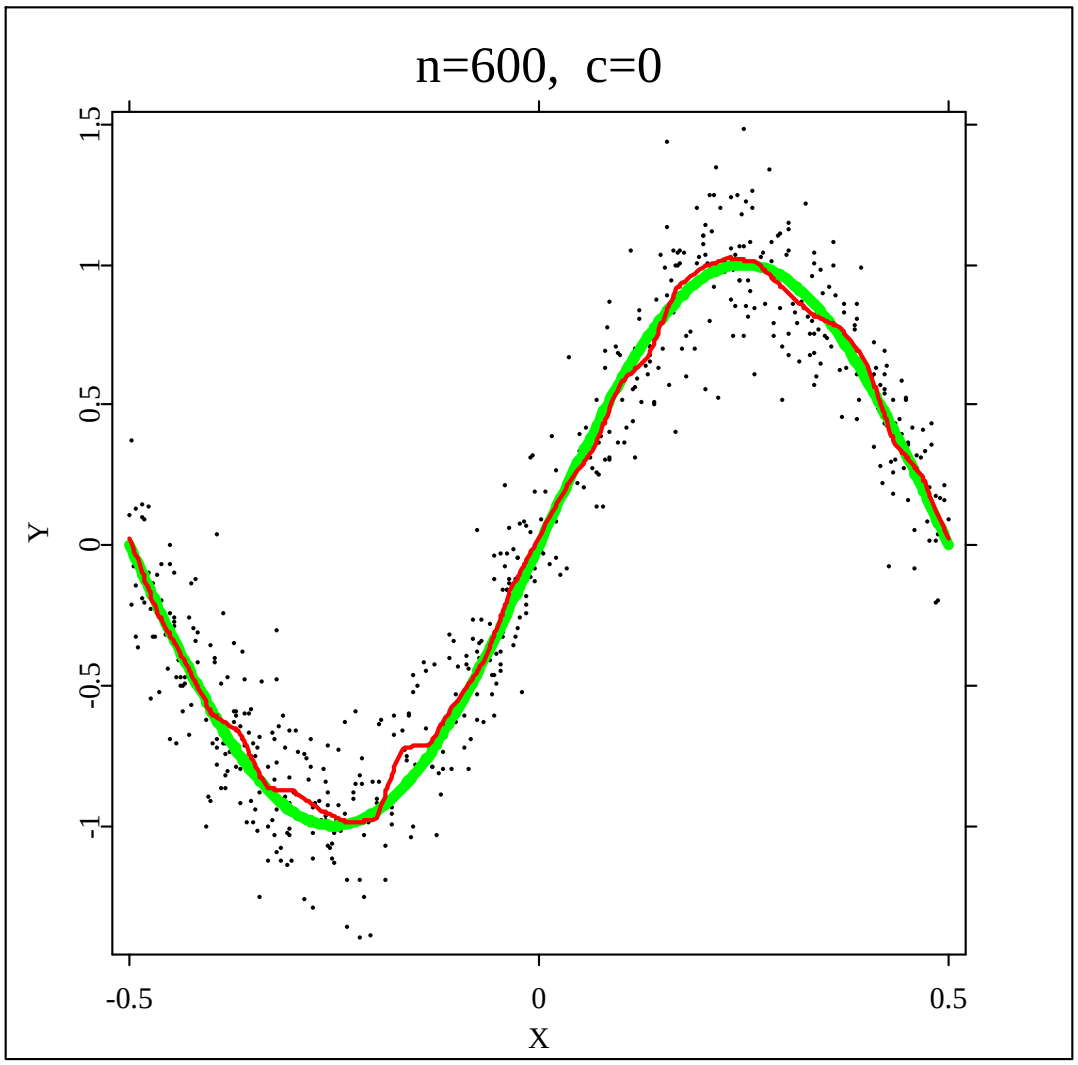


Figure 2.3: Plots of the spline estimator for $c = 0$ in Chapter 2

Note: plots of the true function $m(x)$ (thick solid curve), spline estimator $\hat{m}_2(x)$ (solid curve) and the data scatter plots at $\sigma = 0.2$.

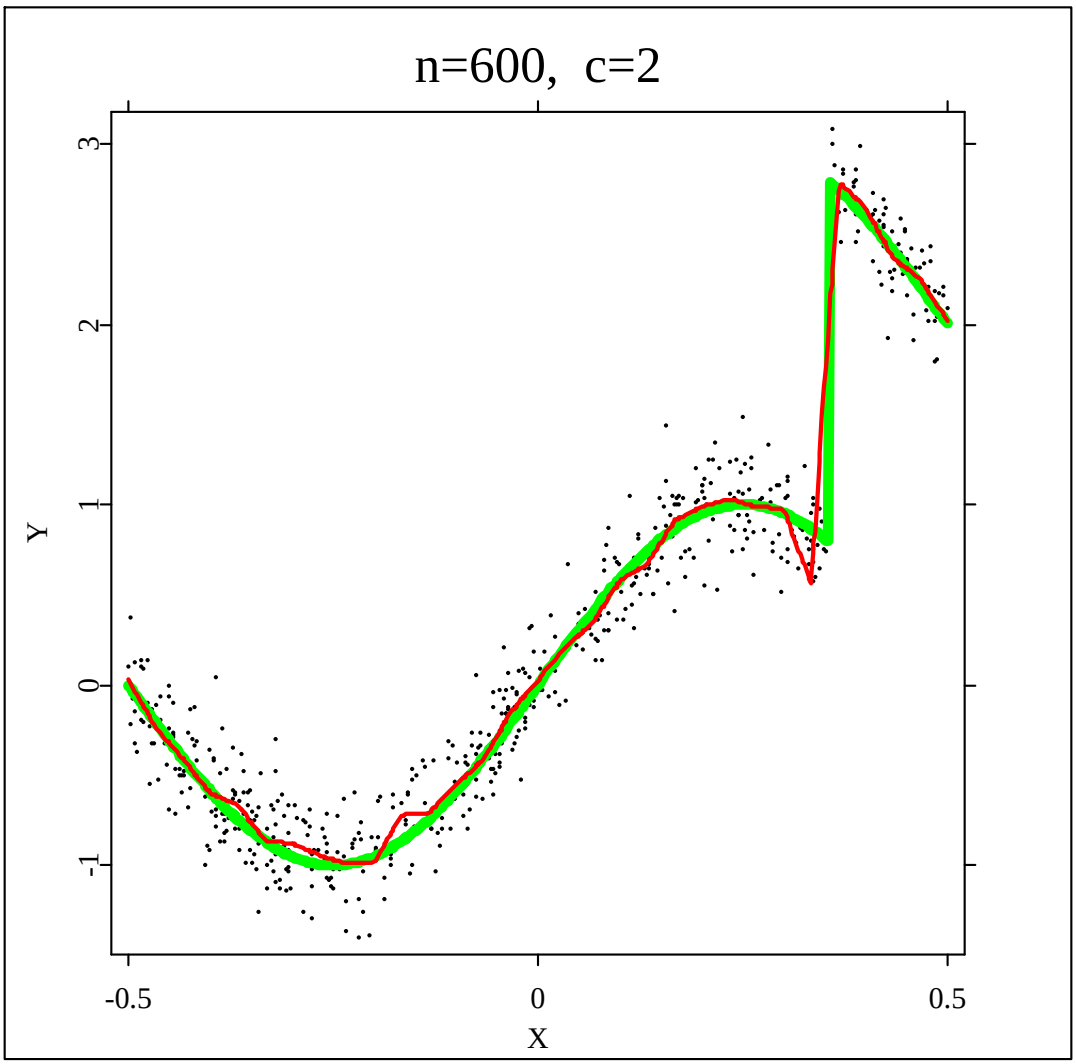


Figure 2.4: Plots of the spline estimator $c = 2$ in Chapter 2

Note: plots of the true function $m(x)$ (thick solid curve), spline estimator $\hat{m}_2(x)$ (solid curve) and the data scatter plots at $\sigma = 0.2$.

from 1945 through 1989. Penny thickness was reduced in World War II and restored to its original thickness sometime around 1960 and reduced again in the 70's. The asymptotic p-value $p_{\text{value},2} < 10^{-20}$. Two jump points are detected with the p-values 0.014468 and 0.00077337, located around the year 1958 with increased magnitude 2.80, and around the year 1974 with decreased magnitude 3.75, respectively, which confirms the result in Gijbels et al. (2007). Figure 2.5 depicts the data points and the spline estimator $\hat{m}_2(x)$ (solid line), which confirms visually these findings. Findings from both simulated and real data demonstrate the effectiveness of our approach in detecting the existence of jumps. The plots of $\hat{m}_2(x)$ in Figures 2.3 and 2.5 give an outline of the true function, without breaking the curve at the jumps. Obtaining jump-preserving spline estimator of the true non-smooth function is beyond the scope of this chapter, but makes an interesting topic for further research.

## 2.5 Appendix A

### 2.5.1 Preliminaries

Denote by $\|\cdot\|_\infty$ the supremum norm of a function $r$ on $[0,1]$, i.e. $\|r\|_\infty = \sup_{x\in[0,1]} |r(x)|$. We denote by the same letters $c, C$, any positive constants without distinction. The following extension of Leadbetter, Lindgren and Rootzén, H. (1983), Theorem 6.2.1 is a key result on the absolute maximum of discrete time Gaussian processes.

LEMMA 2.1. *Let* $\xi_1^{(n)}, ..., \xi_n^{(n)}$ *have jointly normal distribution with* $E\xi_i^{(n)} \equiv 0$, $E\left(\xi_i^{(n)}\right)^2 \equiv 1$, $1 \le i \le n$ *and there exists constants* $C > 0$, $a > 1, r \in (0,1)$ *such that the correlations* $r_{ij} = r_{ij}^{(n)} = E\xi_i^{(n)}\xi_j^{(n)}$, $1 \le i \ne j \le n$ *satisfy* $\left|r_{ij}\right| \le \min\left(r, Ca^{-|i-j|}\right)$ *for*

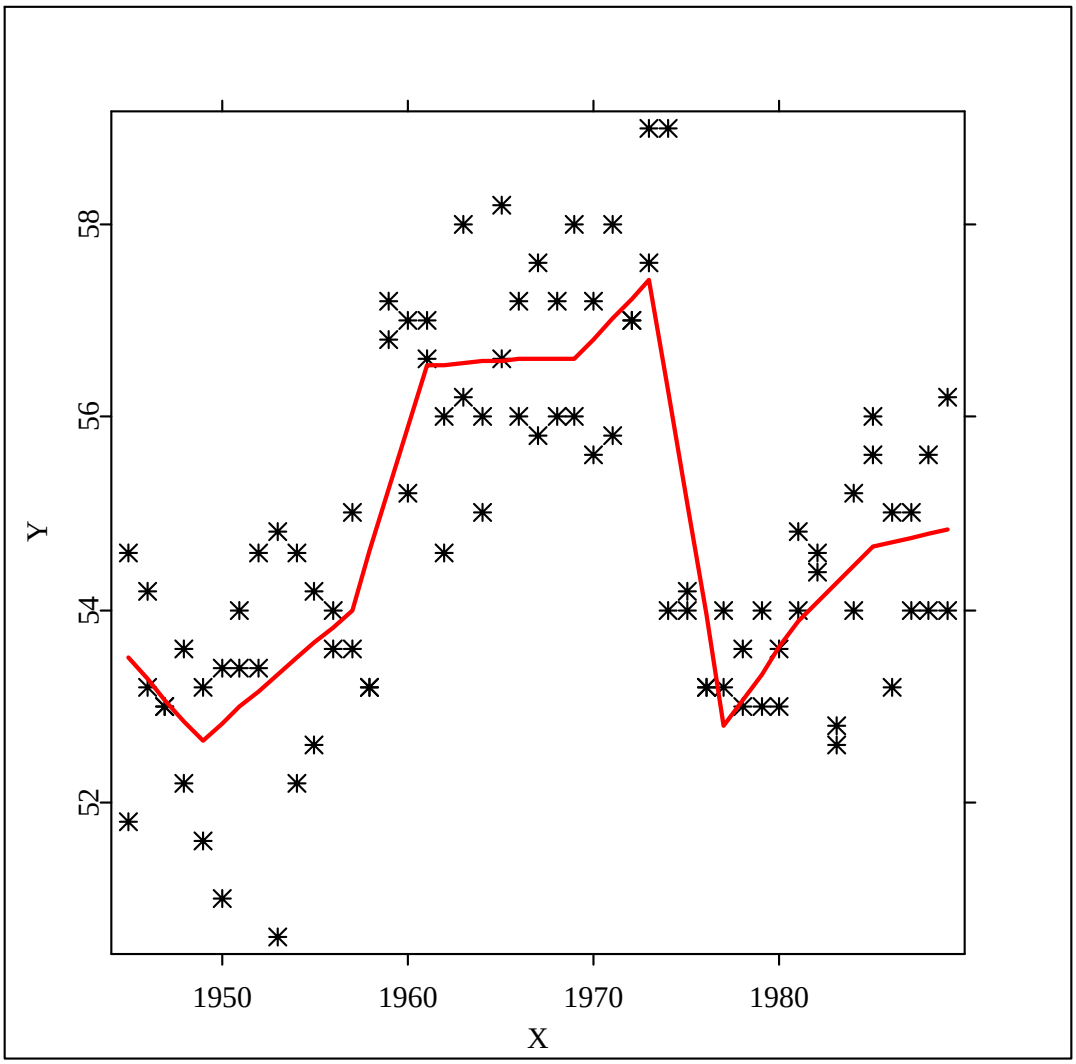


Figure 2.5: Spline estimator for the thickness of pennies data

Note: the thickness of pennies data (points) and the spline estimator $\hat{m}_2(x)$.

$1 \leq i \neq j \leq n$. *Then the absolute maximum* $M_{n,\xi} = \max\left\{\left|\xi_1^{(n)}\right|, ..., \left|\xi_n^{(n)}\right|\right\}$ *satisfies for any* $\tau \in R$, $P\left(M_{n,\xi} \leq \tau/a_n + b_n\right) \to \exp\left(-2e^{-\tau}\right)$, *as* $n \to \infty$, *where* $a_n = (2\log n)^{1/2}$, $b_n = a_n - \frac{1}{2}a_n^{-1}(\log\log n + \log 4\pi)$.

*Proof.* Take $\varepsilon > 0$ such that $(2-\varepsilon)(1+r)^{-1} = 1+\delta$, for some $\delta > 0$. Let $\tau_n = \tau/a_n + b_n$, then $\tau_n^2/(2\log n) \to 1$, as $n \to \infty$, so for large $n$, $\tau_n^2 > (2-\varepsilon)\log n$. By the condition $\left|r_{ij}\right| \leq \min\left(r, Ca^{-|i-j|}\right) < 1$ for $i \neq j$, one has $\left|r_{ij}\right|\left(1-\left|r_{ij}\right|^2\right)^{-1/2} \leq Ca^{-|i-j|}\left(1-r^2\right)^{-1/2}$. Let $M_{n,\eta} = \max\{|\eta_1|, \ldots, |\eta_n|\}$, where $\eta_1, \ldots, \eta_n$ are i.i.d. copies of $N(0,1)$. By Leadbetter, Lindgren and Rootzén (1983), Theorem 1.5.3, $P\left(M_{n,\eta} \leq \tau_n\right) \to \exp\left(-2e^{-\tau}\right)$, as $n \to \infty$. The Normal Comparison Lemma (Leadbetter, Lindgren and Rootzén, H. (1983), Lemma 11.1.2) entails that

$$\left|P\left(-\tau_n < \xi_j^{(n)} \leq \tau_n \text{ for } j = 1, \ldots, n\right) - P\left(-\tau_n < \eta_j \leq \tau_n \text{ for } j = 1, \ldots, n\right)\right|$$
$$\leq (4/2\pi)\sum\nolimits_{1\leq j<i\leq n}\left|r_{ij}\right|\left(1-\left|r_{ij}\right|^2\right)^{-1/2}\exp\left\{-\tau_n^2/\left(1+r_{ij}\right)\right\}.$$

$$\left|P\left(M_{n,\xi} \leq \tau_n\right) - P\left(M_{n,\eta} \leq \tau_n\right)\right| \leq \frac{4}{2\pi}\sum_{1\leq i<j\leq n} Ca^{-|i-j|}\left(1-r^2\right)^{-1/2}\exp\left(\frac{-\tau_n^2}{1+r}\right)$$
$$\leq (4/2\pi)\,C\left(1-r^2\right)^{-1/2}\sum\nolimits_{1\leq j<i\leq n} a^{-(i-j)}\exp\left\{-(2-\varepsilon)(1+r)^{-1}\log n\right\}$$
$$= (4/2\pi)\,C\left(1-r^2\right)^{-1/2}\sum\nolimits_{k=1}^{n-1}(n-k)\,a^{-k}n^{-1-\delta} \leq Cn^{-\delta} \to 0, \text{ as } n \to \infty,$$

for large $n$, hence $P\left(M_{n,\xi} \leq \tau_n\right) \to \exp\left(-2e^{-\tau}\right)$, as $n \to \infty$. ∎

We break the estimation error $\hat{m}_p(x) - m(x)$ into bias and noise. $\hat{m}_p(x)$ defined in (2.2) can be written as $\hat{m}p(x) \equiv \sum_{J=1-p}^{N}\hat{\lambda}_{J,p}B_{J,p}(x)$, where coefficients $\left\{\hat{\lambda}_{1-p,p}, ..., \hat{\lambda}_{N,p}\right\}^T$

are solutions of the following least squares problem

$$\left\{\hat{\lambda}_{1-p,p},...,\hat{\lambda}_{N,p}\right\}^T = \underset{\left\{\lambda_{1-p,p},...,\lambda_{N,p}\right\}\in R^{N+p}}{\operatorname{argmin}} \sum_{i=1}^{n}\left\{Y_i - \sum_{J=1-p}^{N}\lambda_{J,p}B_{J,p}\left(X_i\right)\right\}^2. \tag{2.12}$$

Projecting the relationship in model (3.3) leads to the following decomposition in $G^{(p-2)}$

$$\hat{m}_p\left(x\right) = \tilde{m}_p\left(x\right) + \tilde{\varepsilon}_p\left(x\right), \tag{2.13}$$

$$\tilde{m}_p\left(x\right) = \sum\nolimits_{J=1-p}^{N}\tilde{\lambda}_{J,p}B_{J,p}\left(x\right),\ \tilde{\varepsilon}_p\left(x\right) = \sum\nolimits_{J=1-p}^{N}\tilde{a}_{J,p}B_{J,p}\left(x\right). \tag{2.14}$$

The vectors $\left\{\tilde{\lambda}_{1-p,p},...,\tilde{\lambda}_{N,p}\right\}^T$ and $\left\{\tilde{a}_{1-p,p},...,\tilde{a}_{N,p}\right\}^T$ are solutions to (2.12) with $Y_i$ replaced by $m\left(X_i\right)$ and $\sigma\varepsilon_i$ respectively.

Next lemma is from de Boor (2001), p. 149 and Theorem 5.1 of Huang (2003).

LEMMA 2.2. *There are constants $C_p > 0, p \geq 1$ such that for any $m \in C^{(p)}[0,1]$, there exists a function $g \in G^{(p-2)}[0,1]$ such that $\|g-m\|_\infty \leq C_p\left\|m^{(p)}\right\|_\infty h^p$ and $\tilde{m}\left(x\right)$ defined in (2.14), with probability approaching 1, satisfies $\left\|\tilde{m}_p\left(x\right)-m\left(x\right)\right\|_\infty = O\left(h^p\right)$.*

### 2.5.2 Proof of Theorem 2.1 for $p = 1$

For $x \in [0,1]$, define its location and relative position indices $J\left(x\right), \delta\left(x\right)$ as $J\left(x\right) = J_n\left(x\right) = \min\left\{\left[x/h\right], N\right\}, \delta\left(x\right) = \left\{x - t_{J(x)}\right\}h^{-1}$. It is clear that $t_{J_n(x)} \leq x < t_{J_n(x)+1}$, $0 \leq \delta\left(x\right) < 1, \forall x \in [0,1)$, and $\delta\left(1\right) = 1$. Since $\left\langle B_{J',1},B_{J,1}\right\rangle_n = 0$ unless $J = J'$, for $B_{J,1}\left(x\right)$ given in (2.4). $\tilde{\varepsilon}_1\left(x\right)$ in (2.14) can be written as

$$\tilde{\varepsilon}_1\left(x\right) = \sum\nolimits_{J=0}^{N}\varepsilon^*_{J,1}B_{J,1}\left(x\right)\left\|B_J\right\|^{-2}_{2,n},\ \varepsilon^*_{J,1} = n^{-1}\sum\nolimits_{i=1}^{n}B_{J,1}\left(X_i\right)\sigma\varepsilon_i.$$

Let $\hat{\varepsilon}_1(x) = \sum_{J=0}^{N} \varepsilon_{J,1}^* B_{J,1}(x)$, it is easy to prove that $E\{\hat{\varepsilon}_1(x)\}^2 = \sigma^2/(nh)$ and for $0 \leq J \leq N-1$, $E\left\{\hat{\varepsilon}_1\left(t_{J+1}\right) - \hat{\varepsilon}_1\left(t_J\right)\right\}^2 = 2\sigma^2/(nh)$, which is $\sigma_{n,1}^2$ defined in (2.8). Define for $0 \leq J \leq N-1$, $\tilde{\xi}_{n,1,J} = \sigma_{n,1}^{-1}\left\{\tilde{\varepsilon}_1\left(t_{J+1}\right) - \tilde{\varepsilon}_1\left(t_J\right)\right\}$, $\hat{\xi}_{n,1,J} = \sigma_{n,1}^{-1}\left\{\hat{\varepsilon}_1\left(t_{J+1}\right) - \hat{\varepsilon}_1\left(t_J\right)\right\}$.

LEMMA 2.3. *Under Assumptions (A2)-(A4), as* $n \to \infty$, $\sup_{0 \leq J \leq N-1}\left|\tilde{\xi}_{n,1,J} - \hat{\xi}_{n,1,J}\right| = O_{a.s}\left(n^{-1/2}h^{-1/2}\log n\right) = o_{a.s}(1)$.

LEMMA 2.4. *Under Assumptions (A2)-(A4), there exist* $\left\{\hat{\xi}_{n,1,J}^{(k)}\right\}_{J=0}^{N-1}, k = 1,2,3$ *such that as* $n \to \infty$, $\sup_{0 \leq J \leq N-1}\left|\hat{\xi}_{n,1,J} - \hat{\xi}_{n,1,J}^{(1)}\right| + \sup_{0 \leq J \leq N-1}\left|\hat{\xi}_{n,1,J}^{(2)} - \hat{\xi}_{n,1,J}^{(3)}\right| = o_{a.s}(1)$. $\left\{\hat{\xi}_{n,1,J}^{(1)}\right\}_{J=0}^{N-1}$ *has the same probability distribution as* $\left\{\hat{\xi}_{n,1,J}^{(2)}\right\}_{J=0}^{N-1}$, *and* $\left\{\hat{\xi}_{n,1,J}^{(3)}\right\}_{J=0}^{N-1}$ *is a Gaussian process with mean* 0, *variance* 1, *and covariance*

$$\operatorname{cov}\left\{\hat{\xi}_{n,1,J}^{(3)}, \hat{\xi}_{n,1,J'}^{(3)}\right\} = \begin{cases} -1/2, \text{ for } \left|J - J'\right| = 1 \\ 0, \text{ for } \left|J - J'\right| > 1 \end{cases}.$$

Lemmas 2.3 and 2.4 follow from Appendix A of Wang and Yang (2009a).

**Proof of Theorem 2.1 for p = 1:** It is clear from Lemma 2.4 that the Gaussian process $\left\{\hat{\xi}_{n,1,J}^{(3)}\right\}_{J=0}^{N-1}$ satisfies the conditions of Lemma 2.1, hence as $n \to \infty$,

$$P\left\{\left(\sup_{0 \leq J \leq N-1}\left|\hat{\xi}_{n,1,J}^{(3)}\right| \leq \tau/a_N + b_N\right)\right\} \to \exp\left(-2e^{-\tau}\right).$$

By letting $\tau = -\log\{-(1/2)\log(1-\alpha)\}$, and using the definition of $a_N$, $b_N$ and $d_N(\alpha)$

we obtain

$$\lim_{n\to\infty} P\left\{\sup_{0\leq J\leq N-1}\left|\hat{\xi}^{(3)}_{n,1,J}\right| \leq -\log\left\{-(1/2)\log(1-\alpha)\right\}(2\log N)^{-1/2}\right.$$
$$\left.+(2\log N)^{1/2}-(1/2)(2\log N)^{-1/2}(\log\log N+\log 4\pi)\right\}=1-\alpha.$$
$$\lim_{n\to\infty} P\left\{\sup_{0\leq J\leq N-1}\left|\hat{\xi}^{(3)}_{n,1,J}\right| \leq (2\log N)^{1/2} d_N(\alpha)\right\}=1-\alpha.$$

By Lemmas 2.3 and 2.4, we have

$$\lim_{n\to\infty} P\left\{\sup_{0\leq J\leq N-1}\left|\tilde{\xi}_{n,1,J}\right| \leq (2\log N)^{1/2} d_N(\alpha)\right\}=1-\alpha,$$

which implies for $0\leq J\leq N-1$

$$\lim_{n\to\infty} P\left\{d_N(\alpha)^{-1}(2\log N)^{-1/2}\sigma_{n,1}^{-1}\sup_{0\leq J\leq N-1}\left|\tilde{\varepsilon}_1\left(t_{J+1}\right)-\tilde{\varepsilon}_1\left(t_J\right)\right|\leq 1\right\}=1-\alpha.$$

Lemma 2.2 entails that under $\mathcal{H}_0$ $\sup_{0\leq J\leq N-1}\left|\tilde{m}_1\left(t_J\right)-m\left(t_J\right)\right|=O_p(h)$ and

$$\sup_{0\leq J\leq N-1}\left|m\left(t_{J+1}\right)-m\left(t_J\right)\right|=O_p(h),$$

which imply that

$$\begin{aligned}&\sigma_{n,1}^{-1}(\log N)^{-1/2}\sup_{0\leq J\leq N-1}\left|m\left(t_{J+1}\right)-m\left(t_J\right)\right|\\ =\;& O_p\left\{(nh)^{1/2}(\log N)^{-1/2}h\right\}=o_p\left\{(\log n)^{-2}\right\}.\end{aligned}$$

By (2.13), $\hat{m}_1\left(t_{J+1}\right)-\hat{m}_1\left(t_J\right)=\left\{\tilde{m}_1\left(t_{J+1}\right)-m\left(t_{J+1}\right)\right\}-\left\{\tilde{m}_1\left(t_J\right)-m\left\{t_J\right\}\right\}+$

$\left\{m\left(t_{J+1}\right)-m\left(t_J\right)\right\}+\left\{\tilde{\varepsilon}_1\left(t_{J+1}\right)-\tilde{\varepsilon}_1\left(t_J\right)\right\}$, then for $d_N(\alpha)$ defined in (2.9),

$$\lim_{n\to\infty} P\left\{\sup_{0\le J\le N-1}\left|\hat{m}_1\left(t_{J+1}\right)-\hat{m}_1\left(t_J\right)\right|\le\sigma_{n,1}\left(2\log N\right)^{1/2}d_N(\alpha)\right\}=$$

$$\lim_{n\to\infty} P\left\{\sup_{0\le J\le N-1}\left|\tilde{\varepsilon}_1\left(t_{J+1}\right)-\tilde{\varepsilon}_1\left(t_J\right)\right|\le\sigma_{n,1}\left(2\log N\right)^{1/2}d_N(\alpha)\right\}=1-\alpha.\ \square$$

## 2.6 Appendix B

### 2.6.1 Preliminaries

The following lemma from Wang and Yang (2009a) shows that multiplication by $\mathbf{V}$ defined in (2.5) behaves similarly to multiplication by a constant. We use $|T|$ to denote the maximal absolute value of any matrix $T$.

LEMMA 2.5. *Given matrix* $\Omega=\mathbf{V}+\Gamma$, *in which* $\Gamma=\left(\gamma_{jj'}\right)_{J,J'=-1}^{N}$ *satisfies* $\gamma_{jj'}\equiv 0$ *if* $\left|J-J'\right|>1$ *and* $|\Gamma|\xrightarrow{p}0$. *Then there exist constants* $c,C>0$ *independent of* $n$ *and* $\Gamma$, *such that with probability approaching* 1

$$c\left|\boldsymbol{\xi}\right|\le\left|\Omega\boldsymbol{\xi}\right|\le C\left|\boldsymbol{\xi}\right|,\ C^{-1}\left|\boldsymbol{\xi}\right|\le\left|\Omega^{-1}\boldsymbol{\xi}\right|\le c^{-1}\left|\boldsymbol{\xi}\right|,\ \forall\boldsymbol{\xi}\in R^{N+2}. \tag{2.15}$$

To prove Theorem 2.1 for $p=2$, we need the result below (Corollary 16 of Kılıç (2008)), which gives explicit formula for the inverse of symmetric tridiagonal matrix.

LEMMA 2.6. *For any symmetric tridiagonal matrix* $G_n=\begin{pmatrix} x_1 & y_1 & & \\ y_1 & x_2 & \ddots & \\ & \ddots & \ddots & y_{n-1} \\ & & y_{n-1} & x_n \end{pmatrix}$, *the*

*inverse of the matrix* $G_n$, $G_n^{-1} = \left[w_{ij}\right]$ *is given by*

$$w_{ij} = \begin{cases} \left(C_i^b\right)^{-1} + \sum_{k=i+1}^{n} \left(C_k^b\right)^{-1} \prod_{t=i}^{k-1} \left(C_t^b\right)^{-2} y_t^2, \; i = J \\ (-1)^{i+J} \left\{\prod_{t=J}^{i-1} \left(C_t^b\right)^{-1} y_t\right\} w_{ii}, \; i > J \end{cases}$$

*in which* $C_1^b = x_1$, $C_n^b = x_n - \left(C_{n-1}^b\right)^{-1} y_{n-1}^2$, $n = 2, 3, ....$

LEMMA 2.7. *There exists a constant* $C_S > 0$, *such that* $\sum_{J=-1}^{N} \left|s_{J'J}\right| \leq C_S$, *and* $17/16 \leq s_{jj} \leq 5/4$, *where* $s_{J'J}$, $0 \leq J, J' \leq N-1$, *is the element of* $\mathbf{S}$ *defined in (2.6).*

*Proof.* By (2.15), let $\tilde{\boldsymbol{\xi}}_{J'} = \left\{\operatorname{sgn}\left(s_{J'J}\right)\right\}_{J=-1}^{N}$, then $\sum_{J=-1}^{N} \left|s_{J'J}\right| \leq \left|\mathbf{S}\tilde{\boldsymbol{\xi}}_{J'}\right| \leq C_S \left|\tilde{\boldsymbol{\xi}}_{J'}\right| = C_S, \forall J' = -1, 0, ..., N$. Applying Lemma 2.6 to $\mathbf{V}$ with $x_{-1} = \cdots = x_N = 1$, $y_{-1} = y_{N-1} = \sqrt{2}/4$, $y_0 = \cdots = y_{N-1} = 1/4$, $s_{jj} = \left(C_J^b\right)^{-1} + \sum_{k=J+1}^{N} \left(C_k^b\right)^{-1} \prod_{t=J}^{k-1} \left(C_t^b\right)^{-2} y_t^2$. By mathematical induction, one obtains that $9/10 \leq C_J^b \leq 1$, for $0 \leq J \leq N-1$. So, $1 \leq \left(C_J^b\right)^{-1} \leq 10/9$, and for $0 \leq J \leq N-1$,

$$s_{jj} \geq 1 + \sum_{k=J+1}^{N} \prod_{t=J}^{k-1} y_t^2 \geq 1 + \sum_{k=J+1}^{N} (16)^{-(k-J)} \geq 17/16,$$
$$s_{jj} \leq 10/9 + (10/9) \sum_{k=J+1}^{N} (1/9)^{k-J} \leq 5/4.$$

∎

### 2.6.2 Variance calculation

Vector $\tilde{\mathbf{a}}_2 = \left(\tilde{a}_{-1,2}, ..., \tilde{a}_{N,2}\right)^T$ given in (2.14) solves the normal equations,

$$\left(\left\langle B_{J,2}, B_{J',2} \right\rangle_n\right)_{J,J'=-1}^{N} \tilde{\mathbf{a}}_2 = \left(n^{-1} \sum_{i=1}^{n} B_{J,2}\left(X_i\right) \sigma\varepsilon_i\right)_{J=-1}^{N},$$

for $B_{J,2}(x)$ in (2.4). In other words, $\tilde{\mathbf{a}}_2 = \left(V + \tilde{B}\right)^{-1} \left(n^{-1} \sum_{i=1}^{n} B_{J,2}(X_i)\, \sigma\varepsilon_i\right)_{J=-1}^{N}$, where $\tilde{B} = \left(\left\langle B_{J,2}, B_{J',2} \right\rangle_n\right)_{J,J'=-1}^{N} - V$ satisfies $\left|\tilde{B}\right| = O_p\left(\sqrt{n^{-1}h^{-1}\log(n)}\right)$ according to Subsection B.2 of the supplement to Wang and Yang (2009a).

Now define $\hat{\mathbf{a}}_2 = \left(\hat{a}_{-1,2}, ..., \hat{a}_{N,2}\right)^T$ by replacing $\left(\mathbf{V} + \tilde{B}\right)^{-1}$ with $\mathbf{V}^{-1} = \mathbf{S}$ in above formula, i.e., $\hat{\mathbf{a}}_2 = \left(\sum_{J=-1}^{N} s_{J'J} n^{-1} \sum_{i=1}^{n} B_{J,2}(X_i)\, \sigma\varepsilon_i\right)_{J'=-1,..,N}$, and define for $x \in [0,1]$

$$\hat{\varepsilon}_2(x) = \sum_{J=-1}^{N} \hat{a}_{J,2} B_{J,2}(x) = \sum_{J,J'=-1}^{N} s_{J'J} n^{-1} \sum_{i=1}^{n} B_{J,2}(X_i)\, \sigma\varepsilon_i B_{J',2}(x),$$

$$\hat{\xi}_{2,J} = \left\{\hat{\varepsilon}_2\left(t_{J+1}\right) + \hat{\varepsilon}_2\left(t_{J-1}\right)\right\}/2 - \hat{\varepsilon}_2\left(t_J\right), 2 \le J \le N-1, \tag{2.16}$$

$$\mathbf{D} = \begin{pmatrix} 0 & 1 & -2 & 1 & & & & 0 & 0 \\ \vdots & & \ddots & \ddots & \ddots & & & & \vdots \\ \vdots & & & \ddots & \ddots & \ddots & & & \vdots \\ 0 & 0 & & & 1 & -2 & 1 & 0 & \end{pmatrix}_{(N-2)\times(N+2)}. \tag{2.17}$$

LEMMA 2.8. *With* $\mathbf{S}$ *and* $\mathbf{D}$ *given in (2.6) and (2.17),* $\left\{\hat{\xi}_{2,J}\right\}_{J=2}^{N-1}$ *has covariance matrix*

$$\left[\operatorname{cov}\left(\hat{\xi}_{2,J}, \hat{\xi}_{2,J'}\right)\right]_{J,J'=2}^{N-1} = \sigma^2 (8nh/3)^{-1} \mathbf{D}\mathbf{S}\mathbf{D}^T. \tag{2.18}$$

*Proof.* For $0 \leq J, J' \leq N+1$, $\hat{\varepsilon}_2\left(t_J\right) = \sum_{k,k'=-1}^{N} {}^{s}{}_{k'k} n^{-1} \sum_{i=1}^{n} B_{k,2}\left(X_i\right) \sigma\varepsilon_i B_{k',2}\left(t_J\right)$

$$= \sigma \sum_{k=-1}^{N} n^{-1} \sum_{i=1}^{n} B_{k,2}\left(X_i\right) \varepsilon_i {}^{s}{}_{(J-1),k} B_{(J-1),2}\left(t_J\right).$$

$$E\left[\hat{\varepsilon}_2\left(t_J\right) \hat{\varepsilon}_2\left(t_{J'}\right)\right] = \sigma^2 E\left[\sum_{k=-1}^{N} n^{-1} \sum_{i=1}^{n} B_{k,2}\left(X_i\right) \varepsilon_i {}^{s}{}_{J-1,k} B_{J-1,2}\left(t_J\right)\right]$$

$$\times \left[\sum_{k'=-1}^{N} n^{-1} \sum_{i=1}^{n} B_{k',2}\left(X_i\right) \varepsilon_i {}^{s}{}_{J'-1,k'} B_{J'-1,2}\left(t_{J'}\right)\right]$$

$$= \sigma^2 n^{-1} \sum_{k,k'=-1}^{N} B_{J-1,2}\left(t_J\right) B_{J'-1,2}\left(t_{J'}\right) {}^{s}{}_{J-1,k} {}^{s}{}_{J'-1,k'} E B_{k,2}\left(X\right) B_{k',2}\left(X\right)$$

$$= \sigma^2 n^{-1} \sum_{k,k'=-1}^{N} B_{J-1,2}\left(t_J\right) B_{J'-1,2}\left(t_{J'}\right) {}^{s}{}_{J-1,k} {}^{s}{}_{J'-1,k'} {}^{v}{}_{k,k'}$$

$$= \sigma^2 n^{-1} B_{J-1,2}\left(t_J\right) B_{J'-1,2}\left(t_{J'}\right) \sum_{k'=-1}^{N} {}^{s}{}_{J'-1,k'} \sum_{k=-1}^{N} {}^{s}{}_{J-1,k} {}^{v}{}_{k,k'}$$

$$= \sigma^2 n^{-1} B_{J-1,2}\left(t_J\right) B_{J'-1,2}\left(t_{J'}\right) \sum_{k'=-1}^{N} {}^{s}{}_{J'-1,k'} \delta_{J-1,k'}$$

$$= \sigma^2 n^{-1} B_{J-1,2}\left(t_J\right) B_{J'-1,2}\left(t_{J'}\right) {}^{s}{}_{J'-1,J-1} = \sigma^2 n^{-1} d_{J-1,n}^{-1/2} d_{J'-1,n}^{-1/2} {}^{s}{}_{J'-1,J-1}.$$

By definitions of $\hat{\xi}_{2,J}, d_{J,n}$ in (2.16), (2.3), for $2 \leq J, J' \leq N-1$, $E\left(\hat{\xi}_{2,J}\hat{\xi}_{2,J'}\right)$ is

$$\sigma^2 \left(8nh/3\right)^{-1} \left({}^{s}{}_{J',J} + {}^{s}{}_{J'-2,J} - 2{}^{s}{}_{J'-1,J} + {}^{s}{}_{J',J-2} + {}^{s}{}_{J'-2,J-2}\right.$$

$$\left. -2{}^{s}{}_{J'-1,J-2} - 2{}^{s}{}_{J',J-1} - 2{}^{s}{}_{J'-2,J-1} + 4{}^{s}{}_{J'-1,J-1}\right)$$

$$= \sigma^2 \left(8nh/3\right)^{-1} \begin{pmatrix} 1 & -2 & 1 \end{pmatrix} \begin{pmatrix} {}^{s}{}_{J'-2,J-2} & {}^{s}{}_{J'-2,J-1} & {}^{s}{}_{J'-2,J} \\ {}^{s}{}_{J'-1,J-2} & {}^{s}{}_{J'-1,J-1} & {}^{s}{}_{J'-1,J} \\ {}^{s}{}_{J',J-2} & {}^{s}{}_{J',J-1} & {}^{s}{}_{J',J} \end{pmatrix} \begin{pmatrix} 1 & -2 & 1 \end{pmatrix}^T.$$

Therefore, for $2 \leq J, J' \leq N-1$, $\left[\operatorname{cov}\left(\hat{\xi}_{2,J}, \hat{\xi}_{2,J'}\right)\right]_{J,J'=2}^{N-1} = \sigma^2 \left(8nh/3\right)^{-1} \mathbf{DSD}^T$. ∎

LEMMA 2.9. *For $2 \leq J \leq N-1$, $\sigma^2_{n,2,J}$ defined in (2.8) satisfies that $c_\sigma (8nh/3)^{-1} \sigma^2 \leq \sigma^2_{n,2,J} \leq C_\sigma (8nh/3)^{-1} \sigma^2$, for $c_\sigma = (65/8)(17/16)$, $C_\sigma = 100/9$.*

*Proof.* It follows from (2.18) that $\sigma^2_{n,2,J} = E\hat{\xi}^2_{2,J}$. Then by Lemmas 2.6 and 2.8, for $2 \leq J \leq N-1$, $\left\{\sigma^2 (8nh/3)^{-1}\right\}^{-1} \sigma^2_{n,2,J}$ is

$$\begin{gathered} s_{J,J} - 4s_{J,J-1} + 2s_{J,J-2} + 4s_{J-1,J-1} - 4s_{J-1,J-2} + s_{J-2,J-2} \\ = s_{J-2,J-2} + 4\left\{\left(C^b_{J-2}\right)^{-1} y_{J-2} + 1\right\} s_{J-1,J-1} + \\ \left\{2\left(C^b_{J-2}C^b_{J-1}\right)^{-1} y_{J-2}y_{J-1} + 4\left(C^b_{J-1}\right)^{-1} y_{J-1} + 1\right\} s_{J,J}, \end{gathered}$$

$$\begin{aligned} \text{thus, } \sigma^2_{n,2,J} &\leq \{1 + 4(1/3+1) + (2/9+4/3+1)\}(5/4)\sigma^2 (8nh/3)^{-1} \\ &= (100/9)(8nh/3)^{-1}\sigma^2, \\ \sigma^2_{n,2,J} &\geq \{1 + 4(1/4+1) + (2/16+1+1)\}(17/16)\sigma^2 (8nh/3)^{-1} \\ &= (65/8)(17/16)(8nh/3)^{-1}\sigma^2. \end{aligned}$$

∎

### 2.6.3 Proof of Theorem 2.1 for $p = 2$

Several lemmas will be given below for proving Theorem 2.1 for $p = 2$. With $\tilde{\varepsilon}_2(x)$, $\hat{\xi}_{2,J}$ and $\sigma_{n,2,J}$ defined in (2.14), (2.16) and (2.8), define for $2 \leq J \leq N-1$

$$\tilde{\xi}_{n,2,J} = \sigma^{-1}_{n,2,J}\left[\left\{\tilde{\varepsilon}_2(t_{J+1}) + \tilde{\varepsilon}_2\left(t_{J-1}\right)\right\}/2 - \tilde{\varepsilon}_2(t_J)\right], \ \hat{\xi}_{n,2,J} = \sigma^{-1}_{n,2,J}\hat{\xi}_{2,J}, \quad (2.19)$$

LEMMA 2.10. *Under Assumptions (A2)-(A4), as* $n \to \infty$, $\sup_{2 \le J \le N-1} \left| \hat{\xi}_{n,2,J} - \tilde{\xi}_{n,2,J} \right| = O_{a.s}\left( \sqrt{\log n / (nh)} \right) = o_{a.s}(1)$.

LEMMA 2.11. *Under Assumptions (A2)-(A4), there exist* $\left\{ \hat{\xi}^{(k)}_{n,2,J} \right\}_{J=2}^{N-1}, k = 1,2,3,$ *such that as* $n \to \infty$, $\sup_{2 \le J \le N-1} \left| \hat{\xi}_{n,2,J} - \hat{\xi}^{(1)}_{n,2,J} \right| + \sup_{2 \le J \le N-1} \left| \hat{\xi}^{(2)}_{n,2,J} - \hat{\xi}^{(3)}_{n,2,J} \right| = o_{a.s}(1)$. $\hat{\xi}^{(1)}_{n,2,J}$ *has the same probability distribution as* $\hat{\xi}^{(2)}_{n,2,J}$. $\left\{ \hat{\xi}^{(3)}_{n,2,J} \right\}$ *is a Gaussian process with mean* 0, *variance* 1, *covariance* $r^{\xi}_{J,J'} = \operatorname{cov}\left( \hat{\xi}^{(3)}_{n,2,J}, \hat{\xi}^{(3)}_{n,2,J'} \right) = \sigma^{-1}_{n,2,J} \sigma^{-1}_{n,2,J'} E\left( \hat{\xi}_{2,J} \hat{\xi}_{2,J'} \right)$ *for which there exist constants* $0 < C, 0 < r < 1$ *such that for large* $n$, $\left| r^{\xi}_{J,J'} \right| \le \min\left( r, C3^{-\left| J - J' \right|} \right), 2 \le J, J' \le N-1$.

*Proof.* We only prove $\left| r^{\xi}_{J,J'} \right| \le \min\left( r, C3^{-\left| J - J' \right|} \right)$. Lemma 2.10 and the rest of Lemma 2.11 follow from Appendix B of the supplement to Wang and Yang (2009a). By (2.18),

$$\sigma^{-2} (8nh/3)^{-1} E\left( \hat{\xi}_{2,J} \hat{\xi}_{2,J'} \right) = s_{J',J} + s_{J'-2,J} - 2s_{J'-1,J} + s_{J',J-2} + s_{J'-2,J-2}$$
$$-2s_{J'-1,J-2} - 2s_{J',J-1} - 2s_{J'-2,J-1} + 4s_{J'-1,J-1}.$$

By Lemma 2.6, for $-1 \le J' < J \le N$, $s_{J,J'} = (-1)^{J+J'} \prod_{t=J'}^{J-1} \left( C^b_t \right)^{-1} y_t s_{jj}$, then for

$2 \leq J, J' \leq N-1$ and $J - J' > 2$, by Lemma 2.7,

$$\begin{aligned}
&\left\{\sigma^2 (8nh/3)^{-1}\right\}^{-1} \left|E\left[\hat{\xi}_{2,J}\hat{\xi}_{2,J'}\right]\right| \\
&= \left|(-1)^{J+J'} \left\{\left(C^b_{J'-2}\right)^{-1} y_{J'-2} + 2\left(C^b_{J'-1}\right)^{-1} y_{J'-1} + 1\right\}\right. \\
&\quad \times \left\{s_{J-2,J-2} + 2\left(C^b_{J-2}\right)^{-1} y_{J-2} s_{J-1,J-1} + \right. \\
&\quad \left.\left.\left(C^b_{J-2} C^b_{J-1}\right)^{-1} y_{J-2} y_{J-1} s_{J,J}\right\} \prod_{t=J'}^{J-3} \left(C^b_t\right)^{-1} y_t\right| \\
&\leq (5/4)(1/3 + 2/3 + 1)\left\{1 + 2/3 + (1/3)^2\right\} 3^{-\left(J-J'-2\right)} \leq 40 \times 3^{-\left(J-J'\right)}.
\end{aligned}$$

By Lemma 2.9, $\left\{\sigma^2 (8nh/3)^{-1}\right\}^{-1} \sigma^2_{n,2,J} \geq (65/8)(17/16)$, for $2 \leq J \leq N-1$. Therefore, for $2 \leq J, J' \leq N-1$ and $J - J' > 2$, $\left|r^{\xi}_{J,J'}\right| \leq C 3^{-\left(J-J'\right)} \leq r < 1$, with $C = 40(8/65)(16/17)$ and $r = 40(8/65)(16/17)/3^3 < 1$. For $J - J' = 1, 2$, the result can be proved following the same procedure above. ∎

**Proof of Theorem 2.1 for p = 2:** It is clear from Lemma 2.11 that the Gaussian process $\left\{\hat{\xi}^{(3)}_{n,2,J}\right\}_{J=2}^{N-1}$ satisfies the conditions of Lemma 2.1, hence as $n \to \infty$,

$$P\left(\sup_{2\leq J\leq N-1}\left|\hat{\xi}^{(3)}_{n,2,J}\right| \leq \tau/a_N + b_N\right) \to \exp\left(-2e^{-\tau}\right).$$

By Lemmas 2.10, 2.11, with $\tau = -\log\{-(1/2)\log(1-\alpha)\}$ and definitions of $a_N$ and $b_N$,

$$\lim_{n\to\infty} P\left(\sup_{2\leq J\leq N-1}\left|\tilde{\xi}_{n,2,J}\right| \leq \{2\log(N-2)\}^{1/2} d_{N-2}(\alpha)\right) = 1 - \alpha,$$

for any $0 < \alpha < 1$, $\hat{\xi}_{n,2,J}$ and $d_N(\alpha)$ defined in (2.19) and (2.9). By (2.13) and (2.16),

$\left\{\hat{m}_2\left(t_{J+1}\right)+\hat{m}_2\left(t_{J-1}\right)\right\}/2-\hat{m}_2\left(t_J\right)$ is

$$\left\{\tilde{m}_2\left(t_{J+1}\right)-m\left(t_{J+1}\right)\right\}/2+\left\{\tilde{m}_2\left(t_{J-1}\right)-m\left(t_{J-1}\right)\right\}/2-\left\{\tilde{m}_2\left(t_J\right)-m\left(t_J\right)\right\}$$
$$+\left[\left\{m\left(t_{J+1}\right)+m\left(t_{J-1}\right)\right\}/2-m\left(t_J\right)\right]+\hat{\xi}_{2,J+1}$$

Now Lemma 2.2 implies that under $\mathcal{H}_0$, $\left\|\tilde{m}_2-m\right\|_\infty=O_p\left(h^2\right)$, hence

$$(nh)^{1/2}\left\{\log\left(N-2\right)\right\}^{-1/2}\times$$
$$\sup_{2\leq J\leq N-1}\left|\left\{\tilde{m}_2\left(t_{J+1}\right)-m\left(t_{J+1}\right)\right\}/2+\left\{\tilde{m}_2\left(t_{J-1}\right)-m\left(t_{J-1}\right)\right\}/2\right.$$
$$\left.-\left\{\tilde{m}_2\left(t_J\right)-m\left(t_J\right)\right\}\right|=O_p\left[(nh)^{1/2}\left\{\log\left(N-2\right)\right\}^{-1/2}h^2\right]=o_p\left\{\left(\log n\right)^{-3}\right\}.$$

By Taylor expansion, $\sup_{2\leq J\leq N-1}\left|\left\{m\left(t_{J+1}\right)+m\left(t_{J-1}\right)\right\}/2-m\left(t_J\right)\right|=O_p\left(h^2\right)$ under $\mathcal{H}_0$, as $n\to\infty$. Hence

$$(nh)^{1/2}\left\{\log\left(N-2\right)\right\}^{-1/2}\sup_{2\leq J\leq N-1}\left|\left\{m\left(t_{J+1}\right)+m\left(t_{J-1}\right)\right\}/2-m\left(t_J\right)\right|$$
$$=\ O_p\left[\sqrt{nh}\left\{\log\left(N-2\right)\right\}^{-1/2}h^2\right]=o_p\left\{\left(\log n\right)^{-3}\right\}.$$

By above results, for $T_{2n}$ defined in (2.7),

$\lim_{n\to\infty}P\left\{T_{2n}\leq\left\{2\log\left(N-2\right)\right\}^{1/2}d_{N-2}\left(\alpha\right)\right\}=1-\alpha.$ □

# Chapter 3

# A Simultaneous Confidence Band for Sparse Longitudinal Regression

## 3.1 Introduction

This chapter is based on Ma, Yang and Carroll (2011). Functional data analysis (FDA) has in recent years become a focal area in statistics research, and much has been published in this area. An incomplete list includes Cardot, Ferraty, and Sarda (2003), Cardot and Sarda (2005), Ferraty and Vieu (2006), Hall and Heckman (2002), Hall, Müller, and Wang (2006), Izem and Marron (2007), James, Hastie, and Sugar (2000), James (2002), James and Silverman (2005), James and Sugar (2003), Li and Hsing (2007), Li and Hsing (2010), Morris and Carroll (2006), Müller and Stadtmüller (2005), Müller, Stadtmüller, and Yao (2006), Müller and Yao (2008), Ramsay and Silverman (2005), Wang, Carroll, and Lin (2005), Yao and Lee (2006), Yao, Müller, and Wang (2005a), Yao, Müller, and Wang (2005b), Yao (2007), Zhang and Chen (2007), Zhao, Marron, and Wells (2004), and Zhou, Huang, and Carroll

(2008). According to Ferraty and Vieu (2006), a functional data set consists of iid realizations $\left\{\xi_i(x), x \in \mathcal{X}\right\}, 1 \leq i \leq n$, of a smooth stochastic process (random curve) $\{\xi(x), x \in \mathcal{X}\}$ over an entire interval $\mathcal{X}$. A more data oriented alternative in Ramsay and Silverman (2005) emphasizes smooth functional features inherent in discretely observed longitudinal data, so that the recording of each random curve $\xi_i(x)$ is over a finite number of points in $\mathcal{X}$, and contaminated with noise. This second view is taken in this chapter.

A typical functional data set therefore has the form $\left\{X_{ij}, Y_{ij}\right\}, 1 \leq i \leq n, 1 \leq j \leq N_i$, in which $N_i$ observations are taken for the $i^{th}$ subject, with $X_{ij}$ and $Y_{ij}$ the $j^{th}$ predictor and response variables, respectively, for the $i^{th}$ subject. Generally, the predictor $X_{ij}$ takes values in a compact interval $\mathcal{X} = [a, b]$. For the $i^{th}$ subject, its sample path $\left\{X_{ij}, Y_{ij}\right\}$ is the noisy realization of a continuous time stochastic process $\xi_i(x)$ in the sense that

$$Y_{ij} = \xi_i\left(X_{ij}\right) + \sigma\left(X_{ij}\right)\varepsilon_{ij}, \tag{3.1}$$

with errors $\varepsilon_{ij}$ satisfying $E\left(\varepsilon_{ij}\right) = 0, E(\varepsilon_{ij}^2) = 1$, and $\left\{\xi_i(x), x \in \mathcal{X}\right\}$ are iid copies of a process $\{\xi(x), x \in \mathcal{X}\}$ which is $L^2$, i.e., $E\int_{\mathcal{X}} \xi^2(x)dx < +\infty$.

For the standard process $\{\xi(x), x \in \mathcal{X}\}$, one defines the mean function $m(x) = E\{\xi(x)\}$ and the covariance function $G\left(x, x'\right) = \text{cov}\left\{\xi(x), \xi(x')\right\}$. Let sequences $\left\{\lambda_k\right\}_{k=1}^{\infty}$, $\left\{\psi_k(x)\right\}_{k=1}^{\infty}$ be the eigenvalues and eigenfunctions of $G\left(x, x'\right)$, respectively, in which $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0, \sum_{k=1}^{\infty}\lambda_k < \infty$, $\left\{\psi_k\right\}_{k=1}^{\infty}$ form an orthonormal basis of $L^2(\mathcal{X})$ and $G\left(x, x'\right) = \sum_{k=1}^{\infty}\lambda_k\psi_k(x)\psi_k\left(x'\right)$, which implies that $\int G\left(x, x'\right)\psi_k\left(x'\right)dx' = \lambda_k\psi_k(x)$.

The process $\{\xi_i(x), x \in \mathcal{X}\}$ allows the Karhunen-Loève $L^2$ representation

$$\xi_i(x) = m(x) + \sum_{k=1}^{\infty} \xi_{ik}\phi_k(x), \tag{3.2}$$

where the random coefficients $\xi_{ik}$ are uncorrelated with mean 0 and variances 1, and the functions $\phi_k = \sqrt{\lambda_k}\psi_k$. In what follows, we assume that $\lambda_k = 0$, for $k > \kappa$, where $\kappa$ is a positive integer, thus $G(x, x') = \sum_{k=1}^{\kappa} \phi_k(x)\phi_k\left(x'\right)$. Based on (3.2), the data generating process is now written as

$$Y_{ij} = m\left(X_{ij}\right) + \sum_{k=1}^{\kappa} \xi_{ik}\phi_k\left(X_{ij}\right) + \sigma\left(X_{ij}\right)\varepsilon_{ij}. \tag{3.3}$$

The sequences $\{\lambda_k\}_{k=1}^{\kappa}$, $\{\phi_k(x)\}_{k=1}^{\kappa}$ and the random coefficients $\xi_{ik}$ exist mathematically, but are unknown and unobservable.

Two distinct types of functional data have been studied. Li and Hsing (2007), and Li and Hsing (2010) concern dense functional data, which in the context of model (3.1) means $\min_{1\leq i\leq n} N_i \to \infty$ as $n \to \infty$. On the other hand, Yao, Müller, and Wang (2005a), Yao, Müller, and Wang (2005b), and Yao (2007) studied sparse longitudinal data for which $N_i$'s are i.i.d. copies of an integer-valued positive random variable. Pointwise asymptotic distributions were obtained in Yao (2007) for local polynomial estimators of $m(x)$ based on sparse functional data, but without uniform confidence bands. Nonparametric simultaneous confidence bands are a powerful tool of global inference for functions, see Claeskens and Van Keilegom (2003), Fan and Zhang (2000), Hall and Titterington (1988), Härdle (1989), Härdle and Marron (1991), Huang, Wang, Yang, and Kravchenko (2008), Ma and Yang (2010), Song and Yang (2009), Wang and Yang (2009), Wu and Zhao (2007), Zhao and

Wu (2008), and Zhou, Shen, and Wolfe (1998) for its theory and applications. The fact that a simultaneous confidence band has not been established for functional data analysis is certainly not due to lack of interesting applications, but to the greater technical difficulty in formulating such bands for functional data and establishing their theoretical properties. Specifically, the strong approximation results used to establish the asymptotic confidence level in nearly all published works on confidence bands, commonly known as "Hungarian embedding", are unavailable for sparse functional data.

In this chapter, we present simultaneous confidence bands for $m(x)$ in sparse functional data via a piecewise-constant spline smoothing approach. While there exist a number of smoothing methods for estimating $m(x)$ and $G\left(x, x'\right)$ such as kernels (Yao, Müller and, Wang (2005a); Yao, Müller, and Wang (2005b); Yao (2007)), penalized splines (Cardot, Ferraty, and Sarda (2003); Cardot and Sarda (2005); Yao and Lee (2006)), wavelets Morris and Carroll (2006), and parametric splines James (2002), we choose B splines (Zhou, Huang, and Carroll (2008)) for simple implementation, fast computation and explicit expression, see Huang and Yang (2004), Wang and Yang (2007), and Xue and Yang (2006) for discussion of the relative merits of various smoothing methods.

We organize this chapter as follows. In Section 3.2 we state our main results on confidence bands constructed from piecewise constant splines. In Section 3.3 we provide further insights into the error structure of spline estimators. Section 3.4 describes the actual steps to implement the confidence bands. Section 3.5 reports findings of a simulation study. An empirical example in Section 3.6 illustrates how to use the proposed confidence band for inference. Proofs of technical lemmas are in the Appendix.

## 3.2 Main Results

For convenience, we denote the modulus of continuity of a continuous function $r$ on $[a,b]$ by $\omega(r,\delta) = \max_{x,x' \in [a,b], |x-x'| \leq \delta} \left| r(x) - r\left(x'\right) \right|$. Denote the empirical $L^2$ norm as $\|g\|_{2,N_{\mathrm{T}}}^2 = N_{\mathrm{T}}^{-1} \sum_{i=1}^{n} \sum_{j=1}^{N_i} g^2\left(X_{ij}\right)$, where we denote the total sample size by $N_{\mathrm{T}} = \sum_{i=1}^{n} N_i$. Without loss of generality, we take the range of $X$, $\mathcal{X} = [a,b]$, to be $[0,1]$. For any $\beta \in (0,1]$, we denote the collection of order $\beta$ Hõlder continuous function on $[0,1]$ by

$$C^{0,\beta}[0,1] = \left\{ \phi : \|\phi\|_{0,\beta} = \sup_{x \neq x', x, x' \in [0,1]} \frac{\left| \phi(x) - \phi\left(x'\right) \right|}{\left| x - x' \right|^{\beta}} < +\infty \right\},$$

in which $\|\phi\|_{0,\beta}$ is the $C^{0,\beta}$-seminorm of $\phi$. Let $C[0,1]$ be the collection of continuous function on $[0,1]$. Clearly, $C^{0,\beta}[0,1] \subset C[0,1]$ and, if $\phi \in C^{0,\beta}[0,1]$, then $\omega(\phi,\delta) \leq \|\phi\|_{0,\beta} \delta^{\beta}$.

We use the piecewise constant spline functions introduced in Section 1.3 of Chapter 1 with the number of interior knots denoted as $N_{\mathrm{S}}$. Let $h_{\mathrm{S}} = 1/(N_{\mathrm{S}}+1)$ be the distance between neighboring knots. The mean function $m(x)$ is estimated by

$$\widehat{m}(x) = \operatorname*{argmin}_{g \in G^{(-1)}} \sum_{i=1}^{n} \sum_{j=1}^{N_i} \left\{ Y_{ij} - g\left(X_{ij}\right) \right\}^2. \tag{3.4}$$

The technical assumptions we need are as follows

(A1) *The regression function* $m(x) \in C^{0,1}[0,1]$.

(A2) *The functions* $f(x), \sigma(x)$, *and* $\phi_k(x) \in C^{0,\beta}[0,1]$ *for some* $\beta \in (2/3, 1]$ *with* $f(x) \in \left[c_f, C_f\right], \sigma(x) \in [c_\sigma, C_\sigma], x \in [0,1]$, *for constants* $0 < c_f \leq C_f < \infty, 0 < c_\sigma \leq C_\sigma < \infty$.

(A3) *The set of random variables* $\left(N_i\right)_{i=1}^{n}$ *is a subset of* $\left(N_i\right)_{i=1}^{\infty}$ *consisting of independent variables* $N_i$*, the numbers of observations made for the* $i$*-th subject,* $i = 1, 2, ...,$ *with* $N_i \backsim N$*, where* $N > 0$ *is a positive integer-valued random variable with* $E\{N^{2r}\} \leq r! c_N^r, r = 2, 3, ...$ *for some constant* $c_N > 0$*. The set of random variables* $\left(X_{ij}, Y_{ij}, \varepsilon_{ij}\right)_{i=1,j=1}^{n,N_i}$ *is a subset of* $\left(X_{ij}, Y_{ij}, \varepsilon_{ij}\right)_{i=1,j=1}^{\infty,\infty}$ *in which* $\left(X_{ij}, \varepsilon_{ij}\right)_{i=1,j=1}^{\infty,\infty}$ *are iid. The number* $\kappa$ *of nonzero eigenvalues is finite and the random coefficients* $\xi_{ik}, k = 1, ..., \kappa, i = 1, ..., \infty$ *are iid* $N(0,1)$*. The variables* $\left(N_i\right)_{i=1}^{\infty}$*,* $\left(\xi_{ik}\right)_{i=1,k=1}^{\infty,\kappa}$*,* $\left(X_{ij}\right)_{i=1,j=1}^{\infty,\infty}$*,* $\left(\varepsilon_{ij}\right)_{i=1,j=1}^{\infty,\infty}$ *are independent.*

(A4) *As* $n \to \infty$*, the number of interior knots* $N_{\mathrm{S}} = o\left(n^{\vartheta}\right)$ *for some* $\vartheta \in (1/3, 2\beta - 1)$ *while* $N_{\mathrm{S}}^{-1} = o\left\{n^{-1/3} (\log n)^{-1/3}\right\}$*. The subinterval length* $h_{\mathrm{S}} \sim N_{\mathrm{S}}^{-1}$*.*

(A5) *There exists* $r > 2/\{\beta - (1+\vartheta)/2\}$ *such that* $E|\varepsilon_{11}|^r < \infty$*.*

Assumptions (A1), (A2), (A4) and (A5) are similar to (A1)–(A4) in Wang and Yang (2009), with (A1) weaker than its counterpart. Assumption (A3) is the same as (A1.1), (A1.2), and (A5) in Yao, Müller, and Wang (2005b), without requiring joint normality of the measurement errors $\varepsilon_{ij}$.

We use the B-spline basis of $G^{(-1)}$, the space of piecewise constant splines, denoted as $\{b_J(x)\}_{J=0}^{N_{\mathrm{S}}}$ for theoretical analysis. Define

$$c_{J,n} = \|b_J\|_2^2 = \int_0^1 b_J(x) f(x) dx, J = 0, ..., N_{\mathrm{S}}, \tag{3.5}$$

$$\sigma_Y^2(x) = \mathrm{var}(Y | X = x) = G(x,x) + \sigma^2(x), \forall x \in [0,1],$$

$$\sigma_n^2(x) = c_{J(x),n}^{-2}\{nE(N_1)\}^{-1}\left\{\frac{E\{N_1(N_1-1)\}}{EN_1}\sum_{k=1}^{\kappa}\left(\int_{\chi_{J(x)}}\phi_k(u)f(u)du\right)^2 + \int_{\chi_{J(x)}}\sigma_Y^2(u)f(u)du\right\}. \tag{3.6}$$

$$\text{In addition, define } Q_{N_S+1}(\alpha) = b_{N_S+1} - a_{N_S+1}^{-1}\log\{-(1/2)\log(1-\alpha)\},$$

$$a_{N_S+1} = \{2\log(N_S+1)\}^{1/2}, b_{N_S+1} = a_{N_S+1} - \frac{\log\left(2\pi a_{N_S+1}^2\right)}{2a_{N_S+1}}, \tag{3.7}$$

for any $\alpha \in (0,1)$. We now state our main results.

THEOREM 3.1. *Under Assumptions (A1)-(A5), for any* $\alpha \in (0,1)$,

$$\lim_{n\to\infty} P\left\{\sup_{x\in[0,1]}|\widehat{m}(x) - m(x)|/\sigma_n(x) \le Q_{N_S+1}(\alpha)\right\} = 1-\alpha,$$

$$\lim_{n\to\infty} P\left\{|\widehat{m}(x) - m(x)|/\sigma_n(x) \le Z_{1-\alpha/2}\right\} = 1-\alpha, \forall x \in [0,1],$$

*where* $\sigma_n(x)$ *and* $Q_{N_S+1}(\alpha)$ *are given in (3.6) and (3.7), respectively, while* $Z_{1-\alpha/2}$ *is the* $100(1-\alpha/2)^{th}$ *percentile of the standard normal distribution.*

The definition of $\sigma_n(x)$ in (3.6) does not allow for practical use. The next proposition provides two data-driven alternatives

PROPOSITION 3.1. *Under Assumptions (A2), (A3), and (A5), as* $n \to \infty$,

$$\sup_{x\in[0,1]}\left\{\left|\sigma_n^{-1}(x)\sigma_{n,\text{IID}}(x) - 1\right| + \left|\sigma_n^{-1}(x)\sigma_{n,\text{LONG}}(x) - 1\right|\right\} = O\left(h_S^{\beta}\right),$$

*in which for* $x \in [0,1]$, $\sigma_{n,\mathrm{IID}}(x) \equiv \sigma_Y(x)\left\{f(x)h_{\mathrm{S}}nE(N_1)\right\}^{-1/2}$ *and*

$$\sigma_{n,\mathrm{LONG}}(x) \equiv \sigma_{n,\mathrm{IID}}(x)\left\{1+\frac{E\left\{N_1\left(N_1-1\right)\right\}}{EN_1}h_{\mathrm{S}}\frac{G\left(x,x\right)f(x)}{\sigma_Y^2(x)}\right\}^{1/2}.$$

Using $\sigma_{n,\mathrm{IID}}(x)$ instead of $\sigma_n(x)$ means to treat the $(X_{ij}, Y_{ij})$ as iid data rather than as sparse longitudinal data, while using $\sigma_{n,\mathrm{LONG}}(x)$ means to correctly account for the longitudinal correlation structure. The difference of the two approaches, although asymptotically negligible uniformly for $x \in [0,1]$ according to Proposition 3.1, is significant in finite samples, as shown in the simulation results of Section 3.5. For similar phenomenon with kernel smoothing, see Wang, Carroll, and Lin (2005).

Corollary 3.1. *Under Assumptions (A1)-(A5), for any* $\alpha \in (0,1)$*, as* $n \to \infty$*, an asymptotic* $100\left(1-\alpha\right)\%$ *simultaneous confidence band for* $m(x), x \in [0,1]$ *is*

$$\widehat{m}(x) \pm \sigma_n(x) Q_{N_{\mathrm{S}}+1}\left(\alpha\right),$$

*while an asymptotic* $100\left(1-\alpha\right)\%$ *pointwise confidence interval for* $m(x), x \in [0,1]$*, is* $\widehat{m}(x) \pm \sigma_n(x) Z_{1-\alpha/2}$.

## 3.3 Decomposition

In this section, we decompose the estimation error $\widehat{m}(x) - m(x)$ by the representation of $Y_{ij}$ as the sum of $m\left(X_{ij}\right)$, $\sum_{k=1}^{\kappa}\xi_{ik}\phi_k\left(X_{ij}\right)$, and $\sigma\left(X_{ij}\right)\varepsilon_{ij}$.

We introduce the rescaled B-spline basis $\left\{B_J(x)\right\}_{J=0}^{N_{\mathrm{S}}}$ for $G^{(-1)}$, which is $B_J(x) \equiv$

$b_J(x)\left\|b_J\right\|_2^{-1}, J=0,...,N_{\mathrm{S}}$. Therefore,

$$B_J(x) \equiv b_J(x)\left\{c_{J,n}\right\}^{-1/2}, J=0,...,N_{\mathrm{S}}. \tag{3.8}$$

It is easily verified that $\left\|B_J\right\|_2^2 = 1, J=0,1,\ldots,N_{\mathrm{S}}, \left\langle B_J, B_{J'}\right\rangle \equiv 0, J \neq J'$.

The definition of $\widehat{m}(x)$ in (3.4) means that

$$\widehat{m}(x) \equiv \sum\nolimits_{J=0}^{N_{\mathrm{S}}} \widehat{\lambda}'_J b_J(x), \tag{3.9}$$

with coefficients $\left\{\widehat{\lambda}'_0,...,\widehat{\lambda}'_{N_{\mathrm{S}}}\right\}^{\mathrm{T}}$ as solutions of the least squares problem

$$\left\{\widehat{\lambda}'_0,...,\widehat{\lambda}'_{N_{\mathrm{S}}}\right\}^{\mathrm{T}} = \underset{\left\{\lambda_0,...,\lambda_{N_{\mathrm{S}}}\right\}\in R^{N_{\mathrm{S}}+1}}{\operatorname{argmin}} \sum\nolimits_{i=1}^{n}\sum\nolimits_{j=1}^{N_i}\left\{Y_{ij} - \sum\nolimits_{J=0}^{N_{\mathrm{S}}}\lambda_J b_J\left(X_{ij}\right)\right\}^2.$$

Simple linear algebra shows $\widehat{m}(x) \equiv \sum_{J=0}^{N_{\mathrm{S}}}\widehat{\lambda}_J B_J(x)$, where the coefficients $\left\{\widehat{\lambda}_0,...,\widehat{\lambda}_{N_{\mathrm{S}}}\right\}^{\mathrm{T}}$ are solutions of the least squares problem

$$\left\{\widehat{\lambda}_0,...,\widehat{\lambda}_{N_{\mathrm{S}}}\right\}^{\mathrm{T}} = \underset{\left\{\lambda_0,...,\lambda_{N_{\mathrm{S}}}\right\}\in R^{N_{\mathrm{S}}+1}}{\operatorname{argmin}} \sum\nolimits_{i=1}^{n}\sum\nolimits_{j=1}^{N_i}\left\{Y_{ij} - \sum\nolimits_{J=0}^{N_{\mathrm{S}}}\lambda_J B_J\left(X_{ij}\right)\right\}^2. \tag{3.10}$$

Projecting the relationship in model (3.3) onto the linear subspace of $R^{N_{\mathrm{T}}}$ spanned by $\left\{B_J\left(X_{ij}\right)\right\}_{1\leq j\leq N_i, 1\leq i\leq n, 0\leq J\leq N_{\mathrm{S}}}$, we obtain the following crucial decomposition in the space $G^{(-1)}$ of spline functions:

$$\widehat{m}(x) = \widetilde{m}(x) + \widetilde{e}(x) = \widetilde{m}(x) + \widetilde{\varepsilon}(x) + \sum\nolimits_{k=1}^{\kappa}\widetilde{\xi}_k(x), \tag{3.11}$$

$$\widetilde{m}(x) = \sum_{J=0}^{N_{\mathrm{S}}} \widetilde{\lambda}_J B_J(x),\ \widetilde{\varepsilon}(x) = \sum_{J=0}^{N_{\mathrm{S}}} \widetilde{a}_J B_J(x),$$
$$\widetilde{\xi}_k(x) = \sum_{J=0}^{N_{\mathrm{S}}} \widetilde{\tau}_{k,J} B_J(x). \tag{3.12}$$

The vectors $\{\widetilde{\lambda}_0, ..., \widetilde{\lambda}_{N_{\mathrm{S}}}\}^{\mathrm{T}}$, $\left\{\widetilde{a}_0, ..., \widetilde{a}_{N_{\mathrm{S}}}\right\}^{\mathrm{T}}$, and $\left\{\widetilde{\tau}_{k,0}, ..., \widetilde{\tau}_{k,N_{\mathrm{S}}}\right\}^{\mathrm{T}}$ are solutions to (3.10) with $Y_{ij}$ replaced by $m\left(X_{ij}\right)$, $\sigma\left(X_{ij}\right)\varepsilon_{ij}$, and $\xi_{ik}\phi_k\left(X_{ij}\right)$, respectively. We cite next an important result concerning the function $\widetilde{m}(x)$. The first part is from de Boor (2001), p. 149, and the second is from Theorem 5.1 of Huang (2003).

THEOREM 3.2. *There is an absolute constant $C_g > 0$ such that for every $\phi \in C[0,1]$, there exists a function $g \in G^{(-1)}[0,1]$ that satisfies $\|g - \phi\|_\infty \leq C_g \omega(\phi, h_{\mathrm{S}})$. In particular, if $\phi \in C^{0,\beta}[0,1]$ for some $\beta \in (0,1]$, then $\|g - \phi\|_\infty \leq C_g \|\phi\|_{0,\beta} h_{\mathrm{S}}^{\beta}$. Under Assumptions (A1) and (A4), with probability approaching* 1, *the function $\widetilde{m}(x)$ defined in (3.12) satisfies $\|\widetilde{m}(x) - m(x)\|_\infty = O(h_{\mathrm{S}})$.*

The next proposition concerns the function $\widetilde{e}(x)$ given in (3.11).

PROPOSITION 3.2. *Under Assumptions (A2)-(A5), for any $\tau \in R$, and $\sigma_n(x)$, $a_{N_{\mathrm{S}}+1}$, and $b_{N_{\mathrm{S}}+1}$ as given in (3.6) and (3.7),*

$$\lim_{n\to\infty} P\left\{\sup_{x\in[0,1]} \left|\sigma_n(x)^{-1}\widetilde{e}(x)\right| \leq \tau/a_{N_{\mathrm{S}}+1} + b_{N_{\mathrm{S}}+1}\right\} = \exp\left(-2e^{-\tau}\right).$$

## 3.4 Implementation

In this section, we describe procedures to implement the confidence bands and intervals given in Corollary 3.1. Given any data set $\left(X_{ij}, Y_{ij}\right)_{j=1,i=1}^{N_i,n}$ from model (3.3), the spline estimator $\widehat{m}(x)$ is obtained by (3.9), and the number of interior knots in (3.9) is taken

to be $N_{\mathrm{S}} = [cN_{\mathrm{T}}^{1/3}(\log n)]$, in which $[a]$ denotes the integer part of $a$ and $c$ is a positive constant. When constructing the confidence bands, one needs to evaluate the function $\sigma_n^2(x)$ by estimating the unknown functions $f(x)$, $\sigma_Y^2(x)$, and $G(x,x)$, and then plugging in these estimators: the same approach is taken in Wang and Yang (2009).

The number of interior knots for pilot estimation of $f(x)$, $\sigma_Y^2(x)$, and $G(x,x)$ is taken to be $N_{\mathrm{S}}^* = \left[n^{1/3}\right]$, and $h_{\mathrm{S}}^* = 1/\left(1+N_{\mathrm{S}}^*\right)$. The histogram pilot estimator of the density function $f(x)$ is

$$\widehat{f}(x) = \{\sum_{i=1}^{n}\sum_{j=1}^{N_i} b_{J(x)}(X_{ij})\}/\{(\sum_{i=1}^{n} N_i)h_{\mathrm{S}}^*\}.$$

Defining the vector $\mathbf{R} = \{R_{ij}\}^{\mathrm{T}}_{1\le j\le N_i, 1\le i\le n} = \left\{\left(Y_{ij} - \widehat{m}(X_{ij})\right)^2\right\}^{\mathrm{T}}_{1\le j\le N_i, 1\le i\le n}$, the estimator of $\sigma_Y^2(x)$ is $\widehat{\sigma}_Y^2(x) = \sum_{J=0}^{N_{\mathrm{S}}^*}\widehat{\rho}_J b_J(x)$, where the coefficients $\{\widehat{\rho}_0,...,\widehat{\rho}_{N_{\mathrm{S}}^*}\}^{\mathrm{T}}$ are solutions of the least squares problem:

$$\left\{\widehat{\rho}_0,...,\widehat{\rho}_{N_{\mathrm{S}}^*}\right\}^{\mathrm{T}} = \underset{\left\{\widehat{\rho}_0,...,\widehat{\rho}_{N_{\mathrm{S}}^*}\right\}\in R^{N_{\mathrm{S}}+1}}{\operatorname{argmin}} \sum_{i=1}^{n}\sum_{j=1}^{N_i}\left\{R_{ij} - \sum_{J=0}^{N_{\mathrm{S}}^*}\rho_J b_J\left(X_{ij}\right)\right\}^2.$$

The pilot estimator of covariance function $G\left(x,x'\right)$ is

$$\widehat{G}\left(x,x'\right) = \arg\min_{g\in G^{(-1)}\otimes G^{(-1)}} \sum_{i=1}^{n}\sum_{j,j'=1,j\neq j'}^{N_i}\left\{C_{ijj'} - g\left(X_{ij},X_{ij'}\right)\right\}^2,$$

where $C_{ijj'} = \left\{Y_{ij} - \widehat{m}\left(X_{ij}\right)\right\}\left\{Y_{ij'} - \widehat{m}\left(X_{ij'}\right)\right\}, 1\le j,j'\le N_i, 1\le i\le n$. The

function $\sigma_n(x)$ is estimated by either $\widehat{\sigma}_{n,\mathrm{IID}}(x) \equiv \widehat{\sigma}_Y(x)\left\{\widehat{f}(x)h_{\mathrm{S}}N_{\mathrm{T}}\right\}^{-1/2}$ or

$$\widehat{\sigma}_{n,\mathrm{LONG}}(x) \equiv \widehat{\sigma}_{n,\mathrm{IID}}(x)\left\{1+\left(\sum_{i=1}^{n} N_i^2/N_{\mathrm{T}}-1\right)\frac{\widehat{G}(x,x)}{\widehat{\sigma}_Y^2(x)}\widehat{f}(x)h_{\mathrm{S}}\right\}^{1/2}.$$

We now state a result. That is easily proved by standard theory of kernel and spline smoothing, as in Wang and Yang (2009).

PROPOSITION 3.3. *Under Assumptions (A1)-(A5), as* $n \to \infty$

$$\begin{aligned}&\sup_{x\in[0,1]}\left\{\left|\widehat{\sigma}_{n,\mathrm{IID}}(x)\sigma_{n,\mathrm{IID}}^{-1}(x)-1\right|+\left|\widehat{\sigma}_{n,\mathrm{LONG}}(x)\sigma_{n,\mathrm{LONG}}^{-1}(x)-1\right|\right\}\\ &\qquad = O_{a.s.}\left(h_{\mathrm{S}}^{\beta}+n^{-1/2}N_{\mathrm{S}}^{-1}\left(logn\right)^{1/2}\right).\end{aligned}$$

Proposition 3.1, about how $\sigma_{n,\mathrm{IID}}(x)$ and $\sigma_{n,\mathrm{LONG}}(x)$ uniformly approximate $\sigma_n(x)$, and Proposition 3.3 together imply that both $\widehat{\sigma}_{n,\mathrm{IID}}(x)$ and $\widehat{\sigma}_{n,\mathrm{LONG}}(x)$ approximate $\sigma_n(x)$ uniformly at a rate faster than $\left(n^{-1/2+1/3}\left(\log n\right)^{1/2-1/3}\right)$, according to Assumption (A5). Therefore as $n \to \infty$, the confidence bands

$$\widehat{m}(x) \pm \widehat{\sigma}_{n,\mathrm{IID}}(x)Q_{N_{\mathrm{S}}+1}(\alpha), \tag{3.13}$$

$$\widehat{m}(x) \pm \widehat{\sigma}_{n,\mathrm{LONG}}(x)Q_{N_{\mathrm{S}}+1}(\alpha), \tag{3.14}$$

with $Q_{N_{\mathrm{S}}+1}(\alpha)$ given in (3.7), and the pointwise intervals $\widehat{m}(x)\pm\widehat{\sigma}_{n,\mathrm{IID}}(x)Z_{1-\alpha/2}$, $\widehat{m}(x)\pm\widehat{\sigma}_{n,\mathrm{LONG}}(x)Z_{1-\alpha/2}$ have asymptotic confidence level $1-\alpha$.

## 3.5 Simulation

To illustrate the finite-sample performance of the spline approach, we generated data from the model

$$Y_{ij} = m\left(X_{ij}\right) + \sum_{k=1}^{2} \xi_{ik}\phi_k\left(X_{ij}\right) + \sigma\varepsilon_{ij}, 1 \le j \le N_i, 1 \le i \le n,$$

with $X \sim \text{Uniform}[0,1]$, $\xi_k \sim \text{Normal}(0,1), k = 1,2$, $\varepsilon \sim \text{Normal}(0,1)$, $N_i$ having a discrete uniform distribution from $25, \cdots, 35$, for $1 \le i \le n$, and $m(x) = \sin\{2\pi(x-1/2)\}, \phi_1(x) = -2\cos\{\pi(x-1/2)\}/\sqrt{5}, \phi_2(x) = \sin\{\pi(x-1/2)\}/\sqrt{5}$, thus $\lambda_1 = 2/5, \lambda_2 = 1/10$. The noise levels were $\sigma = 0.5, 1.0$, the number of subjects $n$ was taken to be $20, 50, 100, 200$, the confidence levels were $1-\alpha = 0.95, 0.99$, and the constant $c$ in the definition of $N_S$ in Section 3.4 was taken to be $1, 2, 3$. We found that the confidence band (3.13) did not have good coverage rates for moderate sample sizes, and hence in Table 3.1 we report the coverage as the percentage out of the total 200 replications for which the true curve was covered by (3.14) at the 101 points $\{k/100, k = 0, \ldots, 100\}$.

At all noise levels, the coverage percentages for the confidence band (3.14) are very close to the nominal confidence levels 0.95 and 0.99 for $c = 1, 2$, but decline for $c = 3$ when $n = 20, 50$. The coverage percentages thus depend on the choice of $N_S$, and the dependency becomes stronger when sample sizes decrease. For large sample sizes $n = 100, 200$, the effect of the choice of $N_S$ on the coverage percentages is insignificant. Because $N_S$ varies with $N_i$, for $1 \le i \le n$, the data-driven selection of some "optimal" $N_S$ remains an open problem.

We next examine two alternative methods to compute the confidence band, based on the observation that the estimated mean function $\widehat{m}(x)$ and the confidence intervals are

Table 3.1: Uniform coverage rates in Chapter 3

| $\sigma$ | $n$ | $1-\alpha$ | $c=1$ | $c=2$ | $c=3$ |
|---|---|---|---|---|---|
| 0.5 | 20 | 0.950 | 0.920 | 0.930 | 0.800 |
| | | 0.990 | 0.990 | 0.990 | 0.900 |
| | 50 | 0.950 | 0.960 | 0.965 | 0.910 |
| | | 0.990 | 0.995 | 0.995 | 0.965 |
| | 100 | 0.950 | 0.955 | 0.955 | 0.955 |
| | | 0.990 | 1.000 | 1.000 | 0.985 |
| | 200 | 0.950 | 0.950 | 0.965 | 0.975 |
| | | 0.990 | 0.985 | 0.985 | 0.990 |
| 1.0 | 20 | 0.950 | 0.935 | 0.930 | 0.735 |
| | | 0.990 | 0.990 | 0.990 | 0.870 |
| | 50 | 0.950 | 0.975 | 0.960 | 0.895 |
| | | 0.990 | 0.995 | 0.995 | 0.980 |
| | 100 | 0.950 | 0.950 | 0.940 | 0.935 |
| | | 0.990 | 0.995 | 0.990 | 0.990 |
| | 200 | 0.950 | 0.940 | 0.965 | 0.960 |
| | | 0.990 | 0.985 | 0.995 | 0.995 |

Note: uniform coverage rates from 200 replications using the confidence band (3.14). For each sample size $n$, the first row is the coverage of a nominal 95% confidence band, while the second row is for a 99% confidence band.

step functions that remain the same on each subinterval $\chi_J$, $0 \leq J \leq N_{\mathrm{S}}$. Follwing an associate editor's suggestion, locally weighted smoothing was applied to the upper and lower confidence limits to generate a smoothed confidence band. Following a referee's suggestion to treat the number $(N_{\mathrm{S}}+1)$ of subintervals as fixed instead of growing to infinity, a naive parametric confidence band was computed as

$$\widehat{m}(x) \pm \widehat{\sigma}_{n,\mathrm{LONG}}(x) Q_{1-\alpha.N_{\mathrm{S}}+1} \tag{3.15}$$

in which $Q_{1-\alpha.N_{\mathrm{S}}+1} = Z_{\left\{1+(1-\alpha)^{1/(N_{\mathrm{S}}+1)}\right\}/2}$ is the $(1-\alpha)$ quantile of the maximal absolute values of $(N_{\mathrm{S}}+1)$ iid $N(0,1)$ random variables. We compare the performance of

the confidence band in (3.14), the smoothed band and naive parametric band in (3.15). Given $n = 20$ with $N_S = 8, 12$, and $n = 50$ $N_S = 44$ (by taking $c = 1$ in the definition of $N_S$ in Section 3.4), $\sigma = 0.5, 1.0$, and $1 - \alpha = 0.99$, Table 3.2 reports the coverage percentages $\widehat{P}$, $\widehat{P}_{\text{naive}}, \widehat{P}_{\text{smooth}}$ and the average maximal widths $W, W_{\text{naive}}, W_{\text{smooth}}$ of $N_S + 1$ intervals out of 200 replications calculated from confidence bands (3.14), (3.15), and the smoothed confidence bands, respectively.

Table 3.2: Uniform coverage rates and average maximal widths in Chapter 3

| $n$ | $\sigma$ | $N_S$ | $\widehat{P}$ | $\widehat{P}_{\text{naive}}$ | $\widehat{P}_{\text{smooth}}$ | $W$ | $W_{\text{naive}}$ | $W_{\text{smooth}}$ |
|---|---|---|---|---|---|---|---|---|
| 20 | 0.5 | 8 | 0.820 | 0.505 | 0.910 | 1.490 | 1.210 | 1.480 |
| | | 12 | 0.930 | 0.765 | 0.955 | 1.644 | 1.363 | 1.628 |
| | 1.0 | 8 | 0.910 | 0.655 | 0.970 | 1.725 | 1.401 | 1.721 |
| | | 12 | 0.960 | 0.820 | 0.985 | 1.937 | 1.606 | 1.928 |
| 50 | 0.5 | 44 | 0.990 | 0.960 | 0.990 | 1.651 | 1.522 | 1.609 |
| | 1.0 | 44 | 0.990 | 0.975 | 1.000 | 2.054 | 1.893 | 2.016 |

Note: uniform coverage rates and average maximal widths of confidence intervals from 200 replications using the confidence bands (3.14), (3.15), and the smoothed bands respectively, for $1 - \alpha = 0.99$.

In all experiments, one has $\widehat{P}_{\text{smooth}} > \widehat{P} > \widehat{P}_{\text{naive}}$ and $W > W_{\text{smooth}} > W_{\text{naive}}$. The coverage percentages for both the confidence bands in (3.14) and the smoothed bands are much closer to the nominal level than those of the naive bands in (3.15), while the smoothed bands perform slightly better than the constant spline bands in (3.14), with coverage percentages closer to the nominal and smaller widths. Based on these observations, the naive band is not recommended due to poor coverage. As for the smoothed band, although it has slightly better coverage than the constant spline band, its asymptotic property has yet to be established, and the second step smoothing adds to its conceptual complexity and computational burden. Therefore with everything considered, the constant spline band is recommended for

its satisfactory theoretical property, fast computing, and conceptual simplicity.

For visualization of the actual function estimates, at $\sigma = 0.5$ with $n = 20, 50$, Figures 3.1 and 3.2 depict the simulated data points and the true curve, and Figures 3.3, 3.4, 3.5 and 3.6 show the true curve, the estimated curve, the uniform confidence band, and the pointwise confidence intervals.

## 3.6 Empirical Example

In this section, we apply the confidence band procedure of Section 3.4 to the data collected from a study by the AIDS Clinical Trials Group, ACTG 315 (Zhou, Huang, and Carroll (2008)). In this study, 46 HIV 1 infected patients were treated with potent antiviral therapy consisting of ritonavir, 3TC and AZT. After initiation of the treatment on day 0, patients were followed for up to 10 visits. Scheduled visit times common for all patients were 7, 14, 21, 28, 35, 42, 56, 70, 84, and 168 days. Since the patients did not follow exactly the scheduled times and/or missed some visits, the actual visit times $T_{ij}$ were irregularly spaced and varied from day 0 to day 196. The CD4+ cell counts during HIV/AIDS treatments are taken as the response variable $Y$ from day 0 to day 196. Figure 3.7 shows that the data points (dots) are extremely sparse between day 100 and 150, thus we first transform the data by $X_{ij} = T_{ij}^{1/3}$. A histogram (not shown) indicates that the $X_{ij}$-values are distributed fairly uniformly. The number of interior knots in (3.9) is taken to be $N_{\mathrm{S}} = 6$, so that the range for visit time $T$, which is $[0, 196]$, is divided into seven unequal subintervals, and in each subinterval, the mean CD4+ cell counts and the confidence bands remain the same. Table 3.3 gives the mean CD4+ cell counts and the confidence limits on each subinterval at simultaneous confidence level 0.95. For instance, from day 4 to 14, the mean CD4+ cell

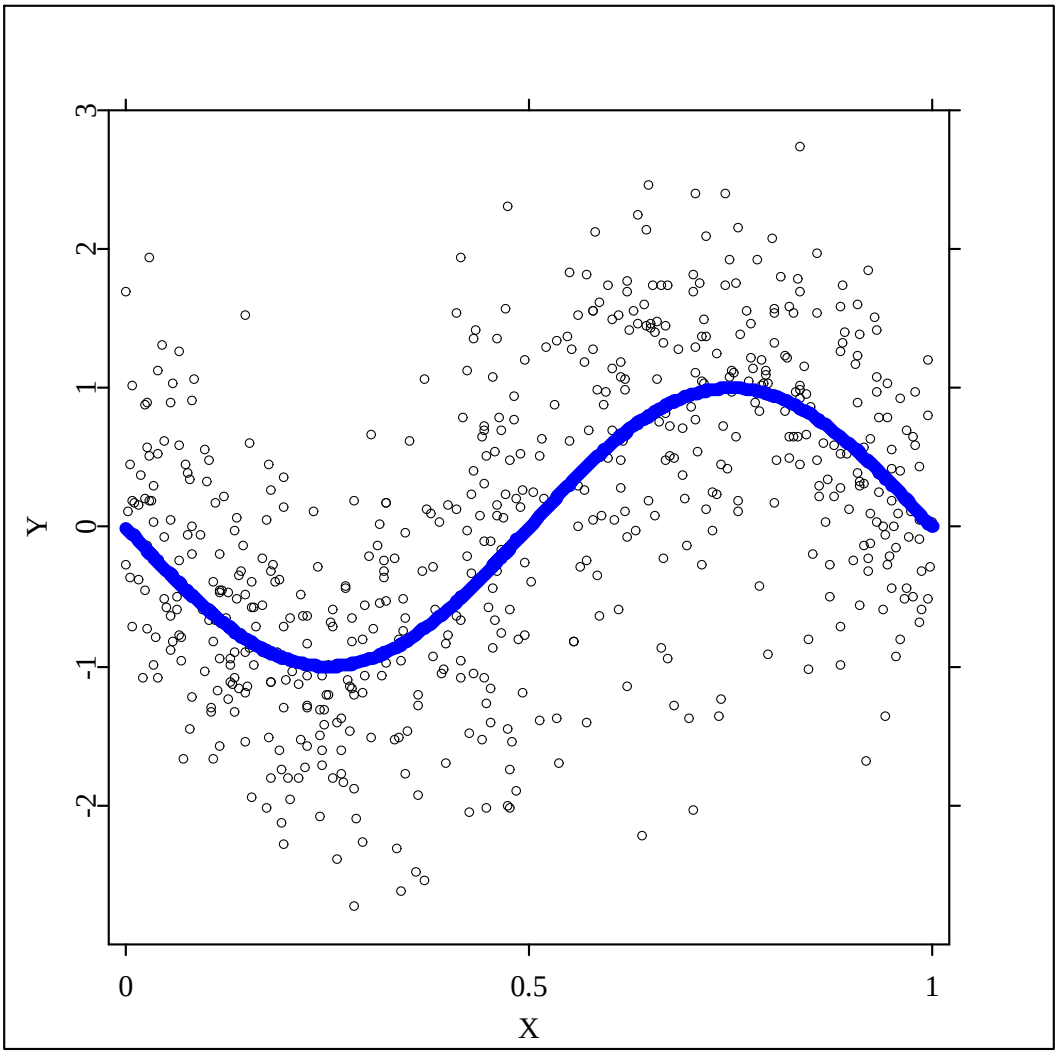


Figure 3.1: Plots of simulated data for $n = 20$ in Chapter 3

Note: plots of simulated data scatter points at $\sigma = 0.5$, $n = 20$, and the true curve.

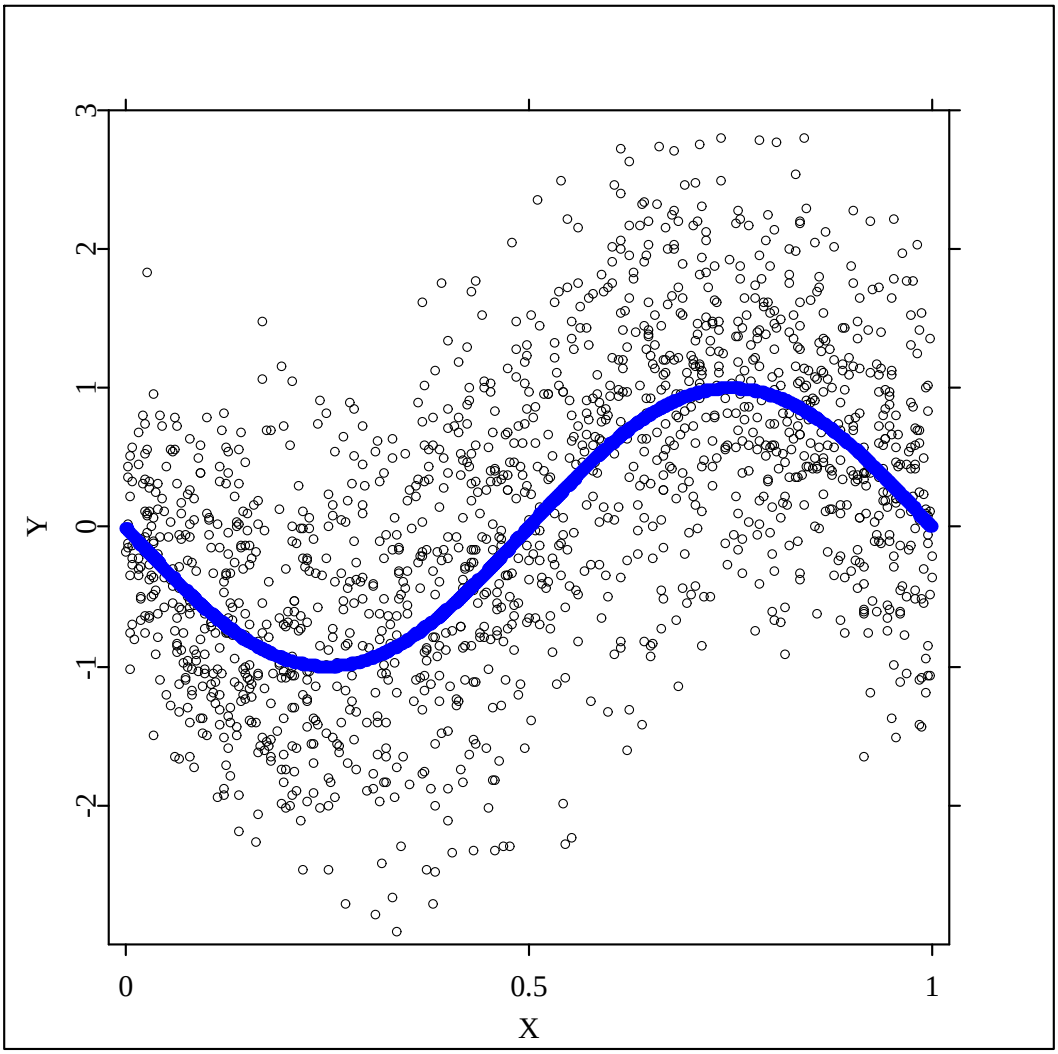


Figure 3.2: Plots of simulated data for $n = 50$ in Chapter 3

Note: plots of simulated data scatter points at $\sigma = 0.5$, $n = 50$, and the true curve.

counts is 241.62 with lower and upper limits 171.81 and 311.43 respectively.

Table 3.3: Confidence limits for CD4 data set

| Days | Mean CD4+ cell counts | Confidence limits |
|---|---|---|
| $[0, 1)$ | 178.23 | $[106.73, 249.72]$ |
| $[1, 4)$ | 200.32 | $[130.51, 270.13]$ |
| $[4, 15)$ | 241.62 | $[171.81, 311.43]$ |
| $[15, 36)$ | 271.87 | $[194.70, 349.04]$ |
| $[36, 71)$ | 299.51 | $[222.34, 376.68]$ |
| $[71, 123)$ | 280.78 | $[203.50, 358.06]$ |
| $[123, 196]$ | 299.27 | $[221.99, 376.55]$ |

Note: the mean CD4+ cell counts and the confidence limits on each subinterval at simultaneous confidence level 0.95.

Figure 3.7 depicts the 95% simultaneous (smoothed) confidence band according to (3.14) in (median) thin lines, and Figure 3.8 depicts the pointwise 95% confidence intervals in thin lines. The center thick line is the piecewise-constant spline fit $\widehat{m}(x)$. It can be seen that the pointwise confidence intervals are of course narrower than the uniform confidence band by the same ratio. Figure 3.7 is essentially a graphical representation of Table 3.3; both confirm that the mean CD4+ cell counts generally increases over time as Zhou, Huang, and Carroll (2008) pointed out. The advantage of the current method is that such inference on the overall trend is made with predetermined type I error probability, in this case 0.05.

## 3.7 Discussion

In this chapter, we have constructed a simultaneous confidence band for the mean function $m(x)$ for sparse longitudinal data via piecewise-constant spline fitting. Our approach extends the asymptotic results in Wang and Yang (2009) for i.i.d. random designs to a much more complicated data structure by allowing dependence of measurements within each subject.

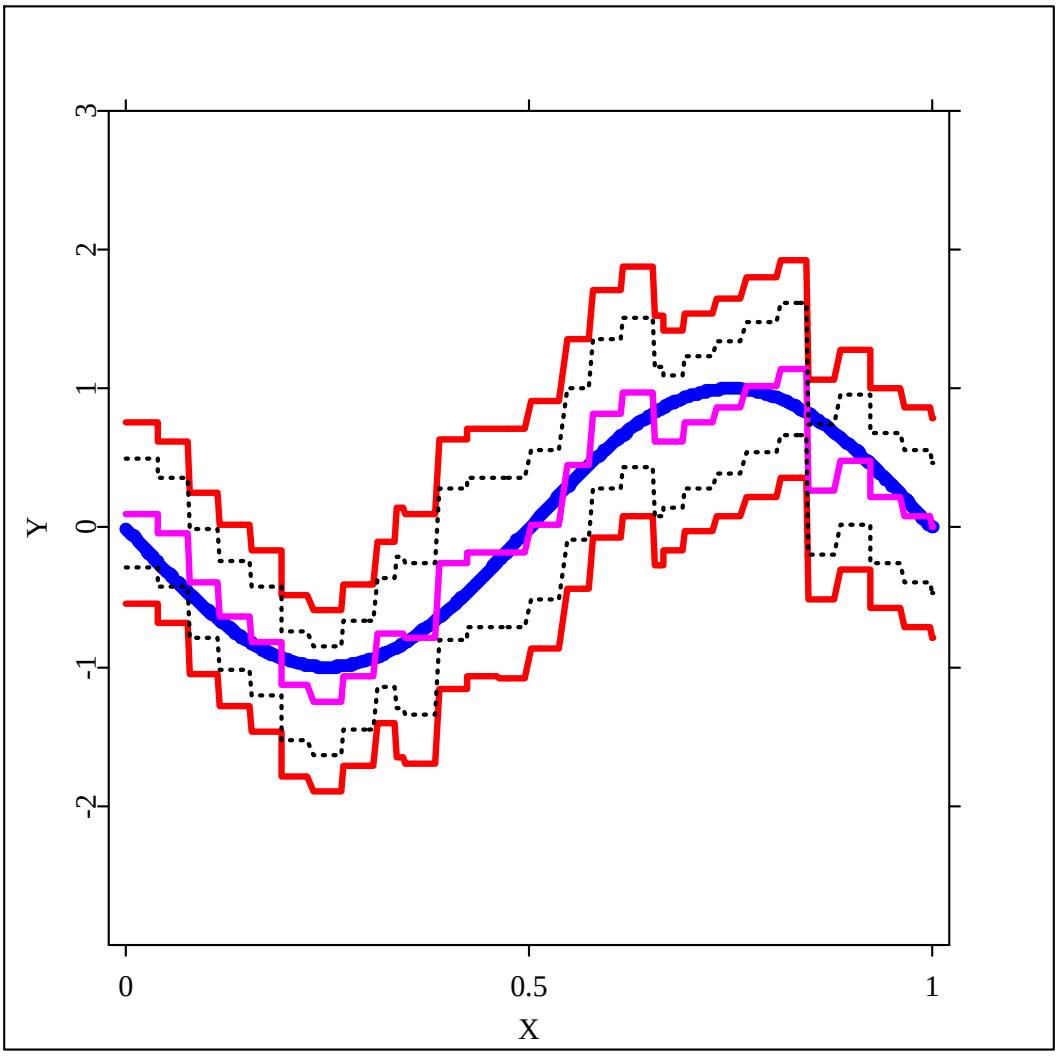


Figure 3.3: Plots of confidence bands at $1-\alpha = 0.95, n = 20$ in Chapter 3

Note: plots of confidence bands (3.14) (upper and lower solid lines), pointwise confidence intervals (upper and lower dashed lines), the spline estimator (middle thin line), and the true function (middle thick line) at $1-\alpha = 0.95, n = 20$.

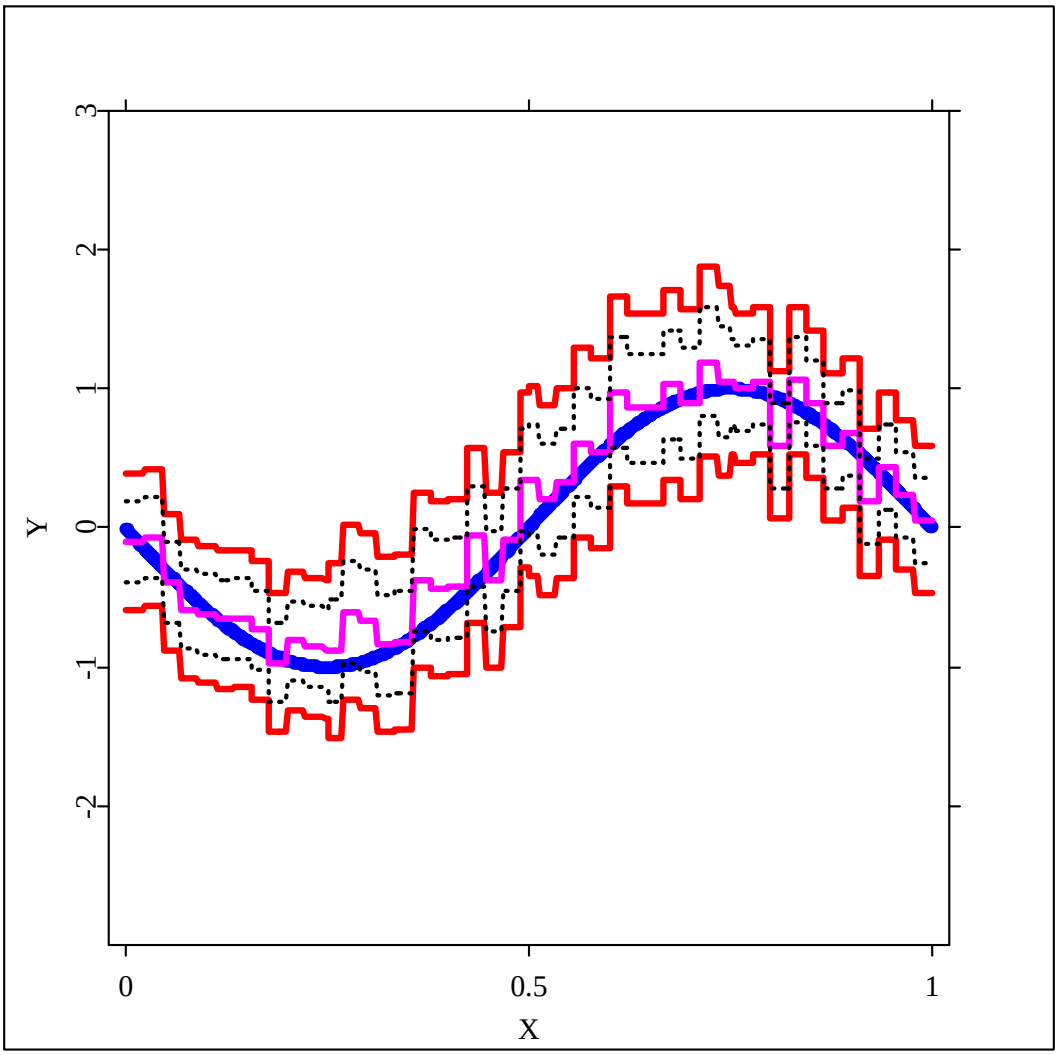


Figure 3.4: Plots of confidence bands at $1-\alpha = 0.95, n = 50$ in Chapter 3

Note: plots of confidence bands (3.14) (upper and lower solid lines), pointwise confidence intervals (upper and lower dashed lines), the spline estimator (middle thin line), and the true function (middle thick line) at $1-\alpha = 0.95, n = 50$.

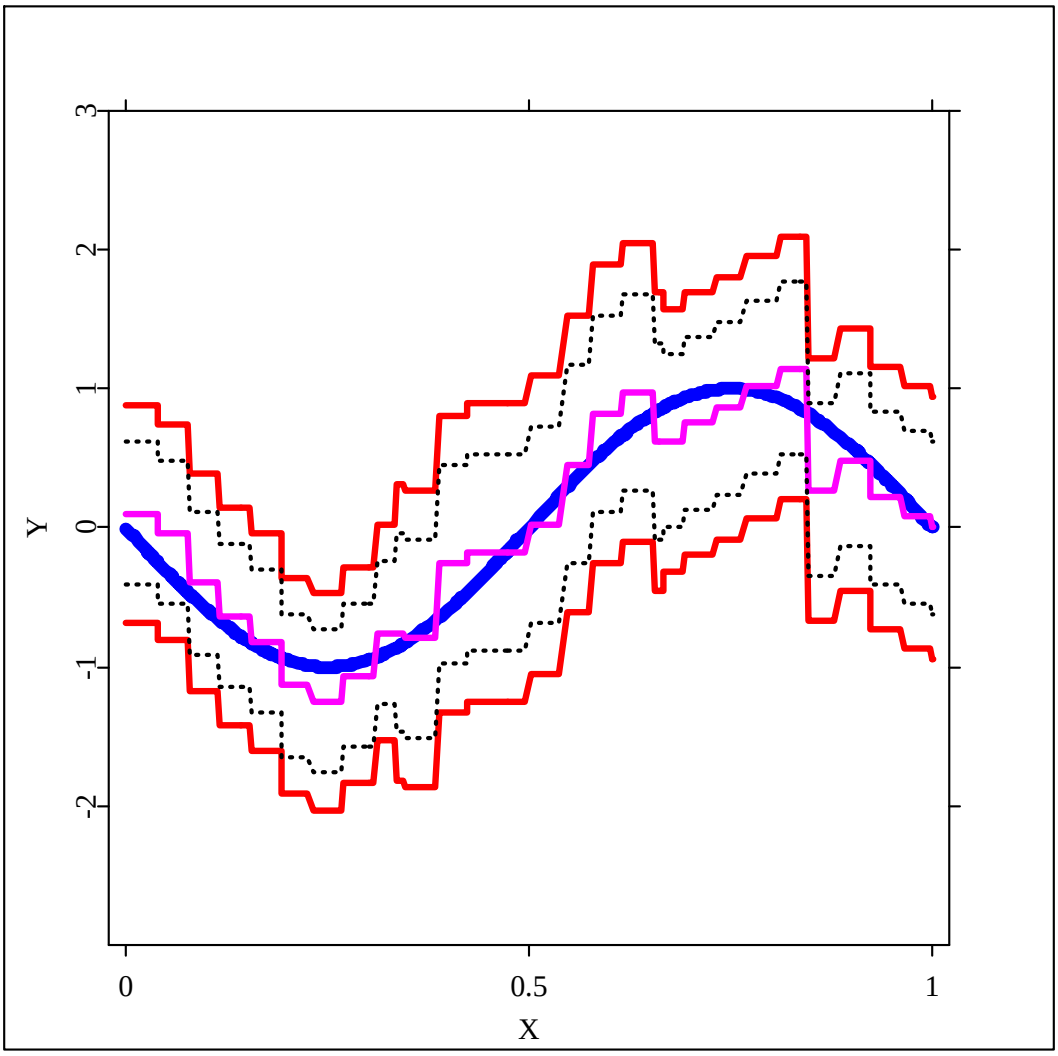


Figure 3.5: Plots of confidence bands at $1-\alpha = 0.99, n = 20$ in Chapter 3

Note: plots of confidence bands (3.14) (upper and lower solid lines), pointwise confidence intervals (upper and lower dashed lines), the spline estimator (middle thin line), and the true function (middle thick line) at $1-\alpha = 0.99, n = 20$.

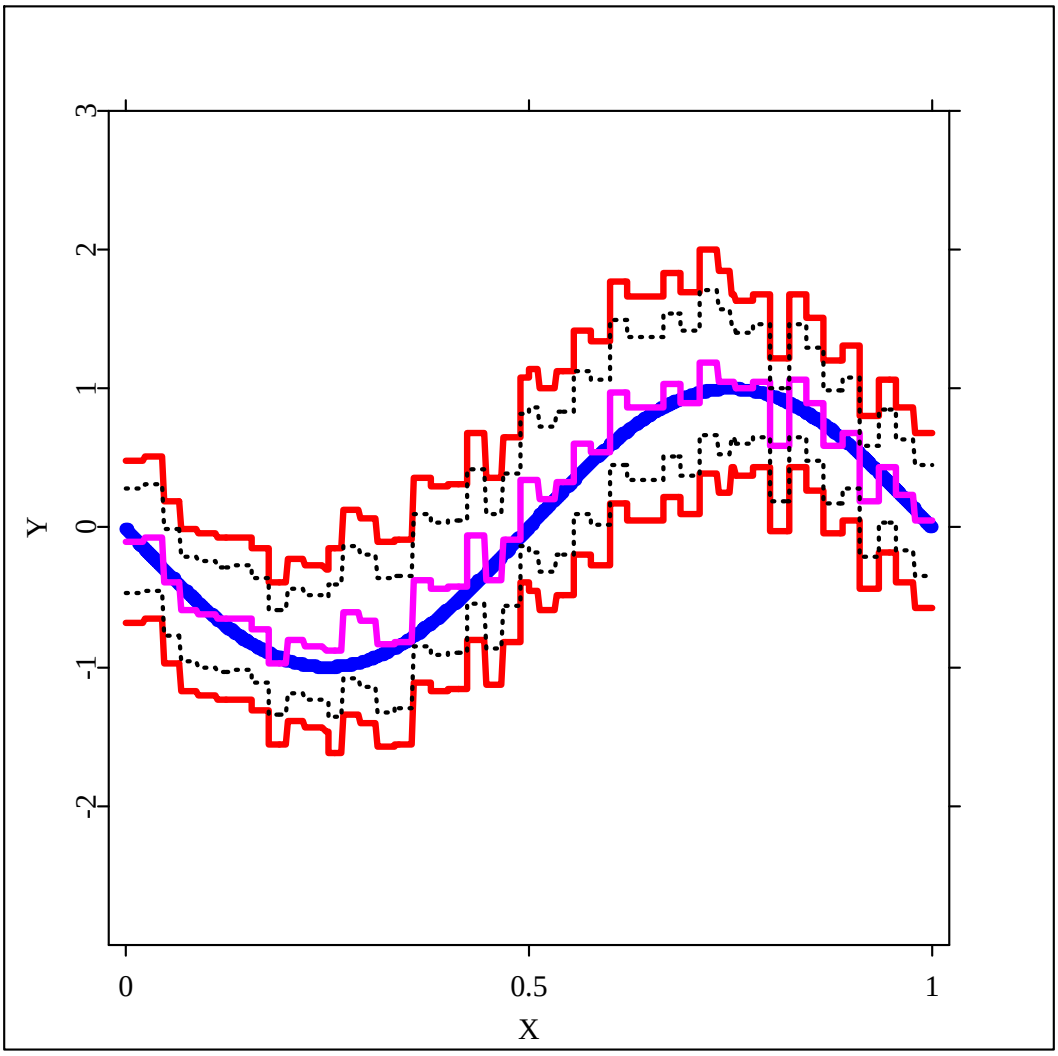


Figure 3.6: Plots of confidence bands at $1-\alpha = 0.99, n = 50$ in Chapter 3

Note: plots of confidence bands (3.14) (upper and lower solid lines), pointwise confidence intervals (upper and lower dashed lines), the spline estimator (middle thin line), and the true function (middle thick line) at $1-\alpha = 0.99, n = 50$.

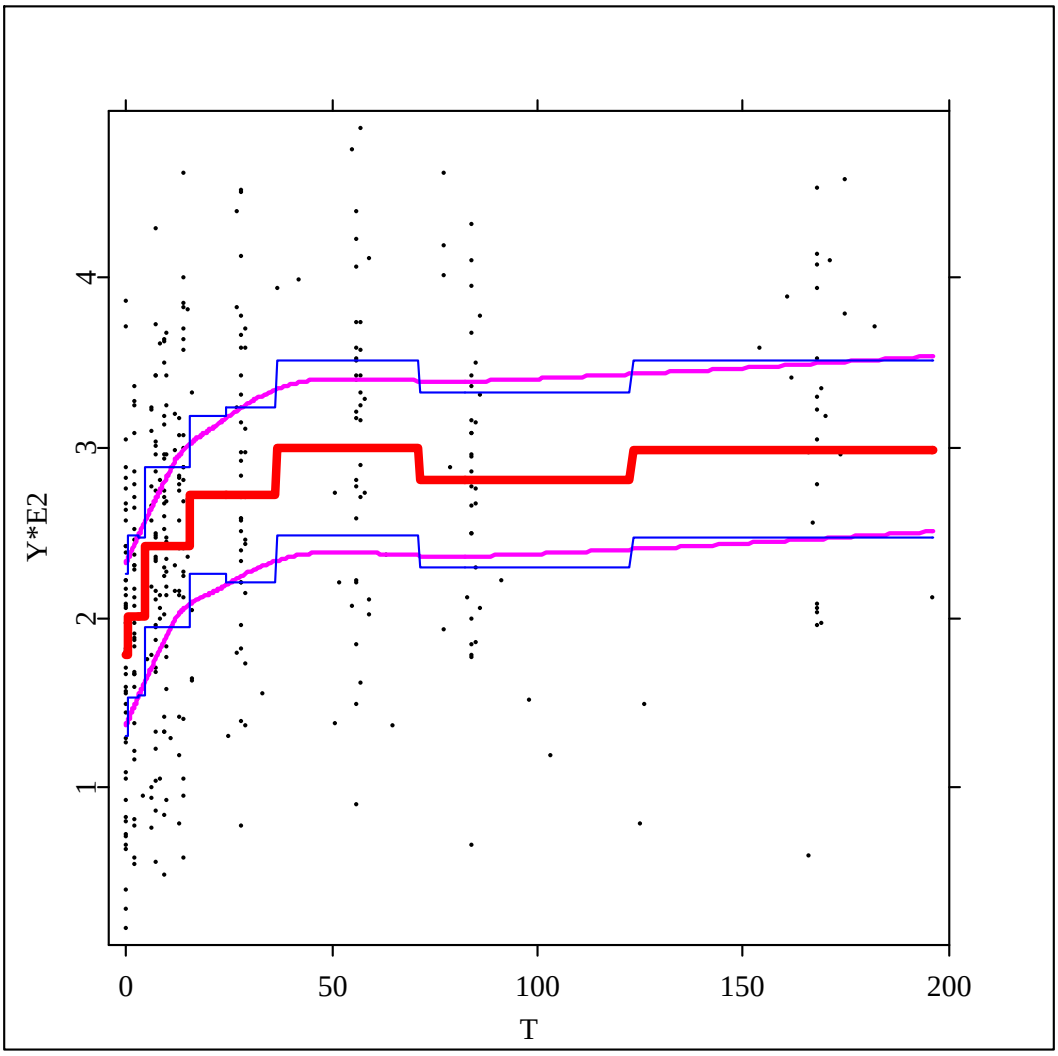


Figure 3.7: Plots of confidence bands for CD4 data

Note: plots of the piecewise-constant spline estimator (thick line), the data (dots), confidence band (3.14) (upper and lower solid lines), and the smoothed band (upper and lower thin lines) at confidence level 0.95.

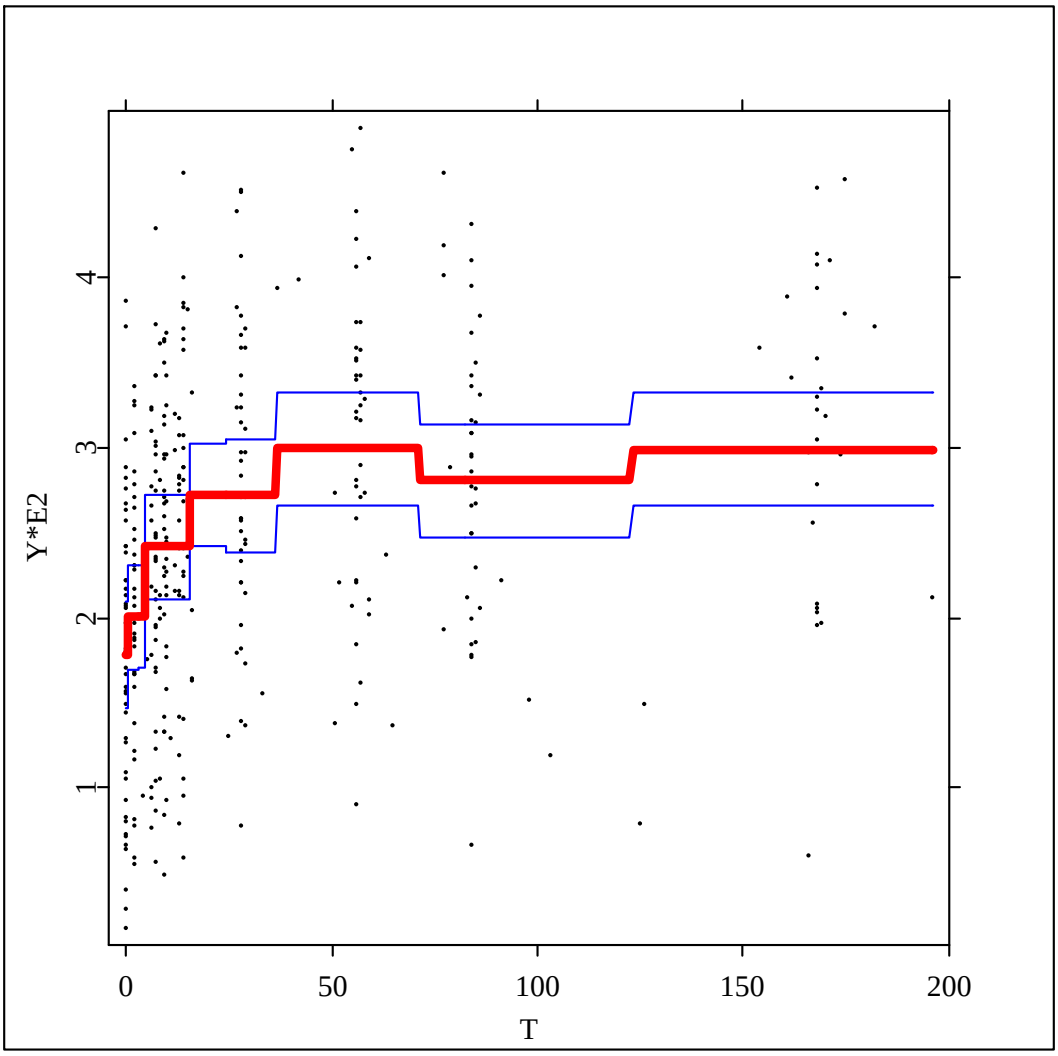


Figure 3.8: Plots of confidence intervals for CD4 data

Note: plots of the piecewise-constant spline estimator (thick line), the data (dots), pointwise confidence intervals (upper and lower thin lines), and the smoothed band (upper and lower thin lines) at confidence level 0.95.

The proposed estimator has good asymptotic behavior, and the confidence band had coverage very close to the nominal in our simulation study. An empirical study for the mean CD4+ cell counts illustrates the practical use of the confidence band.

Clearly the simultaneous confidence band in (3.14) can be improved in terms of both theoretical and numerical performance if higher order spline or local linear estimators are used. Constant piecewise spline estimators are less appealing and have sub-optimal convergence rates in the sense of Hall, Müller, and Wang (2006), which uses local linear approaches. Establishing the asymptotic confidence level for such extensions, however, requires highly sophisticated extreme value theory, for sequences of non-stationary Gaussian processes over intervals growing to infinity. That is much more difficult than the proofs of this chapter. We consider the confidence band in (3.14) significant because it is the first of its kind for the longitudinal case with complete theoretical justification, and with satisfactory numerical performance for commonly encountered data sizes.

Our methodology can be applied to construct simultaneous confidence bands for other functional objects, such as the covariance function $G\left(x, x'\right)$ and its eigenfunctions, see Yao (2007). It can also be adapted to the estimation of regression functions in the functional linear model, as in Li and Hsing (2007). We expect further research along these lines to yield deep theoretical results with interesting applications.

## 3.8 Appendix

Throughout this section, $a_n \sim b_n$ means $\lim_{n\to\infty} b_n/a_n = c$, where $c$ is some nonzero constant, and for functions $a_n(x), b_n(x)$, $a_n(x) = u\{b_n(x)\}$ means $a_n(x)/b_n(x) \to 0$ as $n \to \infty$ uniformly for $x \in [0, 1]$.

### 3.8.1 A.1. Preliminaries

We first state some results on strong approximation, extreme value theory and the classic Bernstein inequality. These are used in the proofs of Lemma 3.7, Theorem 3.1, and Lemma 3.6.

LEMMA 3.1. *(Theorem 2.6.7 of Csőrgő and Révész (1981)) Suppose that* $\xi_i, 1 \leq i \leq n$ *are iid with* $E(\xi_1) = 0, E(\xi_1^2) = 1$, *and* $H(x) > 0$ $(x \geq 0)$ *is an increasing continuous function such that* $x^{-2-\gamma}H(x)$ *is increasing for some* $\gamma > 0$ *and* $x^{-1}logH(x)$ *is decreasing with* $EH(|\xi_1|) < \infty$. *Then there exists a Wiener process* $\{W(t), 0 \leq t < \infty\}$ *that is a Borel function of* $\xi_i, 1 \leq i \leq n$, *and constants* $C_1, C_2, a > 0$ *which depend only on the distribution of* $\xi_1$, *such that for any* $\{x_n\}_{n=1}^{\infty}$ *satisfying* $H^{-1}(n) < x_n < C_1(nlogn)^{1/2}$ *and* $S_k = \sum_{i=1}^{k} \xi_i$,

$$P\left\{\max_{1\leq k\leq n} \left|S_k - W(k)\right| > x_n\right\} \leq C_2 n \left\{H(ax_n)\right\}^{-1}.$$

LEMMA 3.2. *Let* $\xi_i^{(n)}, 1 \leq i \leq n$, *be jointly normal with* $\xi_i^{(n)} \sim N(0,1)$. *Let* $r_{ij}^{(n)} = E\xi_i^{(n)}\xi_j^{(n)}$ *be such that for* $\gamma > 0, C_r > 0$, $\left|r_{ij}^{(n)}\right| < C_r/n^{\gamma}, i \neq j$. *Then for* $\tau \in R$, *as* $n \to \infty$, $P\left\{M_{n,\xi} \leq \tau/a_n + b_n\right\} \to \exp\left(-2e^{-\tau}\right)$, *in which* $M_{n,\xi} = \max\left\{\left|\xi_1^{(n)}\right|, \ldots, \left|\xi_n^{(n)}\right|\right\}$ *and* $a_n, b_n$ *are as in (3.7) with* $N_S + 1$ *replaced by* $n$.

PROOF. Let $\{\eta_i\}_{i=1}^{n}$ be i.i.d. standard normal r.v.'s, $\mathbf{u} = \{u_i\}_{i=1}^{n}, \mathbf{v} = \{v_i\}_{i=1}^{n}$ be vectors of real numbers, and $w = \min(|u_1|, \ldots, |u_n|, |v_1|, \ldots, |v_n|)$. By the Normal Comparison Lemma (Leadbetter, Lindgren and Rootzén (1983), Lemma 11.1.2),

$$\left|P\left\{-v_j < \xi_j^{(n)} \leq u_j \text{ for } j = 1, \ldots, n\right\} - P\left\{-v_j < \eta_j \leq u_j \text{ for } j = 1, \ldots, n\right\}\right|$$

$$\leq \frac{4}{2\pi}\sum\nolimits_{1\leq i<j\leq n}\left|r_{ij}^{(n)}\right|\left(1-\left|r_{ij}^{(n)}\right|^{2}\right)^{-1/2}\exp\left(\frac{-w^{2}}{1+r_{ij}^{(n)}}\right).$$

If $u_1=\cdots=u_n=v_1=\cdots=v_n=\tau/a_n+b_n=\tau_n$, it is clear that $\tau_n^2/\left(2\log n\right)\to 1$, as $n\to\infty$. Then $\tau_n^2>(2-\varepsilon)\log n$, for any $\varepsilon>0$ and large $n$. Since $1-r_{ij}^{(n)2}\geq 1-(C_r/n^{\gamma})^2\to 1$ as $n\to\infty, i\neq j$, for $i\neq j$, $\exists C_{r2}>0$ such that $1-r_{ij}^{(n)2}\geq C_{r2}>0$ and $1+r_{ij}^{(n)}<1+\epsilon$ for any $\epsilon>0$ and large $n$. Let $M_{n,\eta}=\max\left\{|\eta_1|,\ldots,|\eta_n|\right\}$. By Leadbetter, Lindgren and Rootzén (1983), Theorem 1.5.3, $P\left\{M_{n,\eta}\leq\tau_n\right\}\to\exp\left(-2e^{-\tau}\right)$ as $n\to\infty$, while the above results entail

$$\left|P\left(M_{n,\xi}\leq\tau_n\right)-P\left(M_{n,\eta}\leq\tau_n\right)\right|\leq\frac{4}{2\pi}\sum_{i<j}\left|r_{ij}^{(n)}\right|\left(1-\left|r_{ij}^{(n)}\right|^{2}\right)^{-1/2}\exp\left(\frac{-w^{2}}{1+r_{ij}^{(n)}}\right)$$
$$\leq\frac{4}{2\pi}\sum\nolimits_{1\leq i<j\leq n}C_r n^{-\gamma}C_{r2}^{-1/2}\exp\left\{\frac{-\left(2-\varepsilon\right)\log n}{1+\epsilon}\right\}\leq C_r' n^{2-\gamma-(2-\varepsilon)(1+\epsilon)^{-1}}\to 0$$

as $n\to\infty$. Hence $P\left\{M_{n,\xi}\leq\tau_n\right\}\to\exp\left(-2e^{-\tau}\right)$, as $n\to\infty$. □

LEMMA 3.3. *(Theorem 1.2 of Bosq (1998)) Suppose that* $\left\{\xi_i\right\}_{i=1}^n$ *are iid with* $E(\xi_1)=0, \sigma^2=E\xi_1^2$, *and there exists* $c>0$ *such that for* $r=3,4,...,E\left|\xi_1\right|^r\leq c^{r-2}r!E\xi_1^2<+\infty$. *Then for each* $n>1$, $t>0$, $P(|S_n|\geq\sqrt{n}\sigma t)\leq 2\exp\left(-t^2\left(4+2ct/\sqrt{n}\sigma\right)^{-1}\right)$, *in which* $S_n=\sum_{i=1}^n\xi_i$.

LEMMA 3.4. *Under Assumption (A2), as* $n\to\infty$ *for* $c_{J,n}$ *defined in (3.5),* $c_{J,n}=f\left(t_J\right)h_{\mathrm{S}}\left(1+r_{J,n}\right), \left\langle b_J, b_{J'}\right\rangle\equiv 0, J\neq J'$, *where* $\max_{0\leq J\leq N_{\mathrm{S}}}\left|r_{J,n}\right|\leq C\omega\left(f,h_{\mathrm{S}}\right)$. *There exist constants* $C_B>c_B>0$ *such that* $c_B h_{\mathrm{S}}^{1-r/2}\leq E\left\{B_J\left(X_{ij}\right)\right\}^r\leq C_B h_{\mathrm{S}}^{1-r/2}$ *for* $r=1,2,\ldots$ *and* $1\leq J\leq N_{\mathrm{S}}+1, 1\leq j\leq N_i, 1\leq i\leq n$.

PROOF. By the definition of $c_{J,n}$ in (3.5),

$$c_{J,n} = \int b_J(x) f(x) dx = \int_{[t_J, t_{J+1}]} f(x) dx = f\left(t_J\right) h_{\mathrm{S}} + \int_{[t_J, t_{J+1}]} \left\{ f(x) - f\left(t_J\right) \right\} dx.$$

Hence for all $J = 0, ..., N_{\mathrm{S}}$, $\left| c_{J,n} - f\left(t_J\right) h_{\mathrm{S}} \right| \leq \int_{[t_J, t_{J+1}]} \left| f(x) - f\left(t_J\right) \right| dx \leq \omega\left(f, h_{\mathrm{S}}\right) h_{\mathrm{S}}$, or $\left| r_{J,n} \right| = \left| c_{J,n} - f\left(t_J\right) h_{\mathrm{S}} \right| \left\{ f\left(t_J\right) h_{\mathrm{S}} \right\}^{-1} \leq C\omega\left(f, h_{\mathrm{S}}\right), J = 0, ..., N_{\mathrm{S}}$. By (3.8), $E\left\{ B_J\left(X_{ij}\right) \right\}^r = \left(c_{J,n}\right)^{-r/2} \int b_J(x) f(x) dx = \left(c_{J,n}\right)^{1-r/2} \sim h_{\mathrm{S}}^{1-r/2}$. □

PROOF OF PROPOSITION 3.1. By Lemma 3.4 and Assumption (A2) on the continuity of functions $\phi_k^2(x), \sigma^2(x)$ and $f(x)$ on $[0,1]$, for any $x \in [0,1]$

$$\left| \int_{\chi_{J(x)}} \phi_k(x) f(x) du - \int_{\chi_{J(x)}} \phi_k\left(u\right) f\left(u\right) du \right| \leq \omega\left(\phi_k f, h_{\mathrm{S}}\right) h_{\mathrm{S}} = O\left(h_{\mathrm{S}}^{1+\beta}\right),$$

$$\left| \int_{J(x)} \left\{ \sigma_Y^2(x) f(x) - \sigma_Y^2\left(u\right) f\left(u\right) \right\} du \right| \leq \omega\left(\sigma_Y^2 f, h_{\mathrm{S}}\right) h_{\mathrm{S}} = O\left(h_{\mathrm{S}}^{1+\beta}\right).$$

Hence,

$$\sigma_n^2(x) = c_{J(x),n}^{-2} \left(nEN_1\right)^{-1} \int_{J(x)} \sigma_Y^2\left(u\right) f\left(u\right) du \times$$
$$\left\{ 1 + \frac{E\left\{N_1\left(N_1 - 1\right)\right\}}{EN_1} \sum_{k=1}^{\kappa} \left( \int_{\chi_{J(x)}} \phi_k\left(u\right) f\left(u\right) du \right)^2 \left\{ \int_{J(x)} \sigma_Y^2\left(u\right) f\left(u\right) du \right\}^{-1} \right\}$$

$$= \left\{ f(x) h_{\mathrm{S}} + U\left(h_{\mathrm{S}}^{1+\beta}\right) \right\}^{-2} \left(nEN_1\right)^{-1} \left\{ \sigma_Y^2(x) f(x) h_{\mathrm{S}} + U\left(h_{\mathrm{S}}^{1+\beta}\right) \right\} \{1+$$
$$\frac{E\left\{N_1\left(N_1 - 1\right)\right\}}{EN_1} \sum_{k=1}^{\kappa} \left\{ \phi_k(x) f(x) h_{\mathrm{S}} + U\left(h_{\mathrm{S}}^{1+\beta}\right) \right\}^2 \left\{ \sigma_Y^2(x) f(x) h_{\mathrm{S}} + U\left(h_{\mathrm{S}}^{1+\beta}\right) \right\}^{-1} \Bigg\}$$

$$= (f(x)h_{\mathrm{S}}nEN_1)^{-1}\,\sigma_Y^2(x)\left\{1+\frac{E\left\{N_1\left(N_1-1\right)\right\}}{EN_1}\frac{\sum_{k=1}^{\kappa}\phi_k^2(x)f(x)h_{\mathrm{S}}}{\sigma_Y^2(x)}\right\}\left\{1+U\left(h_{\mathrm{S}}^{\beta}\right)\right\}$$

$$= \sigma_{n,\mathrm{LONG}}^2(x)\left\{1+U\left(h_{\mathrm{S}}^{\beta}\right)\right\} = \sigma_{n,\mathrm{IID}}^2(x)\left\{1+U\left(h_{\mathrm{S}}^{\beta}\right)\right\}.\square$$

### A.2. Proof of Theorem 1

Note that $B_{J(x)}(x) \equiv c_{J(x),n}^{-1/2}, x \in [0,1]$, so the terms $\widetilde{\xi}_k(x)$ and $\widetilde{\varepsilon}(x)$ defined in (3.12) are

$$\widetilde{\xi}_k(x) = \sum_{J=0}^{N_{\mathrm{S}}} N_{\mathrm{T}}^{-1}B_J(x)\left\|B_J\right\|_{2,N_{\mathrm{T}}}^{-2}\sum_{i=1}^{n}\sum_{j=1}^{N_i}B_J\left(X_{ij}\right)\phi_k\left(X_{ij}\right)\xi_{ik}$$

$$= c_{J(x),n}^{-1/2}\left\|B_{J(x)}\right\|_{2,N_{\mathrm{T}}}^{-2}N_{\mathrm{T}}^{-1}\sum_{i=1}^{n}\sum_{j=1}^{N_i}B_{J(x)}\left(X_{ij}\right)\phi_k\left(X_{ij}\right)\xi_{ik},$$

$$\widetilde{\varepsilon}(x) = c_{J(x),n}^{-1/2}\left\|B_{J(x)}\right\|_{2,N_{\mathrm{T}}}^{-2}N_{\mathrm{T}}^{-1}\sum_{i=1}^{n}\sum_{j=1}^{N_i}B_{J(x)}\left(X_{ij}\right)\sigma\left(X_{ij}\right)\varepsilon_{ij}.$$

Let

$$\widehat{\xi}_k(x) = \left\|B_{J(x)}\right\|_{2,N_{\mathrm{T}}}^{2}\widetilde{\xi}_k(x) = c_{J(x),n}^{-1/2}N_{\mathrm{T}}^{-1}\textstyle\sum_{i=1}^{n}R_{ik,\xi,J(x)}\xi_{ik},$$

$$\widehat{\varepsilon}(x) = \left\|B_{J(x)}\right\|_{2,N_{\mathrm{T}}}^{2}\widetilde{\varepsilon}(x) = c_{J(x),n}^{-1/2}N_{\mathrm{T}}^{-1}\textstyle\sum_{i=1}^{n}\sum_{j=1}^{N_i}R_{ij,\varepsilon,J(x)}\varepsilon_{ij}, \tag{3.16}$$

where

$$R_{ik,\xi,J} = \textstyle\sum_{j=1}^{N_i}B_J\left(X_{ij}\right)\phi_k\left(X_{ij}\right), R_{ij,\varepsilon,J} = B_J\left(X_{ij}\right)\sigma\left(X_{ij}\right), 0 \le J \le N_{\mathrm{S}}. \tag{3.17}$$

LEMMA 3.5. *Under Assumption (A3), for* $\widetilde{e}(x)$ *given in (3.11) and* $\widehat{\xi}_k(x)$, $\widehat{\varepsilon}(x)$ *given in (3.16), we have*

$$\left|\widetilde{e}(x)-\left\{\sum_{k=1}^{\kappa}\widehat{\xi}_k(x)+\widehat{\varepsilon}(x)\right\}\right|\leq A_n\left(1-A_n\right)^{-1}\left|\sum_{k=1}^{\kappa}\widehat{\xi}_k(x)+\widehat{\varepsilon}(x)\right|, x\in[0,1],$$

*where* $A_n=\sup_{0\leq J\leq N_{\mathrm{S}}}\left|\left\|B_J\right\|_{2,N_{\mathrm{T}}}^2-1\right|$. *There exists* $C_A>0$, *such that for large* $n$, $P\left(A_n\geq C_A\sqrt{log\left(n\right)/\left(nh_{\mathrm{S}}\right)}\right)\leq 2n^{-3}$. $A_n=O_{a.s.}\left(\sqrt{log\left(n\right)/\left(nh_{\mathrm{S}}\right)}\right)$ *as* $n\to\infty$.

See the Supplement of Wang and Yang (2009) for a detailed proof. □

LEMMA 3.6. *Under Assumptions (A2) and (A3), for* $R_{1k,\xi,J}$, $R_{11,\varepsilon,J}$ *in (3.17),*

$$ER_{1k,\xi,J}^2=c_{J,n}^{-1}\left[E\left(N_1\right)\int b_J\left(u\right)\phi_k^2\left(u\right)f\left(u\right)du+\right.$$
$$\left.E\left\{N_1\left(N_1-1\right)\right\}\left(\int b_J\left(u\right)\phi_k\left(u\right)f\left(u\right)du\right)^2\right],$$

$$ER_{11,\varepsilon,J}^2=c_{J,n}^{-1}\int b_J\left(u\right)\sigma^2\left(u\right)f\left(u\right)du, 0\leq J\leq N_{\mathrm{S}},$$

*there exist* $0<c_R<C_R<\infty$, *such that* $ER_{1k,\xi,J}^2, ER_{11,\varepsilon,J}^2\in\left[c_R,C_R\right]$ *for* $0\leq J\leq N_{\mathrm{S}}$, $\sup_{0\leq J\leq N_{\mathrm{S}}}\left|n^{-1}\sum_{i=1}^n R_{ik,\xi,J}^2-ER_{1k,\xi,J}^2\right|=O_{a.s.}\left(\sqrt{logn/\left(nh_{\mathrm{S}}\right)}\right)$, $1\leq k\leq\kappa$, $\sup_{0\leq J\leq N_{\mathrm{S}}}\left|N_{\mathrm{T}}^{-1}\sum_{i=1}^n\sum_{j=1}^{N_i}R_{ij,\varepsilon,J}^2-ER_{11,\varepsilon,J}^2\right|=O_{a.s.}\left(\sqrt{logn/\left(nh_{\mathrm{S}}\right)}\right)$ *as* $n\to\infty$.

PROOF. By independence of $X_{1j}, 1 \le j \le N_1$ and $N_1$ and (3.8),

$$ER_{1k,\xi,J}^2 = E\left\{\sum_{j,j'=1}^{N_1} E\left\{B_J\left(X_{1j}\right) B_J\left(X_{1j'}\right) \phi_k\left(X_{1j}\right) \phi_k\left(X_{1j'}\right) | N_1\right\}\right\}$$
$$= E\left\{\sum_{j=1}^{N_1} E\left\{B_J^2\left(X_{1j}\right) \phi_k^2\left(X_{1j}\right) | N_1\right\}\right\} +$$
$$E\left\{\sum_{j\neq j'}^{N_1} E\left\{B_J\left(X_{1j}\right) B_J\left(X_{1j'}\right) \phi_k\left(X_{1j}\right) \phi_k\left(X_{1j'}\right) | N_1\right\}\right\}$$
$$= c_{J(x),n}^{-1}\left\{E\left(N_1\right) \int b_J\left(u\right) \phi_k^2\left(u\right) f\left(u\right) du +\right.$$

$$\left. E\left\{N_1\left(N_1-1\right)\right\}\left(\int b_J\left(u\right) \phi_k\left(u\right) f\left(u\right) du\right)^2\right\}.$$

It is easily shown that $\exists\ 0 < c_R < C_R < \infty$ such that $c_R \le ER_{1k,\xi,J}^2 \le C_R, 0 \le J \le N_{\mathrm{S}}$. Let $\zeta_{i,J} = \zeta_{i,k,J} = R_{ik,\xi,J}^2, \zeta_{i,J}^* = \zeta_{i,J} - E\left(\zeta_{1,J}\right)$ for $r \ge 1$ and large $n$,

$$E\left(\zeta_{i,J}\right)^r = E\left\{\sum_{j=1}^{N_i} B_J\left(X_{ij}\right) \phi_k\left(X_{ij}\right)\right\}^{2r} \le C_\phi^{2r} E\left\{\sum_{j=1}^{N_i} B_J\left(X_{ij}\right)\right\}^{2r}$$
$$= C_\phi^{2r} E\left\{\sum_{0\le\nu_1\cdots\nu_{N_i}\le 2r}^{\nu_1+\cdots+\nu_{N_i}=2r} \binom{2r}{\nu_1\cdots\nu_{N_i}} \prod_{j=1}^{N_i} E\left\{B_J\left(X_{ij}\right)\right\}^{\nu_j}\right\}$$

$$\le C_\phi^{2r} E\left\{N_1^{2r} \max\left\{\prod_{j=1}^{N_i} E\left\{B_J\left(X_{ij}\right)\right\}^{\nu_j}\right\}\right\} \le C_\phi^{2r}\left(EN_1^{2r}\right) C_B h_{\mathrm{S}}^{1-r}$$
$$\le C_\phi^{2r} C_B c_N^r r! h_{\mathrm{S}}^{1-r} = C_\zeta r! h_{\mathrm{S}}^{1-r},$$
$$E\left(\zeta_{i,J}\right)^r \ge c_\phi^{2r} E\left\{\sum_{j=1}^{N_i} E\left\{B_J\left(X_{ij}\right)\right\}^{2r}\right\} \ge c_\phi^{2r}\left(EN_1\right) c_B h_{\mathrm{S}}^{1-r},$$

by Lemma 3.4. So $\left\{E\left(\zeta_{1,J}\right)\right\}^r \sim 1$, $E\left(\zeta_{i,J}\right)^r \gg \left\{E\left(\zeta_{1,J}\right)\right\}^r$ for $r \ge 2$, and $\exists C_\zeta' >$

$c'_\zeta > 0$ such that $C'_\zeta h_{\rm S}^{-1} \geq \sigma^2_{\zeta^*} \geq c'_\zeta h_{\rm S}^{-1}$, for $\sigma_{\zeta^*} = \left\{E\left(\zeta^*_{i,J}\right)^2\right\}^{1/2}$. We obtain $E\left|\zeta^*_{i,J}\right|^r \leq c_*^{r-2} r! E\left(\zeta^*_{i,J}\right)^2$ with $c_* = \left(C_\zeta/c'_\zeta\right)^{\frac{1}{r-2}} h_{\rm S}^{-1}$, which implies that $\left\{\zeta^*_{i,J}\right\}_{i=1}^n$ satisfies Cramér's condition. Applying Lemma 3.3 to $\sum_{i=1}^n \zeta^*_{i,J}$, for $r > 2$ and any large enough $\delta > 0$, $P\left\{n^{-1}\left|\sum_{i=1}^n \zeta^*_{i,J}\right| \geq \delta\sqrt{\log n/(nh_{\rm S})}\right\}$ is bounded by

$$2\exp\left\{\frac{-\delta^2\left(C'_\zeta\right)^{-1}(\log n)}{4 + 2\left(C_\zeta/c'_\zeta\right)^{\frac{1}{r-2}}\delta\left(c'_\zeta\right)^{-1}h_{\rm S}^{1/2}(\log n)^{1/2}n^{-1/2}}\right\}$$
$$\leq 2\exp\left\{\frac{-\delta^2(\log n)}{4C'_\zeta}\right\} \leq 2n^{-3}.$$

Hence $\sum_{n=1}^\infty P\left\{\sup_{0\leq J\leq N_{\rm S}}\left|\frac{1}{n}\sum_{i=1}^n R^2_{ik,\xi,J} - ER^2_{1k,\xi,J}\right| \geq \delta\sqrt{\log n/(nh_{\rm S})}\right\} \leq \sum_{n=1}^\infty \frac{2N_{\rm S}}{n^3} < \infty$.

Thus, $\sup_{0\leq J\leq N_{\rm S}}\left|n^{-1}\sum_{i=1}^n R^2_{ik,\xi,J} - ER^2_{1k,\xi,J}\right| = O_{\text{a.s.}}\left(\sqrt{\log n/(nh_{\rm S})}\right)$ as $n\to\infty$ by Borel-Cantelli Lemma. The properties of $R_{ij,\varepsilon,J}$ are obtained similarly. □

Order all $X_{ij}, 1\leq j\leq N_i, 1\leq i\leq n$ from large to small as $X_{(t)}$, $X_{(1)}\geq\cdots\geq X_{(N_{\rm T})}$, and denote the $\varepsilon_{ij}$ corresponding to $X_{(t)}$ as $\varepsilon_{(t)}$. By (3.16),

$$\begin{aligned}\widehat{\varepsilon}(x) &= c^{-1}_{J(x),n}N_{\rm T}^{-1}\sum_{t=1}^{N_{\rm T}} b_{J(x)}\left(X_{(t)}\right)\sigma\left(X_{(t)}\right)\varepsilon_{(t)}\\ &= c^{-1}_{J(x),n}N_{\rm T}^{-1}\sum_{t=1}^{N_{\rm T}} b_{J(x)}\left(X_{(t)}\right)\sigma\left(X_{(t)}\right)\{S_t - S_{t-1}\},\end{aligned}$$

where $S_q = \sum_{t=1}^q \varepsilon_{(t)}, q\geq 1$ and $S_0 = 0$.

LEMMA 3.7. *Under Assumptions (A2)-(A5), there is a Wiener process* $\{W(t), 0\leq t<\infty\}$ *independent of* $\left\{N_i, X_{ij}, 1\leq j\leq N_i, \xi_{ik}, 1\leq k\leq\kappa, 1\leq i\leq n\right\}$, *such that as* $n\to\infty$,

$\sup_{x\in[0,1]}\left|\widehat{\varepsilon}^{(0)}(x)-\widehat{\varepsilon}(x)\right|=o_{a.s.}\left(n^{t}\right)$ *for some* $t<-\left(1-\vartheta\right)/2<0$, *where* $\widehat{\varepsilon}^{(0)}(x)$ *is*

$$\left(c_{J(x),n}N_{\mathrm{T}}\right)^{-1}\sum_{t=1}^{N_{\mathrm{T}}}b_{J(x)}\left(X_{(t)}\right)\sigma\left(X_{(t)}\right)\left\{W\left(t\right)-W\left(t-1\right)\right\},x\in[0,1]. \quad (3.18)$$

Proof. Define $M_{N_{\mathrm{T}}}=\max_{1\leq q\leq N_{\mathrm{T}}}\left|Sq-W\left(q\right)\right|$, in which $\left\{W\left(t\right),0\leq t<\infty\right\}$ is the Wiener process in Lemma 3.1 that a Borel function of the set of variables $\left\{\varepsilon_{(t)}\ 1\leq t\leq N_{\mathrm{T}}\right\}$ is independent of $\left\{N_i,X_{ij},1\leq j\leq N_i,\xi_{ik},1\leq k\leq\kappa,1\leq i\leq n\right\}$ since $\left\{\varepsilon_{(t)}\ 1\leq t\leq N_{\mathrm{T}}\right\}$ is. Further, $\sup_{x\in[0,1]}\left|\widehat{\varepsilon}^{(0)}(x)-\widehat{\varepsilon}(x)\right|$ equals to

$$\begin{aligned}
&\sup_{x\in[0,1]}c_{J(x),n}^{-1}N_{\mathrm{T}}^{-1}\left|b_{J(x)}\left(X_{\left(N_{\mathrm{T}}\right)}\right)\sigma\left(X_{\left(N_{\mathrm{T}}\right)}\right)\left\{W\left(N_{\mathrm{T}}\right)-S_{N_{\mathrm{T}}}\right\}\right.\\
&\left.+\sum_{t=1}^{N_{\mathrm{T}}-1}\left\{b_{J(x)}\left(X_{(t)}\right)\sigma\left(X_{(t)}\right)-b_{J(x)}\left(X_{(t+1)}\right)\sigma\left(X_{(t+1)}\right)\right\}\left\{W\left(t\right)-S_t\right\}\right|\\
&\leq\max_{0\leq J\leq N_{\mathrm{S}}+1}c_{Jn}^{-1}N_{\mathrm{T}}^{-1}\left\{b_J\left(X_{\left(N_{\mathrm{T}}\right)}\right)\sigma\left(X_{\left(N_{\mathrm{T}}\right)}\right)+\right.\\
&\left.\sum_{t=1}^{N_{\mathrm{T}}-1}\left|b_J\left(X_{(t)}\right)\sigma\left(X_{(t)}\right)-b_J\left(X_{(t+1)}\right)\sigma\left(X_{(t+1)}\right)\right|\right\}M_{N_{\mathrm{T}}}
\end{aligned}$$

$$\leq\max_{0\leq J\leq N_{\mathrm{S}}+1}c_{J,n}^{-1}N_{\mathrm{T}}^{-1}M_{N_{\mathrm{T}}}\left\{3C_\sigma+\sum_{1\leq t\leq N_{\mathrm{T}}-1,X_{(t)}\in b_J}\left|\sigma\left(X_{(t)}\right)-\sigma\left(X_{(t+1)}\right)\right|\right\}$$

which, by the Hölder continuity of $\sigma$ in Assumption (A2), is bounded by

$$N_{\mathrm{T}}^{-1}M_{N_{\mathrm{T}}}\max_{0\leq J\leq N_{\mathrm{S}}+1}c_{J,n}^{-1}\left\{3C_\sigma+\left\|\sigma\right\|_{0,\beta}\sum_{1\leq t\leq N_{\mathrm{T}}-1,X_{(t)}\in b_J}\left|X_{(t)}-X_{(t+1)}\right|^{\beta}\right\}\leq$$

$$N_{\mathrm{T}}^{-1}M_{N_{\mathrm{T}}}\max_{J}c_{J,n}^{-1}\left\{3C_\sigma+\left\|\sigma\right\|_{0,\beta}n_J^{1-\beta}\left(\sum_{1\leq t\leq N_{\mathrm{T}}-1,X_{(t)}\in b_J}\left|X_{(t)}-X_{(t+1)}\right|\right)^{\beta}\right\}$$

$$\leq N_{\mathrm{T}}^{-1} M_{N_{\mathrm{T}}} \left( \max_{0 \leq J \leq N_{\mathrm{S}}+1} c_{J,n}^{-1} \right) \left\{ 3C_\sigma + \|\sigma\|_{0,\beta} \, h_{\mathrm{S}}^{\beta} \left( \max_{0 \leq J \leq N_{\mathrm{S}}+1} n_J \right)^{1-\beta} \right\}$$

where $n_J = \sum_{t=1}^{N_{\mathrm{T}}} I\left(X_{(t)} \in \chi_J\right), 0 \leq J \leq N_{\mathrm{S}}+1$, has a binomial distribution with parameters $\left(N_{\mathrm{T}}, p_{J,n}\right)$, where $p_{J,n} = \int_{\chi_J} f(x)dx$. Simple application of Lemma 3.3 entails $\max_{0 \leq J \leq N_{\mathrm{S}}+1} n_J = O_{\mathrm{a.s.}}\left(N_{\mathrm{T}} N_{\mathrm{S}}^{-1}\right)$. Meanwhile, by letting $H(x) = x^r, x_n = n^{t'}$, $t' \in (2/r, \beta - (1+\vartheta)/2)$, the existence of which is due to the Assumption (A4) that $r > 2/\{\beta - (1+\vartheta)/2\}$. It is clear that $\left\{\varepsilon_{(t)}\right\}_{t=1}^{N_{\mathrm{T}}}$ satisfies the conditions in Lemma 3.1. Since $\frac{n}{H(ax_n)} = a^{-r} n^{1-rt'} = O\left(n^{-\gamma_1}\right)$ for some $\gamma_1 > 1$, one can use the probability inequality in Lemma 3.1 and the Borel-Cantelli Lemma to obtain $M_{N_{\mathrm{T}}} = O_{\mathrm{a.s.}}(x_n) = O_{\mathrm{a.s.}}\left(n^{t'}\right)$. Hence Lemma 3.4 and the above imply

$$\sup_{x \in [0,1]} \left| \widehat{\varepsilon}^{(0)}(x) - \widehat{\varepsilon}(x) \right| = O_{\mathrm{a.s.}}\left(N_{\mathrm{S}} n^{t'-1}\right) \left\{ 1 + N_{\mathrm{S}}^{-\beta} \left( N_{\mathrm{T}} N_{\mathrm{S}}^{-1} \right)^{1-\beta} \right\}$$

$$\begin{aligned} &= O_{\mathrm{a.s.}}\left( N_{\mathrm{S}} n^{t'-1} + N_{\mathrm{S}} n^{t'-1} \times N_{\mathrm{S}}^{-1} n^{1-\beta} \right) \\ &= O_{\mathrm{a.s.}}\left( N_{\mathrm{S}} n^{t'-1} + N_{\mathrm{S}} n^{t'-\beta} \right) = o_{\mathrm{a.s.}}\left( n^{t'-\beta+\vartheta} \right) \end{aligned}$$

since $t' < \beta - (1+\vartheta)/2$ by definition, implying $t' - 1 \leq t' - \beta < -(1+\vartheta)/2$. The Lemma follows by setting $t = t' - \beta + \vartheta$. □

Now

$$\begin{aligned} \widehat{\varepsilon}^{(0)}(x) &= c_{J(x),n}^{-1} N_{\mathrm{T}}^{-1} \sum_{t=1}^{N_{\mathrm{T}}} b_{J(x)}\left(X_{(t)}\right) \sigma\left(X_{(t)}\right) Z_{(t)} \\ &= c_{J(x),n}^{-1} N_{\mathrm{T}}^{-1} \sum_{i=1}^{n} \sum_{j=1}^{N_i} b_{J(x)}\left(X_{ij}\right) \sigma\left(X_{ij}\right) Z_{ij}, \end{aligned} \tag{3.19}$$

where $Z_{(t)} = W(t) - W(t-1), 1 \le t \le N_{\mathrm{T}}$, are i.i.d $N(0,1)$, $\xi_{ik}, Z_{ij}, X_{ij}, N_i$ are independent, for $1 \le k \le \kappa, 1 \le j \le N_i, 1 \le i \le n$, and $\widehat{\xi}_k(x), \widehat{\varepsilon}^{(0)}(x)$ are conditional independent of $X_{ij}, N_i, 1 \le j \le N_i, 1 \le i \le n$. If the conditional variances of $\widehat{\xi}_k(x), \widehat{\varepsilon}^{(0)}(x)$ on $\left(X_{ij}, N_i\right)_{1 \le j \le N_i, 1 \le i \le n}$ are $\sigma^2_{\xi_k,n}(x), \sigma^2_{\varepsilon,n}(x)$, we have

$$\begin{aligned} \sigma_{\xi_k,n}(x) &= \left\{ c^{-1}_{J(x),n} N_{\mathrm{T}}^{-2} \sum_{i=1}^{n} R^2_{ik,\xi,J(x)} \right\}^{1/2} \\ \sigma_{\varepsilon,n}(x) &= \left\{ c^{-1}_{J(x),n} N_{\mathrm{T}}^{-2} \sum_{i=1}^{n} \sum_{j=1}^{N_i} R^2_{ij,\varepsilon,J(x)} \right\}^{1/2}, \end{aligned} \tag{3.20}$$

where $R_{ik,\xi,J(x)}, R_{ij,\varepsilon,J(x)}$, and $c_{J(x),n}$ are given in (3.17) and (3.5).

LEMMA 3.8. *Under Assumptions (A2) and (A3), let*

$$\eta(x) = \left\{ \sum_{k=1}^{\kappa} \sigma^2_{\xi_k,n}(x) + \sigma^2_{\varepsilon,n}(x) \right\}^{-1/2} \left\{ \sum_{k=1}^{\kappa} \widehat{\xi}_k(x) + \widehat{\varepsilon}^{(0)}(x) \right\}, \tag{3.21}$$

*with $\sigma_{\xi_k,n}(x), \sigma_{\varepsilon,n}(x), \widehat{\xi}_k(x), \widehat{\varepsilon}^{(0)}(x)$, and $c_{J(x),n}$ given in (3.20), (3.16), (3.18), and (3.5). Then $\eta(x)$ is a Gaussian process consisting of $(N_{\mathrm{S}}+1)$ standard normal variables $\{\eta_J\}_{J=0}^{N_{\mathrm{S}}}$ such that $\eta(x) = \eta_{J(x)}$ for $x \in [0,1]$, and there exists a constant $C > 0$ such that for large $n$, $\sup_{0 \le J \ne J' \le N_{\mathrm{S}}} \left| E\eta_J \eta_{J'} \right| \le C h_{\mathrm{S}}$.*

PROOF. It is apparent that $\mathcal{L}\left\{ \eta_J \,\middle|\, \left(X_{ij}, N_i\right), 1 \le j \le N_i, 1 \le i \le n \right\} = N(0,1)$ for $0 \le J \le N_{\mathrm{S}}$, so $\mathcal{L}\left\{\eta_J\right\} = N(0,1)$, for $0 \le J \le N_{\mathrm{S}}$. For $J \ne J'$, by (3.17) and (3.8), $R_{ij,\varepsilon,J} R_{ij,\varepsilon,J'} = B_J\left(X_{ij}\right) B_{J'}\left(X_{ij}\right) \sigma^2\left(X_{ij}\right) = 0$, along with (3.19), (3.18), the conditional independence of $\widehat{\xi}_k(x), \widehat{\varepsilon}^{(0)}(x)$ on $X_{ij}, N_i, 1 \le j \le N_i, 1 \le i \le n$, and independence

of $\xi_{ik}, Z_{ij}, X_{ij}, N_i, 1 \le k \le \kappa, 1 \le j \le N_i, 1 \le i \le n$, $E\left(\eta_J \eta_{J'}\right)$ is

$$E\left\{\left\{\sum_{i=1}^{n}\left\{\sum_{k=1}^{\kappa} R_{ik,\xi,J}^2 + \sum_{j=1}^{N_i} R_{ij,\varepsilon,J}^2\right\}\right\}^{-1/2}\left\{\sum_{i=1}^{n}\left\{\sum_{k=1}^{\kappa} R_{ik,\xi,J'}^2 + \sum_{j=1}^{N_i} R_{ij,\varepsilon,J'}^2\right\}\right\}^{-1/2} E\left\{\sum_{k=1}^{\kappa}\left\{\sum_{i=1}^{n} R_{ik,\xi,J}\xi_{ik}\right\}\left\{\sum_{i=1}^{n} R_{ik,\xi,J'}\xi_{ik}\right\} + \left\{\sum_{i=1}^{n}\sum_{j=1}^{N_i} R_{ij,\varepsilon,J} Z_{ij}\right\}\left\{\sum_{i=1}^{n}\sum_{j=1}^{N_i} R_{ij,\varepsilon,J'} Z_{ij}\right\} \Big| \left(X_{ij}, N_i\right)\right\}\right\}$$

$$= EC_{n,J,J'}$$

$$\text{in which} \quad C_{n,J,J'} = \left\{N_{\mathrm{T}}^{-1}\sum_{i=1}^{n}\left\{\sum_{k=1}^{\kappa} R_{ik,\xi,J}^2 + \sum_{j=1}^{N_i} R_{ij,\varepsilon,J}^2\right\}\right\}^{-1/2} \times \left\{N_{\mathrm{T}}^{-1}\sum_{i=1}^{n}\left\{\sum_{k=1}^{\kappa} R_{ik,\xi,J'}^2 + \sum_{j=1}^{N_i} R_{ij,\varepsilon,J'}^2\right\}\right\}^{-1/2}\left\{N_{\mathrm{T}}^{-1}\sum_{k=1}^{\kappa}\sum_{i=1}^{n} R_{ik,\xi,J} R_{ik,\xi,J'}\right\}.$$

Note that according to definitions of $R_{ik,\xi,J}, R_{ij,\varepsilon,J}$, and Lemma 3.5,

$$N_{\mathrm{T}}^{-1}\sum_{i=1}^{n}\left\{\sum_{k=1}^{\kappa} R_{ik,\xi,J}^2 + \sum_{j=1}^{N_i} R_{ij,\varepsilon,J}^2\right\}$$

$$\ge c_\sigma^2 N_{\mathrm{T}}^{-1}\sum_{i=1}^{n}\sum_{j=1}^{N_i} B_J^2\left(X_{ij}\right) = c_\sigma^2 \left\|B_J\right\|_{2,N_{\mathrm{T}}}^2 \ge c_\sigma^2\left(1 - A_n\right), \text{ for } 0 \le J \le N_{\mathrm{S}},$$

$$P\left\{\inf_{0 \le J \ne J' \le N_{\mathrm{S}}}\left\{N_{\mathrm{T}}^{-1}\sum_{i=1}^{n}\left(\sum_{k=1}^{\kappa} R_{ik,\xi,J}^2 + \sum_{j=1}^{N_i} R_{ij,\varepsilon,J}^2\right)\right\} \times \left\{N_{\mathrm{T}}^{-1}\sum_{i=1}^{n}\left(\sum_{k=1}^{\kappa} R_{ik,\xi,J'}^2 + \sum_{j=1}^{N_i} R_{ij,\varepsilon,J'}^2\right)\right\} \ge c_\sigma^4\left(1 - C_A\sqrt{\frac{\log(n)}{nh_{\mathrm{S}}}}\right)^2\right] \ge 1 - 2n^{-3},$$

by Lemma 3.5. Thus for large $n$, with probability $\geq 1-2n^{-3}$, the numerator of $C_{n,J,J'}$ is uniformly greater than $c_\sigma^2/2$. Applying Bernstein's inequality to $N_{\mathrm{T}}^{-1}\left\{\sum_{k=1}^{\kappa}\sum_{i=1}^{n}R_{ik,\xi,J}R_{ik,\xi,J'}\right\}$, there exists $C_0>0$ such that, for large $n$,

$$P\left(\sup_{0\leq J\neq J'\leq N_{\mathrm{S}}}\left|N_{\mathrm{T}}^{-1}\sum_{k=1}^{\kappa}\sum_{i=1}^{n}R_{ik,\xi,J}R_{ik,\xi,J'}\right|\leq C_0h_{\mathrm{S}}\right)\geq 1-2n^{-3}.$$

Putting the above together, for large $n$, $C_1=C_0\left(c_\sigma^2/2\right)^{-1}$,

$$P\left(\sup_{0\leq J\neq J'\leq N_{\mathrm{S}}}\left|C_{n,J,J'}\right|\leq C_1h_{\mathrm{S}}\right)\geq 1-4n^{-3}.$$

Note that as a continuous random variable, $\sup_{0\leq J\neq J'\leq N_{\mathrm{S}}}\left|C_{n,J,J'}\right|\in[0,1]$, thus

$$E\left(\sup_{0\leq J\neq J'\leq N_{\mathrm{S}}}\left|C_{n,J,J'}\right|\right)=\int_0^1 P\left(\sup_{0\leq J\neq J'\leq N_{\mathrm{S}}}\left|C_{n,J,J'}\right|>t\right)dt.$$

For large $n$, $C_1h_{\mathrm{S}}<1$ and then $E\left(\sup_{0\leq J\neq J'\leq N_{\mathrm{S}}}\left|C_{n,J,J'}\right|\right)$ is

$$\int_0^{C_1h_{\mathrm{S}}}P\left\{\sup_{0\leq J\neq J'\leq N_{\mathrm{S}}}\left|C_{n,J,J'}\right|>t\right\}dt+\int_{C_1h_{\mathrm{S}}}^1 P\left\{\sup_{0\leq J\neq J'\leq N_{\mathrm{S}}}\left|C_{n,J,J'}\right|>t\right\}dt$$

$$\leq\int_0^{C_1h_{\mathrm{S}}}1dt+\int_{C_1h_{\mathrm{S}}}^1 4n^{-3}dt\leq C_1h_{\mathrm{S}}+4n^{-3}\leq Ch_{\mathrm{S}}$$

for some $C>0$ and large enough $n$. The lemma now follows from

$$\sup_{0\leq J\neq J'\leq N_{\mathrm{S}}}\left|E\left(C_{n,J,J'}\right)\right|\leq E\left(\sup_{0\leq J\neq J'\leq N_{\mathrm{S}}}\left|C_{n,J,J'}\right|\right)\leq Ch_{\mathrm{S}}.\quad\square$$

By Lemma 3.8, the $(N_{\mathrm{S}}+1)$ standard normal variables $\eta_0, ..., \eta_{N_{\mathrm{S}}}$ satisfy the conditions of Lemma 3.2 Hence for any $\tau \in R$,

$$\lim_{n\to\infty} P\left(\sup_{x\in[0,1]} |\eta(x)| \leq \tau/a_{N_{\mathrm{S}}+1} + b_{N_{\mathrm{S}}+1}\right) = \exp\left(-2e^{-\tau}\right). \tag{3.22}$$

For $x \in [0,1]$, $R_{ik,\xi,J}, R_{ij,\varepsilon,J}$ given in (3.17), define the ratio of population and sample quantities as $r_n(x) = \left\{nE\left(N_1\right)/N_{\mathrm{T}}\right\}^{1/2} \left\{\bar{R}_n(x)/\bar{R}(x)\right\}^{1/2}$, with

$$\bar{R}_n(x) = N_{\mathrm{T}}^{-1}\left\{\sum_{i=1}^{n}\left(\sum_{k=1}^{\kappa} R_{ik,\xi,J(x)}^2 + \sum_{j=1}^{N_i} R_{ij,\varepsilon,J(x)}^2\right)\right\}$$
$$\bar{R}(x) = \left(EN_1\right)^{-1}\sum_{k=1}^{\kappa} ER_{1k,\xi,J(x)}^2 + ER_{11,\varepsilon,J(x)}^2.$$

Lemma 3.9. *Under Assumptions (A2), (A3), for $\eta(x), \sigma_n(x)$ in (3.21), (3.6),*

$$\left|\sigma_n(x)^{-1}\left\{\sum_{k=1}^{\kappa}\widehat{\xi}_k(x) + \widehat{\varepsilon}^{(0)}(x)\right\} - \eta(x)\right| = |r_n(x)-1|\,|\eta(x)| \tag{3.23}$$

*as $n\to\infty$, $\sup_{x\in[0,1]}\left\{a_{N_{\mathrm{S}}+1}\,|r_n(x)-1|\right\} = O_{a.s.}\left(\sqrt{\{log\left(N_{\mathrm{S}}+1\right)\}\left(logn\right)/\left(nh_{\mathrm{S}}\right)}\right)$.*

Proof. Equation (3.23) follows from the definitions of $\eta(x)$ and $\sigma_n(x)$. By Lemma 3.6,
$\sup_{x\in[0,1]}\left|N_{\mathrm{T}}^{-1}\sum_{i=1}^{n}\sum_{j=1}^{N_i} R_{ij,\varepsilon,J(x)}^2 - ER_{11,\varepsilon,J(x)}^2\right| = O_{\text{a.s.}}\left(\sqrt{\log n/\left(nh_{\mathrm{S}}\right)}\right)$,

$$\sup_{x\in[0,1]}\left|N_{\mathrm{T}}^{-1}\sum_{k=1}^{\kappa}\sum_{i=1}^{n} R_{ik,\xi,J(x)}^2 - \left(EN_1\right)^{-1}\sum_{k=1}^{\kappa} ER_{1k,\xi,J(x)}^2\right|$$

$$\leq \sup_{x\in[0,1]} (EN_1)^{-1} \sum_{k=1}^{\kappa} \left| n^{-1} \sum_{i=1}^{n} R^2_{ik,\xi,J(x)} - ER^2_{1k,\xi,J(x)} \right|$$

$$+ \sup_{x\in[0,1]} (EN_1)^{-1} \sum_{k=1}^{\kappa} \left| n (EN_1) N_{\mathrm{T}}^{-1} - 1 \right| \left| n^{-1} \sum_{i=1}^{n} R^2_{ik,\xi,J(x)} \right|$$

$$= O_{\text{a.s.}}\left(\sqrt{\log n/(nh_{\mathrm{S}})}\right) + O_{\text{a.s.}}\left(n^{-1/2}\right) = O_{\text{a.s.}}\left(\sqrt{\log n/(nh_{\mathrm{S}})}\right),$$

and there exist constants $0 < c_{\bar{R}} < C_{\bar{R}} < \infty$ such that for all $x \in [0,1]$, $c_{\bar{R}} < \bar{R}(x) < C_{\bar{R}}$. Thus, $\sup_{x\in[0,1]} \left|\bar{R}n(x) - \bar{R}(x)\right|$ is bounded by

$$\sup_{x\in[0,1]} \left| N_{\mathrm{T}}^{-1} \sum_{k=1}^{\kappa} \sum_{i=1}^{n} R^2_{ik,\xi,J(x)} - (EN_1)^{-1} \sum_{k=1}^{\kappa} ER^2_{1k,\xi,J(x)} \right| +$$
$$\sup_{x\in[0,1]} \left| N_{\mathrm{T}}^{-1} \sum_{i=1}^{n} \sum_{j=1}^{N_i} R^2_{ij,\varepsilon,J(x)} - ER^2_{11,\varepsilon,J(x)} \right| = O_{\text{a.s.}}\left(\sqrt{\log n/(nh_{\mathrm{S}})}\right).$$

Thus $\sup\limits_{x\in[0,1]} \left| \left\{\bar{R}_n(x)\right\}^{1/2} - \left\{\bar{R}(x)\right\}^{1/2} \right| \leq \sup\limits_{x\in[0,1]} \left|\bar{R}_n(x) - \bar{R}(x)\right| \sup\limits_{x\in[0,1]} \left\{\bar{R}(x)\right\}^{-1/2}$
$= O_{\text{a.s.}}\left(\sqrt{\log n/(nh_{\mathrm{S}})}\right)$. Then $\sup_{x\in[0,1]} \left\{a_{N_{\mathrm{S}}+1} |rn(x) - 1|\right\}$ is bounded by

$$a_{N_{\mathrm{S}}+1} \left\{ \left\{ nE(N_1)/N_{\mathrm{T}} \right\}^{1/2} \sup_{x\in[0,1]} \left| \left\{\bar{R}_n(x)/\bar{R}(x)\right\}^{1/2} - 1 \right| + \left| 1 - \left\{ nE(N_1)/N_{\mathrm{T}} \right\}^{1/2} \right| \right\}$$

$$\leq a_{N_{\mathrm{S}}+1} \left\{ \left\{ nE(N_1)/N_{\mathrm{T}} \right\}^{1/2} \sup_{x\in[0,1]} \left\{\bar{R}(x)\right\}^{-1/2} \sup_{x\in[0,1]} \left| \left\{\bar{R}_n(x)\right\}^{1/2} - \left\{\bar{R}(x)\right\}^{1/2} \right| \right.$$
$$\left. + \left| 1 - \left\{ nE(N_1)/N_{\mathrm{T}} \right\}^{1/2} \right| \right\} = O_{\text{a.s.}}\left(\sqrt{\{\log(N_{\mathrm{S}}+1)\}(\log n)/(nh_{\mathrm{S}})}\right). \quad \square$$

Proof of Proposition 3.2. The proof follows from Lemmas 3.5, 3.7, 3.9, (3.22), and Slutsky's Theorem. $\square$

PROOF OF THEOREM 3.1. By Theorem 3.2, $\|\widetilde{m}(x) - m(x)\|_\infty = O_p(h_{\mathrm{S}})$, so

$$a_{N_{\mathrm{S}}+1}\left(\sup_{x\in[0,1]}\sigma_n^{-1}(x)\left|\widetilde{m}(x) - m(x)\right|\right) = O_p\left\{(nh_{\mathrm{S}})^{1/2}\sqrt{\log(N_{\mathrm{S}}+1)}h_{\mathrm{S}}\right\} = o_p(1),$$

$$a_{N_{\mathrm{S}}+1}\left(\sup_{x\in[0,1]}\sigma_n^{-1}(x)\left|\widehat{m}(x) - m(x)\right| - \sup_{x\in[0,1]}\sigma_n^{-1}(x)\left|\sum_{k=1}^{\kappa}\widetilde{\xi}_k(x) + \widetilde{\varepsilon}(x)\right|\right) = o_p(1).$$

Meanwhile, (3.11) and Proposition 3.2 entail that, for any $\tau \in R$,

$$\lim_{n\to\infty} P\left\{a_{N_{\mathrm{S}}+1}\left(\sup_{x\in[0,1]}\sigma_n^{-1}(x)\left|\sum_{k=1}^{\kappa}\widetilde{\xi}_k(x) + \widetilde{\varepsilon}(x)\right| - b_{N_{\mathrm{S}}+1}\right) \le \tau\right\} = \exp\left(-2e^{-\tau}\right).$$

Thus Slutsky's Theorem implies that

$$\lim_{n\to\infty} P\left\{a_{N_{\mathrm{S}}+1}\left(\sup_{x\in[0,1]}\sigma_n^{-1}(x)\left|\widehat{m}(x) - m(x)\right| - b_{N_{\mathrm{S}}+1}\right) \le \tau\right\} = \exp\left(-2e^{-\tau}\right).$$

Let $\tau = -\log\left\{-\frac{1}{2}\log(1-\alpha)\right\}$, definitions of $a_{N_{\mathrm{S}}+1}, b_{N_{\mathrm{S}}+1}$, and $Q_{N_{\mathrm{S}}+1}(\alpha)$ in (3.7) entail

$$\begin{aligned}&\lim_{n\to\infty} P\left\{m(x) \in \widehat{m}(x) \pm \sigma_n(x)Q_{N_{\mathrm{S}}+1}(\alpha), \forall x \in [0,1]\right\}\\ =& \lim_{n\to\infty} P\left\{Q_{N_{\mathrm{S}}+1}^{-1}(\alpha)\sup_{x\in[0,1]}\sigma_n^{-1}(x)\left|\widetilde{e}(x) + \widetilde{m}(x) - m(x)\right| \le 1\right\} = 1-\alpha.\end{aligned}$$

by (3.11). That $\sigma_n(x)^{-1}\{\widehat{m}(x) - m(x)\} \to_d N(0,1)$ for any $x \in [0,1]$ follows by directly using $\eta(x) \sim N(0,1)$, without reference to $\sup_{x\in[0,1]}|\eta(x)|$. □

# Chapter 4

# Spline-backfitted Kernel Smoothing of Partially Linear Additive Model

## 4.1 Introduction

This chapter is based on Ma and Yang (2011b). Since the 1980's, 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, Thurston, Ruppert, Apanasovich and Hauser (2008)

$$Y_i = m\left(\mathbf{X}_i, \mathbf{T}_i\right) + \sigma\left(\mathbf{X}_i, \mathbf{T}_i\right)\varepsilon_i, m(\mathbf{x}, \mathbf{t}) = c_{00} + \sum_{l=1}^{d_1} c_{0l} t_l + \sum_{\alpha=1}^{d_2} m_\alpha\left(x_\alpha\right) \quad (4.1)$$

in which the sequence $\left\{Y_i, \mathbf{X}_i^T, \mathbf{T}_i^T\right\}_{i=1}^n = \left\{Y_i, X_{i1}, ..., X_{id_2}, T_{i1}, \ldots T_{id_1}\right\}_{i=1}^n$. 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., $Em_\alpha\left(X_{i\alpha}\right) \equiv 0, 1 \leq \alpha \leq d_2, ET_{il} = 0, 1 \leq l \leq d_1$.

If parameters $c_{0l} \equiv 0, 1 \leq l \leq d_1$, $\left(T_{i1}, \ldots T_{id_1}\right)$ are redundant. $\left\{Y_i, X_{i1}, ..., X_{id_2}\right\}_{i=1}^n$ follow an additive model. For applications of additive model, see, Nácher, Ojeda, Cadarso-Suárez, Roca-Pardiñas and Acuña (2006), Roca-Pardiñas, Cadarso-Suárez, Nácher and Acuña (2006), González-Manteiga, Martínez-Miranda and Raya-Miranda (2008). Additive model, however, is only appropriate to model nonparametric effects of continuous predictors $\left(X_{i1}, ..., X_{id_2}\right)$ supported on compact intervals. The effects of possibly discrete and/or unbounded predictors can be neatly modeled as some of the variables $\left(T_{i1}, \ldots T_{id_1}\right)$ in the PLAM (4.1), see the simulation example in Section 4.3 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 4.4 which shows that the simpler PLAM fits the Boston housing data much better than an additive model. For general references on partially linear 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, Wang and Carroll (2007) and Liang et al. (2008).

Satisfactory estimators of functions $\{m_\alpha(x_\alpha)\}_{\alpha=1}^{d_2}$ and constants $\{c_{0l}\}_{l=0}^{d_1}$ in model (4.1) based on $\left\{Y_i, \mathbf{X}_i^T, \mathbf{T}_i^T\right\}_{i=1}^n$ 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). 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 (4.1). If the regression coefficients $\{c_{0l}\}_{l=0}^{d_1}$ and the component functions $\left\{m_\beta\left(x_\beta\right)\right\}_{\beta=1,\beta\neq\alpha}^{d_2}$ were known by "oracle", one could create $\{Y_{i\alpha}, X_{i\alpha}\}_{i=1}^{n}$ with $Y_{i\alpha} = Y_i - c_{00} - \sum_{l=1}^{d_1} c_{0l}T_{il} - \sum_{\beta=1,\beta\neq\alpha}^{d_2} m_\beta\left(X_{i\beta}\right)$
$= m_\alpha\left(X_{i\alpha}\right) + \sigma\left(\mathbf{X}_i, \mathbf{T}_i\right)\varepsilon_i$, from which one could compute an "oracle smoother" to estimate the only unknown function $m_\alpha\left(x_\alpha\right)$, bypassing the "curse of dimensionality". The idea was to obtain approximations to the unobservable variables $Y_{i\alpha}$ by substituting $m_\beta\left(X_{i\beta}\right)$, $1 \leq i \leq n$, $1 \leq \beta \leq d_2$, $\beta \neq \alpha$, with spline estimates and argue that the error incurred by this "cheating" is of smaller magnitude than the rate $O\left(n^{-2/5}\right)$ for estimating $m_\alpha\left(x_\alpha\right)$ from the unobservable data. Lemmas 4.9, 4.14, 4.17 and 4.18 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).

This chapter is organized as follows. The SBK estimators are introduced in Section 4.2 with theoretical properties. Section 4.3 contains Monte Carlo results to demonstrate the asymptotic properties of SBK estimators for moderate dimensions. The SBK estimator is applied to the Boston housing data in Section 4.4. Proofs of technical lemmas are in the Appendix.

## 4.2 The SBK Estimators

For convenience, we denote vectors as $\mathbf{x} = \left(x_1, ..., x_d\right)^T$ 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., $\|\mathbf{x}\|_\infty = \sup_{1\leq\alpha\leq d} |x_\alpha|$. 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 $\text{diag}(a, b)$ the $2 \times 2$ diagonal matrix with diagonal entries $a, b$. Let $\left\{Y_i, \mathbf{X}_i^T, \mathbf{T}_i^T\right\}_{i=1}^{n}$ be a sequence of strictly stationary observations from a geometrically $\alpha$-mixing process following model (4.1), where $Y_i$ and $\left(\mathbf{X}_i, \mathbf{T}_i\right) = \{\left(X_{i1}, ..., X_{id_2}\right)^T, (T_{i1}, \ldots T_{id_1})^T\}$ are the $i$-th response and predictor vector. Denote $\mathbf{Y} = (Y_1, ..., Y_n)^T$ the response vector. Without loss of generality, we assume $X_\alpha$ is distributed on $[0, 1], 1 \leq \alpha \leq d_2$. An integer $N = N_n \sim n^{1/4} \log n$ is pre-selected. Denote the class of Lipschitz continuous functions for a constant $C > 0$ as $\text{Lip}([0, 1], C) = \left\{m | \left|m(x) - m(x')\right| \leq C \left|x - x'\right|, \forall x, x' \in [0, 1]\right\}$.

We use the second order B spline (or linear B spline) basis $b_J(x) = b_{J,2}(x)$, $0 \leq J \leq N + 1$ which is defined in Section 1.3 of Chapter 1. Let the distance between neighboring interior or boundary knots be $H = H_n = (N + 1)^{-1}$. Define next the space $G$ of partially

linear additive spline functions as the linear space spanned by

$$\left\{1, t_l, b_J\left(x_\alpha\right), 1 \leq l \leq d_1, 1 \leq \alpha \leq d_2, 1 \leq J \leq N+1\right\},$$

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

$$\hat{m}\left(\mathbf{x}, \mathbf{t}\right) = \hat{c}_{00} + \sum_{l=1}^{d_1} \hat{c}_{0l} t_l + \sum_{\alpha=1}^{d_2} \sum_{J=1}^{N+1} \hat{c}_{J,\alpha} b_J\left(x_\alpha\right), \tag{4.2}$$

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

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

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

$$\begin{aligned}
\hat{Y}_{i\alpha} &= Y_i - \hat{c}_{00} - \sum_{l=1}^{d_1} \hat{c}_{0l} T_{il} - \sum_{\beta=1, \beta \neq \alpha}^{d_2} \hat{m}_\beta\left(X_{i\beta}\right), \\
Y_{i\alpha} &= Y_i - c_{00} - \sum_{l=1}^{d_1} c_{0l} T_{il} - \sum_{\beta=1, \beta \neq \alpha}^{d_2} m_\beta\left(X_{i\beta}\right).
\end{aligned} \tag{4.3}$$

Based on $\left\{\hat{Y}_{i\alpha}, X_{i\alpha}\right\}_{i=1}^n$, the SBK estimator $\hat{m}_{\mathrm{SBK},\alpha}\left(x_\alpha\right)$ of $m_\alpha\left(x_\alpha\right)$ 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

$$\begin{aligned}\hat{m}_{\mathrm{SBK},\alpha}(x_\alpha) &= \left\{n^{-1}\sum_{i=1}^n K_h\left(X_{i\alpha}-x_\alpha\right)\hat{Y}_{i\alpha}\right\}/\hat{f}_\alpha(x_\alpha),\\ \tilde{m}_{\mathrm{K},\alpha}(x_\alpha) &= \left\{n^{-1}\sum_{i=1}^n K_h\left(X_{i\alpha}-x_\alpha\right)Y_{i\alpha}\right\}/\hat{f}_\alpha(x_\alpha),\end{aligned} \tag{4.4}$$

with $\hat{Y}_{i\alpha}, Y_{i\alpha}$ in (4.3), $\hat{f}_\alpha(x_\alpha) = n^{-1}\sum_{i=1}^n K_h\left(X_{i\alpha}-x_\alpha\right)$ 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 $n \to \infty$,

$$\begin{aligned}&\sup_{x_1\in[h,1-h]}\left|\tilde{m}_{\mathrm{K},1}(x_1) - m_1(x_1)\right| = o_p\left(n^{-2/5}\log n\right),\\ &\sqrt{nh}\left\{\tilde{m}_{\mathrm{K},1}(x_1) - m_1(x_1) - b_1(x_1)h^2\right\} \xrightarrow{D} N\left\{0, v_1^2(x_1)\right\},\\ &\text{where, } b_1(x_1) = \int u^2 K(u)\,du\left\{m_1''(x_1)f_1(x_1)/2 + m_1'(x_1)f_1'(x_1)\right\}f_1^{-1}(x_1),\\ &v_1^2(x_1) = \int K^2(u)\,du E\left[\sigma^2(\mathbf{X},\mathbf{T})\,|X_1 = x_1\right]f_1^{-1}(x_1).\end{aligned} \tag{4.5}$$

It is shown in Li (2000) and Schimek (2000) that the spline estimator $\hat{m}_1(x_1)$ in the first step uniformly converges to $m_1(x_1)$ with certain convergence rate, but lacks asymptotic distribution. Theorem 4.1 below states that the difference between $\hat{m}_{\mathrm{SBK},1}(x_1)$ and $\tilde{m}_{\mathrm{K},1}(x_1)$ is $o_p\left(n^{-2/5}\right)$ uniformly, dominated by the asymptotic uniform size of $\tilde{m}_{\mathrm{K},1}(x_1) - m_1(x_1)$. So $\hat{m}_{\mathrm{SBK},1}(x_1)$ has identical asymptotic distribution as $\tilde{m}_{\mathrm{K},1}(x_1)$.

THEOREM 4.1. *Under Assumptions A1-A7, as $n \to \infty$, the SBK estimator $\hat{m}_{\mathrm{SBK},1}(x_1)$ given in (4.4) satisfies* $\sup_{x_1\in[0,1]}\left|\hat{m}_{\mathrm{SBK},1}(x_1) - \tilde{m}_{\mathrm{K},1}(x_1)\right| = o_p\left(n^{-2/5}\right)$. *Hence with*

$b_1(x_1)$ *and* $v_1^2(x_1)$ *as defined in (4.5), for any* $x_1 \in [h, 1-h]$,

$$\sqrt{nh}\left\{\hat{m}_{\mathrm{SBK},1}(x_1) - m_1(x_1) - b_1(x_1)h^2\right\} \xrightarrow{D} N\left\{0, v_1^2(x_1)\right\}.$$

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

$$\hat{m}_{\mathrm{SBK},1}(x_1) - \hat{b}_1(x_1)h^2 \pm z_{\alpha/2}\hat{v}_1(x_1)(nh)^{-1/2} \tag{4.6}$$

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 4.1. *Under Assumptions A1-A7 and the additional assumption* $m_\alpha \in C^{(2)}[0,1], 2 \leq \alpha \leq d_2$. *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* $\hat{m}_{\mathrm{SBK},\alpha}(x_\alpha), 1 \leq \alpha \leq d_2$, *defined in (4.4), and* $b_\alpha(x_\alpha)$, $v_\alpha^2(x_\alpha)$ *similarly defined as in (4.5), as* $n \to \infty$,

$$\sqrt{nh}\left\{\hat{m}_{\mathrm{SBK}}(\mathbf{x}) - \sum_{\alpha=1}^{d_2} m_\alpha(x_\alpha) - b(\mathbf{x})h^2\right\} \xrightarrow{D} N\left\{0, v^2(\mathbf{x})\right\}$$

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

Theorem 4.2. *Under Assumptions A1-A6, as* $n \to \infty$, $\|\hat{\mathbf{c}} - \mathbf{c}\| = O_p\left(n^{-1/2}\right)$. *With the additional Assumption A8,* $\sqrt{n}(\hat{\mathbf{c}} - \mathbf{c}) \to_d N\left(0, \sigma_0^2 \begin{Bmatrix} 1 & 0_{d_1}^T \\ 0_{d_1} & \Sigma^{-1} \end{Bmatrix}\right)$, *for* $\Sigma = \mathrm{cov}\left(\tilde{\mathbf{T}}\right)$

*with random vector* $\tilde{\mathbf{T}}$ *defined in (4.10).*

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 $b^*_{J,\alpha}(x_\alpha)\equiv b_J(x_\alpha)-n^{-1}\sum_{i=1}^{n}b_J\left(X_{i\alpha}\right)$ is the empirical centering of $b_J(x)$ for the $\alpha$-th variable $X_\alpha$, defined in the Appendix and $\left(\hat{a}_{J,\alpha}\right)_{1\leq J\leq N+1,1\leq\alpha\leq d_2}$ minimize

$$\left\|T_l-\sum_{\alpha=1}^{d_2}\sum_{J=1}^{N+1}a_{J,\alpha}b^*_{J,\alpha}(X_\alpha)\right\|_n^2.$$

## 4.3 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 (4.2) is given by $N=\min\left(\left[c_1n^{1/4}\log n\right]+c_2,\left[(n/2-1-d_1)\,d_2^{-1}-1\right]\right)$, in which $[a]$ denotes the integer part of $a$. In our implementation, we have used $c_1=c_2=1$. The additional constraint that $N\leq(n/2-1-d_1)\,d_2^{-1}-1$ ensures that the number of terms in the linear least squares problem (4.2), $1+d_1+d_2(N+1)$, is no greater than $n/2$, which is necessary when the sample size $n$ is moderate.

The i.i.d. data $\{Y_i,\mathbf{X}_i,\mathbf{T}_i\}_{i=1}^n$ is generated according to the partially linear additive model (4.1), which satisfies Assumptions A1-A5, and A8

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

where $\sigma_0=2$, $\varepsilon_i\sim N(0,1)$ is independent of $\left(\mathbf{X}_i,\mathbf{T}_i\right)$, $\mathbf{T}_i=\left(T_{i1},T_{i2},T_{i3},T_{i4}\right)$ such that $T_{i3},T_{i4},\left(T_{i1},T_{i2}\right)$ are independent, $T_{i3}=\pm1$ with probability $1/2$, $T_{i4}\sim U(-0.5,0.5)$,

$\left(T_{i1}, T_{i2}\right)' \sim N\left((0,0)', \text{diag}\left(a\left(X_{i1}\right), a\left(X_{i2}\right)\right)\right)$, $a(x) = \frac{5-\sin(2\pi x)}{5+\sin(2\pi x)}$.
$\mathbf{X}_i = \left\{\left(X_{i\alpha}\right)_{\alpha=1}^{d_2}\right\}^T$ is generated from the vector autoregression (VAR) equation $X_{i\alpha} = \Phi\left\{\left(1-a^2\right)^{1/2} Z_{i\alpha}\right\} - 1/2, 1 \le \alpha \le d_2$ with stationary distribution
$\mathbf{Z}_i = \left(Z_{i1}, ..., Z_{id_2}\right)^T \sim N\left(0_{d_2}, \left(1-a^2\right)^{-1} \mathbf{\Sigma}\right)$

$$\begin{aligned} \mathbf{Z}_1 &\sim N\left(0_{d_2}, \left(1-a^2\right)^{-1}\mathbf{\Sigma}\right), \mathbf{Z}_i = a\mathbf{Z}_{i-1} + \boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_i \sim N\left(0, \mathbf{\Sigma}\right), 2 \le i \le n, \\ \mathbf{\Sigma} &= (1-r)\,\mathbf{I}_{d_2 \times d_2} + r\mathbf{1}_{d_2}\mathbf{1}_{d_2}^T, 0 < a < 1, 0 < r < 1, \end{aligned}$$

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

$$\text{eff}_\alpha = \left[\frac{\sum_{i=1}^n \left\{\tilde{m}_{\text{K},\alpha}(x_\alpha) - m_\alpha\left(X_{i\alpha}\right)\right\}^2}{\sum_{i=1}^n \left\{\hat{m}_{\text{SBK},\alpha}(x_\alpha) - m_\alpha\left(X_{i\alpha}\right)\right\}^2}\right]^{1/2}. \tag{4.7}$$

Theorem 4.1 indicates $\text{eff}_\alpha$ should be close to 1 for $1 \le \alpha \le d_2$. Figures 4.1, 4.2, 4.3, and 4.4 provide 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 $\sigma_0 = 2$, $d_2 = 3$, and $d_2 = 30$. 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 4.1.

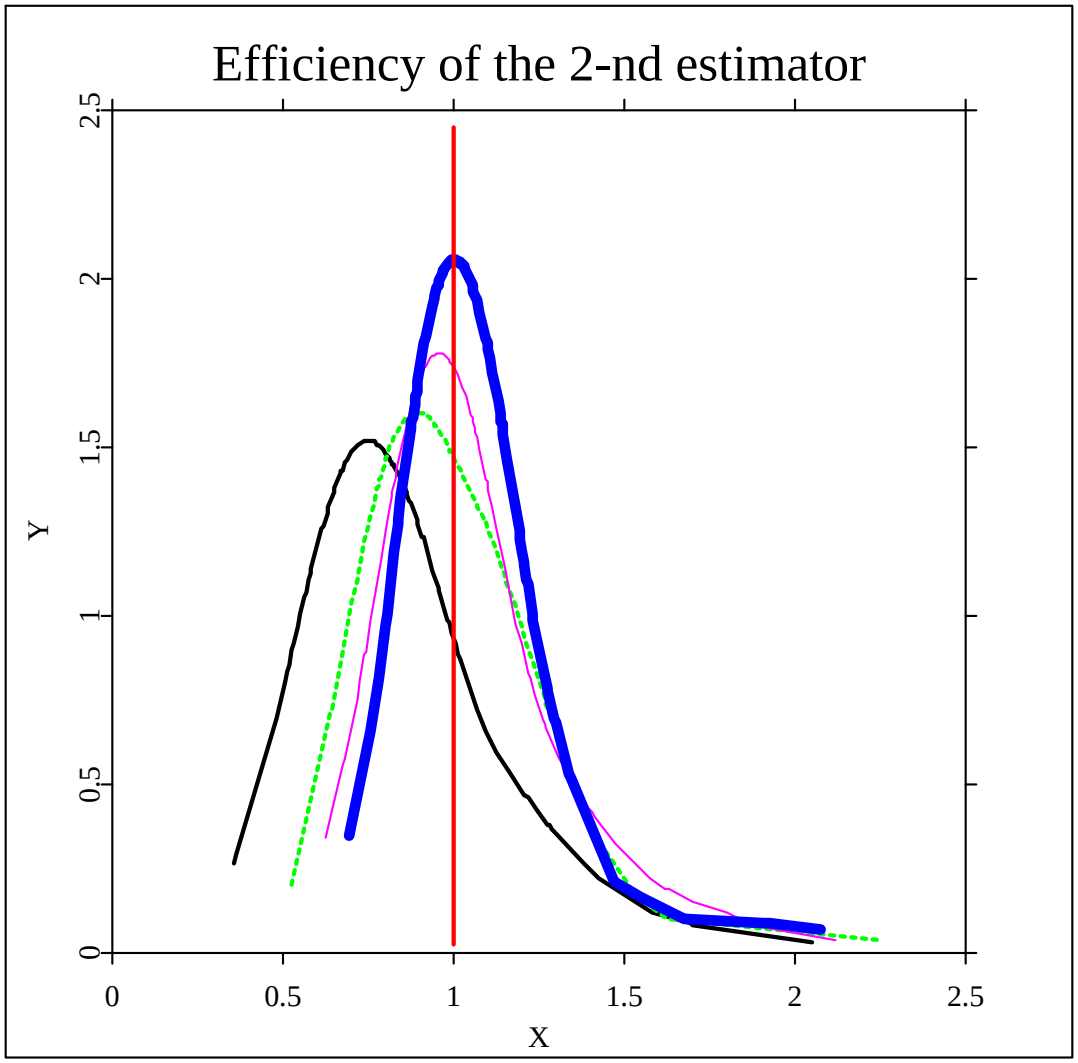


Figure 4.1: Kernel density plots for $\alpha = 2, d_1 = 4, d_2 = 3$ in Chapter 4

Note: kernel density plots of the 100 empirical efficiencies of $\hat{m}_{\mathrm{SBK},\alpha}(x_\alpha)$ to $\tilde{m}_{\mathrm{K},\alpha}(x_\alpha)$, computed according to (4.7) for sample sizes $n = 100$ (solid lines), 200 (dashed lines), 500 (thin lines) and 1000 (thick lines) at $\alpha = 2, d_1 = 4, d_2 = 3$.

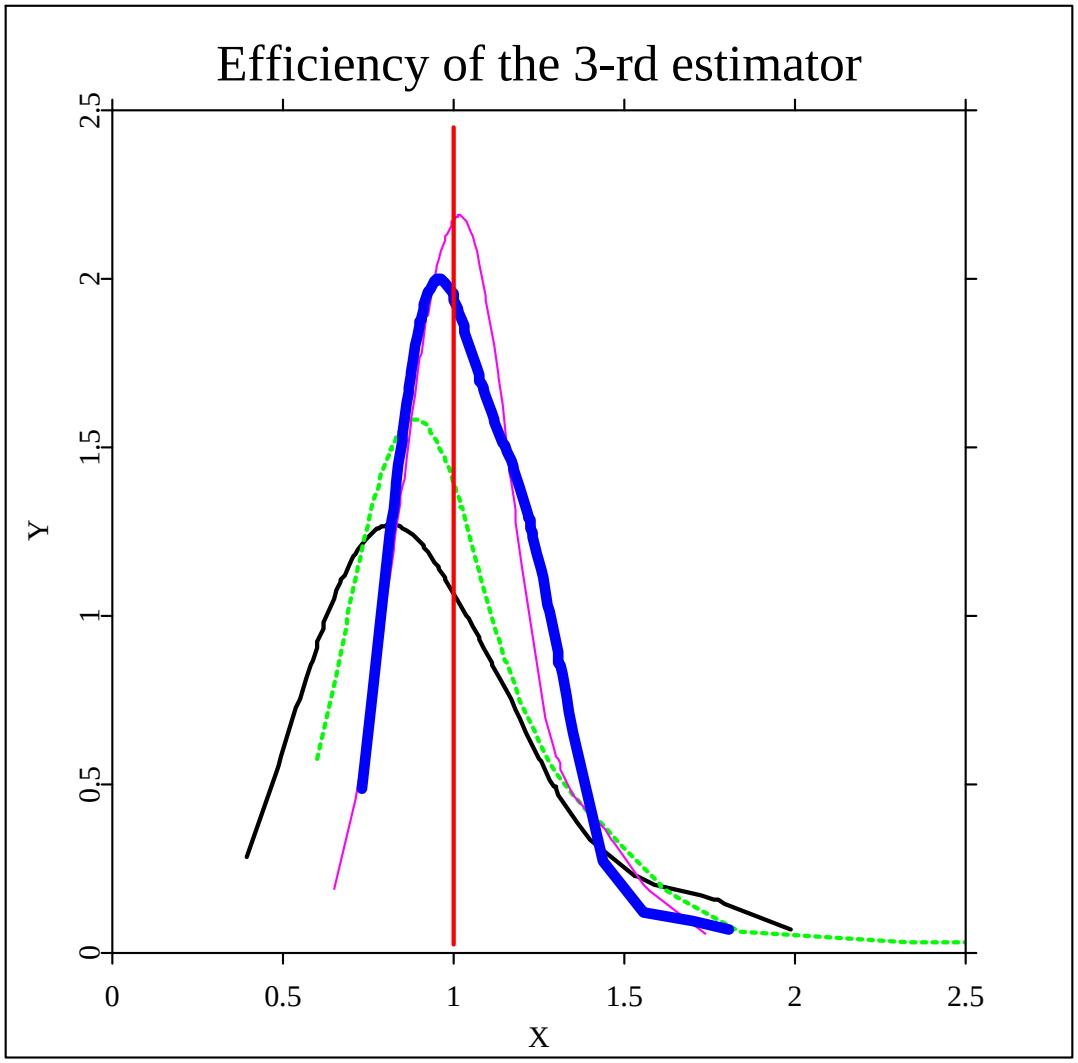


Figure 4.2: Kernel density plots in Chapter 4

Note: kernel density plots of the 100 empirical efficiencies of $\hat{m}_{\mathrm{SBK},\alpha}(x_\alpha)$ to $\tilde{m}_{\mathrm{K},\alpha}(x_\alpha)$, computed according to (4.7) for sample sizes $n = 100$ (solid lines), 200 (dashed lines), 500 (thin lines) and 1000 (thick lines) at $\alpha = 3, d_1 = 4, d_2 = 3$.

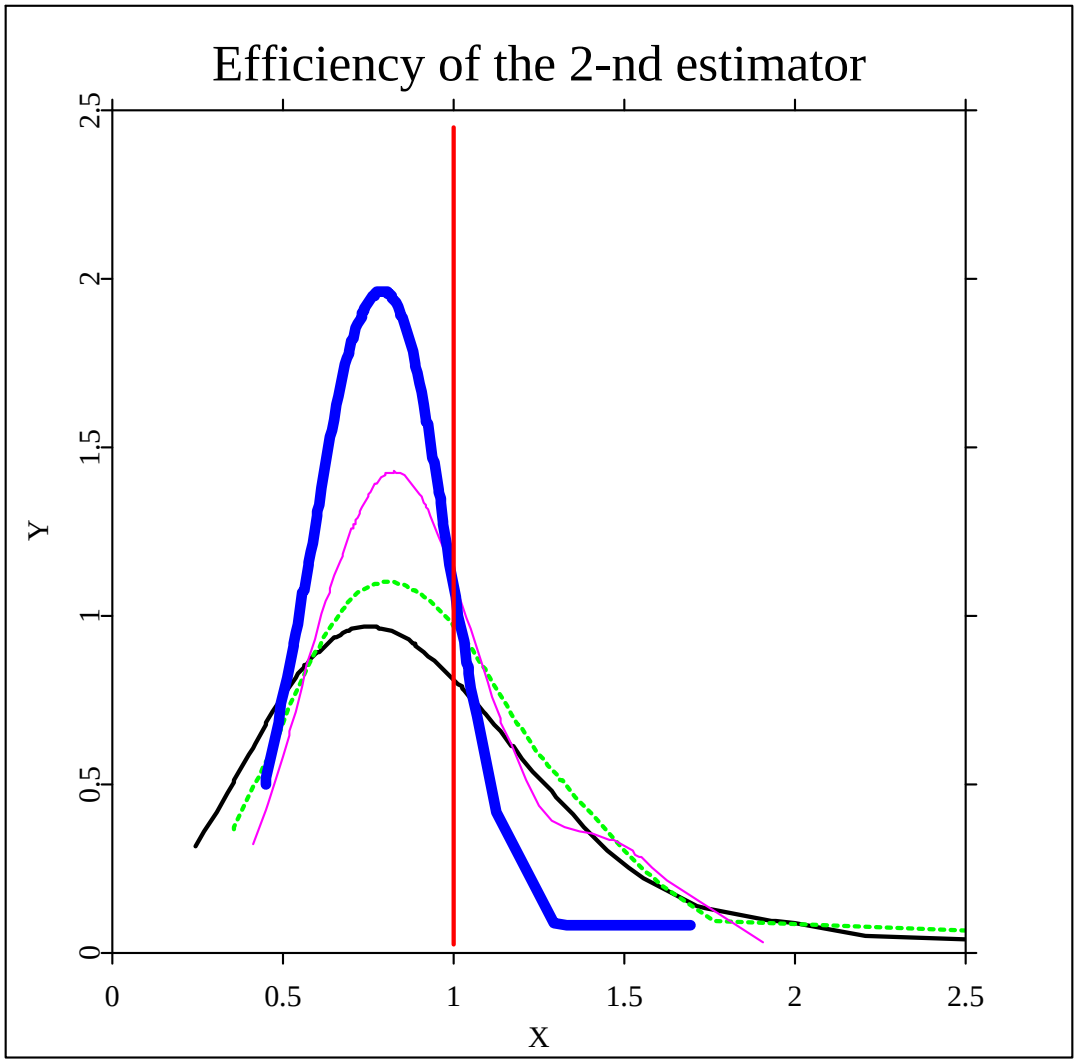


Figure 4.3: Kernel density plots for $\alpha = 2, d_1 = 4, d_2 = 30$ in Chapter 4

Note: kernel density plots of the 100 empirical efficiencies of $\hat{m}_{\mathrm{SBK},\alpha}(x_\alpha)$ to $\tilde{m}_{\mathrm{K},\alpha}(x_\alpha)$, computed according to (4.7) for sample sizes $n = 100$ (solid lines), 200 (dashed lines), 500 (thin lines) and 1000 (thick lines) at $\alpha = 2, d_1 = 4, d_2 = 30$.

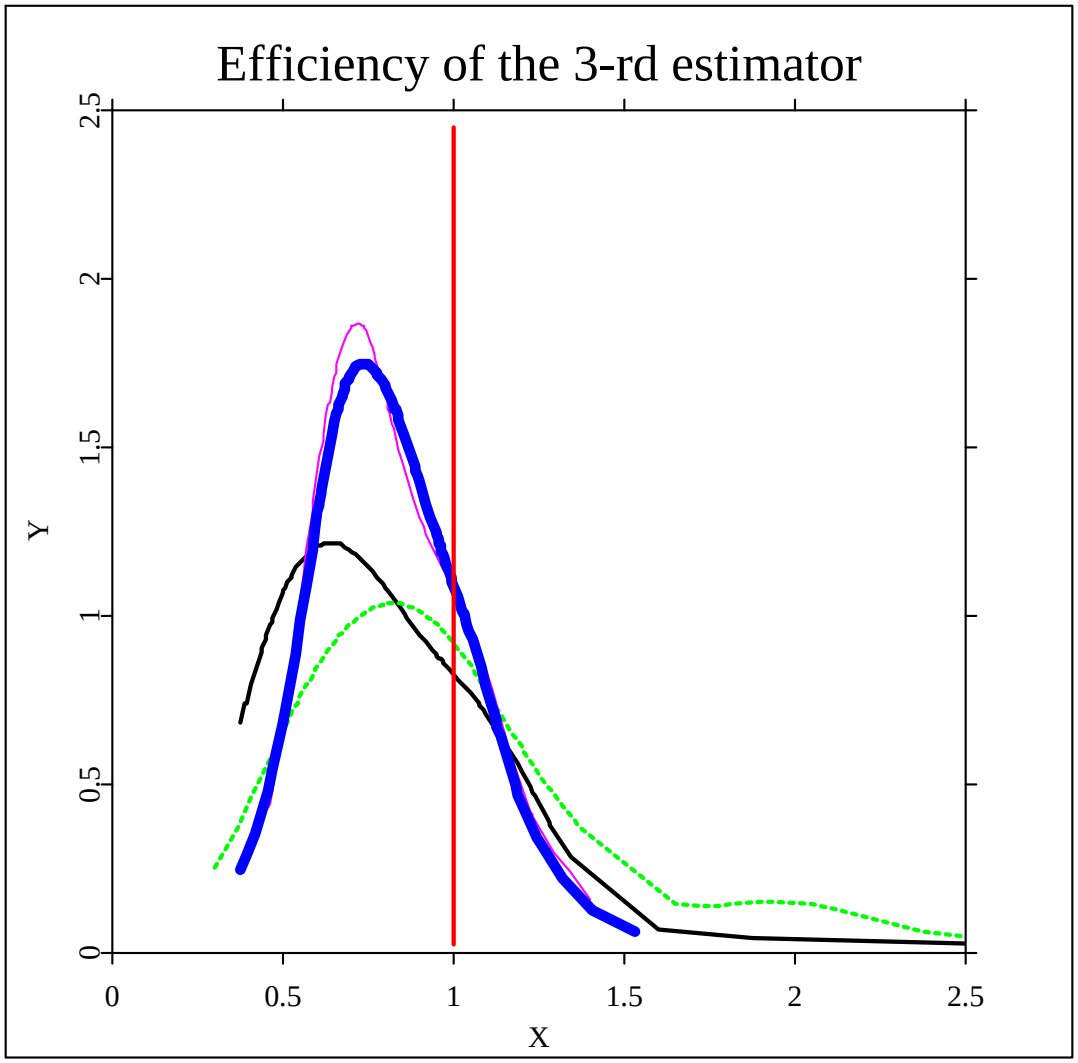


Figure 4.4: Kernel density plots for $\alpha = 3, d_1 = 4, d_2 = 30$ in Chapter 4

Note: kernel density plots of the 100 empirical efficiencies of $\hat{m}_{\mathrm{SBK},\alpha}(x_\alpha)$ to $\tilde{m}_{\mathrm{K},\alpha}(x_\alpha)$, computed according to (4.7) for sample sizes $n = 100$ (solid lines), 200 (dashed lines), 500 (thin lines) and 1000 (thick lines) at $\alpha = 3, d_1 = 4, d_2 = 30$.

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 $\text{eff}_l = \left[\frac{\sum_{t=1}^{100}\left\{\tilde{c}_{0l,t}-c_{0l}\right\}^2/100}{\sum_{t=1}^{100}\left\{\hat{c}_{0l,t}-c_{0l}\right\}^2/100}\right]^{1/2}$, 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 4.1 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.

Table 4.1: Estimation of parameters for the linear part in Chapter 4

| | $r$ | $a$ | 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) |

Note: 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).

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 Figures 4.5, 4.6, 4.7, and 4.8. The SBK estimators seem rather satisfactory and their performance improves with increasing $n$.

## 4.4 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 \$1000's

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 (2009b) 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 (2009b), we fitted a partial linear additive

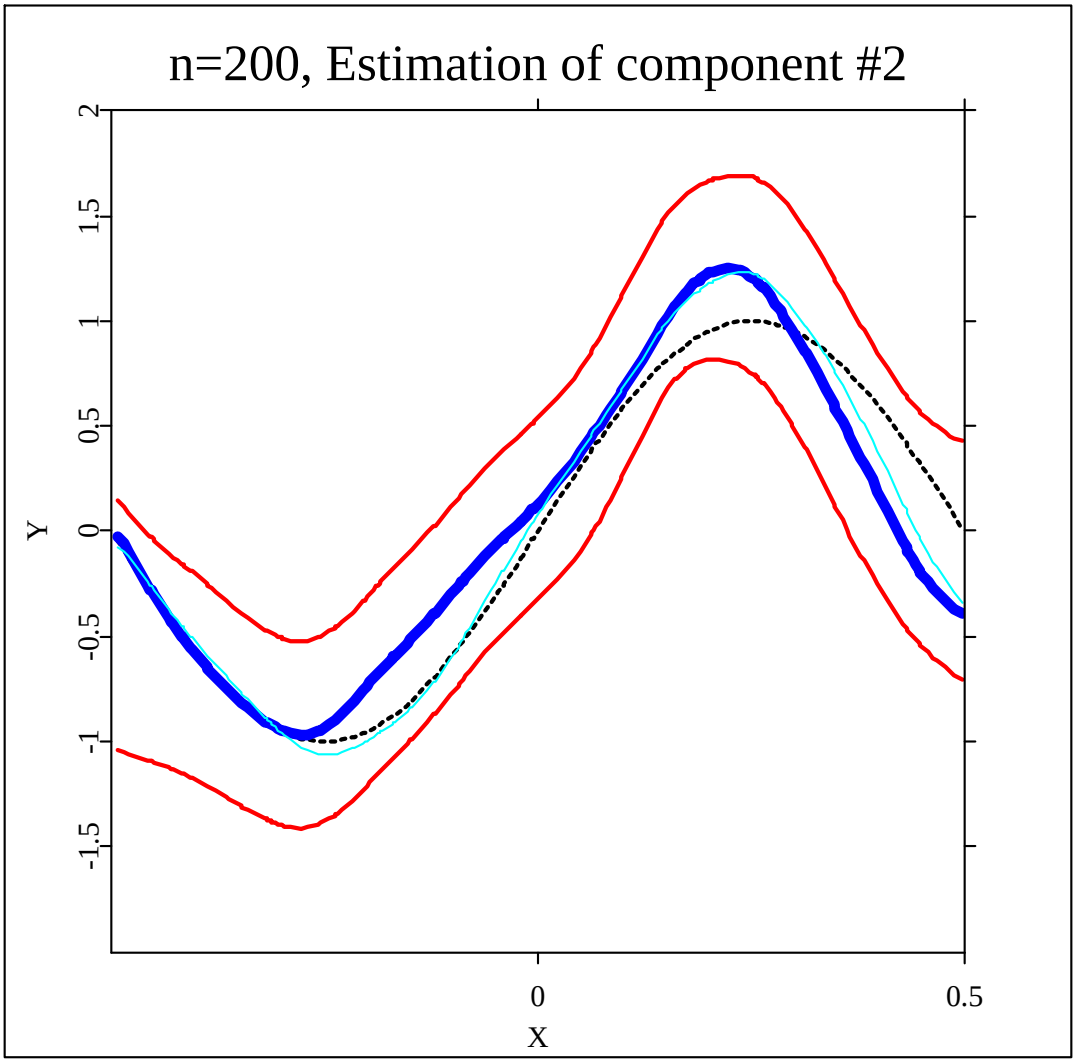


Figure 4.5: Plots of the estimator for the nonparametric part at $\alpha = 2, n = 200$

Note: plots of the oracle smoother $\tilde{m}_{\mathrm{K},\alpha}(x_\alpha)$ (thin curve), the SBK estimator $\hat{m}_{\mathrm{SBK},\alpha}(x_\alpha)$ (thick curve) defined in (4.4), and the 95% pointwise confidence intervals constructed by (4.6) (upper and lower medium curves) of the function components $m_\alpha(x_\alpha)$, $\alpha = 2$ (dashed curve), $d_1 = 4, d_2 = 3$.

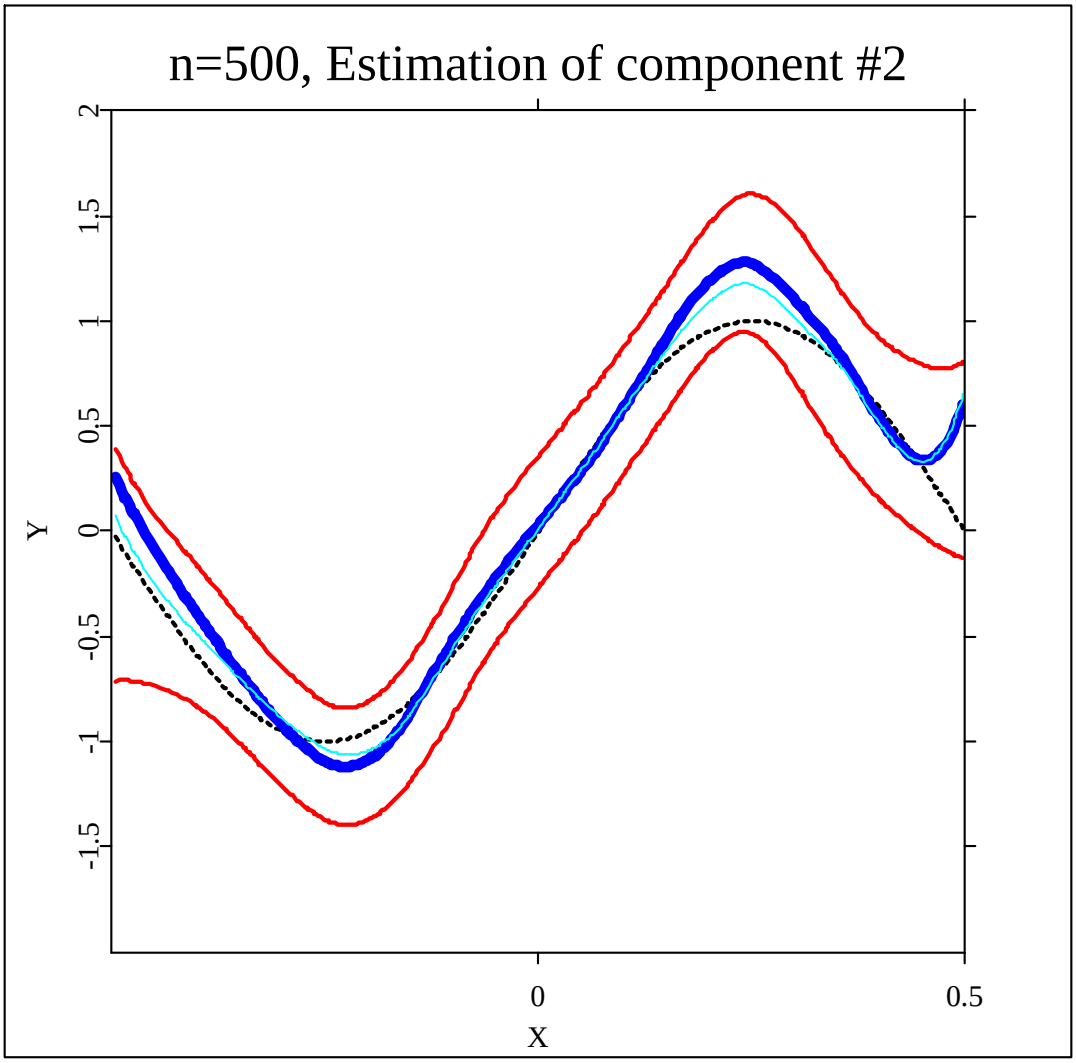


Figure 4.6: Plots of the estimator for the nonparametric part at $\alpha = 2, n = 500$

Note: plots of the oracle smoother $\tilde{m}_{\mathrm{K},\alpha}(x_\alpha)$ (thin curve), the SBK estimator $\hat{m}_{\mathrm{SBK},\alpha}(x_\alpha)$ (thick curve) defined in (4.4), and the 95% pointwise confidence intervals constructed by (4.6) (upper and lower medium curves) of the function components $m_\alpha(x_\alpha)$, $\alpha = 2$ (dashed curve), $d_1 = 4, d_2 = 3$.

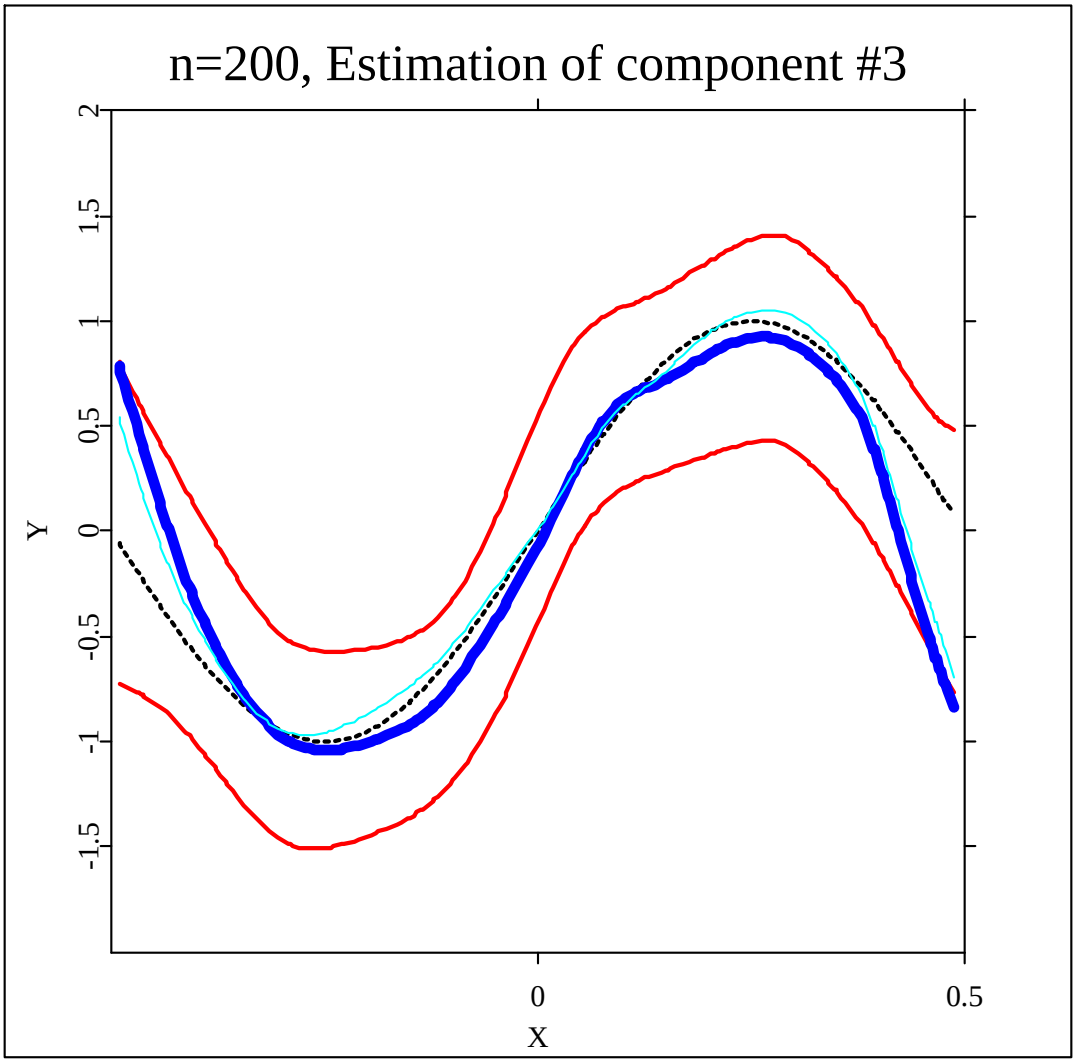


Figure 4.7: Plots of the estimator for the nonparametric part at $\alpha = 3, n = 200$

Note: plots of the oracle smoother $\tilde{m}_{\mathrm{K},\alpha}(x_\alpha)$ (thin curve), the SBK estimator $\hat{m}_{\mathrm{SBK},\alpha}(x_\alpha)$ (thick curve) defined in (4.4), and the 95% pointwise confidence intervals constructed by (4.6) (upper and lower medium curves) of the function components $m_\alpha(x_\alpha)$, $\alpha = 3$ (dashed curve), $d_1 = 4, d_2 = 3$.

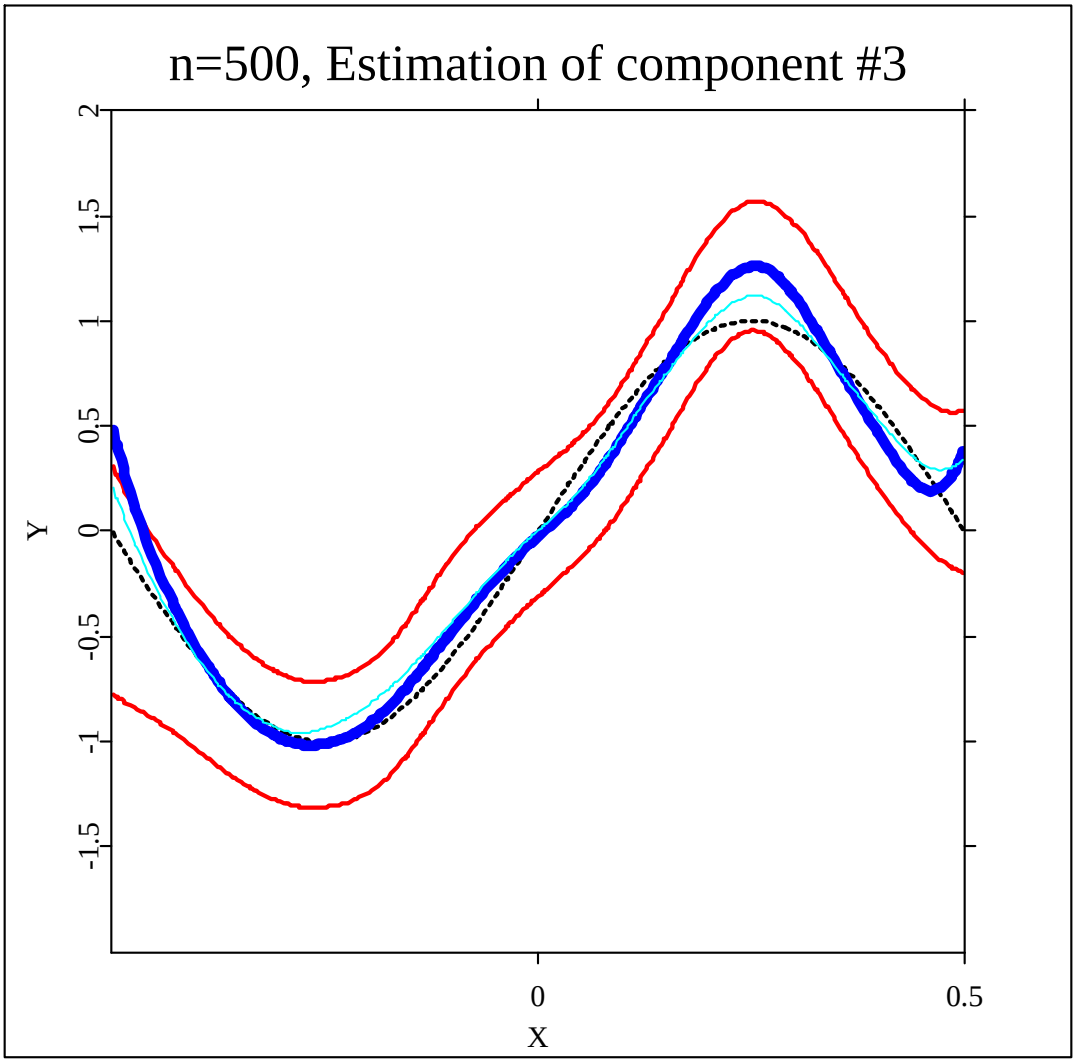


Figure 4.8: Plots of the estimator for the nonparametric part at $\alpha = 3, n = 500$

Note: plots of the oracle smoother $\tilde{m}_{\mathrm{K},\alpha}(x_\alpha)$ (thin curve), the SBK estimator $\hat{m}_{\mathrm{SBK},\alpha}(x_\alpha)$ (thick curve) defined in (4.4), and the 95% pointwise confidence intervals constructed by (4.6) (upper and lower medium curves) of the function components $m_\alpha(x_\alpha)$, $\alpha = 3$ (dashed curve), $d_1 = 4, d_2 = 3$.

model as follows:

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

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

In Figure 4.9, 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 (4.3). The confidence bands are used to test the linearity of the nonparametric components. In Figures 4.9, 4.10 and 4.11 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 (2009b). 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 (2009b).

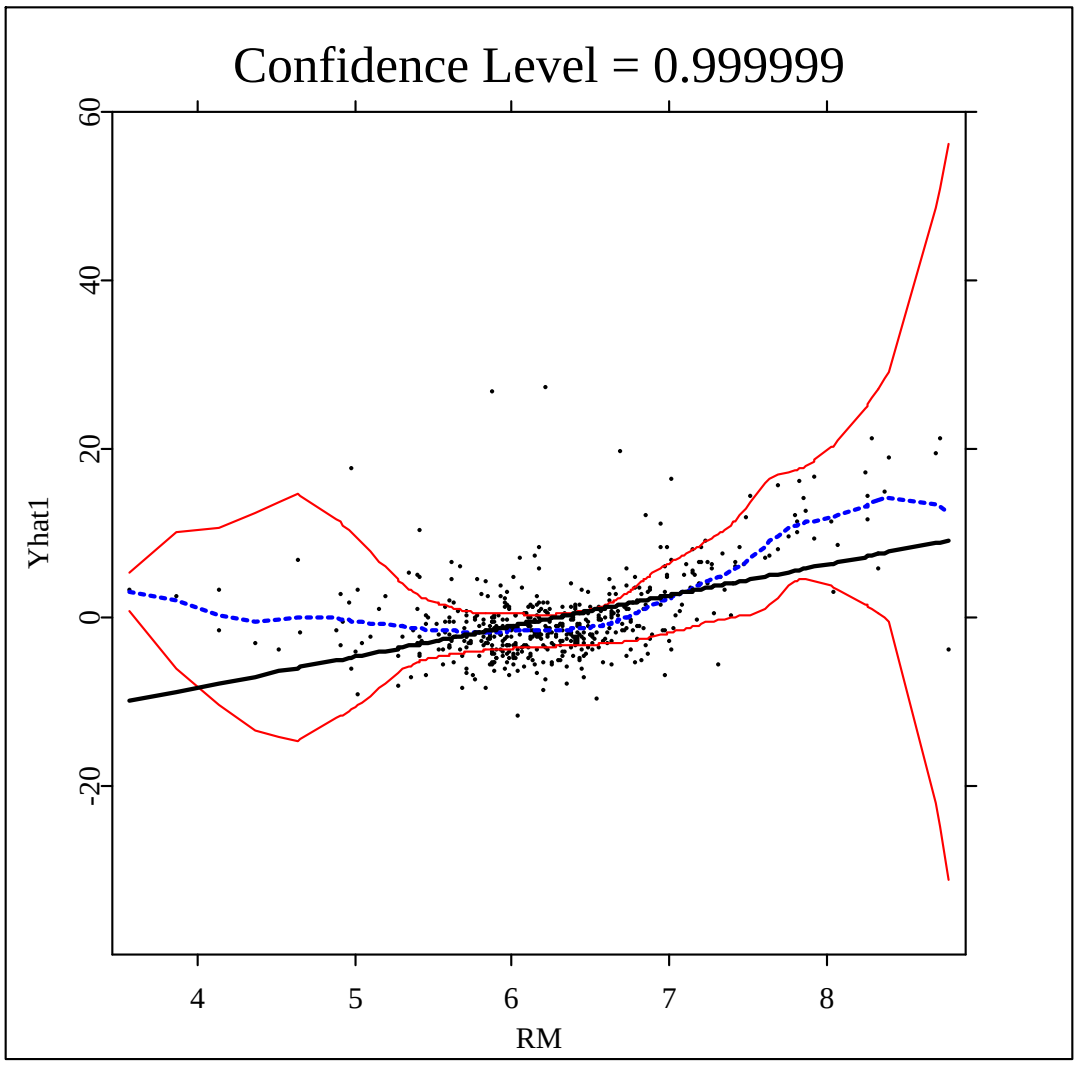


Figure 4.9: Plots of the estimators for RM for Boston housing data

Note: 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).

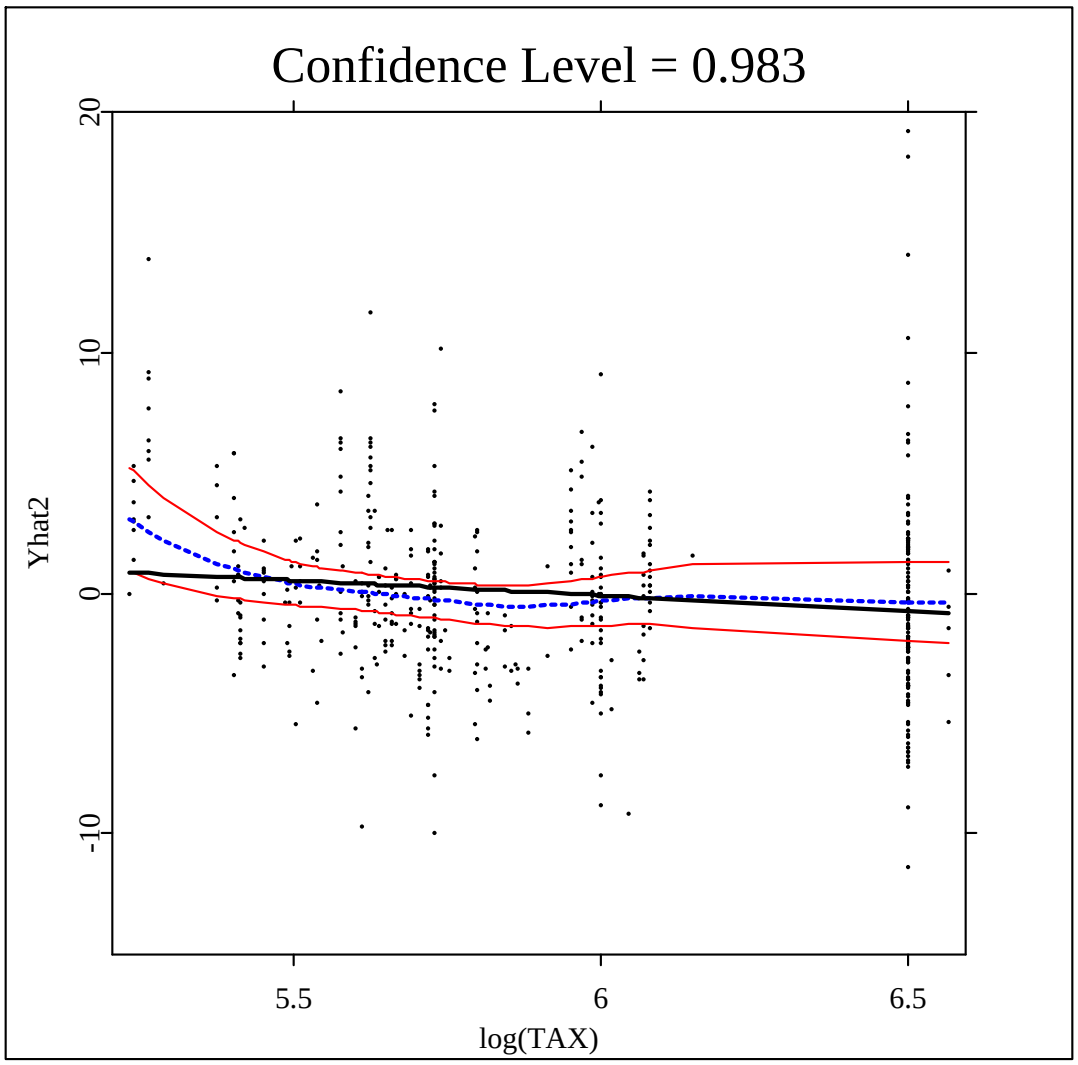


Figure 4.10: Plots of the estimators for log(TAX) for Boston housing data

Note: 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).

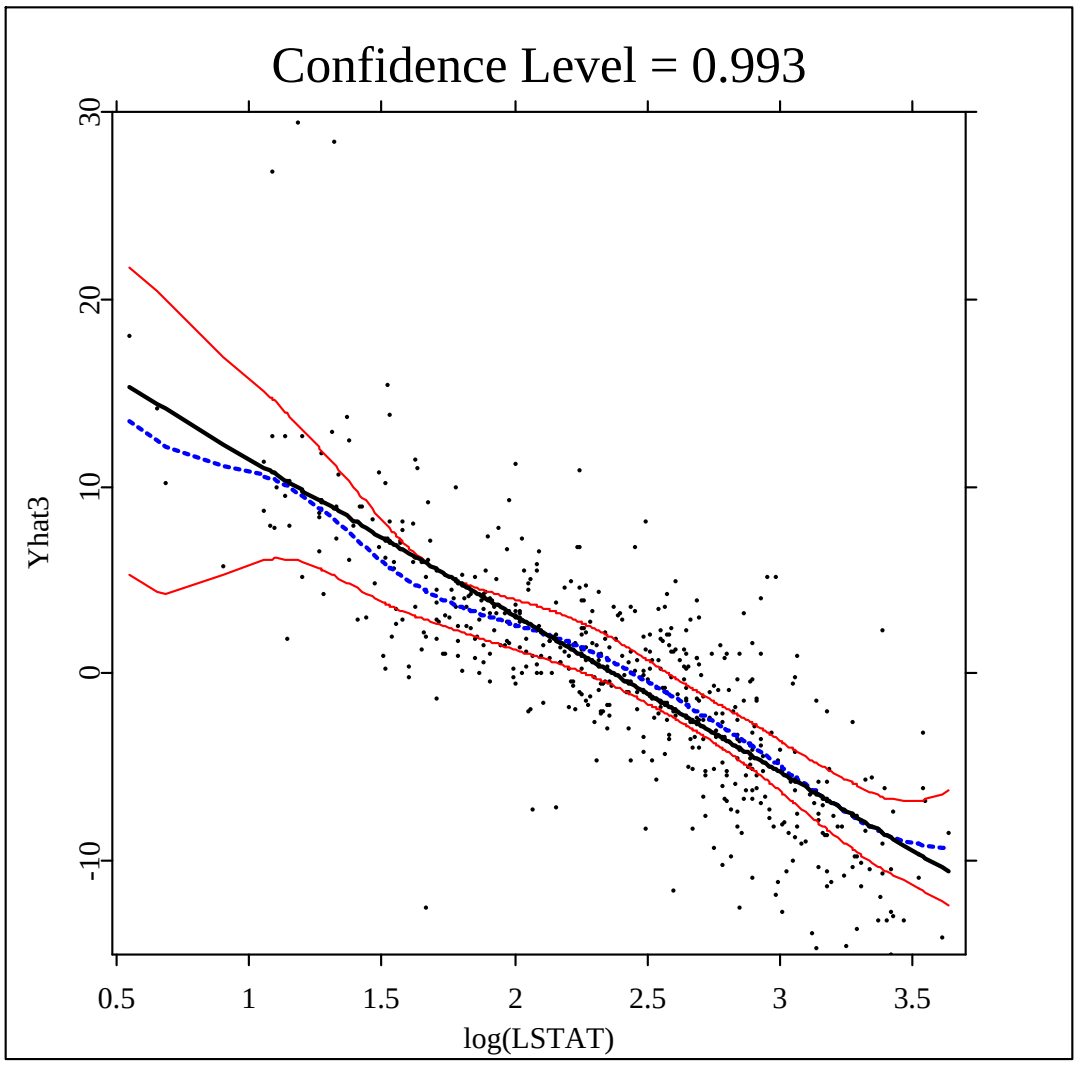


Figure 4.11: Plots of the estimators for log(LSTAT) for Boston housing data

Note: 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).

## 4.5 Appendix

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

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

$$\alpha(k) = \sup_{B\in\sigma\{\mathbf{Z}_t, t\leq 0\}, C\in\sigma\{\mathbf{Z}_t, t\geq k\}} |P(B\cap C) - P(B)P(C)|. \tag{4.8}$$

A3. The noise $\varepsilon_i$ satisfies $E\left(\varepsilon_i \left| \mathcal{F}_i\right.\right) = 0, E\left(\varepsilon_i^2 \left| \mathcal{F}_i\right.\right) = 1$, $E\left(\left|\varepsilon_i\right|^{2+\delta} \left| \mathcal{F}_i\right.\right) < M_\delta < +\infty$ for some $\delta > 2/3, M_\delta > 0$, and $\sigma$-fields $\mathcal{F}_i = \sigma\left\{\left(\mathbf{X}_{i'}, \mathbf{T}_{i'}\right), i' \leq i; \varepsilon_{i'}, i' \leq i-1, 1 \leq i \leq n\right\}$. Function $\sigma(\mathbf{x}, \mathbf{t})$ is continuous with

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

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

A5. There exist constants $0 < c_\delta \leq C_\delta < +\infty, 0 < c_{\mathbf{Q}} \leq C_{\mathbf{Q}} < +\infty$ such that

$c_\delta \leq E\left(\left|T_l\right|^{2+\delta}|\mathbf{X}=\mathbf{x}\right) \leq C_\delta, \forall\, x \in [0,1]^{d2}, 1 \leq l \leq d_1$. and $c_{\mathbf{Q}} I_{(d_1+1)\times(d_1+1)} \leq Q(\mathbf{x}) \leq C_{\mathbf{Q}} I_{(d_1+1)\times(d_1+1)}$, where $\mathbf{Q}(\mathbf{x}) = E\left\{\begin{pmatrix} 1 & \mathbf{T}^T \end{pmatrix}^T \begin{pmatrix} 1 & \mathbf{T}^T \end{pmatrix} |\mathbf{X}=\mathbf{x}\right\}$.

A6. The number of interior knots $N = N_n \sim n^{1/4}\log n$, i.e., $c_N n^{1/4}\log n \leq N \leq C_N n^{1/4}\log n$ 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., $c_h n^{-1/5} \leq h \leq C_h n^{-1/5}$ 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 $[0,1]^{d2} \times \mathbf{R}^{d1}$, 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 $\langle \phi, \varphi \rangle = E\left\{\phi(\mathbf{X}_i, \mathbf{T}_i)\varphi(\mathbf{X}_i, \mathbf{T}_i)\right\}$, $\|\phi\|^2 = E\left\{\phi^2(\mathbf{X}_i, \mathbf{T}_i)\right\}$ 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 $b_{J,\alpha}(x_\alpha) = b_J(x_\alpha) - \frac{c_{J,\alpha}}{c_{J-1,\alpha}} b_{J-1}(x_\alpha), \forall 1 \leq \alpha \leq d_2, 1 \leq J \leq N+1$, where $c_{J,\alpha} = E b_J(X_\alpha) = \int b_J(x_\alpha) f_\alpha(x_\alpha)\, dx_\alpha$. The standardized basis is

$$B_{J,\alpha}(x_\alpha) = b_{J,\alpha}(x_\alpha) / \left\|b_{J,\alpha}\right\|, \forall 1 \leq \alpha \leq d_2, 1 \leq J \leq N+1. \tag{4.9}$$

For the proof of Theorem 4.2, 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 $\left\{ B_{J,\alpha}(x_\alpha), 1 \le \alpha \le d_2, 1 \le J \le N+1 \right\}$. Denote

$$\begin{aligned} \mathrm{Proj}_{\mathcal{H}} T_l &= p_l(\mathbf{X}) = \mathrm{argmin}_{p \in \mathcal{H}} E\left\{T_l - p(\mathbf{X})\right\}^2, \tilde{T}_l = T_l - \mathrm{Proj}_{\mathcal{H}} T_l, \\ \mathrm{Proj}_{\mathcal{H}_n} T_l &= \mathrm{argmin}_{p \in \mathcal{H}_n} E\left\{T_l - p(\mathbf{X})\right\}^2, \tilde{T}_{l,n} = T_l - \mathrm{Proj}_{\mathcal{H}_n} T_l, \end{aligned}$$

for $1 \le l \le d_1$, where $\mathrm{Proj}_{\mathcal{H}} \mathrm{T}_l$ 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

$$\tilde{\mathbf{T}}_n = \left\{ \tilde{T}_{l,n} \right\}_{1 \le l \le d_1}, \tilde{\mathbf{T}} = \left\{ \tilde{T}_l \right\}_{1 \le l \le d_1}. \tag{4.10}$$

The next assumption is needed for the second part of Theorem 4.2.

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

$$\hat{m}(\mathbf{x}, \mathbf{t}) = \hat{c}_{00} + \sum_{l=1}^{d_1} \hat{c}_{0l} t_l + \sum_{\alpha=1}^{d_2} \sum_{J=1}^{N+1} \hat{c}_{J,\alpha} B_{J,\alpha}(x_\alpha), \tag{4.11}$$

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

$$\sum_{i=1}^{n} \left\{ Y_i - c_0 - \sum_{l=1}^{d_1} c_l T_{il} - \sum_{\alpha=1}^{d_2} \sum_{J=1}^{N+1} c_{J,\alpha} B_{J,\alpha}(X_{i\alpha}) \right\}^2. \tag{4.12}$$

While (4.2) is used for statistical implementation, algebraically equivalent (4.11) is for mathematical analysis. Pilot estimators of $m_\alpha(x_\alpha)$ and $\mathbf{c}^T$ are

$$\hat{m}_\alpha(x_\alpha) = \sum_{J=1}^{N+1} \hat{c}_{J,\alpha} B^*_{J,\alpha}(x_\alpha), \hat{\mathbf{c}}^T = \left\{\hat{c}_{00}, \hat{c}_{0l}\right\}_{l=1}^{d_1} \tag{4.13}$$

where $B^*_{J,\alpha}(x_\alpha) \equiv B_{J,\alpha}(x_\alpha) - E_n B_{J,\alpha} = B_{J,\alpha}(x_\alpha) - n^{-1}\sum_{i=1}^n B_{J,\alpha}\left(X_{i\alpha}\right)$ 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 $\left\{\hat{m}\left(\mathbf{X}_i,\mathbf{T}_i\right)\right\}_{1\le i\le n}^T$, 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 $\mathbf{\Lambda} = \left\{\Lambda_i\right\}_{1\le i\le n}^T$, 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 $\left\{\hat{\lambda}_0, \hat{\lambda}_l, \hat{\lambda}_{J,\alpha}\right\}$ given in (4.12). 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 $\left(\mathbf{P}_{n,c}\mathbf{\Lambda}\right)^T = \left\{\hat{\lambda}_0, \hat{\lambda}_l\right\}_{1\le l\le d_1}$. Rewrite estimators $\hat{m}(\mathbf{x})$, $\hat{m}_\alpha(x_\alpha)$, $\hat{\mathbf{c}}$ defined in (4.11) and (4.13) 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

$$\begin{aligned}
\tilde{m}(\mathbf{x},\mathbf{t}) &= \mathbf{P}_n\mathbf{m}(\mathbf{x},\mathbf{t}), \tilde{m}_\alpha(x_\alpha) = \mathbf{P}_{n,\alpha}\mathbf{m}(x_\alpha), \tilde{\mathbf{c}}_m = \mathbf{P}_{n,c}\mathbf{m},\\
\tilde{\varepsilon}(\mathbf{x},\mathbf{t}) &= \mathbf{P}_n\mathbf{E}, \tilde{\varepsilon}_\alpha(x_\alpha) = \mathbf{P}_{n,\alpha}\mathbf{E}(x_\alpha), \tilde{\mathbf{c}}_\varepsilon = \mathbf{P}_{n,c}\mathbf{E},
\end{aligned} \tag{4.14}$$

for $\mathbf{m} = \left\{m\left(\mathbf{X}_i,\mathbf{T}_i\right)\right\}_{i=1}^n, \mathbf{E} = \left\{\sigma\left(\mathbf{X}_i,\mathbf{T}_i\right)\varepsilon_i\right\}_{i=1}^n$. 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

$$\hat{m}(\mathbf{x},\mathbf{t}) = \tilde{m}(\mathbf{x},\mathbf{t}) + \tilde{\varepsilon}(\mathbf{x},\mathbf{t}), \hat{m}_\alpha(x_\alpha) = \tilde{m}_\alpha(x_\alpha) + \tilde{\varepsilon}_\alpha(x_\alpha), \hat{\mathbf{c}} = \tilde{\mathbf{c}}_m + \tilde{\mathbf{c}}_\varepsilon. \tag{4.15}$$

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

$$\sum_{i=1}^n \left\{\sigma\left(\mathbf{X}_i, \mathbf{T}_i\right)\varepsilon_i - a_0 - \sum_{l=1}^{d_1} a_l T_{il} - \sum_{\alpha=1}^{d_2}\sum_{J=1}^{N+1} a_{J,\alpha} B_{J,\alpha}\left(X_{i\alpha}\right)\right\}^2. \tag{4.16}$$

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

$$\sum_{i=1}^n \left\{m\left(\mathbf{X}_i, \mathbf{T}_i\right) - c_0 - \sum_{l=1}^{d_1} c_l T_{il} - \sum_{\alpha=1}^{d_2}\sum_{J=1}^{N+1} c_{J,\alpha} B_{J,\alpha}\left(X_{i\alpha}\right)\right\}^2. \tag{4.17}$$

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

$$\mathbf{B}\left(\mathbf{x}, \mathbf{t}\right) = \left\{1, t_l, B_{J,\alpha}\left(x_\alpha\right)\right\}^T_{1\le l\le d_1, 1\le J\le N+1, 1\le\alpha\le d_2}, \mathbf{B} = \left\{\mathbf{B}\left(\mathbf{X}_i, \mathbf{T}_i\right)\right\}^T_{1\le i\le n}, \tag{4.18}$$

and $\tilde{\mathbf{a}}$, the solution of (4.16), equals to

$$\left\{\begin{array}{ccc} 1 & E_n T_l & E_n B_{J,\alpha} \\ \left(E_n T_{l'}\right)^T & \left\langle T_{l'}, T_l\right\rangle_n & \left\langle T_{l'}, B_{J,\alpha}\right\rangle_n \\ \left(E_n B_{J',\alpha'}\right)^T & \left\langle B_{J',\alpha'}, T_l\right\rangle_n & \left\langle B_{J',\alpha'}, B_{J,\alpha}\right\rangle_n \end{array}\right\}^{-1}_{\substack{1\le l,l'\le d_1, 1\le\alpha,\alpha'\le d_2,\\ 1\le J,J'\le N+1}}$$
$$\times \left\{\begin{array}{c} n^{-1}\sum_{i=1}^n \sigma\left(\mathbf{X}_i, \mathbf{T}_i\right)\varepsilon_i \\ n^{-1}\sum_{i=1}^n T_{il}\sigma\left(\mathbf{X}_i, \mathbf{T}_i\right)\varepsilon_i \\ n^{-1}\sum_{i=1}^n B_{J,\alpha}\left(X_{i\alpha}\right)\sigma\left(\mathbf{X}_i, \mathbf{T}_i\right)\varepsilon_i \end{array}\right\}_{1\le l\le d_1, 1\le J\le N+1, 1\le\alpha\le d_2} \tag{4.19}$$

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

LEMMA 4.1. *[Theorem 1.4, page 31 of Bosq (1998)] 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* $t = 1, ..., n, k = 3, 4, ..., E|\xi_t|^k \leq c^{k-2} k! E\xi_t^2 < +\infty$ *(Cramér's condition) then for* $n > 1$*, integer* $q \in [1, n/2]$*,* $\varepsilon > 0$ *and* $k \geq 3$

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

*where* $\alpha(\cdot)$ *is the* $\alpha$*-mixing coefficient defined in (4.8) and*

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

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

LEMMA 4.2. *Under Assumptions A2, A4, A6,* $c_f H/2 \leq c_{J,\alpha} \leq C_f H$ *and (i)* $\exists$ *constants* $c_0(f), C_0(f)$ *depending on* $f_\alpha(x_\alpha), 1 \leq \alpha \leq d_2$*, such that* $c_0(f) H \leq \left\|b_{J,\alpha}\right\|^2 \leq C_0(f)H$*. (ii) uniformly for* $1 \leq J, J' \leq N+1, 1 \leq \alpha, \alpha' \leq d_2, E\left|B_{J,\alpha}(X_\alpha)\right| \leq CH^{1/2}$*,* $E\left|B_{J,\alpha}(X_{i\alpha})\, B_{J',\alpha'}\left(X_{i\alpha'}\right)\right|^2 \geq c_f C_f^{-2} > 0, E\left|B_{J,\alpha}(X_{i\alpha})\, B_{J',\alpha'}\left(X_{i\alpha'}\right)\right|^k \leq C^k H^{2-k}$*,* $k \geq 1$*. (iii) uniformly for* $1 \leq J, J' \leq N+1, 1 \leq \alpha \leq d_2$*,*

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

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

LEMMA 4.3. *Under Assumptions A2, A4, A6 and A7, denote*

$$\omega_{J,\alpha}\left(\mathbf{X}_l, x_1\right) = K_h\left(X_{l1} - x_1\right) B_{J,\alpha}\left(X_{l\alpha}\right), \mu_{J,\alpha}\left(x_1\right) = E\omega_{J,\alpha}\left(\mathbf{X}_l, x_1\right), \tag{4.20}$$

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

LEMMA 4.4. *Under Assumptions A2, A4, A6 and A7, as* $n \to \infty$,

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

*where* $\omega_{J,\alpha}\left(\mathbf{X}_l, x_1\right)$ *and* $\mu_{J,\alpha}\left(x_1\right)$ *are given in (4.20).*

LEMMA 4.5. *Under Assumptions A2, A4-A6, as* $n \to \infty$,

$$A_{n,1} = \sup_{J,\alpha} \left|E_n B_{J,\alpha}\right| = O_p\left(n^{-1/2} \log n\right), \tag{4.21}$$

$$A_{n,2} = \sup_{J,J',\alpha} \left|\left\langle B_{J,\alpha}, B_{J',\alpha}\right\rangle_n - \left\langle B_{J,\alpha}, B_{J',\alpha}\right\rangle\right| = O_p\left\{(nH)^{-1/2} \log n\right\}, \tag{4.22}$$

$$A_{n,3} = \sup_{\alpha \ne \alpha'} \left|\left\langle B_{J,\alpha}, B_{J',\alpha'}\right\rangle_n - \left\langle B_{J,\alpha}, B_{J',\alpha'}\right\rangle\right| = O_p\left(n^{-1/2} \log n\right), \tag{4.23}$$

$$A_{n,4} = \sup_{l,J,\alpha} \left|\left\langle T_l, B_{J,\alpha}\right\rangle_n - \left\langle T_l, B_{J,\alpha}\right\rangle\right| = O_p\left(n^{-1/2} \log n\right). \tag{4.24}$$

LEMMA 4.6. *Under Assumptions A2, A4-A7, denote*

$$\zeta_l\left(X_{i1}, T_{il}, x_1\right) = K_h\left(X_{i1} - x_1\right) T_{il},\ \mu_l\left(x_1\right) = E\zeta_l\left(X_{i1}, T_{il}, x_1\right), \tag{4.25}$$

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

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

For proofs of Lemmas 4.2-4.6, see Ma and Yang (2011b). Let $\left(v_{(J',\alpha'),(J,\alpha)}\right) = \left\langle B_{J',\alpha'}, B_{J,\alpha}\right\rangle, \left(v_{l',l}\right) = \left\langle T_{l'}, T_l\right\rangle, \left(v_{l',(J,\alpha)}\right) = \left\langle T_{l'}, B_{J,\alpha}\right\rangle, \left(v_{(J',\alpha'),l}\right) = \left\langle B_{J',\alpha'}, T_l\right\rangle$. Denote by $\mathbf{V}$ the theoretical inner product matrix of the standardized basis

$$\left\{1, t_l, B_{J,\alpha}(x_\alpha), 1 \le l \le d_1, 1 \le J \le N+1, 1 \le \alpha \le d_2\right\}, \textit{i.e.}$$

$$\mathbf{V} = \begin{pmatrix} 1 & \mathbf{0}_{d_1}^T & \mathbf{0}_{d_2(N+1)}^T \\ \mathbf{0}_{d_1} & \left(v_{l',l}\right) & \left(v_{l',(J,\alpha)}\right) \\ \mathbf{0}_{d_2(N+1)} & \left(v_{(J',\alpha'),l}\right) & \left(v_{(J',\alpha'),(J,\alpha)}\right) \end{pmatrix}_{\substack{1\le l,l'\le d_1, 1\le\alpha\le\alpha'\le d_2, \\ 1\le J,J'\le N+1}}. \quad (4.26)$$

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

$$\mathbf{S} = \begin{pmatrix} 1 & \mathbf{0}_{d_1}^T & \mathbf{0}_{d_2(N+1)}^T \\ \mathbf{0}_{d_1} & \left(s_{l',l}\right) & \left(s_{l',(J,\alpha)}\right) \\ \mathbf{0}_{d_2(N+1)} & \left(s_{(J',\alpha'),l}\right) & \left(s_{(J',\alpha'),(J,\alpha)}\right) \end{pmatrix}_{\substack{1\le l,l'\le d_1, 1\le\alpha\le\alpha'\le d_2, \\ 1\le J,J'\le N+1}}. \quad (4.27)$$

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

$$\hat{\mathbf{V}} = \begin{pmatrix} 0 & E_n T_l & E_n B_{J,\alpha} \\ \left(E_n T_{l'}\right)^T & \left\langle T_{l'}, T_l\right\rangle_n & \left\langle T_{l'}, B_{J,\alpha}\right\rangle_n \\ \left(E_n B_{J',\alpha'}\right)^T & \left\langle B_{J',\alpha'}, T_l\right\rangle_n & \left\langle B_{J',\alpha'}, B_{J,\alpha}\right\rangle_n \end{pmatrix}_{\substack{1\le l,l'\le d_1, 1\le\alpha\le\alpha'\le d_2, \\ 1\le J,J'\le N+1}}. \quad (4.28)$$

LEMMA 4.7. *Under Assumptions A2, A4-A7, for matrices* $\mathbf{V}$, $\mathbf{S}$ *and* $\hat{\mathbf{V}}$ *defined in (4.26), (4.27), and (4.28) (i) there exist constants* $C_V > c_V > 0$, $C_S = c_V^{-1}, c_S = C_V^{-1}$ *such that*

$$\begin{aligned} c_V \mathbf{I}_{1+d_1+d_2(N+1)} &\leq \mathbf{V} \leq C_V \mathbf{I}_{1+d_1+d_2(N+1)}, \\ c_S \mathbf{I}_{1+d_1+d_2(N+1)} &\leq \mathbf{S} \leq C_S \mathbf{I}_{1+d_1+d_2(N+1)}. \end{aligned} \tag{4.29}$$

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

$$\begin{aligned} c_V \mathbf{I}_{1+d_1+d_2(N+1)} &\leq \hat{\mathbf{V}} \leq C_V \mathbf{I}_{1+d_1+d_2(N+1)}, \\ c_S \mathbf{I}_{1+d_1+d_2(N+1)} &\leq \hat{\mathbf{V}}^{-1} \leq C_S \mathbf{I}_{1+d_1+d_2(N+1)}. \end{aligned} \tag{4.30}$$

LEMMA 4.8. *Under Assumptions A2-A7, as* $n \to \infty$,

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

For proofs of Lemmas 4.7 and 4.8, see Ma and Yang (2011b).

COROLLARY 4.2. *Under Assumptions A2-A7, as* $n \to \infty$,
$\left\| n^{-1} \mathbf{B}^T \mathbf{E} \right\| = O_p(n^{-1/2} N^{1/2} \log n)$, $\left\| n^{-1} \mathbf{B}^T \mathbf{E} \right\|_\infty = O_p(n^{-1/2} \log n)$.

Corollary 4.2 follows from Lemma 4.8 directly.

Next we study the difference between $\hat{m}_{\mathrm{SBK}}(x_1)$ and $\tilde{m}_{\mathrm{K},1}(x_1)$, both given in (4.4). Denote $\mathbf{c} = \{c_{0l}\}_{l=0}^{d_1}$, the decomposition (4.15) implies that $\tilde{m}_{\mathrm{K},1}(x_1) - \hat{m}_{\mathrm{SBK}}(x_1) =$

$\left\{\Psi_{Tb}(x_1)+\Psi_{Tv}(x_1)+\Psi_b(x_1)+\Psi_v(x_1)\right\}/\hat{f}_1(x_1)$, where

$$\Psi_{Tb}(x_1)=n^{-1}\sum_{i=1}^{n}K_h\left(X_{i1}-x_1\right)\left(1,\mathbf{T}_i^T\right)\left(\tilde{\mathbf{c}}_m-\mathbf{c}\right), \tag{4.31}$$

$$\Psi_{Tv}(x_1)=n^{-1}\sum_{i=1}^{n}K_h\left(X_{i1}-x_1\right)\left(1,\mathbf{T}_i^T\right)\tilde{\mathbf{c}}_\varepsilon,$$

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

$$\Psi_v(x_1)=n^{-1}\sum_{i=1}^{n}K_h\left(X_{i1}-x_1\right)\sum_{\alpha=2}^{d_2}\tilde{\varepsilon}_\alpha\left(X_{i\alpha}\right). \tag{4.32}$$

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

LEMMA 4.9. *Under Assumptions A1, A2, A4-A7, as $n\to\infty$,*

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

LEMMA 4.10. *Under Assumptions A1, A2 and A6, there exist functions $g_\alpha\in G, 1\le\alpha\le d_2$, such that as $n\to\infty$, $\left\|\tilde{m}-g+\sum_{\alpha=1}^{d_2}E_n g_\alpha(X_\alpha)\right\|_n=O_p\left(n^{-1/2}+H^2\right)$, 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 (4.14).*

For proofs of Lemma 4.10 and Lemma 4.9, see Ma and Yang (2011b). Next we show $\Psi_v(x_1)$ in (4.32) is uniformly of order $o_p\left(n^{-2/5}\right)$. For $\tilde{a}_{J,\alpha}$ given in (4.19), define an auxiliary entity

$$\tilde{\varepsilon}_\alpha^*=\sum_{J=1}^{N+1}\tilde{a}_{J,\alpha}B_{J,\alpha}(x_\alpha), \tag{4.33}$$

The $\tilde{\varepsilon}_\alpha(x_\alpha)$ in (4.14) is the empirical centering of $\tilde{\varepsilon}_2^*(x_2)$, i.e.

$$\tilde{\varepsilon}_\alpha(x_\alpha)\equiv\tilde{\varepsilon}_\alpha^*(x_\alpha)-n^{-1}\sum_{i=1}^{n}\tilde{\varepsilon}_\alpha^*\left(X_{i\alpha}\right). \tag{4.34}$$

According to (4.34), we can write $\Psi_v(x_1) = \Psi_v^{(2)}(x_1) - \Psi_v^{(1)}(x_1)$, where

$$\Psi_v^{(1)}(x_1) = n^{-1}\sum_{l=1}^{n} K_h(X_{l1} - x_1)\sum_{\alpha=2}^{d_2} n^{-1}\sum_{i=1}^{n}\tilde{\varepsilon}_\alpha^*(X_{i\alpha}), \tag{4.35}$$

$$\Psi_v^{(2)}(x_1) = n^{-1}\sum_{l=1}^{n} K_h(X_{l1} - x_1)\sum_{\alpha=2}^{d_2}\tilde{\varepsilon}_\alpha^*(X_{l\alpha}), \tag{4.36}$$

for $\tilde{\varepsilon}_\alpha^*(X_{i\alpha})$ in (4.33). By (4.19) and (4.33), $\Psi_v^{(2)}(x_1)$ can be rewritten as

$$\Psi_v^{(2)}(x_1) = n^{-1}\sum_{\alpha=2}^{d_2}\sum_{l=1}^{n}\sum_{J=1}^{N+1}\tilde{a}_{J,\alpha}\omega_J(\mathbf{X}_l, x_1), \tag{4.37}$$

for $\omega_{J,\alpha}(\mathbf{X}_l, x_1)$ given in (4.20). $\Psi_v^{(1)}(x_1)$ and $\Psi_v^{(2)}(x_1)$ are of order $O_p\left\{Nn^{-1}(\log n)^2\right\}$ and $O_p\left(n^{-1/2}\log n\right)$ uniformly, given in Lemmas 4.12 and 4.13. 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 (4.16).

LEMMA 4.11. *Under Assumptions A2-A6, as $n \to \infty$,*

$$\tilde{\mathbf{a}}^T\tilde{\mathbf{a}} = \tilde{a}_0^2 + \sum_{l=1}^{d_1}\tilde{a}_l^2 + \sum_{\alpha=1}^{d_2}\sum_{J=1}^{N+1}\tilde{a}_{J,\alpha}^2 = O_p\left\{Nn^{-1}(\log n)^2\right\}. \tag{4.38}$$

*Proof.* By (4.18), (4.19), (4.30), with probability approaching 1 as $n \to \infty$, $\|\tilde{\mathbf{a}}\|\left\|n^{-1}\mathbf{B}^T\mathbf{E}\right\| \geq \tilde{\mathbf{a}}^T\left(n^{-1}\mathbf{B}^T\mathbf{E}\right) = \tilde{\mathbf{a}}^T\hat{\mathbf{V}}\tilde{\mathbf{a}} \geq c_V\|\tilde{\mathbf{a}}\|^2$ with Corollary 4.2 imply $\|\tilde{\mathbf{a}}\|^2 \leq c_V^{-1}\|\tilde{\mathbf{a}}\|\left\|n^{-1}\mathbf{B}^T\mathbf{E}\right\| = c_V^{-1}\|\tilde{\mathbf{a}}\| \times O_p\left\{N^{1/2}n^{-1/2}\log n\right\}$. Thus $\|\tilde{\mathbf{a}}\| = O_p\left\{N^{1/2}n^{-1/2}\log n\right\}$, $n \to \infty$. ∎

LEMMA 4.12. *Under Assumptions A2-A7, as $n \to \infty$, $\Psi_v^{(1)}(x_1)$ in (4.35) satisfies* $\sup_{x_1\in[0,1]}\left|\Psi_v^{(1)}(x_1)\right| = O_p\left\{Nn^{-1}(\log n)^2\right\}$.

For proof, see Ma and Yang (2011b). The vector $\tilde{\mathbf{a}}$ in (4.19) is

$$\tilde{\mathbf{a}} = \left(\hat{\mathbf{V}}\right)^{-1}\left(n^{-1}\mathbf{B}^T\mathbf{E}\right). \tag{4.39}$$

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 (4.37) as

$$\hat{\mathbf{a}} = \mathbf{V}^{-1}\left(n^{-1}\mathbf{B}^T\mathbf{E}\right) = \mathbf{S}\left(n^{-1}\mathbf{B}^T\mathbf{E}\right), \tag{4.40}$$

$$\hat{\Psi}_v^{(2)}(x_1) = n^{-1}\sum_{i=1}^{n}\sum_{\alpha=1}^{d_2}\sum_{J=1}^{N+1}\hat{a}_{J,\alpha}\omega_{J,\alpha}\left(\mathbf{X}_i, x_1\right). \tag{4.41}$$

LEMMA 4.13. *Under Assumptions A2-A7, as $n \to \infty$, $\Psi_v^{(2)}(x_1)$ in (4.36) satisfies* $\sup_{x_1\in[0,1]}\left|\Psi_v^{(2)}(x_1)\right| = O_p\left(n^{-1/2}\log n\right) = o_p\left(n^{-2/5}\right)$.

*Proof.* According to (4.39) and (4.40), one has $\mathbf{V}\hat{\mathbf{a}} = \left(\hat{\mathbf{V}}\right)\tilde{\mathbf{a}}$, which implies $\left(\hat{\mathbf{V}} - \mathbf{V}\right)\tilde{\mathbf{a}} = \mathbf{V}\left(\hat{\mathbf{a}} - \tilde{\mathbf{a}}\right)$. Using (4.21), (4.22), (4.23), (4.24) one obtains that $\|\mathbf{V}\left(\hat{\mathbf{a}} - \tilde{\mathbf{a}}\right)\| = \left\|\left(\hat{\mathbf{V}} - \mathbf{V}\right)\tilde{\mathbf{a}}\right\|$ $\leq O_p\left(n^{-1/2}H^{-1}\log n\right)\|\tilde{\mathbf{a}}\|$. According to Lemma 4.11, $\|\tilde{\mathbf{a}}\| = O_p\left(n^{-1/2}N^{1/2}\log n\right)$, so as $n \to \infty$, $\|\mathbf{V}\left(\hat{\mathbf{a}} - \tilde{\mathbf{a}}\right)\| \leq O_p\left\{n^{-1}N^{3/2}(\log n)^2\right\}$. By Lemmas 4.7 and 4.11, as $n \to \infty$,

$$\|\hat{\mathbf{a}} - \tilde{\mathbf{a}}\| = O_p\left\{n^{-1}N^{3/2}(\log n)^2\right\}, \tag{4.42}$$

$$\|\hat{\mathbf{a}}\| \leq \|\hat{\mathbf{a}} - \tilde{\mathbf{a}}\| + \|\tilde{\mathbf{a}}\| = O_p\left(n^{-1/2}N^{1/2}\log n\right). \tag{4.43}$$

$$\sup_{x_1\in[0,1]}\left|\Psi_v^{(2)}(x_1) - \hat{\Psi}_v^{(2)}(x_1)\right| \leq \sup_{x_1\in[0,1]}\left|\sum_{\alpha=2}^{d_2}\sum_{J=1}^{N+1}\left(\tilde{a}_{J,\alpha} - \hat{a}_{J,\alpha}\right)n^{-1}\sum_{l=1}^{n}\omega_J\left(\mathbf{X}_l, x_1\right)\right|$$

$$= \sqrt{d_2(N+1)}O_p\left(n^{-1}H^{-3/2}\log^2 n\right)O_p\left(H^{1/2}\right) = O_p\left(n^{-1}H^{-3/2}\log^2 n\right). \tag{4.44}$$

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

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

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 $\sup_{x_1 \in [0,1]} Q_1(x_1) \leq T_1 + T_2$, where $T_1 = \max_{1 \leq k \leq M_n} \left| \sum_{\alpha=2}^{d_2} \sum_{J=1}^{N+1} \hat{a}_{J,\alpha} \mu_{J,\alpha}\left(x_{1,k}\right) \right|$, $T_2$ equals to

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

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

$$\sum_{J=1}^{N+1} \hat{a}_{J,\alpha} \mu_{J,\alpha}\left(x_{1,k}\right) = n^{-1} \sum_{i,l,J} \mu_{J,\alpha}\left(x_{1,k}\right) s_{(J,\alpha),l} T_{il} \sigma\left(\mathbf{X}_i, \mathbf{T}_i\right) \varepsilon_i$$

$$+ n^{-1} \sum_{i,\alpha',J,J'} \mu_{J,\alpha}\left(x_{1,k}\right) s_{(J,\alpha),\left(J',\alpha'\right)} B_{J',\alpha'}\left(X_{i\alpha'}\right) \sigma\left(\mathbf{X}_i, \mathbf{T}_i\right) \varepsilon_i.$$

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

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

$$= O_p\left(n^{-1/2} \log n\right) O_p(1) = O_p\left(n^{-1/2} \log n\right),$$

which follows from Lemma 4.8 and Lemma 4.3. Let $D_n = n^{\theta} 0, (2+\delta)^{-1} < \theta_0 < 3/8$, where $\delta$ is the same as in Assumption A3. Define

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

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

Denote $W^D_{\alpha,\alpha'} = \max\limits_{1\leq k\leq M_n}\left|n^{-1}\sum_{i=1}^n U_{i,k}\right|$ as truncated centered version of $W_{\alpha,\alpha'}$. To show $\left|W_{\alpha,\alpha'} - W^D_{\alpha,\alpha'}\right| = Up\left(n^{-1/2}\log n\right)$, note $\left|W_{\alpha,\alpha'} - W^D_{\alpha,\alpha'}\right| \leq \Lambda_1 + \Lambda_2$,

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

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

Let $\boldsymbol{\mu}_\alpha\left(x_{1,k}\right) = \left\{\mu_{1,\alpha}\left(x_{1,k}\right),\cdots,\mu_{N+1,\alpha}\left(x_{1,k}\right)\right\}^T$, then

$$\begin{aligned}\Lambda_1 &= \max_{1\leq k\leq M_n}\left[\boldsymbol{\mu}_\alpha\left(x_{1,k}\right)^T\left\{s_{(J,\alpha),\left(J',\alpha'\right)}\right\}_{J,J'}\right.\\ &\quad\left.\left\{n^{-1}\sum_{i=1}^n B_{J',\alpha'}\left(X_{i\alpha'}\right)\sigma\left(\mathbf{X}_i,\mathbf{T}_i\right) E\left(\varepsilon^-_{i,D}\left|\mathbf{X}_i,\mathbf{T}_i\right.\right)\right\}_{J'}\right]\end{aligned}$$

$$\begin{aligned}&\leq \max_{1\leq k\leq M_n}\left[\boldsymbol{\mu}_\alpha\left(x_{1,k}\right)^T\boldsymbol{\mu}_\alpha\left(x_{1,k}\right)\left[\left\{s_{(J,\alpha),\left(J',\alpha'\right)}\right\}_{1\leq J,J'\leq N+1}\times\right.\right.\\ &\left\{n^{-1}\sum_{i=1}^n B_{J',\alpha'}\left(X_{i\alpha'}\right)\sigma\left(\mathbf{X}_i,\mathbf{T}_i\right) E\left(\varepsilon^-_{i,D}\left|\mathbf{X}_i,\mathbf{T}_i\right.\right)\right\}_{J'}\Big]^T\left[\left\{s_{(J,\alpha),\left(J',\alpha'\right)}\right\}_{J,J'}\right.\\ &\left.\left.\times\left\{n^{-1}\sum_{i=1}^n B_{J',\alpha'}\left(X_{i\alpha'}\right)\sigma\left(\mathbf{X}_i,\mathbf{T}_i\right) E\left(\varepsilon^-_{i,D}\left|\mathbf{X}_i,\mathbf{T}_i\right.\right)\right\}_{J'=1}^{N+1}\right]\right]^{1/2}\leq\end{aligned}$$

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

By Assumption A3, $\left|E\left(\varepsilon^-_{i,D}\left|\mathbf{X}_i,\mathbf{T}_i\right.\right)\right| = \left|E\left(\varepsilon^+_{i,D}\left|\mathbf{X}_i,\mathbf{T}_i\right.\right)\right| \leq M_\delta D_n^{-(1+\delta)}$. By Lemma

4.1, $\sup_{J',\alpha'} \left| n^{-1} \sum_{i=1}^{n} B_{J',\alpha'} \left( X_{i\alpha'} \right) \sigma \left( \mathbf{X}_i, \mathbf{T}_i \right) \right| = O_p \left( n^{-1/2} \log n \right)$. So

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

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

$$\sum_{n=1}^{\infty} P \left( |\varepsilon_n| \geq D_n \right) \leq \sum_{n=1}^{\infty} \frac{E \left| \varepsilon_n \right|^{2+\delta}}{D_n^{2+\delta}} = \sum_{n=1}^{\infty} \frac{E \left( E \left| \varepsilon_n \right|^{2+\delta} | \mathbf{X}n, \mathbf{T}n \right)}{D_n^{2+\delta}} \leq \sum_{n=1}^{\infty} \frac{M_\delta}{D_n^{2+\delta}} \leq \infty,$$

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

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

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

$$\boldsymbol{\mu}_\alpha \left( x_{1,k} \right)^T \left\{ s_{(J,\alpha),\left(J',\alpha'\right)} \right\}_{J,J'} \operatorname{var} \left( \left\{ B_{J,\alpha'}(X_{i\alpha'}) \right\}_{1 \leq J \leq N+1}^{T} \sigma \left( \mathbf{X}_i, \mathbf{T}_i \right) \varepsilon_{i,D}^{*} \right)$$
$$\times \left\{ s_{(J,\alpha),\left(J',\alpha'\right)} \right\}_{1 \leq J,J' \leq N+1}^{T} \boldsymbol{\mu}_\alpha \left( x_{1,k} \right).$$

By Assumption A3, $c_\sigma^2 \left( v_{(J,\alpha'),\left(J',\alpha'\right)} \right)_{1 \leq J,J' \leq N+1}$ is bounded by

$$\operatorname{var} \left( \left\{ B_{J,\alpha'}(X_{i\alpha'}) \right\}_{1 \leq J \leq N+1}^{T} \sigma \left( \mathbf{X}_i, \mathbf{T}_i \right) \right) \leq C_\sigma^2 \left( v_{(J,\alpha'),\left(J',\alpha'\right)} \right)_{1 \leq J,J' \leq N+1}.$$

$$\text{var}\left(U_{i,k}\right) \sim \boldsymbol{\mu}_\alpha\left(x_{1,k}\right)^T \left\{s_{(J,\alpha),(J',\alpha')}\right\}_{1\le J,J'\le N+1} \left(v_{(J,\alpha'),(J',\alpha')}\right)_{1\le J,J'\le N+1}$$

$$\times \left\{s_{(J,\alpha),(J',\alpha')}\right\}^T_{1\le J,J'\le N+1} \boldsymbol{\mu}_\alpha\left(x_{1,k}\right) V_{\varepsilon,D} \sim$$

$$\boldsymbol{\mu}_\alpha\left(x_{1,k}\right)^T \left\{s_{(J,\alpha),(J',\alpha')}\right\}_{J,J'} \left\{s_{(J,\alpha),(J',\alpha')}\right\}^T_{J,J'} \boldsymbol{\mu}_\alpha\left(x_{1,k}\right) V_{\varepsilon,D},$$

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

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

Since $E\left|U_{i,k}\right|^r \le \left\{c_0\kappa\left(x_{1,k}\right) D_n H^{-1/2}\right\}^{r-2} r! E\left|U_{i,k}\right|^2 < +\infty$, for $r \ge 3$, $\left\{U_{i,k}\right\}_{i=1}^n$ satisfies the Cramér's condition with Cramér's constant

$c_* = c_0\kappa\left(x_{1,k}\right) D_n H^{-1/2}$, hence by Lemma 4.1

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

where $m_2^2 \sim \kappa^2\left(x_{1,k}\right) V_{\varepsilon,D}$, $m_3 \le \left\{c\kappa^3\left(x_{1,k}\right) H^{-1/2} D_n V_{\varepsilon,D}\right\}^{1/3}$, $\rho_n = \rho \log n/\sqrt{n}$, $a_1 = 2\frac{n}{q} + 2\left(1 + \frac{\rho_n^2}{25m_2^2 + 5c_*\rho_n}\right)$, $a_2(3) = 11n\left(1 + \frac{5m_3^{6/7}}{\rho_n}\right)$. Similar as in Lemma 4.4,

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

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

$$P\left(\frac{1}{n}\left|\sum_{i=1}^{n} U_{i,k}\right| > \rho n^{-1/2}\log n\right) \leq c\log n\exp\left\{-c_2\rho^2\log n\right\} + Cn^{2-6\lambda_0 c_0/7} \leq n^{-3},$$

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

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

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

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

Employing Cauchy-Schwartz inequality and Lipschitz continuity of kernel $K$, Assumption A5, Lemma 4.2 (ii) and (4.43) lead to

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

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

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

by Cauchy-Schwartz inequality, (4.43), Lemma 4.4, Assumptions A6 and A7. Thus

$$\sup_{x_1\in[0,1]}\left|\hat{\Psi}_v^{(2)}(x_1)\right| \leq \sup_{x_1\in[0,1]}\{Q_1(x_1)+Q_2(x_1)\} = O_p\left(n^{-1/2}\log n\right) = o_p\left(n^{-2/5}\right).$$

The desired result follows from the above result and (4.44). ∎

LEMMA 4.14. *Under Assumptions A2-A4, A6 and A7, as* $n\to\infty$,

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

Lemma 4.14 follows from Lemmas 4.12 and 4.13. Next we bound $\tilde{\mathbf{c}}_m^T - \mathbf{c}^T, \tilde{\mathbf{c}}_\varepsilon^T$ defined in (4.17), (4.16). Denote by $\mathbf{I}_{r\times d}$ the matrix $\begin{pmatrix} \mathbf{I}_r & \mathbf{0}_{r\times d}\end{pmatrix}$.

LEMMA 4.15. *Under Assumptions A1, A2, A4-A7, as* $n\to\infty$, $\|\tilde{\mathbf{c}}_m - \mathbf{c}\| = o_p\left(n^{-1/2}\right)$.

*Proof.* By the result on page 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 $Eg_\alpha(X_\alpha) = 0, \|g_\alpha - m_\alpha\|_\infty \leq C_\infty \left\|m_\alpha^{'}\right\| H^2$, $1\leq\alpha\leq d_2$. Then

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

with $\hat{\mathbf{V}}$ defined in (4.28). So by Lemma 4.7, $\|\tilde{\mathbf{c}}_m - \mathbf{c}\|^2$ is bounded with probability ap-

proaching 1 by $\left\|\hat{\mathbf{V}}^{-1}n^{-1}\mathbf{B}^T\left[\sum_{\alpha=1}^{d_2}\left\{m_\alpha\left(X_{i\alpha}\right)-g_\alpha\left(X_{i\alpha}\right)\right\}\right]_{1\le i\le n}\right\|^2$

$$\le C_S^2\left\|n^{-1}\mathbf{B}^T\left[\sum_{\alpha=1}^{d_2}\left\{m_\alpha\left(X_{i\alpha}\right)-g_\alpha\left(X_{i\alpha}\right)\right\}\right]_{1\le i\le n}\right\|^2$$

$$\le C_S^2\left(\sum_{\alpha=1}^{d_2}\|g_\alpha-m_\alpha\|_\infty\right)^2+C_S^2\sum_{l=1}^{d_1}\left(\sum_{\alpha=1}^{d_2}\|g_\alpha-m_\alpha\|_\infty n^{-1}\sum_{i=1}^{n}\left|T_{il}\right|\right)^2$$

$$+C_S^2\sum_{\alpha'=1}^{d_2}\sum_{J=1}^{N+1}\left(\sum_{\alpha=1}^{d_2}\|g_\alpha-m_\alpha\|_\infty n^{-1}\sum_{i=1}^{n}\left|B_{J,\alpha'}\left(X_{i\alpha'}\right)\right|\right)$$

$$\le C_S^2\left(\sum_{\alpha=1}^{d_2}\|g_\alpha-m_\alpha\|_\infty\right)^2\left\{1+d_1\left(\sup_{1\le l\le d_1}n^{-1}\sum_{i=1}^{n}\left|T_{il}\right|\right)^2\right.$$

$$\left.+\left(N+1\right)d_2\left(\sup_{1\le\alpha'\le d_2,1\le J\le N+1}n^{-1}\sum_{i=1}^{n}\left|B_{J,\alpha'}\left(X_{i\alpha'}\right)\right|\right)^2\right\}.$$

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

LEMMA 4.16. *Under Assumptions A1-A7, as* $n\to\infty$, $\|\tilde{\mathbf{c}}_\varepsilon\|=O_p\left(n^{-1/2}\right)$.

*Proof.* Let $(\mathbf{Q}_1,\mathbf{Q}_2)=\mathbf{I}_{(1+d_1)\times\{d_2(N+1)\}}\left(\hat{\mathbf{a}},\tilde{\mathbf{a}}-\hat{\mathbf{a}}\right)$. By (4.40), (4.42), (4.27)

$$\tilde{\mathbf{c}}_\varepsilon^T=\mathbf{I}_{(1+d_1)\times\{d_2(N+1)\}}\tilde{\mathbf{a}}=\mathbf{Q}_1+\mathbf{Q}_2, \tag{4.45}$$

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

$$\mathbf{Q}_1=\left(n^{-1}\sum_{i=1}^{n}\sigma\left(\mathbf{X}_i,\mathbf{T}_i\right)\varepsilon_i,n^{-1}\sum_{i=1}^{n}\xi_{il'}\right)_{1\le l'\le d_1}^T \tag{4.46}$$

in which $\xi_{il'}=\left\{\sum_l s_{l'l}T_{il}+\sum_{\alpha,J}s_{l',(J,\alpha)}B_{J,\alpha}\left(X_{i\alpha}\right)\right\}\sigma\left(\mathbf{X}_i,\mathbf{T}_i\right)\varepsilon_i$. Clearly $E\xi_{il'}=0$ while $E\xi_{il'}^2\le C_\sigma^2E\left[\left\{\sum_l s_{l'l}T_l+\sum_{\alpha,J}s_{l',(J,\alpha)}B_{J,\alpha}\left(X_\alpha\right)\right\}\right]^2\le C_\sigma^2\left(0,s_{l'l},s_{l',(J,\alpha)}\right)\times$

$\mathbf{V}\left(0, s_{l'l}, s_{l',(J,\alpha)}\right)^T \leq C_\sigma^2 C_V \left\|\left(0, s_{l'l}, s_{l',(J,\alpha)}\right)\right\|^2 \leq C_\sigma^2 C_V C_S^2$. It is easily checked $E\left(\xi_{il'}\xi_{jl'}\right) = 0$ for $i \neq i'$ thus by Markov Inequality, $\sup_{1\leq l'\leq d_1}\left|n^{-1}\sum_{i=1}^n \xi_{il'}\right| = O_p(n^{-1/2})$. Likewise $n^{-1}\sum_{i=1}^n \sigma\left(\mathbf{X}_i, \mathbf{T}_i\right)\varepsilon_i = O_p(n^{-1/2})$. Then, Lemma 4.16 follows from the above results. ∎

LEMMA 4.17. *Under Assumptions A1, A2, A4-A7, as* $n \to \infty$, $\sup_{x_1\in[0,1]}\left|\Psi_{Tb}(x_1)\right| = o_p\left(1/\sqrt{n}\right)$.

For proof, see Ma and Yang (2011b). Define a theoretical version of
$\Psi_{Tv}(x_1) = n^{-1}\sum_{i=1}^n K_h\left(X_{i1} - x_1\right)\left(\tilde{a}_0 + \sum_{l=1}^{d_1}\tilde{a}_l T_{il}\right)$ as
$\hat{\Psi}_{Tv}(x_1) = \hat{a}_0 n^{-1}\sum_{i=1}^n K_h\left(X_{i1} - x_1\right) + \sum_{l=1}^{d_1}\hat{a}_l n^{-1}\sum_{i=1}^n \zeta_l\left(X_{i1}, T_{il}, x_1\right)$.

LEMMA 4.18. *Under Assumptions A1-A7, as* $n \to \infty$,
$\sup_{x_1\in[0,1]}\left|\Psi_{Tv}(x_1)\right| = O_p\left\{n^{-1/2}(\log n)^4\right\} = o_p\left(n^{-2/5}\right)$.

For proof, see Ma and Yang (2011b).

LEMMA 4.19. *Under Assumptions A1-A6 and A8,* $\left(s_{l',l}\right) = \operatorname{cov}\left(\tilde{\mathbf{T}}_n\right)^{-1}$ *and as* $n \to \infty$, $\operatorname{cov}\left(\tilde{\mathbf{T}}_n\right)^{-1} \to \operatorname{cov}\left(\tilde{\mathbf{T}}\right)^{-1}$, *where* $\left(s_{l',l}\right)$ *defined in (4.27),* $\tilde{\mathbf{T}}_n$ *and* $\tilde{\mathbf{T}}$ *defined in (4.10).*

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

**Proof of Theorem** 4.1. The term $\Psi_{Tb}(x_1)$ in (4.31) has order $o_p(n^{-1/2})$ and other terms have order $o_p(n^{-2/5})$ by Lemmas 4.17, 4.18, 4.9, 4.14. Standard theory ensures that

$\hat{f}_1(x_1) = n^{-1}\sum_{i=1}^{n} K_h\left(X_{i1} - x_1\right)$ has a positive lower bound. Theorem 4.1 then follows. □

**Proof of Theorem** 4.2. The first part of Theorem 4.2 follows from Lemmas 4.15 and 4.16. By (4.15), (4.45) and Lemma 4.15, $\sqrt{n}\left(\hat{\mathbf{c}} - \mathbf{c}\right) = \sqrt{n}\tilde{\mathbf{c}}_{\varepsilon} + \sqrt{n}\left(\tilde{\mathbf{c}}_m - \mathbf{c}\right) = \sqrt{n}\left(\mathbf{Q}_1 + \mathbf{Q}_2\right) + \sqrt{n}\left(\tilde{\mathbf{c}}_m - \mathbf{c}\right) = \sqrt{n}\mathbf{Q}_1 + o_p\{1\}$. It is easily verified $E\left(\sqrt{n}\mathbf{Q}_1\right) = 0$. By (4.46), Lemma 4.19,

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

Theorem 4.2 then follows by applying Theorem 1 of Sunklodas (1984). □

# Chapter 5

# Spline Regression in the Presence of Categorical Predictors

## 5.1 Background

This chapter is based on Ma, Racine and Yang (2011). Applied researchers must frequently model relationships involving both continuous and categorical predictors, and a range of nonparametric kernel regression methods have recently been proposed for such settings. These developments have extended the reach of kernel smoothing methods beyond the traditional continuous-only predictor case and have had a marked impact on applied nonparametric research; see Li and Racine (2007) for examples and an overview. Though kernel methods hold much appeal for practitioners, many in the applied community continue to resist their use, often for valid reasons. In particular, nonparametric kernel methods are local in nature, bandwidth selection can be fragile and numerically demanding, interpretation can be challenging, while imposing constraints on the resulting estimate can be difficult.

Regression spline methods, on the other hand, are global in nature and involve straightforward least squares solutions hence unconstrained and constrained estimation is much easier to handle and faster to compute. Furthermore, their least squares underpinnings render the methods immediately accessible to those who routinely use least squares approaches. As such, regression splines provide an immediately accessible and attractive alternative to kernel methods for the nonparametric estimation of regression models. For excellent overviews of spline modeling we direct the interested reader to Stone (1985), Stone (1994), Huang (2003), and Wang and Yang (2009a). For applications of spline approaches, see Huang (1998) for functional ANOVA models, Huang and Yang (2004), Wang and Yang (2007) and Xue (2009) for additive models, Wang and Yang (2009a) and Wang (2009) for single-index models, Liu and Yang (2010) and Xue and Liang (2010) for additive coefficient models, and Ma and Yang (2011) for jump detection in regression functions. However, just like their traditional kernel-based continuous-only predictor kin, regression splines are limited by their inability to handle the presence of categorical predictors without resorting to sample-splitting which can entail a substantial loss in efficiency. In this chapter we consider a regression spline alternative motivated by recent developments in the kernel smoothing of relationships involving categorical covariates. The proposed spline approach possesses intuitive appeal by producing a piecewise polynomial, computational expedience as discussed before and theoretical reliability according to the mean square and uniform convergence rates, and the pointwise asymptotic distribution results established in this chapter.

The remainder of this chapter proceeds as follows. Section 5.2 outlines the framework and presents theorems detailing rates of convergence and the asymptotic distribution of the proposed approach. Section 5.3 considers cross-validatory selection of the spline knot vector

and kernel bandwidth vector. Section 5.4 examines the finite-sample performance of the proposed method versus the traditional 'frequency' (i.e. 'sample-splitting') estimator, the additive regression spline estimator, and the cross-validated local linear kernel regression estimator, while proofs are to be found in the appendix.

## 5.2 Methods and Main Results

In what follows we presume that the reader is interested in the unknown conditional mean in the following location-scale model,

$$Y = g\left(\mathbf{X}, \mathbf{Z}\right) + \sigma\left(\mathbf{X}, \mathbf{Z}\right)\varepsilon, \tag{5.1}$$

where $g(\cdot)$ is an unknown function, $\mathbf{X} = \left(X_1, \ldots, X_q\right)^{\mathrm{T}}$ is a $q$-dimensional vector of continuous predictors, and $\mathbf{Z} = \left(Z_1, \ldots, Z_r\right)^{\mathrm{T}}$ is an $r$-dimensional vector of categorical predictors. Letting $\mathbf{z} = \left(z_s\right)_{s=1}^{r}$, we assume that $z_s$ takes $c_s$ different values in $D_s \equiv \{0, 1, \ldots, c_s - 1\}$, $s = 1, \ldots, r$, and let $c_s$ be a finite positive constant. Let $\left(Y_i, \mathbf{X}_i^{\mathrm{T}}, \mathbf{Z}_i^{\mathrm{T}}\right)_{i=1}^{n}$ be an i.i.d copy of $\left(Y, \mathbf{X}^{\mathrm{T}}, \mathbf{Z}^{\mathrm{T}}\right)$. Assume for $1 \leq l \leq q$, each $X_l$ is distributed on a compact interval $\left[a_l, b_l\right]$, and without loss of generality, we take all intervals $\left[a_l, b_l\right] = [0, 1]$.

In order to handle the presence of categorical predictors, we define

$$l\left(Z_s, z_s, \lambda_s\right) = \begin{cases} 1, \text{when } Z_s = z_s \\ \lambda_s, \text{ otherwise.} \end{cases},$$

$$L\left(\mathbf{Z}, \mathbf{z}, \boldsymbol{\lambda}\right) = \prod_{s=1}^{r} l\left(Z_s, z_s, \lambda_s\right) = \prod_{s=1}^{r} \lambda_s^{1(Z_s \neq z_s)}, \tag{5.2}$$

where $l(\cdot)$ is a variant of Aitchison and Aitken (1976) univariate categorical kernel function, $L(\cdot)$ is a product categorical kernel function, and $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_r)^{\mathrm{T}}$ is the vector of bandwidths for each of the categorical predictors. We use the tensor product splines introduced in Section 1.4 of Chapter 1. Let $\mathbf{B} = \left[\{\mathcal{B}(\mathbf{X}_1), \ldots, \mathcal{B}(\mathbf{X}_n)\}^{\mathrm{T}}\right]_{n \times \mathbf{K}_n}$, where $\mathcal{B}(\mathbf{x})$ is defined in (1.3). Then $g(\mathbf{x}, \mathbf{z})$ can be approximated by $\mathcal{B}(\mathbf{x})^{\mathrm{T}} \beta(\mathbf{z})$, where $\beta(\mathbf{z})$ is a $\mathbf{K}_n \times 1$ vector. We estimate $\beta(\mathbf{z})$ by minimizing the following weighted least squares criterion,

$$\widehat{\beta}(\mathbf{z}) = \arg \min_{\beta(\mathbf{z}) \in \mathbb{R}\mathbf{K}n} \sum_{i=1}^{n} \left\{ Y_i - \mathcal{B}\left(\mathbf{X}_i\right)^{\mathrm{T}} \beta(\mathbf{z}) \right\}^2 L\left(\mathbf{Z}_i, \mathbf{z}, \boldsymbol{\lambda}\right).$$

Let $\mathcal{L}_z = \operatorname{diag}\{L(\mathbf{Z}_1, \mathbf{z}, \boldsymbol{\lambda}), \ldots, L(\mathbf{Z}_n, \mathbf{z}, \boldsymbol{\lambda})\}$ be a diagonal matrix with $L\left(\mathbf{Z}_i, \mathbf{z}, \boldsymbol{\lambda}\right)$, $1 \leq i \leq n$ as the diagonal entries. Then $\widehat{\beta}(\mathbf{z})$ can be written as

$$\widehat{\beta}(\mathbf{z}) = \left(n^{-1}\mathbf{B}^{\mathrm{T}}\mathcal{L}_z\mathbf{B}\right)^{-1}\left(n^{-1}\mathbf{B}^{\mathrm{T}}\mathcal{L}_z\mathbf{Y}\right), \tag{5.3}$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$. $g(\mathbf{x}, \mathbf{z})$ is estimated by $\widehat{g}(\mathbf{x}, \mathbf{z}) = \mathcal{B}(\mathbf{x})^{\mathrm{T}} \widehat{\beta}(\mathbf{z})$.

Given any $\mathbf{z} \in \mathcal{D}$, for any $\mu \in (0, 1]$, we denote by $C^{0,\mu}[0,1]^q$ the space of order $\mu$ Hölder continuous functions on $[0,1]^q$, i.e.,

$$C^{0,\mu}[0,1]^q = \left\{ \phi : \|\phi\|_{0,\mu,\mathbf{z}} = \sup_{\mathbf{x} \neq \mathbf{x}', \mathbf{x}, \mathbf{x}' \in [0,1]^q} \frac{\left|\phi(\mathbf{x}, \mathbf{z}) - \phi\left(\mathbf{x}', \mathbf{z}\right)\right|}{\left\|\mathbf{x} - \mathbf{x}'\right\|_2^{\mu}} < +\infty \right\}$$

in which $\|\mathbf{x}\|_2 = \left(\sum_{l=1}^{q} x_l^2\right)^{1/2}$ is the Euclidean norm of $\mathbf{x}$ , and $\|\phi\|_{0,\mu,\mathbf{z}}$ is the $C^{0,\mu}$ -norm of $\phi$. Let $C[0,1]^q$ be the space of continuous functions on $[0,1]^q$. Given a $q$-tuple $\boldsymbol{\alpha} = \left(\alpha_1, \ldots, \alpha_q\right)$ of nonnegative integers, let $[\boldsymbol{\alpha}] = \alpha_1 + \cdots + \alpha_q$ and let $D^{\boldsymbol{\alpha}}$ denote the differential operator defined by $D^{\boldsymbol{\alpha}} = \dfrac{\partial^{[\boldsymbol{\alpha}]}}{\partial x_1^{\alpha_1} \cdots \partial x_q^{\alpha_q}}$. For positive numbers $a_n$ and $b_n$,

$n \geq 1$, let $a_n \asymp b_n$ mean that $\lim_{n\to\infty} a_n/b_n = c$, where $c$ is some nonzero constant. The assumptions employed for the asymptotic results are listed below:

(A1) *For any given* $\mathbf{z} \in \mathcal{D}$, *the regression function satisfies* $D^{\boldsymbol{\alpha}} g \in C^{0,1}[0,1]^q$ , *for all* $\boldsymbol{\alpha}$ *with* $[\boldsymbol{\alpha}] = p-1$ *and* $1 \leq p \leq \min(m_1, \ldots, m_q)$.

(A2) *The marginal density* $f(\mathbf{x})$ *of* $\mathbf{X}$ *satisfies* $f(\mathbf{x}) \in C[0,1]^q$ *and* $f(\mathbf{x}) \in \left[c_f, C_f\right]$ *for constants* $0 < c_f \leq C_f < \infty$. *There exists a constant* $c_P > 0$, *such that* $P(\mathbf{Z} = \mathbf{z} \,|\mathbf{X}) \geq c_P$ *for all* $\mathbf{z} \in \mathcal{D}$.

(A3) *The noise* $\varepsilon$ *satisfies* $E(\varepsilon \,|\mathbf{X}, \mathbf{Z}) = 0$, $E\left(\varepsilon^2 \,|X, \mathbf{Z}\right) = 1$. *There exists a positive value* $\delta > 0$ *and finite positive* $M_\delta$ *such that* $\sup_{\mathbf{x}\in[0,1]^q, \mathbf{z}\in\mathcal{D}} E\left(|\varepsilon|^{2+\delta} \,|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}\right) < M_\delta$ *and* $E\left(|\varepsilon|^{2+\delta}\right) < M_\delta$. *The standard deviation function* $\sigma(\mathbf{x}, \mathbf{z})$ *is continuous on* $[0,1]^q \times \mathcal{D}$ *for* $\mathcal{D} = D_1 \times \cdots \times D_r$ *and* $0 < c_\sigma \leq \inf_{\mathbf{x}\in[0,1]^q, \mathbf{z}\in\mathcal{D}} \sigma(\mathbf{x}, \mathbf{z}) \leq \sup_{\mathbf{x}\in[0,1]^q, \mathbf{z}\in\mathcal{D}} \sigma(\mathbf{x}, \mathbf{z}) \leq C_\sigma < \infty$.

(A4) *The number of interior knots* $N_l, 1 \leq l \leq q$ *satisfy as* $n \to \infty$, $\max_{1\leq l\leq q}\left(N_l^{-1}\right) = o\left\{n^{-1/(2p+q)}\right\}$, $\prod_{l=1}^q N_l = o\left\{n(\log n)^{-1}\right\}$ , *and the bandwidths* $\lambda_s, 1 \leq s \leq r$ *satisfy as* $n \to \infty$, $\sum_{s=1}^r \lambda_s = o\left\{\left(n^{-1}\prod_{l=1}^q N_l\right)^{1/2}\right\}$.

THEOREM 5.1. *Under assumptions (A1)-(A4), as* $n \to \infty$,

$$\sup_{\mathbf{x}\in[0,1]^q, \mathbf{z}\in\mathcal{D}} |\widehat{g}(\mathbf{x}, \mathbf{z}) - g(\mathbf{x}, \mathbf{z})| = O_{a.s.}\left[\sum_{l=1}^q h_l^p + \sum_{s=1}^r \lambda_s + \left\{\frac{1}{n}\left(\prod_{l=1}^q h_l\right)^{-1} \log n\right\}^{1/2}\right].$$

THEOREM 5.2. *Under assumptions (A1)-(A4), for* $\widehat{\sigma}_n^2(\mathbf{x}, \mathbf{z})$ *in (5.14), as* $n \to \infty$, $\widehat{\sigma}_n^{-1}(\mathbf{x}, \mathbf{z})\{\widehat{g}(\mathbf{x}, \mathbf{z}) - g(\mathbf{x}, \mathbf{z})\} \longrightarrow \mathbf{N}(0,1)$. *For any given* $(\mathbf{x}, \mathbf{z}) \in [0,1]^q \times \mathcal{D}$, $c_\sigma^* n^{-1} \times \left(\prod_{l=1}^q h_l\right)^{-1} \leq \widehat{\sigma}_n^2(\mathbf{x}, \mathbf{z}) \leq C_\sigma^* n^{-1}\left(\prod_{l=1}^q h_l\right)^{-1}$, *for some constants* $0 < c_\sigma^* < C_\sigma^* < \infty$.

Proofs of these theorems are presented in the appendix. Having outlined the theoretical underpinnings of the proposed method, we now consider an illustrative simulated example that demonstrates how smoothing the categorical predictors in the manner prescribed above impacts the resulting estimator $\hat{g}(\mathbf{x}, \mathbf{z})$.

## 5.3 Cross-Validated Choice of $N$ and $\lambda$

Cross-validation Stone (1977) has a rich pedigree in the regression spline arena and has been used for decades to choose the appropriate number of interior knots and is the basis for Friedman (1991) multivariate adaptive regression spline (MARS) methodology among others; see Wahba (199) for an overview in the spline context. It has also been used extensively for bandwidth selection for kernel estimators such as the local linear kernel estimator proposed by Li and Racine (2004) that appears in the simulations in Section 5.4 (see also Racine and Li (2004) for the local constant counterpart). Following in this tradition we choose the number of interior knots (i.e. the vector $(N)$) and smoothing parameters (i.e. the bandwidth vector $\lambda$) by minimizing the cross-validation function defined by

$$CV(N, \lambda) = n^{-1} \sum_{i=1}^{n} (Y_i - B_m(X_i)^T \hat{\beta}_{-i}(Z_i))^2, \tag{5.4}$$

where $\hat{\beta}_{-i}(Z_i)$ denotes the leave-one-out estimate of $\beta$.

To illustrate the behavior of the data-driven cross-validated selection of $N$ and $\lambda$, we consider two simple data generating processes (DGPs) and plot the resulting cross-validated regression estimate based on the popular cubic spline basis. For each figure we regress $Y_i$ on $X_i$ and $Z_i$ using the proposed method. Figure 5.1 presents four simulated data sets and the

cross-validated values of $N$ and $\lambda$ (maximum $N$ is 15, $n = 500$, $\lambda \in [0, 1]$).

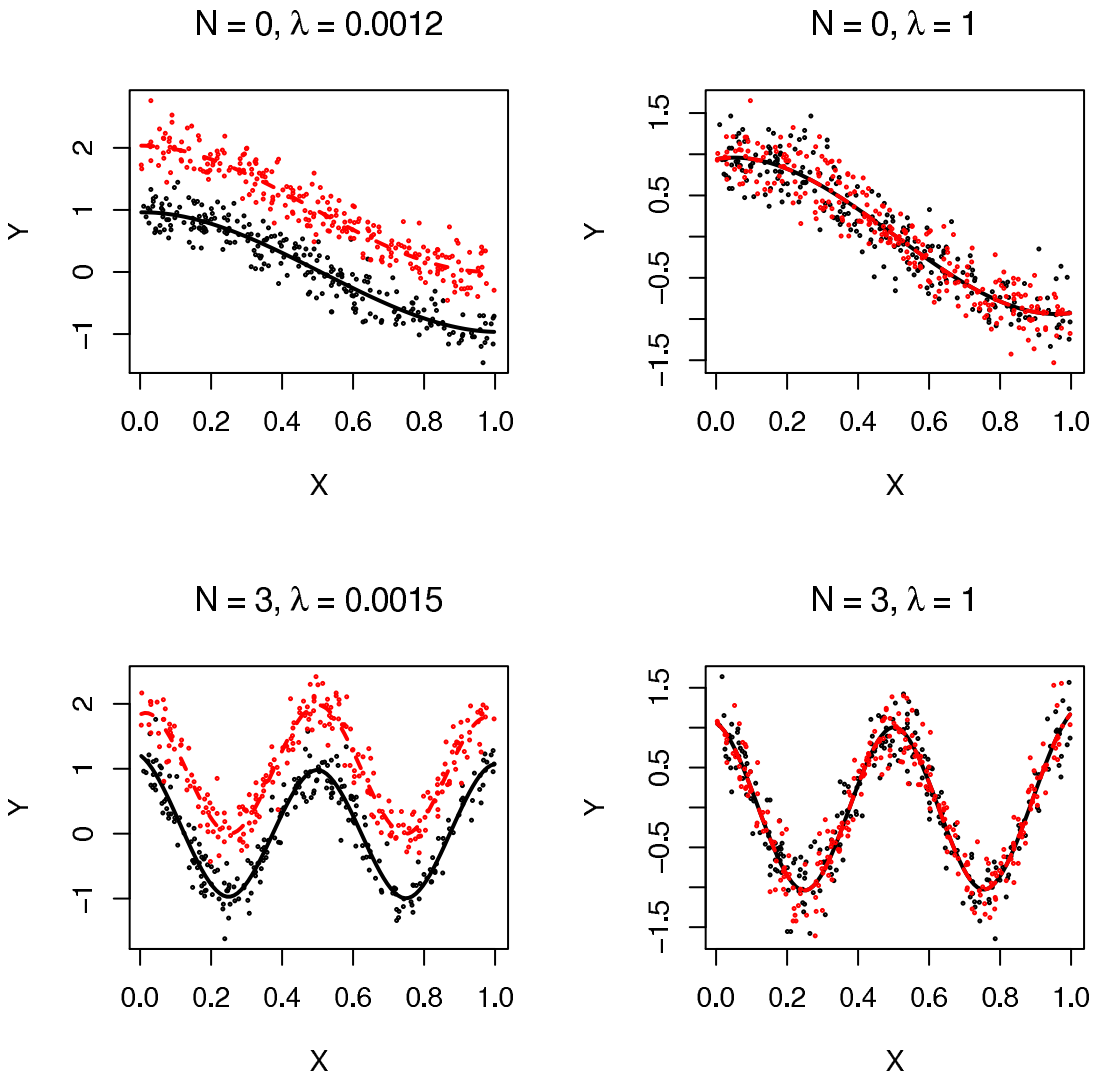


Figure 5.1: Example with $n = 500$ and a variety of DGPs in Chapter 5

Note: for the lower left DGP there is no difference in the function when $Z = 0$ versus $Z = 1$ and cross-validation ought to select $\lambda = 1$ ($N$ the number of interior knots, $\lambda$ the bandwidth).

Figure 5.1 illustrates how the cross-validated choices of $N$ and $\lambda$ differ depending on the underlying DGP. For instance, the plot on the upper left is one for which $Y_i = \cos(\pi X_i) + \varepsilon_i$ if $Z_i = 0$ and $Y_i = 1 + \cos(\pi X_i) + \varepsilon_i$ if $Z_i = 1$. It is evident that $N \approx 0$ and $\lambda \approx 0.0012$ is appropriate here. The figure on the upper right is one for which $Y_i = \cos(\pi X_i) + \varepsilon_i$

regardless of the value taken on by $Z_i$. Those on the lower left and right are similar but use $\cos(4\pi X_i)$ instead. For both the lower and upper figures on the right, $N \approx 0$ and $N \approx 3$ appear to be appropriate while $\lambda \approx 1$ and $\lambda \approx 1$ are also appropriate since $Z_i$ is independent of $Y_i$. These simple examples serve to illustrate how cross-validation is delivering values of $N$ and $\lambda$ that are appropriate for the DGP at hand.

Before proceeding, a few words on the numerical optimization of (5.4) are in order. Search takes place over $N_1, \ldots, N_q$ and $\lambda_1, \ldots, \lambda_r$ where the $\lambda$ are continuous lying in $[0, 1]$ and the $N$ are integers. Clearly this is a mixed integer combinatorial optimization procedure which would render exhaustive search infeasible when facing a non-trivial number of predictors. However, in settings such as these one could leverage recent advances in mixed integer search algorithms which is the avenue we pursue in the Monte Carlo simulations reported below. In particular, we adopt the 'Nonsmooth Optimization by Mesh Adaptive Direct Search' (NOMAD) approach (Abramson, Audet, Couture, Dennis Jr., and Le Digabel (2011)). Given that the objective function can be trivially computed for large sample sizes as it involves nothing more than computing the hat matrix for weighted least squares, it turns out that the computational burden is in fact nowhere near as costly as, say, cross-validated kernel regression for moderate to large data sets. As such, the proposed approach constitutes a computationally attractive alternative to multivariate cross-validated kernel regression. In addition, in the next section we shall also see that the proposed approach constitutes a statistically attractive alternative as well, at least from the perspective of finite-sample square error loss.

## 5.4 Monte Carlo Simulations

In this section we consider a modest Monte Carlo experiment designed to assess the finite-sample performance of the proposed method. We consider two DGPs with $q = 2$ continuous predictors and $r = 2$ categorical predictors given by

$$\begin{aligned} \text{DGP-M:} \qquad & Y_i = \cos(2\pi X_{i1}) \times \sin(2\pi X_{i2}) + (Z_{i1} + Z_{i2})/10 + \varepsilon_i, \\ \text{DGP-A:} \qquad & Y_i = \cos(2\pi X_{i1}) + \sin(2\pi X_{i2}) + (Z_{i1} + Z_{i2})/10 + \varepsilon_i, \end{aligned} \tag{5.5}$$

where the continuous predictors are drawn from the uniform ($X_j \sim U[0,1]$, $j = 1, 2$), the categorical predictors ($Z_j$, $j = 1, 2$) are drawn from the rectangular distribution with equal probability ($z_s \in \{0, 1, \ldots, c_s - 1\}$ where $c_s$ is the number of categorical outcomes, $c_s \geq 2$, $s = 1, 2$), and $\varepsilon \sim N(0, \sigma^2)$ with $\sigma = 0.1$. For what follows we set $c_s = c$, $s = 1, 2$. Observe that DGP-M is multiplicative in the continuous components while DGP-A is additive.

We draw $M = 1,000$ Monte Carlo replications and for each replication we compute the cross-validated frequency estimator (i.e. that based only on the $(Y, \mathbf{X})$ pairs lying in each 'cell' defined by $\mathbf{Z}$), the proposed cross-validated categorical regression spline estimator, the cross-validated additive categorical regression spline estimator Ma and Racine (2011) and the cross-validated local linear kernel estimator that is often used to model continuous and categorical predictors in a manner similar to that undertaken here (note that the local linear kernel estimator is minimax efficient and has the best boundary bias correction properties of the class of kernel regression estimators; see Li and Racine (2004) for details). For the regression spline estimators we set the spline degree vector equal to $(3, 3)$ (a popular choice) and cross-validate the number of knots vector $(N_1, N_2)$ and the bandwidth vector $(\lambda_1, \lambda_2)$.

We then compute the MSE of each estimator based upon (5.5) for each replication and report the relative median MSE over all $M$ replications in Table 5.1. Table 5.2 reports a summary of the smoothing parameters chosen by cross-validation.

Table 5.1: Relative median MSE in Chapter 5

| DGP-M: Multiplicative Specification | | | | | DGP-A: Additive Specification | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $n$ | $c$ | Frequency | Additive | Kernel | $n$ | $c$ | Frequency | Additive | Kernel |
| 500 | 2 | 0.591 | 0.008 | 0.487 | 500 | 2 | 0.554 | 2.028 | 0.489 |
| 500 | 3 | 0.572 | 0.014 | 0.618 | 500 | 3 | 0.488 | 1.939 | 0.593 |
| 500 | 4 | 0.519 | 0.020 | 0.718 | 500 | 4 | 0.434 | 1.788 | 0.702 |
| 1000 | 2 | 0.745 | 0.004 | 0.403 | 1000 | 2 | 0.776 | 2.072 | 0.417 |
| 1000 | 3 | 0.591 | 0.008 | 0.494 | 1000 | 3 | 0.518 | 2.096 | 0.494 |
| 1000 | 4 | 0.609 | 0.013 | 0.585 | 1000 | 4 | 0.523 | 2.072 | 0.571 |
| 1500 | 2 | 0.815 | 0.003 | 0.359 | 1500 | 2 | 0.822 | 2.053 | 0.381 |
| 1500 | 3 | 0.636 | 0.006 | 0.446 | 1500 | 3 | 0.660 | 2.119 | 0.446 |
| 1500 | 4 | 0.626 | 0.010 | 0.518 | 1500 | 4 | 0.520 | 2.176 | 0.514 |
| 2000 | 2 | 0.838 | 0.002 | 0.337 | 2000 | 2 | 0.852 | 2.087 | 0.368 |
| 2000 | 3 | 0.722 | 0.005 | 0.408 | 2000 | 3 | 0.754 | 2.143 | 0.418 |
| 2000 | 4 | 0.628 | 0.008 | 0.480 | 2000 | 4 | 0.572 | 2.226 | 0.478 |

Note: Relative median MSE of the proposed proposed spline regression estimator versus the frequency regression spline, additive regression spline, and local linear kernel regression estimator. Numbers less than one indicate superior performance of the proposed spline estimator. $c$ denotes the number of outcomes for the discrete predictors, $n$ the sample size.

Table 5.1 illustrates how, for a given sample size, the relative performance of the proposed approach that smooths the categorical predictors versus the frequency approach that breaks the data into $c_1 \times c_2 = (4, 9, 16)$ subsets improves as $c$ increases, as expected (each categorical predictor takes on $c_1 = c_2 = c = (2, 3, 4)$ values). Furthermore, for a given $c$, as $n$ increases the proposed estimator approaches the frequency estimator since $\lambda \to 0$ as $n \to \infty$. Table 5.1 further illustrates how the proposed cross-validated method dominates the popular local linear kernel estimator for all sample sizes and values of $c$ considered.

Often additive spline models are recommended in applied settings due to the curse of

dimensionality (the property that the multiplicative tensor regression spline has a rate of convergence that deteriorates with the number of continuous predictors, $q$). Observe, however, that even in small sample settings such as $n = 500$, if the additive model is used when additivity is not present, the square error properties of the additive regression spline model can be orders of magnitude worse than the multiplicative tensor regression spline model (the tensor model has roughly 1/100 the MSE of the additive model for DGP-M). Naturally, if additivity is appropriate the additive model that incorporates this information will have better finite-sample properties (the tensor model has roughly 2 times the MSE of the additive model for DGP-A, the additive DGP). Simulations not reported here for space considerations indicate that the finite-sample mean square error improvement over the kernel regression estimator holds a) whether or not there exist categorical predictors, and b) in higher-dimension settings than those reported here.

Table 5.2 reveals how the cross-validated bandwidths tend to zero as $n$ increases. These findings are consistent with the theoretical properties detailed in the appendix.

In the above simulations the tensor-based multivariate regression spline approach dominates the popular local linear kernel regression approach for the range of sample sizes and number of predictors considered. However, we caution the reader that this is not guaranteed to always be the case. The dimension of the tensor spline basis grows exponentially with the number of continuous predictors for a given order/knot combination for each predictor. Thus, for a fixed sample size $n$, as the number of continuous predictors $q$ increases degrees of freedom will quickly fall and the square error properties of the resulting estimator will naturally deteriorate. So, in small $n$ large $q$ low degrees of freedom settings one could readily construct instances in which the kernel regression approach will display better finite-sample

Table 5.2: Median values for smoothing parameters in Chapter 5

| DGP-M: Multiplicative Specification | | | | | | DGP-A: Additive Specification | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | $c$ | $N_1$ | $N_2$ | $\lambda_1$ | $\lambda_2$ | $n$ | $c$ | $N_1$ | $N_2$ | $\lambda_1$ | $\lambda_2$ |
| 500 | 2 | 2 | 3 | 0.16 | 0.16 | 500 | 2 | 2 | 3 | 0.16 | 0.16 |
| 500 | 3 | 2 | 3 | 0.11 | 0.11 | 500 | 3 | 2 | 3 | 0.11 | 0.11 |
| 500 | 4 | 2 | 1 | 0.07 | 0.08 | 500 | 4 | 2 | 3 | 0.09 | 0.09 |
| 1000 | 2 | 2 | 3 | 0.10 | 0.10 | 1000 | 2 | 2 | 3 | 0.10 | 0.10 |
| 1000 | 3 | 2 | 3 | 0.07 | 0.07 | 1000 | 3 | 2 | 3 | 0.07 | 0.07 |
| 1000 | 4 | 2 | 3 | 0.05 | 0.05 | 1000 | 4 | 2 | 3 | 0.05 | 0.05 |
| 1500 | 2 | 2 | 3 | 0.07 | 0.07 | 1500 | 2 | 2 | 3 | 0.07 | 0.07 |
| 1500 | 3 | 2 | 3 | 0.05 | 0.05 | 1500 | 3 | 2 | 3 | 0.05 | 0.05 |
| 1500 | 4 | 2 | 3 | 0.04 | 0.04 | 1500 | 4 | 2 | 3 | 0.04 | 0.04 |
| 2000 | 2 | 2 | 3 | 0.06 | 0.06 | 2000 | 2 | 2 | 3 | 0.06 | 0.06 |
| 2000 | 3 | 2 | 3 | 0.04 | 0.04 | 2000 | 3 | 2 | 3 | 0.04 | 0.04 |
| 2000 | 4 | 2 | 3 | 0.03 | 0.03 | 2000 | 4 | 2 | 3 | 0.03 | 0.03 |

Note: Median values for the number of interior knots and bandwidths for the proposed spline regression estimator. $c = c_1 = c_2$ denotes the number of outcomes for the discrete predictors, $n$ the sample size, $N_j$ the number of interior knots for continuous predictor $X_j$, $j = 1, 2$, and $\lambda_j$ the bandwidth for continuous predictor $Z_j$, $j = 1, 2$.

behavior than the regression spline approach. We therefore offer the following advice for the sound practical application of the methods proposed herein:

1. The proposed methods are best suited to settings in which $q$ is not overly large and $n$ not overly small.

   Our experience shows that for a range of DGPs the regression spline outperforms kernel regression when $n \geq 500$ and $q \leq 5$.

   One practical advantage is the reduced computational burden of cross-validation for regression splines versus their kernel counterpart, and in large sample settings (say, $n \geq 10,000$) one can push the dimension of $q$ much higher than that considered here.

2. Of course, when the dimension of the multivariate tensor spline becomes a practical barrier to their sound application, one can always resort to additive spline models; see

Ma and Racine (2011) for details. The drawback of the additive spline approach is that if the DGP is non-additive, the inefficiency of the additive spline approach can be much worse than the multivariate kernel approach as clearly demonstrated above. Of course, it is a simple matter to compare the value of the cross-validation function for each of the tensor-based, additive-based, and kernel-based cross-validated approaches and it is perfectly sensible to use this as a guide in applied settings. But our experience is that the tensor-based multivariate regression spline will indeed be competitive and ought to be part of every practitioners toolkit.

In summary, the simulations outlined above indicate that the proposed method is capable of outperforming the frequency estimator that breaks the sample into subsets, while it provides a compelling alternative to kernel methods when faced with a mix of categorical and continuous predictors and to additive regression spline models for general nonlinear DGPs for which additivity is not fully present.

## 5.5 Concluding Remarks

Applied researchers frequently must model relationships containing categorical predictors, yet may require nonparametric estimators of, say, regression functions. The traditional kernel and spline estimators break the data into subsets defined by the categorical predictors and then model the resulting relationship involving continuous predictors only. Though consistent, these approaches are acknowledged to be inefficient. In this chapter we provide an approach that combines regression splines with categorical kernel functions that is capable of overcoming the efficiency losses present in the traditional sample-splitting approach. Furthermore, the proposed approach constitutes an attractive alternative to cross-validated

kernel estimators that admit categorical predictors. Theoretical underpinnings are provided and simulations are undertaken to assess the finite-sample performance of the proposed method. We hope that these methods are of interest to those modelling regression functions nonparametrically when faced with both continuous and categorical predictors.

## 5.6 Appendix

For any vector $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_s) \in \mathbb{R}^s$, denote the norm $\|\boldsymbol{\zeta}\|_r = \left(|\zeta_1|^r + \cdots + |\zeta_s|^r\right)^{1/r}, 1 \leq r < +\infty$, $\|\boldsymbol{\zeta}\|_\infty = \max\left(|\zeta_1|, \ldots, |\zeta_s|\right)$. For any functions $\phi, \varphi$, define the empirical inner product and norm as $\langle \phi, \varphi \rangle_{n,\mathcal{L}_z} = n^{-1} \sum_{i=1}^n \phi\left(\mathbf{X}_i\right) \varphi\left(\mathbf{X}_i\right) L\left(\mathbf{Z}_i, \mathbf{z}, \boldsymbol{\lambda}\right)$, $\|\phi\|^2_{n,\mathcal{L}_z} = n^{-1} \sum_{i=1}^n \phi^2\left(\mathbf{X}_i\right) L\left(\mathbf{Z}_i, \mathbf{z}, \boldsymbol{\lambda}\right)$. If the functions $\phi, \varphi$ are $L_2$-integrable, we define the theoretical inner product and the corresponding norm as $\langle \phi, \varphi \rangle_{\mathcal{L}_z} = E\left\{\phi\left(\mathbf{X}\right) \varphi\left(\mathbf{X}\right) L\left(\mathbf{Z}, \mathbf{z}, \boldsymbol{\lambda}\right)\right\}$, $\|\phi\|^2_{\mathcal{L}_z} = E\left\{\phi^2\left(\mathbf{X}\right) L\left(\mathbf{Z}, \mathbf{z}, \boldsymbol{\lambda}\right)\right\}$. We denote by the same letters $c$, $C$, any positive constants without distinction. For any $s \times s$ symmetric matrix $\mathbf{A}$, denote its $L_r$ norm as $\|\mathbf{A}\|_r = \max_{\boldsymbol{\zeta} \in \mathbb{R}^s, \boldsymbol{\zeta} \neq \mathbf{0}} \|\mathbf{A}\boldsymbol{\zeta}\|_r \|\boldsymbol{\zeta}\|_r^{-1}$. Let $\|\mathbf{A}\|_\infty = \max_{1 \leq i \leq s} \sum_{j=1}^s \left|A_{ij}\right|$. Let $\mathbf{I}_s$ be the $s \times s$ identity matrix.

$\widehat{\beta}\left(\mathbf{z}\right)$ can be decomposed into $\widehat{\beta}_g\left(\mathbf{z}\right)$ and $\widehat{\beta}_\varepsilon\left(\mathbf{z}\right)$, such that $\widehat{\beta}\left(\mathbf{z}\right) = \widehat{\beta}_g\left(\mathbf{z}\right) + \widehat{\beta}_\varepsilon\left(\mathbf{z}\right)$, and

$$\widehat{\beta}_g\left(\mathbf{z}\right) = \left(n^{-1}\mathbf{B}^{\mathrm{T}}\mathcal{L}_z\mathbf{B}\right)^{-1}\left(n^{-1}\mathbf{B}^{\mathrm{T}}\mathcal{L}_z\mathbf{g}\right), \widehat{\beta}_\varepsilon\left(\mathbf{z}\right) = \left(n^{-1}\mathbf{B}^{\mathrm{T}}\mathcal{L}_z\mathbf{B}\right)^{-1}\left(n^{-1}\mathbf{B}^{\mathrm{T}}\mathcal{L}_z\mathbf{E}\right), \tag{5.6}$$

where $\mathbf{g} = \left\{g\left(\mathbf{X}_1, \mathbf{Z}_1\right), \ldots, g\left(\mathbf{X}_n, \mathbf{Z}_n\right)\right\}^{\mathrm{T}}$, $\mathbf{E} = \left\{\boldsymbol{\sigma}\left(\mathbf{X}_1, \mathbf{Z}_n\right)\varepsilon_1, \ldots, \boldsymbol{\sigma}\left(\mathbf{X}_n, \mathbf{Z}_n\right)\varepsilon_n\right\}^{\mathrm{T}}$. Then $\widehat{g}\left(\mathbf{x}, \mathbf{z}\right) = \widehat{g}_g\left(\mathbf{x}, \mathbf{z}\right) + \widehat{g}_\varepsilon\left(\mathbf{x}, \mathbf{z}\right)$, in which

$$\widehat{g}_g\left(\mathbf{x}, \mathbf{z}\right) = \mathcal{B}\left(\mathbf{x}\right)^{\mathrm{T}}\widehat{\beta}_g\left(\mathbf{z}\right), \quad \widehat{g}_\varepsilon\left(\mathbf{x}, \mathbf{z}\right) = \mathcal{B}\left(\mathbf{x}\right)^{\mathrm{T}}\widehat{\beta}_\varepsilon\left(\mathbf{z}\right). \tag{5.7}$$

Next we cite a result in the Appendix of Huang (2003).

LEMMA 5.1. *Under assumptions (A2), for any* $\boldsymbol{\alpha} \in \mathbb{R}^{\mathbf{K}n}$, *there exist constants* $0 < c_{\mathcal{B}} < C_{\mathcal{B}} < \infty$ *that do not depend on* $n$, *such that for any* $z \in \mathcal{D}$

$$c_{\mathcal{B}} \left( \prod_{l=1}^{q} h_l \right) \|\boldsymbol{\alpha}\|_2^2 \leq E \left[ \left\{ \boldsymbol{\alpha}^{\mathrm{T}} \mathcal{B}(\mathbf{X}) \right\}^2 \right] \leq C_{\mathcal{B}} \left( \prod_{l=1}^{q} h_l \right) \|\boldsymbol{\alpha}\|_2^2 .$$

Since

$$E \left\{ L^k (\mathbf{Z}, \mathbf{z}, \boldsymbol{\lambda}) \,|\mathbf{X} \right\} = E \left\{ \prod_{s=1}^{r} \lambda_s^{k(\mathbf{Z}_s \neq z_s)} \,|X \right\} \geq P(\mathbf{Z} = \mathbf{z} \,|X) \geq c_p > 0,$$

$$E \left\{ L^k (\mathbf{Z}, \mathbf{z}, \lambda) \,|X \right\} \leq 1, \tag{5.8}$$

for any integer $k \geq 1$. Thus, for $C_B = C_{\mathcal{B}}$ and $c_B = c_p c_{\mathcal{B}}$, one has

$$\left\| \boldsymbol{\alpha}^{\mathrm{T}} \mathcal{B} \right\|_{\mathcal{L}_z}^2 \leq E \left[ \left\{ \boldsymbol{\alpha}^{\mathrm{T}} \mathcal{B}(\mathbf{X}) \right\}^2 \right] \leq C_B \left( \prod_{l=1}^{q} h_l \right) \|\boldsymbol{\alpha}\|_2^2 ,$$

$$\left\| \boldsymbol{\alpha}^{\mathrm{T}} \mathcal{B} \right\|_{\mathcal{L}_z}^2 \geq c_p E \left[ \left\{ \boldsymbol{\alpha}^{\mathrm{T}} \mathcal{B}(\mathbf{X}) \right\}^2 \right] \geq c_B \left( \prod_{l=1}^{q} h_l \right) \|\boldsymbol{\alpha}\|_2^2 . \tag{5.9}$$

LEMMA 5.2. *Under assumptions (A2) and (A4), as* $n \to \infty$ ,

$$\max_{\mathbf{z} \in \mathcal{D}} \max_{j_1, \ldots, j_q, j_1', \ldots j_q'} \left| \left\langle \mathcal{B}_{j_1, \ldots, j_q}, \mathcal{B}_{j_1', \ldots, j_q'} \right\rangle_{n, \mathcal{L}_z} - \left\langle \mathcal{B}_{j_1, \ldots, j_q}, \mathcal{B}_{j_1', \ldots, j_q'} \right\rangle_{\mathcal{L}_z} \right|$$

$$= O_{a.s.} \left[ \left\{ n^{-1} \left( \prod_{l=1}^{q} h_l \right) \log n \right\}^{1/2} \right].$$

*Proof.* Let $\zeta_{j_1, \ldots, j_q, j_1', \ldots, j_q', i} = \mathcal{B}_{j_1, \ldots, j_q}(\mathbf{X}_i) \, \mathcal{B}_{j_1', \ldots, j_q'}(\mathbf{X}_i) \, L(\mathbf{Z}_i, \mathbf{z}, \boldsymbol{\lambda}) -$
$E \left\{ \mathcal{B}_{j_1, \ldots, j_q}(\mathbf{X}_i) \, \mathcal{B}_{j_1', \ldots, j_q'}(\mathbf{X}_i) \, L(\mathbf{Z}_i, \mathbf{z}, \boldsymbol{\lambda}) \right\}$. When $\left| j_l - j_l' \right| \leq m_l - 1$ for all $1 \leq l \leq q$,

by the properties of the B-spline basis, there exist constants $0 < c_{B,k} < C_{B,k} < \infty$ and $0 < c'_B < C'_B < \infty$, such that $c_{B,k}\left(\prod_{l=1}^{q} h_l\right) \leq E\left|\mathcal{B}_{j_1,...,j_q}(\mathbf{X}_i)\,\mathcal{B}_{j'_1,...,j'_q}(\mathbf{X}_i)\right|^k \leq C_{B,k}\left(\prod_{l=1}^{q} h_l\right)$ and $c'_B\left(\prod_{l=1}^{q} h_l\right)^k \leq \left|E\left\{\mathcal{B}_{j_1,...,j_q}(\mathbf{X}_i)\,\mathcal{B}_{j'_1,...,j'_q}(\mathbf{X}_i)\right\}\right|^k \leq C'_B\left(\prod_{l=1}^{q} h_l\right)^k$, thus by (5.8),

$$\begin{aligned}
E\zeta^2_{j_1,...,j_q,j'_1,...,j'_q,i} &\geq c_p E\left\{\mathcal{B}^2_{j_1,...,j_q}(\mathbf{X}_i)\,\mathcal{B}^2_{j'_1,...,j'_q}(\mathbf{X}_i)\right\} \\
&\quad - \left[E\left\{\mathcal{B}_{j_1,...,j_q}(\mathbf{X}_i)\,\mathcal{B}_{j'_1,...,j'_q}(\mathbf{X}_i)\right\}\right]^2 \\
&\geq c_p c_{B,2}\left(\prod_{l=1}^{q} h_l\right) - C'_B\left(\prod_{l=1}^{q} h_l\right)^2 \geq c_{\zeta^2}\left(\prod_{l=1}^{q} h_l\right), \\
E\zeta^2_{j_1,...,j_q,j'_1,...,j'_q,i} &\leq E\left\{\mathcal{B}^2_{j_1,...,j_q}(\mathbf{X}_i)\,\mathcal{B}^2_{j'_1,...,j'_q}(\mathbf{X}_i)\right\} \leq C_{B,2}\left(\prod_{l=1}^{q} h_l\right),
\end{aligned}$$

$$\begin{aligned}
E\left|\zeta_{j_1,...,j_q,j'_1,...,j'_q,i}\right|^k &\leq 2^{k-1}\left[E\left|\mathcal{B}_{j_1,...,j_q}(\mathbf{X}_i)\,\mathcal{B}_{j'_1,...,j'_q}(\mathbf{X}_i)\,L(\mathbf{Z}_i,z,\lambda)\right|^k\right. \\
&\quad \left. + \left|E\left\{\mathcal{B}_{j_1,...,j_q}(\mathbf{X}_i)\,\mathcal{B}_{j'_1,...,j'_q}(\mathbf{X}_i)\,L(\mathbf{Z}_i,z,\lambda)\right\}\right|^k\right] \\
&\leq 2^{k-1}\left(C_{B,k}\left(\prod_{l=1}^{q} h_l\right) + C'_B\left(\prod_{l=1}^{q} h_l\right)^k\right) \leq c_{\zeta^k}\left(\prod_{l=1}^{q} h_l\right).
\end{aligned}$$

There exists a constant $c = c_{\zeta^k} c_{\zeta^2}^{-1}$, such that $E\left|\zeta_{j_1,...,j_q,j'_1,...,j'_q,i}\right|^k \leq ck!E\zeta^2_{j_1,...,j_q,j'_1,...,j'_q,i} < \infty$, for $k \geq 3$. Then by Bernstein's inequality (p.24, Theorem 1.2, BOSQ(1998)),

$$P\left(n^{-1}\left|\sum_{i=1}^{n}\zeta_{j_1,...,j_q,j'_1,...,j'_q,i}\right| \geq \left\{c'n^{-1}\left(\prod_{l=1}^{q} h_l\right)\log n\right\}^{1/2}\right)$$

$$\leq 2\exp\left\{-\frac{c'n\left(\prod_{l=1}^{q}h_l\right)\log n}{4C_{B,2}n\left(\prod_{l=1}^{q}h_l\right)+2c\left\{c'n\left(\prod_{l=1}^{q}h_l\right)\log n\right\}^{1/2}}\right\}$$

$$=2n^{-c'\left(4C_{B,2}\right)^{-1}}\leq 2n^{-4},\text{ for any } c'\geq 16C_{B,2},$$

which implies $\sum_{n=1}^{\infty}P\left[\max_{z\in\mathcal{D}}\max_{j_1,...,j_q,j'_1,...,j'_q}\left|n^{-1}\sum_{i=1}^{n}\zeta_{j_1,...,j_q,j'_1,...,j'_q,i}\right|\right.$

$$\left.\geq\left\{c'n^{-1}\left(\prod_{l=1}^{q}h_l\right)\log n\right\}^{1/2}\right]\leq 2\sum_{n=1}^{\infty}\left(\max_{1\leq s\leq r}c_s\right)^r\mathbf{K}_n^2n^{-4}<\infty.$$

Thus, the Borel-Cantelli Lemma entails that

$$\max_{\mathbf{z}\in\mathcal{D}}\max_{j_1,...,j_q,j'_1,...,j'_q}\left|n^{-1}\sum_{i=1}^{n}\zeta_{j_1,...,j_q,j'_1,...,j'_q,i}\right|=O_{a.s.}\left[\left\{n^{-1}\left(\prod_{l=1}^{q}h_l\right)\log n\right\}^{1/2}\right].$$

When $\left|j_l-j'_l\right|>m_l-1$ for some $1\leq l\leq q$,

$$\left\langle\mathcal{B}_{j_1,...,j_q},\mathcal{B}_{j'_1,...,j'_q}\right\rangle_{n,\mathcal{L}_z}-\left\langle\mathcal{B}_{j_1,...,j_q},\mathcal{B}_{j'_1,...,j'_q}\right\rangle_{\mathcal{L}_z}=0.$$

∎

LEMMA 5.3. *Under assumptions (A2) and (A4), as $n\to\infty$,*

$$A_n=\sup_{\mathbf{z}\in\mathcal{D}}\sup_{\gamma_1,\gamma_2\in\mathcal{G}}\left|\frac{\langle\gamma_1,\gamma_2\rangle_{n,\mathcal{L}_z}-\langle\gamma_1,\gamma_2\rangle_{\mathcal{L}_z}}{\|\gamma_1\|_{\mathcal{L}_z}\|\gamma_2\|_{\mathcal{L}_z}}\right|=O_{a.s.}\left[\left\{n^{-1}\left(\prod_{l=1}^{q}h_l\right)^{-1}\log n\right\}^{1/2}\right].\tag{5.10}$$

*Proof.* Any $\gamma_1,\gamma_2\in\mathcal{G}$ can be written as $\gamma_1(\mathbf{x})=\boldsymbol{\alpha}_1\mathcal{B}(\mathbf{x})$, $\gamma_2(\mathbf{x})=\boldsymbol{\alpha}_2\mathcal{B}(\mathbf{x})$ for some

vectors $\boldsymbol{\alpha}_1 = \left(\alpha_{j_1,...,j_q,1}\right) \in \mathbb{R}^{\mathbf{K}n}$, $\boldsymbol{\alpha}_2 = \left(\alpha_{j_1,...,j_q,2}\right) \in \mathbb{R}^{\mathbf{K}n}$. $\langle \gamma_1, \gamma_2 \rangle_{n,\mathcal{L}_z}$ is

$$n^{-1} \sum_{i=1}^{n} \left( \sum_{j_1,...,j_q} \alpha_{j_1,...,j_q,1} \mathcal{B}_{j_1,...,j_q} \right) \left( \sum_{j_1,...,j_q} \alpha_{j_1,...,j_q,2} \mathcal{B}_{j_1,...,j_q} \right) L\left(\mathbf{Z}_i, \mathbf{z}, \boldsymbol{\lambda}\right)$$

$$= \sum\nolimits_{j_1,...,j_q,j'_1,...,j'_q} \alpha_{j_1,...,j_q,1} \alpha_{,j'_1,...,j'_q,2} \left\langle \mathcal{B}_{j_1,...,j_q}, \mathcal{B}_{j'_1,...,j'_q} \right\rangle_{n,\mathcal{L}_z}$$

$$\langle \gamma_1, \gamma_2 \rangle_{\mathcal{L}_z} = \sum\nolimits_{j_1,...,j_q,j'_1,...,j'_q} \alpha_{j_1,...,j_q,1} \alpha_{,j'_1,...,j'_q,2} \left\langle \mathcal{B}_{j_1,...,j_q}, \mathcal{B}_{j'_1,...,j'_q} \right\rangle_{\mathcal{L}_z}.$$

According to (5.9), one has for any $\mathbf{z} \in \mathcal{D}$, there exist constants $0 < c_1 < C_1 < \infty$ and $0 < c_2 < C_2 < \infty$, such that $c_1 \left(\prod_{l=1}^{q} h_l\right)^{1/2} \|\boldsymbol{\alpha}_1\|_2 \leq \|\gamma_1\|_{\mathcal{L}_z} \leq C_1 \left(\prod_{l=1}^{q} h_l\right)^{1/2} \|\boldsymbol{\alpha}_1\|_2$, $c_2 \left(\prod_{l=1}^{q} h_l\right)^{1/2} \|\boldsymbol{\alpha}_2\|_2 \leq \|\gamma_2\|_{\mathcal{L}_z} \leq C_2 \left(\prod_{l=1}^{q} h_l\right)^{1/2} \|\boldsymbol{\alpha}_2\|_2$, thus
$c_1 c_2 \left(\prod_{l=1}^{q} h_l\right) \|\boldsymbol{\alpha}_1\|_2 \|\boldsymbol{\alpha}_2\|_2 \leq \|\gamma_1\|_{\mathcal{L}_z} \|\gamma_1\|_{\mathcal{L}_z} \leq C_1 C_2 \left(\prod_{l=1}^{q} h_l\right) \|\boldsymbol{\alpha}_1\|_2 \|\boldsymbol{\alpha}_2\|_2$. Hence

$$\begin{aligned} A_n \leq & \frac{\|\boldsymbol{\alpha}_1\|_2 \|\boldsymbol{\alpha}_2\|_2}{c_1 c_2 \left(\prod_{l=1}^{q} h_l\right) \|\boldsymbol{\alpha}_1\|_2 \|\boldsymbol{\alpha}_2\|_2} \times \\ & \max_{\mathbf{z} \in \mathcal{D}} \max_{j_1,...,j_q,j'_1,...j'_q} \left| \left\langle \mathcal{B}_{j_1,...,j_q}, \mathcal{B}_{j'_1,...,j'_q} \right\rangle_{n,\mathcal{L}_z} - \left\langle \mathcal{B}_{j_1,...,j,q}, \mathcal{B}_{j'_1,...,j'_q} \right\rangle_{\mathcal{L}_z} \right| \\ = & O_{a.s.} \left[ \left\{ n^{-1} \left(\prod_{l=1}^{q} h_l\right)^{-1} \log n \right\}^{1/2} \right]. \end{aligned}$$

∎

Let

$$\widehat{\mathbf{V}}_{\mathbf{z},n} = n^{-1} \mathbf{B}^{\mathrm{T}} \mathcal{L}_z \mathbf{B} = \left\{ \left\langle \mathcal{B}_{j_1,...,j_q}, \mathcal{B}_{j'_1,...,j'_q} \right\rangle_{n,\mathcal{L}_z} \right\}_{\mathbf{K}n \times \mathbf{K}n}, \tag{5.11}$$

$$\mathbf{V}_{\mathbf{z},n} = \left\{ \left\langle \mathcal{B}_{j_1,\ldots,j_q}, \mathcal{B}_{j'_1,\ldots,j'_q} \right\rangle_{\mathcal{L}_z} \right\}_{\mathbf{K}_n \times \mathbf{K}_n}. \tag{5.12}$$

A result in Demko (1986) is described as follows, which plays an essential role in the proof of Lemma 5.5.

LEMMA 5.4. *For positive definite Hermitian matrices* $\mathbf{A}$ *such that* $\mathbf{A}$ *has no more than* $k$ *nonzero entries in each row, then* $\left\|\mathbf{A}^{-1}\right\|_\infty \leq 33\sqrt{k} \left\|\mathbf{A}^{-1}\right\|_2^{5/4} \|\mathbf{A}\|_2^{1/4}$.

LEMMA 5.5. *Under assumptions (A2) and (A4), there exists a constant* $0 < C_m < \infty$ *depending only on* $m_l, 1 \leq l \leq q$, *such that* $\sup_{z\in\mathcal{D}} \left\|\mathbf{V}_{\mathbf{z},n}^{-1}\right\|_\infty \leq C_m \left(\prod_{l=1}^q h_l\right)^{-1}$, *where* $\mathbf{V}_{\mathbf{z},n}$ *is defined in (5.12)*

*Proof.* For any vector $\mathbf{w} \in \mathbb{R}^{\mathbf{K}n}$, $c_B \left(\prod_{l=1}^q h_l\right) \|\mathbf{w}\|_2^2 \leq \mathbf{w}^{\mathrm{T}} \mathbf{V}_{\mathbf{z},n} \mathbf{w} \leq C_B \left(\prod_{l=1}^q h_l\right) \|\mathbf{w}\|_2^2$ which follows directly from (5.9), and it is clear that $\mathbf{V}_{\mathbf{z},n}$ is symmetric, thus $\mathbf{V}_{\mathbf{z},n}$ is a positive definite Hermitian matrix .

$$\begin{aligned}
\left\|\mathbf{V}_{\mathbf{z},n}\right\|_2 &= \sup_{\mathbf{w}} \left\{ \left(\mathbf{V}_{\mathbf{z},n}\mathbf{w}\right)^{\mathrm{T}} \left(\mathbf{V}_{\mathbf{z},n}\mathbf{w}\right) / \|\mathbf{w}\|_2^2 \right\}^{1/2} \\
&\leq \sup_{\mathbf{w}} \left\{ C_B \left(\prod_{l=1}^q h_l\right) \left(\mathbf{V}_{\mathbf{z},n}\mathbf{w}\right)^{\mathrm{T}} \mathbf{V}_{\mathbf{z},n}^{-1} \left(\mathbf{V}_{\mathbf{z},n}\mathbf{w}\right) / \|\mathbf{w}\|_2^2 \right\}^{1/2} \\
&= C_B^{1/2} \left(\prod_{l=1}^q h_l\right)^{1/2} \sup_{\mathbf{w}} \left\{ \mathbf{w}^{\mathrm{T}} \mathbf{V}_{\mathbf{z},n} \mathbf{w} / \|\mathbf{w}\|_2^2 \right\}^{1/2} \leq C_B \left(\prod_{l=1}^q h_l\right)
\end{aligned}$$

Similarly $\left\|\mathbf{V}_{\mathbf{z},n}^{-1}\right\|_2 \leq c_B^{-1} \left(\prod_{l=1}^q h_l\right)^{-1}$. By tensor spline properties, $\mathbf{V}_{\mathbf{z},n}$ has no more than $\prod_{l=1}^q (2m_l - 1)$ nonzero entries in each row, thus $\sup_{z\in\mathcal{D}} \left\|\mathbf{V}_{\mathbf{z},n}^{-1}\right\|_\infty \leq 33\sqrt{\prod_{l=1}^q (2m_l - 1)} \left\|\mathbf{V}_{\mathbf{z},n}^{-1}\right\|_2^{5/4} \left\|\mathbf{V}_{\mathbf{z},n}^{-1}\right\|_2^{1/4} = C_m \left(\prod_{l=1}^q h_l\right)^{-1}$, where $C_m = 33\sqrt{\prod_{l=1}^q (2m_l - 1) c_B C_B^{-1}}$. ∎

LEMMA 5.6. *For* $\widehat{\mathbf{V}}_{\mathbf{z},n}$ *defined in (5.11), and for* $n$ *large enough,* $\sup_{\mathbf{z}\in\mathcal{D}} \left\|\widehat{\mathbf{V}}_{\mathbf{z},n}^{-1}\right\|_\infty \leq 2C_m \left(\prod_{l=1}^q h_l\right)^{-1}$ *in which* $C_m > 0$ *is the constant in Lemma 5.5.*

*Proof.* By Lemma 5.2, as $n \to \infty$,

$$\sup_{\mathbf{z}\in\mathcal{D}} \left\|\widehat{\mathbf{V}}_{\mathbf{z},n} - \mathbf{V}_{\mathbf{z},n}\right\|_\infty =$$

$$\sup_{\mathbf{z}\in\mathcal{D}} \max_{j_1,\ldots,j_q} \sum_{j_1',\ldots j_q'} \left| \left\langle \mathcal{B}_{j_1,\ldots,j_q}, \mathcal{B}_{j_1',\ldots,j_q'} \right\rangle_{n,\mathcal{L}_z} - \left\langle \mathcal{B}_{j_1,\ldots,j_q}, \mathcal{B}_{j_1',\ldots,j_q'} \right\rangle_{\mathcal{L}_z} \right| \leq$$

$$\prod_{l=1}^q (2m_l - 1) \sup_{\mathbf{z}\in\mathcal{D}} \max_{j_1,\ldots,j_q,j_1',\ldots j_q'} \left| \left\langle \mathcal{B}_{j_1,\ldots,j_q}, \mathcal{B}_{j_1',\ldots,j_q'} \right\rangle_{n,\mathcal{L}_z} - \left\langle \mathcal{B}_{j_1,\ldots,j_q}, \mathcal{B}_{j_1',\ldots,j_q'} \right\rangle_{\mathcal{L}_z} \right|$$
$$= O_{a.s.}\left[ \left\{ n^{-1} \left(\prod_{l=1}^q h_l\right) \log n \right\}^{1/2} \right]. \tag{5.13}$$

Let $\boldsymbol{\xi} = \mathbf{V}_{\mathbf{z},n}\boldsymbol{\eta}$, for any given vector $\boldsymbol{\eta}$ with dimension $\mathbf{K}n \times 1$, then for any given $z \in \mathcal{D}$, $\left\|\mathbf{V}_{\mathbf{z},n}^{-1}\boldsymbol{\xi}\right\|_\infty \leq \left\|\mathbf{V}_{\mathbf{z},n}^{-1}\right\|_\infty \|\boldsymbol{\xi}\|_\infty \leq C_m \left(\prod_{l=1}^q h_l\right)^{-1} \|\boldsymbol{\xi}\|_\infty$ by Lemma 5.5, and thus $\left\|\mathbf{V}_{\mathbf{z},n}\boldsymbol{\eta}\right\|_\infty \geq C_m^{-1} \left(\prod_{l=1}^q h_l\right) \|\boldsymbol{\eta}\|_\infty$. By (5.13) and $\left\|\left(\widehat{\mathbf{V}}_{\mathbf{z},n} - \mathbf{V}_{\mathbf{z},n}\right)\boldsymbol{\eta}\right\|_\infty \leq \left\|\widehat{\mathbf{V}}_{\mathbf{z},n} - \mathbf{V}_{\mathbf{z},n}\right\|_\infty \|\boldsymbol{\eta}\|_\infty$, for $n$ large enough $\left\|\widehat{\mathbf{V}}_{\mathbf{z},n}\boldsymbol{\eta}\right\|_\infty \geq (1/2)\, C_m^{-1} \left(\prod_{l=1}^q h_l\right) \|\boldsymbol{\eta}\|_\infty$. Let $\boldsymbol{\xi}_1 = \widehat{\mathbf{V}}_{\mathbf{z},n}\boldsymbol{\eta}$, then we have $\left\|\widehat{\mathbf{V}}_{\mathbf{z},n}^{-1}\boldsymbol{\xi}_1\right\|_\infty \leq 2C_m \left(\prod_{l=1}^q h_l\right)^{-1} \|\boldsymbol{\xi}_1\|_\infty$, for any given $z \in \mathcal{D}$ and $n$ large enough. Therefore the result follows. ∎

LEMMA 5.7. *Under assumptions (A2)-(A4), as* $n \to \infty$,

$$\sup_{\mathbf{z}\in\mathcal{D}} \left\|n^{-1}\mathbf{B}^{\mathrm{T}}\mathcal{L}_z\mathbf{E}\right\|_\infty = O_{a.s.}\left[ \left\{ n^{-1} \left(\prod_{l=1}^q h_l\right) \log n \right\}^{1/2} \right].$$

*Proof.* Let $D_n = n^\vartheta$, with $\vartheta < 1/2$, $\vartheta\,(2+\delta) > 1$, $\vartheta\,(1+\delta) > 1/2$, which are satisfied

by $\delta > 0$. We decompose the noise variable $\varepsilon_i$ into a truncated part and a tail part $\varepsilon_i = \varepsilon_{i,1}^{D_n} + \varepsilon_{i,2}^{D_n} + \varepsilon_{i,3}^{D_n}$, where $\varepsilon_{i,1}^{D_n} = \varepsilon_i I\left(|\varepsilon_i| > D_n\right)$, $\varepsilon_{i,2}^{D_n} = \varepsilon_i I\left(|\varepsilon_i| \leq D_n\right) - \varepsilon_{i,3}^{D_n}$, $\varepsilon_{i,3}^{D_n} = E\left\{\varepsilon_i I\left(|\varepsilon_i| \leq D_n\right) | \mathbf{X}_i, \mathbf{Z}_i\right\}$. Since $\left|\varepsilon_{i,3}^{D_n}\right| \leq \left(E\left|\varepsilon_i\right|^{2+\delta} | \mathbf{X}_i, \mathbf{Z}_i\right) / D_n^{1+\delta} = o\left(n^{-1/2}\right)$, then

$$\sup_{\mathbf{z}\in\mathcal{D}, j_1,\ldots,j_q} \left| n^{-1} \sum_{i=1}^n \mathcal{B}_{j_1,\ldots,j_q}\left(\mathbf{X}_i\right) L\left(\mathbf{Z}_i, \mathbf{z}, \boldsymbol{\lambda}\right) \sigma\left(\mathbf{X}_i, \mathbf{Z}_i\right) \varepsilon_{i,3}^{D_n} \right| = o\left(n^{-1/2}\right).$$

The tail part vanishes almost surely, since $\sum_{n=1}^{\infty} P\left(|\varepsilon_n| > D_n\right) \leq M_\delta \sum_{n=1}^{\infty} n^{-\vartheta(2+\delta)} < \infty$. The Borel Cantelli Lemma implies that

$$\sup_{\mathbf{z}\in\mathcal{D}, j_1,\ldots,j_q} \left| n^{-1} \sum_{i=1}^n \mathcal{B}_{j_1,\ldots,j_q}\left(\mathbf{X}_i\right) L\left(\mathbf{Z}_i, \mathbf{z}, \boldsymbol{\lambda}\right) \sigma\left(\mathbf{X}_i, \mathbf{Z}_i\right) \varepsilon_{i,l}^{D_n} \right| = O\left(n^{-k}\right), \text{ for any } k > 0.$$

For the truncated part, using Bernstein's inequality in Theorem 1.2 of Bosq (1998), one has

$$\begin{aligned} & \sup_{\mathbf{z}\in\mathcal{D}, j_1,\ldots,j_q} \left| n^{-1} \sum_{i=1}^n \mathcal{B}_{j_1,\ldots,j_q}\left(\mathbf{X}_i\right) L\left(\mathbf{Z}_i, \mathbf{z}, \boldsymbol{\lambda}\right) \sigma\left(\mathbf{X}_i, \mathbf{Z}_i\right) \varepsilon_{i,2}^{D_n} \right| \\ = & \left[ \left\{ n^{-1} \left( \prod_{l=1}^q h_l \right) \log n \right\}^{1/2} \right] \end{aligned}$$

as $n \to \infty$. Therefore the result of Lemma 5.7 follows from above. ∎

LEMMA 5.8. *Under assumptions (A2)-(A4), for $\widehat{g}_\varepsilon\left(\mathbf{x}, \mathbf{z}\right)$ in (5.7), as $n \to \infty$,* $\sup_{\mathbf{x}\in[0,1]^q, \mathbf{z}\in\mathcal{D}} \left|\widehat{g}_\varepsilon\left(\mathbf{x}, \mathbf{z}\right)\right| = O_{a.s.}\left[ \left\{ n^{-1} \left( \prod_{l=1}^q h_l \right)^{-1} \log n \right\}^{1/2} \right]$.

*Proof.* By Theorem 5.4.2 in DeVore and Lorentz (1993), Lemma 5.6, Lemma 5.7 and the

definition of $\widehat{g}_{\varepsilon}(x,\mathbf{z})$, one has

$$\begin{aligned}\sup_{\mathbf{z}\in\mathcal{D}}\left\|\widehat{\beta}_{\varepsilon}\right\|_{\infty} &= \sup_{\mathbf{z}\in\mathcal{D}}\left\|\widehat{\mathbf{V}}_{\mathbf{Z},n}^{-1}\left(n^{-1}\mathbf{B}^{\mathrm{T}}\mathcal{L}_{z}\mathbf{E}\right)\right\|_{\infty} \leq \sup_{\mathbf{z}\in\mathcal{D}}\left\|\widehat{\mathbf{V}}_{\mathbf{Z},n}^{-1}\right\|_{\infty}\sup_{\mathbf{z}\in\mathcal{D}}\left\|n^{-1}\mathbf{B}^{\mathrm{T}}\mathcal{L}_{z}\mathbf{E}\right\|_{\infty} \\ &= O_{a.s.}\left[\left\{n^{-1}\left(\prod_{l=1}^{q}h_{l}\right)^{-1}\log n\right\}^{1/2}\right].\end{aligned}$$

$$\sup_{\mathbf{x}\in[0,1]^{q},\mathbf{z}\in\mathcal{D}}\left|\widehat{g}_{\varepsilon}(\mathbf{x},\mathbf{z})\right| \leq \sup_{\mathbf{z}\in\mathcal{D}}\left\|\widehat{\beta}_{\varepsilon}\right\|_{\infty} = O_{a.s.}\left[\left\{n^{-1}\left(\prod_{l=1}^{q}h_{l}\right)^{-1}\log n\right\}^{1/2}\right].$$

∎

Let $\boldsymbol{\Sigma}_{\mathbf{Z}} = \left\{\left\langle \mathcal{B}_{j_1,\ldots,j_q}, \mathcal{B}_{j_1',\ldots,j_q'}\right\rangle_{\mathbf{W}_z}\right\}_{\mathbf{K}n\times\mathbf{K}n}$, where
$\left\langle \mathcal{B}_{j_1,\ldots,j_q}, \mathcal{B}_{j_1',\ldots,j_q'}\right\rangle_{\mathbf{W}_z} = E\left\{\mathcal{B}_{j_1,\ldots,j_q}(\mathbf{X})\,\mathcal{B}_{j_1',\ldots,j_q'}(\mathbf{X})\,L^{2}(\mathbf{Z},\mathbf{z},\boldsymbol{\lambda})\,\sigma^{2}(\mathbf{X},\mathbf{Z})\right\}$, and

$$\begin{aligned}\mathbf{W}_z &= \mathcal{L}_z \mathrm{diag}\left\{\sigma^{2}(\mathbf{X}_1,\mathbf{Z}_1),\ldots,\sigma^{2}(\mathbf{X}_n,\mathbf{Z}_n)\right\}\mathcal{L}_z \\ &= \mathrm{diag}\left\{L^{2}(\mathbf{Z}_1,\mathbf{z},\boldsymbol{\lambda})\,\sigma^{2}(\mathbf{X}_1,\mathbf{Z}_1),\ldots,L^{2}(\mathbf{Z}_1,\mathbf{z},\boldsymbol{\lambda})\,\sigma^{2}(\mathbf{X}_n,\mathbf{Z}_n)\right\}.\end{aligned}$$

For $(\mathbf{x},\mathbf{z})\in[0,1]^{q}\times\mathcal{D}$ define

$$\widehat{\sigma}_{n}^{2}(\mathbf{x},\mathbf{z}) = n^{-1}\mathcal{B}(\mathbf{x})^{\mathrm{T}}\mathbf{V}_{\mathbf{Z},n}^{-1}\boldsymbol{\Sigma}_{\mathbf{Z}}\mathbf{V}_{\mathbf{Z},n}^{-1}\mathcal{B}(\mathbf{x}). \tag{5.14}$$

LEMMA 5.9. *Under assumptions (A2)-(A4), for $\widehat{g}_{\varepsilon}(\mathbf{x},\mathbf{z})$ in (5.7) and $\widehat{\sigma}_{n}^{2}(\mathbf{x},\mathbf{z})$ in (5.14), as $n\to\infty$, $\widehat{\sigma}_{n}^{-1}(\mathbf{x},\mathbf{z})\left\{\widehat{g}_{\varepsilon}(\mathbf{x},\mathbf{z})\right\} \longrightarrow \mathrm{N}(0,1)$. For any given $(\mathbf{x},\mathbf{z})\in[0,1]^{q}\times\mathcal{D}$,*
*$c_{\sigma}^{*}n^{-1}\left(\prod_{l=1}^{q}h_{l}\right)^{-1} \leq \widehat{\sigma}_{n}^{2}(x,\mathbf{z}) \leq C_{\sigma}^{*}n^{-1}\left(\prod_{l=1}^{q}h_{l}\right)^{-1}$, for some constants $0<c_{\sigma}^{*}<C_{\sigma}^{*}<\infty$.*

*Proof.* For any given $\mathbf{z}\in\mathcal{D}$, by the definition of $\widehat{\beta}_{\varepsilon}(\mathbf{z})$ in (5.6), for any $\mathbf{c}\in\mathbb{R}^{Kn}$ with

$\|\mathbf{c}\|_2 = 1$, we can write $\mathbf{c}^{\mathrm{T}}\widehat{\beta}_{\varepsilon}(\mathbf{z}) = \sum_{i=1}^{n} a_i\varepsilon_i$, where

$$a_i^2 = n^{-2}\mathbf{c}^{\mathrm{T}}\widehat{\mathbf{V}}_{\mathbf{z},n}^{-1}\mathcal{B}(\mathbf{X}_i)\,\mathcal{B}(\mathbf{X}_i)^{\mathrm{T}}\,L^2(\mathbf{Z}_i,\mathbf{z},\boldsymbol{\lambda})\,\sigma^2(\mathbf{X}_i)\,\widehat{\mathbf{V}}_{\mathbf{z},n}^{-1}\mathbf{c}.$$

For any given $\mathbf{z} \in \mathcal{D}$, by Lemma 5.6, as $n \to \infty$ with probability 1, $\max_{1\le i\le n} a_i^2 \le C_\sigma^2 (2C_m)^2 n^{-2}\left(\prod_{l=1}^{q} h_l\right)^{-2}$. As $n \to \infty$ with probability 1,

$$\begin{aligned}\sum_{i=1}^{n} a_i^2 &\ge c_\sigma^2 C_B^{-2}\left(\prod_{l=1}^{q} h_l\right)^{-2} n^{-2}\sum_{i=1}^{n}\left\{\mathbf{c}^{\mathrm{T}}\mathcal{B}(\mathbf{X}_i)\right\}^2 L^2(\mathbf{Z}_i,\mathbf{z},\boldsymbol{\lambda})\\ &\ge c_\sigma^2 C_B^{-2}\left(\prod_{l=1}^{q} h_l\right)^{-2} n^{-1} E\left[\left\{\mathbf{c}^{\mathrm{T}}\mathcal{B}(\mathbf{X}_i)\right\}^2 L^2(\mathbf{Z}_i,\mathbf{z},\boldsymbol{\lambda})\right](1-A_n),\end{aligned}$$

the proof of which follows the same pattern as in Lemmas 5.1 and 5.3, thus as $n \to \infty$ with probability 1, $\sum_{i=1}^{n} a_i^2 \ge C' n^{-1}\left(\prod_{l=1}^{q} h_l\right)^{-1}$. Hence for any given $\mathbf{z} \in \mathcal{D}$, $\max\limits_{1\le i\le n} a_i^2 / \sum_{i=1}^{n} a_i^2 = O_{a.s.}\left\{n^{-1}\left(\prod_{l=1}^{q} h_l\right)^{-1}\right\} = O_{a.s.}(1)$. Given any $\xi > 0$, one has

$$\begin{aligned}&\lim_{n\to\infty}\|\mathbf{a}\|_2^{-2}\sum_{k=1}^{n} a_k^2 E\left\{\varepsilon^2\mathbf{I}\left(|a_k\varepsilon| > \xi\|\mathbf{a}\|_2\right)\right\}\\ &\le \lim_{n\to\infty}\|\mathbf{a}\|_2^{-2}\sum_{k=1}^{n} a_k^2\left(E|\varepsilon|^{2+\delta}\right)^{2/(2+\delta)}\left\{P\left(|\varepsilon| > \xi a_k^{-1}\|\mathbf{a}\|_2\right)\right\}^{\delta/(2+\delta)}\\ &\le M_\delta^{2/(2+\delta)}\lim_{n\to\infty}\max_{1\le k\le n}\left\{P\left(|\varepsilon| > \xi a_k^{-1}\|\mathbf{a}\|_2\right)\right\}^{\delta/(2+\delta)} = 0,\end{aligned}$$

thus, the Lindeberg condition is satisfied. By the Lindeberg-Feller CLT, as $n \to \infty$, $\sum_{i=1}^{n} a_i\varepsilon_i / \left(\sum_{i=1}^{n} a_i^2\right)^{-1/2} \to N(0,1)$. Therefore, $[\mathrm{Var}\{\widehat{g}_\varepsilon(\mathbf{x},\mathbf{z})\,|\mathbf{X},\mathbf{Z}\}]^{-1/2}\,\widehat{g}_\varepsilon(\mathbf{x},\mathbf{z}) \to N(0,1)$.

$$\mathrm{Var}\{\widehat{g}_\varepsilon(\mathbf{x},\mathbf{z})\,|\mathbf{X},\mathbf{Z}\} = n^{-1}\mathcal{B}(\mathbf{x})^{\mathrm{T}}\mathbf{V}_{n,z}^{-1}\boldsymbol{\Sigma}_{n,z}\mathbf{V}_{n,z}^{-1}\mathcal{B}(\mathbf{x})^{\mathrm{T}},$$

where

$$\begin{aligned}&\left\langle \mathcal{B}_{j_1,...,j_q}, \mathcal{B}_{j_1',...,j_q'} \right\rangle_{\mathbf{W}_z,n} \\ = \;& n^{-1} \sum_{i=1}^{n} \mathcal{B}_{j_1,...,j_q}(\mathbf{X}_i)\, \mathcal{B}_{j_1',...,j_q'}(\mathbf{X}_i)\, L^2(\mathbf{Z}_i, \mathbf{z}, \boldsymbol{\lambda})\, \sigma^2(\mathbf{X}_i, \mathbf{Z}_i)\end{aligned}$$

. By Lemma 5.3, one can prove that

$$\mathbf{w}^{\mathrm{T}} \boldsymbol{\Sigma}_z \mathbf{w} (1 - A_n) \leq \mathbf{w}^{\mathrm{T}} \boldsymbol{\Sigma}_{n,z} \mathbf{w} \leq \mathbf{w}^{\mathrm{T}} \boldsymbol{\Sigma}_z \mathbf{w} (1 + A_n),$$

and $c_\Sigma \left( \prod_{l=1}^{q} h_l \right) \|\mathbf{w}\|^2 \leq \mathbf{w}^{\mathrm{T}} \boldsymbol{\Sigma}_z \mathbf{w} \leq C_\Sigma \left( \prod_{l=1}^{q} h_l \right) \|\mathbf{w}\|^2$, thus

$$\begin{aligned}\mathbf{w}^{\mathrm{T}} \mathbf{V}_{\mathbf{z},n}^{-1} \boldsymbol{\Sigma}_z \mathbf{V}_{\mathbf{z},n}^{-1} \mathbf{w} &\leq C_{\boldsymbol{\Sigma}} \left( \prod_{l=1}^{q} h_l \right) \mathbf{w}^{\mathrm{T}} \mathbf{V}_{\mathbf{z},n}^{-1} \mathbf{V}_{\mathbf{z},n}^{-1} \mathbf{w} \leq C_{\boldsymbol{\Sigma}} c \left( \prod_{l=1}^{q} h_l \right)^{-1} \|\mathbf{w}\|^2, \\ \mathbf{w}^{\mathrm{T}} \mathbf{V}_{\mathbf{z},n}^{-1} \boldsymbol{\Sigma}_z \mathbf{V}_{\mathbf{z},n}^{-1} \mathbf{w} &\geq c_{\boldsymbol{\Sigma}} \left( \prod_{l=1}^{q} h_l \right) \mathbf{w}^{\mathrm{T}} \mathbf{V}_{\mathbf{z},n}^{-1} \mathbf{V}_{\mathbf{z},n}^{-1} \mathbf{w} \geq c_{\boldsymbol{\Sigma}} C_{\mathbf{V}}^{-\mathbf{2}} \left( \prod_{l=1}^{q} h_l \right)^{-1} \|\mathbf{w}\|^2.\end{aligned}$$

For any given $(\mathbf{x}, \mathbf{z}) \in [0,1]^q \times \mathcal{D}$, $\widehat{\sigma}_n^2(\mathbf{x}, \mathbf{z}) \leq n^{-1} C_{\boldsymbol{\Sigma}} c_{\mathbf{V}}^{-\mathbf{2}} \left( \prod_{l=1}^{q} h_l \right)^{-1} \left\| \mathcal{B}_{j_1,...,j_q}(\mathbf{x}) \right\|_2^2 \leq C_\sigma^* n^{-1} \left( \prod_{l=1}^{q} h_l \right)^{-1}$, where $\widehat{\sigma}_n^2(\mathbf{x}, \mathbf{z})$ is defined in (5.14), and similarly one has $\widehat{\sigma}_n^2(\mathbf{x}, \mathbf{z}) \geq c_\sigma^* n^{-1} \left( \prod_{l=1}^{q} h_l \right)^{-1}$, for some constants $0 < c_\sigma^* < C_\sigma^* < \infty$. For any given $(\mathbf{x}, \mathbf{z}) \in [0,1]^q \times \mathcal{D}$,

$\lim_{n\to\infty} \left[ \left[ \mathrm{Var} \left\{ \widehat{g}_\varepsilon(\mathbf{x}, \mathbf{z}) | \mathbf{X}, \mathbf{Z} \right\} \right]^{1/2} \left\{ \widehat{\sigma} n(\mathbf{x}, \mathbf{z}) \right\}^{-1} \right] = 1$, since

$$\begin{aligned}\mathrm{Var} \left\{ \widehat{g}_\varepsilon(\mathbf{x}, \mathbf{z}) | \mathbf{X}, \mathbf{Z} \right\} &\leq \mathcal{B}(\mathbf{x})^{\mathrm{T}} \mathbf{V}_{\mathbf{z},n}^{-1} \boldsymbol{\Sigma}_z \mathbf{V}_{\mathbf{z},n}^{-1} \mathcal{B}(\mathbf{x}) (1 + A_n) (1 - A_n)^{-2}, \\ \mathrm{Var} \left\{ \widehat{g}_\varepsilon(\mathbf{x}, \mathbf{z}) | \mathbf{X}, \mathbf{Z} \right\} &\geq \mathcal{B}(\mathbf{x})^{\mathrm{T}} \mathbf{V}_{\mathbf{z},n}^{-1} \boldsymbol{\Sigma}_z \mathbf{V}_{\mathbf{z},n}^{-1} \mathcal{B}(\mathbf{x}) (1 - A_n) (1 + A_n)^{-2}.\end{aligned}$$

Thus, the result in Lemma 5.9 follows. ∎

LEMMA 5.10. *Under assumptions (A1), (A2) and (A4), for* $\widehat{g}_g(\mathbf{x},\mathbf{z})$ *in (5.7), as* $n\to\infty$, $\sup_{\mathbf{x}\in[0,1]^q,\mathbf{z}\in\mathcal{D}}\left|\widehat{g}_g(\mathbf{x},\mathbf{z})-g(\mathbf{x},\mathbf{z})\right| = O_{a.s.}\left(\sum_{l=1}^{q}h_l^p+\sum_{s=1}^{r}\lambda_s\right)$.

*Proof.* For $1\le i\le n$, $1\le s\le r$, let $\mathbf{Z}_{-is}$ be the leave-one out vector of $\mathbf{Z}_i$, then

$$L(\mathbf{Z}_i,\mathbf{z},\boldsymbol{\lambda}) = \prod_{s=1}^{r}\lambda_s^{1\left(\mathbf{Z}_{is}\neq z_s\right)} = \mathbf{1}(\mathbf{Z}_i=\mathbf{z}) + \sum_{s=1}^{r}\lambda_s\mathbf{1}(\mathbf{Z}_{is}\neq z_s,\mathbf{Z}_{-is}=\mathbf{z}_{-is}) + o\left(\sum_{s=1}^{r}\lambda_s\right).$$

Denote $\mathcal{L}_z=\mathcal{L}_{z,1}+\mathcal{L}_{z,2}+\mathcal{L}_{z,3}$, where $\mathcal{L}_{z,1}=\operatorname{diag}\{\mathbf{1}(\mathbf{Z}_1=\mathbf{z}),\ldots,\mathbf{1}(\mathbf{Z}_n=\mathbf{z})\}$, $\mathcal{L}_{z,2}=\operatorname{diag}\left\{\sum_{s=1}^{r}\lambda_s\mathbf{1}(\mathbf{Z}_{1s}\neq z_s,\mathbf{Z}_{-1s}=\mathbf{z}_{-1s}),\ldots,\sum_{s=1}^{r}\lambda_s\mathbf{1}(\mathbf{Z}_{ns}\neq z_s,\mathbf{Z}_{-ns}=\mathbf{z}_{-ns})\right\}$ and $\mathcal{L}_{z,3}=o\left(\sum_{s=1}^{r}\lambda_s\right)\mathbf{I}_n$. Thus by the definition of $\widehat{g}_g(\mathbf{x},\mathbf{z})$ in (5.7),

$$\widehat{g}_g(\mathbf{x},\mathbf{z})-g(\mathbf{x},\mathbf{z}) = \mathcal{B}(\mathbf{x})^{\mathrm{T}}\widehat{\mathbf{V}}_{\mathbf{z},n}^{-1}\left(n^{-1}\mathbf{B}^{\mathrm{T}}\mathcal{L}_z\mathbf{g}\right)-g(\mathbf{x},\mathbf{z}) = \Pi_1+\Pi_2+\Pi_3,$$

where $\Pi_1=\mathcal{B}(\mathbf{x})^{\mathrm{T}}\widehat{\mathbf{V}}_{\mathbf{z},n}^{-1}\left(n^{-1}\mathbf{B}^{\mathrm{T}}\mathcal{L}_{z,1}\mathbf{g}\right)-g(\mathbf{x},\mathbf{z})$,

$$\Pi_2=\mathcal{B}(\mathbf{x})^{\mathrm{T}}\widehat{\mathbf{V}}_{\mathbf{z},n}^{-1}\left(n^{-1}\mathbf{B}^{\mathrm{T}}\mathcal{L}_{z,2}\mathbf{g}\right), \Pi_3=\mathcal{B}(\mathbf{x})^{\mathrm{T}}\widehat{\mathbf{V}}_{\mathbf{z},n}^{-1}\left(n^{-1}\mathbf{B}_m^{\mathrm{T}}\mathcal{L}_{z,3}\mathbf{g}\right).$$

By Theorem 12.8 and (13.69) on p. 149 of de Boor (2001), for any $\mathbf{z}\in\mathcal{D}$, there exists $\beta(\mathbf{z})\in\mathbb{R}^{Kn}$, such that $\sup_{\mathbf{x}\in[0,1]^q}\left|\mathcal{B}(\mathbf{x})^{\mathrm{T}}\beta(\mathbf{z})-g(\mathbf{x},\mathbf{z})\right| = O\left(\sum_{l=1}^{q}h_l^p\right)$. Let $\mathbf{g}_\mathbf{z}=\{g(\mathbf{X}_1,\mathbf{z}),\ldots,g(\mathbf{X}_n,\mathbf{z})\}^{\mathrm{T}}$, $\Pi_1=\mathcal{B}(\mathbf{x})^{\mathrm{T}}\widehat{\mathbf{V}}_{\mathbf{z},n}^{-1}\left(n^{-1}\mathbf{B}^{\mathrm{T}}\mathcal{L}_{z,1}\mathbf{g}_\mathbf{z}\right)-g(\mathbf{x},\mathbf{z})=\Pi_{11}+\Pi_{12}$, where $\Pi_{11}=\mathcal{B}(\mathbf{x})^{\mathrm{T}}\widehat{\mathbf{V}}_{\mathbf{z},n}^{-1}\left[n^{-1}\mathbf{B}^{\mathrm{T}}\mathcal{L}_{z,1}\{\mathbf{g}_\mathbf{z}-\mathbf{B}\beta(\mathbf{z})\}\right]$,

$$\Pi_{12}=\mathcal{B}(\mathbf{x})^{\mathrm{T}}\widehat{\mathbf{V}}_{\mathbf{z},n}^{-1}\left\{n^{-1}\mathbf{B}^{\mathrm{T}}\mathcal{L}_{z,1}\mathbf{B}\beta(\mathbf{z})\right\}-g(\mathbf{x},\mathbf{z}).$$

$$\sup_{\mathbf{z}\in\mathcal{D}}\left\|\widehat{\mathbf{V}}_{\mathbf{z},n}^{-1}\left[n^{-1}\mathbf{B}^{\mathrm{T}}\mathcal{L}_{z,1}\left\{\mathbf{g}_{\mathbf{Z}}-\mathbf{B}\beta\left(\mathbf{z}\right)\right\}\right]\right\|_{\infty}$$

$$\leq \sup_{\mathbf{z}\in\mathcal{D}}\left\|\widehat{\mathbf{V}}_{\mathbf{z},n}^{-1}\right\|_{\infty}\sup_{\mathbf{z}\in\mathcal{D}}\left\|n^{-1}\mathbf{B}^{\mathrm{T}}\mathcal{L}_{z,1}\left\{\mathbf{g}_{\mathbf{Z}}-\mathbf{B}\beta\left(\mathbf{z}\right)\right\}\right\|_{\infty}\leq$$

$$\left(\sup_{\mathbf{z}\in\mathcal{D}}\left\|\widehat{\mathbf{V}}_{\mathbf{z},n}^{-1}\right\|_{\infty}\right)\left\{\sup_{\mathbf{z}\in\mathcal{D}}\left\|\mathbf{g}_{\mathbf{Z}}-\mathbf{B}\beta\left(\mathbf{z}\right)\right\|_{\infty}\right\}\left(\left\|n^{-1}\mathbf{B}^{\mathrm{T}}\mathcal{L}_{z,1}\right\|_{\infty}\right)=O_{a.s.}\left(\sum_{l=1}^{q}h_{l}^{p}\right),$$

$$\sup_{\mathbf{x}\in[0,1]^{q},\mathbf{z}\in\mathcal{D}}\left|\Pi_{11}\right|\leq\sup_{\mathbf{z}\in\mathcal{D}}\left\|\widehat{\mathbf{V}}_{\mathbf{z},n}^{-1}\left[n^{-1}\mathbf{B}^{\mathrm{T}}\left\{\mathbf{g}_{\mathbf{Z}}-\mathbf{B}\beta\left(\mathbf{z}\right)\right\}\right]\right\|_{\infty}=O_{a.s.}\left(\sum_{l=1}^{q}h_{l}^{p}\right),$$

by Theorem 5.4.2 in Devore and Lorentz (1993), Lemma 5.6 and properties of the B-spline.

$$\sup_{\mathbf{x}\in[0,1]^{q},\mathbf{z}\in\mathcal{D}}\left|\Pi_{12}\right|\leq\sup_{\mathbf{x}\in[0,1]^{q},\mathbf{z}\in\mathcal{D}}\left|\mathcal{B}\left(\mathbf{x}\right)^{\mathrm{T}}\beta\left(\mathbf{z}\right)-g\left(\mathbf{x},\mathbf{z}\right)\right|+$$

$$\sup_{\mathbf{x}\in[0,1]^{q},\mathbf{z}\in\mathcal{D}}\left|\mathcal{B}\left(\mathbf{x}\right)^{\mathrm{T}}\widehat{\mathbf{V}}_{\mathbf{z},n}^{-1}\left\{n^{-1}\mathbf{B}^{\mathrm{T}}\left(\mathcal{L}_{z,2}+\mathcal{L}_{z,3}\right)\mathbf{B}\beta\left(\mathbf{z}\right)\right\}\right|$$

$$\leq O\left(\sum_{l=1}^{q}h_{l}^{p}\right)+\sup_{\mathbf{z}\in\mathcal{D}}\left\|\widehat{\mathbf{V}}_{\mathbf{z},n}^{-1}\right\|_{\infty}\sup_{\mathbf{z}\in\mathcal{D}}\left\|n^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{B}\beta\left(\mathbf{z}\right)\right\|_{\infty}O\left(\sum_{s=1}^{r}\lambda_{s}\right).$$

By Lemmas 5.1 and 5.2 and , one has $\sup_{\mathbf{z}\in\mathcal{D}}\left\|n^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{B}\right\|_{\infty}=O_{a.s.}\left(\prod_{l=1}^{q}h_{l}\right)$ , then

$$\sup_{\mathbf{x}\in[0,1]^{q},\mathbf{z}\in\mathcal{D}}\left|\Pi_{12}\right|\leq O\left(\sum_{l=1}^{q}h_{l}^{p}\right)+$$

$$\sup_{\mathbf{z}\in\mathcal{D}}\left\|\widehat{\mathbf{V}}_{\mathbf{z},n}^{-1}\right\|_{\infty}\sup_{\mathbf{z}\in\mathcal{D}}\left\|n^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{B}\right\|_{\infty}\sup_{\mathbf{z}\in\mathcal{D}}\left\|\beta\left(\mathbf{z}\right)\right\|_{\infty}O\left(\sum_{s=1}^{r}\lambda_{s}\right)$$

$$=O_{a.s.}\left(\sum_{l=1}^{q}h_{l}^{p}+\sum_{s=1}^{r}\lambda_{s}\right).$$

$$\sup_{\mathbf{x}\in[0,1]^{q},\mathbf{z}\in\mathcal{D}}\left|\Pi_{1}\right|\leq\sup_{\mathbf{x}\in[0,1]^{q},\mathbf{z}\in\mathcal{D}}\left(\left|\Pi_{11}\right|+\left|\Pi_{12}\right|\right)=O_{a.s.}\left(\sum_{l=1}^{q}h_{l}^{p}+\sum_{s=1}^{r}\lambda_{s}\right).$$

$$\sup_{\mathbf{z}\in\mathcal{D}}\left\|\mathbf{V}_{n,z}^{-1}\left(n^{-1}\mathbf{B}^{\mathrm{T}}\mathcal{L}_{z,2}\mathbf{g}\right)\right\|_{\infty}\leq\sup_{\mathbf{z}\in\mathcal{D}}\left\|\mathbf{V}_{n,z}^{-1}\right\|_{\infty}\sup_{\mathbf{z}\in\mathcal{D}}\left\|n^{-1}\mathbf{B}^{\mathrm{T}}\mathcal{L}_{z,2}\mathbf{g}\right\|_{\infty}$$

$$\leq \sup_{\mathbf{z}\in\mathcal{D}} \left\|\mathbf{V}_{n,z}^{-1}\right\|_\infty \sup_{\mathbf{z}\in\mathcal{D}} \left\|n^{-1}\mathbf{B}^{\mathrm{T}}\right\|_\infty \sup_{\mathbf{z}\in\mathcal{D}} \|\mathbf{g}\|_\infty O\left(\sum_{s=1}^{r}\lambda_s\right) = O_{a.s.}\left(\sum_{s=1}^{r}\lambda_s\right),$$

thus, $\sup_{\mathbf{x}\in[0,1]^q,\mathbf{z}\in\mathcal{D}} \left|\widehat{g}_g(\mathbf{x},\mathbf{z}) - g(\mathbf{x},\mathbf{z})\right| \leq \sup_{\mathbf{x}\in[0,1]^q,\mathbf{z}\in\mathcal{D}} (|\Pi_1| + |\Pi_2| + |\Pi_3|)$

$= O_{a.s.}\left(\sum_{l=1}^{q} h_l^p + \sum_{s=1}^{r}\lambda_s\right)$. ∎

*Proofs of Theorems 5.1 and 5.2.* Theorem 5.1 follows from Lemmas 5.8 and 5.10 directly, while Theorem 5.2 follows from Lemmas 5.9 and 5.10 and assumption (A4) directly. ∎

# BIBLIOGRAPHY

# BIBLIOGRAPHY


[1] Aitchison, J. and Aitken, C. G. G. (1976). Multivariate binary discrimination by the kernel method. *Biometrika*, 63, 413–420.

[2] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* , 57, 289–300.

[3] Bickel, P. J. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. *Annals of Statistics*, 1, 1071–1095.

[4] Bosq, D. (1998). *Nonparametric Statistics for Stochastic Processes*. Springer-Verlag, New York.

[5] Breiman, L. & Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80, 580–619.

[6] Cai, Z., Fan, J. and Li, R. (2000). Efficient estimation and inferences for varying-coefficient models. *Journal of American Statistical Association*, 95, 888–902.

[7] Caspi, A. and Moffitt, T. E. (2006). Gene-environment interactions in psychiatry: joining forces with neuroscience. *Nature Reviews Neuroscience*, 7, 583–590.

[8] Caspi, A., Sugden, K., Moffitt, T. E., Taylor, A., Craig, I. W., Harrington, H., McClay, J., Mill, J., Martin, J., Braithwaite,A. and Poulton, R. (2003). Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene. *Science*, 301, 386–389.

[9] Cardot, H., Ferraty, F., and Sarda, P. (2003). Spline estimators for the functional linear model. *Statistica Sinica*, 13, 571–591.

[10] Cardot, H. and Sarda, P. (2005). Estimation in generalized linear models for functional data via penalized likelihood. *Journal of Multivariate Analysis*, 92, 24–41.

[11] Chatterjee, N. and Carroll, R. J. (2005). Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika*, 92, 399–418.

[12] Chiang, C. T., Rice, J. A. and Wu, C. O. (2001). Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables. *Journal of American Statistical Association*, 96, 605–619.

[13] Claeskens, G., and Van Keilegom, I. (2003). Bootstrap confidence bands for regression curves and their derivatives. *The Annals of Statistics*, 31, 1852–1884.

[14] Cleveland, W. S., Grosse, E. and Shyu, W. M. (1991). *Local regression models.* In Statistical Models in S, J. M. Chambers and T. J. Hastie, 309–376. Pacific Grove: Wadsworth & Brooks.

[15] Cockerham, C. C. (1954). An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. *Genetics*, 39, 859–882.

[16] Costa, L. G. and Eaton, D. L. (2006) *Gene-Environment Interactions: Fundamentals of Ecogenetics*, Hoboken, NJ: John Wiley & Sons.

[17] Csőrgő, M. and Révész, P. (1981). *Strong Approximations in Probability and Statistics.* Academic Press, New York-London.

[18] de Boor, C. (2001). *A Practical Guide to Splines.* Springer-Verlag, New York.

[19] Demko, S. (1986). Spectral bounds for $\left\|A^{-1}\right\|_{\infty}$. *Journal of Approximation Theory*, 48, 207–212.

[20] DeVore, R. A. and Lorentz, G. G. (1993). *Constructive Approximation*, Springer-Verlag, New York.

[21] Falconer, D. S. (1952). The problem of environment and selection. *Natural American*, 86, 293–298.

[22] Fan, J. (1993). Local linear regression smoothers and their minimax efficiency. *The Annals of Statistics*, 21, 196–216.

[23] Fan, J. Q., and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications.* Chapman and Hall, London.

[24] Fan, J. and Zhang, W. (1999). Statistical estimation in varying coefficient models. *The Annals of Statistics*, 27, 1491–1518.

[25] Fan, J. and Zhang, W. Y. (2000). Simultaneous confidence bands and hypothesis testing in varying-coefficient models. *Scandinavian Journal of Statistics*. 27, 715–731.

[26] Fan, J. and Zhang, W. (2008). Statistical methods with varying coefficient models. *Statistics and its Interface*, 1, 179–195.

[27] Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice.* Springer; Berlin.

[28] Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. *The Annals of Statistics*, 19, 1–67.

[29] Gijbels, I., Lambert, A., and Qiu, P. H. (2007). Jump-preserving regression and smoothing using local linear fitting: a compromise. *Annals of the Institute of Statistical Mathematics*, 59, 235–272.

[30] González-Manteiga, W., Martínez-Miranda, MD. & Raya-Miranda, R. (2008). SiZer Map for inference with additive models. *Statistics and Computing*, 18, 297–312.

[31] Guo, S. W. (2000). Gene-environment interaction and the mapping of complex traits: some statistical models and their implications. *Human Heredity*, 50, 286–303.

[32] Hall, P. and Heckman, N. (2002). Estimating and depicting the structure of a distribution of random functions. *Biometrika*, 89, 145–158.

[33] Hall, P., Müller, H. G., and Wang, J. L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *Annals of Statistics*, 34, 1493–1517.

[34] Hall, P., and Patil, P. (1995). Formulae for mean integrated squared error of nonlinear wavelet-based density estimators. *Annals of Statistics*, 23, 905–928.

[35] Hall, P. and Titterington, D. M. (1988). On confidence bands in nonparametric density estimation and regression. *Journal of Multivariate Analysis*, 27, 228–254.

[36] Härdle, W. (1989). Asymptotic maximal deviation of M-smoothers. *Journal of Multivariate Analysis*, 29, 163–179.

[37] Härdle, W., Hlávka, Z., and Klinke, S. (2000). *XploRe Application Guide*, Springer-Verlag, Berlin.

[38] Härdle, W. and Marron, J. S. (1991). Bootstrap simultaneous error bars for nonparametric regression. *Annals of Statistics*, 19, 778-796.

[39] Härdle, W. and Mammen, E. (1993) Comparing nonparmetric versus parametric regression fits. *Annals of Statistics*, 21, 1926–1947.

[40] Harrison, D. & Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for cleaning air. *Journal of Economics and Management*, 5, 81–102.

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

[42] Hastie, T. J. and Tibshirani, R. J. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society, Series B*, 55, 757–796.

[43] Hahn, L. W., Ritchie, M. D., Moore, J. H. (2003). Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics*, 19, 376–382.

[44] Hoover, D., Rice, J., Wu, C. O. and Yang, L. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, 85, 809–822.

[45] Huang, J. Z. (1998). Projection estimation in multiple regression with application to functional ANOVA models. *Annals of Statistics*, 26, 242–272.

[46] Huang, J. Z. (2003). Local asymptotics for polynomial spline regression. *The Annals of Statistics*, 31, 1600–1635.

[47] Huang, J. Z. & Yang, L. (2004). Identification of nonlinear additive autoregression models. *Journal of the Royal Statistical Society, Series B*, 66, 463–477.

[48] Huang, X., Wang, L., Yang, L., and Kravchenko, A. N. (2008). Management practice effects on relationships of grain yields with topography and precipitation. *Agronomy Journal*, 100, 1463–1471.

[49] Huang, J., Wu, C. and Zhou, L. (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica*, 14, 763–788.

[50] Izem, R. and Marron, J. S. (2007). Analysis of nonlinear modes of variation for functional data. *Electronic Journal of Statistics*, 1, 641–676.

[51] James, G. M., Hastie, T., and Sugar, C. (2000). Principal component models for sparse functional data. *Biometrika*, 87, 587–602.

[52] James, G. M. (2002). Generalized linear models with functional predictors. *Journal of the Royal Statistical Society B*, 64, 411–432.

[53] James, G. M. and Silverman, B. W. (2005). Functional adaptive model estimation. *Journal of the American Statistical Association*, 100, 565–576.

[54] James, G. M. and Sugar, C. A. (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association*, 98, 397–408.

[55] Joo, J., and Qiu, P. (2009). Jump detection in a regression curve and its derivative. *Technometrics*, 51, 289–305.

[56] Kang, K. H., Koo, J. Y., and Park, C. W. (2000). Kernel estimation of discontinuous regression function. *Statistics and Probability Letters*, 47, 277–285.

[57] Kılıç, E. (2008). Explicit formula for the inverse of a tridiagonal matrix by backward continued fractions. *Applied Mathematics and Computation*, 197, 345–357.

[58] Koo, J. Y. (1997). Spline estimation of discontinuous regression functions. *Journal of Computational and Graphical Statistics*, 6, 266–284.

[59] Kraft, P., Yen, Y. C., Stram, D. O., Morrison, J. and Gauderman, W. J. (2007). Exploiting gene-environment interaction to detect genetic associations. *Human Heredity*, 63, 111–119.

[60] Lamb, K., and Rizzino, A. (1998). Effects of differentiation on the transcriptional regulation of the FGF-4 gene: Critical roles played by a distal enhancer. *Molecular Reproduction and Development*, 51, 218–224.

[61] Leadbetter, M. R., Lindgren, G., and Rootzén, H. (1983), *Extremes and Related Properties of Random Sequences and Processes*, Springer-Verlag, New York.

[62] Li, S.Y., Lu, Q., Fu, W., Romero, R., and Cui, Y. (2009). A regularized regression approach for dissecting genetic conflicts that increase disease risk in pregnancy. *Statistical Applications in Genetics and Molecular Biology* Vol. 8, Iss. 1, Article 45.

[63] Li, Q. (2000). Efficient estimation of additive partially linear models. *International Economic Review*, 41, 1073–1092.

[64] Li, Q. and J. S. Racine (2004). Cross-validated local linear nonparametric regression. *Statistica Sinica*, 14, 485–512.

[65] Li, Y. and Hsing, T. (2007). On rates of convergence in functional linear regression. *Journal of Multivariate Analysis*, 98, 1782–1804.

[66] Li, Y. and Hsing, T. (2010). Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *Annals of Statistics*, 6, 3321–3351.

[67] Liang, H. (2006). Estimation in partially linear models and numerical comparisons. *Computational Statistics & Data Analysis*, 50, 675–687.

[68] Liang, H., Wang, S. & Carroll, R. (2007). Partially linear models with missing response variables and error-prone covariates. *Biometrika*, 94, 185–198.

[69] Liang, H., Thurston, S., Ruppert, D., Apanasovich, T. & Hauser, R. (2008). Additive partial linear models with measurement errors. *Biometrika*, 95, 667–678.

[70] Li, Q. and J. Racine (2007). *Nonparametric Econometrics: Theory and Practice*, Princeton University Press.

[71] Liu, R. and Yang, L. (2010). Spline-backfitted kernel smoothing of additive coefficient model. *Econometric Theory*, 26, 29–59.

[72] Ma, S. and Yang, L. (2011a). A Jump-detecting Procedure based on Polynomial Spline Estimation. *Journal of Nonparametric Statistics*, 23, 67-81.

[73] Ma, S. and Yang, L. (2011b). Spline-backfitted Kernel Smoothing of Partially Linear Additive Model. *Journal of Statistical Planning and Inference*, 141, 204-219.

[74] Ma, S., Yang, L. and Carroll, R. (2011). A Simultaneous Confidence Band for Sparse Longitudinal Regression. *Statistica Sinica*, accepted.

[75] Ma, S. and Racine, J. (2011). Additive Regression Splines With Irrelevant Regressors. *Manuscript*.

[76] Ma, S., Racine, J. and Yang, L. (2011). Spline Regression in the Presence of Categorical Predictors. *Manuscript*.

[77] Mack, Y. P., and Silverman, B. W. (1982). Weak and strong uniform consistency of kernel regression estimates. *Wahrscheinlichkeitstheorie verm. Gebiete*, 61, 405–415.

[78] Maity, A., Carroll, R. J., Mammen, E. and Chatterjee, N. (2009). Testing in semi-parametric models with interaction, with applications to gene-environment interactions. *Journal of the Royal Statistical Society B*, 71, 75–96.

[79] Morris, J. S. and Carroll, R. J. (2006). Wavelet-based functional mixed models. *Journal of the Royal Statistical Society B*, 68 , 179–199.

[80] Müller, H. G. (1992). Change-points in nonparametric regression analysis.*The Annals of Statistics*, 20, 737–761.

[81] Müller, H. G. and Stadtmüller, U. (2005). Generalized functional linear models. *Annals of Statistics*, 33, 774–805.

[82] Müller, H. G., Stadtmüller, U., and Yao, F. (2006). Functional variance processes. *Journal of the American Statistical Association*, 101, 1007–1018.

[83] Müller, H. G., and Song, K. S. (1997). Two-stage change-point estimators in smooth regression models. *Statistics & Probability Letters*, 34, 323–335.

[84] Müller, H. G. and Yao, F. (2008). Functional additive models. *Journal of American Statistical Association*, 103, 1534–1544.

[85] Nácher, V., Ojeda, S., Cadarso-Suárez, C., Roca-Pardiñas, J. & Acuña, C. (2006). Neural correlates of memory retrieval in the prefrontal cortex. *European Journal of Neuroscience*, 24, 925–936.

[86] Abramson, M.A., Audet, C., Couture, G., Dennis Jr., J.E. and Le Digabel, S. (2011). The NOMAD project. *"Software available at http://www.gerad.ca/nomad".*

[87] Park, C., and Kim, W. (2004). Estimation of a regression function with a sharp change point using boundary wavelets. *Statistics & Probability Letters*, 66, 435–448.

[88] Park, C., and Kim, W. (2006). Wavelet estimation of a regression function with a sharp change point in a random design. *Journal of Statistical Planning and Inference*, 136, 2381–2394.

[89] Peacock, M., Turner, C. H., Econs, M. J. and Foroud, T. (2002). Genetics of osteoporosis. *Endocrine Reviews*, 23, 303–326.

[90] Qiu, P. H., Asano, C., and Li, X. (1991). Estimation of jump regression function. *Bulletin of Informatics and Cybernetics*, 24, 197–212.

[91] Qiu, P. H. (1994). Estimation of the number of jumps of the jump regression functions. *Communications in Statistics-Theory and Methods*, 23, 2141–2155.

[92] Qiu, P. H., and Yandell, B. (1998). A local polynomial jump detection algorithm in nonparametric regression. *Technometrics*, 40, 141–152.

[93] Qiu, P. H. (2003). A jump-preserving curve fitting procedure based on local piecewise-linear kernel estimation. *Journal of Nonparametric Statistics*, 15, 437–453.

[94] Qiu, P. H. (2005). *Image Processing and Jump Regression Analysis*, John Wiley & Sons, New York.

[95] Qiu, P. H. (2007). Jump surface estimation, edge detection, and image restoration. *Journal of the American Statistical Association*, 102, 745–756.

[96] Racine, J. S. and Q. Li (2004). Nonparametric estimation of regression functions with both categorical and continuous Data. *Journal of Econometrics*, 119, 99–130.

[97] Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis, Second Edition.* Springer; New York.

[98] 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. *Statistics in Medicine*, 25, 2483–2501.

[99] Schimek, M. (2000). Estimation and inference in partially linear models with smoothing splines. *Journal of Statistical Planning and Inference*, 91, 525–540.

[100] Scott, D. W. (1992). *Multivariate Density Estimation. Theory, Practice, and Visualization*, Wiley, New York.

[101] Shiau, J. (1987). A note on MSE coverage intervals in a partial spline model. *Communications in Statistics-Theory and Methods*, 16, 1851–1866.

[102] Song, Q. and Yang, L. (2009). Spline confidence bands for variance function. *Journal of Nonparametric Statistics*, 21, 589–609.

[103] Stone, C. J. (1977). An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion. *Journal of the Royal Statistical Society. Series B*, 39, 44–47.

[104] Stone, C. J. (1985). Additive regression and other nonparametric models. *Annals of Statistics*, 13, 689–705.

[105] Stone, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. *Annals of Statistics*, 22, 118–184.

[106] Su, L. & Ullah, A. (2006). Profile likelihood estimation of partially linear panel data models with fixed effects. *Economics Letters*, 92, 75–81.

[107] Sunklodas, J. (1984). On the rate of convergence in the central limit theorem for strongly mixing random variables. *Lithuanian Mathematical Journal*, 24, 182–190.

[108] Tjøstheim, D. & Auestad, B. (1994). Nonparametric identification of nonlinear time series: projections. *Journal of the American Statistical Association*, 89, 1398–1409.

[109] Ulrich, C.M., Kampman, E., Bigler, J., Schwartz, S. M., Chen C, Bostick, R., Fosdick, L., Beresford, S., Yasui, Y. and Potter, J. (1999). Colorectal adenomas and the C677T MTHFR polymorphism: evidence for gene-environment interaction. *Cancer Epidemiology, Biomarkers & Prevention* 8, 659–668.

[110] Xia, Y. C. (1998). Bias-corrected confidence bands in nonparametric regression. *Journal of the Royal Statistical Society Series B*, 60, 797–811.

[111] Xia, Y. and Li, W. K. (1999). On the estimation and testing of functional-coefficient linear models. *Statistica Sinica*, 3, 735–757.

[112] Xue, L., and Yang, L. J. (2006). Additive coefficient modeling via polynomial spline. *Statistica Sinica*, 16, 1423–1446.

[113] Xue, L. (2006). Variable selection in additive models. *Statistica Sinica*, 19, 1281–1296.

[114] Xue, L. and Liang, H. (2010). Polynomial spline estimation for the generalized additive coefficient model. *Scandinavian Journal of Statistics*, 37, 26–46.

[115] Wahba, G. (1990). *Spline Models for Observational Data*, SIAM, Philadelphia.

[116] Wang, N., Carroll, R. J., and Lin, X. (2005). Efficient semiparametric marginal estimation for longitudinal/clustered data. *Journal of the American Statistical Association*, 100, 147–157.

[117] Wang, L. and Yang, L. (2007). Spline-backfitted kernel smoothing of nonlinear additive autoregression model. *Annals of Statistics*, 35, 2474–2503.

[118] Wang, L. (2009). Single-index model-assisted estimation in survey sampling. *Journal of Nonparametric Statistics*, 21, 487–504.

[119] Wang, J., and Yang, L. J. (2009a). Polynomial spline confidence bands for regression curves. *Statistica Sinica*, 19, 325–342.

[120] Wang, J. & Yang, L. (2009b). Efficient and fast spline-backfitted kernel smoothing of additive models. *Annals of the Institute of Statistical Mathematics*, 61, 663–690.

[121] Wang, Y. (1995). Jump and sharp cusp detection by wavelets. *Biometrika*, 82, 385–397.

[122] Wu, J. S., and Chu, C. K. (1993). Kernel-type estimators of jump points and values of a regression function. *The Annals of Statistics*, 3, 1545–1566.

[123] Wu, W. and Zhao, Z. (2007). Inference of trends in time series. *Journal of the Royal Statistical Society B*, 69, 391–410.

[124] Yao, F. and Lee, T. C. M. (2006). Penalized spline models for functional principal component analysis. *Journal of the Royal Statistical Society B*, 68, 3–25.

[125] Yao, F., Müller, H. G., and Wang, J. L. (2005a). Functional linear regression analysis for longitudinal data. *Annals of Statistics*, 33, 2873–2903.

[126] Yao, F., Müller, H. G., and Wang, J. L. (2005b). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100, 577–590.

[127] Yao, F. (2007). Asymptotic distributions of nonparametric regression estimators for longitudinal or functional data. *Journal of Multivariate Analysis*, 98, 40–56.

[128] Zhang, J. T. and Chen, J. (2007). Statistical inferences for functional data. *Annals of Statistics*, 35, 1052–1079.

[129] Zhao, X., Marron, J. S., and Wells, M. T. (2004). The functional data analysis view of longitudinal data. *Statistica Sinica*, 14, 789–808.

[130] Zhao, Z. and Wu, W. (2008). Confidence bands in nonparametric time series regression. *Annals of Statistics*, 36, 1854–1878.

[131] Zhou, L., Huang, J., and Carroll, R. J. (2008). Joint modelling of paired sparse functional data using principal components. *Biometrika*, 95, 601–619.

[132] Zhou, S., Shen, X. and Wolfe, D. A. (1998). Local asymptotics of regression splines and confidence regions. *The Annals of Statistics*, 26, 1760–1782.

[133] Zhou, X. and You, J. (2004). Wavelet estimation in varying-coefficient partially linear regression models. *Statistics & Probability Letters*, 68, 91–104.