The Application of B-Spline Smoothing: Confidence Bands and Additive Modelling

By

Jing Wang

A DISSERTATION

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

DOCTOR OF PHILOSOPHY

Department of Statistics and Probability

2006

ABSTRACT

The Application of B-Spline Smoothing: Confidence Bands and Additive Modelling

By

Jing Wang

Asymptotically exact and conservative confidence bands are obtained for nonparametric regression function, based on constant and linear polynomial spline estimation, respectively. Compared to the pointwise nonparametric confidence interval of Huang (2003), the confidence bands are inflated only by a factor of $\{\log(n)\}^{1/2}$, similar to the Nadaraya-Watson confidence bands of Härdle (1989), and the local polynomial bands of Xia (1998) and Claeskens and Van Keilegom (2003). Simulation experiments have provided strong evidence that corroborates with the asymptotic theory.

A great deal of effort has been devoted to the inference of additive model in the last decade. Among the many existing procedures, the kernel type are too costly to implement for large number of variables or for large sample sizes, while the spline type provide no asymptotic distribution or any measure of uniform accuracy. We propose a synthetic estimator of the component function in an additive regression model, using a one-step backfitting, with spline smoothing in the first stage and kernel smoothing in the second stage. Under very mild conditions, the proposed SBK estimator of the component function is asymptotically equivalent to an ordinary univariate Nadaraya-

Watson estimator, hence the dimension is effectively reduced to one at any point. This dimension reduction holds uniformly over an interval under stronger assumptions of normal errors, and asymptotic simultaneous confidence bands are provided for the component functions. Monte Carlo evidence supports the asymptotic results for dimensions ranging from low to very high, and sample sizes ranging from moderate to large. The proposed simultaneous confidence bands are applied to the Boston housing data for linearity diagnosis.

Phenological information reflecting seasonal changes in vegetation is an important input variable in climate models such as the Regional Atmospheric Modeling System (RAMS). It varies not only among different vegetation types but also with geographic locations (latitude and longitude). In the current version of RAMS, phenologies are treated as a simple sine function that is solely related to the day of year and latitude, in spite of major seasonal variability in precipitation and temperature. In short, the sine curves of phenology are far different from the observed. Via linear spline smoothing we developed more realistic phenological functions of all land covers in the East Africa to improve RAMS model based on remote sensing observations. In addition, we quantify the differences between the RAMS's default phenological curves and those linear spline estimates derived from remote sensing observations.

To my grandma, my parents, and Yuming

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisor Professor Lijian Yang. He is always willing to answer all kinds of questions with great patience and share his profound insights with me. I am very appreciative to his innumerous encouragement and support during my research and job search. With enduring enthusiasm and dedication to the academia and thoughtful attention to the students, Professor Yang sets an example for being an excellent faculty.

I am truly grateful to Professors Jiaguo Qi, R. V. Ramamoorthi and Yijun Zuo for taking time to serve in my dissertation committee. Especially, I would like to thank Professor Qi and the CLIP group for providing me financial support and sharing their knowledge with me on the project. I really appreciate Professors Dennis Gilliland and Connie Page for their guidance for two years at CSTAT, I obtained valuable experience on consulting service.

My special thanks goes to Professor James Stapleton for continuous help and encouragement from the very beginning. I also want to thank Professor Vince Melfi, Cathy Sparks and Laurie Secord for their assistance, and I thank all the professors and friends who helped me at MSU over five years.

This dissertation research has been supported in part by NSF grants DMS 0405330 and BCS 0308420.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

1.1 Introduction

For the past three decades, nonparametric regression has been widely used in many statistical applications, from biostatistics to econometrics, from engineering to geography. This is due to its flexibility in modelling complex relationships among variables by "letting the data speak for themselves". To fix the idea, we begin with the univariate regression models. 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, ε) satisfying the regression model

$$Y = m(X) + \sigma(X)\varepsilon, . \tag{1.1}$$

The unknown mean and standard deviation functions $m(x)$ and $\sigma(x)$, defined on a compact interval $[a, b]$, need not to be of any specific form.

Two popular nonparametric smoothing techniques are local polynomial/kernel and polynomial spline. The kernel type estimators are "local", treated comprehensively

in Fan and Gijbels (1996) and Härdle (1990). The polynomial spline estimators, on the other hand, are global, see Stone (1985, 1994) and Huang (2003).

The fidelity of a nonparametric regressor is measured in terms of its rate of convergence to the unknown regression function. The type of convergence rates can be pointwise, or uniform. For kernel type estimators, rates of convergence of all three types have been established by Mack and Silverman (1982), Fan and Gijbels (1996).For kernel smoothing of univariate regression function, Hall and Titterington (1988), Härdle (1989), and Xia (1998) made significant contributions on the confidence bands. All of these are based on strong approximation of some empirical processes by the 2-dimensional Brownian bridge, as in Tusnády (1977), which is the same idea used in Bickel and Rosenblatt (1973) for confidence band of probability density function. More recently, Claeskens and Van Keilegom (2003) improved upon Xia (1998) by using smoothed bootstrap, and by extending the confidence band to derivatives of the regression function. Härdle, Huet, Mammen and Sperlich (2004) introduced the bootstrap bands with corrected bias.

For polynomial splines, least squares rates of convergence have been obtained by Stone (1985, 1994), while pointwise convergence rates and asymptotic distribution have been recently established in Huang (2003). Confidence band for polynomial spline regression, however, remains unavailable except under the strong restriction of homoscedastic normal errors, see Zhou, Shen and Wolfe (1998). Since the confidence bands is one of the most important ways to do the model diagnosis, in another words testing the validity of the parametric model, the confidence bands for the heteroscedastic model is in great demand because of its generality.

1.2 Confidence Bands

An asymptotic exact (conservative) $100(1-\alpha)\%$ confidence band for the unknown $m(x)$ over interval $[a,b]$ consists of an estimator $\hat{m}(x)$ of $m(x)$, lower and upper confidence limit $\hat{m}(x) - l_n(x)$, $\hat{m}(x) + l_n(x)$ at every $x \in [a,b]$ such that

$$\begin{aligned} \lim_{n\to\infty} P\{m(x) \in \hat{m}(x) \pm l_n(x), \forall x \in [a,b]\} &= 1-\alpha, \text{ exact,} \\ \liminf_{n\to\infty} P\{m(x) \in \hat{m}(x) \pm l_n(x), \forall x \in [a,b]\} &\geq 1-\alpha, \text{ conservative.} \end{aligned}$$

Confidence band of kernel type estimators are computationally intensive since a least squares estimation has to be done at every point. In contrast, it is enough to solve only one least square problem to get the polynomial spline estimator. The greatest advantages of polynomial spline estimation are its simplicity of implementation and fast computation. But so far the asymptotics property of the spline smoothing is not complete as the kernel type.

To introduce the spline functions, divide the finite interval $[a,b]$ into $(N+1)$ subintervals $J_j = [t_j, t_{j+1}), j = 0, ..., N-1, J_N = [t_N, b]$. A sequence of equally-spaced points $\{t_j\}_{j=1}^{N}$, called interior knots, are given as

$$t_0 = a < t_1 < \cdots < t_N < b = t_{N+1}, t_j = a + jh, \ j = 0, 1, ..., N+1,$$

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

Our first objective is get the following polynomial spline estimator based on data $\{(X_i, Y_i)\}_{i=1}^n$ drawn from model (1.1)

$$\hat{m}_p(x) = \text{argmin}_{g \in G^{(p-2)}[a,b]} \sum_{i=1}^{n} \{Y_i - g(X_i)\}^2, p = 1, 2, \tag{1.2}$$

and then construct the error bound function $l_n(x)$ around this spline estimator.

We now state our main results in the next two theorems.

Theorem 1.2.1. *Under Assumptions (AC1)-(AC4) on Page 16, if $p = 1$ (constant), then an asymptotic $100(1-\alpha)\%$ exact confidence band for $m(x)$ over interval $[a, b]$ is*

$$\hat{m}_1(x) \pm \sigma_{n,1}(x) \{2 \log(N+1)\}^{1/2} d_n,$$

in which $\sigma_{n,1}(x)$ is the pointwise variance function of $\hat{m}_1(x)$, and can be replaced by $\sigma(x) \{f(x) nh\}^{-1/2}$, d_n is defined in (2.2.13) with limit 1 as $n \to \infty$.

Theorem 1.2.2. *Under Assumptions (AC1)-(AC4) on page 16, if $p = 2$ (linear), then an asymptotic $100(1-\alpha)\%$ conservative confidence band for $m(x)$ over interval $[a, b]$ is*

$$\hat{m}_2(x) \pm \sigma_{n,2}(x) \{2 \log(N+1) - 2 \log \alpha\}^{1/2},$$

in which $\sigma_{n,2}(x)$ is the pointwise variance function of $\hat{m}_2(x)$, is defined in (2.2.11).

The construction in Theorem 1.2.1 is similar to the connected error bar of Hall and Titterington (1988). Ours is superior in two aspects: first, we treat not only equally-spaced designs, but random designs; second, by applying the strong approximation of Tusnády (1977), our confidence band is asymptotically exact rather than conservative. The error bars of Hall and Titterington (1988) are based on a kernel estimator while

ours is based on a regressogram. The upcrossing results used in the proof of Theorem 1.2.2 is also different from that used in Bickel and Rosenblatt (1973), Rosenblatt (1976) and Härdle (1989). The theorem on linear confidence band, however, bears no similarity to the local polynomial bands in Xia (1998), Claeskens and Van Keilegom (2003). It is instructive to point out that the asymptotic variance function $\sigma_{n,2}(x)$ of $\hat{m}_2(x)$ is a special unconditional version of equation (6.2), in [Huang (2003), Remark 6.1, page 1624]. Thus, as we have mentioned in the abstract, the linear confidence band localized at any given point x, is only a factor of $(\log n)^{1/2}$ wider than the pointwise normal confidence interval of Huang (2003).

1.3 Additive Component Eastimation

While in practice we have to deal with the high dimensional data in most times. Much effort has been devoted to addressing the issue of the "curse of dimensionality". One popular choice for such purpose is the additive model popularized by the book of Hastie and Tibshirani (1990). Stone (1985) proposed estimators for component functions and their derivatives, and established optimal rates of convergence. These were later called polynomial spline estimators in the extended context of functional ANOVA model in Stone (1994), Huang (1998). Huang and Yang (2004) further extended these estimators to weakly dependent data and developed consistent BIC model selection procedure based on such estimation.

Hastie and Tibshirani (1990) proposed backfitting estimators for components functions without theoretical justifications, while Opsomer and Ruppert (1997) offered

partial asymptotic results for the case of $d = 2$ under some strong assumptions. Opsomer (2000) extended the theoretical results to a general case with more than 2 covariates. Mammen, Linton and Nielsen (1999) proposed a projection based modification of the backfitting algorithm and established its theoretical properties, which was implemented in Nielsen and Sperlich (2005) and called smooth backfitting estimator. Another viable alternative is the so-called marginal integration method, as first proposed in Tjøstheim and Auestad (1994), Linton and Nielsen (1995), Linton and Härdle (1996), and further developed in various contexts by Fan, Härdle and Mammen (1998), Yang, Härdle and Nielsen (1999), Sperlich, Tjøstheim and Yang (2002), Yang, Sperlich and Härdle (2003), Xue and Yang (2006). Using the wavelet transformation, Härdle, Sperlich and Spokoiny (2001) developed the additivity and the polynomial structural tests. Series estimator in Andrews and Whang (1990) circumvented the curse of dimensionality when interactions are present in the model.

Let $\left\{Y_i, \mathbf{X}_i^T\right\}_{i=1}^n = \{Y_i, X_{i1}, ..., X_{id}\}_{i=1}^n$ be an i.i.d. sample following the additive model

$$Y = m(\mathbf{X}) + \sigma(\mathbf{X})\varepsilon, \mathbf{X} = (X_1, ..., X_d), m(\mathbf{x}) = c + \sum_{\alpha=1}^{d} m_\alpha(x_\alpha), \tag{1.3}$$

where the noise satisfies $E(\varepsilon|\mathbf{X}) = 0, \text{var}(\varepsilon|\mathbf{X}) = 1$ and the component functions satisfy the identification conditions $Em_\alpha(X_\alpha) \equiv 0, \alpha = 1, ..., d$. In addition, one assumes that each predictor X_α is distributed on a compact interval $[a_\alpha, b_\alpha]$.

If the last $d - 1$ of the component functions were known by "oracle", then one could define a new variable $Y_1 = Y - c - \sum_{\alpha=2}^{d} m_\alpha(X_\alpha) = m_1(X_1) + \sigma(\mathbf{X})\varepsilon$ which one can use to regress on the numerical variable X_1 to estimate the only

unknown function $m_1(x_1)$, without the "curse of dimensionality". The basic idea of Linton (1997) was to obtain an approximation to the variable Y_1 by substituting $m_\alpha(X_\alpha), \alpha = 2, ..., d$ with the marginal integration pilot estimates (kernel-based) and establishing that the error caused by this "cheating" is negligible for estimating function $m_1(x_1)$. The two-step idea for nonparametric regression also later appeared in Fan and Chen (1999) for local quasi-likelihood estimation. It is well known that the kernel estimation in high dimension would be extremely computationally intensive. Kim, Linton and Hengartner (1999) provided an computationally efficient two-step estimator, a reduction in computation of order n compared with marginal integration. The spline method, on the other hand, is very fast, but the rate of convergence is only established in mean squares sense, and there is no pointwise confidence interval or even consistency in additive models. In particular, Härdle, Marron and Yang (1997) demonstrated that the adaptive spline method could lack uniform consistency.

We propose to pre-estimate the functions $\{m_\alpha(x_\alpha)\}_{\alpha=1}^d$ by an under smoothed constant spline procedure. These function estimates are then used as if they were the true functions for constructing the "oracle"estimator. The greatest advantage of our approach over that of Linton (1997) is that ours is much faster, and can be applied to cases of extremely high dimension data (e.g., the number of predictors, d, can be as large as 50 or 100). One may wonder how one could have all these good features in one method. The success of our method is due to the well-known "reducing bias by undersmoothing" and "averaging out the variance" principles, both goals are accomplished with the joint asymptotics of kernel and spline functions, which is the new feature of our proofs.

In addition to those features, uniform confidence bands are provided for all function estimates under mild conditions. For additive regression model, however, it seems that this present work is the one of the few to offer the measure of uniform accuracy with theoretical justifications. The good news is that the confidence band we provide for $m_\alpha(x_\alpha)$ with any $\alpha = 1, ..., d$, is asymptotically the same confidence band that Härdle (1989) established for univariate regression with kernel smoother, regardless how many regressors there are and what other functions $m_\alpha(x_\alpha), \alpha = 1, ..., d$ are. Hence neither the dimension d nor other function components play any role in forming the band for $m_\alpha(x_\alpha)$, at least according to the asymptotic theory. In this sense, our estimator of $m_\alpha(x_\alpha)$ possesses what we would like to call "uniform oracle efficiency", which is much stronger than the "pointwise oracle efficiency" of Linton (1997).

Without loss of generality, we take all intervals $[a_\alpha, b_\alpha] = [0, 1], \alpha = 1, ..., d$. Define for any $\alpha = 1, ..., d$, the indicator function $I_{J,\alpha}(x_\alpha)$ of the $(N+1)$ equally-spaced subintervals of the finite interval $[0, 1]$, that is

$$I_{J,\alpha}(x_\alpha) = \begin{cases} 1 & JH \leq x_\alpha < (J+1)H, \\ 0 & \text{otherwise,} \end{cases} \quad H = H_n = (N_n + 1)^{-1}, J = 0, 1, ..., N. \tag{1.4}$$

Define the $(1 + dN)$-dimensional space G of additive spline functions as the linear space spanned by $\{1, I_{J,\alpha}(x_\alpha), \alpha = 1, ..., d, J = 1, ..., N\}$. The spline estimator of additive function $m(\mathbf{x})$ is the unique element $\hat{m}(\mathbf{x}) = \hat{m}_n(\mathbf{x})$ from the space G so that the vector $\{\hat{m}(\mathbf{X}_1), ..., \hat{m}(\mathbf{X}_n)\}^T$ best approximates the response vector $\mathbf{Y} = (Y_1, ..., Y_n)^T$. To be precise, we define

$$\hat{m}(\mathbf{x}) = \hat{\lambda}_0 + \sum_{\alpha=1}^{d} \sum_{J=1}^{N} \hat{\lambda}_{J,\alpha} I_{J,\alpha}(x_\alpha), \tag{1.5}$$

where the coefficients $\hat{\lambda}_0, \hat{\lambda}_{1,1}, ..., \hat{\lambda}_{N,d}$ are the least square solution given by

$$\left\{\hat{\lambda}_0, \hat{\lambda}_{1,1}, ..., \hat{\lambda}_{N,d}\right\}^T = \text{argmin}_{R^{dN+1}} \sum_{i=1}^{n} \left\{ Y_i - \lambda_0 - \sum_{\alpha=1}^{d} \sum_{J=1}^{N} \lambda_{J,\alpha} I_{J,\alpha}(X_{i\alpha}) \right\}^2. \tag{1.6}$$

The pilot estimators of each component function and the constant are defined as

$$\begin{aligned} \hat{m}_\alpha(x_\alpha) &= \sum_{J=1}^{N} \hat{\lambda}_{J,\alpha} I_{J,\alpha}(x_\alpha) - n^{-1} \sum_{i=1}^{n} \sum_{J=1}^{N} \hat{\lambda}_{J,\alpha} I_{J,\alpha}(X_{i\alpha}), \\ \hat{m}_c &= \hat{\lambda}_0 + n^{-1} \sum_{\alpha=1}^{d} \sum_{i=1}^{n} \sum_{J=1}^{N} \hat{\lambda}_{J,\alpha} I_{J,\alpha}(X_{i\alpha}). \end{aligned} \tag{1.7}$$

These pilot estimators are then used to define a set of new pseudo-responses $\hat{Y}_{i1}$ which are estimated versions of the unobservable "oracle" responses Y_{i1},

$$\hat{Y}_{i1} = Y_i - \hat{c} - \sum_{\alpha=2}^{d} \hat{m}_\alpha(X_{i\alpha}), Y_{i1} = Y_i - c - \sum_{\alpha=2}^{d} m_\alpha(X_{i\alpha}), i = 1, ..., n, \hat{c} = n^{-1} \sum_{i=1}^{n} Y_i. \tag{1.8}$$

The proposed spline-backfitted kernel (SBK) estimator of $m_1(x_1)$ as $\hat{m}_{s,1}(x_1)$ based on $\left\{\hat{Y}_{i1}, X_{i1}\right\}_{i=1}^{n}$, which is an attempt to mimick the would-be Nadaraya-Watson estimator $\tilde{m}_{s,1}(x_1)$ of $m_1(x_1)$ based on $\{Y_{i1}, X_{i1}\}_{i=1}^{n}$, had the unobservable "oracle" responses $\{Y_{i1}\}_{i=1}^{n}$ been available.

$$\hat{m}_{s,1}(x_1) = \frac{\sum_{i=1}^{n} K_h(X_{i1} - x_1) \hat{Y}_{i1}}{\sum_{i=1}^{n} K_h(X_{i1} - x_1)}, \tilde{m}_{s,1}(x_1) = \frac{\sum_{i=1}^{n} K_h(X_{i1} - x_1) Y_{i1}}{\sum_{i=1}^{n} K_h(X_{i1} - x_1)}, \tag{1.9}$$

where $\hat{Y}_{i1}$ and Y_{i1} are defined above. Similar constructions can be based on local linear instead of Nadaraya-Watson estimator, which is called spline-backfitted local linear estimator (SBLL).

The asymptotic property of the kernel smoother $\tilde{m}_{s,1}(x_1)$ is well-developed according to Theorem 4.2.1 of Härdle (1990), one has

$$\sqrt{nh}\left\{\tilde{m}_{s,1}(x_1) - m_1(x_1) - b(x_1) h^2\right\} \xrightarrow{D} N\left(0, v^2(x_1)\right),$$

where

$$\begin{aligned} b(x_1) &= \mu_2(K)\left\{m_1''(x_1) f_1(x_1)/2 + m_1'(x_1) f_1'(x_1)\right\} f_1^{-1}(x_1), \\ v^2(x_1) &= \|K\|_2^2 E\left\{\sigma^2(x_1, X_2, ..., X_d)\right\} f_1^{-1}(x_1). \end{aligned} \tag{1.10}$$

In contrast, the bias coefficient of the SBLL estimator would simply be $b(x_1) = \mu_2(K) m_1''(x_1)/2$, without the additional term of the SBK estimator, while the variance coefficients of SBLL and SBK are the same.

Härdle (1989)provide the uniform asymptotics for kernel smoother. For any $\alpha \in (0,1)$, an asymptotic $100(1-\alpha)\%$ confidence band for $m_1(x_1)$ over interval $[0,1]$ is

$$\lim_{n\to\infty} P\left\{m_1(x_1) \in \tilde{m}_{s,1}(x_1) \pm l_n(x_1), \forall x_1 \in [0,1]\right\} = 1-\alpha$$

where $l_n(x_1)$ is defined in (3.2.9).

Theorem 1.3.1. *Under Assumptions (AS1) to (AS6) on page 57, for any $x_1 \in [0,1]$, the SBK/SBLL estimator $\hat{m}_{s,1}(x_1)$ given in (1.9) satisfies*

$$\left|\hat{m}_{s,1}(x_1) - \tilde{m}_{s,1}(x_1)\right| = o_p\left(n^{-2/5}\right)$$

Theorem 1.3.2. *Under Assumptions (AS1) to (AS6) and (AS2') on page 57, the SBK/SBLL estimator $\hat{m}_{s,1}(x_1)$ given in (1.9) satisfies*

$$\sup_{x_1\in[0,1]} \left|\hat{m}_{s,1}(x_1) - \tilde{m}_{s,1}(x_1)\right| = o_p\left(n^{-2/5}\right).$$

The two theorems state that the asymptotic magnitude of difference between $\hat{m}_{s,1}(x_1)$ and $\tilde{m}_{s,1}(x_1)$ is dominated by the asymptotic size of $\tilde{m}_{s,1}(x_1) - m_1(x_1)$. Hence $\hat{m}_{s,1}(x_1)$ will have the same asymptotic distribution as $\tilde{m}_{s,1}(x_1)$, pointwise and uniformly. Higher order local polynomials can also be used, with obvious modifications. For more on the properties of local linear estimators, in particular, its minimax efficiency, see Fan and Gijbels (1996).

1.4 Application to Seasonality Analysis

Many studies demonstrate the influence of land use and land cover change on local and regional climate. The Climate and Land use Interaction Project, or CLIP (http://clip.msu.edu) attempts to understand the nature and magnitude of the interactions of climate and land use/cover change across East Africa.

Phenological information reflecting the seasonal variability of vegetation is an important input variable in regional climate models such as Regional Atmosphere Simulation System (RAMS). It varies not only among different vegetation types but also with geographic locations (latitude and longitude).

Many climate models use simple functions for vegetation parameters since, to first order, the planet is warmer and wetter as you approach the equator. However, east Africa is unique in having semiarid grasslands along the equator, and drastically different surface conditions govern the radiationbudget in this region. Climate models are dependent on an accurate representation of the surface radiation budget to replicate atmospheric development. Thus, modeling climate for a unique area like east Africa requires a different treatment of vegetation characteristics.

RAMS version 4.4 (Cotton et al. 2003), a state-of-the-art three dimensional atmospheric model, includes a representation of vegetation called the Land-Ecosystem-Atmosphere Feedback, version 2 (LEAF-2) (Walko et al. 2000). For a given land cover class, LEAF-2 provides functions for several vegetation characteristics including LAI, fractional cover, roughness length, and displacement height. Although these characteristics are interrelated, we will consider only LAI here.

Based on the observations of LAI of MODIS data, the polynomial spline regression is employed to fit the function of each land type in East Africa. We develop the function first temporally and then further investigate the spatial influence. In other words, the estimate function of LAI will rely on the time and the spatial index (latitude and longitude). Four major land cover types are chosen to display the trend of the LAI.

Let Z =LAI, x = latitude, y = longitude, t =Julian day. For each LC type we develop the LAI function as follows,

$$Z(x,y,t) = \hat{a}_0(x,y) + \sum_{j=1}^{11} \hat{a}_j(x,y) \cdot (t - t_j)_+ + \hat{a}_{12}(x,y)\, t, \tag{1.11}$$

The coefficients $\hat{a}_j(x,y)$ for $j = 0, 1, ..., 12$, are estimated based on the MODIS data at each individual grid. Different LC type will have different coefficients set, see Tables 4.5 - 4.8.

Figure 4.11 and 4.12 illustrates two examples of the seasonal variation in LAI for common classes in the study area, "Rainfed Herbaceous Crop" and "Open to Very Open Trees". The observed LAI and resultant splines are distinctly different from the RAMS/LEAF-2 default parameterization, with the LEAF-2 parameterization completely failing to capture the seasonality at the equator (solid) or in the regions +/- 5 (dashed/dotted) away. The spline parameterizations accurately capture bimodal greening events at the equator, unimodal features away from the equator, and the very low LAI for maize regions following harvest.

Figures 4.17 shows LAI values at 8 May 2000 for three combinations of land cover and LAI phenology, along with a MODIS image for comparison. The profound

difference in LAI from 4.17 (a) to (d) at the Equator shows that the LEAF-2 function is essentially treating the semidesert of eastern Kenya as having high LAI with no variation. These successive improvements have helped to give a more precise surface parameterization while keeping the flexibility needed to accommodate projected land use change.

The hypotheses for each land type is

$$H_0 \ : \ \text{LAI trend curve follows the RAMS Curve}$$

$$H_a \ : \ \text{Not follow the RAMS Curve.}$$

The test illustrates that the RAMS curves overestimate the LAI, with the difference being significantly large indicated from the small p-value< 0.001, see Figures 4.13 to 4.16.

The dissertation is organized as follows. In Chapter 2, we develop the exact confidence bands via constant spline regression and the conservative ones via linear spline regression. Chapter 3 the spline-backfitted kernel estimator is proposed to estimate the component function in an additive model under mild conditions. We applied the linear spline estimator and its uniform asymptotics to estimate and test the Leaf Area Index trend for CLIP (Climate Land Interaction Project) in Chapter 4.

CHAPTER 2

Spline Confidence Bands

2.1 Introduction

In this chapter, we present confidence bands of univariate regression function based on polynomial spline smoothing. 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, ε) satisfying the regression model

$$Y = m(X) + \sigma(X)\varepsilon, \tag{2.1.1}$$

where the joint distribution of (X, ε) satisfies Assumption (AC4) in Section 2.2. The unknown mean and standard deviation functions $m(x)$ and $\sigma(x)$, defined on interval $[a, b]$, need not to be of any specific form. If the data actually follows a polynomial regression model, $m(x)$ would be a polynomial and $\sigma(x)$, a constant.

We organize this chapter as follows. In Section 2.2 we state our main results on confidence bands constructed from (piecewise) constant/linear splines. In Section 2.3 we provide further insights into the error structure of spline estimators. Section 2.4

describes the actual steps to implement the confidence bands. Section 2.5 reports findings in an extensive simulation study and the application to the testing of polynomial trend hypothesis for the well-known motorcycle data. Section 2.6 concludes. All technical proofs are contained in Section 2.7.

2.2 Main Results

To introduce the spline functions, divide the finite interval $[a, b]$ into $(N+1)$ subintervals $J_j = \left[t_j, t_{j+1}\right), j = 0, ..., N-1, J_N = [t_N, b]$. A sequence of equally-spaced points $\{t_j\}_{j=1}^{N}$, called interior knots, are given as

$$t_0 = a < t_1 < \cdots < t_N < b = t_{N+1}, t_j = a + jh,\ j = 0, 1, ..., N+1,$$

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

In what follows, $\|\cdot\|_\infty$ denotes the supremum norm of a function r on $[a, b]$, i.e. $\|r\|_\infty = \sup_{x \in [a,b]} |r(x)|$, and the moduli of continuity of a continuous function r on $[a, b]$ is denoted as $\omega(r, h) = \max_{x, x' \in [a,b], |x-x'| \le h} \left|r(x) - r\left(x'\right)\right|$. One has $\lim_{h \to 0} \omega(r, h) = 0$ by the uniform continuity of r on a compact interval $[a, b]$.

Our approach is to get the following polynomial spline estimator based on data

$\{(X_i, Y_i)\}_{i=1}^n$ drawn from model (2.1.1)

$$\hat{m}_p(x) = \text{argmin}_{g \in G^{(p-2)}[a,b]} \sum_{i=1}^{n} \{Y_i - g(X_i)\}^2, p = 1, 2, \tag{2.2.1}$$

and then construct the error bound function $l_n(x)$ around this spline estimator. The technical assumptions we need are as follows:

(AC1) *The regression function* $m(\cdot) \in C^{(p)}[a, b]$, $p = 1, 2$.

(AC2) *The density function* $f(x)$ *of* X *is continuous and positive on interval* $[a, b]$. *The standard deviation function* $\sigma(x) \in C[a, b]$ *has bounded variation and positive lower bound on* $[a, b]$.

(AC3) *The subinterval length* $h \sim n^{-1/(2p+1)}$. *I.e., the number of interior knots* $N \sim n^{1/(2p+1)}$.

(AC4) *The joint distribution* $F(x, \varepsilon)$ *of random variables* (X, ε) *satisfies the following:*

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

(b) *There exists a positive value* $\delta 1/p$ *and finite positive* M_δ *such that* $E|\varepsilon|^{2+\delta} < M_\delta$ *and*

$\sup_{x \in [a,b]} E\left(|\varepsilon|^{2+\delta} | X = x\right) < M_\delta$.

Assumptions (AC1)-(AC3) are the same as in Huang (2003), while Assumption (AC4) is the same as (C2) (a) of Mack and Silverman (1982). All are typical assumptions for nonparametric regression, with (AC1), (AC2) and (AC4) weaker than the corresponding assumptions in Härdle (1989).

To properly define the confidence bands, we introduce some additional notations. For any $x \in [a,b]$, define its location and relative position indices $j(x), \delta(x)$ as

$$j(x) = j_n(x) = \min\left\{\left[\frac{x-a}{h}\right], N\right\}, \delta(x) = \frac{x - t_{j(x)}}{h}. \tag{2.2.2}$$

It is clear that $t_{j_n(x)} \leq x < t_{j_n(x)+1}$, $0 \leq \delta(x) < 1, \forall x \in [a,b)$, and $\delta(b) = 1$. Denote by $\|\phi\|_2$ the theoretical L^2 norm of a function ϕ on $[a,b]$, $\|\phi\|_2^2 = E\{\phi^2(X)\} = \int_a^b \phi^2(x) f(x)\, dx$, 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

$$\langle \phi, \varphi \rangle = \int_a^b \phi(x)\varphi(x) f(x)\, dx = E\{\phi(X)\varphi(X)\}, \langle \phi, \varphi \rangle_n = \frac{1}{n}\sum_{i=1}^n \phi(X_i)\varphi(X_i).$$

for any L^2-integrable functions ϕ, φ on $[a,b]$. Clearly $E\langle \phi, \varphi \rangle_n = \langle \phi, \varphi \rangle$.

Although the truncated power basis is used in implementation (see Section 2.4), it is more convenient to work with the B-spline basis for theoretical analysis. The B-spline basis of $G^{(-1)}$, the space of piecewise constant splines, are indicator functions of intervals J_j, $b_{j,1}(x) = I_j(x) = I_{J_j}(x), j = 0, 1, ..., N$. The B-spline basis of $G^{(0)}$, the space of piecewise linear splines, are $\{b_{j,2}(x)\}_{j=-1}^N$

$$b_{j,2}(x) = K\left(\frac{x - t_{j+1}}{h}\right), j = -1, 0, ..., N, \text{ for } K(u) = (1 - |u|)_+.$$

Define next their theoretical norms

$$c_{j,n} = \|b_{j,1}\|_2^2 = \int_a^b I_j(x) f(x)\, dx, d_{j,n} = \|b_{j,2}\|_2^2 = \int_a^b K^2\left(\frac{x - t_{j+1}}{h}\right) f(x)\, dx. \tag{2.2.3}$$

We introduce the rescaled B-spline basis $\{B_{j,1}(x)\}_{j=0}^N$ and $\{B_{j,2}(x)\}_{j=-1}^N$ for

$G^{(-1)}$ and $G^{(0)}$

$$\begin{aligned} B_{j,1}(x) &\equiv b_{j,1}(x)\left\{c_{j,n}\right\}^{-1/2}, j=0,...,N, \\ B_{j,2}(x) &\equiv b_{j,2}(x)\left\{d_{j,n}\right\}^{-1/2}, j=-1,...,N. \end{aligned} \tag{2.2.4}$$

It is straightforward to see that

$$\left\|B_{j,1}\right\|_2^2 = 1, j=0,1,\ldots,N, \left\langle B_{j,1}, B_{j',1}\right\rangle \equiv 0, j \neq j'. \tag{2.2.5}$$

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

$$V = \left(v_{j'j}\right)_{j,j'=-1}^{N} = \left(\left\langle B_{j',2}, B_{j,2}\right\rangle\right)_{j,j'=-1}^{N}, \tag{2.2.6}$$

whose inverse S and 2×2 diagonal submatrices of S are expressed as

$$S = \left(s_{j'j}\right)_{j,j'=-1}^{N} = V^{-1}, S_j = \begin{pmatrix} s_{j-1,j-1} & s_{j-1,j} \\ s_{j,j-1} & s_{j,j} \end{pmatrix}, j=0,...,N. \tag{2.2.7}$$

Next define matrices $\Sigma, \boldsymbol{\Delta}(x)$ and Ξ_j as

$$\Sigma = \left(\sigma_{jl}\right)_{j,j'=-1}^{N} = \left\{\int \sigma^2(v) B_{j,2}(v) B_{l,2}(v) f(v)\, dv\right\}_{j,j'=-1}^{N}. \tag{2.2.8}$$

$$\begin{aligned} \boldsymbol{\Delta}(x) &= \begin{pmatrix} c_{j(x)-1}\{1-\delta(x)\} \\ c_{j(x)}\delta(x) \end{pmatrix}, c_j = \begin{cases} \sqrt{2} & j=-1,N \\ 1 & j=0,...,N-1 \end{cases}, \\ \Xi_j &= \begin{pmatrix} l_{j+1,j+1} & l_{j+1,j+2} \\ l_{j+2,j+1} & l_{j+2,j+2} \end{pmatrix}, j=0,1,...,N, \end{aligned} \tag{2.2.9}$$

with terms $l_{ik}, |i-k| \leq 1$ defined through the following matrix inversion

$$M_{N+2} = \begin{pmatrix} 1 & \sqrt{2}/4 & & & & 0 \\ \sqrt{2}/4 & 1 & 1/4 & & & \\ & 1/4 & 1 & \ddots & & \\ & & \ddots & \ddots & 1/4 & \\ & & & 1/4 & 1 & \sqrt{2}/4 \\ 0 & & & & \sqrt{2}/4 & 1 \end{pmatrix}_{(N+2)\times(N+2)} = \left(l_{ik}\right)_{(N+2)\times(N+2)}^{-1}, \tag{2.2.10}$$

and computed via (2.4.14), (2.4.17), and (2.4.18).

We define now

$$\sigma_{n,1}^2(x) = \frac{\int I_{j(x)}(v)\,\sigma^2(v)\,f(v)\,dv}{nc_{j(x),n}^2}, \sigma_{n,2}^2(x) = \frac{1}{n}\sum_{j,j',l,l'=-1}^{N} B_{j',2}(x)\,B_{l',2}(x)\,s_{jj'}s_{ll'}\sigma_{jl}, \tag{2.2.11}$$

with $j(x)$ defined in (2.2.2), $c_{j,n}$ in (2.2.3), $B_{j',2}(x)$ in (2.2.4), and $s_{ll'}$ and σ_{jl} in (2.2.7), (2.2.8). These $\sigma_{n,p}^2(x)$ are shown in Lemmas 2.7.4, 2.7.4 to be the pointwise variance functions of $\hat{m}_p(x), p = 1, 2$.

We now state our main results in the next two theorems.

Theorem 1. *Under Assumptions (AC1)-(AC4), if* $p = 1$, *then an asymptotic* $100(1-\alpha)\%$ *exact confidence band for* $m(x)$ *over interval* $[a, b]$ *is*

$$\hat{m}_1(x) \pm \sigma_{n,1}(x)\{2\log(N+1)\}^{1/2} d_n, \tag{2.2.12}$$

in which $\sigma_{n,1}(x)$ *is given in (2.2.11) and can be replaced by* $\sigma(x)\{f(x)nh\}^{-1/2}$, *according to (2.7.7) in Lemma* 2.7.4, *and*

$$d_n = 1 - \{2\log(N+1)\}^{-1}\left[\log\left\{-\frac{1}{2}\log(1-\alpha)\right\} + \frac{1}{2}\{\log\log(N+1) + \log 4\pi\}\right]. \tag{2.2.13}$$

Theorem 2. *Under Assumptions (AC1)-(AC4), if* $p = 2$, *then an asymptotic* $100(1-\alpha)\%$ *conservative confidence band for* $m(x)$ *over interval* $[a, b]$ *is*

$$\hat{m}_2(x) \pm \sigma_{n,2}(x)\{2\log(N+1) - 2\log\alpha\}^{1/2}, \tag{2.2.14}$$

in which $\sigma_{n,2}(x)$ *is given in (2.2.11) and can be replaced by* $\sigma(x)\{2f(x)nh/3\}^{-1/2}\Delta^T(x)S_{j(x)}\Delta(x)$, *according to Lemma* 2.7.4, *and by* $\sigma(x)\{2f(x)nh/3\}^{-1/2}\Delta^T(x)\Xi_{j(x)}\Delta(x)$ *according to Lemma* 2.7.3.

The construction in Theorem 1 is similar to the connected error bar of Hall and Titterington (1988). Ours is superior in two aspects: first, we treat not only equally-spaced designs, but random designs; second, by applying the strong approximation theorem of Tusnády (1977), our confidence band is asymptotically exact rather than conservative. The error bars of Hall and Titterington (1988) are based on a kernel estimator while ours regressogram. The upcrossing results (Theorem 2.3.4) used in the proof of Theorem 1 is also different from that used in Bickel and Rosenblatt (1973), Rosenblatt (1976) and Härdle (1989). Theorem 2 on linear confidence band, however, bears no similarity to the local polynomial bands in Xia (1998), Claeskens and Van Keilegom (2003), except the width of the band being of the same order $n^{-1/5}(\log n)^{1/2}$. The asymptotic variance function $\sigma_{n,2}^2(x)$ of $\hat{m}_2(x)$ in (2.2.11) is a special unconditional version of equation (6.2), in Huang (2003), Remark 6.1, page 1624. Thus, the linear band localized at any given point x, is only a factor of $(\log n)^{1/2}$ wider than the pointwise confidence interval of Huang (2003).

2.3 Error Decomposition

In this section, we break the estimation error $\hat{m}_p(x) - m(x)$ into a bias term and a noise term. To understand this decomposition, we begin by discussing the spline space $G^{(p-2)}$ and the representation of the spline estimator $\hat{m}_p(x)$ in (2.2.1).

The first fact to note is that the empirical inner products of the B-spline basis $\{B_{j,1}(x)\}_{j=0}^{N}$ and $\{B_{j,2}(x)\}_{j=-1}^{N}$ defined in (2.2.4) approximate the theoretical inner products uniformly at the rate of $\sqrt{n^{-1}h^{-1}\log(n)}$, according to the following lemma.

Lemma 2.3.1. *As $n \to \infty$, the B-spline basis $\{B_{j,1}(x)\}_{j=0}^{N}$ and $\{B_{j,2}(x)\}_{j=-1}^{N}$ defined in (2.2.4) satisfy*

$$A_{n,1} = \sup_{0 \le j \le N} \left| \|B_{j,1}\|_{2,n}^2 - 1 \right| = O_p\left(\sqrt{\log n/(nh)}\right), \tag{2.3.1}$$

$$A_{n,2} = \sup_{g_1, g_2 \in G^{(0)}} \left| \frac{\langle g_1, g_2 \rangle_n - \langle g_1, g_2 \rangle}{\|g_1\|_2 \|g_2\|_2} \right| + \sup_{g \in G^{(0)}} \left| \frac{\|g\|_{2,n}}{\|g\|_2} - 1 \right| = O_p\left(\sqrt{\log n/(nh)}\right). \tag{2.3.2}$$

To express the estimator $\hat{m}_p(x)$ in $\{B_{j,p}(x)\}_{j=1-p}^{N}$, we introduce the following vectors in R^n for $p = 1, 2$

$$\mathbf{Y} = (Y_1, ..., Y_n)^T, \mathbf{B}_{j,p}(\mathbf{X}) = \{B_{j,p}(X_1), ..., B_{j,p}(X_n)\}^T, j = 1-p, ..., N.$$

The definition of $\hat{m}_p(x)$ in (2.2.1) entails that $\hat{m}_p(x) \equiv \sum_{j=1-p}^{N} \hat{\lambda}_{j,p} B_{j,p}(x)$ where the coefficients $\left\{\hat{\lambda}_{1-p,p}, ..., \hat{\lambda}_{N,p}\right\}^T$ are solutions of the following least squares problem

$$\left\{\hat{\lambda}_{1-p,p}, ..., \hat{\lambda}_{N,p}\right\}^T = \operatorname{argmin} \sum_{i=1}^{n} \left\{ Y_i - \sum_{j=1-p}^{N} \lambda_{j,p} B_{j,p}(X_i) \right\}^2. \tag{2.3.3}$$

We write $\mathbf{Y}$ as the sum of a signal vector $\mathbf{m}$ and a noise vector $\mathbf{E}$

$$\mathbf{Y} = \mathbf{m} + \mathbf{E}, \mathbf{m} = \{m(X_1), ..., m(X_n)\}^T, \mathbf{E} = \{\sigma(X_1)\varepsilon_1, ..., \sigma(X_n)\varepsilon_n\}^T.$$

Projecting this relationship into the linear space spanned by $G_n^{(p-2)} = \{\mathbf{B}_{j,p}(\mathbf{X})\}_{j=1-p}^{N}$, a subspace of R^n, one gets

$$\hat{\mathbf{m}}_p = \{\hat{m}_p(X_1), ..., \hat{m}_p(X_n)\}^T = \operatorname{Proj}_{G_n^{(p-2)}} \mathbf{Y} = \operatorname{Proj}_{G_n^{(p-2)}} \mathbf{m} + \operatorname{Proj}_{G_n^{(p-2)}} \mathbf{E}.$$

It entails that in the space $G^{(p-2)}$ of spline functions

$$\hat{m}_p(x) = \tilde{m}_p(x) + \tilde{\varepsilon}_p(x), \tag{2.3.4}$$

where

$$\tilde{m}_p(x) = \sum_{j=1-p}^{N} \tilde{\lambda}_{j,p} B_{j,p}(x), \tilde{\varepsilon}_p(x) = \sum_{j=1-p}^{N} \tilde{a}_{j,p} B_{j,p}(x). \tag{2.3.5}$$

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 (2.3.3) with Y_i replaced by $m(X_i)$ and $\sigma(X_i)\varepsilon_i$ respectively.

We cite next two important results. The first one from de Boor (2001), page 149, the second one from Theorem 5.1 of Huang (2003).

Theorem 2.3.1. *There is an absolute constant $C_p > 0, p \geq 1$ such that for every $m \in C^{(p)}[a,b]$, there exists a function $g \in G^{(p-2)}[a,b]$ such that*

$$\|g - m\|_\infty \leq C_p \left\|\omega\left(m^{(p-1)}, h\right)\right\|_\infty h^{p-1} \leq C_p \left\|m^{(p)}\right\|_\infty h^p$$

Theorem 2.3.2. *There is an absolute constant $C_p > 0, p \geq 1$ such that for any $m \in C^{(p)}[a,b]$ and the function $\tilde{m}_p(x)$ defined in (2.3.5),*

$$\left\|\tilde{m}_p(x) - m(x)\right\|_\infty \leq C_p \inf_{g \in G^{(p-2)}} \|g - m\|_\infty = O_p(h^p). \tag{2.3.6}$$

According to (2.3.4), the estimation error $\hat{m}_p(x) - m(x) = \left\{\tilde{m}_p(x) - m(x)\right\} + \tilde{\varepsilon}_p(x)$ where according to Theorem 2.3.2, the bias term $\tilde{m}_p(x) - m(x)$ is of order $O_p(h^p)$. Hence the main hurdle of proving Theorems 1 and 2 is the noise term $\tilde{\varepsilon}_p(x)$. This is handled by the next two propositions.

Proposition 2.3.1. *With $\sigma_{n,1}(x)$ given in (2.2.11), the process $\sigma_{n,1}(x)^{-1}\tilde{\varepsilon}_1(x), x \in [a,b]$ is almost surely uniformly approximated by a Gaussian process $U(x), x \in [a,b]$ with covariance structure*

$$EU(x)U(y) = \sum_{j=0}^{N} I_j(x) \cdot I_j(y) = \delta_{j(x),j(y)}, \forall x, y \in [a,b],$$

where δ_{jl} is the Kronecker symbol, i.e., $\delta_{j,l} = 1$ if $j = l$ and 0 otherwise.

Proposition 2.3.2. *For a given* $0 < \alpha < 1$, *and* $\sigma_{n,2}(x)$ *as given in (2.2.11)*

$$\liminf_{n\to\infty} P\left[\sup_{x\in[a,b]}\left|\sigma_{n,2}^{-1}(x)\,\tilde{\varepsilon}_2(x)\right| \le \{2\log(N+1) - 2\log\alpha\}^{1/2}\right] \ge 1-\alpha. \quad (2.3.7)$$

We state next the strong approximation theorem of Tusnády (1977), which will be used later in the proof of Lemmas 2.7.6 and 2.7.6, key steps in proving Proposition 2.3.1 and Proposition 2.3.2.

Theorem 2.3.3. *Let* $U_1, \ldots, U_n$ *be i.i.d. r.v.'s on the* 2 *-dimensional unit square with*

$$P(U_i < \mathbf{t}) = \lambda(\mathbf{t}), \mathbf{0} \le \mathbf{t} \le \mathbf{1},$$

where $\mathbf{t} = (t_1, t_2)$ *and* $\mathbf{1} = (1,1)$ *are* 2*-dimensional vectors,* $\lambda(\mathbf{t}) = t_1 t_2$. *The empirical distribution function* $F_n^u(\mathbf{t})$ *based on sample* $(U_1, \ldots, U_n)$ *is* $F_n^u(\mathbf{t}) = n^{-1}\sum_{i=1}^n I_{\{U_i<\mathbf{t}\}}$ *for* $\mathbf{0} \le \mathbf{t} \le \mathbf{1}$. *The* 2*-dimensional Brownian bridge* $B(\mathbf{t})$ *is defined by* $B(\mathbf{t}) = W(\mathbf{t}) - \lambda(\mathbf{t})W(\mathbf{1})$ *for* $\mathbf{0} \le \mathbf{t} \le \mathbf{1}$, *where* $W(\mathbf{t})$ *is a* 2*-dimensional Wiener process. Then there is a version of* $F_n^u(\mathbf{t})$ *and* $B(\mathbf{t})$ *such that*

$$P\left[\sup_{\mathbf{0}\le\mathbf{t}\le\mathbf{1}}\left|n^{1/2}\{F_n^u(\mathbf{t}) - \lambda(\mathbf{t})\} - B(\mathbf{t})\right| > n^{-1/2}(C\log n + x)\log n\right] < Ke^{-\lambda x} \quad (2.3.8)$$

holds for all x, *where* C, K, λ *are positive constants.*

For the rest of the paper, we denote the well-known Rosenblatt quantile transformation as

$$(X', \varepsilon') = M(X, \varepsilon) = \left\{F_X(x), F_{\varepsilon|X}(\varepsilon|x)\right\}, \quad (2.3.9)$$

which produces random variables X' and ε' with independent and identical uniform distribution on the interval $[0,1]$. This transformation had been used in, for instance,

Bickel and Rosenblatt (1973), Härdle (1989). Substituting the vector $\mathbf{t} = (t_1, t_2)$ in Theorem 2.3.3 with (X', ε'), and the stochastic process $n^{1/2}\{F_n^u(\mathbf{t}) - \lambda(\mathbf{t})\}$ with

$$Z_n\left\{M^{-1}(x', \varepsilon')\right\} = Z_n(x, \varepsilon) = \sqrt{n}\{F_n(x, \varepsilon) - F(x, \varepsilon)\}, \tag{2.3.10}$$

where $F_n(x, \varepsilon)$ denotes the empirical distribution of (X, ε), then (2.3.8) implies that there exists a version of 2-dimensional Brownian bridge B such that

$$\sup_{x,\varepsilon} |Z_n(x, \varepsilon) - B\{M(x, \varepsilon)\}| = O\left(n^{-1/2}\log^2 n\right), \text{w.p.1.} \tag{2.3.11}$$

The next result on upcrossing probability is from Leadbetter, Lindgren and Rootzén (1983), Theorem 1.5.3, page 14. In our proof of Theorem 1, it plays the role of Theorem A1 in Bickel and Rosenblatt (1973) or Theorem C in Rosenblatt (1976).

Theorem 2.3.4. *If $\xi_1, ..., \xi_n$ are i.i.d. standard normal r.v.'s, then for $M_n = \max\{\xi_1, ..., \xi_n\}$, $\tau \in R$, as $n \to \infty$*

$$P\{a_n(M_n - b_n) \leq \tau\} \to \exp\left(-e^{-\tau}\right), P\{|M_n| \leq \tau/a_n + b_n\} \to \exp\left(-2e^{-\tau}\right),$$

where $a_n = (2\log n)^{1/2}, b_n = (2\log n)^{1/2} - \frac{1}{2}(2\log n)^{-1/2}(\log\log n + \log 4\pi)$.

2.4 Implementation

In this section, we describe the procedures to implement the confidence bands in Theorems 1 and 2. We have written our codes in XploRe due to the convenience of using certain kernel type estimators. Information on XploRe is in Härdle, Hlávka and Klinke (2000).

Given any sample $\{(X_i, Y_i)\}_{i=1}^n$ from model (2.1.1), we use $\min(X_1, ..., X_n)$ and $\max(X_1, ..., X_n)$ respectively as the endpoints of interval $[a, b]$. Minor adjustments could be made for outliers. The number of interior knots is taken to be $N = \left[c_1 n^{1/(2p+1)}\right] + c_2$, where c_1 and c_2 are positive integers. Since explicit formula of coverage probability does not exist for the bands, there is no optimal method to select (c_1, c_2). In simulation, the simple choice of 5 for c_1 and 1 for c_2 seems to work well, so these are set as default values.

The least squares problem in (2.2.1) can be solved via the truncated power basis

$1, x, ..., x^{p-1}$,

$(x - t_j)_+^{p-1}, j = 1, ..., N$. In other words

$$\hat{m}_p(x) = \sum_{k=0}^{p-1} \hat{\gamma}_k x^k + \sum_{j=1}^{N} \hat{\gamma}_{j,p} (x - t_j)_+^{p-1}, \tag{2.4.1}$$

where the coefficients $\{\hat{\gamma}_0, ..., \hat{\gamma}_{p-1}, \hat{\gamma}_{1,p}, ..., \hat{\gamma}_{N,p}\}^T$ are solutions of the following least squares problem

$$\{\hat{\gamma}_0, ..., \hat{\gamma}_{p-1}, \hat{\gamma}_{1,p}, ..., \hat{\gamma}_{N,p}\}^T = \text{argmin} \sum_{i=1}^{n} \left\{ Y_i - \sum_{k=0}^{p-1} \gamma_k X_i^k + \sum_{j=1}^{N} \gamma_{j,p} (X_i - t_j)_+^{p-1} \right\}^2.$$

When constructing the confidence bands, one needs to evaluate the functions $\sigma_{n,p}^2(x)$ in (2.2.11). This is done differently for the exact and conservative bands, and the description is separated into two subsections. For both constant and linear bands, according to Lemmas 2.7.4, 2.7.4, one needs the unknown functions $f(x)$ and $\sigma^2(x)$. Let $\widetilde{K}(u) = 15(1 - u^2)^2 I\{|u| \leq 1\}/16$ be the quartic kernel, s_n =the sample standard deviation of $(X_i)_{i=1}^n$ and

$$\hat{f}(x) = n^{-1} \sum_{i=1}^{n} h_{\text{rot},f}^{-1} \widetilde{K}\left(\frac{X_i - x}{h_{\text{rot},f}}\right), h_{\text{rot},f} = (4\pi)^{1/10} (140/3)^{1/5} n^{-1/5} s_n \tag{2.4.2}$$

where $h_{\text{rot},f}$ is the rule-of-thumb bandwidth of Silverman (1986). Define next matrices $\mathbf{Z}_p = \{Z_{1,p},..,Z_{n,p}\}^T, p = 1,2$ with $Z_{i,p} = \{Y_i - \hat{m}_p(X_i)\}^2$ and

$$\mathbf{X} = \mathbf{X}(\mathbf{x}) = \begin{pmatrix} X_1 - x & \cdots & X_n - x \end{pmatrix}^T, \mathbf{W} = \mathbf{W}(x) = \text{diag}\left\{\widetilde{K}\left(\frac{X_i - x}{h_{\text{rot},\sigma}}\right)\right\}_{i=1}^n,$$

where $h_{\text{rot},\sigma}$ =the rule-of-thumb bandwidth of Fan and Gijbels (1996) based on data $(X_i, Z_{i,p})_{i=1}^n$. Then one defines the following estimators of $\sigma^2(x)$

$$\hat{\sigma}_p^2(x) = \left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Z}_p, p = 1,2. \tag{2.4.3}$$

Bickel and Rosenblatt (1973), Fan and Gijbels (1996) provide the following uniform consistency results

$$\max_{p=1,2}\sup_{x\in[a,b]}\left|\hat{\sigma}_p(x) - \sigma(x)\right| + \sup_{x\in[a,b]}\left|\hat{f}(x) - f(x)\right| = o_p(1). \tag{2.4.4}$$

2.4.1 Implementing Exact Bands

The function $\sigma_{n,1}(x)$ is approximated by either one of the following, with $\hat{f}(x)$ and $\hat{\sigma}_1(x)$ defined in (2.4.2) and (2.4.3), $j(x)$ defined in (2.2.2)

$$\hat{\sigma}_{n,1}(x,1) = \hat{\sigma}_1\left(t_{j(x)}\right)\hat{f}^{-1/2}\left(t_{j(x)}\right)n^{-1/2}h^{-1/2}, \tag{2.4.5}$$

$$\hat{\sigma}_{n,1}(x,2) = \hat{\sigma}_1(x)\hat{f}^{-1/2}(x)\,n^{-1/2}h^{-1/2}, \tag{2.4.6}$$

where the additional parameter value 1 or 2 indicating the estimation at each value x or at the nearest left knot. Since $\sup_{x\in[a,b]}\left|x - t_{j(x)}\right| \leq h \to 0$, as $n \to \infty$,(2.4.4) entails that both of the bands below are asymptotically exact with $\hat{m}_1(x)$ given in (2.4.1) and d_n in (2.2.13)

$$\hat{m}_1(x) \pm \hat{\sigma}_{n,1}(x,\text{opt})\left\{2\log(N+1)\right\}^{1/2}d_n, \text{opt} = 1,2. \tag{2.4.7}$$

2.4.2 Implementing Conservative Bands

According to Lemma 2.7.3, for $0 \leq j \leq N$, the matrix Ξ_j approximates matrix S_j uniformly. Hence both of the bands below are asymptotically conservative, with $\hat{m}_2(x)$ given in (2.4.1)

$$\hat{m}_2(x) \pm \hat{\sigma}_{n,2}(x, \text{opt}) \{2\log(N+1) - 2\log\alpha\}^{1/2}, \text{opt} = 1, 2, \tag{2.4.8}$$

where the function $\sigma_{n,2}(x)$ in (2.2.11) for the linear band is estimated consistently by either one of the next two formulae

$$\hat{\sigma}_{n,2}(x,1) = \left\{\Delta^T(x)\Xi_{j(x)}\Delta(x)\right\}^{1/2} \sqrt{3/2}\hat{\sigma}_2\left(t_{j(x)}\right)\hat{f}^{-1/2}\left(t_{j(x)}\right)n^{-1/2}h^{-1/2}, \tag{2.4.9}$$

$$\hat{\sigma}_{n,2}(x,2) = \left\{\Delta^T(x)\Xi_{j(x)}\Delta(x)\right\}^{1/2} \sqrt{3/2}\hat{\sigma}_2(x)\hat{f}^{-1/2}(x)n^{-1/2}h^{-1/2}, \tag{2.4.10}$$

with $\Delta(x)$ and Ξ_j defined in (2.2.9), $j(x)$ defined in (2.2.2), and $\hat{f}(x)$ and $\hat{\sigma}_2(x)$ defined in (2.4.2) and (2.4.3).

In order to calculate the matrix M_{N+2}^{-1}, which is needed for (2.2.9), we introduce two theorems from matrix theory.

Theorem 2.4.1. *[Gantmacher and Krein (1960), page 95, equation (43)] For a symmetric Jacobi matrix J given as follows*

$$J = \begin{pmatrix} a_1 & b_1 & & 0 \\ b_1 & \ddots & \ddots & \\ & \ddots & \ddots & b_{N+1} \\ 0 & & b_{N+1} & a_{N+2} \end{pmatrix}_{(N+2)\times(N+2)},$$

its inverse matrix $J^{-1} = (l_{ik})_{(N+2)\times(N+2)}$ satisfies

$$l_{i,k} = \psi_i\chi_k, i \leq k, l_{i,k} = \psi_k\chi_i, k \leq i, \tag{2.4.11}$$

where

$$\psi_i = \frac{(-1)^i \det\left(J_{(1,...,i-1)}\right) b_i b_{i+1} \cdots b_{N+1}}{\det(J)}, \chi_k = \frac{(-1)^k \det\left(J_{(k+1,...,N+2)}\right)}{b_k b_{k+1} \cdots b_{N+1}}, \tag{2.4.12}$$

and $J_{(1,...,i-1)}$ *is defined as the upper left* $(i-1) \times (i-1)$ *submatrix of* J, $\det(J)$ *is the determinant of matrix* J, *while* $J_{(k+1,...,N+2)}$ *is the corresponding lower right* $(N+2-k) \times (N+2-k)$ *submatrix.*

Theorem 2.4.2. *[Zhang (1999), page 101, Theorem 4.5] For a tridiagonal matrix given as*

$$T_N = \begin{pmatrix} a & b & & 0 \\ c & a & \ddots & \\ & \ddots & \ddots & b \\ 0 & & c & a \end{pmatrix}_{N \times N}, N \geq 1, \tag{2.4.13}$$

if $a^2 \neq 4bc$, *then the determinant of* T_N *is*

$$\det T_N = \frac{\alpha^{N+1} - \beta^{N+1}}{\alpha - \beta}, \alpha = \frac{a + \sqrt{a^2 - 4bc}}{2}, \beta = \frac{a - \sqrt{a^2 - 4bc}}{2}.$$

To apply Theorem 2.4.1 and Theorem 2.4.2, we let

$$z_1 = \frac{2+\sqrt{3}}{4}, z_2 = \frac{2-\sqrt{3}}{4}, \theta = \frac{z_2}{z_1} = \left(2-\sqrt{3}\right)^2 = 7 - 4\sqrt{3}. \tag{2.4.14}$$

For any $N \geq 1$, Theorem 2.4.2 entails that $\det(T_N) = \left(z_1^{N+1} - z_2^{N+1}\right) / (z_1 - z_2)$, if one takes $a = 1, b = c = 1/4$ in (2.4.13). Next, denote for any $N \geq 1$

$$\tilde{M}_{N+1} = \begin{pmatrix} T_N & \tilde{T}_N^T \\ \tilde{T}_N & 1 \end{pmatrix}_{(N+1)\times(N+1)}, \tilde{T}_N = \left(0, ..., 0, \sqrt{2}/4\right)_{1 \times N}$$

with the convention that $\tilde{M}_1 \equiv 1$. By the expansion of determinant of matrix $\tilde{M}_i$ along the last row and then the last column, $\forall i = 1, ..., N+1$

$$\det\left(\tilde{M}_i\right) = \det(T_{i-1}) - 8^{-1} \det(T_{i-2}) = \frac{8 z_1^{i-1} \left\{z_1 \left(1 - \theta^i\right) - \left(1 - \theta^{i-1}\right)\right\}}{8(z_1 - z_2)}.$$

The determinant of matrix M_{N+2} can be expanded along the first row and then the first column:

$$\begin{aligned}\det(M_{N+2}) &= \det\left(\tilde{M}_{N+1}\right) - 8^{-1}\det\left(\tilde{M}_N\right) \\ &= z_1^{N-1}\left\{64z_1^2\left(1-\theta^{N+1}\right) - 16z_1\left(1-\theta^N\right) + \left(1-\theta^{N-1}\right)\right\} / \left\{64(z_1-z_2)\right\}.\end{aligned}$$

Applying (2.4.12) to matrix M_{N+2} yields

$$\psi_i = \begin{cases} (-1)(1/4)^{N-1}\left(\sqrt{2}/4\right)^2 / \det(M_{N+2}), & i = 1, \\ (-1)^i (1/4)^{N+1-i}\left(\sqrt{2}/4\right)\det\left(\tilde{M}_{i-1}\right) / \det(M_{N+2}), & 2 \le i \le N, \end{cases} \tag{2.4.15}$$

$$\chi_k = \begin{cases} (-1)\left\{(1/4)^{N-1}\left(\sqrt{2}/4\right)^2\right\}^{-1}\det\left(\tilde{M}_{N+1}\right), & k = 1, \\ (-1)^k\left\{(1/4)^{N+1-k}\left(\sqrt{2}/4\right)\right\}^{-1}\det\left(\tilde{M}_{(N+2)-k}\right), & 2 \le k \le N. \end{cases} \tag{2.4.16}$$

Next, we apply (2.4.11) from Theorem 2.4.1 together with (2.4.15) and (2.4.16), for all $i, k = 1, ..., N+2$. Then the principle diagonal entries are

$$l_{k,k} = \begin{cases} \det\left(\tilde{M}_{N+1}\right) / \det(M_{N+2}), & k = 1, N+2 \\ \det\left(\tilde{M}_{(N+2)-k}\right)\det\left(\tilde{M}_{k-1}\right) / \det(M_{N+2}), & k = 2, ..., N+1 \end{cases}$$

which, after some algebra, becomes

$$l_{11} = l_{N+2,N+2} = \frac{8z_1^2\left(1-\theta^{N+1}\right) - z_1\left(1-\theta^N\right)}{8z_1^2\left(1-\theta^{N+1}\right) - 2z_1\left(1-\theta^N\right) + 8\left(1-\theta^{N-1}\right)},$$

$$l_{k,k} = \frac{\left\{8z_1\left(1-\theta^{N+2-k}\right) - \left(1-\theta^{N+1-k}\right)\right\}\left\{8z_1\left(1-\theta^{k-1}\right) - \left(1-\theta^{k-2}\right)\right\}}{(z_1-z_2)\left\{64z_1^2\left(1-\theta^{N+1}\right) - 16z_1\left(1-\theta^N\right) + 64\left(1-\theta^{N-1}\right)\right\}}. \tag{2.4.17}$$

where $2 \le k \le N+1$. Similarly, the upper diagonal entries are

$$l_{i,i+1} = l_{i+1,i} = \begin{cases} \left(-\sqrt{2}/4\right)\det\left(\tilde{M}_N\right) / \det(M_{N+2}), & i = 1, N+1 \\ (-1/4)\det\left(\tilde{M}_{(N+1)-i}\right)\det\left(\tilde{M}_{i-1}\right) / \det(M_{N+2}), & i = 2, ..., N \end{cases}$$

which, by applying again (2.4.11), (2.4.15) and (2.4.16), becomes

$$l_{12} = l_{N+1,N+2} = \frac{\left(-2\sqrt{2}\right)z_1\left(1-\theta^N\right) - \left(1-\theta^{N-1}\right)}{8z_1^2\left(1-\theta^{N+1}\right) - 2z_1\left(1-\theta^N\right) + 8\left(1-\theta^{N-1}\right)},$$

$$l_{i,i+1} = \frac{\left\{8z_1\left(1-\theta^{N+1-i}\right)-\left(1-\theta^{N-i}\right)\right\}\left\{8z_1\left(1-\theta^{i-1}\right)-\left(1-\theta^{i-2}\right)\right\}}{(-4)(z_1-z_2)\left\{64z_1^2\left(1-\theta^{N+1}\right)-16z_1\left(1-\theta^{N}\right)+64\left(1-\theta^{N-1}\right)\right\}}, \tag{2.4.18}$$

in which $2 \leq i \leq N$. By the symmetry of matrix M_{N+2}, the lower diagonal entries are $l_{i+1,i} = l_{i,i+1}$, for all $i = 1, ..., N+1$.

2.5 Simulation and Examples

2.5.1 Simulation

To illustrate the finite-sample behavior of our confidence bands, we present some simulation results. The data set is generated from model (2.1.1), with

$$m(x) = \sin(2\pi x), \sigma(x) = \sigma_0 \frac{100-\exp(x)}{100+\exp(x)}, X \sim U[-.5,.5], \varepsilon \sim N(0,1) \tag{2.5.1}$$

The noise level $\sigma_0 = 0.2, 0.5$ while sample size $n = 100, 200, 500, 10000$. Confidence level $1-\alpha = 0.99, 0.95$. Tables 4.1 and 4.2 contain the coverage probabilities as the percentage of coverage of the true curve at all data points by the confidence bands in (2.4.7) and (2.4.8), over 500 replications of sample size n. We have also computed the coverage probabilities of the confidence bands in (2.2.12) by plugging in the true value of density function $f(x) = I_{[-1/2,1/2]}(x)$ and the variance function $\sigma(x)$ in (2.5.1). These bands are called "oracle bands" as they use quantities that are unknown but for "oracles"; whereas the bands in (2.4.7) are called "estimated bands".

In Table 4.1 the surprising outcome is that all four bands have the same coverage with noise level 0.5. At noise level 0.2, the performance of all four bands becomes much closer with sample sizes increasing, whereas for small sample sizes the oracle

bands are slightly better. In Table 4.2, the coverage percentages show very positive confirmation of Theorem 2. At sample size 200, regardless of noise level, both of the two candidate bands in (2.4.8) achieve at least 95.6% and 90% for confidence level $1-\alpha = 0.99, 0.95$ respectively.

From both tables, it is obvious that larger sample size guarantees improved coverage, with reasonable coverage achieved at moderate sample sizes. Under the same circumstances, the linear band performs much better than the constant band, which corroborates with the theory. The noise level has more influence to the constant bands than the linear ones.

For the linear bands, we have also carried out simulation for sample size $n = 10000$ and opt = 1. Regardless of the noise level, the coverage is always 99.4% for $\alpha = 0.01$ and 97.6% for $\alpha = 0.05$, both higher than the nominal coverage of 99% and 95%, consistent with their conservative definitions. Remarkably, it takes merely 88 minutes to run 500 simulations with sample size as large as 10000 on a Pentium 4 PC. This is extremely fast considering that nonparametric regression is done without WARPing [Härdle, Hlávka and Klinke (2000)].

The graphs in Figure 2.4.8 are created based on two samples of size 100 and 500 respectively, each with four types of symbols: points (data), center thin solid line (true curve), center dashed line (the estimated curve), upper and lower thick solid line (confidence bands). In all figures, the confidence bands of $n = 500$ are thinner and fits better than those of $n = 100$.

2.5.2 Fossil Example

The fossil data reflect global climate millions of years ago through ratios of strontium isotopes found in fossil shell, it was studied by Chaudhuri and Marron (1999) to detect the structure via kernel smoothing. Ruppert, Wand and Carroll (2003) provide penalized spline smoothing fits to the data. In this section we test the polynomial form of the fossil data regression curve. The null hypothesis is $H_0 : m(x) = \sum_{k=1}^{d} a_k x^k$, with polynomial degree $d = 2, 3, 5, 6$. The response Y is the strontium isotopes ratio after linear transformation, $Y = 0.70715 + \text{ratio} * 10^{-5}$, since all the value are very close to 0.707, while the predictor X is the fossil shell age in million years.

In Figure 4.5, the center dotted line is the linear spline fit. The upper/lower thin lines represent linear bands based on Theorem 2, implemented according to (2.4.8). The solid line is the least square polynomial fit with degrees $2, 3, 5, 6$. Clearly, the oversmoothed quadratic null curve ($d = 2$) is rejected at significance level 0.01 since it is far away from being totally covered by the confidence bands with confidence 0.99. Though when $d = 3, 5$ the null solid curves can capture the big dip at the range of $110 - 115$ million years old, it is not a good fit even visually. Thus both null parametric models H_0 are rejected at the level 0.01. While in the case $d = 6$, all significant features are shown in the null polynomial curve, the relative high ratio before 105 million years old, the substantial dip around 115 million years old, the relative flat stage between 95 and 105. Given a 80% confidence bands the entire null curve falls between the upper and lower limits even though the bands are narrower than the those with confidence 90%, in other words for the testing we obtain a p-value

greater than 0.20. The shape of the polynomial curve with $d = 6$ is consistent with the nonparametric structure given in , Chaudhuri and Marron (1999) and Ruppert, Wand and Carroll (2003).

2.6 Conclusions

We provide exact forms of two confidence bands constructed from polynomial spline regression. Asymptotic properties have been established for equally spaced, nonadaptive selection of knots. Extension to adaptive design is infeasible, as Härdle, Marron and Yang (1997) had shown that adaptive knots selection could lead to inconsistency in L_∞ norm.

It is possible, however, to extend the constant spline band in Theorem 1 to unequally spaced deterministic knots subject to mesh constraints as in Huang (2003). The linear band in Theorem 2 does not allow such direct extension. This is one of the two reasons that the constant band remains viable despite the fact that the linear band has much better theoretical property and practical performance. The constant band is kept also for its simplicity. When implemented according to (2.4.7) with estimation on equally-spaced knots, the confidence limits at point x is the exact same as those at the nearest knot $t_{j(x)}$, so the constant band is in fact $(N + 1)$ independently inflated confidence intervals. In contrast, the piecewise linear band has to be calibrated at each new point x. That is, the confidence limits at x and the ones at $t_{j(x)}$ are different.

Extension to multivariate regression is difficult for lack of sharp approximation of

the kind in (2.3.8). This limitation is also in Xia (1998), Claeskens and Van Keilegom (2003). The main hurdle of generalizing our method to higher order splines is the inversion of the inner product matrix of B-spline basis, for which close form solutions exist in the case of linear spline with the aid of (2.4.11) and (2.4.12). The inner product matrices of the two basis in (2.2.4) are diagonal and tridiagonal respectively, while for higher order splines it becomes multi-diagonal.

2.7 Proof of Theorems

2.7.1 Preliminaries for Theorem 1

Throughout Appendices A and B, we denote by the same letters c, C, any positive constants, without distinction in each case. The detailed proof is given at http://www.msu.edu/~yangli/bandfull.pdf.

Lemma 2.7.1. *Under Assumptions (AC3) and (AC4), there exists a sequence* $\{D_n\} = \{n^{\alpha_0}\}$ *for some* $\alpha_0 > 0$, *such that the following conditions are fulfilled*

$$\sum_{n=1}^{\infty} D_n^{-(2+\delta)} < \infty, (nh)^{-1/2} \log^2 n D_n, \sqrt{nh} D_n^{-(1+\delta)}, D_n^{-\delta} h^{-1/2} \to 0. \tag{2.7.1}$$

And for any sequence $\{D_n\}$ that satisfies the above four conditions, we have

$$P\{\omega \,|\exists N(\omega), \ni |\varepsilon_i| \leq D_n, i = 1, ..., n, n > N(\omega)\} = 1.$$

Lemma 2.7.2. *As* $n \to \infty$, *for* $c_{j,n}$ *and* $d_{j,n}$ *defined in (2.2.3)*

$$c_{j,n} = f(t_j) h (1 + r_{j,n,1}), \left\langle b_{j,1}, b_{j',1} \right\rangle \equiv 0, j \neq j' \tag{2.7.2}$$

$$d_{j,n} = \frac{2}{3} f(t_{j+1}) h \begin{cases} 1 + r_{j,n,2} & j = 0, ..., N-1, \\ 1/2 + r_{j,n,2} & j = -1, N, \end{cases} \tag{2.7.3}$$

$$\left\langle b_{j,2}, b_{j',2} \right\rangle = \frac{1}{6} f\left(t_{j+1}\right) h \begin{cases} 1 + \tilde{r}_{j,n,2} & \left|j' - j\right| = 1, \\ 0 & \left|j' - j\right| > 1, \end{cases} \tag{2.7.4}$$

where

$$\max_{0 \le j \le N} \left|r_{j,n,1}\right| + \max_{-1 \le j \le N} \left|r_{j,n,2}\right| + \max_{-1 \le j \le N-1} \left|\tilde{r}_{j,n,2}\right| \le C\omega\left(f, h\right). \tag{2.7.5}$$

In particular,

$$\frac{1}{3} f\left(t_{j+1}\right) h \left\{1 - C\omega\left(f, h\right)\right\} \le d_{j,n} \le \frac{2}{3} f\left(t_{j+1}\right) h \left\{1 + C\omega\left(f, h\right)\right\}. \tag{2.7.6}$$

PROOF OF LEMMA 2.3.1. For brevity, we give only the proof of (2.3.1) for $A_{n,1}$. Take any $j = 0, 1, ..., N$

$$\left| \left\| B_{j,1} \right\|_{2,n}^2 - 1 \right| = \left| \sum_{i=1}^{n} \xi_i \right|, \xi_i = \left\{ B_{j,1}^2 \left(X_i\right) - 1 \right\} n^{-1}$$

with $E\xi_i = 0$ and for any $k \ge 2$, Minkowski's inequality implies that

$$E\left|\xi_i\right|^k = n^{-k} E \left| B_{j,1}^2 \left(X_i\right) - 1 \right|^k \le \left(2/n\right)^k 2^{-1} E\left[B_{j,1}^{2k} \left(X_i\right) + 1 \right] \le \left\{ \frac{2}{nh} \right\}^k C_0 h,$$

while (2.7.2) entails that $E\xi_i^2 \ge n^{-2} E\left[\frac{1}{2} B_{j,1}^4 \left(X_i\right) - 1 \right] \ge \left\{2/\left(nh\right)\right\}^2 C_1 h$.

It is then clear that one can find a constant $c > 0$ such that for all $k > 2$, $E\left|\xi_i\right|^k \le \left(cn^{-1}h^{-1}\right)^{k-2} k! E\left|\xi_i\right|^2$. Applying Bernstein's inequality to $\sum_{i=1}^{n} \xi_i$, for any large enough $\delta > 0$

$$P\left\{ \left| \sum_{i=1}^{n} \xi_i \right| \ge \delta \sqrt{\left(nh\right)^{-1} \log\left(n\right)} \right\} \le 2n^{-3}$$

$$\Rightarrow \quad \sum_{n=1}^{\infty} P\left\{ \sup_{0 \le j \le N} \left| \left\| B_{j,1} \right\|_{2,n}^2 - 1 \right| \ge \delta \sqrt{\left(nh\right)^{-1} \log\left(n\right)} \right\} < \infty$$

for such $\delta > 0$, then (2.3.1) follows. □

2.7.2 Proof of Theorem 1

In this section, we will investigate the asymptotic behavior of $\tilde{\varepsilon}_1(x)$ defined in (2.3.5). Since

$\left\langle \mathbf{B}_{j',1}(\mathbf{X}), \mathbf{B}_{j,1}(\mathbf{X}) \right\rangle_n = 0$ unless $j = j'$, $\tilde{\varepsilon}_1(x)$ can be written as

$$\tilde{\varepsilon}_1(x) = \sum_{j=0}^{N} \varepsilon_j^* B_{j,1}(x) \left\| B_{j,1} \right\|_{2,n}^{-2}$$

in which

$$\varepsilon_j^* = \left\langle \mathbf{E}, \mathbf{B}_{j,1}(\mathbf{X}) \right\rangle_n = \frac{1}{n} \sum_{i=1}^{n} B_{j,1}(X_i) \sigma(X_i) \varepsilon_i.$$

Lemma 2.7.3. *Let* $\hat{\varepsilon}_1(x) = \sum_{j=0}^{N} \varepsilon_j^* B_{j,1}(x), x \in [a,b]$ *then*

$$\left|\tilde{\varepsilon}_1(x) - \hat{\varepsilon}_1(x)\right| \leq A_{n,1}\left(1 - A_{n,1}\right)^{-1} \left|\hat{\varepsilon}_1(x)\right|, x \in [a,b],$$

where $A_{n,1}$ *is defined in (2.3.1).*

The asymptotic behavior of $\sup_{x\in[a,b]} \left|\tilde{\varepsilon}_1(x)\right|$ therefore is the same as that of $\sup_{x\in[a,b]} \left|\hat{\varepsilon}_1(x)\right|$.

Lemma 2.7.4. *The pointwise variance of* $\hat{\varepsilon}_1(x)$ *is the function* $\sigma_{n,1}^2(x)$ *defined in (2.2.11) which satisfies*

$$E\left\{\hat{\varepsilon}_1(x)\right\}^2 \equiv \sigma_{n,1}^2(x) = \frac{\sigma^2(x)}{f(x)nh}\left\{1 + r_{n,1}(x)\right\}, x \in [a,b] \tag{2.7.7}$$

with $\sup_{x\in[a,b]} \left|r_{n,1}(x)\right| \to 0$.

PROOF. The term $E\left\{\hat{\varepsilon}_1(x)\right\}^2$ has the expression for $\sigma_{n,1}^2(x)$ in (2.2.11). By (2.7.5) and the continuity of functions $\sigma^2(x)$ and $f(x)$, $\sigma_{n,1}^2(x)$ can be expressed as

$$\frac{\sigma^2(x) f(x) h + \int_{J_{j(x)}} \left\{\sigma^2(v) f(v) - \sigma^2(x) f(x)\right\} dv}{n\left\{f\left(t_{j(x)}\right) h + r_{j(x),n,1}\right\}^2} = \frac{\sigma^2(x)}{nf(x)h}\left\{1 + r_{n,1}(x)\right\},$$

with $\sup_{x\in[a,b]}\left|r_{n,1}(x)\right|\to 0$, establishing (2.7.7). □

Lemma 2.7.5. *Let the sequence $\{D_n\}$ satisfy (2.7.1) and define for $x\in[a,b]$*

$$\begin{aligned}\hat{\varepsilon}_{n,1}(x) &= \sigma_{n,1}(x)^{-1}\sum_{j=0}^{N}B_{j,1}(x)\,\varepsilon_j^* = \sigma_{n,1}(x)^{-1}\sum_{j=0}^{N}B_{j,1}(x)\left(\varepsilon_j^*-E\varepsilon_j^*\right),\\ \hat{\varepsilon}_{n,1}^{D}(x) &= \sigma_{n,1}(x)^{-1}\sum_{j=0}^{N}B_{j,1}(x)\left(\varepsilon_j^*-E\varepsilon_j^*\right)I_{\{|\varepsilon|<D_n\}}\end{aligned}\tag{2.7.8}$$

then with probability 1

$$\left\|\hat{\varepsilon}_{n,1}(x)-\hat{\varepsilon}_{n,1}^{D}(x)\right\|_{\infty}=O\left(D_n^{-(1+\delta)}\sqrt{nh}\right)=o(1).$$

PROOF. Notice that $E\varepsilon_j^*=E\{\frac{1}{n}\sum_{i=1}^{n}B_{j,1}(X_i)\,\sigma(X_i)\,\varepsilon_i\}=0$ since $E(\varepsilon_i|X_i)=0$, then

$$\hat{\varepsilon}_{n,1}(x)=\left\{\sigma_{n,1}(x)\sqrt{n}c_{j(x),n}\right\}^{-1}\int\int I_{j(x)}(v)\,\sigma(v)\,\varepsilon dZ_n(v,\varepsilon)$$

according to the definition of $Z_n(v,\varepsilon)$ in (2.3.10). The process $\hat{\varepsilon}_{n,1}(x)$ is separated into two parts $\hat{\varepsilon}_{n,1}(x)=\hat{\varepsilon}_{n,1}^{D}(x)+\left\{\hat{\varepsilon}_{n,1}(x)-\hat{\varepsilon}_{n,1}^{D}(x)\right\}$. The truncated part $\hat{\varepsilon}_{n,1}^{D}(x)$ is defined in (2.7.8). The tail part $\hat{\varepsilon}_{n,1}(x)-\hat{\varepsilon}_{n,1}^{D}(x)$ is bounded uniformly over $[a,b]$ by

$$\begin{aligned}&\sup_{x\in[a,b]}\left|\left\{\sigma_{n,1}(x)\sqrt{n}c_{j(x),n}\right\}^{-1}\int\int I_{j(x)}(v)\,\sigma(v)\,\varepsilon I_{\{|\varepsilon|\geq D_n\}}dZ_n(v,\varepsilon)\right|\\ &\leq\sup_{x\in[a,b]}\left|\left\{\sigma_{n,1}(x)\,c_{j(x),n}\right\}^{-1}\frac{1}{n}\sum_{i=1}^{n}I_{j(x)}(X_i)\,\sigma(X_i)\,\varepsilon_i I_{\{|\varepsilon_i|\geq D_n\}}\right| \qquad (2.7.9)\\ &+\sup_{x\in[a,b]}\left|\left\{\sigma_{n,1}(x)\,c_{j(x),n}\right\}^{-1}\int\int I_{j(x)}(v)\,\sigma(v)\,\varepsilon I_{\{|\varepsilon|\geq D_n\}})dF(v,\varepsilon)\right| \qquad (2.7.10)\end{aligned}$$

By Lemma 2.7.1, the term in (2.7.9) is 0 almost surely. The term in (2.7.10) is

bounded by

$$\sup_{x\in[a,b]}\left\{\sigma_{n,1}(x)\,c_{j(x),n}\right\}^{-1}\int I_{j(x)}(v)\,\sigma(v)\,f(v)\left[\int|\varepsilon|\,I_{\{|\varepsilon|\geq D_n\}}dF(\varepsilon\,|v)\right]dv$$

$$\leq \sup_{x\in[a,b]}\left\{\sigma_{n,1}(x)\,c_{j(x),n}\right\}^{-1}\int I_{j(x)}(v)\,\sigma(v)\,f(v)\,dv\frac{M_\delta}{D_n^{1+\delta}}\leq C\frac{\sqrt{nh}}{D_n^{1+\delta}}.$$

The lemma follows immediately by the third condition in (2.7.1). □

Lemma 2.7.6. *Define for* $x\in[a,b]$

$$\hat{\varepsilon}_{n,1}^{(0)}(x)=\left\{\sigma_{n,1}(x)\sqrt{n}c_{j(x),n}\right\}^{-1}\int\int I_{j(x)}(v)\,\sigma(v)\,\varepsilon I_{\{|\varepsilon|<D_n\}}dB\{M(v,\varepsilon)\} \tag{2.7.11}$$

then with probability 1

$$\sup_{x\in[a,b]}\left|\hat{\varepsilon}_{n,1}^{(0)}(x)-\hat{\varepsilon}_{n,1}^{D}(x)\right|=O\left(h^{-1/2}n^{-1/2}D_n\log^2 n\right)=o(1).$$

PROOF. First, $\sup_{x\in[a,b]}\left|\hat{\varepsilon}_{n,1}^{(0)}(x)-\hat{\varepsilon}_{n,1}^{D}(x)\right|$ can be written as

$$\sup_{x\in[a,b]}\left|\left\{\sigma_{n,1}(x)\sqrt{n}c_{j(x),n}\right\}^{-1}\int\int I_{j(x)}(v)\,\sigma(v)\,\varepsilon I_{\{|\varepsilon|<D_n\}}d\left[Z_n(v,\varepsilon)-B\{M(v,\varepsilon)\}\right]\right|,$$

which the double integration becomes the following via integration by parts

$$\sup_{x\in[a,b]}\left|\int\int\left[Z_n(v,\varepsilon)-B\{M(v,\varepsilon)\}\right]d\left\{I_{j(x)}(v)\,\sigma(v)\,\varepsilon I_{\{|\varepsilon|<D_n\}}\right\}\right|$$

$$\leq \sup_{x\in[a,b]}$$

$$\times\int\int\left|Z_n(v,\varepsilon)-B\{M(v,\varepsilon)\}\right|d\left\{\varepsilon I_{\{|\varepsilon|<D_n\}}\right\}d\left\{I_{j(x)}(v)\,\sigma(v)\right\}.$$

Next, by Lemma 2.7.4, the bounded variation of the function $\sigma(x)$ in Assumption (AC2), the strong approximation result (2.3.11) and the first condition in (2.7.1), the above term is bounded as

$$O\left\{(nh)^{1/2}n^{-1/2}h^{-1}\left(n^{-1/2}\log^2 n\right)D_n\right\}=O\left(n^{-1/2}h^{-1/2}D_n\log^2 n\right)=o(1)\text{ w. p. 1},$$

thus completing the proof of the lemma. □

The next lemma finds a process $\hat{\varepsilon}_{n,1}^{(1)}(x)$ defined in terms of the 2-dimensional Brownian motion to approximate $\hat{\varepsilon}_{n,1}^{(0)}(x)$ in (2.7.11).

Lemma 2.7.7. *Define for* $x \in [a,b]$

$$\hat{\varepsilon}_{n,1}^{(1)}(x) = \left\{\sigma_{n,1}(x)\sqrt{n}c_{j(x),n}\right\}^{-1} \int\int I_{j(x)}(v)\,\sigma(v)\,\varepsilon I_{\{|\varepsilon|<D_n\}} dW\{M(v,\varepsilon)\}$$

then with probability 1

$$\left\|\hat{\varepsilon}_{n,1}^{(1)}(x) - \hat{\varepsilon}_{n,1}^{(0)}(x)\right\|_{\infty} = O\left(h^{1/2}D_n^{-(1+\delta)}\right) = o(1).$$

PROOF. Based on the Rosenblatt transformation $M(x,\varepsilon)$ defined in (2.3.9), and $\frac{\partial M(x,\varepsilon)}{\partial(x,\varepsilon)} = f(x,\varepsilon)$, then the term $\left\|\hat{\varepsilon}_{n,1}^{(1)}(x) - \hat{\varepsilon}_{n,1}^{(0)}(x)\right\|_{\infty}$ is bounded by

$$\begin{aligned}
&\sup_{x\in[a,b]}\left|\left\{\sigma_{n,1}(x)\sqrt{n}c_{j(x),n}\right\}^{-1}\int\int I_{j(x)}(v)\,\sigma(v)\,|\varepsilon|\,I_{\{|\varepsilon|<D_n\}}dM(v,\varepsilon)\,W(1,1)\right| \\
\leq\ &\sup_{x\in[a,b]}\left\{\sigma_{n,1}(x)\sqrt{n}c_{j(x),n}\right\}^{-1}\int I_{j(x)}(v)\,\sigma(v)\,f(v)\,dv \\
&\times\left\{\int|\varepsilon|\,I_{\{|\varepsilon|<D_n\}}f_{\varepsilon|v}(\varepsilon|v)d\varepsilon\right\}|W(1,1)| \\
\leq\ &C\left(\frac{\sqrt{nh}}{\sqrt{n}h}\right)h\frac{M_\delta}{D_n^{1+\delta}}|W(1,1)| = O\left(h^{1/2}D_n^{-(1+\delta)}\right) = o(1)\ \text{w. p. 1}
\end{aligned}$$

The last step is obtained by applying the third condition in (2.7.1). □

The next lemma expresses the distribution of $\hat{\varepsilon}_{n,1}^{(1)}(x)$ in terms of 1-dimensional Brownian motion.

Lemma 2.7.8. *The process* $\hat{\varepsilon}_{n,1}^{(1)}(x)$ *has the same probability structure as the process*

$$\hat{\varepsilon}_{n,1}^{(2)}(x) = \left\{\sigma_{n,1}(x)\sqrt{n}c_{j(x),n}\right\}^{-1}\int I_{j(x)}(v)\,\sigma(v)\,s_n(v)\,f^{\frac{1}{2}}(v)\,dW(v), x\in[a,b]$$

where

$$s_n^2(v) = \int \varepsilon^2 I_{\{|\varepsilon|<D_n\}} f_{\varepsilon|v}(\varepsilon|v) d\varepsilon. \tag{2.7.12}$$

PROOF. By applying Itô's Isometry Theorem, it is obtained that var $\left\{\hat{\varepsilon}_{n,1}^{(1)}(x)\right\}$ and var $\left\{\hat{\varepsilon}_{n,1}^{(2)}(x)\right\}$ are exactly the same for any $x \in [a,b]$. Hence, the two Gaussian processes $\hat{\varepsilon}_n^{(1)}(x)$ and $\hat{\varepsilon}_n^{(2)}(x)$ have the same probability structure. □

Lemma 2.7.9. *Define for any* $x \in [a,b]$

$$\hat{\varepsilon}_{n,1}^{(3)}(x) = \left\{\sigma_{n,1}(x)\sqrt{n}c_{j(x),n}\right\}^{-1} \int I_{j(x)}(v)\sigma(v) f^{\frac{1}{2}}(v) dW(v)$$

then

$$\left\|\hat{\varepsilon}_{n,1}^{(2)}(x) - \hat{\varepsilon}_{n,1}^{(3)}(x)\right\|_\infty = O\left(\frac{1}{D_n^\delta \sqrt{h}}\right) = o(1) \ \ w.\ p.\ 1.$$

PROOF. By the fourth condition in (2.7.1) , $\sup_{x\in[a,b]} \left|\hat{\varepsilon}_{n,1}^{(2)}(x) - \hat{\varepsilon}_{n,1}^{(3)}(x)\right|$ is almost surely bounded by

$$\begin{aligned}
&\sup_{v\in[a,b]} \left|s_n^2(v) - 1\right| \sup_{x\in[a,b]} \left|\sigma_{n,1}^{-1}(x) c_{j(x),n}^{-1} n^{-1/2} \int I_{j(x)}(v)\sigma(v) f^{\frac{1}{2}}(v) dW(v)\right| \\
&= O\left(D_n^{-\delta} h^{-1/2}\right) = o(1)
\end{aligned}$$

Lemma 2.7.10. *The process* $\hat{\varepsilon}_{n,1}^{(3)}(x)$ *is a Gaussian process with mean* 0, *variance* 1, *and covariance*

$$cov\left\{\hat{\varepsilon}_{n,1}^{(3)}(x), \hat{\varepsilon}_{n,1}^{(3)}(y)\right\} = \delta_{j(x),j(y)}, \forall x, y \in [a,b].$$

PROOF. The variance and covariance are given by Itô's Isometry Theorem

$$\text{var}\left\{\hat{\varepsilon}_{n,1}^{(3)}(x)\right\} = \left\{\sigma_{n,1}(x)\sqrt{n}c_{j(x),n}\right\}^{-2} \int I_{j(x)}(v)\sigma^2(v) f(v) dv = 1$$

according to (2.7.7). Likewise the covariance $\operatorname{cov}\left\{\hat{\varepsilon}_{n,1}^{(3)}(x), \hat{\varepsilon}_{n,1}^{(3)}(y)\right\}$ is

$$\begin{aligned}
&\left\{\sigma_{n,1}(x)\,\sigma_{n,1}(y)\,nc_{j(x),n}c_{j(y),n}\right\}^{-1} \\
&\times E\left\{\int_{J_{j(x)}} \sigma(v)\, f^{\frac{1}{2}}(v)\, dW(v) \int_{J_{j(y)}} \sigma(v)\, f^{\frac{1}{2}}(v)\, dW(v)\right\} \\
&= \left\{\sigma_{n,1}(x)\,\sigma_{n,1}(y)\,nc_{j(x),n}c_{j(y),n}\right\}^{-1} \int_{J_{j(x)}\cap J_{j(y)}} \sigma^2(v)\, f(v)\, dv = \delta_{j(x),j(y)}
\end{aligned}$$

which completes the proof. □

PROOF OF PROPOSITION 2.3.1. The proof follows immediately from Lemmas 2.7.3, 2.7.5, 2.7.6, 2.7.7, 2.7.8, 2.7.9 and 2.7.10. □

PROOF OF THEOREM 1. It is clear from Proposition 2.3.1 that the Gaussian process $U(x)$ consists of $(N+1)$ i.i.d. standard normal variables $U(t_0), ..., U(t_N)$, hence Theorem 2.3.4 implies that as $n \to \infty$

$$P\left\{\sup_{x\in[a,b]} |U(x)| \le \tau/a_{N+1} + b_{N+1}\right\} \to \exp\left(-2e^{-\tau}\right).$$

By letting $\tau = -\log\left\{-\frac{1}{2}\log(1-\alpha)\right\}$, and using the definition of a_{N+1} and b_{N+1}, we obtain

$$\begin{aligned}
\lim_{n\to\infty} P\Bigg[\sup_{x\in[a,b]} |U(x)| \le -\log\left\{-\frac{1}{2}\log(1-\alpha)\right\}\{2\log(N+1)\}^{-1/2} \\
+ \{2\log(N+1)\}^{1/2} - \frac{1}{2}\{2\log(N+1)\}^{-1/2}\{\log\log(N+1) + \log 4\pi\}\Bigg] = 1-\alpha.
\end{aligned}$$

Replacing $U(x)$ with $\sigma_{n,1}(x)^{-1}\tilde{\varepsilon}_1(x)$ (Proposition 2.3.1), and the definition of d_n in (2.2.13) entail that

$$\lim_{n\to\infty} P\left[\sup_{x\in[a,b]} \left|\sigma_{n,1}(x)^{-1}\tilde{\varepsilon}_1(x)\right| \le \{2\log(N+1)\}^{1/2} d_n\right] = 1-\alpha.$$

According to (2.3.6), it implies that $(nh)^{-1/2}\sqrt{\log(N+1)}\,\|\tilde{m}_1(x)-m(x)\|_\infty = o_p(1)$. Thus according to (2.3.4)

$$\begin{aligned}
&\lim_{n\to\infty} P\left[m(x)\in\hat{m}_1(x)\pm\sigma_{n,1}(x)\{2\log(N+1)\}^{1/2}d_n,\forall x\in[a,b]\right]\\
=&\lim_{n\to\infty} P\left[\{2\log(N+1)\}^{-1/2}d_n^{-1}\sup_{x\in[a,b]}\sigma_{n,1}^{-1}(x)\,|\tilde{\varepsilon}_1(x)+\tilde{m}_1(x)-m(x)|\le 1\right]\\
=&\lim_{n\to\infty} P\left[\{2\log(N+1)\}^{-1/2}d_n^{-1}\sup_{x\in[a,b]}\sigma_{n,1}^{-1}(x)\,|\tilde{\varepsilon}_1(x)|\le 1\right]=1-\alpha. \quad \square
\end{aligned}$$

2.7.1 Preliminaries for Theorem 2

In this subsection we examine some matrices used in the construction of confidence band in (2.2.14) and in the proof of Theorem 2.

The next lemma corresponds to (2.2.5) for piecewise constant basis. In what follows, we use $|T|$ to denote the maximal absolute value of any matrix T, and M_{N+2} is the tridiagonal matrix as defined in (2.2.10).

Lemma 2.7.1. *The inner product matrix V of the B-spline basis $\{B_{j,2}(x)\}_{j=-1}^{N}$ defined in (2.2.6) has the following decomposition*

$$V = M_{N+2} + \left(\tilde{v}_{j'j}\right)_{j,j'=-1}^{N} = M_{N+2} + \tilde{V} \tag{2.7.1}$$

where $\tilde{v}_{j'j}\equiv 0$ if $|j-j'|\ge 1$, and

$$\left|\tilde{V}\right| \le C\omega(f,h). \tag{2.7.2}$$

PROOF. By (2.7.3), (2.7.4) and (2.7.5), the inner product of $\left\langle b_{j',2}, b_{j,2}\right\rangle$ can be replaced by $\frac{1}{6}f(t_{j+1})h$ if $|j'-j|=1$, and $\frac{1}{3}f(t_{j+1})h$ or $\frac{2}{3}f(t_{j+1})h$ when $j'=j$, plus some uniformly infinitesimal differences dominated by $\omega(f,h)$. Then based on the definition of $B_{j,2}(x)$, the lemma follows immediately. $\square$

The next lemma shows that multiplication by M_{N+2} behaves similarly to multiplication by a constant.

Lemma 2.7.2. *Given matrix* $\Omega = M_{N+2} + \Gamma$, *in which* $\Gamma = \left(\gamma_{jj'}\right)_{j,j'=-1}^{N}$ *satisfies* $\gamma_{jj'} \equiv 0$ *if* $\left|j - j'\right| \geq 1$ *and* $|\Gamma| \xrightarrow{p} 0$. *Then there exist constants* $c, C > 0$ *independent of* n *and* Γ, *such that in probability*

$$c\,|\xi| \leq |\Omega\xi| \leq C\,|\xi|\,, C^{-1}\,|\xi| \leq \left|\Omega^{-1}\xi\right| \leq c^{-1}\,|\xi|\,, \forall \xi \in R^{N+2}. \tag{2.7.3}$$

PROOF. Since each row of M_{N+2} has diagonal element equal to 1, and one or two nonzero off-diagonal terms whose total absolute values do not exceed $2\sqrt{2}/4 = 1/\sqrt{2}$, hence

$$\left(1 - 1/\sqrt{2} - 3\,|\Gamma|\right)|\xi| \leq |\Omega\xi| \leq 3\,(1 + |\Gamma|)\,|\xi|\,,$$

which entails the left inequality of (2.7.3), and the right one follows by switching the roles of ξ and $\Omega\xi$. □

As an application of Lemma 2.7.2, consider the matrix $S = V^{-1}$ defined in (2.2.7). Let $\tilde{\xi}_{j'} = \left\{\operatorname{sgn}\left(s_{j'j}\right)\right\}_{j=-1}^{N}$, then there exists a positive C_s such that

$$\sum_{j=-1}^{N}\left|s_{j'j}\right| \leq \left|S\tilde{\xi}_{j'}\right| \leq C_s\left|\tilde{\xi}_{j'}\right| = C_s, \forall j' = -1, 0, ..., N. \tag{2.7.4}$$

The matrix S appears in the construction of the confidence band, but it can not be computed exactly as it involves the unknown density $f(x)$. We approximate S with the inverse of M_{N+2}, with a simpler, distribution-free form in (2.2.10). This approximation is uniform for S_j in (2.2.7) and Ξ_j (2.2.9) as well.

Lemma 2.7.3. *As* $n \to \infty$, $\left|M_{N+2}^{-1} - S\right| \to 0$ *and* $\max_{0 \leq j \leq N}\left|\Xi_j - S_j\right| \to 0$.

PROOF. By definition,

$$M_{N+2}M_{N+2}^{-1} = I = VS = \left(M_{N+2} + \tilde{V}\right) S.$$

Denote by e_i the unit vector with i-th element 1, then applying Lemma 2.7.2 with $\Omega = M_{N+2}$, one derives

$$\begin{aligned} & c\left|M_{N+2}^{-1} - S\right| = c \max_{i=1}^{N+2} \left|\left(M_{N+2}^{-1} - S\right) e_i\right| \\ \leq & \max_{i=1}^{N+2} \left|M_{N+2}\left(M_{N+2}^{-1} - S\right) e_i\right| \leq \left|\tilde{V}\right| \left(\left|M_{N+2}^{-1} - S\right| + \left|M_{N+2}^{-1}\right|\right) \end{aligned}$$

Since (2.7.2) makes $\left|\tilde{V}\right| \leq C\omega(f,h)$, as $n \to \infty$

$$\left|M_{N+2}^{-1} - S\right| \leq \frac{C\omega(f,h)}{c - C\omega(f,h)} \left|M_{N+2}^{-1}\right| = O\{\omega(f,h)\} \to 0.$$

Now by definition of submatrices S_j and Ξ_j, $\max_{0 \leq j \leq N} |\Xi_j - S_j| \leq \left|M_{N+2}^{-1} - S\right|$, the lemma follows. □

2.7.2 Variance Calculation

We now examine the asymptotic behavior of $\text{Proj}_{G_n^{(0)}}\mathbf{E}$, which is

$$\tilde{\varepsilon}_2(x) = \text{Proj}_{G_n^{(0)}}\mathbf{E} = \sum_{j=-1}^{N} \tilde{a}_j B_{j,2}(x), x \in [a,b] \tag{2.7.5}$$

where the spline coefficient vector $\tilde{\mathbf{a}} = (\tilde{a}_{-1}, ..., \tilde{a}_N)^T$ are solutions to the normal equations

$$\left(\left\langle B_{j,2}, B_{j',2}\right\rangle_n\right)_{j,j'=-1}^{N} \begin{pmatrix} \tilde{a}_{-1} \\ \vdots \\ \tilde{a}_N \end{pmatrix} = \left(\frac{1}{n}\sum_{i=1}^{n} B_{j,2}(X_i)\,\sigma(X_i)\,\varepsilon_i\right)_{j=-1}^{N}.$$

In other words

$$\tilde{\mathbf{a}} = \begin{pmatrix} \tilde{a}_{-1} \\ \vdots \\ \tilde{a}_N \end{pmatrix} = \left(V + \tilde{B}\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} B_{j,2}(X_i)\,\sigma(X_i)\,\varepsilon_i\right)_{j=-1}^{N}, \tag{2.7.6}$$

where $\left|\tilde{B}\right| \leq A_{n,2} = O_p\left(\sqrt{n^{-1}h^{-1}\log(n)}\right)$ by (2.3.2).

Now define $\hat{a}_j$'s by replacing $\left(V + \tilde{B}\right)^{-1}$ with $V^{-1} = S$ in above formula, i.e.

$$\hat{\mathbf{a}} = \begin{pmatrix} \hat{a}_{-1} \\ \vdots \\ \hat{a}_N \end{pmatrix} = \left(\sum_{j=-1}^{N} s_{j'j}\frac{1}{n}\sum_{i=1}^{n} B_{j,2}(X_i)\,\sigma(X_i)\,\varepsilon_i\right)_{j'=-1,..,N} \tag{2.7.7}$$

and define for $x \in [a,b]$

$$\hat{\varepsilon}_2(x) = \sum_{j=-1}^{N} \hat{a}_j B_{j,2}(x) = \sum_{j,j'=-1}^{N} s_{j'j}\frac{1}{n}\sum_{i=1}^{n} B_{j,2}(X_i)\,\sigma(X_i)\,\varepsilon_i B_{j',2}(x). \tag{2.7.8}$$

In order to calculate the variance of $\hat{\varepsilon}_2(x)$, we express the matrix Σ defined in (2.2.8) as

$$\Sigma = \Theta_n V \Theta_n + \left(\tilde{\sigma}_{jl}\right)_{j,j'=-1}^{N} = \Theta_n V \Theta_n + \tilde{\Sigma}, \Theta_n = \operatorname{diag}\left\{\sigma(t_0), \ldots, \sigma(t_{N+1})\right\}, \tag{2.7.9}$$

where

$$\tilde{\sigma}_{jl} \equiv 0 \text{ if } \left|j - j'\right| 1, \sup_{j,l=-1}^{N} \left|\tilde{\sigma}_{jl}\right| \leq C\left\{\omega(f,h) + \omega\left(f\sigma^2, h\right)\right\}. \tag{2.7.10}$$

The next lemma is a special case of the unconditional version of equation (6.2) in Huang (2003).

Lemma 2.7.4. *The pointwise variance of $\hat{\varepsilon}_2(x)$ is the function $\sigma_{n,2}^2(x)$ defined in (2.2.11), which satisfies*

$$E\left\{\hat{\varepsilon}_2^2(x)\right\} \equiv \sigma_{n,2}^2(x) = \frac{3\sigma^2(x)}{2f(x)nh}\Delta^T(x)\,S_{j(x)}\Delta(x)\left\{1 + r_{n,2}(x)\right\} \tag{2.7.11}$$

with $\sup_{x\in[a,b]}\left|r_{n,2}(x)\right|\to 0$, $j(x)$ *is as defined in (2.2.2),* $\Delta(x)$ *as defined in (2.2.9) and matrix* S_j *in (2.2.7). Consequently, there exist positive constants* c_σ *and* C_σ *such that for large enough* n

$$c_\sigma (nh)^{-1/2} \leq \sigma_{n,2}(x) \leq C_\sigma (nh)^{-1/2}, \forall x \in [a,b]. \tag{2.7.12}$$

PROOF. See Wang and Yang (2005). □

2.7.3 Proof of Theorem 2

Several lemmas will be given below for the proof of Proposition 2.3.2.

Lemma 2.7.5. *Define for* $x \in [a,b]$

$$\begin{aligned}\hat{\varepsilon}_{n,2}(x) &= \sigma_{n,2}^{-1}(x)\hat{\varepsilon}_2(x) = \sigma_{n,2}^{-1}(x)\sum_{j'=-1}^{N}\hat{a}_{j'}B_{j',2}(x),\\ \hat{\varepsilon}_{n,2}^{D}(x) &= \sigma_{n,2}^{-1}(x)\sum_{j'=-1}^{N}\hat{a}_{j'}B_{j',2}(x)I_{\{|\varepsilon|<D_n\}}.\end{aligned} \tag{2.7.13}$$

where D_n *satisfies (2.7.1). Then with probability* 1

$$\left\|\hat{\varepsilon}_{n,2}(x)-\hat{\varepsilon}_{n,2}^{D}(x)\right\|_\infty = O\left(n^{1/2}h^{1/2}D_n^{-(1+\delta)}\right) = o(1).$$

PROOF. Since obviously $E\hat{\varepsilon}_{n,2}(x)=0, \forall x\in[a,b]$,

$$\hat{\varepsilon}_{n,2}(x) = \sigma_{n,2}^{-1}(x)n^{-1/2}\sum_{j'=j(x)-1}^{j(x)}B_{j',2}(x)\sum_{j=-1}^{N}s_{j'j}\int\int B_{j,2}(v)\sigma(v)\varepsilon dZ_n(v,\varepsilon)$$

where $Z_n(x,\varepsilon)$ is defined in (2.3.10). The technical proof is very similar to Lemma 2.7.5, except that we employ (2.7.4) to deal with $\sum_{j=-1}^{N}s_{j'j}$. The same order is also achieved. □

Lemma 2.7.6. *Let M be the Rosenblatt transformation given in (2.3.9) and define for $x \in [a,b]$*

$$\hat{\varepsilon}_{n,2}^{(0)}(x) = \left\{\sqrt{n}\sigma_{n,2}(x)\right\}^{-1} \sum_{j'j=-1}^{N} B_{j',2}(x)\, s_{j'j} \int\int B_{j,2}(v)\,\sigma(v)\,\varepsilon I_{\{|\varepsilon|<D_n\}} dB\left\{M(v,\varepsilon)\right\}.$$

Then with probability 1

$$\sup_{x\in[a,b]} \left|\hat{\varepsilon}_{n,2}^{(0)}(x) - \hat{\varepsilon}_{n,2}^{D}(x)\right| = O\left(n^{-1/2}h^{-1/2}D_n \log^2 n\right) = o(1).$$

PROOF. See Lemma 2.7.6. □

Lemma 2.7.7. 2.7.7*Define for $x \in [a,b]$*

$$\hat{\varepsilon}_{n,2}^{(1)}(x) = \frac{\sigma_{n,2}^{-1}(x)}{\sqrt{n}} \sum_{j'j=-1}^{N} B_{j',2}(x)\, s_{j'j} \int\int B_{j,2}(v)\,\sigma(v)\,\varepsilon I_{\{|\varepsilon|<D_n\}} dW\left\{M(v,\varepsilon)\right\},$$

then with probability 1

$$\sup_{x\in[a,b]} \left|\hat{\varepsilon}_{n,2}^{(1)}(x) - \hat{\varepsilon}_{n,2}^{(0)}(x)\right| = O\left(h^{1/2}D_n^{-(1+\delta)}\right) = o(1).$$

Lemma 2.7.8. *The process $\hat{\varepsilon}_{n,2}^{(1)}(x), x \in [a,b]$ has the same probability structure as*

$$\hat{\varepsilon}_{n,2}^{(2)}(x) = \frac{\sigma_{n,2}^{-1}(x)}{\sqrt{n}} \sum_{j'j=-1}^{N} B_{j',2}(x)\, s_{j'j} \int\int B_{j,2}(v)\,\sigma(v)\,s_n(v)\,f^{\frac{1}{2}}(v)\,dW(v), x \in [a,b]$$

where $s_n^2(v)$ is as defined in (2.7.12).

PROOF. Use Itô's Isometry Theorem again. □

Lemma 2.7.9. *Define for any $x \in [a,b]$*

$$\hat{\varepsilon}_{n,2}^{(3)}(x) = \frac{\sigma_{n,2}^{-1}(x)}{\sqrt{n}} \sum_{j'j=-1}^{N} B_{j',2}(x)\, s_{j'j} \int B_{j,2}(v)\,\sigma(v)\,f^{\frac{1}{2}}(v)\,dW(v)$$

then var$\left\{\hat{\varepsilon}_{n,2}^{(3)}(x)\right\} \equiv 1, \forall x \in [a,b]$, and with probability 1

$$\left\|\hat{\varepsilon}_{n,2}^{(2)}(x) - \hat{\varepsilon}_{n,2}^{(3)}(x)\right\|_{\infty} = O\left(h^{-1/2}D_n^{-\delta}\right) = o(1).$$

PROOF. Using (2.7.1) in the last step, the term $\sup_{x\in[a,b]}\left|\hat{\varepsilon}_{n,2}^{(2)}(x)-\hat{\varepsilon}_{n,2}^{(3)}(x)\right|$ is bounded by

$$\sup_{x\in[a,b]}\left|1-s_n^2(x)\right|\sup_{x\in[a,b]}\left\{\frac{\sigma_{n,2}^{-1}(x)}{\sqrt{n}}\sum_{j'j=-1}^{N}B_{j',2}(x)\left|s_{j'j}\right|\int B_{j,2}(v)\,\sigma(v)\,f^{\frac{1}{2}}(v)\,dW(v)\right\}$$
$$\leq M_\delta D_n^{-\delta}h^{1/2}C\left|\int\sigma(v)\,f^{\frac{1}{2}}(v)\,dW(v)\right|=O\left(h^{-1/2}D_n^{-\delta}\right)=o(1)\text{ w. p. 1.}$$

Meanwhile, for any $x\in[a,b]$

$$\operatorname{var}\left\{\hat{\varepsilon}_{n,2}^{(3)}(x)\right\}=E\left\{\frac{\sigma_{n,2}^{-1}(x)}{\sqrt{n}}\sum_{j'j=-1}^{N}B_{j',2}(x)\,s_{j'j}\int B_{j,2}(v)\,\sigma(v)\,f^{\frac{1}{2}}(v)\,dW(v)\right\}^2$$
$$=\frac{\sigma_{n,2}^{-2}(x)}{n}\left\{\sum_{j,j',l,l'=-1}^{N}B_{j',2}(x)\,B_{l',2}(x)\,s_{jj'}s_{ll'}\int B_{j,2}(v)\,B_{l,2}(v)\,\sigma^2(v)\,f(v)\,dv\right\}=1$$

directly from (2.2.8) and (2.2.11). □

Now define for any $j'=-1,...,N$ and $x\in[a,b]$, the functions

$$\zeta_{j'}(x)=n^{-1/2}\sigma_{n,2}^{-1}(x)\,B_{j',2}(x),\tilde{\zeta}(x)=\left(\zeta_{j(x)-1}(x),\zeta_{j(x)}(x)\right)^T$$

and the random vector $\Lambda=(\Lambda_{-1},\Lambda_0,\ldots,\Lambda_N)^T$ where

$$\Lambda_{j'}=\sum_{j=-1}^{N}s_{j'j}\int\int B_{j,2}(v)\,\sigma(v)\,f^{\frac{1}{2}}(v)\,dW(v).$$

Then $\Lambda\sim\mathbf{N}(\mathbf{0},S\sum S)$ as $E\Lambda_{j'}=0,\forall j'=-1,..,N$, and the covariance is $E\Lambda_{j'}\Lambda_{l'}=\sum_{j,l=-1}^{N}s_{j'j}\sigma_{jl}s_{ll'}$, for any $j',l'=-1,..,N$, and σ_{jl} is defined in (2.2.8). Notice that

$$\hat{\varepsilon}_{n,2}^{(3)}(x)\equiv\sum_{j'=j(x)-1,j(x)}\zeta_{j'}(x)\,\Lambda_{j'}=\tilde{\zeta}(x)^T\Lambda_{j(x)},\Lambda_j=\left(\Lambda_{j-1},\Lambda_j\right)^T,j=0,...,N$$

and since Lemma 2.7.9 states that the term $\hat{\varepsilon}_{n,2}^{(3)}(x)$ always has variance 1, it means that

$$\hat{\varepsilon}_{n,2}^{(3)}(x)=\frac{\tilde{\zeta}(x)^T\Lambda_{j(x)}}{\sqrt{\tilde{\zeta}(x)^T\left\{\operatorname{cov}\left(\Lambda_{j(x)}\right)\right\}\tilde{\zeta}(x)}}.\tag{2.7.14}$$

Lemma 2.7.10. *For any given* $0 < \alpha < 1$, *one has*

$$\liminf_{n\to\infty} P\left(\sup_{x\in[a,b]} \left|\hat{\varepsilon}_{n,2}(x)\right| \leq \left[2\left\{\log(N+1) - \log\alpha\right\}\right]^{1/2}\right) \geq 1-\alpha. \qquad (2.7.15)$$

PROOF. Define for any $j = 0, ..., N$

$$Q_j = \Lambda_j^T \left\{\operatorname{cov}\left(\Lambda_j\right)\right\}^{-1} \Lambda_j.$$

Result 4.7 (a), page 140 of Johnson and Wichern (1992) ensures that Q_j is distributed as χ_2^2 for any $j = 0, .., N$, hence

$$P\left[Q_j > 2\left\{\log(N+1) - \log\alpha\right\}\right] = \frac{\alpha}{N+1}, \forall 0 \leq j \leq N.$$

Then (2.7.14) and the Maximization Lemma of Johnson and Wichern (1992), page 66 ensure that for any $x \in [a, b]$

$$\left\{\hat{\varepsilon}_{n,2}^{(3)}(x)\right\}^2 = \frac{\left|\tilde{\zeta}(x)^T \Lambda_{j(x)}\right|^2}{\tilde{\zeta}(x)^T \left\{\operatorname{cov}\left(\Lambda_{j(x)}\right)\right\} \tilde{\zeta}(x)} \leq \Lambda_{j(x)}^T \left\{\operatorname{cov}\left(\Lambda_{j(x)}\right)\right\}^{-1} \Lambda_{j(x)} = Q_{j(x)}.$$

One has therefore $\sup_{x\in[a,b]} \left|\hat{\varepsilon}_{n,2}^{(3)}(x)\right|^2 \leq \max_{0\leq j\leq N}\left\{Q_j\right\}$ and

$$\begin{aligned} & P\left[\sup_{x\in[a,b]} \left|\hat{\varepsilon}_{n,2}^{(3)}(x)\right|^2 \leq 2\left\{\log(N+1) - \log\alpha\right\}\right] \\ \geq \quad & P\left[\max_{0\leq j\leq N}\left\{Q_j\right\} > 2\left\{\log(N+1) - \log\alpha\right\}\right] \geq 1-\alpha. \end{aligned}$$

Now (2.7.15) follows from Lemmas 2.7.5, 2.7.6 2.7.7, 2.7.8, 2.7.9. □

Lemma 2.7.11.

$$\left|\sup_{x\in[a,b]} \left|\frac{\hat{\varepsilon}_2(x)}{\sigma_{n,2}(x)}\right| - \sup_{x\in[a,b]} \left|\frac{\tilde{\varepsilon}_2(x)}{\sigma_{n,2}(x)}\right|\right| = O_p\left(\sqrt{\frac{\log n}{nh}}\right) = o_p(1).$$

PROOF. Recall the definition for $\tilde{\mathbf{a}} = (\tilde{a}_{-1}, \tilde{a}_0, ..., \tilde{a}_N)^T$ and $\hat{\mathbf{a}} = (\hat{a}_{-1}, \hat{a}_0, \ldots, \hat{a}_N)^T$ in (2.7.6) and (2.7.7), one has $\left(V + \tilde{B}\right)\tilde{\mathbf{a}} = V\hat{\mathbf{a}}$. Based on Lemma 2.7.2 and (2.3.2), there exists a constant c such that

$$c\left|\hat{\mathbf{a}} - \tilde{\mathbf{a}}\right| \leq \left|V\left(\hat{\mathbf{a}} - \tilde{\mathbf{a}}\right)\right| = \left|\tilde{B}\tilde{\mathbf{a}}\right| \leq A_{n,2}\left(\left|\hat{\mathbf{a}} - \tilde{\mathbf{a}}\right| + \left|\hat{\mathbf{a}}\right|\right) \Rightarrow \left|\hat{\mathbf{a}} - \tilde{\mathbf{a}}\right| \leq \frac{A_{n,2}}{c - A_{n,2}}\left|\hat{\mathbf{a}}\right|. \tag{2.7.16}$$

From the definitions of $\tilde{\varepsilon}_2(x)$ in (2.7.5) and $\hat{\varepsilon}_2(x)$ in (2.7.8), plus (2.7.12), (2.7.16) and (2.7.6), as $n \to \infty$

$$\sup_{x\in[a,b]}\left|\frac{\hat{\varepsilon}_2(x)}{\sigma_{n,2}(x)} - \frac{\tilde{\varepsilon}_2(x)}{\sigma_{n,2}(x)}\right| \leq \sup_{x\in[a,b]}\left|\sum_{j=-1}^{N}\sigma_{n,2}^{-1}(x)\left|\hat{\mathbf{a}} - \tilde{\mathbf{a}}\right| B_{j,2}(x)\right| \leq Cn^{1/2}\frac{A_{n,2}}{c - A_{n,2}}\left|\hat{\mathbf{a}}\right|. \tag{2.7.17}$$

Use (2.7.6) again, it implies that as $n \to \infty$

$$\sup_{x\in[a,b]}\left|\frac{\hat{\varepsilon}_2(x)}{\sigma_{n,2}(x)}\right| \geq \frac{\sqrt{nh}}{C_\sigma}\sup_{x\in[a,b]}\left|\sum_{j=-1}^{N}\hat{a}_j B_{j,2}(x)\right| = \frac{\sqrt{nh}}{C_\sigma}\sup_{x\in[a,b]}\left|\hat{\mathbf{a}}\mathbf{B}_2^T(x)\right| \geq C\sqrt{n}\left|\hat{\mathbf{a}}\right| \tag{2.7.18}$$

where $\mathbf{B}_2(x) = \left\{B_{-1,2}(x), ..., B_{N,2}(x)\right\}^T$, $\mathbf{b}_2(x) = \left\{b_{-1,2}(x), ..., b_{N,2}(x)\right\}^T$.

Then the desired result follows from (2.7.17) and (2.7.18), i.e.

$$\sup_{x\in[a,b]}\left|\frac{\hat{\varepsilon}_2(x)}{\sigma_{n,2}(x)} - \frac{\tilde{\varepsilon}_2(x)}{\sigma_{n,2}(x)}\right| \leq C\frac{A_{n,2}}{c - A_{n,2}}\sup_{x\in[a,b]}\left|\frac{\hat{\varepsilon}_2(x)}{\sigma_{n,2}(x)}\right| = O_p\left(\sqrt{\frac{\log n}{nh}}\right) = o_p(1). \quad \square$$

PROOF OF PROPOSITION 2.3.2. It follows from Lemma 2.7.10 and Lemma 2.7.11 automatically. $\square$

PROOF OF THEOREM 2. Now (2.3.6) implies that $\|\tilde{m}_2(x) - m(x)\|_\infty = O_p\left(h^2\right)$, and hence

$$(nh)^{-1/2}\sqrt{\log(N+1)}\,\|\tilde{m}_2(x) - m(x)\|_\infty = O_p\left\{(nh)^{-1/2}\sqrt{\log(N+1)}h^2\right\} = o_p(1).$$

Applying (2.3.7) in Proposition 2.3.2

$$\begin{aligned}
&\liminf_{n\to\infty} P\left[m(x) \in \hat{m}_2(x) \pm \sigma_{n,2}(x)\left\{2\log(N+1) - 2\log\alpha\right\}^{1/2}, \forall x \in [a,b]\right] \\
= &\liminf_{n\to\infty} P\left[\sup_{x\in[a,b]} \sigma_{n,2}^{-1}(x)\left|\tilde{\varepsilon}_2(x) + \tilde{m}_2(x) - m(x)\right| \le \left\{2\log(N+1) - 2\log\alpha\right\}^{1/2}\right] \\
= &\liminf_{n\to\infty} P\left[\sup_{x\in[a,b]} \left|\frac{\tilde{\varepsilon}_2(x)}{\sigma_{n,2}(x)}\right| \le \left\{2\log(N+1) - 2\log\alpha\right\}^{1/2}\right] \ge 1-\alpha. \square
\end{aligned}$$

CHAPTER 3

Spline-Backfitted Kernel Regression

3.1 Introduction

One popular choice to addressing the issue of the "curse of dimensionality" is the additive model popularized by the book of Hastie and Tibshirani (1990)

$$Y = m(\mathbf{X}) + \sigma(\mathbf{X})\varepsilon, \mathbf{X} = (X_1, ..., X_d), m(\mathbf{x}) = c + \sum_{\alpha=1}^{d} m_\alpha(x_\alpha), \qquad (3.1.1)$$

where the noise satisfies $E(\varepsilon|X) = 0, \text{var}(\varepsilon|X) = 1$ and the component functions satisfy the identification conditions $Em_\alpha(X_\alpha) \equiv 0, \alpha = 1, ..., d$. In addition, we assume that the predictor X_α is distributed on a compact interval $[a_\alpha, b_\alpha], \alpha = 1, ..., d$. The goal is the efficient and fast estimation of the d unknown component functions $\{m_\alpha(x_\alpha)\}_{\alpha=1}^{d}$ based on an i.i.d. sample $\left\{Y_i, \mathbf{X}_i^T\right\}_{i=1}^{n} = \{Y_i, X_{i1}, ..., X_{id}\}_{i=1}^{n}$ following model (3.1.1).

If the last $d-1$ of the component functions were known by "oracle", then one could define a new variable $Y_1 = Y - c - \sum_{\alpha=2}^{d} m_\alpha(X_\alpha) = m_1(X_1) + \sigma(\mathbf{X})\varepsilon$ which one can use to regress on the numerical variable X_1 to estimate the only unknown function $m_1(x_1)$, without the "curse of dimensionality". The basic idea of Linton (1997) was to obtain an approximation to the variable Y_1 by substituting $m_\alpha(X_\alpha), \alpha = 2, ..., d$ with the marginal integration pilot estimates (kernel-based) and establishing that the error caused by this "cheating" is negligible for estimating function $m_1(x_1)$.

In this chapter we propose to pre-estimate the functions $\{m_\alpha(x_\alpha)\}_{\alpha=1}^{d}$ by an under smoothed constant spline procedure. These function estimates are then used as as if they were the true functions for constructing the "oracle" estimator. The greatest advantage of our approach over that of Linton (1997) is that ours is much faster, and can be applied to cases of extremely high dimension data (e.g., the number of predictors, d, can be as large as 50 or 100). We believe that our approach is the first example of marrying the traditionally parallel spline smoothing and kernel smoothing techniques, leading to an estimator with asymptotically normal distribution like a typical kernel estimator, without the formidable computational burden of high dimensional kernel smoothing. Figuratively speaking, spline smoothing can be compared to a sledge-hammer capable of breaking any huge chunk of material (i.e., a regression problem from data of very high dimension and very large sample size), in one slam (i.e., solving only one linear least squares problem), but does not guarantee the fine shapes of the broken pieces (i.e., the estimates are not guaranteed to converge at any point or uniformly over an interval, only in the L^2 sense). In contrast, kernel smoothing works like a sharp knife that cuts anything into pieces of

precise shapes (i.e., confidence intervals are available at any point based on asymptotic normal distribution, and confidence bands are available over compact intervals), but is too tedious to use for a large chunk of material (i.e., the computation cost is intolerable when dimension is high and/or sample size is large). Our proposed new tool can be described as a hammer-knife capable of first slamming any huge clump into many much smaller pieces (i.e., univariate regression problems) in one hit (the spline backfitting step), and then cutting all the smaller pieces into the exact desired shapes (one dimensional kernel smoothing of backfitted pseudo data). In this sense, the method we propose combines the best features of both spline and kernel methods.

Smoothing experts may wonder how one could have all these good features in one method. The success of our method is due to the well-known "reducing bias by undersmoothing" and "averaging out the variance" principles, see Propositions 3.3.1, 3.3.2 and 3.3.3. Both goals are accomplished with the joint asymptotics of kernel and spline functions, which is the new feature of our proofs. For more details, see Lemmas 3.6.3, 3.6.4 and 3.6.5.

In addition to the above features, uniform confidence bands are provided for all function estimates under mild conditions. Literature on nonparametric confidence bands has been scarce, and as far as we know, is lacking in multivariate regression setting. For additive regression model, however, it seems that the present work is the one of the few to offer the measure of uniform accuracy with theoretical justifications. The good news is that the confidence band we provide for $m_\alpha(x_\alpha)$ with any $\alpha = 1, ..., d$, is asymptotically the same confidence band that Härdle (1989) established for univariate regression with kernel smoother, regardless how many regressors there are

and what other functions $m_\alpha(x_\alpha), \alpha = 1, ..., d$ are. Hence neither the dimension d nor other function components play any role in forming the band for $m_\alpha(x_\alpha)$, at least according to the asymptotic theory. In this sense, our estimator of $m_\alpha(x_\alpha)$ possesses what we would like to call "uniform oracle efficiency", which is much stronger than the "pointwise oracle efficiency" of Linton (1997). Furthermore, components in directions not of interests are only required to be Lipschitz continuous (see Remark 3 at the end of Section 3.2). Compared to all existing methods, this feature makes admissible the broadest class of additive model.

The rest of the chpater is organized as follows. In Section 3.2 we introduce the spline-backfitted kernel estimator, and state their asymptotic "oracle efficiency" under appropriate assumptions, both pointwise and uniform. In Section 3.3 we provide some insights into the ideas behind our proofs of the main theoretical results, by decomposing the estimator's "cheating" error into a bias and a noise part, which will be shown separately to be of negligible order. In Section 3.4, we present extensive Monte Carlo results to demonstrate that the proposed estimator does indeed possess the claimed asymptotic properties. The simulated examples cover a wide range of sample sizes with correlated structure and some very high dimensions, which would have been either infeasible to handle with kernel smoothing methods, or lacking any measure of confidence, pointwise or global, by spline method. The proposed estimator are applied to the Boston Housing data in Section 3.4.2. Section 3.5 concludes, and all technical proofs are contained in the 3.6.

3.2 SBK and SBLL Estimators

In this section, we describe the spline-backfitted kernel estimation procedure. Let $\left\{Y_i, \mathbf{X}_i^T\right\}_{i=1}^n = \{Y_i, X_{i1}, ..., X_{id}\}_{i=1}^n$ be an i.i.d. sample following model (3.1.1). In what follows, we write all responses as $\mathbf{Y} = (Y_1, ..., Y_n)^T$, and denote by $\tilde{\mathbf{X}}$ the design matrix $(\mathbf{X}_1, ..., \mathbf{X}_n)^T$. Without loss of generality, we take all intervals $[a_\alpha, b_\alpha] = [0, 1], \alpha = 1, ..., d$. We pre select an integer $N_n \sim n^{2/5} \log(n)$, see Assumption (AS6) below. Next, we define for any $\alpha = 1, ..., d$, the indicator function $I_{J,\alpha}(x_\alpha)$ of the $(N+1)$ equally-spaced subintervals of the finite interval $[0, 1]$, that is

$$I_{J,\alpha}(x_\alpha) = \begin{cases} 1 & JH \leq x_\alpha < (J+1)H, \\ 0 & \text{otherwise,} \end{cases} \quad H = H_n = (N_n + 1)^{-1}, J = 0, 1, ..., N. \tag{3.2.1}$$

Define next the $(1 + dN)$-dimensional space G of additive spline functions as the linear space spanned by $\left\{1, I_{J,\alpha}(x_\alpha), \alpha = 1, ..., d, J = 1, ..., N\right\}$, while denote by G_n the subspace of R^n spanned by $\left\{\{1\}_{i=1}^n, \left\{I_{J,\alpha}(X_{i\alpha})\right\}_{i=1}^n, \alpha = 1, ..., d, J = 1, ..., N\right\}$. As $n \to \infty$, the dimension of G_n becomes $1 + dN$ with probability approaching one.

The spline estimator of additive function $m(\mathbf{x})$ is the unique element $\hat{m}(\mathbf{x}) = \hat{m}_n(\mathbf{x})$ from the space Gso that the vector $\{\hat{m}(\mathbf{X}_1), ..., \hat{m}(\mathbf{X}_n)\}^T$ best approximates the response vector $\mathbf{Y} = (Y_1, ..., Y_n)^T$. To be precise, we define

$$\hat{m}(\mathbf{x}) = \hat{\lambda}_0 + \sum_{\alpha=1}^{d} \sum_{J=1}^{N} \hat{\lambda}_{J,\alpha} I_{J,\alpha}(x_\alpha), \tag{3.2.2}$$

where the coefficients $\hat{\lambda}_0, \hat{\lambda}_{1,1}, ..., \hat{\lambda}_{N,d}$ are the solution of the following least squares

problem

$$\left\{\hat{\lambda}_0, \hat{\lambda}_{1,1}, ..., \hat{\lambda}_{N,d}\right\}^T = \text{argmin}_{R^{dN+1}} \sum_{i=1}^{n} \left\{ Y_i - \lambda_0 - \sum_{\alpha=1}^{d} \sum_{J=1}^{N} \lambda_{J,\alpha} I_{J,\alpha}(X_{i\alpha}) \right\}^2. \tag{3.2.3}$$

The pilot estimators of each component function and the constant are defined as

$$\begin{aligned} \hat{m}_\alpha(x_\alpha) &= \sum_{J=1}^{N} \hat{\lambda}_{J,\alpha} I_{J,\alpha}(x_\alpha) - n^{-1} \sum_{i=1}^{n} \sum_{J=1}^{N} \hat{\lambda}_{J,\alpha} I_{J,\alpha}(X_{i\alpha}), \\ \hat{m}_c &= \hat{\lambda}_0 + n^{-1} \sum_{\alpha=1}^{d} \sum_{i=1}^{n} \sum_{J=1}^{N} \hat{\lambda}_{J,\alpha} I_{J,\alpha}(X_{i\alpha}). \end{aligned} \tag{3.2.4}$$

These pilot estimators are then used to define a set of new pseudo-responses $\hat{Y}_{i1}$ which are estimated versions of the unobservable "oracle" responses Y_{i1}, to be specific,

$$\hat{Y}_{i1} = Y_i - \hat{c} - \sum_{\alpha=2}^{d} \hat{m}_\alpha(X_{i\alpha}), Y_{i1} = Y_i - c - \sum_{\alpha=2}^{d} m_\alpha(X_{i\alpha}), i = 1, ..., n, \hat{c} = n^{-1} \sum_{i=1}^{n} Y_i, \tag{3.2.5}$$

where by Central Limit Theorem $\hat{c}$ is a $\sqrt{n}$-consistent estimator of c. Next, we define the spline-backfitted kernel (SBK) estimator of $m_1(x_1)$ as $\hat{m}_{s,1}(x_1)$ based on $\left\{\hat{Y}_{i1}, X_{i1}\right\}_{i=1}^{n}$, which is an attempt to mimick the would-be Nadaraya-Watson estimator $\tilde{m}_{s,1}(x_1)$ of $m_1(x_1)$ based on $\{Y_{i1}, X_{i1}\}_{i=1}^{n}$, had the unobservable "oracle" responses $\{Y_{i1}\}_{i=1}^{n}$ been available.

$$\hat{m}_{s,1}(x_1) = \frac{\sum_{i=1}^{n} K_h(X_{i1} - x_1) \hat{Y}_{i1}}{\sum_{i=1}^{n} K_h(X_{i1} - x_1)}, \tilde{m}_{s,1}(x_1) = \frac{\sum_{i=1}^{n} K_h(X_{i1} - x_1) Y_{i1}}{\sum_{i=1}^{n} K_h(X_{i1} - x_1)}, \tag{3.2.6}$$

where $\hat{Y}_{i1}$ and Y_{i1} are defined in (3.2.5).

Throughout this paper, on any fixed interval $[a, b]$, we denote the space of second order smooth functions as $C^{(2)}[a, b] = \left\{m \middle| m'' \in C[a, b]\right\}$, and the class of Lipschitz continuous functions for any fixed constant $C > 0$ as $\text{Lip}([a, b], C) = \left\{m \middle| |m(x) - m(x')| \leq C|x - x'|, \forall x, x' \in [a, b]\right\}$.

Before presenting the main theoretical results, we state the following assumptions.

(AS1) *The component function* $m_1 \in C^{(2)}[0,1]$, *while there is a constant* $0 < C_\infty < \infty$ *such that* $m_\beta \in \text{Lip}([0,1], C_\infty)$, $\forall \beta = 1, ..., d$.

(AS2) *The noise* ε_i *given* $\mathbf{X}_i$ *are i. i. d. with mean* 0 *and variance* 1, for $i = 1, ..., n$, *while the conditional standard deviation function* $\sigma(\mathbf{x})$ *is continuous on* $[0,1]^d$. *Denote* $C_\sigma = \max_{\mathbf{x} \in [0,1]^d} \sigma(\mathbf{x})$.

(AS2') *The conditional distribution of noise* $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)$ *given* $\tilde{\mathbf{X}} = (\mathbf{X}_1, ..., \mathbf{X}_n)^T$ *is n-dimensional standard normal.*

(AS3) *The density function* $f(\mathbf{x})$ *of* $\mathbf{X}$ *is continuous and*

$$0 < c_f \leq \inf_{\mathbf{x} \in [0,1]^d} \{f(\mathbf{x})\} \leq \sup_{\mathbf{x} \in [0,1]^d} \{f(\mathbf{x})\} \leq C_f < \infty.$$

The marginal densities $f_\alpha(x_\alpha)$ *of* X_α *have continuous derivatives on* $[0,1]$.

(AS4) *The kernel density function* $K \in \text{Lip}([-1,1], C_K)$ *for some constant* $C_K > 0$, *and is bounded, nonnegative, symmetric, and supported on* $[-1,1]$

(AS5) *The bandwidth* h *of the kernel* K *is assumed to be of order* $n^{-1/5}$, *i.e.,* $c_h n^{-1/5} \leq h \leq C_h n^{-1/5}$ *for some positive constants* c_h, C_h.

(AS6) *The number of interior knots* $N_n \sim n^{2/5} \log(n)$, *i.e.,* $c_N n^{2/5} \log(n) \leq N_n \leq C_N n^{2/5} \log(n)$ *for some positive constants* c_N, C_N, *and the interval width* $H = (N_n + 1)^{-1}$.

The asymptotic property of the kernel smoother $\tilde{m}_{s,1}(x_1)$ is well-developed. Under Assumptions (AS1)-(AS5), according to Theorem 4.2.1 of Härdle (1990), one has

$$\sqrt{nh}\left\{\tilde{m}_{s,1}(x_1) - m_1(x_1) - b(x_1)h^2\right\} \xrightarrow{D} N\left(0, v^2(x_1)\right),$$

where

$$\begin{aligned} b(x_1) &= \mu_2(K)\left\{m_1''(x_1)f_1(x_1)/2 + m_1'(x_1)f_1'(x_1)\right\}f_1^{-1}(x_1), \\ v^2(x_1) &= \|K\|_2^2 E\left\{\sigma^2(x_1, X_2, ..., X_d)\right\}f_1^{-1}(x_1). \end{aligned} \tag{3.2.7}$$

Härdle (1989)provide the uniform asymptotics for kernel smoother. For any $\alpha \in (0,1)$, an asymptotic $100(1-\alpha)\%$ confidence band for $m_1(x_1)$ over interval $[0,1]$ is

$$\lim_{n\to\infty} P\left\{m_1(x_1) \in \tilde{m}_{s,1}(x_1) \pm l_n(x_1), \forall x_1 \in [0,1]\right\} = 1-\alpha$$

where

$$l_n(x_1) = \frac{v(x_1)}{\sqrt{nh}}\left[d_n - \left\{\log\left(h^{-2}\right)\right\}^{-1/2}\log\left\{-\frac{1}{2}\log(1-\alpha)\right\}\right] \tag{3.2.8}$$

$$d_n = \left\{\log\left(h^{-2}\right)\right\}^{1/2}\left[1 + \frac{1}{2\left\{\log\left(h^{-2}\right)\right\}}\log\left\{\frac{\int K'^2(u)\,du}{4\pi^2\int K^2(u)\,du}\right\}\right] \tag{3.2.9}$$

The next two theorems state that the asymptotic magnitude of difference between $\hat{m}_{s,1}(x_1)$ and $\tilde{m}_{s,1}(x_1)$ is of order $o_p\left(n^{-2/5}\right)$, which is dominated by the asymptotic size of $\tilde{m}_{s,1}(x_1) - m_1(x_1)$. Hence $\hat{m}_{s,1}(x_1)$ will have the same asymptotic distribution as $\tilde{m}_{s,1}(x_1)$.

Theorem 3.2.1. *Under Assumptions (AS1) to (AS6), for any $x_1 \in [0,1]$, the SBK estimator $\hat{m}_{s,1}(x_1)$ given in (3.2.6) satisfies*

$$\left|\hat{m}_{s,1}(x_1) - \tilde{m}_{s,1}(x_1)\right| = o_p\left(n^{-2/5}\right) \quad or \quad n^{2/5}\left\{\hat{m}_{s,1}(x_1) - \tilde{m}_{s,1}(x_1)\right\} \xrightarrow{P} 0.$$

Hence with $b(x_1)$ and $v^2(x_1)$ as defined in (3.2.7)

$$\sqrt{nh}\left\{\hat{m}_{s,1}(x_1) - m_1(x_1) - b(x_1)h^2\right\} \xrightarrow{D} N\left(0, v^2(x_1)\right).$$

Theorem 3.2.2. *Under Assumptions (AS1) to (AS6) and (AS2'), the SBK estimator* $\hat{m}_{s,1}(x_1)$ *given in (3.2.6) satisfies*

$$\sup_{x_1 \in [0,1]} \left| \hat{m}_{s,1}(x_1) - \tilde{m}_{s,1}(x_1) \right| = o_p\left(n^{-2/5}\right).$$

Hence for any z

$$\lim_{n\to\infty} P\left[\left\{ \log\left(h^{-2}\right) \right\}^{1/2} \left(\sup_{x_1 \in [0,1]} \frac{\sqrt{nh}}{v(x_1)} \left| \hat{m}_{s,1}(x_1) - m_1(x_1) \right| - d_n \right) < z \right]$$
$$= \exp\left\{ -2\exp(-z) \right\},$$

For any $\alpha \in (0,1)$, *an asymptotic* $100(1-\alpha)\%$ *confidence band for* $m_1(x_1)$ *over interval* $[0,1]$ *is*

$$\hat{m}_{s,1}(x_1) \pm v(x_1) \left(\sqrt{nh}\right)^{-1} \left[d_n - \left\{ \log\left(h^{-2}\right) \right\}^{-1/2} \log\left\{ -\frac{1}{2}\log(1-\alpha) \right\} \right]. \tag{3.2.10}$$

in which d_n *equals to*

$$\left\{ \log\left(h^{-2}\right) \right\}^{1/2} + \frac{1}{2} \left\{ \log\left(h^{-2}\right) \right\}^{-1/2} \left[\log\left\{ \int K'^2(u)\, du \right\} - \log\left\{ 4\pi^2 \int K^2(u)\, du \right\} \right].$$

Remark 1. Similar estimators $\hat{m}_{s,\alpha}(x_\alpha)$ can be constructed for any $\alpha = 2, ..., d$ with same oracle properties. Also, similar constructions can be based on local linear instead of Nadaraya-Watson estimator in (3.2.6). In contrast, the bias coefficient of the spline-backfitted local linear (SBLL) estimator would simply be $b(x_1) = \mu_2(K) m_1''(x_1)/2$, without the additional term of the SBK estimator, while the variance coefficients of SBLL and SBK are the same. Higher order local polynomials can also be used, with obvious modifications. For more on the properties of local linear estimators, in particular, its minimax efficiency, see Fan and Gijbels (1996).

Remark 2. The proofs of Theorems 3.2.1 and 3.2.2 will make it clear that the number of knots can be of the more general form $N_n \sim n^{2/5} N_n'$, where the sequence N_n' satisfies $N_n' \to \infty$, $n^{-\theta} N_n' \to 0$ for any $\theta > 0$. There is no optimal way to choose N_n', however, at least to us at this time. The fact that $N_n^{-1} = o\left(n^{-2/5}\right)$ ensures that the bias in the spline pilot estimators is negligible compared to the bias of h^2 in the kernel/local linear smoothing stage. On the other hand, one does not allow N_n to be too large for practical reasons: the number of terms in (3.2.3), $1 + dN_n$ has to be small relative to n. Hence we select N_n to be of barely larger order than $n^{2/5}$.

Remark 3. Assumption A1 requires only the Lipschitz continuity for the components except for the component of interest. Obviously all m_α are required to be second order smooth if one needs to estimate all components.

3.3 Decomposition

In this section, we introduce some additional notations in order to shed some light on the ideas behind the proofs of Theorems 3.2.1 and 3.2.2. Denote by $\|\phi\|_2$ the theoretical L_2 norm of a function ϕ on $[0,1]^d$, $\|\phi\|_2^2 = E\left\{\phi^2(\mathbf{X})\right\} = \int_{[0,1]^d} \phi^2(\mathbf{x}) f(\mathbf{x})\, d\mathbf{x}$, and the empirical L_2 norm as $\|\phi\|_{2,n}^2 = n^{-1} \sum_{i=1}^n \phi^2(\mathbf{X}_i)$. For any L_2-integrable functions ϕ, φ on $[0,1]^d$, the corresponding inner products are defined by

$$\begin{aligned} \langle \phi, \varphi \rangle_2 &= \int_{[0,1]^d} \phi(\mathbf{x}) \varphi(\mathbf{x}) f(\mathbf{x})\, d\mathbf{x} = E\left\{\phi(\mathbf{X}) \varphi(\mathbf{X})\right\}, \\ \langle \phi, \varphi \rangle_{2,n} &= n^{-1} \sum_{i=1}^n \phi(\mathbf{X}_i) \varphi(\mathbf{X}_i). \end{aligned} \tag{3.3.1}$$

A function ϕ on $[0,1]^d$ is called theoretically (empirically) centered if $\langle 1, \varphi \rangle_2 = 0$ ($\langle 1, \varphi \rangle_{2,n} = 0$). Define the following theoretically centered spline basis

$$b_{J,\alpha}(x_\alpha) = I_{J+1,\alpha}(x_\alpha) - \frac{\|I_{J+1,\alpha}\|_2}{\|I_{J,\alpha}\|_2} I_{J,\alpha}(x_\alpha), \forall \alpha = 1, ..., d, J = 1, ..., N, \quad (3.3.2)$$

where the functions $I_{J,\alpha}(x_\alpha)$'s as defined in (3.2.1) are indicators on the subintervals $[JH, (J+1)H)$. The standardized one is given for any $\alpha = 1, ..., d$,

$$B_{J,\alpha}(x_\alpha) = \frac{b_{J,\alpha}(x_\alpha)}{\|b_{J,\alpha}\|_2}, \forall J = 1, ..., N. \quad (3.3.3)$$

The additive function space G defined earlier can also be spanned by the linearly independent basis $\{1, B_{J,\alpha}(x_\alpha), J = 1, ..., N, \alpha = 1, ..., d\}$, although these new basis involve unknown quantities and therefore can not be computed from the data, they are more convenient for mathematical analysis than the truncated power basis in (3.2.1). Similarly G_n can be spanned linearly by the basis $\left\{1, \{B_{J,\alpha}(X_{i\alpha})\}_{i=1}^n, \alpha = 1, ..., d, J = 1, ..., N\right\}$.

For better understanding, we use the projection idea to elaborate the constant spline estimators. The evaluation of constant spline estimator $\hat{m}(\mathbf{x})$ at the n observations results in an n-dimensional vector, $\hat{m}\left(\tilde{\mathbf{X}}\right) = \{\hat{m}(\mathbf{X}_1), ..., \hat{m}(\mathbf{X}_n)\}^T$, which can be considered as the projection of $\mathbf{Y}$ on the space G_n with respect to the empirical inner product $\langle \cdot, \cdot \rangle_{2,n}$ defined in (3.3.1). In general, for any n-dimensional vector $\mathbf{V} = \{V_1, ..., V_n\}^T$, we define $\mathbf{P}_n\mathbf{V}(\mathbf{x})$ as the spline function constructed from

the projection of $\mathbf{V}$ on the inner product space $\left(G_n, \langle \cdot, \cdot \rangle_{2,n}\right)$

$$\begin{aligned}
\mathbf{P}_n\mathbf{V}(\mathbf{x}) &= \hat{v}_0 + \sum_{\alpha=1}^{d}\sum_{J=1}^{N} \hat{v}_{J,\alpha} B_{J,\alpha}(x_\alpha), \\
\left\{\hat{v}_0, \hat{v}_{1,1}, ..., \hat{v}_{N,d}\right\}^T &= \operatorname{argmin}_{R^{dN+1}} \sum_{i=1}^{n} \left\{ V_i - v_0 - \sum_{\alpha=1}^{d}\sum_{J=1}^{N} v_{J,\alpha} B_{J,\alpha}(X_{i\alpha}) \right\}^2,
\end{aligned}$$

which is similar to (3.2.2) and (3.2.3) except the basis. Next, the multivariate function $\mathbf{P}_n\mathbf{V}(\mathbf{x})$ is decomposed into empirically centered additive components $\mathbf{P}_{n,\alpha}\mathbf{V}(x_\alpha), \alpha = 1, ..., d$ and the constant component $\mathbf{P}_{n,c}\mathbf{V}$

$$\mathbf{P}_{n,\alpha}\mathbf{V}(x_\alpha) = \mathbf{P}^*_{n,\alpha}\mathbf{V}(x_\alpha) - n^{-1}\sum_{i=1}^{n} \mathbf{P}^*_{n,\alpha}\mathbf{V}(X_{i\alpha}) \tag{3.3.4}$$

$$\mathbf{P}^*_{n,\alpha}\mathbf{V}(x_\alpha) = \sum_{J=1}^{N} \hat{v}_{J,\alpha} B_{J,\alpha}(x_\alpha), \tag{3.3.5}$$

$$\mathbf{P}_{n,c}\mathbf{V} = \hat{v}_0 + n^{-1}\sum_{\alpha=1}^{d}\sum_{i=1}^{n} \mathbf{P}^*_{n,\alpha}\mathbf{V}(X_{i\alpha}), \tag{3.3.6}$$

in which the centering procedure is the same as (3.2.4).

With these new notations, we can rewrite the constant spline estimators $\hat{m}(\mathbf{x}), \hat{m}_\alpha(x_\alpha), \hat{m}_c$ defined in (3.2.2) and (3.2.4) as

$$\hat{m}(\mathbf{x}) = \mathbf{P}_n\mathbf{Y}(\mathbf{x}), \hat{m}_\alpha(x_\alpha) = \mathbf{P}_{n,\alpha}\mathbf{Y}(x_\alpha), \hat{m}_c = \mathbf{P}_{n,c}\mathbf{Y}.$$

Based on the relation $\mathbf{Y} = m\left(\tilde{\mathbf{X}}\right) + \sigma\left(\tilde{\mathbf{X}}\right)\varepsilon = m\left(\tilde{\mathbf{X}}\right) + \mathbf{E}$, with noise vector $\mathbf{E} = \{\sigma(\mathbf{X}_i)\varepsilon_i\}_{i=1}^n$, similarly define the noiseless spline smoothers

$$\tilde{m}(\mathbf{x}) = \mathbf{P}_n\left\{m\left(\tilde{\mathbf{X}}\right)\right\}(\mathbf{x}), \tilde{m}_\alpha(x_\alpha) = \mathbf{P}_{n,\alpha}\left\{m\left(\tilde{\mathbf{X}}\right)\right\}(x_\alpha), \tilde{m}_c = \mathbf{P}_{n,c}\left\{m\left(\tilde{\mathbf{X}}\right)\right\},$$

and the noise spline components

$$\tilde{\varepsilon}(\mathbf{x}) = \mathbf{P}_n\mathbf{E}(\mathbf{x}), \tilde{\varepsilon}_\alpha(x_\alpha) = \mathbf{P}_{n,\alpha}\mathbf{E}(x_\alpha), \tilde{\varepsilon}_c = \mathbf{P}_{n,c}\mathbf{E}. \tag{3.3.7}$$

Due to the linearity of operators $\mathbf{P}_n, \mathbf{P}_{n,\alpha}, \mathbf{P}_{n,c}, \alpha = 1, ..., d$, one has the following decomposition, which is crucial to prove Theorems 3.2.1 and 3.2.2

$$\hat{m}(\mathbf{x}) = \tilde{m}(\mathbf{x}) + \tilde{\varepsilon}(\mathbf{x}), \hat{m}_\alpha(x_\alpha) = \tilde{m}_\alpha(x_\alpha) + \tilde{\varepsilon}_\alpha(x_\alpha), \hat{m}_c = \tilde{m}_c + \tilde{\varepsilon}_c, \alpha = 1, ..., d. \tag{3.3.8}$$

As closer examination is needed later for $\tilde{\varepsilon}(\mathbf{x})$ and $\tilde{\varepsilon}_\alpha(x_\alpha)$, one define that

$$\tilde{\mathbf{a}} = \{\tilde{a}_0, \tilde{a}_{1,1}, ..., \tilde{a}_{N,d}\}^T = \operatorname{argmin} \sum_{i=1}^{n} \left\{ \sigma(\mathbf{X}_i)\varepsilon_i - a_0 - \sum_{\alpha=1}^{d}\sum_{J=1}^{N} a_{J,\alpha} B_{J,\alpha}(X_{i\alpha}) \right\}^2, \tag{3.3.9}$$

then $\tilde{\varepsilon}(\mathbf{x})$ in (3.3.7) can be rewritten as $\tilde{\mathbf{a}}^T \mathbf{B}(\mathbf{x})$, where $\tilde{\mathbf{a}} = \left(\mathbf{B}^T\mathbf{B}\right)^{-1}\mathbf{B}^T\mathbf{E}$ is the solution of equation (3.3.9), and matrices $\mathbf{B}(\mathbf{x})$ and $\mathbf{B}$ are defined as

$$\mathbf{B}(\mathbf{x}) = \{1, B_{1,1}(x_1), ..., B_{N,d}(x_d)\}^T, \mathbf{B} = \{\mathbf{B}(\mathbf{X}_1), ..., \mathbf{B}(\mathbf{X}_n)\}^T. \tag{3.3.10}$$

To be specific, the least square solution of the noise is

$$\tilde{\mathbf{a}} = \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \left\langle B_{J,\alpha}, B_{J',\alpha'} \right\rangle_{2,n} \end{pmatrix}^{-1}_{\substack{1\le\alpha,\alpha'\le d,\\ 1\le J,J'\le N}} \begin{pmatrix} n^{-1}\sum_{i=1}^n \sigma(\mathbf{X}_i)\varepsilon_i \\ n^{-1}\sum_{i=1}^n B_{J,\alpha}(X_{i\alpha})\sigma(\mathbf{X}_i)\varepsilon_i \end{pmatrix}_{\substack{1\le J\le N,\\ 1\le\alpha\le d}}. \tag{3.3.11}$$

Our main objective is to study the difference between smoothed backfitted estimator $\hat{m}_{s,1}(x_1)$ and the smoothed "oracle" estimator $\tilde{m}_{s,1}(x_1)$, both given in (3.2.6). From now on, we assume without loss of generality that $d = 2$ for notational brevity. Denote the projection matrix $\mathbf{P}_{0_{N+1},I_N} = \begin{pmatrix} 0_{N+1} & \\ & I_N \end{pmatrix}$, we define another auxiliary entity

$$\tilde{\varepsilon}_2^*(x_2) = \mathbf{P}_{n,2}^*\mathbf{E}(x_2) = \left\{\left(\mathbf{B}^T\mathbf{B}\right)^{-1}\mathbf{B}^T\mathbf{E}\right\}^T \mathbf{P}_{0_{N+1},I_N}(\mathbf{B}(\mathbf{x}))^T = \sum_{J=1}^{N} \tilde{a}_{J,2} B_{J,2}(x_2),$$

which, in particular, entails that

$$\tilde{\varepsilon}_2^*(X_{i2}) = \left\{\left(\mathbf{B}^T\mathbf{B}\right)^{-1}\mathbf{B}^T\mathbf{E}\right\}^T \mathbf{P}_{0_{N+1},I_N}\left(e_i^T\mathbf{B}\right)^T = \sum_{J=1}^{N}\tilde{a}_{J,2}B_{J,2}(X_{i2}), \tag{3.3.12}$$

in which e_i is the n-dimensional unit vector with i-th element 1 and else 0 and hence the i-th row of matrix $\mathbf{B}$, $e_i^T\mathbf{B} = \mathbf{B}(\mathbf{X}_i)$, is the basis functions corresponding to the i-th observation $\mathbf{X}_i$. Definitions (3.3.5) and (3.3.6) imply that $\tilde{\varepsilon}_2(x_2)$ is simply the empirical centering of $\tilde{\varepsilon}_2^*(x_2)$, i.e.

$$\tilde{\varepsilon}_2(x_2) \equiv \tilde{\varepsilon}_2^*(x_2) - n^{-1}\sum_{i=1}^{n}\tilde{\varepsilon}_2^*(X_{i2}) = \sum_{J=1}^{N}\tilde{a}_{J,2}B_{J,2}(X_{i2}) - n^{-1}\sum_{i=1}^{n}\sum_{J=1}^{N}\tilde{a}_{J,2}B_{J,2}(X_{i2}). \tag{3.3.13}$$

Making use of the signal noise decomposition (3.3.8), the difference $\tilde{m}_{s,1}(x_1) - \hat{m}_{s,1}(x_1) + \hat{c} - c$ can be treated as the sum of two terms

$$\frac{n^{-1}\sum_{i=1}^{n}K_h(X_{i1}-x_1)\{\hat{m}_2(X_{i2}) - m_2(X_{i2})\}}{n^{-1}\sum_{i=1}^{n}K_h(X_{i1}-x_1)} = \frac{I(x_1) + II(x_1)}{n^{-1}\sum_{i=1}^{n}K_h(X_{i1}-x_1)}, \tag{3.3.14}$$

where

$$I(x_1) = n^{-1}\sum_{i=1}^{n}K_h(X_{i1}-x_1)\cdot\tilde{\varepsilon}_2(X_{i2}), \tag{3.3.15}$$

$$II(x_1) = n^{-1}\sum_{i=1}^{n}K_h(X_{i1}-x_1)\cdot\{\tilde{m}_2(X_{i2}) - m_2(X_{i2})\}. \tag{3.3.16}$$

The term $I(x_1)$ is related to the noise terms $\tilde{\varepsilon}_2(X_{i2})$, while $II(x_1)$ is induced by the bias terms $\tilde{m}_2(X_{i2}) - m_2(X_{i2})$. Propositions 3.3.1 and 3.3.2 below show respectively that the term $I(x_1)$ is of order $o_p\left(n^{-2/5}\right)$, either at a given point or over an interval. This is the most challenging part to be proved, mostly done in Subsection 3.6.1. On the other hand, Proposition 3.3.3 below shows that the bias term $II(x_1)$ is uniformly

of order $o_p\left(n^{-2/5}\right)$ for $x_1 \in [0,1]$, to be proved in Subsection 3.6.2. Standard theory of kernel density estimation ensures that the denominator term in (3.3.14), $n^{-1}\sum_{i=1}^{n} K_h\left(X_{i1}-x_1\right)$, has a positive lower bound for $x_1 \in [0,1]$. The additional nuisance term $\hat{c}-c$ is of clearly order $O\left(n^{-1/2}\right)$ and thus $o_p\left(n^{-2/5}\right)$, which needs no further arguments for the proofs. Hence both Theorems 3.2.1 and 3.2.2 follow from Propositions 3.3.1, 3.3.2 and 3.3.3. Section 3.6, therefore, is devoted exclusively to the proofs of these three propositions, rather than of the main theoretical results, Theorems 3.2.1 and 3.2.2 themselves.

The next three propositions follow respectively from Lemmas 3.6.10 and 3.6.11, Lemmas 3.6.11 and 3.6.12, Lemmas 3.6.1 and 3.6.2.

Proposition 3.3.1. *Under Assumptions (AS1) to (AS6), for any* $x_1 \in [0,1]$

$$|I(x_1)| = O_p\left(n^{-1/2}\right) = o_p\left(n^{-2/5}\right).$$

Proposition 3.3.2. *Under Assumptions (AS1) to (AS6) and (AS2')*

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

Proposition 3.3.3. *Under Assumptions (AS1), and (AS3) to (AS6)*

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

3.4 Simulation and Examples

3.4.1 Simulation

In this section, we present simulated results to illustrate the finite-sample behavior of the spline backfitted kernel estimators $\hat{m}_{s,\alpha}(x_\alpha)$ for any $\alpha = 1, ...d$.

The data set is generated from the regression model $Y = \sum_{\alpha=1}^{d} m_\alpha(X_\alpha) + \sigma(\mathbf{X}) \cdot \varepsilon$. The additive elements are assumed to be

$$m_\alpha(x_\alpha) = \sin(2\pi x_\alpha), \forall \alpha = 1, ..., d.$$

Similar to Nielsen and Sperlich (2005), the predictors X_α are obtained through the transformation $X_\alpha = 2.5 * \{\Phi(Z_\alpha) - 0.5\}$, where Φ is the standard normal distribution function and the variable $Z_\alpha \sim N(0,1), \alpha = 1, ..., d$ with the correlation coefficients $\rho_{\alpha\beta} = \rho, \alpha \neq \beta$ for any pair of Z's. Now the correlation between X's is not ρ any more, it will depend on ρ. In order to validate the assumption that the density is bounded below from 0, we will focus on the estimation inside $[-1, 1]^d$.

Meanwhile, the error term ε follows standard normal distribution and is independent of $\mathbf{X}$. The conditional standard deviation function is defined by

$$\sigma(\mathbf{x}) = \frac{\sqrt{d}}{2} \cdot \frac{100 - \exp\left\{\sum_{\alpha=1}^{d} |x_\alpha| / d\right\}}{100 + \exp\left\{\sum_{\alpha=1}^{d} |x_\alpha| / d\right\}}.$$

By this choice of $\sigma(\mathbf{x})$, we ensure that our design is heteroscedastic, and the variance is roughly proportional to dimension d. This proportionality is intended to mimic the case when independent copies of the same kind of univariate regression problems are simply added together.

We now describe how the SBLL estimator are implemented. The first step is to obtain the spline estimator of $\sum_{\alpha=1}^{d} m_\alpha(X_\alpha)$, using the truncated power B-spline basis as in (3.2.3). The selection of knots will uniquely define the basis. The knots number N_n will be determined by the sample size and two tuning constants, to be specific

$$N_n = \min\left(\left[c_1 n^{2/5}\log n\right] + c_2, \left[(n/4-1)\,d^{-1}\right]\right),$$

in which $[c]$ denotes the integer part of c. In our simulation study, we have used $c_1 = 1 = c_2$. The choice of these constants c_1 and c_2 makes little difference for a large sample. But for small sample size, it does affect the performance to a degree. The additional constraint that $N \leq (n/4-1)\,d^{-1}$ ensures that the number of terms in the linear least squares problem (3.2.3), $1 + dN_n$, is no greater than $n/4$, which is necessary when the sample size n is moderate and dimension d is high.

The oracle smoother $\tilde{m}_{s,1}(x_1)$ for comparison is obtained by local linear regression of the unobservable $m_1(X_1)+\sigma(\mathbf{X})\,\varepsilon$ on X_1 directly, while the oracle SBLL estimators $\hat{m}_{s,1}(x_1)$ are obtained by local linear regression of $\left\{\hat{Y}_{i1}, X_{i1}\right\}_{i=1}^{n}$. To save space, we only implement the local linear version of $\hat{m}_{s,1}(x_1)$, i.e., the SBLL estimator, using the XploRe quantlet "lpregxest". For information on XploRe, see Härdle, Hlávka and Klinke (2000) or visit http://www.xplore-stat.de.

We have run $S = 500$ replications for sample sizes $n = 100, 200, 500$ and 1000 with correlation coefficient $\rho = 0, 0.3$ respectively. The dimensions are taken at $d = 4, 10$. The major objective of this section is to compare the relative efficiency of $\hat{m}_{s,\alpha}$ with

respect to $\tilde{m}_{s,\alpha}$

$$\text{eff}_{\alpha,l} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left\{\tilde{m}_{s,\alpha}\left(X_{i\alpha,l}\right)-m_{\alpha}\left(X_{i\alpha,l}\right)\right\}^{2} I_{\left\{|X_{i\alpha,l}|\leq 1\right\}}}{\frac{1}{n}\sum_{i=1}^{n}\left\{\hat{m}_{s,\alpha}\left(X_{i\alpha,l}\right)-m_{\alpha}\left(X_{i\alpha,l}\right)\right\}^{2} I_{\left\{|X_{i\alpha,l}|\leq 1\right\}}}, \alpha = 1,...,d, l = 1,...S$$

$$\text{eff}_{\alpha} = \frac{1}{S}\sum_{l=1}^{S}\text{eff}_{\alpha,l}, \alpha = 1,...,d,$$

in which $\left\{X_{i1,l},...,X_{id,l}\right\}_{i=1}^{n}$ is the l-th sample, $l = 1,...,S$. Theorems 3.2.1 and 3.2.2 indicate that the efficiency should be close to 1. In particular, when we have an efficiency value bigger than 1, $\hat{m}_{s,\alpha}\left(x_{\alpha}\right)$ is a better estimator in the sense of mean square distance.

The corresponding mean and the standard error (in the parenthesis) of the relative efficiencies for first and third dimension ($\alpha = 1, 3$) is given in Table 4.3. For the case of $\rho = 0$, almost of all the mean values are around 1 without noticeable influence from the sample size and the correlation. The trend of standard errors confirm the comparability of SBLL $\tilde{m}_{s,\alpha}$ to the oracle estimator $\hat{m}_{s,\alpha}$, with faster convergence for a larger sample. At $\rho = 0$ and all the random selected directions, the SBLL performs better than the oracle local linear estimator in most cases because the independent components can be well-estimated at the first stage, then univariate local linear smoothing at the second stage will treat less noise than the case of direct oracle estimator, the local linear estimator.

In the cases of $\rho = 0.3$, the trend to relative efficiency 1 is very clear regardless of the dimension d. All the means are becoming larger accordingly and approaching to 1 steadily when the sample size becomes bigger. Typically, the relative efficiencies are greater than 0.97 for $d = 4$ with sample size 200, and for $d = 10$ with sample size

500 respectively. We believe that in high dimensional cases the convergence rate is slower than in lower dimensional cases when the predictors are strongly correlated. The standard errors in the parenthesis follow the same trend that less variation is with larger sample size, though it shows slower convergence compared to the case of $\rho = 0$, which is not unexpected.

In addition, several figures display the features of the relative efficiencies in details. In Figures 4.6 and 4.7 four types of line characteristics which correspond to the four sample sizes, the solid line (100), the dotted line (200), the thin line (500) and the thick line (1000). The vertical line at efficiency 1 is the standard line for the comparison of $\hat{m}_{s,1}(x_1)$ and $\tilde{m}_{s,1}(x_1)$. More efficiency values distributed around the vertical line would be confirmative to the conclusions of Theorems 3.2.1 and 3.2.2.

All the curves in Figures 4.6 and 4.7 are the density estimates of relative efficiency distributions for specific sample size n, correlation coefficient ρ and dimension d. With increasing sample sizes, we found that the relative efficiency are becoming closer to the vertical standard line, with narrower spread out. In addition, the curve with $\rho = 0$ shows a faster convergence to the vertical line than those with $\rho = 0.3$ in all cases. An interesting point is that almost of all the peak points of the thick line (with the largest sample size) fall very close to the vertical lines. All above confirms the theorem that SBLL behaves similarly like the oracle local linear estimator.

We have done some more simulation with $d = 50$, and $S = 100$ replications for $\rho = 0, 0.3$, and $n = 500, 1000, 1500, 2000$, the results of which are graphically represented in Figures 4.8 and 4.9. The basic graphic pattern is similar to that for the lower dimensions $d = 4, 10$, though with slower convergence rate and relatively

lower efficiency. The corresponding statistics are listed in Table 4.4.

3.4.2 Boston Housing Example

In this section we apply our method to the Boston Housing Data. The data files bostonh.dat is available in the software of Xplore. The data set 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 among the independent variables. By using a stepwise method, they proposed the alternating conditional expectation method to select a subset of the variables in order to maximize the correlation between the fitted value and the selected covariates. Four variables were selected by penalizing for overfitting. Opsomer and Ruppert (1998) illustrated their automated bandwidth selection for fitting additive models based on the selected four variables. We will use the same four covariates for our model fitting and current analysis. 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 %

One major concern is the big gap in the domain of variables TAX and LSTAT, which will cause severe trouble at the first stage of spline estimation. So logarithmic transformation is done for these two variables before fitting the model. We will fit an additive model as follows:

$$\text{MEDV} = \mu + m_1(\text{RM}) + m_2(\log(\text{TAX})) + m_3(\text{PTRATIO}) + m_4(\log(\text{LSTAT})) + \varepsilon.$$

Although the transformation has shrunk the gap in the domain, some compromise will be necessary to estimate the components since we select the same knots number for each direction. In this case we choose a large number of knots, $N = 5$. In the smoothing step, we use the SBLL estimator to get the final function estimate of each input variable.

In Figure 4.10, the univariate function estimates and corresponding confidence bands are displayed together with the "pseudo data points" with pseudo response as the backfitted response after subtracting the sum function of the remaining three covariates as in (3.2.5). All the function estimates are represented by the dotted lines, "data points" by circles, and confidence bands by upper and lower thin lines. The kernel used in SBLL estimator is Quartic kernel, $K(u) = \frac{15}{16}\left(1-u^2\right)^2$ for $-1 < u < 1$.

Besides the estimation of the component functions, we also use our proposed confidence bands to test the linearity of the components. In Figure 4.10 the straight solid lines are the regression lines with the least square coefficients. The first figure shows that the linearity null hypothesis $H_0 : m_1(\text{RM}) = a_1 + b_1 \cdot \text{RM}$, will be rejected since the confidence bands with 0.99 confidence couldn't totally cover the straight regression line, i.e the p-value is less than 0.01. Similarly the linearity of

the component functions for log (TAX) and log (LSTAT) are not accepted at the significance level 0.01. While the least square straight line of variable PTRATIO in the upper right figure totally falls between the upper and lower 95% confidence bands, thus the linearity null hypothesis $H_0 : m_3(\text{PTRATIO}) = a_3 + b_3 \cdot \text{PTRATIO}$ is accepted at the significance level 0.05.

In addition we add up all the SBLL estimates of component functions and the mean response as a estimate for the response (MEDV). The correlation between the estimate and the raw value of MEDV is as high as 0.80112, implying rather satisfactory fit.

3.5 Conclusions

In this paper we have proposed SBK and SBLL estimators for the component functions in an additive regression model. These estimators behave asymptotically like the standard Nadaraya-Watson and local linear estimators in one dimension, thus breaking the problem of d-dimensional additive regression to d univariate regression problems. This is achieved by approximating the unobservable sample $\{Y_{i1}, X_{i1}\}_{i=1}^{n}$ with the spline estimated sample $\left\{\hat{Y}_{i1}, X_{i1}\right\}_{i=1}^{n}$. Although much mathematics is devoted to proving that this approximation works, the implementation is very easy. To give some idea of how fast the procedure is, to run 100 replications for sample sizes $n = 500, 100, 1500, 2000$ and dimension as high as $d = 50$ takes about 40 minutes on a Dell notebook. In other words, within this time span, a total of $100 \times 4 = 400$ SBLL estimators $\hat{m}_{s,\alpha}(x_\alpha)$ and the same number of oracle smoothers $\tilde{m}_{s,1}(x_1)$

are computed. In addition, the SBK and SBLL estimators inherit the asymptotic confidence bands (3.2.10) of univariate Nadaraya-Watson and local linear estimators. The combination of speed and global accuracy for very high dimension regression is very appealing.

3.6 Proof of Theorems

3.6.1 Variance Reduction

In this subsection we prove Propositions 3.3.1 and 3.3.2. The magnitude of the variance term $I(x_1)$ in (3.3.15) can be measured by its conditional second moment given $\mathbf{X}_1, ..., \mathbf{X}_n$. Based on (3.3.13) and (3.3.15), the conditional second moment $E\left\{I(x_1)|\tilde{\mathbf{X}}\right\}^2$ of $I(x_1)$ given $\tilde{\mathbf{X}} = \{\mathbf{X}_1, ..., \mathbf{X}_n\}$ is

$$E\left[\left\{n^{-1}\sum_{l=1}^{n} K_h(X_{l1}-x_1)\tilde{\varepsilon}_2^*(X_{l2}) - n^{-1}\sum_{l=1}^{n} K_h(X_{l1}-x_1)\cdot n^{-1}\sum_{i=1}^{n}\tilde{\varepsilon}_2^*(X_{i2})\right\}^2\middle|\tilde{\mathbf{X}}\right].$$

It is clear that

$$E\left\{I(x_1)|\tilde{\mathbf{X}}\right\}^2 = E\left\{I_1^2(x_1)\middle|\tilde{\mathbf{X}}\right\} - E\left\{I_2^2(x_1)\middle|\tilde{\mathbf{X}}\right\},$$

where for brevity, we write

$$I_1(x_1) = n^{-1}\sum_{l=1}^{n} K_h(X_{l1}-x_1)\tilde{\varepsilon}_2^*(X_{l2}) \tag{3.6.1}$$

$$I_2(x_1) = n^{-1}\sum_{l=1}^{n} K_h(X_{l1}-x_1)\cdot n^{-1}\sum_{i=1}^{n}\tilde{\varepsilon}_2^*(X_{i2}). \tag{3.6.2}$$

If further one denotes

$$\xi_J(\mathbf{X}_l, x_1) = K_h(X_{l1}-x_1)\,B_{J,2}(X_{l2}), \tag{3.6.3}$$

then

$$I_1(x_1) = n^{-1}\sum_{l=1}^{n} K_h(X_{l1} - x_1)\sum_{J=1}^{N}\tilde{a}_{J,2}B_{J,2}(X_{l2}) = n^{-1}\sum_{l=1}^{n}\sum_{J=1}^{N}\tilde{a}_{J,2}\xi_J(\mathbf{X}_l, x_1). \tag{3.6.4}$$

In order to obtain the order of the conditional second moment of $I_1(x_1)$, we first find the supremum magnitudes of $E\xi_J(\mathbf{X}_l, x_1)$, $\xi_J(\mathbf{X}_l, x_1) - E\xi_J(\mathbf{X}_l, x_1)$ and the size of $\sum_{J=1}^{N}|\tilde{a}_{J,2}|$, in Lemma 3.6.3, 3.6.4 and 3.6.7. Consequently, Lemma 3.6.10 shows that $\sup_{x_1\in[0,1]} E\left\{I_1^2(x_1)\middle|\tilde{\mathbf{X}}\right\} = O_p(n^{-1})$. In Lemma 3.6.11 we have $\sup_{x_1\in[0,1]}|I_2(x_1)| = O_p\left(Nn^{-1}\sqrt{\log n}\right)$. Based on the selection of $N \sim n^{2/5}\log n$, Proposition 3.3.1 is thus proved.

There is one more Assumption (AS2') in addition to Assumptions (AS1) to (AS6) in Lemma 3.6.12. The order of $I_1(x_1)$ under the new restrictions is obtained uniformly over $[0,1]$ inflated only by a factor of $\{\log(n)\}^{1/2}$ compared with the pointwise case, one has $\sup_{x_1\in[0,1]}|I_1(x_1)| = O_p\left(\sqrt{\log(n)/n}\right)$. Now again, due to the selection of the interval width $H \backsim \left(n^{2/5}\log n\right)^{-1}$, the order $O_p\left(Nn^{-1}\sqrt{\log n}\right)$ of $\sup_{x_1\in[0,1]}|I_2(x_1)|$ in Lemma 3.6.11 is negligible compared with order of $\sup_{x_1\in[0,1]}|I_1(x_1)|$. So under the Assumptions (AS1) to (AS6) and (AS2'), we have established the uniform bound over $[0,1]$ of Proposition 3.3.2.

3.6.2 Bias Reduction

Now we prove Proposition 3.3.3 by bounding the bias term $II(x_1)$ in (3.3.16). We first cite one important result from page 149 of de Boor (2001).

Theorem 3.6.1. *Under Assumption (A1)* $m_\alpha \in Lip([0,1], C_\infty)$, *then there exists a*

function $g_\alpha \in G[0,1]$ *such that* $\forall \alpha = 1, ..., d$

$$\|g_\alpha - m_\alpha\|_\infty \leq C_\infty H. \tag{3.6.5}$$

Lemma 3.6.1. *Under Assumptions (AS1), (AS3) and (AS6), for the spline function* g_2 *satisfying (3.6.5), one has*

$$\sup_{x_1 \in [0,1]} \left| \frac{\sum_{i=1}^n K_h(X_{i1} - x_1) \{g_2(X_{i2}) - m_2(X_{i2})\}}{\sum_{i=1}^n K_h(X_{i1} - x_1)} \right| \leq C_\infty H, \tag{3.6.6}$$

and for $\alpha = 1, 2$

$$|E_n g_\alpha(X_\alpha)| = \left| n^{-1} \sum_{i=1}^n g_\alpha(X_{i\alpha}) \right| = O_p\left(n^{-1/2} + H\right). \tag{3.6.7}$$

PROOF. The first inequality (3.6.6) follows trivially from (3.6.5). To prove the second inequality, define a function $g(\mathbf{x}) = c + \sum_{\alpha=1}^2 g_\alpha(x_\alpha)$, then $\|g - m\|_\infty \leq 2C_\infty H$ and hence $\|g - m\|_{2,n} \leq 2C_\infty H$. The definition of projection in Hilbert space then implies that

$$\|\tilde{m} - m\|_{2,n} \leq \|g - m\|_{2,n} \leq 2C_\infty H$$

where $\tilde{m}$ is the projection of m to the space G with respect to $\langle \cdot, \cdot \rangle_{2n}$, the triangular inequality implies that

$$\|\tilde{m} - g\|_{2,n} \leq 4C_\infty H. \tag{3.6.8}$$

Now (3.6.5) leads to $|E_n g_\alpha(X_\alpha) - E_n m_\alpha(X_\alpha)| \leq C_\infty H$, while $E m_\alpha(X_\alpha) = 0$ leads to $E_n m_\alpha(X_\alpha) = O_p\left(n^{-1/2}\right)$. Putting these together, one has

$$|E_n g_\alpha(X_\alpha)| \leq |E_n g_\alpha(X_\alpha) - E_n m_\alpha(X_\alpha)| + |E_n m_\alpha(X_\alpha)| \leq C_\infty H + O_p\left(n^{-1/2}\right), \tag{3.6.9}$$

which establishes (3.6.7). □

In order to show that the bias term $II(x_1)$ defined in (3.3.16) is uniformly $o_p\left(n^{-2/5}\right)$, the following lemma suffices.

Lemma 3.6.2. *Under Assumptions (AS1) to (AS6), as* $n \to \infty$

$$\sup_{x_1 \in [0,1]} \left| \frac{\sum_{i=1}^n K_h(X_{i1} - x_1) \{\tilde{m}_2(X_{i2}) - g_2(X_{i2}) + E_n g_2(X_2)\}}{\sum_{i=1}^n K_h(X_{i1} - x_1)} \right| = O_p\left(n^{-1/2} + H\right). \tag{3.6.10}$$

PROOF. Using the same notations as in the proof of Lemma 3.6.1, (3.6.8) and (3.6.9) now give

$$\|\tilde{m} - g + E_n g_1(X_1) + E_n g_2(X_2)\|_{2,n} \leq 6C_\infty H + O_p\left(n^{-1/2}\right),$$

and Lemma 3.6.8 would then entail that

$$\|\tilde{m} - g + E_n g_1(X_1) + E_n g_2(X_2)\|_2 = O_p\left(n^{-1/2} + H\right). \tag{3.6.11}$$

To complete the proof of the lemma, we write

$$(\tilde{m} - g)(\mathbf{x}) + E_n g_1(X_1) + E_n g_2(X_2) = a + \sum_{\alpha=1}^{2} \sum_{J=1}^{N} a_{J,\alpha} B^*_{J,\alpha}(x_\alpha),$$

where the empirically centered spline basis are

$$B^*_{J,\alpha}(x_\alpha) = B_{J,\alpha}(x_\alpha) - E_n B_{J,\alpha}(X_\alpha) = B_{J,\alpha}(x_\alpha) - n^{-1} \sum_{i=1}^{n} B_{J,\alpha}(X_{i\alpha}),$$

for any $1 \leq J \leq N, 1 \leq \alpha \leq 2$. Then for $\alpha = 1, 2$,

$$\tilde{m}_\alpha(x_\alpha) - g_\alpha(x_\alpha) + E_n g_\alpha(X_\alpha) = \sum_{J=1}^{N} a_{J,\alpha} B^*_{J,\alpha}(x_\alpha),$$

and according to (3.6.19) one has

$$\begin{aligned} &\|\tilde{m} - g + E_n g_1(X_1) + E_n g_2(X_2)\|_2^2 \\ \geq\ & c_0 \left[\left\{ a + \sum_{\alpha=1}^{2} \sum_{J=1}^{N} a_{J,\alpha} E_n B_{J,\alpha}(X_\alpha) \right\}^2 + \sum_{\alpha=1}^{2} \sum_{J=1}^{N} a_{J,\alpha}^2 \right]. \end{aligned} \tag{3.6.12}$$

Now

$$
\begin{aligned}
& n^{-1}\sum_{i=1}^{n} K_h\left(X_{i1}-x_1\right)\left\{\tilde{m}_2\left(X_{i2}\right)-g_2\left(X_{i2}\right)+E_n g_2\left(X_2\right)\right\} \\
= \; & n^{-1}\sum_{i=1}^{n} K_h\left(X_{i1}-x_1\right)\sum_{J=1}^{N} a_{J,2}B^*_{J,2}\left(X_{i2}\right),
\end{aligned}
$$

which is bounded by

$$
\begin{aligned}
& \sum_{J=1}^{N}\left|a_{J,2}\right| \sup_{1\le J\le N}\left|n^{-1}\sum_{i=1}^{n} K_h\left(X_{i1}-x_1\right)B^*_{J,2}\left(X_{i2}\right)\right| \\
& \le \sum_{J=1}^{N}\left|a_{J,2}\right|\left\{\sup_{1\le J\le N}\left|n^{-1}\sum_{i=1}^{n} K_h\left(X_{i1}-x_1\right)B_{J,2}\left(X_{i2}\right)\right|\right. \\
+ \; & \left.\left|n^{-1}\sum_{i=1}^{n} K_h\left(X_{i1}-x_1\right)\right| \sup_{1\le J\le N}\left|E_n B_{J,2}\left(X_2\right)\right|\right\}
\end{aligned}
$$

which can be rewritten as the following according to the definitions of $\xi_J\left(\mathbf{X}_l, x_1\right)$ in (3.6.3) and of $A^*_{n,1}$ in (3.6.28)

$$
\sum_{J=1}^{N}\left|a_{J,2}\right|\left\{\sup_{1\le J\le N}\left|n^{-1}\sum_{l=1}^{n}\xi_J\left(\mathbf{X}_l, x_1\right)\right| + A^*_{n,1}\left|n^{-1}\sum_{i=1}^{n} K_h\left(X_{i1}-x_1\right)\right|\right\}.
$$

Minkowski inequality, Lemma 3.6.5, (3.6.29) and standard properties of kernel density estimator now imply that

$$
\begin{aligned}
& \sup_{x_1\in[0,1]}\left|n^{-1}\sum_{i=1}^{n} K_h\left(X_{i1}-x_1\right)\left\{\tilde{m}_2\left(X_{i2}\right)-g_2\left(X_{i2}\right)+E_n g_2\left(X_2\right)\right\}\right| \\
\le \; & \sqrt{N\sum_{J=1}^{N} a^2_{J,2}}\left\{O_p\left(\sqrt{H}\right)+O_p\left(\sqrt{\frac{\log n}{n}}\right)\right\} \\
= \; & O_p\left(H^{1/2}\sqrt{N\sum_{J=1}^{N} a^2_{J,2}}\right) = O_p\left(\sqrt{\sum_{J=1}^{N} a^2_{J,2}}\right) \\
= \; & O_p\left(\left[\left\{\hat{a}+\sum_{\alpha=1}^{2}\sum_{J=1}^{N}\hat{a}_{J,\alpha}E_n B_{J,\alpha}\left(X_\alpha\right)\right\}^2+\sum_{\alpha=1}^{2}\sum_{J=1}^{N} a^2_{J,\alpha}\right]^{1/2}\right),
\end{aligned}
$$

which according to (3.6.11) and (3.6.12) is

$$= O_p\left(\|\tilde{m} - g + E_n g_1(X_1) + E_n g_2(X_2)\|_2\right) = O_p\left(n^{-1/2} + H\right),$$

thus proving (3.6.10). □

Now combining Lemmas 3.6.1 and 3.6.2, one immediately gets

$$\sup_{x_1 \in [0,1]} \left| n^{-1} \sum_{i=1}^{n} K_h(X_{i1} - x_1) \{\tilde{m}_2(X_{i2}) - m_2(X_{i2})\} \right| = O_p\left(n^{-1/2} + H\right) = o_p\left(n^{-2/5}\right),$$

which establishes Proposition 3.3.3.

3.6.3 Technical Lemmas

In this subsection we have collected all the auxiliary results used in Subsections 3.6.1 and 3.6.2.

Lemma 3.6.3. *Under Assumptions (AS3) to (AS6), one has*

$$\sup_{x_1 \in [0,1]} \sup_{1 \le J \le N} |E\xi_J(\mathbf{X}_l, x_1)| = O\left(H^{1/2}\right).$$

PROOF. Define for $\alpha = 1, 2, J = 1, ..., N+1$

$$c_{J,\alpha} = \|I_{J,\alpha}\|_2^2 = \int I_{J,\alpha}^2(x_\alpha) f_\alpha(x_\alpha) dx_\alpha,$$

then $b_{J,\alpha}(x_\alpha)$ in (3.3.2) can be written as $b_{J,\alpha}(x_\alpha) = I_{J+1,\alpha}(x_\alpha) - c_{J+1,\alpha} I_{J,\alpha}(x_\alpha) / c_{J,\alpha}$ and

$$\|b_{J,\alpha}\|_2^2 = c_{J+1,\alpha}\left(1 + c_{J+1,\alpha}/c_{J,\alpha}\right), \forall \alpha = 1, 2, J = 1, ..., N.$$

In Assumption (AS3) the two positive constants c_f, C_f are the upper and lower bounds of all the marginal densities $f_\alpha(x_\alpha)$, then for all $J = 1, ..., N+1, \alpha = 1, 2$

$$c_f H \le c_{J,\alpha} \le C_f H. \tag{3.6.13}$$

Then for all $\alpha = 1, 2, J = 1, ..., N, \left\|b_{J,\alpha}\right\|_2^2 \backsim H$, or specifically

$$c_f \left(1 + c_f/C_f\right) H \leq \left\|b_{J,\alpha}\right\|_2^2 \leq C_f \left(1 + C_f/c_f\right) H. \tag{3.6.14}$$

The absolute expected value of $\xi_J \left(\mathbf{X}_l, x_1\right)$ is

$$\begin{aligned}
& \left|E\xi_J \left(\mathbf{X}_l, x_1\right)\right| = \left|E\left\{K_h \left(X_{l1} - x_1\right) B_{J,2} \left(X_{l2}\right)\right\}\right| \\
\leq & \int\int K_h \left(u_1 - x_1\right) \left|B_{J,2} \left(u_2\right)\right| f \left(u_1, u_2\right) du_1 du_2 \\
= & \int\int K \left(v_1\right) \frac{\left|b_{J,2} \left(u_2\right)\right|}{\left\|b_{J,2}\right\|_2} f \left(hv_1 + x_1, u_2\right) dv_1 du_2 \\
= & \left(\left\|b_{J,2}\right\|_2\right)^{-1} \int\int K \left(v_1\right) \left\{I_{J+1,2} \left(u_2\right) + \left(\frac{c_{J+1,2}}{c_{J,2}}\right)^{1/2} I_{J,2} \left(u_2\right)\right\} \\
& \times f \left(hv_1 + x_1, u_2\right) dv_1 du_2 \\
= & \left(\left\|b_{J,2}\right\|_2\right)^{-1} \left\{\int\int K \left(v_1\right) I_{J+1,2} \left(u_2\right) f \left(hv_1 + x_1, u_2\right) dv_1 du_2 \right. \\
& \left. + \left(\frac{c_{J+1,2}}{c_{J,2}}\right)^{1/2} \int\int K \left(v_1\right) I_{J,2} \left(u_2\right) f \left(hv_1 + x_1, u_2\right) dv_1 du_2\right\}.
\end{aligned}$$

The boundedness of the joint density f and the Lipschitz continuity of the kernel K will then imply that

$$\sup_{x_1 \in [0,1]} \sup_{1 \leq J \leq N} \int\int K \left(v_1\right) I_{J,2} \left(u_2\right) f \left(hv_1 + x_1, u_2\right) dv_1 du_2 \leq C_K C_f H,$$

the proof of the lemma is then completed. □

Lemma 3.6.4. *Denote by D_n a set of endpoints in $[0, 1]$, with cardinality $M_n = |D_n|$ of order n^6, i.e. there exist constants $0 < c_D < C_D$ such that $c_D n^6 \leq M_n \leq C_D n^6$, then under Assumptions (AS3) to (AS6)*

$$\sup_{x_1 \in D_n} \sup_{1 \leq J \leq N} \left|n^{-1} \sum_{l=1}^{n} \left\{\xi_J \left(\mathbf{X}_l, x_1\right) - E\xi_J \left(\mathbf{X}_l, x_1\right)\right\}\right| = O_p \left(\sqrt{\frac{\log n}{nh}}\right). \tag{3.6.15}$$

PROOF. For simplicity, denote $\xi_J^*(\mathbf{X}_l, x_1) = \xi_J(\mathbf{X}_l, x_1) - E\xi_J(\mathbf{X}_l, x_1)$. First we will compute the moments of the theoretical centered random variable $\xi_J^*(\mathbf{X}_l, x_1)$ for later use in Bernstein's inequality

$$E\left\{\xi_J^*(\mathbf{X}_l, x_1)\right\}^2 = E\xi_J^2(\mathbf{X}_l, x_1) - \left\{E\xi_J(\mathbf{X}_l, x_1)\right\}^2,$$

in which the first term

$$\begin{aligned} & E\xi_J^2(\mathbf{X}_l, x_1) = E\left\{K_h(X_{l1} - x_1) B_{J,2}(X_{l2})\right\}^2 \\ = & \int\int \frac{K^2}{h\left\|b_{J,2}\right\|_2^2}(v_1)\left\{I_{J+1,2}(u_2) + \frac{c_{J+1,2}}{c_{J,2}} I_{J,2}(u_2)\right\} f(hv_1 + x_1, u_2)\, dv_1 du_2, \end{aligned}$$

so there exist constants $c', C' > 0$, such that $c'h^{-1} \leq E\xi_J^2(\mathbf{X}_l, x_1) \leq C'h^{-1}$. Then $E\xi_J^2(\mathbf{X}_l, x_1) \gg \left\{E\xi_J(\mathbf{X}_l, x_1)\right\}^2$ where $a_n \gg b_n$ means $\lim_{n\to\infty} b_n/a_n = 0$. Hence

$$E\left\{\xi_J^*(\mathbf{X}_l, x_1)\right\}^2 = E\xi_J^2(\mathbf{X}_l, x_1) - \left\{E\xi_J(\mathbf{X}_l, x_1)\right\}^2 \geq c^* h^{-1},$$

for positive constant $c^* < c'$.

When $k \geq 3$, the k-th moment $E\left|\xi_J(\mathbf{X}_l, x_1)\right|^k$ is

$$\left\{\left\|b_{J,2}\right\|_2\right\}^{-k} \int\int K_h^k(u_1 - x_1)\left\{I_{J+1,2}(u_2) + \left(\frac{c_{J+1,2}}{c_{J,2}}\right)^k I_{J,2}(u_2)\right\} f(u_1, u_2)\, du_1 du_2,$$

and it can be bounded as follows

$$c_k' h^{(1-k)} H^{(1-k/2)}\left\{1 + \left(\frac{c_f}{C_f}\right)^k\right\} \leq E\left|\xi_J(\mathbf{X}_l, x_1)\right|^k \leq C_k' h^{(1-k)} H^{(1-k/2)}\left\{1 + \left(\frac{C_f}{c_f}\right)^k\right\}.$$

Lemma 3.6.3 implies $\left|E\xi_J(\mathbf{X}_l, x_1)\right|^k \leq CH^{k/2}$, then $E\left|\xi_J(\mathbf{X}_l, x_1)\right|^k \gg \left|E\xi_J(\mathbf{X}_l, x_1)\right|^k$. $E\left|\xi_J^*(\mathbf{X}_l, x_1)\right|^k$ can be expressed as

$$\begin{aligned} & E\left|\xi_J(\mathbf{X}_l, x_1) - E\xi_J(\mathbf{X}_l, x_1)\right|^k \leq 2^{k-1}\left(E\left|\xi_J(\mathbf{X}_l, x_1)\right|^k + \left|E\xi_J(\mathbf{X}_l, x_1)\right|^k\right) \\ \leq & C_1 2^{k-1} h^{(1-k)} H^{(1-k/2)}\left(\frac{C_f}{c_f}\right)^k k! = C_1\left\{2h^{-1}H^{-1/2}\left(\frac{C_f}{c_f}\right)\right\}^{(k-2)} k!\left(h^{-1}\right) \\ \leq & \left\{C_2 2h^{-1}H^{-1/2}\right\}^{(k-2)} k! E\left|\xi_J^*(\mathbf{X}_l, x_1)\right|^2, \end{aligned}$$

then there exists such a constant $c = C_2 2h^{-1}H^{-1/2}$ such that

$$E\left|\xi_J^*\left(\mathbf{X}_l, x_1\right)\right|^k \leq c^{k-2}k!E\left|\xi_J^*\left(\mathbf{X}_l, x_1\right)\right|^2,$$

that means the sequence of random variables $\left\{\xi_J^*\left(\mathbf{X}_l, x_1\right)\right\}_{l=1}^n$ satisfies the Cramér's condition, hence by the Bernstein's inequality we have

$$P\left\{\left|n^{-1}\sum_{l=1}^n \xi_J^*\left(\mathbf{X}_l, x_1\right)\right| \geq \delta\sqrt{\frac{\log n}{(nh)}}\right\} \leq 2\exp\left\{\frac{-\delta^2 \log n}{c^* + 2C_2\delta H^{-1/2}\sqrt{\log n/(nh)}}\right\},$$

there exists large enough value $\delta > 0$ such that $-\delta^2/\left\{c^* + 2C_2\delta H^{-1/2}\sqrt{\log n/(nh)}\right\} \leq -10$, then

$$\begin{aligned} &\sum_{n=1}^{\infty} P\left\{\sup_{x_1\in D_n}\sup_{1\leq J\leq N}\left|n^{-1}\sum_{l=1}^n \xi_J^*\left(\mathbf{X}_l, x_1\right)\right| \geq \delta\sqrt{\frac{\log n}{(nh)}}\right\} \\ \leq\ & 2\sum_{n=1}^{\infty} NM_n n^{-10} \leq 2C_D\sum_{n=1}^{\infty} n^{-3} < \infty. \end{aligned}$$

Borel-Cantelli Lemma implies (3.6.15). □

Lemma 3.6.5. *Under Assumptions (AS3) to (AS6)*

$$\sup_{x_1\in[0,1]}\sup_{1\leq J\leq N}\left|n^{-1}\sum_{l=1}^n \xi_J\left(\mathbf{X}_l, x_1\right)\right| = O_p\left(H^{1/2}\right).$$

PROOF. Denote for $x \in [0,1]$, $\Lambda(x) = \sup_{1\leq J\leq N}\left|n^{-1}\sum_{l=1}^n \xi_J\left(\mathbf{X}_l, x\right)\right|$. If we choose the subset D_n as in Lemma 3.6.4 to consist of equally spaced endpoints in $[0,1]$, specifically

$$D_n = \left\{x_{1,k}, 0 \leq k \leq M_n; 0 = x_{1,0} < x_{1,1} < \ldots < x_{1,M_n} = 1\right\},$$

then the consecutive endpoints make a total of M_n subintervals with length M_n^{-1}. Employing the discretization method, we have

$$\sup_{x_1\in[0,1]}\left|\Lambda\left(x_1\right)\right| = \sup_{0\leq k\leq M_n}\left|\Lambda\left(x_{1,k}\right)\right| + \sup_{1\leq k\leq M_n}\sup_{x_1\in\left[x_{1,k-1}, x_{1,k}\right]}\left|\Lambda\left(x_1\right) - \Lambda\left(x_{1,k}\right)\right|. \tag{3.6.16}$$

We only need to bound the second term, as Lemmas 3.6.3 and 3.6.4, and the fact $H^{1/2} \gg \sqrt{\log n/(nh)}$ yield

$$\sup_{0\le k\le M_n} |\Lambda(x_{1,k})| = \sup_{x_1\in D_n} \sup_{1\le J\le N} \left| n^{-1}\sum_{l=1}^{n} \xi_J(\mathbf{X}_l, x_1)\right| = O_p\left(H^{1/2}\right). \quad (3.6.17)$$

Employing Lipschitz continuity of kernel K, one has

$$\begin{aligned} &\sup_{1\le k\le M_n} \sup_{x_1\in\left[x_{1,k-1},x_{1,k}\right]} \left|K_h(X_{l1}-x_1) - K_h(X_{l1}-x_{1,k})\right| \\ \le &\sup_{1\le k\le M_n} \sup_{x_1\in\left[x_{1,k-1},x_{1,k}\right]} C_K\left|\frac{X_{l1}-x_1}{h^2} - \frac{X_{l1}-x_{1,k}}{h^2}\right| \le C_K M_n^{-1} h^{-2}. \end{aligned} \quad (3.6.18)$$

Hence we have

$$\begin{aligned} &\sup_{1\le k\le M_n} \sup_{x_1\in\left[x_{1,k-1},x_{1,k}\right]} \left|\Lambda(x_1) - \Lambda(x_{1,k})\right| \\ \le &\sup_{1\le k\le M_n} \sup_{x_1\in\left[x_{1,k-1},x_{1,k}\right]} \sup_{1\le J\le N} \left| n^{-1}\sum_{l=1}^{n} \xi_J(\mathbf{X}_l, x_1) - n^{-1}\sum_{l=1}^{n} \xi_J(\mathbf{X}_l, x_{1,k})\right| \\ \le &\sup_{1\le k\le M_n} \sup_{x_1\in\left[x_{1,k-1},x_{1,k}\right]} \left|K_h(X_{l1}-x_1) - K_h(X_{l1}-x_{1,k})\right| \\ &\times \sup_{1\le J\le N} n^{-1}\sum_{l=1}^{n} \left|B_{J,2}(X_{l2})\right| \\ \le &\ C_K \frac{1}{M_n h^2} \cdot \sup_{x_2\in[0,1]} \sup_{1\le J\le N} \left|B_{J,2}(x_2)\right| = O\left(M_n^{-1}h^{-2}H^{-1/2}\right) = o\left(n^{-1}\right), \end{aligned}$$

since $c_D n^6 \le M_n \le C_D n^6$ in Lemma 3.6.4. The lemma follows instantly from (3.6.16), (3.6.17) and the above result. □

Lemma 3.6.6. *Under Assumptions (AS3) and (AS6), there exist constants* $C_0 > c_0 > 0$ *such that*

$$c_0\left(a_0^2 + \sum_{J,\alpha} a_{J,\alpha}^2\right) \le \left\|a_0 + \sum_{J,\alpha} a_{J,\alpha} B_{J,\alpha}\right\|_2^2 \le C_0\left(a_0^2 + \sum_{J,\alpha} a_{J,\alpha}^2\right), \quad (3.6.19)$$

for any $\mathbf{a} = (a_0, a_{1,1}, ..., a_{N,1}, a_{1,2}, ..., a_{N,2})^T \in R^{2N+1}$.

PROOF. According to Lemma 1 in Stone (1985), there exists a constant $c_0 > 0$ such that

$$\left\| a_0 + \sum_{J,\alpha} a_{J,\alpha} B_{J,\alpha} \right\|_2^2 \geq c_0 \left(a_0^2 + \left\| \sum_{J=1}^{N} a_{J,1} B_{J,1} \right\|_2^2 + \left\| \sum_{J=1}^{N} a_{J,2} B_{J,2} \right\|_2^2 \right),$$

If it can be proved that there exist constants $C_0' > c_0' > 0$ such that for $\alpha = 1, 2$

$$c_0' \sum_{J=1}^{N} a_{J,\alpha}^2 \leq \left\| \sum_{J=1}^{N} a_{J,\alpha} B_{J,\alpha} \right\|_2^2 \leq C_0' \sum_{J=1}^{N} a_{J,\alpha}^2, \tag{3.6.20}$$

then (3.6.19) follows. To prove (3.6.20), the original B-Spline basis is employed. Without loss of generality we only provide the proof for $\alpha = 1$. We pick the constant basis $\left\{ I_{J,1}(x_1) \right\}_{J=1}^{N+1}$ and represent the term $\sum_{J=1}^{N} a_{J,1} B_{J,1}(x_1)$ as follows

$$\sum_{J=1}^{N} a_{J,1} B_{J,1}(x_1) = \sum_{J=1}^{N+1} d_{J,1} I_{J,1}(x_1). \tag{3.6.21}$$

Theorem 5.4.2 in Devore & Lorentz (1993) says that there is an equivalent relationship between the L_p $(p > 0)$ norm of a B-spline function and the sequence of B-spline coefficients. To be specific, in our case

$$\left\| \sum_{J=1}^{N+1} d_{J,1} I_{J,1} \right\|_{L_2}^2 = \int \left\{ \sum_{J=1}^{N+1} d_{J,1} I_{J,1}(x_1) \right\}^2 dx_1 = \sum_{J=1}^{N+1} d_{J,1}^2 H.$$

As in Assumption (AS3) the joint density bounded between c_f and C_f, we have

$$c_f \left\| \sum_{J=1}^{N+1} d_{J,1} I_{J,1} \right\|_{L_2}^2 \leq \left\| \sum_{J=1}^{N+1} d_{J,1} I_{J,1} \right\|_2^2 \leq C_f \left\| \sum_{J=1}^{N+1} d_{J,1} I_{J,1} \right\|_{L_2}^2.$$

The equality (3.6.21) and (3.6.14) leads to

$$\begin{aligned} \sum_{J=1}^{N+1} d_{J,1}^2 &= \sum_{J=1}^{N} \frac{a_{J,1}^2}{\|b_{J,1}\|_2^2} \left\{ \left(\frac{c_{J+1,1}}{c_{J,1}} \right)^2 + 1 \right\} \\ \Rightarrow \quad c_d \sum_{J=1}^{N} a_{J,1}^2 H^{-1} &\leq \sum_{J=1}^{N+1} d_{J,1}^2 \leq C_d \sum_{J=1}^{N} a_{J,1}^2 H^{-1}, \end{aligned}$$

for positive constants c_d and C_d. Therefore,

$$c_f c_d \sum_{J=1}^{N} a_{J,1}^2 \leq \left\| \sum_{J=1}^{N} a_{J,1} B_{J,1} \right\|_2^2 = \left\| \sum_{J=1}^{N+1} d_{J,1} I_{J,1} \right\|_2^2 \leq C_f C_d \sum_{J=1}^{N} a_{J,1}^2,$$

i.e. (3.6.20) holds given $c_0' = c_f c_d$, $C_0' = C_f C_d$. □

Lemma 3.6.7. *Under Assumptions (AS1) to (AS6), the least square solution* $\tilde{\mathbf{a}}$ *defined in (3.3.9) satisfies*

$$\tilde{\mathbf{a}}^T \tilde{\mathbf{a}} = \tilde{a}_0^2 + \sum_{J=1}^{N} \sum_{\alpha=1}^{2} \tilde{a}_{J,\alpha}^2 = O_p\left(\frac{N}{n}\right). \tag{3.6.22}$$

PROOF. According to (3.3.9), $\tilde{\mathbf{a}} = \left(\mathbf{B}^T\mathbf{B}\right)^{-1}\mathbf{B}^T\mathbf{E}$, then

$$\tilde{\mathbf{a}}^T\mathbf{B}^T\mathbf{B}\tilde{\mathbf{a}} = \left(\tilde{\mathbf{a}}^T\mathbf{B}^T\mathbf{B}\right)\left(\mathbf{B}^T\mathbf{B}\right)^{-1}\mathbf{B}^T\mathbf{E} = \tilde{\mathbf{a}}^T\left(\mathbf{B}^T\mathbf{E}\right).$$

Replacing $\mathbf{B}^T\mathbf{B}$ with matrix of the inner products $\left\langle B_{J,\alpha}, B_{J',\alpha'} \right\rangle_{2,n}$, as the matrix $\mathbf{B}$ is given in (3.3.10), one has

$$\|\mathbf{B}\tilde{\mathbf{a}}\|_{2,n}^2 = \tilde{\mathbf{a}}^T \begin{pmatrix} 1 & \\ & \left\langle B_{J,\alpha}, B_{J',\alpha'} \right\rangle_{2,n} \end{pmatrix} \tilde{\mathbf{a}} = \tilde{\mathbf{a}}^T\left(n^{-1}\mathbf{B}^T\mathbf{E}\right). \tag{3.6.23}$$

Based on (3.6.19), the left hand side of (3.6.23) is bounded below by

$$(1 - A_n)\|\mathbf{B}\tilde{\mathbf{a}}\|_2^2 = (1 - A_n)\left\| \tilde{a}_0 + \sum_{J,\alpha} \tilde{a}_{J,\alpha} B_{J,\alpha} \right\|_2^2 \geq c_0 (1 - A_n)\left(\tilde{a}_0^2 + \sum_{J,\alpha} \tilde{a}_{J,\alpha}^2\right), \tag{3.6.24}$$

where A_n is of order $o_p(1)$ in Lemma 3.6.8. While the last step in (3.6.24) is obtained from (3.6.19). Meanwhile by the Cauchy-Schwartz inequality and the expression of $\tilde{\mathbf{a}}$ in (3.3.11), the right hand side of (3.6.23) is bounded from above by

$$\left(\tilde{a}_0^2 + \sum_{J,\alpha} \tilde{a}_{J,\alpha}^2\right)^{1/2} \left[\left\{n^{-1}\sum_{i=1}^{n} \sigma(\mathbf{X}_i)\varepsilon_i\right\}^2 + \sum_{J,\alpha}\left\{n^{-1}\sum_{i=1}^{n} B_{J,\alpha}(X_{i\alpha})\sigma(\mathbf{X}_i)\varepsilon_i\right\}^2\right]^{1/2}. \tag{3.6.25}$$

Now (3.6.23), (3.6.24) and (3.6.25) will lead implies that $\tilde{a}_0^2 + \sum_{J,\alpha} \tilde{a}_{J,\alpha}^2$ is less than

$$c_0^{-2}(1-A_n)^{-2}\left[\left\{n^{-1}\sum_{i=1}^{n}\sigma(\mathbf{X}_i)\varepsilon_i\right\}^2 + \sum_{J,\alpha}\left\{n^{-1}\sum_{i=1}^{n}B_{J,\alpha}(X_{i\alpha})\sigma(\mathbf{X}_i)\varepsilon_i\right\}^2\right].$$

Note next that it is trivial to verify that

$$E\left[\left\{n^{-1}\sum_{i=1}^{n}\sigma(\mathbf{X}_i)\varepsilon_i\right\}^2 + \sum_{J,\alpha}\left\{n^{-1}\sum_{i=1}^{n}B_{J,\alpha}(X_{i\alpha})\sigma(\mathbf{X}_i)\varepsilon_i\right\}^2\right] = O\left(n^{-1}N\right).$$

Therefore (3.6.22) holds. □

Lemma 3.6.8. *Under Assumptions (AS3) and (AS4), the uniform supremum of the rescaled difference between* $\langle g_1, g_2\rangle_{2,n}$ *and* $\langle g_1, g_2\rangle_2$ *is*

$$A_n = \sup_{g_1,g_2\in G^{(-1)}} \frac{\left|\langle g_1,g_2\rangle_{2,n} - \langle g_1,g_2\rangle_2\right|}{\|g_1\|_2\,\|g_2\|_2} = O_p\left(\sqrt{\frac{\log n}{nH}}\right) = o_p(1). \qquad (3.6.26)$$

PROOF. Let

$$\begin{aligned} g_1(X_1,X_2) &= a_0 + \sum_{J=1}^{N}\sum_{\alpha=1}^{2} a_{J,\alpha}B_{J,\alpha}(X_\alpha), \\ g_2(X_1,X_2) &= a_0' + \sum_{J'=1}^{N}\sum_{\alpha'=1}^{2} a'_{J',\alpha'}B_{J',\alpha'}(X_{\alpha'}), \end{aligned}$$

in which for any $J, J' = 1,...,N, \alpha, \alpha' = 1,2$, $a_{J,\alpha}$ and $a'_{J',\alpha'}$ are real constants.

The difference between the empirical and theoretical inner products of g_1 and g_2 is

$$\begin{aligned} \left|\langle g_1,g_2\rangle_{2,n} - \langle g_1,g_2\rangle_2\right| &= \left|\left\langle a_0 + \sum_{J=1}^{N}\sum_{\alpha=1}^{2} a_{J,\alpha}B_{J,\alpha}, a_0' + \sum_{J'=1}^{N}\sum_{\alpha'=1}^{2} a'_{J',\alpha'}B_{J',\alpha'}\right\rangle_{2,n}\right. \\ &\quad \left. - \left\langle a_0 + \sum_{J=1}^{N}\sum_{\alpha=1}^{2} a_{J,\alpha}B_{J,\alpha}, a_0' + \sum_{J'=1}^{N}\sum_{\alpha'=1}^{2} a'_{J',\alpha'}B_{J',\alpha'}\right\rangle_{2}\right| \end{aligned}$$

$$\leq \left|\sum_{J,\alpha}\langle a_0', a_{J,\alpha}B_{J,\alpha}\rangle_{2,n}\right| + \left|\sum_{J',\alpha'}\left\langle a_0, a'_{J',\alpha'}B_{J',\alpha'}\right\rangle_{2,n}\right|$$
$$+ \sum_{J,J',\alpha,\alpha'} |a_{J,\alpha}|\left|a'_{J',\alpha'}\right|\left|\left\langle B_{J,\alpha}, B_{J',\alpha'}\right\rangle_{2,n} - \left\langle B_{J,\alpha}, B_{J',\alpha'}\right\rangle_2\right|. \quad (3.6.27)$$

The equivalence of norms given in equation (3.6.19) leads to

$$\left|\sum_{J,\alpha}\langle a_0', a_{J,\alpha}B_{J,\alpha}\rangle_{2,n}\right| \leq A^*_{n,1}\cdot|a_0'|\cdot\left\|\sum_{J,\alpha}a_{J,\alpha}B_{J,\alpha}\right\|_2$$
$$\leq C_0 A^*_{n,1}\,|a_0'|^{1/2}\left(\sum_{J,\alpha}\tilde{a}^2_{J,\alpha}\right)^{1/2} \leq C_{A,1}A^*_{n,1}\,\|g_1\|_2\,\|g_2\|_2\,,$$

where

$$A^*_{n,1} = \sup_{J,\alpha}\left|\langle 1, B_{J,\alpha}\rangle_{2,n} - \langle 1, B_{J,\alpha}\rangle_2\right| = \sup_{J,\alpha}\left|\langle 1, B_{J,\alpha}\rangle_{2,n}\right|. \quad (3.6.28)$$

Similarly it holds for the second term in (3.6.27) that

$$\left|\sum_{J,\alpha}\left\langle a_0, a'_{J',\alpha'}B_{J',\alpha'}\right\rangle_{2,n}\right| \leq C'_{A,1}A^*_{n,1}\,\|g_1\|_2\,\|g_2\|_2\,.$$

It is easy to show by Bernstein's inequality that

$$A^*_{n,1} = \sup_{J,\alpha}\left|n^{-1}\sum_{i=1}^{n}B_{J,\alpha}(X_{i\alpha})\right| = O_p\left(\sqrt{\log n/n}\right). \quad (3.6.29)$$

The third term in (3.6.27) will be in probability less than

$$\sum_{J,J',\alpha,\alpha'}|a_{J,\alpha}|\left|a'_{J',\alpha'}\right|\left|\left\langle B_{J,\alpha}, B_{J',\alpha'}\right\rangle_{2,n} - \left\langle B_{J,\alpha}, B_{J',\alpha'}\right\rangle_2\right|$$
$$\leq \sum_{J,J',\alpha,\alpha'}|a_{J,\alpha}|\left|a'_{J',\alpha'}\right|A^*_{n,2} \leq C_{A,2}A^*_{n,2}\left\{\sum_{J,\alpha}a^2_{J,\alpha}\right\}^{1/2}\left\{\sum_{J',\alpha'}a'^2_{J',\alpha'}\right\}^{1/2}$$
$$\leq C_{A,2}A^*_{n,2}\,\|g_1\|_2\,\|g_2\|_2\,,$$

where

$$A^*_{n,2} = \sup_{J,J',\alpha,\alpha'}\left|\left\langle B_{J,\alpha}, B_{J',\alpha'}\right\rangle_{2,n} - \left\langle B_{J,\alpha}, B_{J',\alpha'}\right\rangle_2\right|.$$

Now since

$$\left|\langle g_1, g_2\rangle_{2,n} - \langle g_1, g_2\rangle_2\right| \leq \left\{\left(C_{A,1} + C'_{A,1}\right) A^*_{n,1} + C_{A,2} A^*_{n,2}\right\} \|g_1\|_2 \|g_2\|_2,$$

if we can show that

$$A^*_{n,2} = O_p\left(\sqrt{\log n/(nH)}\right), \tag{3.6.30}$$

plus the fact that $\sqrt{\log n/(nH)} \gg \sqrt{\log n/n}$, based on the selection of $H^{-1} \sim n^{2/5} \log n$, then there exists a constant $C_A > 0$

$$\frac{\left|\langle g_1, g_2\rangle_{2,n} - \langle g_1, g_2\rangle_2\right|}{\|g_1\|_2 \|g_2\|_2} \leq \left(C_{A,1} + C'_{A,1}\right) A^*_{n,1} + C_{A,2} A^*_{n,2} \leq C_A A^*_{n,2},$$

the order $O_p\left(\sqrt{\log n/(nH)}\right)$ of A_n will be established as in the statement (3.6.26).

The proof of (3.6.30) will be provided case by case with various α, α', J and J', via Bernstein's inequality. For brevity, we set $\eta_i = n^{-1}\left[B_{J,\alpha}(X_{i\alpha}) B_{J',\alpha'}(X_{i\alpha'}) - E\left\{B_{J,\alpha}(X_{i\alpha}) B_{J',\alpha'}(X_{i\alpha'})\right\}\right]$, then $A^*_{n,2} = \sup_{1\leq J\leq N, \alpha=1,2} \left|\sum_{i=1}^n \eta_i\right|$.

We will consider $\alpha = \alpha' = 1$ in the CASE 1.1 to CASE 1.3.

CASE 1.1 when $|J - J'| > 1$. The definition of $B_{J,1}$ in (3.3.3) will guarantee that in probability $B_{J,1}(X_{i1}) B_{J',1}(X_{i1}) = 0$ if $|J - J'| > 1$.

CASE 1.2 when $J = J'$. The variable η_i and its second moment can be simplified as follows

$$\eta_i = n^{-1}\left\{B^2_{J,1}(X_{i1}) - 1\right\}, E\eta_i^2 = \frac{1}{n^2} E\left\{B^2_{J,1}(X_{i1}) - 1\right\}^2 = \frac{1}{n^2}\left\{EB^4_{J,1}(X_{i1}) - 1\right\},$$

in which $EB^4_{J,1}(X_{i1}) = \|b_{J,1}\|_2^{-4}\left(c_{J+1,1} + c^4_{J+1,1}/c^3_{J,1}\right)$. The selection of H will make $EB^4_{J,1}(X_{i1})$ the major term of $\left\{EB^4_{J,1}(X_{i1}) - 1\right\}$, then there exist constants

$c_{\eta,2}$ and $C'_{\eta,2} > 0$ such that

$$c_{\eta,2} n^{-2} H^{-1} \leq E\eta_i^2 \leq C'_{\eta,2} n^{-2} H^{-1}.$$

In terms of the Minkowski's inequality, the k-th absolute moment has the following upper bound

$$E\left|\eta_i\right|^k = n^{-k} E\left|B_{J,1}^2\left(X_{i1}\right) - 1\right|^k \leq n^{-k} 2^{k-1}\left\{EB_{J,1}^{2k}\left(X_{i1}\right) + 1\right\}.$$

where $EB_{J,1}^{2k}\left(X_{i1}\right) = \left\|b_{J,1}\right\|_2^{-2k}\left(c_{J+1,1} + c_{J+1,1}^{2k}/c_{J,1}^{2k-1}\right)$. Hence there exist constants c_{B^2} and C_{B^2} such that

$$c_{B^2}^k H^{1-k} \leq EB_{J,1}^{2k}\left(X_{i1}\right) \leq C_{B^2}^k H^{1-k},$$

then the term $EB_{J,1}^{2k}\left(X_{i1}\right)$ will be the dominant one compared with 1. Hence there exists a constant $C_{\eta,2} > 0$ such that

$$E\left|\eta_i\right|^k \leq C_{\eta,2}^k n^{-k} 2^{k-1} H^{1-k}.$$

Next step is to verify the Cramér's condition

$$\begin{aligned} E\left|\eta_i\right|^k &\leq C_{\eta,2}^k n^{-k} 2^{k-1} H^{1-k} = C_{\eta,2}^k n^{-(k-2)} 2^{k-1} H^{-(k-2)} n^{-2} H^{-1} \\ &\leq \frac{2C_{\eta,2}^2}{c_{\eta,2}}\left(\frac{2C_{\eta,2}}{nH}\right)^{(k-2)} c_{\eta,2} n^{-2} H^{-1} \leq \left\{C_{\eta,2}^*\right\}^{k-2} k! E\eta_i^2, \end{aligned}$$

in which $C_{\eta,2}^* = \left(2C_{\eta,2} n^{-1} H^{-1}\right) \max\left(1, 2C_{\eta,2}^2 c_{\eta,2}^{-1}\right)$. For a large value $\delta > 0$, we have

$$\begin{aligned} P\left\{\left|\sum_{l=1}^n \eta_i\right| \geq \delta\sqrt{\log n/(nH)}\right\} &\leq 2\exp\left[\frac{-\delta^2 \log n/(nH)}{4\sum_{i=1}^n E\eta_i^2 + 2C_{\eta,2}^* \delta\sqrt{\log n/(nH)}}\right] \\ \leq\ 2\exp\left[\frac{-\delta^2 \log n/(nH)}{4n\left\{C'_{\eta,2} n^{-2} H^{-1}\right\} + 2C_{\eta,2}^* \delta\sqrt{\log n/(nH)}}\right]. \end{aligned}$$

If the large enough value δ is taken such that $-\delta^2/\left\{4C'_{\eta,2}+2C^*_{\eta,2}\delta\sqrt{\log n/(nH)}\right\}\leq -3$, then

$$\sum_{n=1}^{\infty}P\left\{\sup_{1\leq J\leq N}\left|\sum_{l=1}^{n}\eta_i\right|\geq\delta\sqrt{\frac{\log n}{nH}}\right\}\leq 2\sum_{n=1}^{\infty}Nn^{-3}\leq 2\sum_{n=1}^{\infty}n^{-2}<\infty.$$

Applying Borel -Cantelli lemma, when $J=J',\alpha=\alpha'=1$ we have

$$A^*_{n,2}=\sup_{1\leq J\leq N}\left|\sum_{i=1}^{n}\eta_i\right|=O_p\left(\sqrt{\log n/(nH)}\right).$$

CASE 1.3 when $\left|J-J'\right|=1$. Without loss of generality we only prove the case that $J'=J+1$. Now $\eta_i=n^{-1}B_{J,1}(X_{i1})B_{J+1,1}(X_{i1})$ has the second moment

$$E\eta_i^2=n^{-2}\left[EB^2_{J,1}(X_{i1})B^2_{J+1,1}(X_{i1})-\left\{EB_{J,1}(X_{i1})B_{J+1,1}(X_{i1})\right\}^2\right],$$

where

$$\begin{aligned}
&\left\{EB_{J,1}(X_{i1})B_{J+1,1}(X_{i1})\right\}^2\\
=\;&\left\|b_{J,1}\right\|_2^{-2}\left\|b_{J+1,1}\right\|_2^{-2}\left[\int\left\{I_{J+1,1}(x_1)-\frac{c_{J+1,1}}{c_{J,1}}I_{J,1}(x_1)\right\}\right.\\
&\left.\times\left\{I_{J+2,1}(x_1)-\frac{c_{J+2,1}}{c_{J+1,1}}I_{J+1,1}(x_1)\right\}f_1(x_1)\,dx_1\right]^2\\
=\;&\left\|b_{J,1}\right\|_2^{-2}\left\|b_{J+1,1}\right\|_2^{-2}\left\{-\frac{c_{J+2,1}}{c_{J+1,1}}\int I_{J+1,1}(x_1)f_1(x_1)\,dx_1\right\}^2\\
=\;&c^2_{J+2,1}\left\|b_{J,1}\right\|_2^{-2}\left\|b_{J+1,1}\right\|_2^{-2},
\end{aligned}$$

and

$$\begin{aligned}
&EB^2_{J,1}(X_{i1})B^2_{J+1,1}(X_{i1})\\
=\;&\left\|b_{J,1}\right\|_2^{-2}\left\|b_{J+1,1}\right\|_2^{-2}\int\left\{I_{J+1,1}(x_1)-\frac{c_{J+1,1}}{c_{J,1}}I_{J,1}(x_1)\right\}^2\\
&\times\left\{I_{J+2,1}(x_1)-\frac{c_{J+2,1}}{c_{J+1,1}}I_{J+1,1}(x_1)\right\}^2f_1(x_1)\,dx_1
\end{aligned}$$

$$= \|b_{J,1}\|_2^{-2} \|b_{J+1,1}\|_2^{-2} \left\{ \left(\frac{c_{J+2,1}}{c_{J+1,1}} \right)^2 \int I_{J+1,1}(x_1) f_1(x_1) dx_1 \right\}$$
$$= \left(c_{J+2,1}^2 \|b_{J,1}\|_2^{-2} \|b_{J+1,1}\|_2^{-2} \right) / c_{J+1,1}.$$

According to (3.6.13), $c_f H \leq c_{J+1,1} \leq C_f H$, so $E\eta_i^2$ will be with the same order as the major term $n^{-2} E B_{J,1}^2(X_{i1}) B_{J+1,1}^2(X_{i1})$, i.e. there exist constants $c_{\eta,3}, C'_{\eta,3} > 0$ such that

$$c_{\eta,3} n^{-2} H^{-1} \leq E\eta_i^2 \leq C'_{\eta,3} n^{-2} H^{-1}.$$

The k-th moment is given by

$$E|\eta_i|^k = n^{-k} E \left| B_{J,1}(X_{i1}) B_{J+1,1}(X_{i1}) - E B_{J,1}(X_{i1}) B_{J+1,1}(X_{i1}) \right|^k$$
$$\leq n^{-k} 2^{k-1} \left[E \left| B_{J,1}(X_{i1}) B_{J+1,1}(X_{i1}) \right|^k + \left| E B_{J,1}(X_{i1}) B_{J+1,1}(X_{i1}) \right|^k \right],$$

where

$$\left| E B_{J,1}(X_{i1}) B_{J+1,1}(X_{i1}) \right|^k = c_{J+2,1}^k \|b_{J,1}\|_2^{-k} \|b_{J+1,1}\|_2^{-k} \sim 1$$

$$E \left| B_{J,1}(X_{i1}) B_{J+1,1}(X_{i1}) \right|^k = \left(c_{J+2,1}^k \|b_{J,1}\|_2^{-k} \|b_{J+1,1}\|_2^{-k} \right) / c_{J+1}^{k-1} \sim H^{1-k}.$$

Hence there exists a constant $C_{\eta,3} > 0$ such that

$$E|\eta_i|^k \leq C_{\eta,3}^k n^{-k} 2^{k-1} H^{1-k}.$$

Similar as in Case 1.2, the conclusion follows by using Bernstein's inequality

$$A_{n,2}^* = \sup_{1 \leq J \leq N} \left| \sum_{i=1}^n \eta_i \right| = O_p\left(\sqrt{\log n / (nH)} \right).$$

CASE 2 when $\alpha = \alpha' = 2$, all the above discussion applies without extra modifications.

CASE 3 when $\alpha \neq \alpha'$. Without of loss generality, suppose $\alpha = 1, \alpha' = 2$.

First we still need to calculate the order of second moment $E\eta_i^2$,

$$E\eta_i^2 = n^{-2}\left[E\left\{B_{J,1}(X_{i1})\,B_{J',2}(X_{i2})\right\}^2 - \left\{EB_{J,1}(X_{i1})\,B_{J',2}(X_{i2})\right\}^2\right].$$

The boundedness of the density function $f(x_1, x_2)$ implies the order $O(H)$ of the absolute mean

$$\begin{aligned}
&\left|EB_{J,1}(X_{i1})\,B_{J',2}(X_{i2})\right| \le E|\eta_i| \\
\le\ & \|b_{J,1}\|_2^{-1}\left\|b_{J',2}\right\|_2^{-1}\int\int\left|b_{J,1}(x_{i1})\,b_{J',2}(x_{i2})\right| f(x_1,x_2)\,dx_1dx_2 \\
\le\ & C_f\left\{\|b_{J,1}\|_2^{-1}\int|b_{J,1}(x_{i1})|\,dx_1\right\}\left\{\left\|b_{J',2}\right\|_2^{-1}\int\left|b_{J',2}(x_{i2})\right|dx_2\right\} \\
\le\ & C_f\left\{1+\frac{c_{J+1,1}}{c_{J,1}}\right\}\left\{\|b_{J,1}\|_2^{-1}H\right\}\left\{1+\frac{c_{J'+1,2}}{c_{J',2}}\right\}\left\{\left\|b_{J',2}\right\|_2^{-1}H\right\} \le C_{B,1}H,
\end{aligned}$$

for some constant $C_{B,1} > 0$, where the last step is derived from the equations (3.6.13) and (3.6.14). As a consequence, $\left|E\left\{B_{J,1}(X_{i1})\,B_{J',2}(X_{i2})\right\}\right|^k \le C_{B,1}^k H^k$. Meanwhile the uniform order of the mean square $O(1)$ will be obtained by Assumption (AS3), and (3.6.13) and (3.6.14),

$$\begin{aligned}
&E\left\{B_{J,1}(X_{i1})\,B_{J',2}(X_{i2})\right\}^2 \\
=\ & \|b_{J,1}\|_2^{-2}\left\|b_{J',2}\right\|_2^{-2}\int\int b_{J,1}^2(x_{i1})\,b_{J',2}^2(x_{i2})\,f(x_1,x_2)\,dx_1dx_2 \\
\ge\ & c_f\left\{\|b_{J,1}\|_2^{-2}\int b_{J,1}^2(x_{i1})\,dx_1\right\}\left\{\left\|b_{J',2}\right\|_2^{-2}\int b_{J',2}^2(x_{i2})\,dx_2\right\} \\
=\ & c_f\left\{1+c_{J+1,1}^2/c_{J,1}^2\right\}\left\{\|b_{J,1}\|_2^{-2}H\right\}\left\{1+c_{J'+1,2}^2/c_{J',2}^2\right\}\left\{\left\|b_{J',2}\right\|_2^{-2}H\right\} \ge c_{B,2}.
\end{aligned}$$

Hence there exist constants $c_\eta, C'_\eta > 0$ such that

$$c_\eta n^{-2} \le E\eta_i^2 \le C'_\eta n^{-2}.$$

First we still need to calculate the order of second moment $E\eta_i^2$,

$$E\eta_i^2 = n^{-2}\left[E\left\{B_{J,1}(X_{i1})\,B_{J',2}(X_{i2})\right\}^2 - \left\{EB_{J,1}(X_{i1})\,B_{J',2}(X_{i2})\right\}^2\right].$$

The boundedness of the density function $f(x_1, x_2)$ implies the order $O(H)$ of the absolute mean

$$\begin{aligned}
&\left|EB_{J,1}(X_{i1})\,B_{J',2}(X_{i2})\right| \le E|\eta_i| \\
\le\ & \|b_{J,1}\|_2^{-1}\left\|b_{J',2}\right\|_2^{-1}\int\int\left|b_{J,1}(x_{i1})\,b_{J',2}(x_{i2})\right| f(x_1,x_2)\,dx_1dx_2 \\
\le\ & C_f\left\{\|b_{J,1}\|_2^{-1}\int|b_{J,1}(x_{i1})|\,dx_1\right\}\left\{\left\|b_{J',2}\right\|_2^{-1}\int\left|b_{J',2}(x_{i2})\right|dx_2\right\} \\
\le\ & C_f\left\{1+\frac{c_{J+1,1}}{c_{J,1}}\right\}\left\{\|b_{J,1}\|_2^{-1}H\right\}\left\{1+\frac{c_{J'+1,2}}{c_{J',2}}\right\}\left\{\left\|b_{J',2}\right\|_2^{-1}H\right\} \le C_{B,1}H,
\end{aligned}$$

for some constant $C_{B,1} > 0$, where the last step is derived from the equations (3.6.13) and (3.6.14). As a consequence, $\left|E\left\{B_{J,1}(X_{i1})\,B_{J',2}(X_{i2})\right\}\right|^k \le C_{B,1}^k H^k$. Meanwhile the uniform order of the mean square $O(1)$ will be obtained by Assumption (AS3), and (3.6.13) and (3.6.14),

$$\begin{aligned}
&E\left\{B_{J,1}(X_{i1})\,B_{J',2}(X_{i2})\right\}^2 \\
=\ & \|b_{J,1}\|_2^{-2}\left\|b_{J',2}\right\|_2^{-2}\int\int b_{J,1}^2(x_{i1})\,b_{J',2}^2(x_{i2})\,f(x_1,x_2)\,dx_1dx_2 \\
\ge\ & c_f\left\{\|b_{J,1}\|_2^{-2}\int b_{J,1}^2(x_{i1})\,dx_1\right\}\left\{\left\|b_{J',2}\right\|_2^{-2}\int b_{J',2}^2(x_{i2})\,dx_2\right\} \\
=\ & c_f\left\{1+c_{J+1,1}^2/c_{J,1}^2\right\}\left\{\|b_{J,1}\|_2^{-2}H\right\}\left\{1+c_{J'+1,2}^2/c_{J',2}^2\right\}\left\{\left\|b_{J',2}\right\|_2^{-2}H\right\} \ge c_{B,2}.
\end{aligned}$$

Hence there exist constants $c_\eta, C'_\eta > 0$ such that

$$c_\eta n^{-2} \le E\eta_i^2 \le C'_\eta n^{-2}.$$

For any $k > 2$, the k-th moment of $|\eta_i|$ is given by

$$\begin{aligned} E|\eta_i|^k &= n^{-k} E\left|B_{J,1}(X_{i1}) B_{J',2}(X_{i2}) - E B_{J,1}(X_{i1}) B_{J',2}(X_{i2})\right|^k \\ &\leq n^{-k} 2^{k-1} \left[E\left|B_{J,1}(X_{i1}) B_{J',2}(X_{i2})\right|^k + \left|E B_{J,1}(X_{i1}) B_{J',2}(X_{i2})\right|^k\right] \end{aligned}$$

where there exists a constant $C_{B'} > 0$ such that

$$\begin{aligned} & E\left|B_{J,1}(X_{i1}) B_{J',2}(X_{i2})\right|^k \\ \leq\ & \|b_{J,1}\|_2^{-k} \left\|b_{J',2}\right\|_2^{-k} \int\int \left|b_{J,1}^k(x_{i1}) b_{J',2}^k(x_{i2})\right| f(x_1, x_2)\, dx_1 dx_2 \\ \leq\ & C_f \left\{\|b_{J,1}\|_2^{-k} \int |b_{J,1}(x_{i1})|^k\, dx_1\right\} \left\{\left\|b_{J',2}\right\|_2^{-k} \int \left|b_{J',2}(x_{i2})\right|^k dx_2\right\} \\ \leq\ & C_f \left\{1 + \frac{c_{J+1,1}^k}{c_{J,1}^k}\right\} \left\{1 + \frac{c_{J'+1,2}^k}{c_{J',2}^k}\right\} \left\{\|b_{J,1}\|_2^{-k} \left\|b_{J',2}\right\|_2^{-k}\right\} H^2 \\ \leq\ & C_f \left\{1 + \frac{c_{J+1,1}^k}{c_{J,1}^k}\right\} \left\{1 + \frac{c_{J'+1,2}^k}{c_{J',2}^k}\right\} \left\{c_f\left(1 + c_f/C_f\right)\right\}^{-k} H^{2-k} \leq C_{B'}^k H^{2-k}. \end{aligned}$$

Thus there is a constant $C_\eta > 0$ such that

$$\begin{aligned} E|\eta_i|^k &\leq n^{-k} 2^{k-1} \left[C_{B'}^k H^{2-k} + C_{B,1}^k H^k\right] \leq (C_\eta)^k n^{-k} 2^{k-1} H^{2-k} \\ &\leq \frac{2C_\eta^2}{c_\eta} \left(2C_\eta n^{-1} H^{-1}\right)^{k-2} c_\eta n^{-2} \leq \left\{\frac{2C_\eta}{nH} \max\left(\frac{2C_\eta^2}{c_\eta}, 1\right)\right\}^{k-2} k! E\eta_i^2. \end{aligned}$$

Employing the Bernstein's inequality and the fact that $E\eta_i^2 \sim n^{-2}$, for any

$1 \leq J, J' \leq N$, $\alpha \neq \alpha'$,

$$\sup_{1 \leq J \leq N} \left|\sum_{i=1}^n \frac{B_{J,\alpha}(X_{i\alpha}) B_{J',\alpha'}(X_{i\alpha'}) - E\left\{B_{J,\alpha}(X_{i\alpha}) B_{J,'\alpha'}(X_{i\alpha'})\right\}}{n}\right| = O_p\left(\sqrt{\frac{\log n}{n}}\right).$$

Hence for any $1 \leq J, J' \leq N$, $\alpha, \alpha' = 1, 2$, the proof of (3.6.30) is completed. □

The next lemma on the positive definiteness of matrix $\left(n^{-1}\mathbf{B}^T\mathbf{B}\right)^{-1}$ is a sufficient step to achieve Lemma 3.6.10.

Lemma 3.6.9. *Under Assumptions (AS3) and (AS4), for the matrix* $S = \left(s_{jj'}\right)_{j,j'=1}^{dN+1} = \left(n^{-1}\mathbf{B}^T\mathbf{B}\right)^{-1}$, *there exist constants* $C_S > c_S > 0$ *such that with probability approaching to* 1, *one has*

$$c_S I_{2N+1} \leq S^{-1} \leq C_S I_{2N+1}. \tag{3.6.31}$$

PROOF. Take a real vector $\varsigma = \left(u_0, u_{1,1}, ..., u_{N,1}, u_{1,2}, ..., u_{N,2}\right)^T \in R^{2N+1}$, one has

$$\left\|\varsigma^T B_*\right\|_{2,n}^2 = \varsigma^T \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \left\langle B_{J,\alpha}, B_{J',\alpha'} \right\rangle_{2,n} \end{pmatrix} \varsigma = \varsigma^{\mathbf{T}} S^{-1} \varsigma, \tag{3.6.32}$$

where we denote $B_* = \left\{1, B_{1,1}\left(X_1\right), ..., B_{N,2}\left(X_2\right)\right\}^T$. Meanwhile, the definition of A_n in (3.6.26) entails in particular that

$$\left\|\varsigma^T B_*\right\|_2^2 (1 + A_n) \geq \left\|\varsigma^T B_*\right\|_{2,n}^2 \geq \left\|\varsigma^T B_*\right\|_2^2 (1 - A_n),$$

while (3.6.19) means that there exist constants $C_S > c_S > 0$ such that

$$C_S \left(u_0^2 + \sum_{J,\alpha} u_{J,\alpha}^2\right) \geq \left\|\varsigma^T B_*\right\|_2^2 = u_0^2 + \left\|\sum_{J,\alpha} u_{J,\alpha} B_{J,\alpha}\left(x_\alpha\right)\right\|_2^2 \geq c_S \left(u_0^2 + \sum_{J,\alpha} u_{J,\alpha}^2\right),$$

hence

$$C_S \left(u_0^2 + \sum_{J,\alpha} u_{J,\alpha}^2\right)(1 + A_n) \geq \left\|\varsigma^T B_*\right\|_{2,n}^2 \geq c_S \left(u_0^2 + \sum_{J,\alpha} u_{J,\alpha}^2\right)(1 - A_n). \tag{3.6.33}$$

Putting together (3.6.32), (3.6.33), one concludes that with probability approaching 1

$$C_S \varsigma^{\mathbf{T}} \varsigma = C_S \left(u_0^2 + \sum_{J,\alpha} u_{J,\alpha}^2\right) \geq \varsigma^{\mathbf{T}} S^{-1} \varsigma \geq c_S \left(u_0^2 + \sum_{J,\alpha} u_{J,\alpha}^2\right) = c_S \varsigma^{\mathbf{T}} \varsigma,$$

which gives (3.6.31). □

Lemma 3.6.10. *Under Assumptions (AS1) to (AS6), for any* $x_1 \in [0,1]$ *and* $I_1(x_1)$ *defined in (3.6.2), one has*

$$\sup_{x_1 \in [0,1]} E\left\{ I_1^2(x_1) \middle| \tilde{\mathbf{X}} \right\} = O_p\left(n^{-1}\right). \tag{3.6.34}$$

PROOF. It is known that $\tilde{\mathbf{a}} = \left(\mathbf{B}^T\mathbf{B}\right)^{-1}\mathbf{B}^T\mathbf{E}$, then the conditional mean square of $\tilde{\varepsilon}_2^*(X_{l2})$ given $\tilde{\mathbf{X}}$ is $E\left[\left\{\tilde{\varepsilon}_2^*(X_{l2})\right\}^2 \middle| \tilde{\mathbf{X}}\right]$

$$\begin{aligned}
&= E\left(\left\{ \tilde{\mathbf{a}}^T \mathbf{P}_{0_{N+1}, I_N} \left(e_l^T \mathbf{B}\right)^T \right\}^T \left\{ \tilde{\mathbf{a}}^T \mathbf{P}_{0_{N+1}, I_N} \left(e_{l'}^T \mathbf{B}\right)^T \right\} \middle| \tilde{\mathbf{X}} \right) \\
&= e_l^T \mathbf{B} \mathbf{P}_{0_{N+1}, I_N} \left(\mathbf{B}^T\mathbf{B}\right)^{-1} \mathbf{B}^T \cdot E\left(\mathbf{E} \cdot \mathbf{E}^T \middle| \tilde{\mathbf{X}} \right) \cdot \mathbf{B} \left(\mathbf{B}^T\mathbf{B}\right)^{-1} \mathbf{P}_{0_{N+1}, I_N} \mathbf{B}^T e_{l'}
\end{aligned}$$

Based on Assumption (AS2), we have $E\left\{ \left(\mathbf{E} \cdot \mathbf{E}^T\right) \middle| \mathbf{X}_1, ..., \mathbf{X}_n \right\} \leq C_\sigma^2 I_n$ in the matrix sense, then applying these two matrices to a quadratic form with vector $\left\{ \mathbf{B}\left(\mathbf{B}^T\mathbf{B}\right)^{-1} \mathbf{P}_{0_{N+1}, I_N} \mathbf{B}^T e_{l'} \right\}$, one has $E\left[\left\{\tilde{\varepsilon}_2^*(X_{l2})\right\}^2 \middle| \tilde{\mathbf{X}}\right]$

$$\begin{aligned}
&\leq C_\sigma^2 \cdot \left\{ \left(e_l^T \mathbf{B}\right) \mathbf{P}_{0_{N+1}, I_N} \right\} \cdot \left(\mathbf{B}^T\mathbf{B}\right)^{-1} \cdot \left\{ \mathbf{P}_{0_{N+1}, I_N} \left(e_{l'}^T \mathbf{B}\right)^T \right\} \\
&= n^{-1} C_\sigma^2 \cdot \left\{ 0_{N+1}, B_{1,2}(X_{l2}) ..., B_{N,2}(X_{l2}) \right\} S \left\{ 0_{N+1}, B_{1,2}(X_{l'2}) ..., B_{N,2}(X_{l'2}) \right\}' \\
&= n^{-1} C_\sigma^2 \cdot \sum_{1 \leq J, J' \leq N} B_{J,2}(X_{l2})\, s_{J+N+1, J'+N+1} B_{J',2}(X_{l'2}),
\end{aligned}$$

where the $s_{J+N+1,J'+N+1}$'s are elements of S in Lemma 3.6.9. Plugging in the above term, and employing (3.6.4), the term $E\left\{ I_1^2(x_1) \middle| \tilde{\mathbf{X}} \right\}$

$$\begin{aligned}
&\leq \frac{C_\sigma^2}{n^3} \sum_{l,l'=1}^{n} K_h(X_{l1} - x_1) K_h(X_{l'1} - x_1) \\
&\quad \sum_{1 \leq J, J' \leq N} B_{J,2}(X_{l2})\, s_{J+N+1, J'+N+1} B_{J',2}(X_{l'2}) \\
&\leq \frac{C_\sigma^2}{n} \sum_{1 \leq J, J' \leq N} s_{J+N+1, J'+N+1} \sum_{1 \leq J \leq N} \left\{ n^{-1} \sum_{l=1}^{n} K_h(X_{l1} - x_1) B_{J,2}(X_{l2}) \right\}^2 \\
&\leq \frac{C_\sigma^2}{n} C_S \sum_{1 \leq J \leq N} \left\{ n^{-1} \sum_{l=1}^{n} K_h(X_{l1} - x_1) B_{J,2}(X_{l2}) \right\}^2,
\end{aligned}$$

where C_S is the same as in (3.6.31). Now using Lemma 3.6.5, one has with probability approaching to 1

$$\sup_{x_1\in[0,1]} E\left\{ I_1^2(x_1)\middle|\tilde{\mathbf{X}}\right\} \le \frac{C_\sigma^2}{n} C_S \sum_{1\le J\le N} H = \frac{C}{n},$$

which implies (3.6.34). □

Lemma 3.6.11. *Under Assumptions (AS1) to (AS6), for $I_2(x_1)$ as defined in (3.6.2), one has*

$$\sup_{x_1\in[0,1]} |I_2(x_1)| = \sup_{x_1\in[0,1]} \left| n^{-1}\sum_{l=1}^{n} K_h(X_{l1}-x_1)\cdot n^{-1}\sum_{i=1}^{n}\tilde{\varepsilon}_2^*(X_{i2})\right| = O_p\left(\frac{N}{n}\sqrt{\log n}\right).$$

PROOF. Based on (3.3.12), $n^{-1}\sum_{i=1}^{n}\tilde{\varepsilon}_2^*(X_{i2})$ can be expressed as

$$n^{-1}\sum_{i=1}^{n}\left\{\sum_{J=1}^{N}\tilde{a}_{J,2}B_{J,2}(X_{i2})\right\} = \sum_{J=1}^{N}\tilde{a}_{J,2}\left\{n^{-1}\sum_{i=1}^{n}B_{J,2}(X_{i2})\right\}.$$

Lemma 3.6.7 helps to get

$$\left|\sum_{J=1}^{N}\tilde{a}_{J,2}\right| \le \left\{N\cdot\sum_{J=1}^{N}\tilde{a}_{J,2}^2\right\}^{1/2} \le \left\{N\cdot\tilde{\mathbf{a}}^T\tilde{\mathbf{a}}\right\}^{1/2} = O_p\left(Nn^{-1/2}\right).$$

Now it is clear from (3.6.28) and (3.6.29) that

$$\sup_{1\le J\le N}\left|n^{-1}\sum_{i=1}^{n}B_{J,2}(X_{i2})\right| \le A_{n,1}^* = O_p\left(\sqrt{n^{-1}\log n}\right),$$

hence

$$n^{-1}\sum_{i=1}^{n}\tilde{\varepsilon}_2^*(X_{i2}) \le \left|\sum_{J=1}^{N}\tilde{a}_{J,2}\right|\cdot\sup_{1\le J\le N}\left|n^{-1}\sum_{i=1}^{n}B_{J,2}(X_{i2})\right| = O_p\left(\frac{N}{n}\sqrt{\log n}\right). \tag{3.6.35}$$

By Assumption (AS4) on the kernel function K, standard theory on kernel density estimation entails that $\sup_{x_1\in[0,1]}\left|n^{-1}\sum_{l=1}^{n}K_h(X_{l1}-x_1)\right| = O_p(1)$. Thus with (3.6.35) the lemma follows immediately. □

Lemma 3.6.12. *Under Assumptions (AS1) to (AS6) and (AS2'), and with* $I_1(x_1)$ *defined in (3.6.2), one has*

$$\sup_{x_1\in[0,1]}|I_1(x_1)| = \sup_{x_1\in[0,1]}\left|n^{-1}\sum_{l=1}^{n}K_h(X_{l1}-x_1)\cdot\tilde{\varepsilon}_2^*(X_{l2})\right| = O_p\left(\sqrt{\log n/n}\right). \tag{3.6.36}$$

PROOF. The discretization idea will be employed again in this lemma, by dividing the interval $[0,1]$ into M_n equally spaced intervals with disjoint endpoints $0 = x_{1,0} < x_{1,1} < \ldots < x_{1,M_n} = 1$. As in (3.6.16), we start with

$$\sup_{x_1\in[0,1]}|I_1(x_1)| = \sup_{0\le k\le M_n}|I_1(x_{1,k})| + \sup_{1\le k\le M_n}\sup_{x_1\in\left[x_{1,k-1},x_{1,k}\right]}|I_1(x_1)-I_1(x_{1,k})|. \tag{3.6.37}$$

Note that for any $x_1\in[0,1]$, (3.3.12) and (3.6.2) imply that

$$\tilde{\varepsilon}_2^*(X_{l2}) = \sum_{J=1}^{N}\tilde{a}_{J,2}B_{J,2}(X_{l2}) = \left(e_l^T\mathbf{B}\right)\mathbf{P}_{0_{N+1},I_N}\left(\mathbf{B}^T\mathbf{B}\right)^{-1}\mathbf{B}^T\mathbf{E}.$$

$$\begin{aligned} I_1(x_1) &= n^{-1}\sum_{l=1}^{n}K_h(X_{l1}-x_1)\,\tilde{\varepsilon}_2^*(X_{l2}) \\ &= n^{-1}\sum_{l=1}^{n}K_h(X_{l1}-x_1)\left(e_l^T\mathbf{B}\right)\mathbf{P}_{0_{N+1},I_N}\left(\mathbf{B}^T\mathbf{B}\right)^{-1}\mathbf{B}^T\mathbf{E}. \end{aligned}$$

Since $I_1(x_1)$ is a linear combination of the noise terms in $\mathbf{E}$, its conditional distribution given $\tilde{\mathbf{X}}$ is normal with mean 0, under Assumption (AS2'). Let

$$R\left(\tilde{\mathbf{X}},x_{1,k}\right) = \left(\operatorname{var}\left\{I_1(x_{1,k})\middle|\tilde{\mathbf{X}}\right\}\right)^{-1/2}I_1(x_{1,k}),$$

then the conditional distribution of $R\left(\tilde{\mathbf{X}},x_{1,k}\right)$ given $\tilde{\mathbf{X}}$ is standard normal. In what follows, we use the well-known tail property of the normal distribution, i.e. $1-\Phi(x)\le\phi(x)/x$, for $x\ge 0$, hence there exists some $c>0$, such that $1-\Phi(x)\le c\phi(x)$ for

large x, where $\Phi(x)$ and $\phi(x)$ are the cumulative distribution function and the density function of the standard normal. Take $t_n = \sqrt{16 \log n}$, then there exists a constant c such that for large enough n

$$\sum_{n=1}^{n} P\left\{ \sup_{0\le k\le M_n} \left| R\left(\tilde{\mathbf{X}}, x_{1,k}\right)\right| \ge t_n \middle| \tilde{\mathbf{X}} \right\} = \sum_{n=1}^{n} P\left\{ \sup_{0\le k\le M_n} |Z| \ge t_n \right\}$$
$$\le \sum_{n=1}^{n} M_n \cdot P\{|Z| \ge t_n\} \le c \sum_{n=1}^{n} M_n \cdot \exp\left\{-\frac{t_n^2}{2}\right\} \le c \sum_{n=1}^{n} M_n n^{-8} < \infty,$$

where $Z \sim N(0,1)$. Consequently for a large value $\delta > 0$, we have

$$\sum_{n=1}^{n} P\left\{ \sup_{0\le k\le M_n} \left| R\left(\tilde{\mathbf{X}}, x_{1,k}\right)\right| \ge \delta\sqrt{\log n} \right\} < \infty,$$

the Borel-Cantelli Lemma will then imply that $\sup_{0\le k\le M_n} \left| R\left(\tilde{\mathbf{X}}, x_{1,k}\right)\right| = O_p\left(\sqrt{\log n}\right)$. The conditional variance of $I_1\left(x_{1,k}\right)$ given $\tilde{\mathbf{X}}$ is defined as follows:

$$\operatorname{var}\left\{ I_1\left(x_{1,k}\right) \middle| \tilde{\mathbf{X}} \right\} = E\left[\left\{ I_1\left(x_{1,k}\right) - E I_1\left(x_{1,k}\right) \right\}^2 \middle| \tilde{\mathbf{X}} \right] = E\left\{ I_1^2\left(x_{1,k}\right) \middle| \tilde{\mathbf{X}} \right\}.$$

Now Lemma 3.6.10 implies that $\sup_{0\le k\le M_n} \operatorname{var}\left\{ I_1\left(x_{1,k}\right) \middle| \tilde{\mathbf{X}} \right\} = O_p\left(n^{-1}\right)$. Hence

$$\sup_{0\le k\le M_n} \left| I_1\left(x_{1,k}\right)\right| \le \sup_{0\le k\le M_n} \left| R\left(\tilde{\mathbf{X}}, x_{1,k}\right)\right| \sup_{0\le k\le M_n} \sqrt{\operatorname{var}\left\{ I_1\left(x_{1,k}\right) \middle| \tilde{\mathbf{X}} \right\}} \quad (3.6.38)$$
$$= O_p\left(\sqrt{\frac{\log n}{n}}\right). \quad (3.6.39)$$

Next, with (3.3.12) and (3.6.18), we note that

$$\sup_{1\le k\le M_n} \sup_{x_1 \in \left[x_{1,k-1}, x_{1,k}\right]} \left| I_1(x_1) - I_1\left(x_{1,k}\right)\right|$$
$$= \sup_{1\le k\le M_n} \sup_{x_1 \in \left[x_{1,k-1}, x_{1,k}\right]} \left| n^{-1} \sum_{l=1}^{n} \left\{ K_h\left(X_{l1} - x_1\right) - K_h\left(X_{l1} - x_{1,k}\right) \right\} \cdot \tilde{\varepsilon}_2^*\left(X_{l2}\right) \right|$$

$$\leq \sup_{1\leq k\leq M_n} \sup_{x_1\in\left[x_{1,k-1},x_{1,k}\right]} \left|K_h\left(X_{l1}-x_1\right)-K_h\left(X_{l1}-x_{1,k}\right)\right|$$

$$\times \sup_{1\leq l\leq n}\left|\sum_{J=1}^{N}\tilde{a}_{J,2}B_{J,2}\left(X_{l2}\right)\right|$$

$$\leq CM_n^{-1}h^{-2}H^{-1/2}\sum_{J=1}^{N}\left|\tilde{a}_{J,2}\right| \leq CM_n^{-1}h^{-2}H^{-1/2}N^{1/2}\left(\sum_{J=1}^{N}\tilde{a}_{J,2}^2\right)^{1/2},$$

which, when combined with (3.6.22), leads to

$$\sup_{1\leq k\leq M_n} \sup_{x_1\in\left[x_{1,k-1},x_{1,k}\right]} \left|I_1\left(x_1\right)-I_1\left(x_{1,k}\right)\right| \tag{3.6.40}$$

$$= O_p\left(M_n^{-1}h^{-2}N\cdot N^{1/2}n^{-1/2}\right) = o_p\left(n^{-1}\right). \tag{3.6.41}$$

due to the choice of $c_D n^6 \leq M_n \leq C_D n^6$ in Lemma 3.6.4.

Now (3.6.37), (3.6.39) and (3.6.41) establish the lemma. □

CHAPTER 4

Application to Seasonality Analysis

4.1 Introduction

Many studies demonstrate the influence of land use and land cover change on local and regional climate. The Climate and Land use Interaction Project, or CLIP (http://clip.msu.edu) attempts to understand the nature and magnitude of the interactions of climate and land use/cover change across East Africa.

Phenological information reflecting the seasonal variability of vegetation is an important input variable in regional climate models such as Regional Atmosphere Simulation System (RAMS). It varies not only among different vegetation types but also with geographic locations (latitude and longitude).

Many climate models use simple functions for vegetation parameters since, to first order, the planet is warmer and wetter as you approach the equator. However, east Africa is unique in having semiarid grasslands along the equator, and drastically different surface conditions govern the radiationbudget in this region. Climate models

are dependent on an accurate representation of the surface radiation budget to replicate atmospheric development. Thus, modeling climate for a unique area like east Africa requires a different treatment of vegetation characteristics.

RAMS version 4.4 (Cotton et al. 2003), a state-of-the-art three dimensional atmospheric model, includes a representation of vegetation called the Land-Ecosystem-Atmosphere Feedback, version 2 (LEAF-2) (Walko et al. 2000). For a given land cover class, LEAF-2 provides functions for severalvegetation characteristics including LAI, fractional cover, roughness length, and displacement height. Although these characteristics are interrelated, we will consider only LAI here.

Remote sensing parameterization for land surface schemes in climate models is focusing on the transformation of categorical LULC information into quantitative land surface biophysical parameters (Pitman 2003). The parameters that will result from this analysis, and that will be inputs to the regional climate model, include surface albedo, fractional vegetative cover, leaf area index (both senescence and green) and above ground biomass. In this paper we will investigate the variation of LAI temporally and spatially for each land type.

The phenological discrepancy between the RAMS model and the remote sensing measurement given in Section 2 will show that the pre-assumed relationship is significantly different from the colleted information from MODIS (Moderate Resolution Imaging Spectroradiometer).

Based on the observations of LAI of MODIS data, the polynomial spline regression is employed to fit the function of each land type in East Africa. The fitted curve is a piecewise polynomial joined at knots, which are the equally-spaced time points of

one whole year. The estimated curve is derived from the least square procedure. In this paper, the linear spline is used for simple implementation and reliable theoretical property. The corresponding statistical theory were provided in Huang (2003) and Wang and Yang (2005).

There are two great advantages of spline regression. It is non-parametric, i.e. the estimation only depends on the available data without assuming any specific form of the model. Second, it has a specific expression for the estimated function. Other nonparametric regression methods such as kernel or local polynomial do not produce an overall function formula. Hence the spline function is preferred for data-driven estimation and future prediction.

We will develop the function first temporally and then further investigate the spatial influence. In other words, the estimate function of LAI will rely on the time and the spatial index (latitude and longitude). Compared with the simulation result derived from RAMS, the estimates at the observations will play the role of "observation".

The research objective of this study was to derive spatially explicit phenologies for all LULC types in East Africa for improved parameterization of regional climate methods (such as RAMS). By addressing this objective, the following two questions must be addressed:

What are the differences in LAI between the observations from MODIS censor and the simulated values from RAMS?

Are there any significant differences among the land types and do they if any vary with geographic locations?

4.2 Method

4.2.1 Study Area and Data Description

East Africa is a region that is undergoing rapid land use change and where changes in climate would have serious consequences for people's livelihoods and requiring new coping and land use strategies.

Consequently, uncertainty in climate modeling is expected to be high, partly due to uncertainty related to the use of generic land cover parameters including their phenological functions. The CLIP project also created a new land use land cover (LULC) classification based on the best available international LULC products for the East Africa region (cite Ge et al 2005, Torbick et al 2005a). The new LULC classification (Torbick et al 2005b), labeled "CLIP-cover," was used as the spatial land cover layer for which the LAI remote sensing data were extracted by LULC, or land type.

Two primary data sets are used to develop the phenological curves. The first is a hybrid LULC classification with 34 land types at 1km spatial resolution for the entire study region. The hybrid combines the strengths of Global Land Cover for the year 2000 (GLC2000) (Mayaux et al 2004) and Africover (Africover 2002) LULC products. Assessments determined GLC2000 more accurately classified natural land cover types, while Africover more accurately classified human-managed landscapes (Torbick et al 2005b). The new hybrid CLIP Cover captures these strengths geospatially for a single LULC for the study region.

The second is LAI from the MODIS instrument on the Terra satellite platform.

Briefly, LAI is a description of vegetation structure and the amount of plant canopy relative to a unit on the surface. In climate models, LAI is used to represent components of energy balance equations between the surface and lower atmospheric boundaries. The MODIS LAI product used, MOD15A2 v4.0 (Knyazikhin et al. 1999), is available at 8-day temporal intervals at 1km spatial resolution covering the entire study region in a 2-dimensional tessellation. The data was obtained through the National Aeronautics and Space Administration (NASA) Land Processes Distribution Active Archive Center.

Data was obtained from February 2000 to December 2003 at 8-day intervals. Data preprocessing included mosaicing tiles, rescaling data values, quality control for cloud cover and fill values, and reprojecting data from Integerized Sinusoidal Projection into Lambert Azimuthal Equal Area. Using the hybrid LULC product, LAI data was subset into tables by LULC type. Each table contains 8-day LAI from February 2000 - December 2003 by LULC type with geographic coordinates (latitude / longitude) at each pixel (or LAI value) representing spatial location information.

4.2.2 Polynomial Spline Regression

The imagery data for each land cover type is collected from January 2000 to December 2003, roughly every 8 days for each pixel (solution = 1 kilometer). Some difficulties that have been encountered were empty cells due to cloud cover, small size of some land covers.

First calculate the mean for each grid (0.1 degree) at every available Julian day.

For each specific grid, the LAI of each land cover type can be seen as a series of data points over explanatory variable time (one year). So we treat each series of LAI at each grid as a univariate function of time. The linear spline regression was employed to get the spline estimator of LAI, which is shown in Figures 4.11 and 4.12 .

In order to capture the spatial feature of each land cover type, we combine all the regression coefficient of linear splines. Then for each coefficient we perform the polynomial regression on the spatial index, latitude and longitude. The corresponding outcomes are listed in Tables 4.5 - 4.8.

The dependence of LAI on time is investigated in the framework of nonparametric regression. To introduce this concept, let $\{(T_i, Y_i)\}_{i=1}^{n}$ be identically and independently distributed observations, satisfying

$$Y_i = m(T_i) + \sigma(T_i)\varepsilon_i, i = 1, ..., n.$$

where the errors ε_i have mean zero and variance one. The mean function $m(t)$ and standard deviation function $\sigma(t)$ are not assumed to be of any specific form but have to be estimated from the data directly, see Wang and Yang (2005). If the data actually follows a polynomial regression model, the function $m(t)$ is a polynomial of t and $\sigma(t)$ will typically be a constant.

To introduce the concept of spline, one divides the finite interval $[a, b]$ into $(N+1)$ subintervals $J_j = [t_j, t_{j+1}), j = 0, 1, ..., N-1, J_N = [t_N, b]$. A sequence of equally-spaced points $\{t_j\}_{j=1}^{N}$, called interior knots, are given as

$$t_0 = a < t_1 < \cdots < t_N < b = t_{N+1}, t_j = a + jh, j = 0, 1, ..., N+1,$$

in which $h = (b-a)/(N+1)$ is the distance between neighboring knots. We ap-

proximate $m(t)$ by linear spline. These are piecewise linear functions, linear on J_j each and continuous on the entire interval $[a, b]$.

The linear spline estimator of $m(t)$ based on data $\{(T_i, Y_i)\}_{i=1}^n$ is given by

$$\hat{m}(t) = \hat{a}_0 + \sum_{j=1}^{N} \hat{a}_j (t - t_j)_+ + \hat{a}_{N+1} t \tag{4.2.1}$$

where the coefficient are the solutions of the following least square problem

$$\{\hat{a}_0, ..., \hat{a}_{N+1}\}^T = \text{argmin}_{\,R^{N+2}} \sum_{i=1}^{n} \left\{ Y_i - a_0 - \sum_{j=1}^{N} a_j (T_i - t_j)_+ - a_{N+1} t \right\}$$

in which $(t - t_j)_+ = \max\{0, t - t_j\}$ is the so-called "truncated linear function" with truncation at knot t_j.

4.2.3 Spline Fitting for LAI by LULC Type

At first we resample the LAI pixels within 0.1 latitude degree and 0.1 longitude degree together as one grid block. In order to get the representative LAI values, the spatially averaged LAI at each grid is obtained for each available Julian day. The second step is to get the means of the same Julian days over four years. After the above two-step averages, LAI means of a whole year at each grid is available.

Based on the LAI means, the equation (4.2.1) is established after one step least squared procedure for each grid. To avoid the non-continuity difference between the values of early January and late December, we duplicate the one year data to create a two-year data, hence $[a, b] = [0, 730]$. For uniformity across various LULC types and locations, we pick one knot every two months, i.e. $N = 11$

$$LAI(t) = \hat{a}_0 + \sum_{j=1}^{11} \hat{a}_j (t - t_j)_+ + \hat{a}_{12} t, t_j = 365 \cdot \frac{j}{6}, j = 1, ..., 11 \tag{4.2.2}$$

Let Z =LAI, x = latitude, y = longitude, t =Julian day. For each LC type we develop the LAI function as follows,

$$Z(x,y,t) = \hat{a}_0(x,y) + \sum_{j=1}^{11} \hat{a}_j(x,y) \cdot (t - t_j)_+ + \hat{a}_{12}(x,y)\, t, \tag{4.2.3}$$

The coefficients $\hat{a}_j(x,y)$ for $j = 0, 1, ..., 12$, are estimated based on the MODIS data at each individual grid. Different LC type will have different coefficients set, see Tables 4.5 - 4.8.

4.3 Results

4.3.1 Land Cover Phenologies

In order to show the magnitude of the difference driven by the spatial affect, in particular the latitude, the linear spline curves estimated by formula (4.2.1), the RAMS simulation curve and the difference curve are provided respectively at equator, 5^o north, , and 5^o south. Each grid points covered the area of .1 by .1 squared degrees, the longitudinal of three grid points are chosen to be as close as possible. In Figures 4.11 and 4.12, the green solid line represents the LAI at 5^o North, the red dashed line for the equator, and the blue dotted line for 5^o South.

Figures 4.11 and 4.12 illustrates several examples of the seasonal variation in LAI for common classes in the study area. The lower right graphs are the trigonometric curve of LAI over time for two land types, open to very open trees, and rainfed herbaceous crop. Although the length of vertical axis of the RAMS curve is the same 0.2, the start points of the range are different though. While in the figures of the linear

splines the range of the vertical is 6, from 0 to 6, that is a substantial difference. If the same scale is chosen as the one for the spline estimates, no distinguishable differences occur among the RAMS curve at the three selected latitudes. While there is no longitude effect in RAMS, it plays an unnoticeable role in the system. There is only one valley for northern latitudes and one peak for south latitudes in RAMS, and the valley or peak point is in the exact middle of the year. At the equator it is a flat straight line no matter what land type is represented.

The linear spline estimators have a better fit spatially and temporally compared with RAMS. The green solid line (5^o N) achieves its peak point of LAI around August, while all the blue dotted lines (5^o S) show the largest LAI value in the spring, such as early March for Rainfed Herbaceous Crop. Not surprising are the fact that the northern and southern curves are symmetric about the center, June, for each type because the two locations are symmetric about the equator. For both land types, the LAI at the equator has greater LAI than those far away from the equator. Especially for land type rainfed herbaceous crop, the regression line at the equator is far above both the spline regression lines at 5^o N and 5^oS latitude. The linear spline estimates produce two noticeable valleys at the equator. That is a big difference from the constant LAI value of RAMS. The LAI varies at the equator over time, it is not fixed given the keep-changing weather condition.

The lower right graphs in Figures 4.11 and 4.12 show that the differences between the LAI values from RAMS and the linear spline estimates. From the graph, except there is little overlap between the difference at equator and the "0 line" for land types, all the remaining distance is very large. The statistical testing of the difference

is given in next section.

In summary, the observed LAI and resultant splines are distinctly different from the RAMS/LEAF-2 default parameterization, with the LEAF-2 parameterization completely failing to capture the seasonality at the equator or in the regions +/- 5 away. The spline parameterizations accurately capture bimodal greening events at the equator, unimodal features away from the equator, and the very low LAI for maize regions following harvest.

4.3.2 Sensitivity and Uncertainty

Confidence band of a function estimator is the collection of simultaneous confidence intervals over the range of data. It can be used to test the hypothesized curve. Linear spline confidence bands were developed in Chapter 2. Given a small significance level (less that 0.05), the confidence bands based on the sample information can be obtained. If the null curve is totally covered by the upper and lower confidence bands, then its deviation from the true curve is insignificant and will be accepted as a valid representation of the true curve; otherwise, it should be rejected as the null curve, since it is significantly different from the data pattern.

In this paper, the hypotheses for a land type are:

H_0 :LAI trend curve follows the RAMS Curve Ha :LAI trend curve does not follow the RAMS Curve.

For the test, the same data from the previous four land types for comparison is used in Figure 4.13 to 4.16. The upper right corner figures represent the LAI average

value for each grid block. The three grid blocks are chosen to have almost the same longitude. The triangle is for LAI at equator, the diamond for North 5 degree, the cross for the South 5 degree. The blue solid line represents the LAI value of the RAMS, the green solid line is the linear spline regression line, and the dashed red lines (upper and lower) are the confidence bands derived from the MODIS data given the significance level 0.001.

Although tested with a significance level as low as 0.001, the RAMS curves are above both bands for 5^oN and 5^o S. At the equator there is some overlap for deciduous woodland and deciduous shrubland with sparse trees, however it is still far from being totally covered by the bands. Therefore this test illustrates that the RAMS curves overestimate the LAI, with the difference being significantly large indicated from the small p-value$<$ 0.001.

4.3.3 Phenological Functions of Land Cover

To model the LAI spatially, the coefficients in equation (4.2.3) are further approximated with quadratic functions of x and y. The same four dominant land types are selected for analysis.

From Section 4.2, a coefficient set with 13 coefficient elements $\{\hat{a}_j(x,y)\}_{j=0}^{12}$ is obtained. Each coefficient element $\hat{a}_j(x,y)$ is related to all grid point. For better regression, the outliers (grid points) are first detected and removed from the coefficients based on the screening of the kernel density estimators. Then the corresponding part in the data set will be left out too. The deleted outliers are shown in the following

table , at most 5.234% out of the whole data will not affect the regression.

Outliers	Deciduous with Shrubland Trees	Deciduous Woodland	Open to Very Open Trees	Rainfed Herbaceous Crop
Grid (%)	348 (4.831%)	344 (3.985%)	269 (5.418%)	324 (5.234%)
Data (%)	16254 (3.2%)	16084 (2.672%)	14334 (3.982%)	18448 (4.068%)

The polynomial regression is applied to fit the above trimmed coefficients. The employed function is as follows for

$$\hat{a}_j(x,y) = c_0 + c_1 x + c_2 x^2 + d_1 y + d_2 y^2 + e_1 xy$$

By the ordinary least square procedure, the new set of coefficients $(c_0, c_1, c_2, d_1, d_2, e)$ are obtained for the previous four land cover types and are listed in Tables 4.5 to Table 4.8.

Employ the table coefficients for $\hat{a}_j(x,y)$ in (4.3.3), and further plug into equation (4.2.3), the LAI estimates are obtained based on the parametric regression spatially and spline regression temporally. There is negligible amount of unreasonable estimates

Estimate	Deciduous with Shrubland Trees	Deciduous Woodland	Open to Very Open Trees	Rainfed Herbaceous Crop
Less than 0	699 (0.142%)	369 (0.062%)	122 (0.035%)	110 (0.025%)

We replace all the negatives with 0, then the linear correlation coefficients between the final estimates and the raw LAI is provided in the following table.

Deciduous with Shrubland Trees	Deciduous Woodland	Open to Very Open Trees	Rainfed Herbaceous Crop
0.62814	0.57409	0.59555	0.53253

4.3.4 Implications

Figures 4.17 shows LAI values at 8 May 2000 for three combinations of land cover and LAI phenology, along with a MODIS image for comparison. LAI exerts a strong influence on the radiation budget at the surface, and when incorporated into models it can improve accuracy, see Lu and Shuttleworth (2002). Figure 4.17 (a) shows grid-cell-averaged LAI for OGE with LAI values assigned from LEAF-2. Figure 4.17 (b) shows CLIPCover crosswalked with the same vegetation classes in the LEAF-2 lookup table. Figure 4.17 (c) shows the LAI distribution using the CLIPCover classes, but with LAI values assigned based on the MODIS-derived spline functions. Here, time class-specific curves of LAI (splines) have been estimated for different regions to generate look-up tables for LAI more appropriate for these regions than LEAF-2. Figure 4.17 (d) shows the raw MODIS LAI for the date selected. Since RAMS treats LAI slightly differently from MODIS, the example shown here has been corrected for this discrepancy. The profound difference in LAI from Figure 4.17 (a) to (d) at the Equator shows that the LEAF-2 function is essentially treating the semidesert of eastern Kenya as having high LAI with no variation. These successive improvements have helped to give a more precise surface parameterization while keeping the flexibility needed to accommodate projected land use change.

4.4 Conclusions

In general, we found that this approach resulted in a large improvement over the generic cover parameters in RAMS in the representation of seasonal variability of LAI. This improvement is expected to significantly improve the seasonal precipitation pattern in RAMS scenarios. For certain land cover, the phenological information varies spatially. At the same grid point the phenologies changes for different land covers.

Sensitivity needs to quantify spatially and by type. For better estimation and prediction, the time dependence and the spatial correlation should be considered. There are more influence affects like the elevation and the topology distance to other geographic features such as Ocean, lakes, Mountain and human settlement etc.

Tables

noise level	sample size n	confidence	estimated bands	oracle bands
0.2	100	0.99	0.476 (0.458)	0.606 (0.606)
		0.95	0.256 (0.246)	0.438 (0.436)
	200	0.99	0.704 (0.708)	0.802 (0.802)
		0.95	0.454 (0.456)	0.532 (0.532)
	500	0.99	0.826 (0.834)	0.832 (0.832)
		0.95	0.462 (0.456)	0.468 (0.468)
0.5	100	0.99	0.618 (0.618)	0.618 (0.618)
		0.95	0.504 (0.504)	0.504 (0.504)
	200	0.99	0.860 (0.860)	0.860 (0.860)
		0.95	0.716 (0.716)	0.716 (0.716)
	500	0.99	0.932 (0.932)	0.932 (0.932)
		0.95	0.802 (0.802)	0.802 (0.802)

Table 4.1. Coverage probabilities of constant spline bands.

noise level	sample size n	confidence level 0.99	confidence level 0.95
0.2	100	0.900 (0.896)	0.816 (0.814)
	200	0.956 (0.962)	0.902 (0.904)
	500	0.990 (0.988)	0.954 (0.958)
0.5	100	0.904 (0.904)	0.822 (0.814)
	200	0.956 (0.960)	0.900 (0.902)
	500	0.990 (0.988)	0.956 (0.960)

Table 4.2. Coverage probabilities of linear spline bands.

		eff_1		eff_3	
d	n	$\rho = 0$	$\rho = 0.3$	$\rho = 0$	$\rho = 0.3$
4	100	1.015 (0.287)	0.958 (0.320)	1.000 (0.268)	0.926 (0.266)
	200	0.992 (0.126)	0.974 (0.164)	1.001 (0.133)	0.973 (0.153)
	500	0.993 (0.060)	0.990 (0.083)	0.995 (0.058)	0.990 (0.083)
	1000	0.998 (0.0416)	1.000 (0.060)	0.998 (0.042)	0.997 (0.057)
10	100	0.899 (0.648)	0.666 (0.597)	0.952 (0.832)	0.641 (0.552)
	200	1.026 (0.434)	0.818 (0.361)	1.045 (0.479)	0.826 (0.395)
	500	1.012 (0.145)	0.977 (0.171)	1.002 (0.138)	0.970 (0.182)
	1000	0.999 (0.078)	0.986 (0.104)	0.989 (0.082)	0.988 (0.105)

Table 4.3. Relative efficiency of $\tilde{m}_{s,\alpha}$ against $\hat{m}_{s,\alpha}$ for $d = 4, 10$.

ρ	n	eff_1	eff_{10}	eff_{19}	eff_{50}
0	500	1.030 (0.830)	0.995 (0.778)	0.737 (0.567)	0.861 (0.648)
	1000	1.130 (0.756)	1.015 (0.523)	1.055 (0.467)	1.056 (0.509)
	1500	1.022 (0.318)	1.029 (0.248)	1.107 (0.302)	0.957 (0.205)
	2000	1.029 (0.197)	1.016 (0.194)	1.045 (0.188)	1.061 (0.223)
0.3	500	0.379 (0.297)	0.410 (0.408)	0.352 (0.296)	0.444 (0.721)
	1000	0.618 (0.269)	0.604 (0.290)	0.623 (0.268)	0.607 (0.311)
	1500	0.864 (0.345)	0.843 (0.280)	0.806 (0.254)	0.831 (0.250)
	2000	0.915 (0.247)	0.872 (0.194)	0.917 (0.221)	0.907 (0.221)

Table 4.4. Relative efficiency of $\tilde{m}_{s,\alpha}$ against $\hat{m}_{s,\alpha}$ for $d = 50$.

	c_0	c_1	c_2	d_1	d_2	e_1
$\hat{a}_0$	13.86733	0.189258	−0.01538	−0.61737	0.007717	−0.00977
$\hat{a}_1$	0.501879	0.009621	0.000144	−0.02756	0.00039	−0.00021
$\hat{a}_2$	−0.63643	−0.00351	$-1.1E-05$	0.033409	−0.00044	0.000168
$\hat{a}_3$	0.425755	−0.00089	$-7.2E-05$	−0.02161	0.000266	$-1.6E-05$
$\hat{a}_4$	0.230287	−0.00537	−0.00025	−0.01455	0.000229	0.000037
$\hat{a}_5$	−0.38993	0.001671	$-7.7E-05$	0.022872	−0.00033	$-6.9E-05$
$\hat{a}_6$	−0.17788	$-6.9E-05$	0.000194	0.010029	−0.00015	0.000025
$\hat{a}_7$	0.560264	0.007802	0.000233	−0.03082	0.000431	−0.00013
$\hat{a}_8$	−0.65163	−0.00305	$-3.8E-05$	0.034248	−0.00045	0.000146
$\hat{a}_9$	0.430822	−0.00104	$-6.2E-05$	−0.02189	0.00027	$-8E-06$
$\hat{a}_10$	0.222698	−0.00512	−0.00026	−0.01413	0.000224	0.000024
$\hat{a}_11$	−0.36273	0.000745	$-1.7E-05$	0.021394	−0.00031	$-2.3E-05$
$\hat{a}_12$	−0.2125	−0.00392	0.000027	0.012197	−0.00018	0.00008

Table 4.5. Coefficients table for Deciduous Shrubland with Sparse Trees.

	c_0	c_1	c_2	d_1	d_2	e_1
$\hat{a}_0$	15.60422	0.258968	−0.01681	−0.7006	0.008762	−0.01201
$\hat{a}_1$	0.285465	0.010554	0.000218	−0.01437	0.000196	−0.00021
$\hat{a}_2$	−0.52319	−0.00485	$-3.6E-05$	0.025799	−0.00032	0.000184
$\hat{a}_3$	0.423792	−0.00264	−0.00014	−0.02168	0.000263	0.000014
$\hat{a}_4$	0.165587	−0.00551	−0.00016	−0.01029	0.000164	0.000077
$\hat{a}_5$	−0.44096	0.003254	−0.00018	0.025446	−0.00036	−0.00014
$\hat{a}_6$	0.062666	0.000927	0.000261	−0.00331	0.000032	0.00001
$\hat{a}_7$	0.322273	0.008389	0.000262	−0.01656	0.000226	−0.00012
$\hat{a}_8$	−0.53571	−0.00429	$-4.9E-05$	0.026532	−0.00033	0.000161
$\hat{a}_9$	0.428083	−0.00285	−0.00013	−0.02193	0.000267	0.000022
$\hat{a}_{10}$	0.157467	−0.0052	−0.00017	−0.00982	0.000158	0.000064
$\hat{a}_{11}$	−0.41156	0.002126	−0.00014	0.023738	−0.00034	$-9.7E-05$
$\hat{a}_{12}$	−0.07965	−0.00318	$-1E-06$	0.004623	$-7.5E-05$	0.00005

Table 4.6. Coefficients table for Deciduous Woodland.

	c_0	c_1	c_2	d_1	d_2	e_1
$\hat{a}_0$	21.36797	0.582205	−0.01761	−0.9429	0.011065	−0.01953
$\hat{a}_1$	0.755761	0.026441	0.000046	−0.0398	0.00054	−0.00064
$\hat{a}_2$	−0.40319	−0.00305	−0.00016	0.020365	−0.00027	0.000065
$\hat{a}_3$	−0.52959	−0.02078	$-1.5E-05$	0.030611	−0.00044	0.000568
$\hat{a}_4$	0.583969	0.007367	−0.00017	−0.03334	0.000476	−0.00027
$\hat{a}_5$	−0.20593	−0.00388	−0.00013	0.012463	−0.00018	0.000071
$\hat{a}_6$	−0.4363	−0.00384	0.000293	0.024033	−0.00035	0.000095
$\hat{a}_7$	1.062424	0.023523	0.000225	−0.05847	0.000819	−0.0005
$\hat{a}_8$	−0.49433	−0.00221	−0.00021	0.025909	−0.00035	0.000024
$\hat{a}_9$	−0.49496	−0.02111	0.000005	0.028505	−0.00041	0.000584
$\hat{a}_{10}$	0.529546	0.007892	−0.00021	−0.03003	0.000427	−0.0003
$\hat{a}_{11}$	−0.01689	−0.00576	$-1.4E-05$	0.000928	$-8E-06$	0.000164
$\hat{a}_{12}$	−0.22684	−0.01172	0.000174	0.011508	−0.00016	0.000297

Table 4.7. Coefficients table for Open to Very Open Trees.

	c_0	c_1	c_2	d_1	d_2	e_1
$\hat{a}_0$	27.46197	0.516892	−0.01812	−1.34425	0.017536	−0.01782
$\hat{a}_1$	0.665941	0.016663	0.000098	−0.03529	0.000488	−0.00035
$\hat{a}_2$	−0.30472	0.00122	−0.0002	0.015172	−0.0002	$-7.9E-05$
$\hat{a}_3$	−0.44979	−0.01913	$-2E-06$	0.0253	−0.00035	0.000526
$\hat{a}_4$	0.59182	0.004496	−0.00022	−0.03303	0.000461	−0.0002
$\hat{a}_5$	−0.11557	−0.00089	−0.00021	0.006697	$-9.1E-05$	$-3.5E-05$
$\hat{a}_6$	−0.56834	−0.0029	0.000415	0.031754	−0.00046	0.000102
$\hat{a}_7$	0.902208	0.017373	0.000252	−0.04916	0.000688	−0.0003
$\hat{a}_8$	−0.37557	0.001015	−0.00024	0.019326	−0.00026	$-9.1E-05$
$\hat{a}_9$	−0.4228	−0.01907	0.000015	0.023719	−0.00033	0.000531
$\hat{a}_{10}$	0.54799	0.004403	−0.00024	−0.03046	0.000424	−0.00021
$\hat{a}_{11}$	0.038773	−0.00055	−0.00012	−0.00236	0.000039	$-5E-06$
$\hat{a}_{12}$	−0.28223	−0.00642	0.000175	0.015208	−0.00022	0.000154

Table 4.8. Coefficients table for Rainfed Herbaceous Crop.

Figures

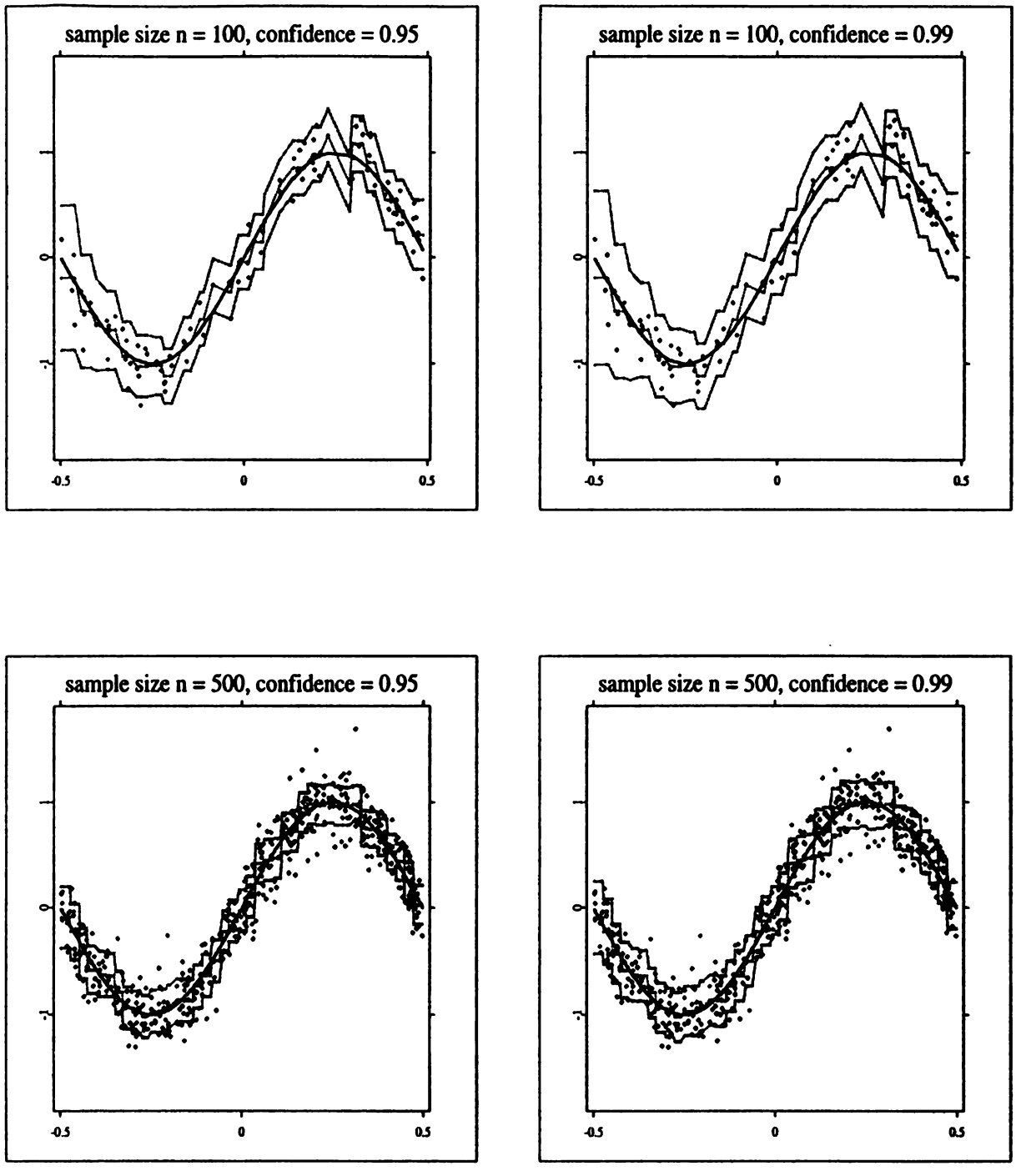

Figure 4.1. Constant spline confidence bands with opt = 1.

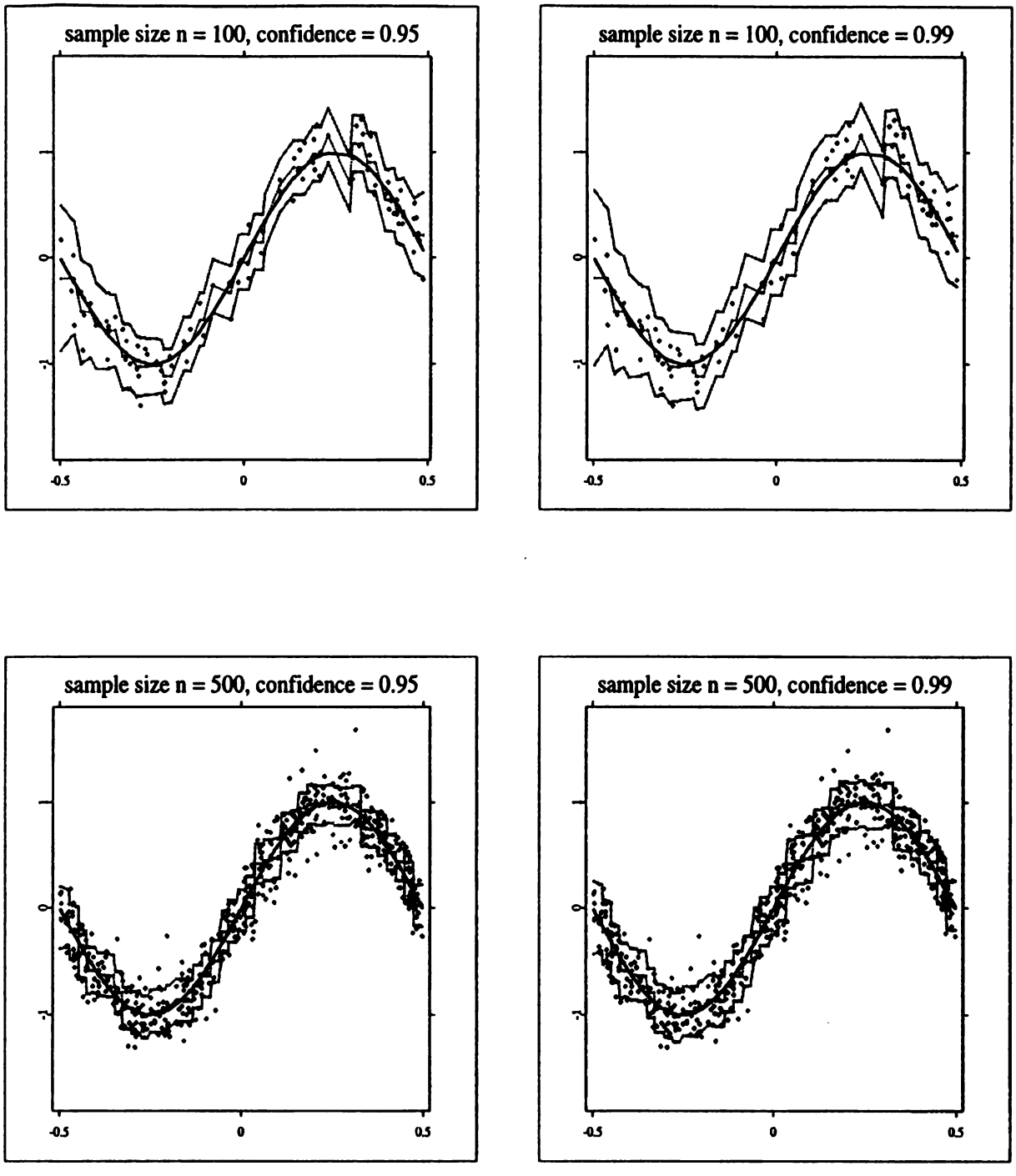

Figure 4.2. Constant spline confidence bands with opt = 2.

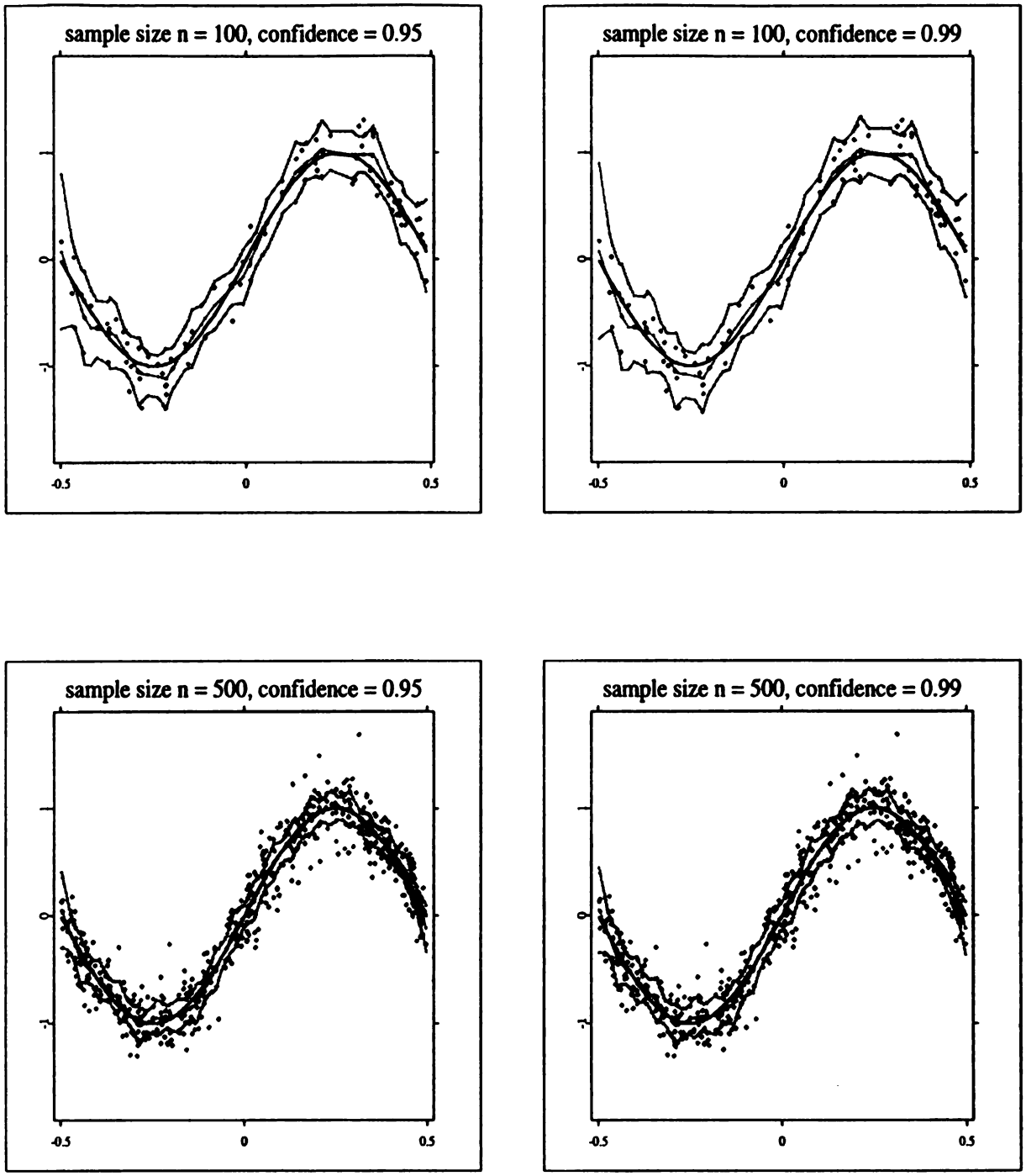

Figure 4.3. Linear spline confidence bands with opt = 1.

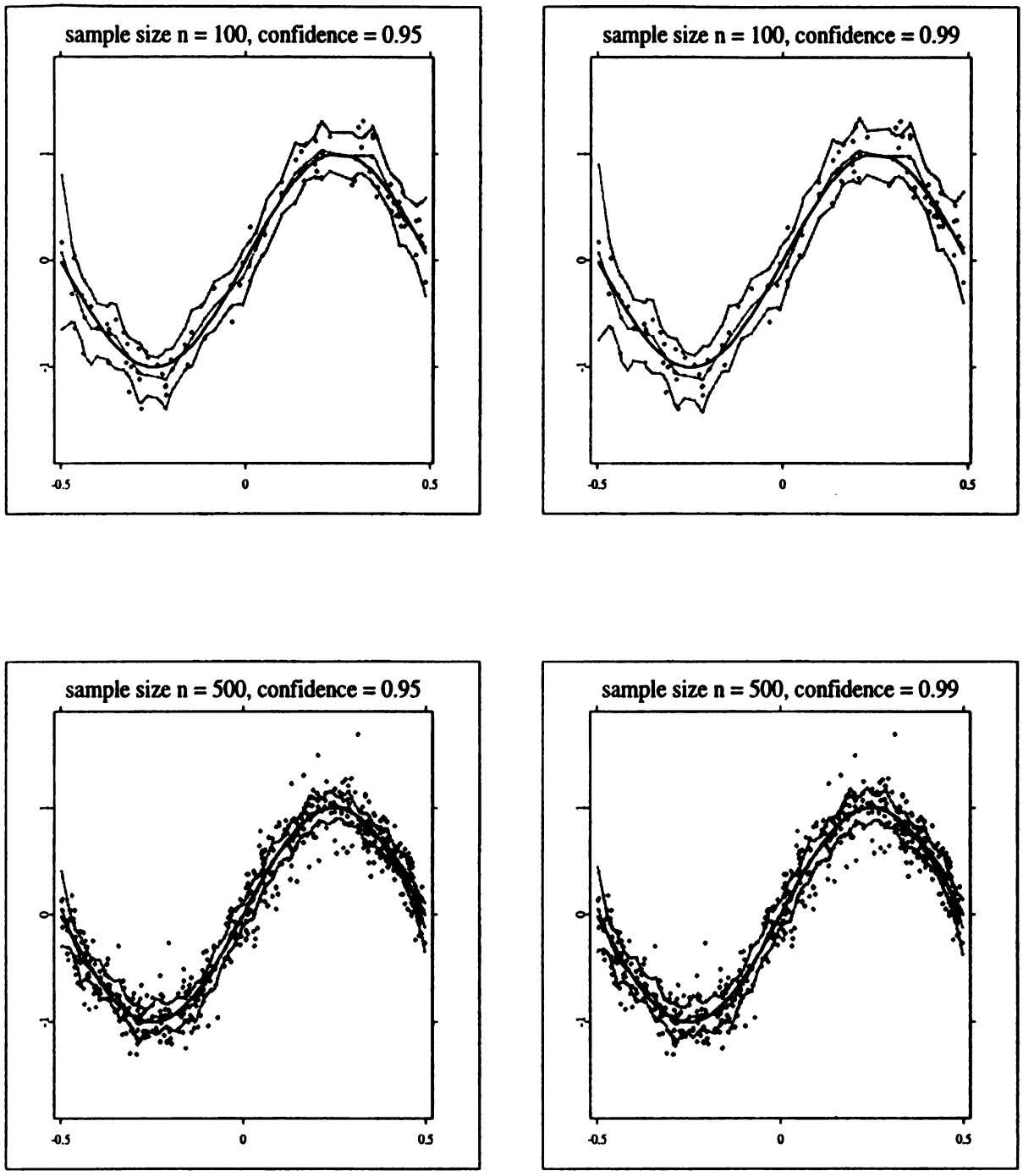

Figure 4.4. Linear spline confidence bands with opt = 2.

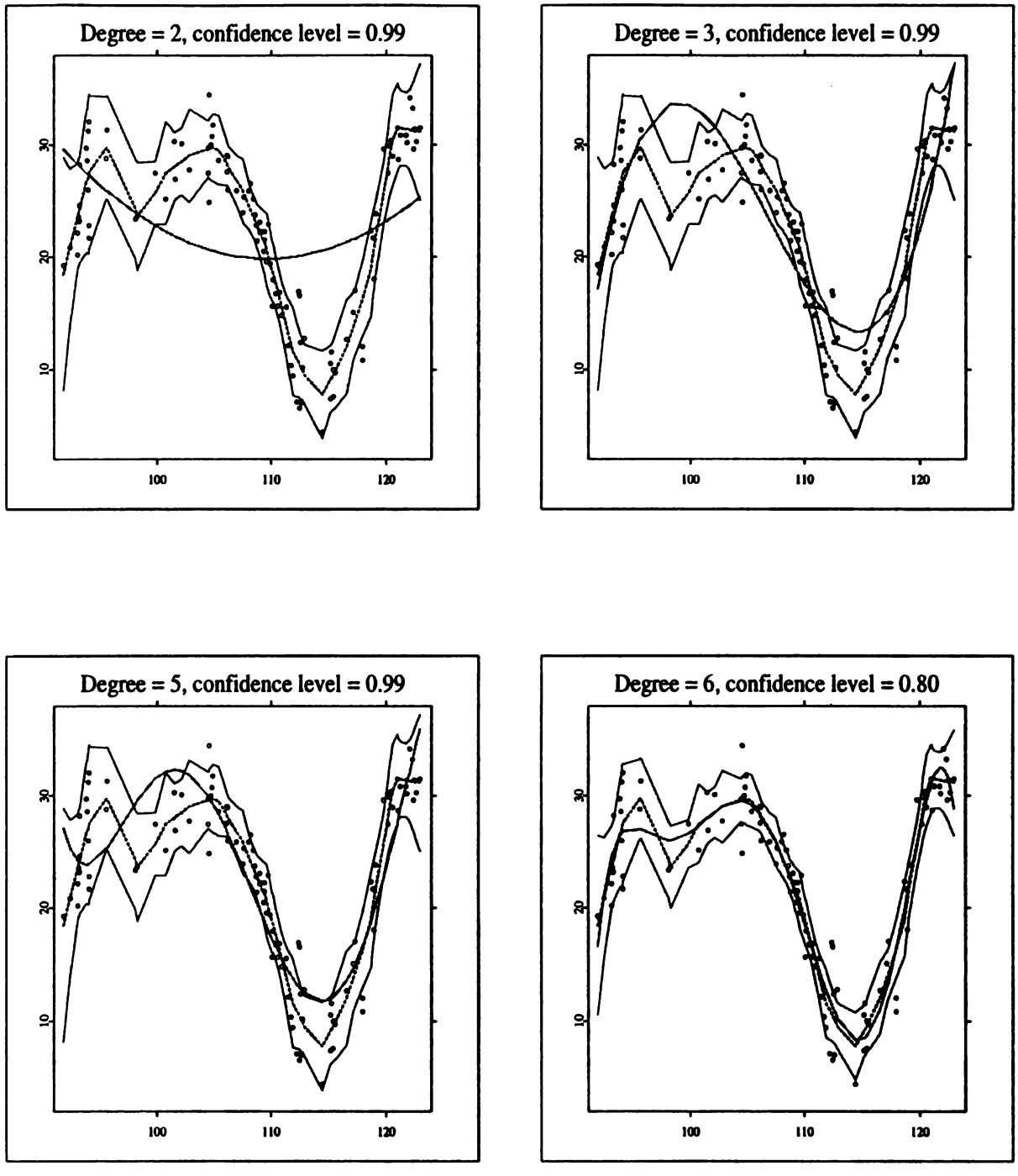

Figure 4.5. Testing $H_0 : m(x) = \sum_{k=1}^{d} a_k x^k, d = 2, 3, 5, 6$ for fossil data.

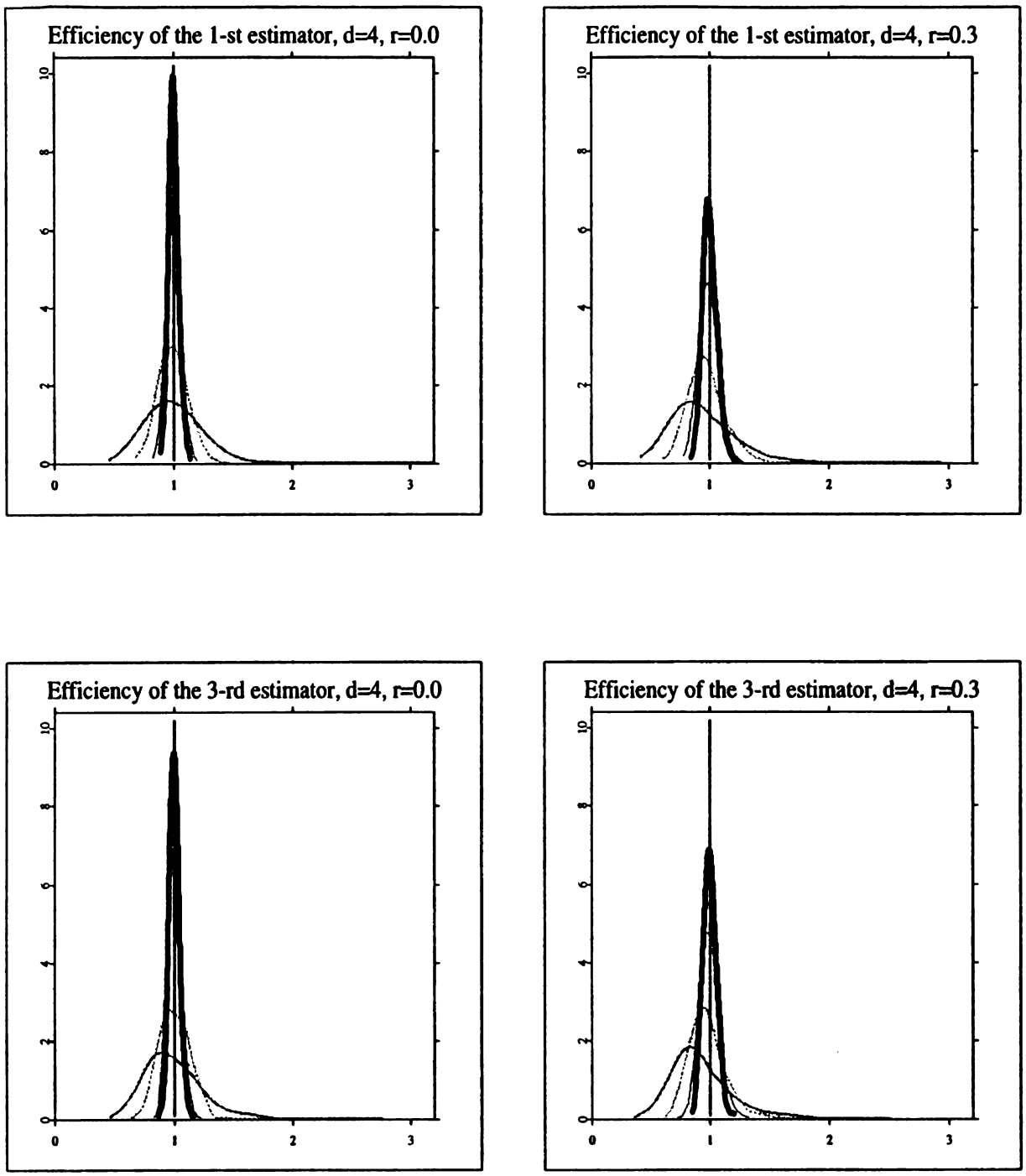

Figure 4.6. Relative efficiency of $\tilde{m}_{s,\alpha}$ against $\hat{m}_{s,\alpha}$, $d = 4$.

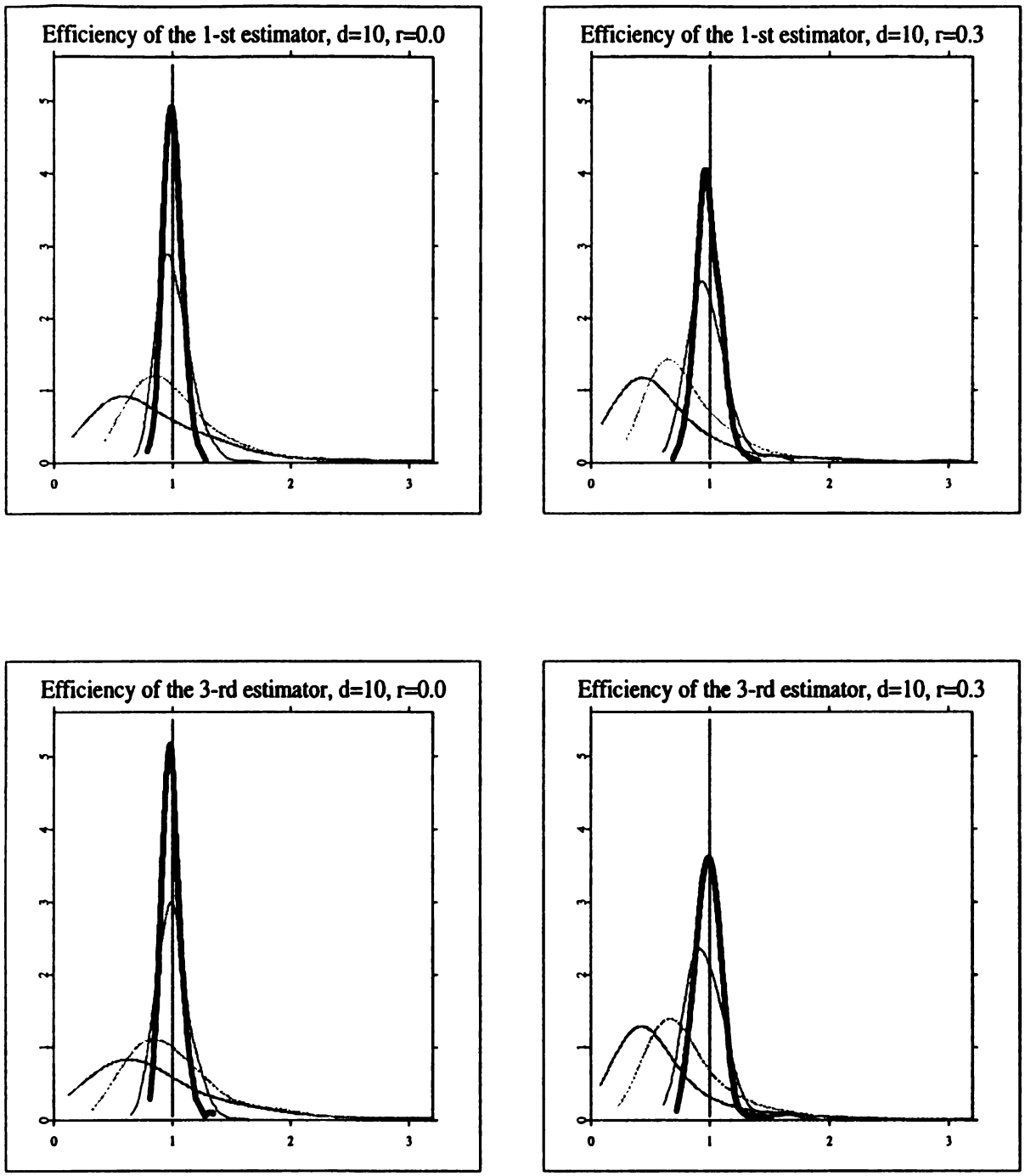

Figure 4.7. Relative efficiency of $\tilde{m}_{s,\alpha}$ against $\hat{m}_{s,\alpha}$, $d = 10$.

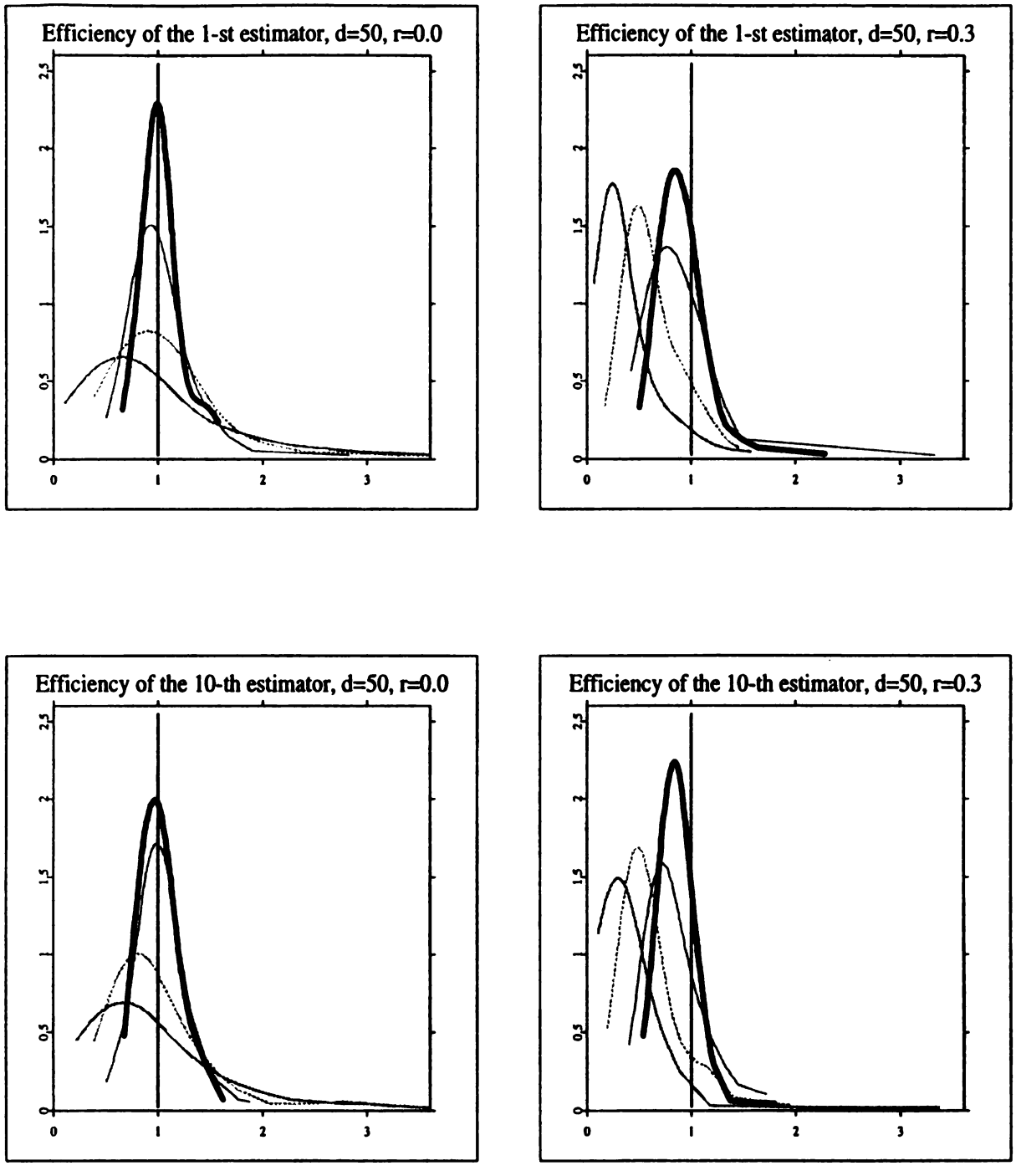

Figure 4.8. Relative efficiency of $\tilde{m}_{s,\alpha}$ against $\hat{m}_{s,\alpha}$, $d = 50, \alpha = 1, 10$.

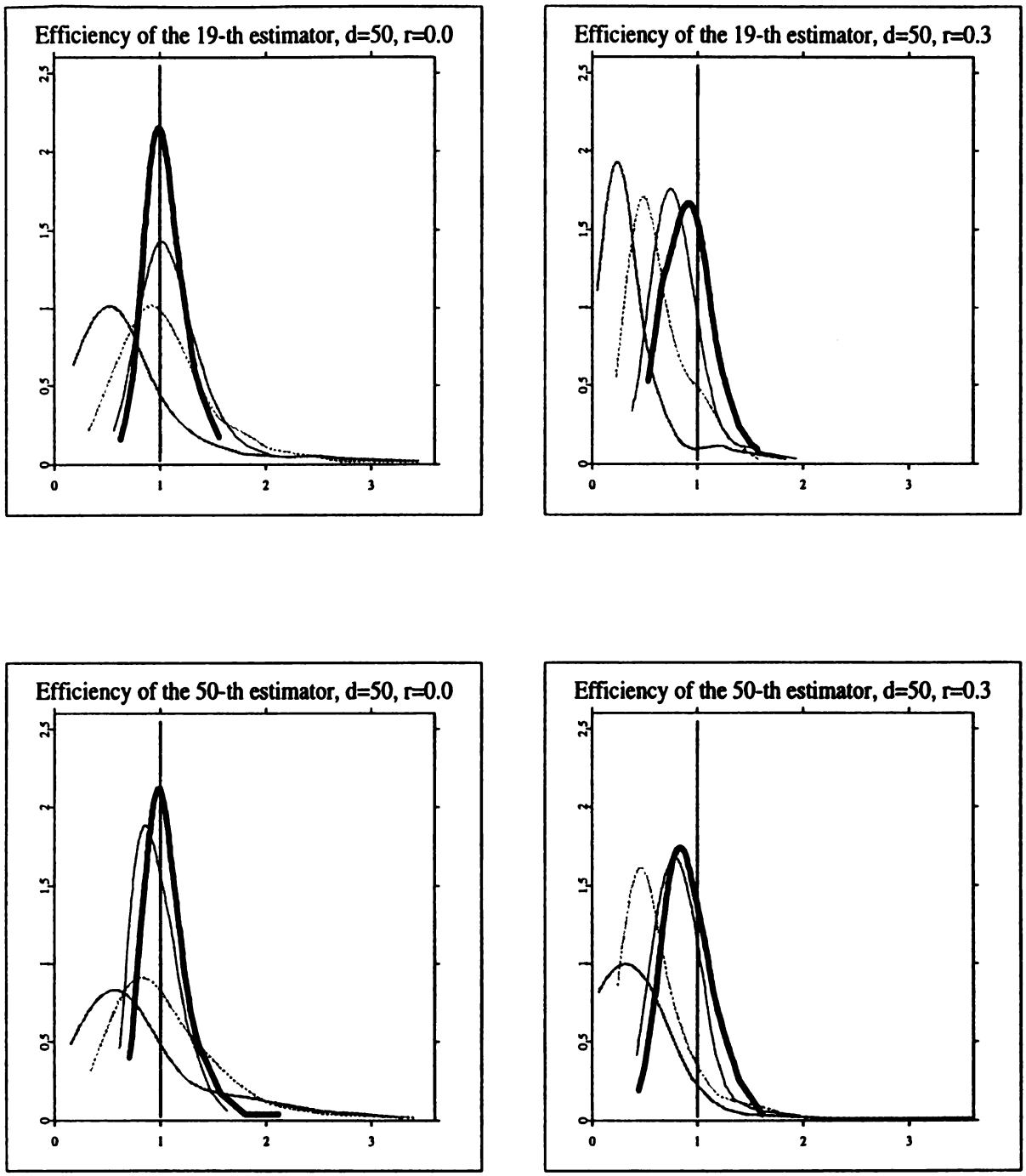

Figure 4.9. Relative efficiency of $\tilde{m}_{s,\alpha}$ against $\hat{m}_{s,\alpha}$, $d = 50, \alpha = 19, 50$.

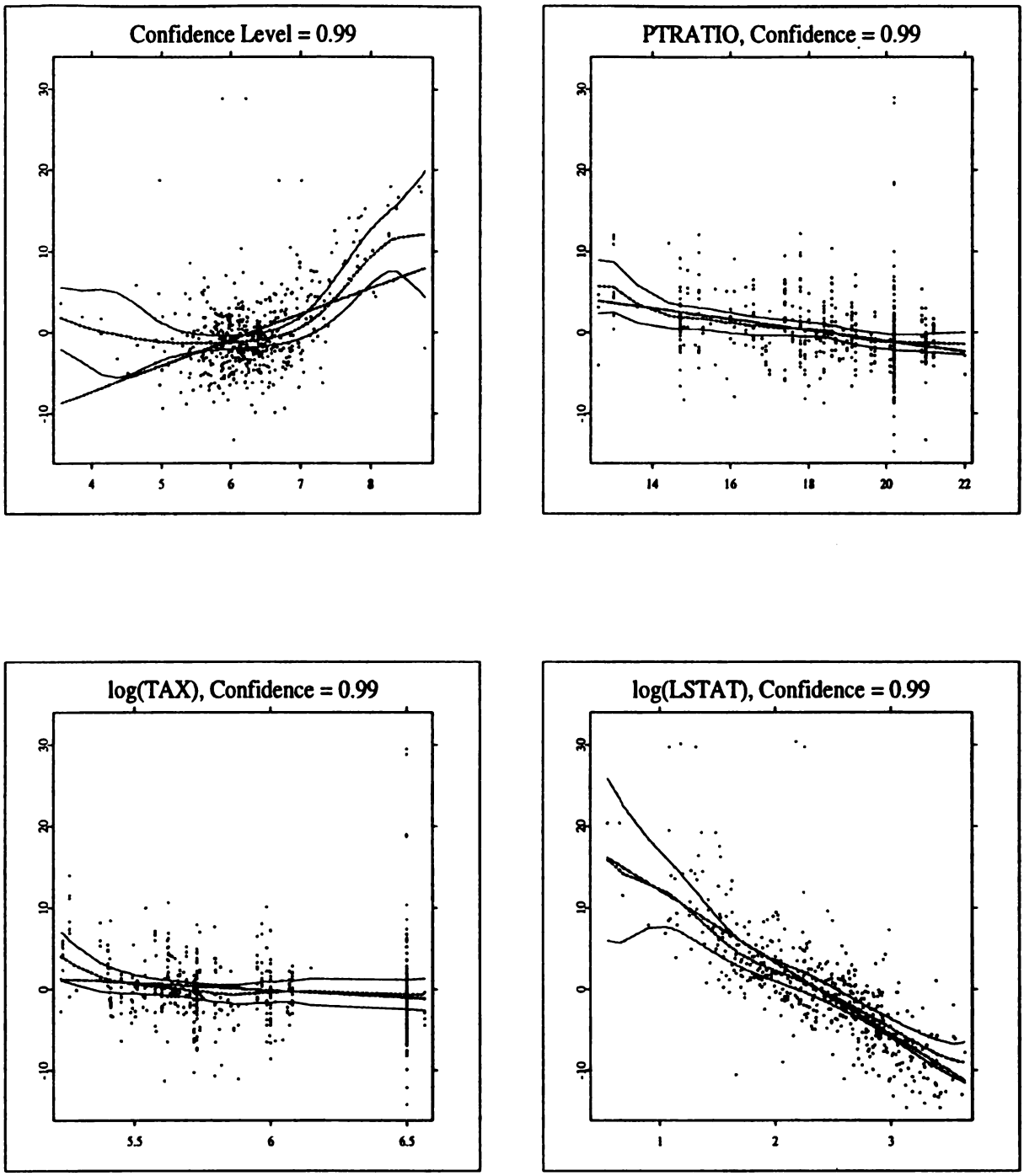

Figure 4.10. Linearity test for the Boston housing data.

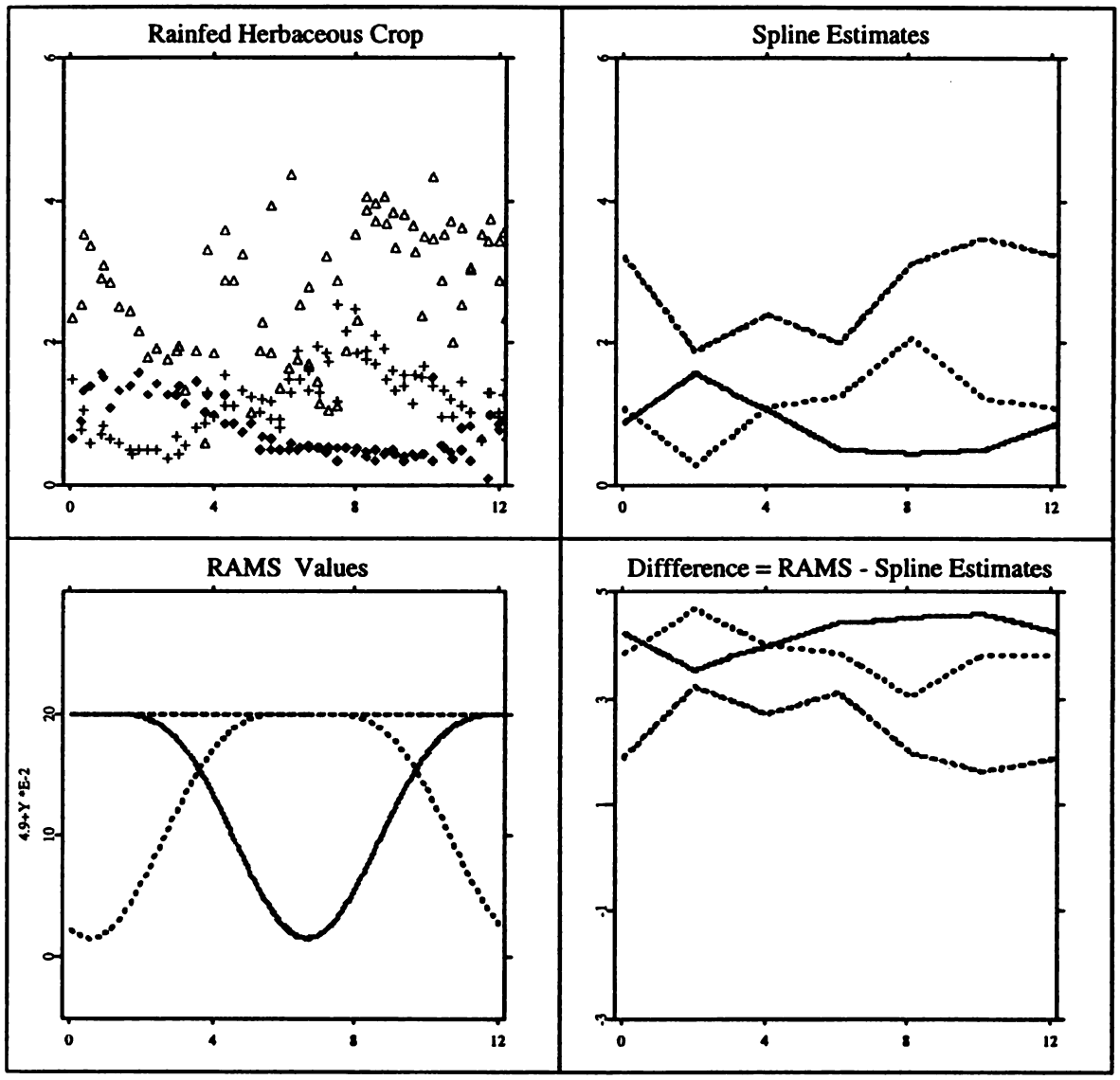

Figure 4.11. LAI trend of rainfed herbaceous crops.

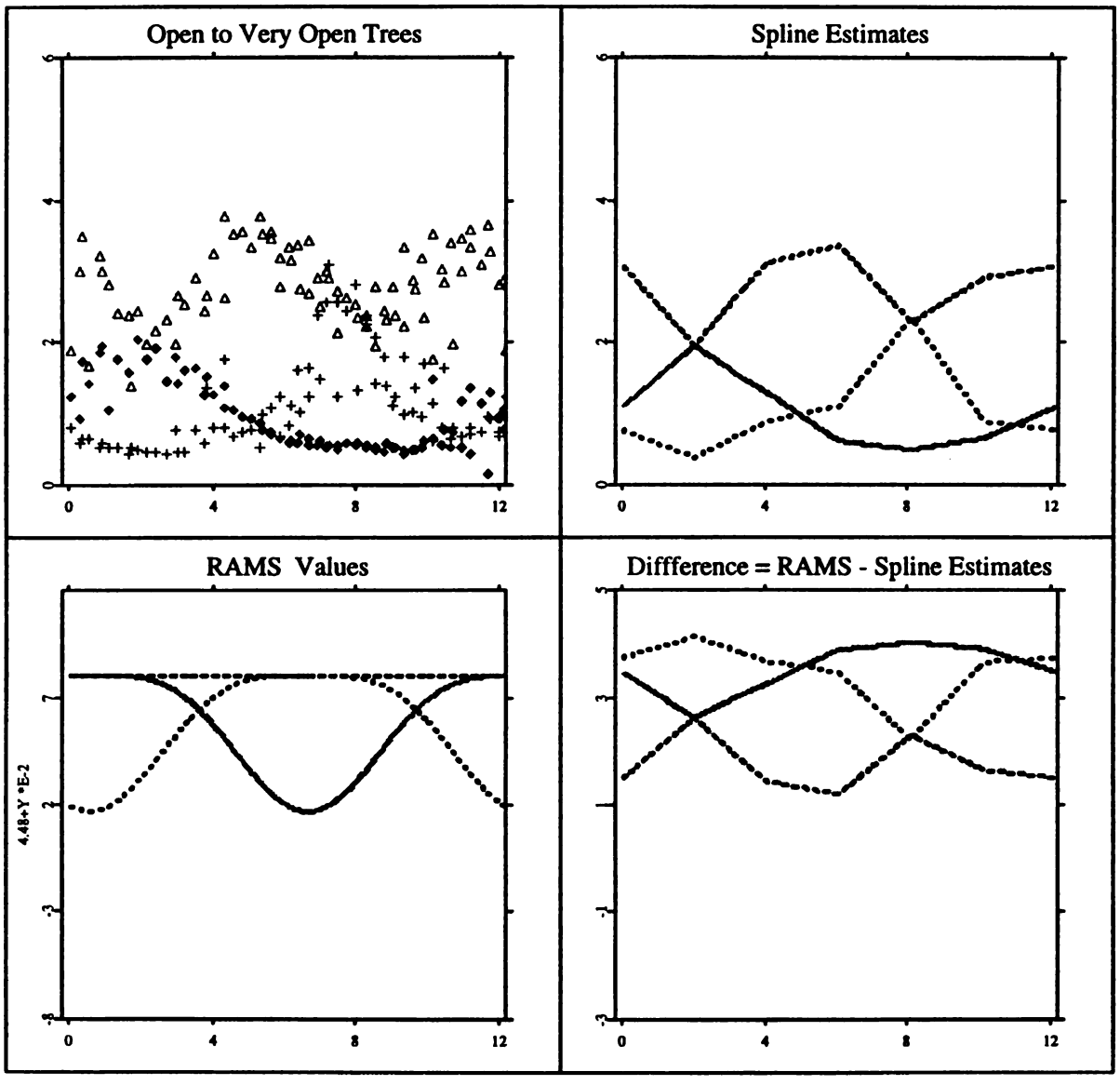

Figure 4.12. LAI trend of open to very open trees.

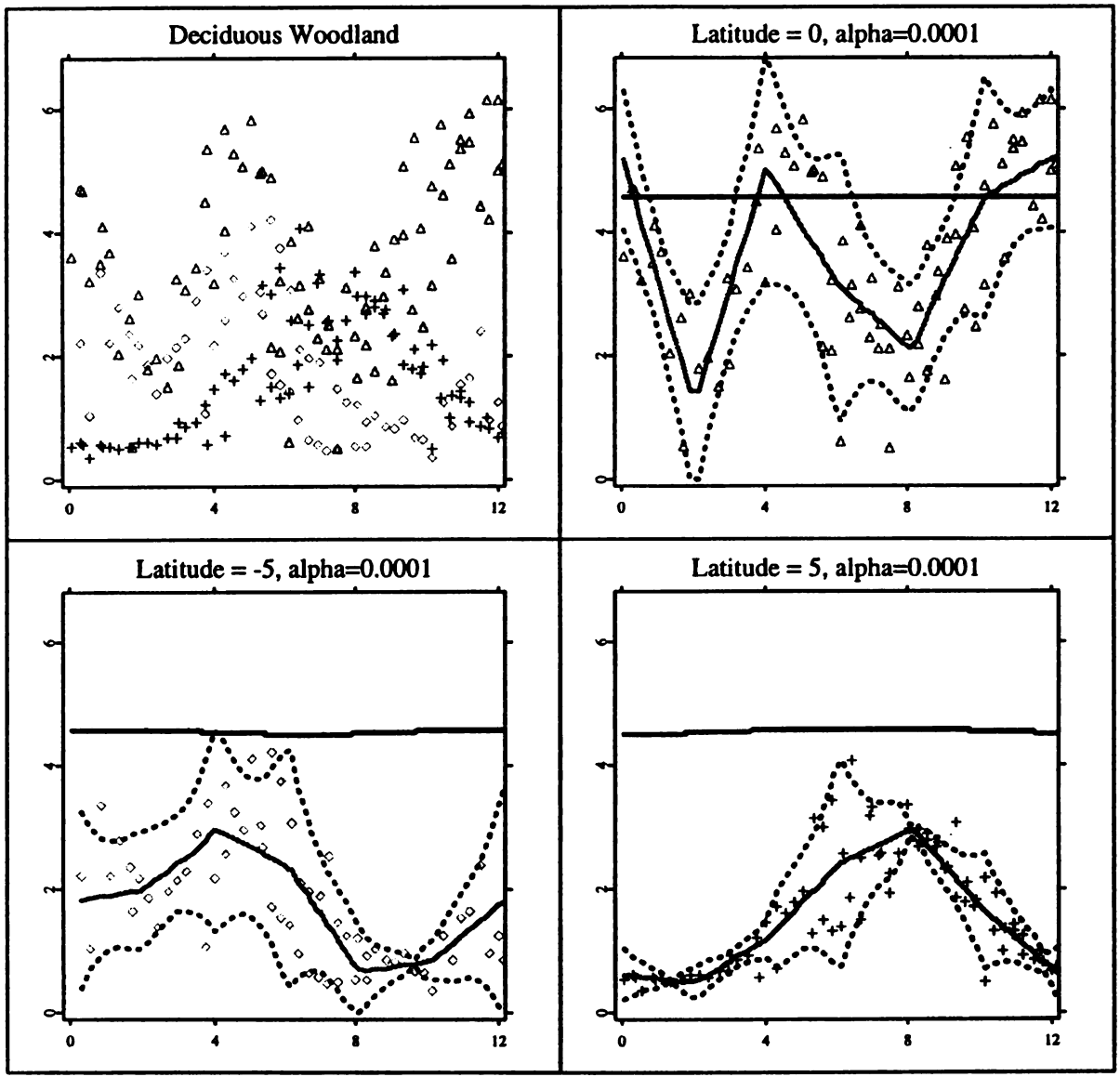

Figure 4.13. Spline confidence bands of LAI of deciduous woodland.

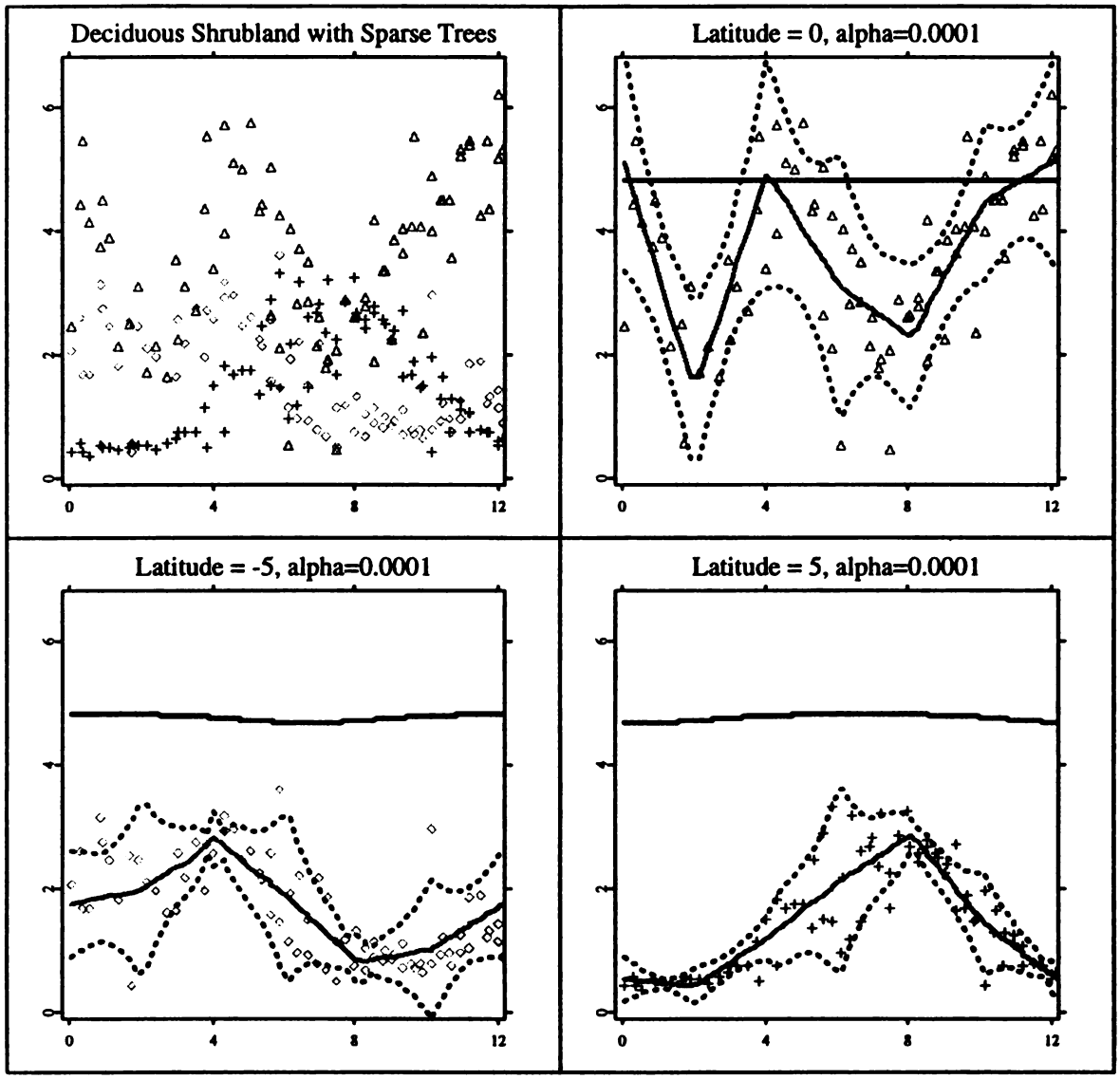

Figure 4.14. Spline confidence bands and RAMS curves of LAI of deciduous shrubland.

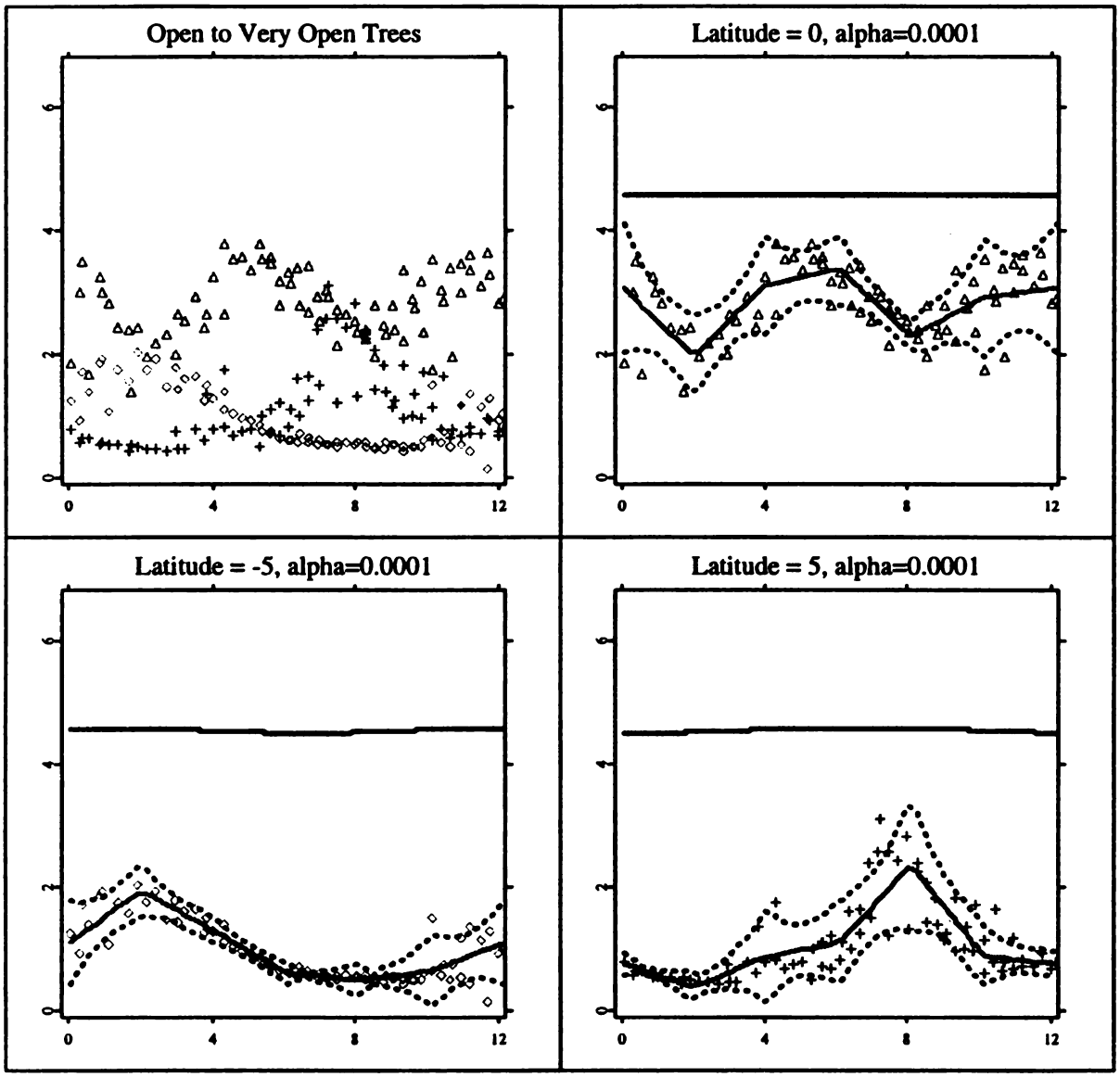

Figure 4.15. Spline confidence bands and RAMS curves of LAI of rainfed herbaceous crop.

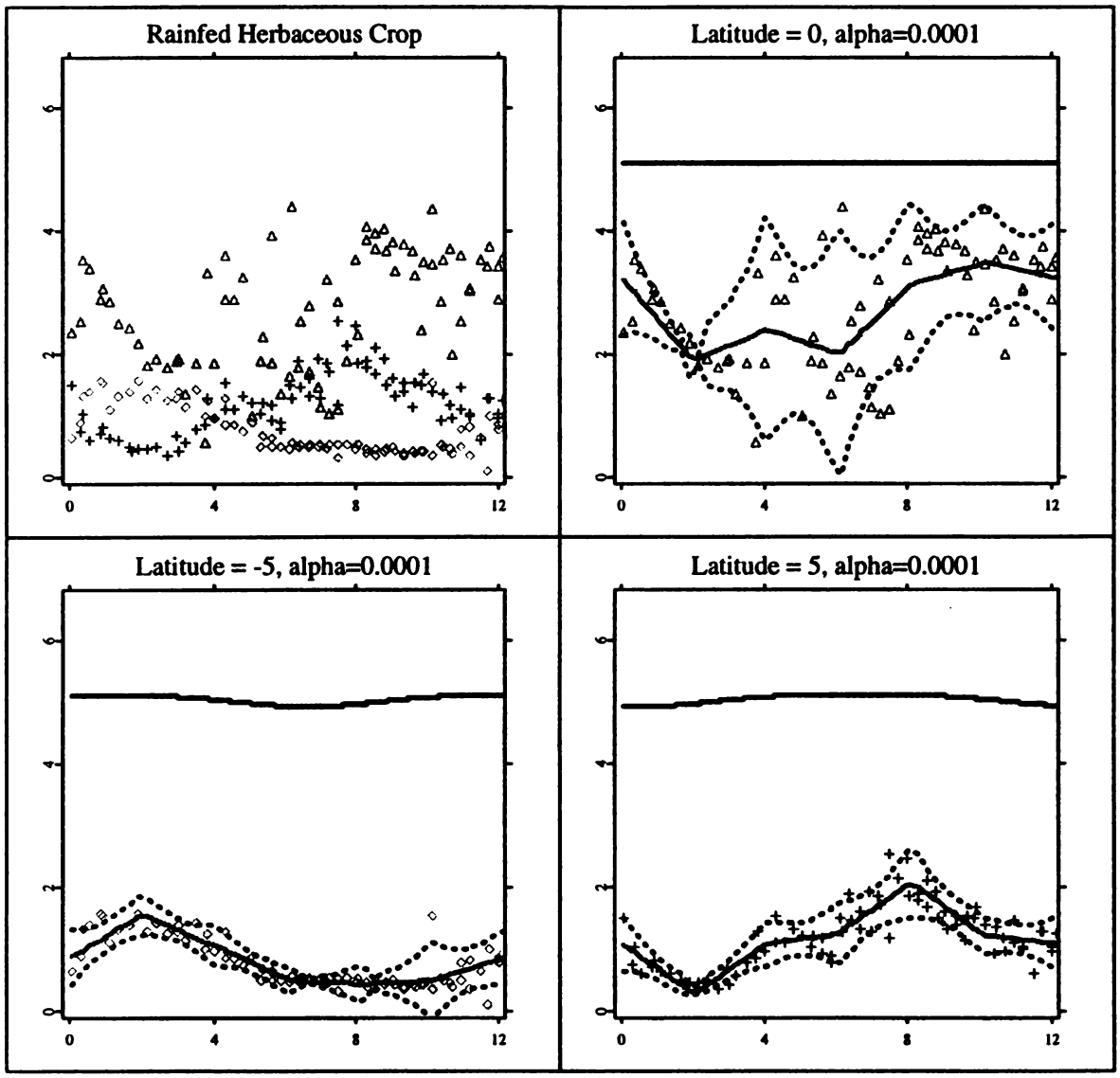

Figure 4.16. Spline confidence bands and RAMS curves of LAI of open to very open trees.

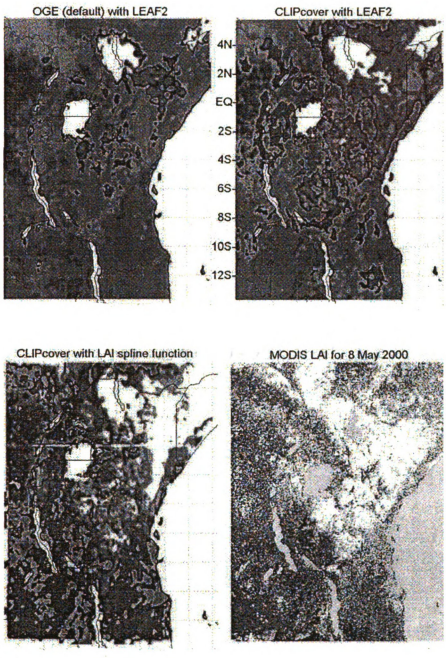

Figure 4.17. Improved representation of land surface in RAMS.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Africover(2002). Africover- Eastern Africa Module. Land cover mapping based on satellite remote sensing. Food and Agriculture Organization of the United Nations.

[2] Andrews, D. and Whang, Y.(1990). Additive interactive regression models: circumvention of the curse of the dimensionality. *Economic Theory.* **6** ,466-479.

[3] Bickel, P. J. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. *The Annals of Statistics.* **1** 1071–1095.

[4] Bralower, T.J., Fullagar, P.D., et al (1997). Mid-cretaceous strontium-isotope stratigraphy of deep-sea sections. *Geological Society of America Bulletin.* **109**, 1421-1442.

[5] Breiman, L. and Friedman, J.H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association.* **80**, 580-619 .

[6] Chaudhuri, P. and Marron, J.S. (1999). SiZer for exploration of structures in curves. *Journal of the American Statistical Association.* **94** 807-823.

[7] Claeskens, G. and Van Keilegom, I. (2003). Bootstrap confidence bands for regression curves and their derivatives. *The Annals of Statistics.* **31** 1852–1884.

[8] Cotton, W. R., et al. (2003). RAMS 2001: Current status and future directions. *Meteorology and Atmospheric Physics.* **82**, 5-29.

[9] de Boor, C. (2001). *A Practical Guide to Splines.* Springer-Verlag, New York.

[10] DeVore, R. A. and Lorentz, G. G. (1993). *Constructive Approximation.* Springer-Verlag, Berlin.

[11] Fan, J. and Chen, J. (1999), One-step local quasi-likelihood estimation. *Journal of the Royal Statistical Society Series B.* **61**, 927-934

[12] Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications.* Chapman and Hall, London.

[13] Fan, J. Härdle, W. and Mammen, E. (1998). Direct estimation of low-dimensional components in additive models. *The Annals of Statistics.* **26**, 943–971.

[14] Gantmacher, F. R. and Krein, M. G. (1960). *Oszillationsmatrizen, Oszillationskerne und kleine Schwingungen mechanischer Systeme.* Akademie-Verlag, Berlin.

[15] Hall, P. and Titterington, D. M. (1988). On confidence bands in nonparametric density estimation and regression. *Journal of Multivariate Analysis.* **27** 228–254.

[16] Härdle, W. (1989). Asymptotic maximal deviation of M-smoothers. *Journal of Multivariate Analysis.* **29** 163–179.

[17] Härdle, W. (1990). *Applied Nonparametric Regression.* Cambridge University Press, Cambridge.

[18] Härdle, W. , Hlávka, Z. and Klinke, S. (2000). *XploRe Application Guide.* Springer-Verlag, Berlin.

[19] Härdle, W., Huet, S. ,Mammen, E., and Sperlich, S.(2004). Bootstrap inference in semiparametric generalized additive models. *Economic Theory.* **20**, 265-300.

[20] Härdle, W., Marron, J. S. and Yang, L. (1997). Discussion of "Polynomial splines and their tensor products in extended linear modeling" by Stone et. al. *The Annals of Statistics.* **25** 1443-1450.

[21] Härdle, W., Sperlich, S. and Spokoiny, V. (2001) Structural tests in additive regression. *Journal of the American Statistical Association.* **96**, 1333-1347.

[22] Harrison, D. and Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for cleaning air. *Journal of Economics and Management.* **5**, 81-102.

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

[24] Huang, J. Z. (1998). Projection estimation in multiple regression with application to functional ANOVA models. *The Annals of Statistics.* **26**, 242–272.

[25] Huang, J. Z. (2003). Local asymptotics for polynomial spline regression. *The Annals of Statistics* . **31**,1600-1635.

[26] Huang, J. Z. and Yang, L. (2004). Identification of nonlinear additive autoregression models. *Journal of the Royal Statistical Society Series B.* **66**, 463-477.

[27] Johnson, R. A. and Wichern, D. W.(1992). *Applied Multivariate Statistical Analysis.* Prentice-Hall, New Jersey.

[28] Kim, W., Linton, O. B., and Hengartner, N.(1999). A Computationally efficient oracle estimator for additive nonparametric regression with bootstrap confidence intervals. *Journal of Computational and Graphical Statistics.* **8**, 278-297.

[29] Knyazikhin, Y., J. Glassy, J. L., Privette, Y.Tian, A. Lotsch, Y. Zhang, Y. Wang, J. T. Morisette, P. otava, R.B. Myneni, R. R. Nemani, S. W. Running,(1999) MODIS Leaf Area Index (LAI) and Fraction of Photosynthetically Active Radiation Absorbed by Vegetation (FPAR) Product (MOD15) *Algorithm Theoretical Basis Document.*

[30] Leadbetter, M. R., Lindgren, G.andRootzén, H.(1983). *Extremes and Related Properties of Random Sequences and Processes.* Springer-Verlag, New York.

[31] Linton, O. B. and Nielsen, J. P.(1995). Estimating structured nonparametric regression models by the kernel method. *Biometrika.* **82**, 93–101.

[32] Linton, O. B. and Härdle, W.(1996). Estimating additive regression models with known links. *Biometrika.* **83**, 529–540.

[33] Linton, O. B.(1997). Efficient estimation of additive nonparametric regression models. *Biometrika.* **84**, 469–473.

[34] Mack, Y. P. and Silverman, B. W.(1982). Weak and strong uniform consistency of kernel regression estimates. *Z. Wahrscheinlichkeitstheorie verm Gebiete.* **61** 405-415.

[35] Mammen, E., Linton, O. and Nielsen, J.(1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *The Annals of Statistics.* **27**, 1443-1490.

[36] Mayaux, P. Bartholome, E., Frtiz, S. and Belward, A. (2004). A new land-cover map of Africa for the year 2000. *Journal of Biogeography.* **31**,861-877.

[37] Nielsen, J. P. and Sperlich, S.(2005), Smooth backfitting in practice, *Journal of the Royal Statistical Society B.* **67**, 43-61.

[38] Olson, J. M., Alagarswamy, G. , Andresen, J., Campbell, D.J., Ge, J., Huebner, M., Brent Lofgren, B., Lusch, D.P., Moore, N., Pijanowski, B.C., Qi, J., Torbick, N., Wang, J. and Yang, L. (2006) Integrating diverse methods to understand climate-land interactions at multiple spatial and temporal scales, *GeoForum.*

[39] Opsomer, J. D.(2000). Asymptotic properties of backfitting estimators. *Journal of Multivariate Analysis.* **73**, 166-179

[40] Opsomer, J. D. and Ruppert, D.(1997). Fitting a bivariate additive model by local polynomial regression. *The Annals of Statistics.* **25** 186–211.

[41] Opsomer, J. D.andRuppert, D.(1998). A Fully automated bandwidth selection method for fitting additive models. *Journal of the American Statistical Association.* **93**, 605–619.

[42] Pitman, A. (2003) The evolution of, and revolution in, land surface schemes designed for climate models. *International Journal of Climatology.* **23**, 479-510.

[43] Rosenblatt, M.(1976). On the maximal deviation of k-dimensional density estimates. *The Annals of Probability.* **41**, 009-1015.

[44] Ruppert, D., Wand, M.P. and Carroll, R.J.(2003) *Semiparametric Regression.* Cambridge University Press, Cambridge; New York .

[45] Silverman, B. W.(1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

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

[47] Stone, C. J.(1985). Additive regression and other nonparametric models. *The Annals of Statistics.* **13**, 689–705.

[48] Stone, C. J.(1994). The use of polynomial splines and their tensor products in multivariate function estimation. *The Annals of Statistics.* **22**, 118–184.

[49] Tjøstheim, D. and Auestad, B.(1994). Nonparametric identification of nonlinear time series: projections. *Journal of the American Statistical Association.* **89**, 1398-1409.

[50] Torbick, N., Lusch, D., Olson, J., Qi, J., Ge, J.. (2005a) An Assessment of Africover and GLC2000 using general agreement and airborne videography. *International Journal of Remote Sensing.* (submitted).

[51] Torbick, N., Qi, J., Lusch, D., Olson, J., Moore, N., Ge, J. (2005b) Developing land use land cover parameterization for climate-land modelling in East Africa. (in progress).

[52] Tusnády, G.(1977). A remark on the approximation of the sample df in the multidimensional case. *Periodica Mathematica Hungarica.* **8**, 53-55.

[53] Walko, R.L., Band, L.E., Baron, J., Kittel, T.G.F., Lammers, R., Lee, T.J., Ojima, D., Pielke Sr., R.A., Taylor, C., Tague, C., Tremback, C.J., Vidale, P.L., (2000). Coupled atmosphere - biophysics - hydrology models for environment modeling. *Journal of Applied Meteorology.* **39**, 931- 944.

[54] Wang, J. and Yang, L.(2006). Polynomial spline confidence bands for regression curves. *The Annals of Statistics.* **tentatively accepted.**

[55] Xia, Y.(1998). Bias-corrected confidence bands in nonparametric regression. *Journal of the Royal Statistical Society Series B.* **60**, 797–811.

[56] Xue, L. and Yang, L.(2006). Estimation of semiparametric additive coefficient model. *Journal of Statistical Planning and Inference.* **136**, 2506-2534.

[57] Yang, L., Härdle, W. and Nielsen, J. P.(1999). Nonparametric autoregression with multiplicative volatility and additive mean. *Journal of Time Series Analysis.* **20**, 579-604.

[58] Yang, L., Sperlich, S.and Härdle, W.(2003). Derivative estimation and testing in generalized additive models. *Journal of Statistical Planning and Inference.* **115**, 521-542.

[59] Zhang, F.(1999). *Matrix Theory. Basic Results and Techniques.* Springer-Verlag, New York.

[60] Zhou, S., Shen, X. and Wolfe, D. A.(1998). Local asymptotics of regression splines and confidence regions. *The Annals of Statistics.* **26**, 1760-1782.