This is to certify that the
dissertation entitled

Polynomial Spline Smoothing for Nonlinear Time Series

presented by

Li Wang

has been accepted towards fulfillment
of the requirements for the

Ph.D. degree in Statistics and Probability

Major Professor's Signature

4/30/07

Date

POLYNOMIAL SPLINE SMOOTHING FOR NONLINEAR TIME SERIES

By

Li Wang

A DISSERTATION

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

DOCTOR OF PHILOSOPHY

Department of Probability and Statistics

2007

ABSTRACT

POLYNOMIAL SPLINE SMOOTHING FOR NONLINEAR TIME SERIES

By

Li Wang

Nonlinear time series analysis has gained much attention in recent years due primarily to the fact that linear time series models have encountered various limitations in real applications and the development in nonparametric regression has established a solid foundation for nonlinear time series analysis. In this dissertation, polynomial spline smoothing is studied for nonlinear time series.

For univariate nonlinear time series, uniform confidence bands of a nonparametric prediction function are constructed using the polynomial spline method. As an application, after removing the environmental Kuznets curve trend effects, the impact of the economic intervention on environmental quality change is quantified for the United States and Japan, with different conclusions.

Application of non- and semiparametric regression techniques to high dimensional time series data have been hampered due to the lack of effective tools to address the "curse of dimensionality". There are essentially two approaches to circumvent this difficulty: function approximation and dimension reduction.

For the function approximation approach, the nonlinear additive autoregression (NAAR) model is examined. Under rather weak conditions, spline-backfitted kernel estimators of the component functions are proposed for weakly dependent samples that are both computationally expedient (so it is usable for analyzing very high dimensional time series), and theoretically reliable (so inference can be made on the component functions with confidence).

For the dimension reduction approach, a single-index prediction (SIP) model based on weakly dependent sample is studied. The single-index is identified by the best approximation to the multivariate prediction function of the response variable, regardless of whether or not the prediction function is a genuine single-index function. A polynomial spline estimator is proposed for the single-index prediction coefficients, and is shown to be root-n consistent and asymptotically normal. An iterative optimization routine is used which is sufficiently fast for the user to analyze large data sets of high dimension within seconds. Application of the proposed procedure to the river flow data of Iceland has yielded superior out-of-sample rolling forecasts.

I dedicate this work to my husband Lei Gao and my parents.

ACKNOWLEDGMENTS

I would like to thank many people who have helped me on the path towards this dissertation. First and foremost, I would like to express my gratitude to my advisor, Professor Lijian Yang. I could never have reached the heights or explored the depths without his generous help, unbreakable support and patient guidance. Every single discussion I had with him was valuable in terms of new ideas and renewed scientific excitement. His infectious enthusiasm and unlimited zeal have been major driving forces through my graduate study at MSU and will keep encouraging me in my future research.

I also wish to express my gratitude to my dissertation committee, Professor Dennis Gilliland, Professor Yiming Xiao, Professor Ana María Herren, for sparing their precious time to serve on my committee and giving valuable comments and suggestions.

I must acknowledge Professor Dennis Gilliland and Professor Connie Page for accepting me as one of the consultants at CSTAT, where I have obtained plenty of opportunities to work with students and faculty from a variety of disciplines.

I am grateful to the entire faculty and staff in the Department of Statistics and Probability who have taught me and assisted me during my study at MSU. And special thanks are given to Professor James Stapleton and Professor Vince Melfi for their numerous help, constant support and encouragement.

Thanks to the graduate school and the Department of Statistics who provided me with the Dissertation Completion Fellowship for working on this dissertation. This dissertation is also supported in part by NSF award DMS 0405330.

Last but not least, I would like to thank my husband, Lei Gao, for his love and support over all these years and two of my academic sisters: Dr. Jing Wang and Dr. Lan Xue for their generous help.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

1.1 Nonlinear Time Series Prediction Model

Classic regression and time series tools such as the generalized linear model and the linear autoregression are known to be inadequate for complex data that exhibit nonlinearity. This recognition has motivated the development of non- and semiparametric regression techniques, with far reaching applications, see, for example, Fan and Gijbels (1996), Bosq (1998), Fan and Yao (2003).

A typical nonparametric problem in time series analysis is the classical decomposition of a realization of a time series into a slowly changing function known as a "trend component", or simply trend, a periodic function referred to as a "seasonal component", and finally a "random noise component", which in terms of the regression theory should be called the time series of residuals. In time series analysis smoothing problems occur of course in the spectral domain when we want to estimate the spectral density, e.g. for model fitting. In the time domain nonparametric prediction is one of the fields where smoothing methods are intensively used. A well-known example is the water flow prediction from a time series of river data, see Section 4.5 in Chapter 4. In the motorcycle crash test, the acceleration of the dummy head after impact follows a complicated instead of a simple polynomial time trend. Another example of the nonlinear time series is the quarterly unemployment rate of U.S. women, which follows a nonlinear instead of a simple linear prediction formula. Effective tools for extracting information from such complex regression data have to be non- and semiparametric in nature.

In the following, let $\left\{\mathbf{X}_i^T, Y_i\right\}_{i=1}^n = \{X_{i,1}, ..., X_{i,d}, Y_i\}_{i=1}^n$ be a $(d+1)$-dimensional strictly stationary process following the stochastic regression model

$$Y_i = m(\mathbf{X}_i) + \sigma(\mathbf{X}_i)\varepsilon_i, m(\mathbf{X}_i) = E(Y_i|\mathbf{X}_i), \tag{1.1.1}$$

in which $E(\varepsilon_i|\mathbf{X}_i) = 0$, $E\left(\varepsilon_i^2|\mathbf{X}_i\right) = 1$, $1 \leq i \leq n$. The d-variate functions m, σ are the unknown mean and standard deviation of the response Y_i conditional on the predictor vector $\mathbf{X}_i$, often estimated nonparametrically.

Two very popular forms of nonparametric regression are kernel/local polynomial type and spline type smoothing. In this work, the polynomial spline smoothing is extensively studied for nonlinear time series. The greatest advantages of spline smoothing, as pointed out in Huang and Yang (2004), Xue and Yang (2006 b) are its simplicity and fast computation.

For model in (1.1.1), when the dimension of the predictor vector $\mathbf{X}_i$ is 1 ($d = 1$), spline confidence bands are obtained in Chapter 2 for time series prediction function m under weak dependence. Application of smoothing techniques to high dimensional time series have been hampered due to the lack of effective tools to address the "curse of dimensionality", which refers to the poor convergence rate of nonparametric estimation of general multivariate function. Much effort has been devoted to methods of circumventing this difficulty. In the words of Xia, Tong, Li and Zhu (2002), there are essentially two approaches: function approximation and dimension reduction. Additive model and single-index model, special cases of model (1.1.1), are good examples to represent these two approaches. Chapter 3 and Chapter 4 discuss these two models separately.

1.2 Spline Confidence Bands

Consider the one dimensional case of model (1.1.1) for strictly stationary bivariate time series $\{(X_i, Y_i)\}_{i=1}^n$

$$Y_i = m(X_i) + \sigma(X_i)\varepsilon_i, i = 1, ..., n, \tag{1.2.1}$$

where the errors $\{\varepsilon_i\}_{i=1}^n$ are white noise, i.e., $E(\varepsilon_i|X_i) = 0, \mathrm{var}(\varepsilon_i|X_i) = 1$ and ε_i is a martingale difference for the σ-field $\mathcal{F}_i = \sigma\left\{X_j, \varepsilon_{j-1}, 1 \leq j \leq i\right\}$ for $i = 1, ..., n$.

To put the discussion in perspective, consider the question of how the adjustment of GDP autonomously influence the change of the environmental quality in Japan, see Section 2.5.2 in Chapter 2. The logarithm of GDP per capita and the emissions per capita of Japan are decomposed as $u(t) + X_t$ and $v(t) + Y_t$, $t = 1, ..., n$ respectively, where the quadratic trends $u(t)$ and $v(t)$ are given in (2.5.3), $\{(X_t, Y_t)\}_{t=1}^n$ are zero mean stationary time series of residuals. The aforementioned question can be formulated in terms of various hypotheses about the prediction function $m(x) = E(Y_t|X_t = x)$. In Figure 4.11 (b), a 99% conservative simultaneous confidence band of $m(x)$ is plotted together with the linear regression line, clearly showing the nonlinear dependence of Y_t on X_t. The corresponding Figure 4.10 (b) for the United States, however, shows a linear and insignificant $m(x)$. Making such inference about the global shape of the prediction function $m(x)$ depends crucially on the construction of simultaneous confidence bands for m using the time series observations $\{(X_i, Y_i)\}_{i=1}^n$.

In Chapter 2, asymptotically conservative simultaneous confidence bands are constructed for nonparametric prediction function m based on piecewise constant and piecewise linear polynomial spline estimation, respectively. Simulation experiments have provided strong evidence that corroborates with the asymptotic theory. As an application, after removing the environmental Kuznets curve trend effects, the impact of the economic intervention on environmental quality change is quantified for the United States and Japan, with different conclusions.

1.3 Nonlinear Additive Autoregression (NAAR) Model

For multi-dimensional strictly stationary time series $\{X_{i,1}, ..., X_{i,d}, Y_i\}_{i=1}^n$, the following additive structure is assumed for model (1.1.1)

$$Y_i = c + \sum_{\alpha=1}^{d} m_\alpha \left(X_{i,\alpha}\right) + \sigma\left(\mathbf{X}_i\right) \varepsilon_i \tag{1.3.1}$$

In nonlinear additive autoregression data-analytical context, each predictor $X_{i,\alpha}, 1 \leq \alpha \leq d$ could be observed lagged values of Y_i, such as $X_{i,\alpha} = Y_{i-\alpha}$, or of a different times series. Model (1.3.1) therefore, is the exact same nonlinear additive autoregression model of

Huang and Yang (2004), which allows for exogenous variables. For identifiability, additive component functions must satisfy the conditions $Em_\alpha\left(X_{i,\alpha}\right) \equiv 0, \alpha = 1, ..., d$.

Application of additive model to high dimensional time series data has been hampered by the scarcity of smoothing tools. The straightforward kernel methods are too computationally intensive for high dimension, thus limiting their applicability to small number of predictors. Spline methods on the other hand, provide only convergence rates but no asymptotic distributions, so no measures of confidence can be assigned to the estimators.

In Chapter 3, a spline-backfitted kernel estimator is proposed for estimating the unknown component functions $\{m_\alpha(\cdot)\}_{\alpha=1}^{d}$ based on a geometrically strong mixing sample following model (1.3.1). under minimal smoothness assumptions. The idea is to employ one step backfitting after the spline pilot estimators, and then follow up with kernel smoothing, which combines the fast computing of polynomial spline smoothing and the good asymptotic property of kernel smoothing. Thus, the spline-backfitted kernel estimator is both computationally expedient for analyzing very high dimensional time series, and theoretically reliable to make inference on the component functions with confidence.

1.4 Single-Index Prediction (SIP) Model

Single-index model, a special case of projection pursuit regression, has proven to be an efficient way of coping with the high dimensional problem in nonparametric regression. Single-index model summarizes the effects of the explanatory variables within a single variable called the index. The basic appeal of single-index model is its simplicity: the d-variate function $m(\mathbf{x}) = m(x_1, ..., x_d)$ is expressed as a univariate function g of $\mathbf{x}^T\boldsymbol{\theta}_0 = \sum_{p=1}^{d} \theta_{0,p}x_p$.

In Chapter 4, a robust single-index prediction (SIP) model is introduced for stochastic regression model 1.1.1 regardless if the underlying function is exactly a single-index function. Applications of SIP models lie in a variety of fields, such as discrete choice analysis in econometrics and dose-response models in biometrics, where high-dimensional regression models are often employed, see Härdle, Hall and Ichimura (1993). The proposed spline estimator of the index coefficient possesses not only the usual strong consistency and $\sqrt{n}$-

rate asymptotically normal distribution, but also is as efficient as if the true link function g is known. By taking advantage of the spline smoothing method and the iterative method, the proposed procedure is much faster than the MAVE method, see Xia, Tong, Li and Zhu (2002). This procedure is especially powerful for large sample size n and high dimension d and unlike the MAVE method, the performance of the SIP remains satisfying in the case $d > n$.

1.5 Polynomial Spline Smoothing

Let $\{X_i, Y_i\}_{i=1}^{n}$ be a strictly stationary process. Assume that X_i, $i = 1, ..., n$, are supported on a compact interval $[a, b]$. Polynomial splines begin by choosing a set of knots (typically, much smaller than the number of data points n), and a set of basis functions spanning a set of piecewise polynomials satisfying continuity and smoothness constraints.

To be specific, divide $[a, b]$ into $(N + 1)$ subintervals $J_j = [t_j, t_{j+1})$, $j = 0, ..., N - 1, J_N = [t_N, 1]$, where $T := \{t_j\}_{j=1}^{N}$ is a sequence of equally-spaced points, called interior knots, given as

$$t_{1-k} = ... = t_{-1} = t_0 = a < t_1 < ... < t_N < b = t_{N+1} = ... = t_{N+k},$$

in which $t_j = jh$, $j = 0, 1, ..., N + 1, h = 1/(N + 1)$ is the distance between neighboring knots. Denote by

$$C^{(k)}[a, b] = \{m | \text{the } kth \text{ order derivative of } m \text{ is continuous on } [a, b]\} \tag{1.5.1}$$

and $G^{(k-2)} = G^{(k-2)}[a, b]$ the space of all $C^{(k-2)}[a, b]$ functions that are polynomials of degree $k - 1$ on each interval. The j-th B-spline of order k for the knot sequence T denoted by $B_{j,k}$ is recursively defined by the de Boor (2001), i.e.

$$B_{j,k}(u) = \frac{(u - t_j) B_{j,k-1}(u)}{t_{j+k-1} - t_j} - \frac{(u - t_{j+k}) B_{j+1,k-1}(u)}{t_{j+k} - t_{j+1}}, 1 - k \leq j \leq N, \tag{1.5.2}$$

for $k > 1$, with

$$B_{j,1}(u) = I_{\{u \in J_j\}} = \begin{cases} 1 & t_j \leq u < t_{j+1} \\ 0 & \text{otherwise} \end{cases}.$$

For model (1.2.1), assume that $m(x)$ belongs to $C^{(k)}[a,b]$, the space of functions that have k-th order continuous derivatives for some integer $k > 0$, on the interval $[a,b]$. The polynomial spline estimator is

$$\hat{m}_k(\cdot) = \underset{g(\cdot)\in G^{(k-2)}[a,b]}{\operatorname{argmin}} \sum_{i=1}^{n} \{Y_i - g(X_i)\}^2, k > 0. \tag{1.5.3}$$

In the rest of this dissertation, spline smoothing is applied for the stochastic regression model (1.1.1) under different conditions.

CHAPTER 2

Spline Confidence Bands for Time Series Prediction Function

2.1 Introduction

Theoretical properties of nonparametric smoothers are typically examined in terms of mean square, pointwise, or uniform rate of convergence, while practical consideration favors methods that are easy to implement and interpret. In addition, fast computing is appealing for users of smoothers. For kernel smoothing of independent data, satisfactory results on rates of convergence have been obtained, see Fan and Gijbels (1996) for pointwise and mean square convergence rates, Müller, Stadtmüller and Schmitt (1987) for confidence intervals of derivative estimates, Neumann (1995, 1997) for bandwidth choice and construction of confidence intervals, Hall and Titterington (1988), Härdle (1989), Xia (1998), Claeskens and Van Keilegom (2003) for uniform confidence bands. Spline smoothers of independent data have been investigated in parallel, see for example, Stone (1985, 1994) for mean square convergence, Huang (2003) for pointwise convergence, and Zhou, Shen and Wolfe (1998) for uniform confidence bands.

Nonparametric smoothing of weakly dependent data has been vigorously pursued in many directions due to its superiority for the modeling and forecasting of nonlinear time series, see, for instance, Fan and Yao (2003) for kernel type autoregression smoothing, and Huang and Yang (2004) for spline type autoregressive smoothing. Confidence bands, however, remain unavailable for all nonparametric smoothers based on dependent observations, due to the lack of Hungarian embedding for dependent random variables, similar to that established by Tusnády (1977) for independent random variables. Existing results on nonparametric smooth confidence bands rely on such strong approximation result of i.i.d. sample, see, for instance, Bickel and Rosenblatt (1973), Rosenblatt (1976), Härdle (1989), Xia (1998), Claeskens and Van Keilegom (2003).

In this chapter, asymptotic simultaneous confidence bands are obtained for the unknown regression function $m(x)$ in (1.1.1) based on the polynomial spline estimator $\hat{m}_k(x)$ defined in (1.5.3), while the observations $\{(X_i, Y_i)\}_{i=1}^n$ are only assumed to have α-mixing coefficient $\alpha(n)$ decaying geometrically (see Assumption (A4) of Section 2.2). Instead of applying the usual Hungarian embedding technique used in most existing works, we make use of the Berry-Esseen bound in Sunklodas (1984) for sequences of mixing random variables to establish that the constructed confidence bands are conservative. The resulting confidence bands are comparable in terms of formula and narrowness to those constructed for i.i.d. sample. Further research will show that these simultaneous confidence bands are very useful for multi-step ahead forecasting of time series data, such as studied in Chen, Yang and Hafner (2004).

The rest of this chapter is organized as follows. The main findings of splines confidence bands are stated in Section 2.2. Section 2.3 provides further insights into the error structure of spline estimators, from which one is able to obtain the asymptotic confidence bands. This is accomplished by establishing simultaneous Berry-Esseen bound for the estimation noise. Section 2.4 describes the actual steps to implement the confidence bands. Section 2.5 reports the findings in an extensive simulation study and the application to the environmental Kuznets curve (EKC) analysis. All technical proofs are contained in Section 2.6.

2.2 Main results

Before stating the main theorems, we formulate some assumptions.

(A1) *The regression function* $m \in C^{(k)}[a,b]$, $k = 1, 2$.

(A2) *The marginal density function* $f(x)$ *of* X *is continuous and positive on its compact* *support, the interval* $[a,b]$. *The standard deviation function* $\sigma(x)$ *is continuous and positive on* $[a,b]$.

(A3) *The number of interior knots* N *satisfies:* $(n/\log n)^{1/(2k+1)} \ll N \ll n^{1/3}$, *hence for* $k=2$, *one can take* $N \sim n^{1/5}$, *while for* $k=1$, *one can take* $N \sim n^{1/3}(\log n)^{-1/6}$.

(A4) *There exist positive constants* K_0 *and* λ_0 *such that* $\alpha(n) \leq K_0 e^{-\lambda_0 n}$ *holds for all* n, *where the strong mixing coefficient of order* n *is defined as*

$$\alpha(n) = \sup_{B \in \sigma\{X_s, Y_s, s \leq t\}, C \in \sigma\{X_s, Y_s, s \geq t+n\}} |P(B \cap C) - P(B)P(C)|, \ n \geq 1.$$

(A5) *The joint distribution 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* $M_0 > 0$ *such that*

$$\sup_{x \in [a,b]} E\left(|\varepsilon|^3\,|X=x\right) < M_0.$$

Assumptions (A1)-(A5) are typical in the nonparametric smoothing literature, see for instance, Fan and Yao (2003), Huang and Yang (2004).

For any $x \in [a,b]$, define its location index $j(x)$ and relative location index $\delta(x)$ as

$$j(x) = j_n(x) = \min\left\{\left[\frac{x-a}{h}\right], N\right\}, \ \delta(x) = \frac{x - t_{j(x)}}{h}. \tag{2.2.1}$$

It is clear that $t_{j(x)} \leq x < t_{j(x)+1}$, $0 \leq \delta(x) < 1$, $\forall x \in [a,b]$, $j(b) = N$, $\delta(b) = 1$. For any L^2-integrable functions ϕ, φ on $[a,b]$, the theoretical and empirical inner products and the corresponding L^2 norms are defined respectively by

$$\langle \phi, \varphi \rangle = \int_a^b \phi(x)\varphi(x)f(x)\,dx = E\{\phi(X_i)\varphi(X_i)\},$$

$$\|\phi\|_2^2 = E\left\{\phi^2(X)\right\} = \int_a^b \phi^2(x)f(x)\,dx,$$

$$\langle \phi, \varphi \rangle_n = n^{-1}\sum_{i=1}^n \{\phi(X_i)\varphi(X_i)\}, \ \|\phi\|_{2,n}^2 = n^{-1}\sum_{i=1}^n \phi^2(X_i).$$

For notation simplicity, we denote by $\|\cdot\|_\infty$ 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 $\omega(r,h) = \max_{x,x'\in[a,b],|x-x'|\le h} |r(x) - r(x')|$. By the uniform continuity of r on an interval $[a,b]$, one has $\lim_{h\to 0} \omega(r,h) = 0$.

We denote the theoretical norms of $B_{j,k}$, $k = 1,2$ in (1.5.2) as follows

$$c_{j,n} = \|B_{j,1}\|_2^2 = \int I_j(x) f(x)\, dx, \tag{2.2.2}$$

$$d_{j,n} = \|B_{j,2}\|_2^2 = \int K\left\{\left(x - t_{j+1}\right) h^{-1}\right\} f(x)\, dx. \tag{2.2.3}$$

For theoretical analysis, in the following of this chapter, we use the rescaled B-spline basis (divided by its theoretical norm $c_{j,n}$, $d_{j,n}$) $\{B_{j,1}(x)\}_{j=0}^N$ and $\{B_{j,2}(x)\}_{j=-1}^N$ for constant spline space $G^{(-1)}$ and linear spline space $G^{(0)}$ defined in Section 1.5. The inner product matrix of the B-spline basis $\{B_{j,1}(x)\}_{j=0}^N$ is obviously the identity matrix $\mathbf{I}_{N+1}$, while the corresponding matrix $\mathbf{V}$ of the B-spline basis $\{B_{j,2}(x)\}_{j=-1}^N$ is denoted as

$$\mathbf{V} = \left(v_{j'j}\right)_{j,j'=-1}^N = \left(\left\langle B_{j',2}, B_{j,2}\right\rangle\right)_{j,j'=-1}^N, \tag{2.2.4}$$

whose inverse matrix $\mathbf{S}$ and its 2×2 diagonal submatrices are expressed as

$$\mathbf{S} = \left(s_{j'j}\right)_{j,j'=-1}^N = \mathbf{V}^{-1},\ \mathbf{S}_j = \begin{pmatrix} s_{j-1,j-1} & s_{j-1,j} \\ s_{j,j-1} & s_{j,j} \end{pmatrix},\ j = 0, ..., N. \tag{2.2.5}$$

Next define matrices Σ, $\Delta(x)$ and Ξ_j as

$$\Sigma = \left(\sigma_{jl}\right)_{j,l=-1}^N = \left\{\int \sigma^2(v) B_{j,2}(v) B_{l,2}(v) f(v)\, dv\right\}_{j,l=-1}^N, \tag{2.2.6}$$

$$\Delta(x) = \begin{pmatrix} c_{j(x)-1}\{1-\delta(x)\} \\ c_{j(x)}\delta(x) \end{pmatrix},\ c_j = \begin{cases} \sqrt{2} & j = -1, N \\ 1 & 0 \le j \le N-1 \end{cases},$$

$$\Xi_j = \begin{pmatrix} l_{j+1,j+1} & l_{j+1,j+2} \\ l_{j+2,j+1} & l_{j+2,j+2} \end{pmatrix}, j = 0, 1, ..., N, \tag{2.2.7}$$

where terms $\{l_{ik}\}_{|i-k|\le 1}$ are the entries of the inverse of the $(N+2)\times(N+2)$ matrix $\mathbf{M}_{N+2}$ and can be computed by Lemma 2.6.10

$$\mathbf{M}_{N+2} = \begin{pmatrix} 1 & \sqrt{2}/4 & & & & 0 \\ \sqrt{2}/4 & 1 & 1/4 & & & \\ & 1/4 & 1 & \ddots & & \\ & & \ddots & \ddots & 1/4 & \\ & & & 1/4 & 1 & \sqrt{2}/4 \\ 0 & & & & \sqrt{2}/4 & 1 \end{pmatrix}. \tag{2.2.8}$$

Define next

$$\sigma_{n,1}^{2}(x) = \frac{\int I_{j(x)}(v)\,\sigma^{2}(v)\,f(v)\,dv}{nc_{j(x),n}^{2}}, \tag{2.2.9}$$

$$\sigma_{n,2}^{2}(x) = \frac{1}{n}\sum_{j,j',l,l'=-1}^{N} B_{j',2}(x)\,B_{l',2}(x)\,s_{jj'}s_{ll'}\sigma_{jl}, \tag{2.2.10}$$

with $j(x)$ defined in (2.2.1), $c_{j,n}$ in (2.2.2) and $s_{ll'}$ and σ_{jl} in (2.2.5), (2.2.6). These $\sigma_{n,k}^{2}(x)$ are shown in Lemmas 2.6.5, 2.6.11 to be the pointwise variance functions of the spline estimators $\hat{m}_k(x)$, $k = 1, 2$. Lastly, define an inflation correction factor, for any $\alpha \in (0,1)$

$$d_n(\alpha) = 1 - \{2\log(N+1)\}^{-1}\left[\log(\alpha/2) + \frac{1}{2}\{\log\log(N+1) + \log 4\pi\}\right]. \tag{2.2.11}$$

THEOREM 2.2.1. *Under Assumptions (A1)-(A5), for any* $\alpha \in (0,1)$, *an asymptotic* $100(1-\alpha)\%$ *conservative confidence band for* $m(x)$ *over interval* $[a, b]$ *is*

$$\hat{m}_k(x) \pm \sigma_{n,k}(x)\{2k\log(N+1)\}^{1/2} d_n(\alpha/k), k = 1, 2. \tag{2.2.12}$$

In other words, for $k = 1, 2$

$$\liminf_{n\to\infty} P\left[m(x) \in \hat{m}_k(x) \pm \sigma_{n,k}(x)\{2k\log(N+1)\}^{1/2} d_n(\alpha/k), \forall x \in [a, b]\right] \geq 1 - \alpha,$$

in which $\sigma_{n,1}(x)$ *is given in (2.2.9), replaceable by* $\sigma(x)\{f(x)nh\}^{-1/2}$ *according to (2.6.6) in Lemma 2.6.5,* $\sigma_{n,2}(x)$ *is given in (2.2.10), replaceable by* $\sigma(x)\{2f(x)nh/3\}^{-1/2}\left\{\Delta^{T}(x)\,\Xi_{j(x)}\Delta(x)\right\}^{1/2}$ *according to Lemma 2.6.9 and (2.6.19) in Lemma 2.6.11, and* $d_n(\alpha)$ *is given in (2.2.11).*

2.3 Error decomposition

In this section, we break the polynomial spline estimation error $\hat{m}_k(x) - m(x)$ into a bias term and a noise term, with $\hat{m}_k(x)$ given in (1.5.3). We first establish the uniform rate at which the empirical inner product approximates the theoretical inner product for all B-splines.

Lemma 2.3.1. *Under Assumptions (A3) and (A5), we have*

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

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

Note that the spline estimator in (1.5.3) is $\hat{m}_k(x)$ that $\hat{m}_k(x) \equiv \sum_{j=1-k}^{N}\hat{\lambda}_{j,k}B_{j,k}(x)$, where

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

With a slight abuse of notation, introduce a function $\mathbf{Y}$ defined only on data points: $\mathbf{Y}(X_i) \equiv Y_i$, $1 \le i \le n$, and write

$$\hat{m}_k(x) = \left\{B_{j,k}(x)\right\}_{1-k\le j\le N}^{T}\left(\left\langle B_{j',k},B_{j,k}\right\rangle_n\right)_{1-k\le j,j'\le N}^{-1}\left\{\left\langle \mathbf{Y},B_{j,k}\right\rangle_n\right\}_{j=1-k}^{N}.$$

Define a similar function $\mathbf{E}$ as $\mathbf{E}(X_i) \equiv \sigma(X_i)\varepsilon_i$, $1 \le i \le n$, then on data points $\mathbf{Y} = \mathbf{m}+\mathbf{E}$ with $\mathbf{m} = \{m(X_1),\ldots,m(X_n)\}^T$. An empirical inner product yields $\hat{m}_k(x) = \tilde{m}_k(x) + \tilde{\varepsilon}_k(x)$, where

$$\tilde{m}_k(x) = \left\{B_{j,k}(x)\right\}_{1-k\le j\le N}^{T}\left(\left\langle B_{j',k},B_{j,k}\right\rangle_n\right)_{1-k\le j,j'\le N}^{-1}\left\{\left\langle \mathbf{m},B_{j,k}\right\rangle_n\right\}_{j=1-k}^{N} \tag{2.3.3}$$

$$\tilde{\varepsilon}_k(x) = \left\{B_{j,k}(x)\right\}_{1-k\le j\le N}^{T}\left(\left\langle B_{j',k},B_{j,k}\right\rangle_n\right)_{1-k\le j,j'\le N}^{-1}\left\{\left\langle \mathbf{E},B_{j,k}\right\rangle_n\right\}_{j=1-k}^{N}. \tag{2.3.4}$$

Thus, the estimation error $\hat{m}_k(x) - m(x)$ consists of a bias term $\tilde{m}_k(x) - m(x)$ and a noise term $\tilde{\varepsilon}_k(x)$, such that

$$\hat{m}_k(x) - m(x) = \left\{\tilde{m}_k(x) - m(x)\right\} + \tilde{\varepsilon}_k(x). \tag{2.3.5}$$

Lemma 2.3.2. *[de Boor (2001) page 149] There exists an absolute constant $C_k > 0$, $k \ge 1$, such that for every $m \in C^{(k)}[a,b]$, there exists a function $g \in G^{(k-2)}[a,b]$, such that*

$$\|g-m\|_\infty \le C_k\left\|\omega\left(m^{(k-1)},h\right)\right\|_\infty h^{k-1} \le C_k\left\|m^{(k)}\right\|_\infty h^{k}.$$

LEMMA 2.3.3. *[Huang (2003) Theorem 5.1] Under Assumptions (A1)-(A4), there exists an absolute constant* $C_k > 0,\ k \geq 1$, *such that for any* $m \in C^{(k)}[a,b]$ *and the function* $\tilde{m}_k(x)$ *as in (2.3.3), with probability approaching* 1

$$\|\tilde{m}_k(x) - m(x)\|_\infty \leq C_k \inf_{g \in G^{(k-2)}} \|g - m\|_\infty = O_p\left(h^k\right). \qquad (2.3.6)$$

Lemmas 2.3.2 and 2.3.3 establish that the bias term is of order $O_p\left(h^k\right)$ uniformly over $x \in [a,b]$. Hence the main hurdle of proving Theorem 2.2.1 is the noise term $\tilde{\varepsilon}_k(x)$ defined in (2.3.4). This is handled by the next proposition.

PROPOSITION 2.3.1. *Under Assumptions (A2)-(A5), with* $\sigma_{n,1}(x)$ *given in (2.2.9) and* $\sigma_{n,2}(x)$ *given in (2.2.10), for any* $0 < \alpha < 1,\ k = 1,2$, *one has*

$$\liminf_{n \to \infty} P\left[\sup_{x \in [a,b]} \left|\sigma_{n,k}^{-1}(x)\tilde{\varepsilon}_k(x)\right| \leq \{2k\log(N+1)\}^{1/2} d_n(\alpha/k)\right] \geq 1 - \alpha. \qquad (2.3.7)$$

2.4 Implementation

In this section, we describe in detail the procedures implemented to construct the confidence bands in Theorem 2.2.1. All of the codes have been written in R.

Given any sample $\{(X_i, Y_i)\}_{i=1}^n$, use the minimum and maximum values of $\{X_i\}_{i=1}^n$ as the endpoints of interval $[a,b]$. The number of knots N is taken to be $\left[c_k n^{1/3}(\log n)^{-1/6}\right]$ for $k = 1$ and $\left[c_k n^{1/5}\right]$ for $k = 2$, where c_k $(k = 1,2)$ are positive integers. As with previous works on confidence bands (Härdle 1989, Xia 1998, Claeskens and Van Keilegom 2003), explicit formula of coverage probability for the bands does not exist, hence there is no optimal method to select c_k $(k = 1,2)$. So we have not attempted adaptive knot selection, as Härdle, Marron and Yang (1997) had illustrated that it could lead to uniform inconsistency. We have set $c_1 = 6, c_2 = 3$ for piecewise constant and piecewise linear bands respectively, which works well in all simulations.

The least squares problem in (1.5.3) is solved by writing spline functions as linear combinations of the truncated power base, which are $1, x, ..., x^{k-1}, (x - t_j)_+^{k-1}, j = 1, ..., N$. In other words, we take

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

where the coefficients $\left\{\hat{\gamma}_0, ..., \hat{\gamma}_{k-1}, \hat{\gamma}_{1,k}, ..., \hat{\gamma}_{N,k}\right\}^T$ minimize the following sum of squares

$$\sum_{i=1}^{n} \left\{ Y_i - \sum_{p=0}^{k-1} \gamma_p X_i^p + \sum_{j=1}^{N} \gamma_{j,k} (X_i - t_j)_+^{k-1} \right\}^2 .$$

When constructing the confidence bands, one needs to estimate the unknown functions $f(x)$ and $\sigma^2(x)$ for the evaluation of the functions $\sigma_{n,1}(x)$ in (2.2.9) and $\sigma_{n,2}(x)$ in (2.2.10) according to Lemma 2.6.5 and Lemma 2.6.11.

Let $\tilde{K}(u) = 15\left(1-u^2\right)^2 I\{|u| \leq 1\}/16$ be the quartic kernel, s_n be the sample standard deviation of $\{X_i\}_{i=1}^n$ and

$$\hat{f}(x) = n^{-1} \sum_{i=1}^{n} h_{\mathrm{rot},f}^{-1} \tilde{K}\left(\frac{X_i - x}{h_{\mathrm{rot},f}}\right), \tag{2.4.2}$$

where $h_{\mathrm{rot},f} = (4\pi)^{1/10}(140/3)^{1/5} n^{-1/5} s_n$ is the rule-of-thumb bandwidth of Silverman (1986). Theorem 2.2 of Bosq (1998), page 47, implies the following uniform consistency result

$$\sup_{x \in [a,b]} \left| \hat{f}(x) - f(x) \right| = 0, \text{ a.s.}. \tag{2.4.3}$$

Define vectors $\mathbf{Z}_k = \left\{Z_{1,k}, .., Z_{n,k}\right\}^T$, $k = 1, 2$ with $Z_{i,k} = \{Y_i - \hat{m}_k(X_i)\}^2$, then the spline estimation of $\sigma^2(x)$, $\hat{\sigma}_k^2(x)$, $k = 1, 2$, can be obtained by using the Nadaraya-Watson estimation on data $\left\{X_i, Z_{i,k}\right\}_{i=1}^n$. It is clear from standard theory of kernel smoothing that

$$\max_{k=1,2} \sup_{x \in [a,b]} \left| \hat{\sigma}_k^2(x) - \sigma^2(x) \right| = O_p(h). \tag{2.4.4}$$

With all the above preparation, one can compute the following confidence bands

$$\hat{m}_k(x) \pm \hat{\sigma}_{n,k}(x, \mathrm{opt}) \{2k \log(N+1)\}^{1/2} d_n(\alpha/2), \; k = 1, 2, \mathrm{opt} = 1, 2, \tag{2.4.5}$$

where $\hat{m}_k(x)$ is given in (2.4.1), the additional parameter opt $= 1, 2$ indicating the estimation being at each value x or at the nearest left knot, with $j(x)$ and $\hat{f}(x)$ defined in (2.2.1) and (2.4.2)

$$\begin{aligned}\hat{\sigma}_{n,1}(x,1) &= \hat{\sigma}_1\left(t_{j(x)}\right)\hat{f}^{-1/2}\left(t_{j(x)}\right)n^{-1/2}h^{-1/2}, &(2.4.6)\\ \hat{\sigma}_{n,1}(x,2) &= \hat{\sigma}_1(x)\hat{f}^{-1/2}(x)n^{-1/2}h^{-1/2}, &(2.4.7)\end{aligned}$$

$$\hat{\sigma}_{n,2}(x,1) = \left\{\Delta^T(x)\Xi_{j(x)}\Delta(x)\right\}^{1/2}\left\{nh\hat{f}\left(t_{j(x)}\right)\right\}^{-1/2}\sqrt{3/2}\hat{\sigma}_2\left(t_{j(x)}\right), \qquad (2.4.8)$$

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

Since $\sup_{x\in[a,b]}\left|x - t_{j(x)}\right| \leq h \to 0$, as $n \to \infty$, and according to Lemma 2.6.9, the matrix Ξ_j approximates matrix $\mathbf{S}_j$ uniformly for $0 \leq j \leq N$, (2.4.3) and (2.4.4) entail that all of the four bands above are asymptotically conservative.

2.5 Examples

2.5.1 Simulation example

To illustrate the finite-sample behavior of the proposed confidence bands, some simulation results are presented. The number of interior knots N is chosen according to Section 2.4. The data set in our simulation study is generated from heteroscedastic regression model (1.2.1), with

$$m(x) = \sin(2\pi x), \; \sigma(x) = \sigma_0\frac{100 - \exp(x)}{100 + \exp(x)}, \; \varepsilon \sim N(0,1), \; \sigma_0 = 0.2, 0.5. \qquad (2.5.1)$$

We simulate $\{T_i\}_{i=1}^n$ from a moving average sequence of order q, i.e,

$$T_i = \frac{1}{\sqrt{1+\theta_1^2+...+\theta_q^2}}\left(\xi_i + \theta_1\xi_{i-1} + \theta_2\xi_{i-2} + ... + \theta_q\xi_{i-q}\right),$$

where in the simulation, q is taken to be 4, $\theta_1 = ... = \theta_q = 0.2$ and ξ_i's are i.i.d. r.v.'s $\sim N(0,1)$. We then define $X_i = \Phi(T_i)$, where Φ is the standard normal distribution function, so X_i is uniformly distributed on $[0,1]$.

We choose sample size n to be 100, 200, 500 and 10000, confidence level $1 - \alpha = 0.99, 0.95$ as usual. 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 (2.4.5) with 500 replications of sample size $n = 100$, 200 and 500. The coverage probabilities of the confidence bands in (2.4.5) have also been computed by plugging in the true value of density function $f(x) = I_{[0,1](x)}$ and the variance function $\sigma(x)$ in (2.5.1), called the oracle bands as they use quantities that are unknown but for "oracles".

Table 4.1 shows that the performance of all four bands becomes much closer with larger sample size. When sample size reaches 500, all four bands have nearly the same coverage at noise level 0.2. In Table 4.2, the coverage percentages show very positive confirmation of Theorem 2.2.1 when $k = 2$. At sample size 100, regardless of noise level, both of the two piecewise linear bands in (2.4.5) achieve at least .980 and .948 for confidence level $1-\alpha = .99$ and .95, respectively.

From Tables 4.1 and 4.2, it is obvious that larger sample size guarantees improved coverage, while reasonable coverage has also been achieved at moderate sample sizes. While under the same circumstances, the band by linear spline performs much better than the band by constant spline. We have also observed that the noise level has more influence on the constant bands coverage, and very little on the linear bands'.

Corresponding to opt $= 1, 2$, four figures of constant bands (Figures 4.1 - 4.4) and four figures of linear bands (Figures 4.5 - 4.8) are created for graphical comparison: each with four types of symbols: dots (data), center smooth solid line (true curve), center dotted line (the spline estimated curve), upper and lower thick solid line (confidence bands). Comparing Figures 4.1 - 4.4, one sees that the band widths are very close as sample size reaches 500. This is more evident from Figures 4.5 - 4.8.

In all figures, the confidence bands of $n = 500$ are thinner and fit better than those of $n = 100$. Also the smaller the significance level, the wider the confidence band. Overall, linear bands are superior to constant ones in terms of smoothness and narrowness.

Observing that the estimation of $\sigma_{n,2}(x)$ by $\hat{\sigma}_{n,2}(x, 1)$ at knots as in (2.4.8) or by $\hat{\sigma}_{n,2}(x, 2)$ at all observations as in (2.4.9) does not seem to have much noticeable impact on the widths of the confidence bands, while the estimation at knots seems to produce closer coverage probabilities to the nominal confidence level, we recommend always using estimation by $\hat{\sigma}_{n,2}(x, 1)$ at knots for simpler and faster implementation.

For the linear bands, we have also carried out simulation at noise level 0.2, for sample size $n = 10000$ and opt $= 1$ (estimation on knots). The coverage is always 99.6% 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 365 seconds to run 500 replications with sample size as large as 10000 on a Pentium III PC. This is extremely fast considering that nonparametric regression is done without WARPing, see Härdle, Hlávka and Klinke (2000).

2.5.2 Environmental Kuznets curve (EKC)

The environmental Kuznets curve (EKC), an inverted-U relationship between pollution and income, is an influential generalization about the way environmental quality changes as a country makes the transition from poverty to relative affluence. The EKC predicts that pollution will first increase, but subsequently decline if income growth proceeds far enough. The shape of the relationship between the rate of environmental degradation and GDP per capita has been the subject of much empirical examination. Several studies have attempted to test the EKC hypothesis empirically. The majority of these studies use panel data in conjunction with a static fixed and/or random effects panel estimator. In this section, we examine whether or not countries (here we select US and Japan) actually behave like the EKC, and we further look at the nonparametric time series nature of the data set after elimination of the trend.

One key variable of this study, the environment index is the emissions of sulfur from 1850 to 1990, see Lefohn, Husar and Husar (1999). The other key variable is GDP per capita from 1850 to 1990, which can be obtained in Maddison (2003).

To gain an insight into the model structure, we decompose the logarithm of GDP per capita and Emission per capita into their trend parts and noise parts, respectively, i.e., for $t = 1, ..., n$

$$\{\log(\text{GDP per capita})\}_t = u(t) + X_t, \ \{\log(\text{Emission per capita})\}_t = v(t) + Y_t.$$

We are interested in two sets of hypotheses, given here separately in terms of the relationship between the trends $u(t)$ and $v(t)$, and between the stationary noise $\{X_t\}_{t=1}^n$ and $\{Y_t\}_{t=1}^n$.

EKC hypothesis: There exists an inverted-U relationship between $u(t)$ and $v(t)$. (see Figure 4.9)

Residual/noise hypothesis: There exists a linear relationship between $\{X_t\}_{t=1}^n$ and $\{Y_t\}_{t=1}^n$.

The EKC hypothesis can be tested by performing a routine trend analysis. After detrending, $\{X_t\}_{t=1}^n$ and $\{Y_t\}_{t=1}^n$ are obtained, then one can estimate the regression relation between them and construct an piecewise linear spline confidence band for the testing.

Case 1. United States Example

We get the trends $u(t)$, $v(t)$ of US data by fitting a polynomial regression on time t.

$$u(t) = 0.0051t + 3.3127, \ v(t) = -0.0001t^2 + 0.0261t - 2.1788, \qquad (2.5.2)$$

with the corresponding $R^2 = 0.9814, 0.9256$. So for US, the EKC hypothesis is retained by the trend analysis. After elimination of the trend, $\{X_t\}_{t=1}^n$, $\{Y_t\}_{t=1}^n$ appear to be stationary.

For the residual hypothesis, Figure 4.10 shows that when confidence level is as small as 80%, the linear regression line is still covered by the confidence band. This phenomena implies that the residual hypothesis is retained.

Moreover, we can see that the confidence bands also cover the horizontal line $E(Y_t|X_t) \equiv 0$. So one concludes that Y_t is unpredictable from X_t, that is, the intervention of emission is immune to the intervention of economy.

Case 2. Japan Example

The quadratic trends $u(t)$, $v(t)$ for Japan data are given as

$$u(t) = 0.0003t^2 - 0.0019t + 6.7308, \ v(t) = -0.0005t^2 + 0.0952t - 9.0772 \qquad (2.5.3)$$

with $R^2 = 0.9829, 0.9544$. From the trend relationship curve, one sees that it is not a U shaped curve as EKC predicted. However, we are not sure whether it would succeed to decouple environmental pollution and resource use from economic growth, which will make this a tuning point and U shape later. To test the residual hypothesis, Figure 4.11 shows that neither the linear regression line nor the horizontal line $E(Y_t|X_t) \equiv 0$ can be covered by the confidence bands even when the confidence level reaches 99%. So the residual hypothesis is rejected at significance level smaller than 0.01 given that the confidence band is already conservative. This phenomena implies that the intervention of emission is not immune to the intervention of economy, or say that the adjustment of GDP has autonomous influence on the change of environmental quality, but not in a linear way.

2.6 Proof of Theorem 2.2.1

2.6.1 Preliminaries of Theorem 2.2.1 with $k = 1$

Throughout the following, denote by c, C, any positive constants, without distinction.

The properties of $c_{j,n}$ and $d_{j,n}$ are given in the following lemma, whose proof consists of direct algebraic verifications.

LEMMA 2.6.1. *As $n \to \infty$, for $c_{j,n}$ defined in (2.2.2) and $d_{j,n}$ in (2.2.3)*

$$\begin{aligned}
c_{j,n} &= f\left(t_j\right) h \left(1 + r_{j,n,1}\right), \left\langle b_{j,1}, b_{j',1}\right\rangle \equiv 0, j \neq j', \qquad (2.6.1)\\
d_{j,n} &= \frac{2}{3} f\left(t_{j+1}\right) h \begin{cases} 1 + r_{j,n,2} & j = 0, \ldots, N-1, \\ 1/2 + r_{j,n,2} & j = -1, N, \end{cases}\\
\left\langle b_{j,2}, b_{j',2}\right\rangle &= \frac{1}{6} f\left(t_{j+1}\right) h \begin{cases} 1 + \tilde{r}_{j,n,2} & \left|j' - j\right| = 1, \\ 0 & \left|j' - j\right| > 1, \end{cases}
\end{aligned}$$

where

$$\max_{0 \le j \le N} \left|r_{j,n,1}\right| + \max_{-1 \le j \le N} \left|r_{j,n,2}\right| + \max_{-1 \le j \le N-1} \left|\tilde{r}_{j,n,2}\right| \le C\omega\left(f, h\right), \qquad (2.6.2)$$

$$\frac{1}{3} f\left(t_{j+1}\right) h \left\{1 - C\omega\left(f, h\right)\right\} \le d_{j,n} \le \frac{2}{3} f\left(t_{j+1}\right) h \left\{1 + C\omega\left(f, h\right)\right\}. \qquad (2.6.3)$$

To prove Lemma 2.3.1, we make use of the following Bernstein inequality for geometrically α-mixing sequence.

LEMMA 2.6.2. *[Bosq (1998), page 31, Theorem 1.4] Let $\{\xi_t, t \in \mathbb{Z}\}$ be a zero mean real valued α-mixing process, $S_n = \sum_{i=1}^{n} \xi_i$. Suppose that there exists $c > 0$ such that for $i = 1, \ldots, n$, $k = 3, 4, \ldots$, $E\left|\xi_i\right|^k \le c^{k-2} k! E\xi_i^2 < +\infty$, then for each $n > 1$, integer $q \in [1, n/2]$, each $\varepsilon > 0$ and $k \ge 3$*

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

where $\alpha(\cdot)$ is the α-mixing coefficient defined in (3.2.10) and

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

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

PROOF OF LEMMA 2.3.1. For brevity, we only give the proof of Lemma 2.3.1 for $A_{n,1}$.

For any $0 \le j \le N$, let $\eta_{i,j} = B_{j,1}^2(X_i) - 1$, then $\|B_{j,1}\|_{2,n}^2 - 1 = n^{-1}\sum_{i=1}^n \eta_{i,j}$, with $E\eta_{i,j} = 0$ and for any $r \ge 2$, C_r inequality implies that

$$E\left|\eta_{i,j}\right|^r = E\left|B_{j,1}^2(X_i) - 1\right|^r \le 2^{r-1}E\left[B_{j,1}^{2r}(X_i) + 1\right] \le C_0\left\{2h^{-1}\right\}^{r-1},$$

where $c_{j,n}$ is as (2.2.2) with properties given in (2.6.1) and (2.6.2). On the other hand

$$E\left(\eta_{i,j}^2\right) = E\left|B_{j,1}^2(X_i) - 1\right|^2 \ge E\left[B_{j,1}^4(X_i) - 1\right] = \left\{2c_{j,n}\right\}^{-1} - 1 \ge C_1h^{-1}.$$

So there is a constant c, such that for all $k > 2$, $E\left|\eta_{i,j}\right|^k \le \left(ch^{-1}\right)^{k-2}k!E\eta_{i,j}^2$. Thus Cramer's condition is satisfied with Cramer's constant equal to ch^{-1}. Applying Lemma 2.6.2 to $n^{-1}\sum_{i=1}^n \eta_{i,j}$, for any $\delta > 0$, $q \in [1, n/2]$, one has for $k = 3$

$$P\left\{\frac{1}{n}\left|\sum_{i=1}^n \eta_{i,j}\right| > \delta_n\right\} \le a_1 \exp\left(\frac{-q\delta_n^2}{25m_2^2 + 5c\delta_n}\right) + a_2(3)\,\alpha\left(\left[\frac{n}{q+1}\right]\right)^{6/7},$$

where

$$\delta_n = \frac{\delta \log n}{\sqrt{nh}},\ a_1 = 2\frac{n}{q} + 2\left(1 + \frac{\delta_n^2}{25m_2^2 + 5c\delta_n}\right),\ m_2^2 = E\eta_{i,j}^2 \sim h^{-1},$$

$$a_2(3) = 11n\left(1 + \frac{5m_3^{6/7}}{\delta_n}\right),\ m_3 = \max_{1\le i\le n} \|\eta_{i,j}\|_3 \le \left\{C_0\left(\frac{2}{h}\right)^2\right\}^{1/3}.$$

Observe that $\delta_n = o(1)$, then by taking q such that $\left[\frac{n}{q+1}\right] \ge c_0 \log n$, $q \ge c_1 n/\log n$ for some constants c_0, c_1, one has $a_1 = O(n/q) = O(\log n)$, $a_2(3) = o\left(n^2\right)$. Assumption (A4) yields that

$$\alpha\left(\left[\frac{n}{q+1}\right]\right)^{6/7} \le \left\{K_0 \exp\left(-\lambda_0\left[\frac{n}{q+1}\right]\right)\right\}^{6/7} \le Cn^{-6\lambda_0 c_0/7}.$$

Thus, for n large enough,

$$P\left\{\frac{1}{n}\left|\sum_{i=1}^n \eta_{i,j}\right| > \frac{\delta \log n}{\sqrt{nh}}\right\} \le c\log n \exp\left\{-c_2\delta^2 \log n\right\} + Cn^{2-6\lambda_0 c_0/7}.$$

Taking c_0, δ large enough, one has for large n, $P\left\{\frac{1}{n}\left|\sum_{i=1}^{n}\eta_{i,j}\right| > (nh)^{-1/2}\delta\log n\right\} \le n^{-3}$. Hence (2.3.1) holds because

$$\sum_{n=1}^{\infty} P\left\{\sup_{0\le j\le N}\left|\left\|B_{j,1}\right\|_{2,n}^{2}-1\right| > \frac{\delta\log n}{\sqrt{nh}}\right\} \le \sum_{n=1}^{\infty} n^{-3}N \le \sum_{n=1}^{\infty} 2n^{-2} < \infty. \qquad \Box$$

2.6.2 Proof of Proposition 2.3.1 with $k = 1$

To prove Proposition 2.3.1, the following important lemmas are needed. We denote by Φ the standard normal distribution function.

LEMMA 2.6.3. *[Sunklodas (1984), Theorem 1] Let $\{\xi_i\}_{i=1}^{n}$ be an α-mixing sequence with $E\xi_n = 0$. Denote $d := \max_{1\le i\le n}\left\{E|\xi_i|^{2+\delta}\right\}, 0 < \delta \le 1$, $S_n = \sum_{i=1}^{n}\xi_i$, $\sigma_n^2 := ES_n^2 \ge c_0 n$ for some $c_0 \in (0, +\infty)$. If $\alpha(n) \le K_0 e^{-\lambda_0 n}$, $\lambda_0 > 0$, $K_0 > 0$, then there exist $c_1 = c_1(K, \delta)$, $c_2 = c_2(K, \delta)$, such that*

$$\Delta_n = \sup_{z}\left|P\left\{\sigma_n^{-1}S_n < z\right\} - \Phi(z)\right| \le c_1\frac{d}{c_0\sigma_n^{\delta}}\left\{\log\left(\sigma_n/c_0^{1/2}\right)/\lambda\right\}^{1+\delta}$$

for any λ with $\lambda_1 \le \lambda \le \lambda_2$, where

$$\lambda_1 = c_2\left\{\log\left(\sigma_n/c_0^{1/2}\right)\right\}^{b}/n,\ b > 2(1+\delta)/\delta;\ \lambda_2 = 4\delta^{-1}(2+\delta)\log\left(\sigma_n/c_0^{1/2}\right).$$

LEMMA 2.6.4. *[Leadbetter, Lindgren and Rootzén (1983), Theorem 1.5.3] As $N \to \infty$, one has*

$$\left[\Phi\left(\tau/a_N + b_N\right)\right]^{N} \to \exp\left(-e^{-\tau}\right),$$

where

$$a_N = (2\log N)^{1/2},\ b_N = (2\log N)^{1/2} - (2\log N)^{-1/2}(\log\log N + \log 4\pi)/2.$$

Note that $\tilde{\varepsilon}_1(x)$ in (2.3.4) can be rewritten as

$$\tilde{\varepsilon}_1(x) = \sum_{j=0}^{N}\varepsilon_j^{*}B_{j,1}(x)\left\|B_{j,1}\right\|_{2,n}^{-2}, \tag{2.6.4}$$

with $\varepsilon_j^{*} = \left\langle \mathrm{E}, B_{j,1}\right\rangle_n = \frac{1}{n}\sum_{i=1}^{n}B_{j,1}(X_i)\,\sigma(X_i)\,\varepsilon_i$. Now define

$$\hat{\varepsilon}_1(x) = \sum_{j=0}^{N}\varepsilon_j^{*}B_{j,1}(x), x \in [a, b]. \tag{2.6.5}$$

The next lemma gives the pointwise variance of $\hat{\varepsilon}_1(x)$.

LEMMA 2.6.5. *The pointwise variance of $\hat{\varepsilon}_1(x)$ is the function $\sigma^2_{n,1}(x)$ defined in (2.2.9) which satisfies*

$$E\{\hat{\varepsilon}_1(x)\}^2 \equiv \sigma^2_{n,1}(x) = \frac{\sigma^2(x)}{f(x)nh}\{1 + r_{n,1}(x)\}, x \in [a,b], \tag{2.6.6}$$

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

PROOF. Note that $E(\varepsilon_i|X_i) = 0$, $E[B_{j,1}(X_i)B_{j,1}(X_k)\sigma(X_i)\sigma(X_k)\varepsilon_i\varepsilon_k] = 0, \forall i \neq k$, the rest of the proof follows from Lemma 2.6.1 and the continuity of functions $\sigma(x)$ and $f(x)$. □

The difference between $\tilde{\varepsilon}_1(x)$ in (2.6.4) and $\hat{\varepsilon}_1(x)$ in (2.6.5) is negligible uniformly over $x \in [a,b]$.

LEMMA 2.6.6. *Under Assumptions (A2) and (A5)*

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

PROOF. For any $x \in [a,b]$

$$|\tilde{\varepsilon}_1(x) - \hat{\varepsilon}_1(x)| \leq |\hat{\varepsilon}_1(x)| \left\{\sup_{0\leq j\leq N}\left|\|B_{j,1}\|^2_{2,n} - 1\right| \sup_{0\leq j\leq N}\|B_{j,1}\|^{-2}_{2,n}\right\}.$$

Meanwhile (2.3.1) of Lemma 2.3.1 implies that

$$\sup_{0\leq j\leq N}\left|\|B_{j,1}\|^2_{2,n} - 1\right| \leq A_{n,1}, (1 + A_{n,1})^{-1} \leq \sup_{0\leq j\leq N}\|B_{j,1}\|^{-2}_{2,n} \leq (1 - A_{n,1})^{-1},$$

hence the lemma follows. □

Since the stochastic function $\hat{\varepsilon}_1(x)$ given in (2.6.5) takes constant value on each interval I_j, one only has to bound each of the $N + 1$ rescaled noise terms simultaneously by the Berry-Esseen bound for weakly dependent data. First we verify the conditions in Lemma 2.6.3 for $\xi_{i,j} \equiv B_{j,1}(X_i)\sigma(X_i)\varepsilon_i$, $1 \leq i \leq n$, $j = 0, ..., N$.

LEMMA 2.6.7. *There exist constants $c_0(f,\sigma), C_0(f,\sigma) > 0$, such that for each $j = 0, ..., N$*

$$\sigma^2_{n,j} \equiv E\left(\sum_{i=1}^n \xi_{i,j}\right)^2 = nE\{B_{j,1}(X_i)\sigma(X_i)\varepsilon_i\}^2 = nc^{-1}_{j,n}\int_{I_j}\sigma^2(u)f(u)du = nc_{0,j}, \tag{2.6.7}$$

where $c_{0,j} = c^{-1}_{j,n}\int_{I_j}\sigma^2(u)f(u)du \geq c_0(f,\sigma) > 0$ with $c_{j,n}$ defined in (2.2.2) and

$$d_j \equiv E|\xi_{1,j}|^3 = E\left\{B^3_{j,1}(X_i)\sigma^3(X_i)|\varepsilon_i|^3\right\} \leq C_0(f,\sigma)h^{-1/2}. \tag{2.6.8}$$

Proof. Using the definition of $\sigma_{n,1}^2(x)$ in (2.2.9)

$$\begin{aligned}\sigma_{n,j}^2 &= E\left(\sum_{i=1}^n \xi_{i,j}\right)^2 = n^2 c_{j,n} E\left\{\frac{1}{n}\sum_{i=1}^n B_{j,1}(x) B_{j,1}(X_i)\sigma(X_i)\varepsilon_i\right\}^2 = n^2 c_{j,n}\sigma_{n,1}^2(x) \\ &= nc_{j,n}^{-1}\int I_{j(x)}(u)\sigma^2(u) f(u)\,du = nc_{0,j} \geq nc_0(f,\sigma) > 0.\end{aligned}$$

Next, by Lemma 2.6.1 and the continuity of functions $\sigma^2(x)$ and $f(x)$, one has

$$d_j = E|\xi_{1,j}|^3 \leq c_{j,n}^{-3/2}\int_{I_j}\sigma^3(u) f(u)\,du \leq C_0(f,\sigma) h^{-1/2}. \qquad \square$$

Proof of Proposition 2.3.1 with $p = 1$. Note that for any $j = 0, ..., N$, $x \in I_j$

$$\sigma_{n,1}^{-1}(x)\frac{1}{n}\sum_{i=1}^n B_{j,1}(x) B_{j,1}(X_i)\sigma(X_i)\varepsilon_i = \sigma_{n,j}^{-1}\sum_{i=1}^n \xi_{i,j}, \tag{2.6.9}$$

in which $\sigma_{n,j}^2 = nc_{0,j} \geq c_0(f,\sigma) > 0$ as in (2.6.7) and $d_j \leq C(f,\sigma) h^{-1/2}$ as in (2.6.8). Observing that $\{\xi_{i,j}\}_{i=1}^n$ forms a stationary α-mixing sequence, with $E\xi_{i,j} = 0$. Define

$$\Delta_n \equiv \max_{0\leq j\leq N}\sup_{z\in R}\left|P\left\{\sigma_{n,1}^{-1}(x)\tilde{\varepsilon}_1(x) \leq z, x \in I_j\right\} - \Phi(z)\right|. \tag{2.6.10}$$

i.e.,

$$\Delta_n = \max_{0\leq j\leq N}\sup_z\left|P\left\{\frac{\sum_{i=1}^n B_{j,1}(x) B_{j,1}(X_i)\sigma(X_i)\varepsilon_i}{n\sigma_{n,1}(x)} \leq z, x \in I_j\right\} - \Phi(z)\right|,$$

equations (2.6.9), (2.6.10) and Lemmas 2.6.3, 2.6.7 imply

$$\begin{aligned}\Delta_n &= \max_{0\leq j\leq N}\sup_z\left|P\left\{\sigma_{n,j}^{-1}\sum_{i=1}^n \xi_{i,j} \leq z, x \in I_j\right\} - \Phi(z)\right| \leq \frac{c_1 C_0(f,\sigma)}{h^{1/2} c_0(f,\sigma)\sigma_{n,j}} \\ &\leq \frac{C(f,\sigma)}{\sqrt{nh}} = o\left(N^{-1}\right),\end{aligned}$$

where the last step follows from Assumption (A3). Using the above, for a_N, b_N given in Lemma 2.6.4 and each $j = 1, ..., N$, one has

$$\begin{aligned}&P\left\{\left|\sigma_{n,j}^{-1}\sum_{i=1}^n \xi_{i,j}\right| \leq -\log(\alpha/2)/a_N + b_N, x \in I_j\right\} \\ &= P\left\{\sigma_{n,j}^{-1}\sum_{i=1}^n \xi_{i,j} \leq -\frac{\log(\alpha/2)}{a_N} + b_N, x \in I_j\right\} \\ &- P\left\{\sigma_{n,j}^{-1}\sum_{i=1}^n \xi_{i,j} \leq \frac{\log(\alpha/2)}{a_N} - b_N, x \in I_j\right\} \\ &= \Phi(-\log(\alpha/2)/a_N + b_N) - \Phi(\log(\alpha/2)/a_N - b_N) + o\left(N^{-1}\right).\end{aligned}$$

Applying Lemma 2.6.4, one easily obtains that as $n \to \infty$

$$\begin{aligned} \Phi(\tau/a_N + b_N) &= 1 - e^{-\tau}N^{-1} + o\left(N^{-1}\right), \\ \Phi(\tau/a_N + b_N) - \Phi(-\tau/a_N - b_N) &= 1 - 2e^{-\tau}N^{-1} + o\left(N^{-1}\right). \end{aligned}$$

Letting $2e^{-\tau} = \alpha$ or $\tau = -\log(\alpha/2)$ entails that uniformly in j,

$$P\left\{\left|\sigma_{n,j}^{-1}\sum_{i=1}^{n}\xi_{i,j}\right| \le -\frac{\log(\alpha/2)}{a_{N+1}} + b_{N+1}, x \in I_j\right\} = 1 - \frac{\alpha}{1+N} + o\left(N^{-1}\right).$$

Thus

$$\begin{aligned} & P\left\{\left|\sigma_{n,j}^{-1}\sum_{i=1}^{n}\xi_{i,j}\right| > -\frac{\log(\alpha/2)}{a_{N+1}} + b_{N+1}, x \in I_j, \text{ for some } 0 \le j \le N\right\} \\ \le & \sum_{j=0}^{N} P\left\{\left|\sigma_{n,j}^{-1}\sum_{i=1}^{n}\xi_{i,j}\right| > -\frac{\log(\alpha/2)}{a_{N+1}} + b_{N+1}, x \in I_j\right\} = \alpha + o(1). \end{aligned}$$

So as $n \to \infty$, one has

$$P\left\{\left|\sigma_{n,j}^{-1}\sum_{i=1}^{n}\xi_{i,j}\right| \le -\frac{\log(\alpha/2)}{a_{N+1}} + b_{N+1}, x \in I_j, 0 \le j \le N\right\} \ge 1 - \alpha + o(1).$$

Hence

$$\begin{aligned} & \liminf_{n\to\infty} P\left[\sup_{x\in[a,b]}\left|\sigma_{n,1}^{-1}(x)\hat{\varepsilon}_1(x)\right| \le \{2\log(N+1)\}^{1/2} d_n(\alpha)\right] \\ = & \liminf_{n\to\infty} P\left[\left|\sigma_{n,j}^{-1}\sum_{i=1}^{n}\xi_{i,j}\right| \le -\frac{\log(\alpha/2)}{a_{N+1}} + b_{N+1}, x \in I_j, 0 \le j \le N\right] \ge 1 - \alpha. \end{aligned}$$

Therefore, using Lemma 2.6.6, one has proved (2.3.7) for $k = 1$. □

2.6.3 Proof of Theorem 2.2.1 with $k = 1$

PROOF OF THEOREM 1 WITH $k = 1$. By (2.3.6) and Assumption (A3), one has

$$\|\tilde{m}_1(x) - m(x)\|_\infty = O_p(h) = o_p\left\{n^{-1/2}h^{-1/2}(\log(N+1))^{1/2}\right\},$$

so the uniform bias order is negligible compared to $(nh)^{-1/2}\{\log(N+1)\}^{1/2}$, which is the uniform noise order of

$$\sigma_{n,1}(x)\{-\log(\alpha/2)/a_{N+1} + b_{N+1}\} = \sigma_{n,1}(x)\{2\log(N+1)\}^{1/2} d_n(\alpha).$$

Now (2.3.5) and Proposition 2.3.1 yield the conservativity of the band in (2.2.12) for $k = 1$

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

Therefore, Theorem 2.2.1 has been proved for the case of $k = 1$. □

2.6.4 Preliminaries of Theorem 2.2.1 with $k = 2$

In this subsection we examine some matrices used in the construction of confidence band in (2.2.12) for $k = 2$. In what follows, $|\mathbf{T}|$ is used to denote the maximal absolute value of all the elements in matrix $\mathbf{T}$, $\mathbf{V}$ is the inner product matrix defined in (2.2.4) and $\mathbf{M}_{N+2}$ is the tridiagonal matrix as defined in (2.2.8).

Lemma 2.6.8. *Given matrix $\mathbf{\Omega} = \mathbf{M}_{N+2} + \mathbf{\Gamma}$, in which $\mathbf{\Gamma} = \left(\gamma_{jj'}\right)_{j,j'=-1}^{N}$ satisfies $\gamma_{jj'} \equiv 0$ if $\left|j - j'\right| > 1$ and $|\mathbf{\Gamma}| \xrightarrow{p} 0$. Then there exist constants $c, C > 0$ independent of n and $\mathbf{\Gamma}$, such that with probability approaching one*

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

Proof of the above lemma is trivial. As an application of Lemma 2.6.8, consider the matrix $\mathbf{S} = \mathbf{V}^{-1}$ defined in (2.2.5). 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

$$\sum_{j=-1}^{N} \left|s_{j'j}\right| \le \left|S\tilde{\xi}_{j'}\right| \le C_s \left|\tilde{\xi}_{j'}\right| = C_s, \forall j' = -1, 0, ..., N. \tag{2.6.11}$$

The next lemma follows by applying Lemma 2.6.8 with $\mathbf{\Omega} = \mathbf{M}_{N+2}$. It ensures that one can approximate $\mathbf{S}$ with the inverse of $\mathbf{M}_{N+2}$, with a simpler distribution-free form in (2.2.8). This approximation is uniform for $\mathbf{S}_j$ in (2.2.5) and Ξ_j in (2.2.7) as well.

Lemma 2.6.9. *As $n \to \infty$, $\left|\mathbf{M}_{N+2}^{-1} - \mathbf{S}\right| \to 0$ and $\max_{0\le j\le N} \left|\Xi_j - \mathbf{S}_j\right| \to 0$.*

The tridiagonal terms of the matrix $\mathbf{M}_{N+2}^{-1}$ can be computed through the following lemma, which is a direct result of Zhang (1999), Theorem 4.5, page 101.

LEMMA 2.6.10. *Let*

$$z_1 = \left(2+\sqrt{3}\right)/4,\ z_2 = \left(2-\sqrt{3}\right)/4,\ \theta = z_2/z_1 = 7-4\sqrt{3},$$

one can compute the terms $l_{i,k} = l_{k,i}, |i-k| \le 1$ *defined in (2.2.8) by the following formulae*

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

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

for $2 \le k \le N+1$.

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

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

for $2 \le k \le N$. *In particular, there exists a constant* $c_l > 0$ *such that* $\max\limits_{|i-k|\le 1} |l_{ik}| \le c_l$.

2.6.5 Variance calculation

We examine the behavior of $\tilde{\varepsilon}_2(x)$ defined in (2.3.4), rewritten as

$$\tilde{\varepsilon}_2(x) = \sum_{j=-1}^{N} \tilde{a}_j B_{j,2}(x), x \in [a,b], \tag{2.6.12}$$

where the spline coefficient vector $\tilde{\mathbf{a}} = (\tilde{a}_{-1}, \cdots, \tilde{a}_N)^T$ according to (2.3.4) is

$$(\mathbf{V}+\mathbf{V}^*)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} B_{j,2}(X_i)\,\sigma(X_i)\,\varepsilon_i\right)_{j=-1}^{N},\ \mathbf{V}^* = \left(\langle B_{j,2}, B_{j',2}\rangle_n\right)_{j,j'=-1}^{N} - \mathbf{V},$$

where $\mathbf{V}^*$, the difference between empirical and theoretical inner product matrices, satisfies $|\mathbf{V}^*| \le A_{n,2} = O_p\left\{(nh)^{-1/2}\log^{1/2}(n)\right\}$ by (2.3.2).

Now define $\hat{\mathbf{a}} = (\hat{a}_{-1}, \cdots, \hat{a}_N)^T$ by replacing $(\mathbf{V}+\mathbf{V}^*)^{-1}$ with $\mathbf{V}^{-1} = \mathbf{S}$ in above formula, i.e.

$$\hat{\mathbf{a}} = \mathbf{S}\left\{\frac{1}{n}\sum_{i=1}^{n} B_{j,2}(X_i)\,\sigma(X_i)\,\varepsilon_i\right\}_{j=-1}^{N} = \left\{\sum_{j=-1}^{N} s_{j'j}\frac{1}{n}\sum_{i=1}^{n} B_{j,2}(X_i)\,\sigma(X_i)\,\varepsilon_i\right\}_{j'=-1}^{N},$$

and define, with $j(x)$ in (2.2.1) and

$$\begin{aligned}\hat{\varepsilon}_2(x) &= \sum_{j=-1}^{N} \hat{a}_j B_{j,2}(x) = \sum_{j,j'=-1}^{N} s_{j'j} \frac{1}{n} \sum_{i=1}^{n} B_{j,2}(X_i)\,\sigma(X_i)\,\varepsilon_i B_{j',2}(x) \\ &= \sum_{j'=j(x)-1,j(x)} B_{j',2}(x) \sum_{j=-1}^{N} s_{j',j} \frac{1}{n} \sum_{i=1}^{n} B_{j,2}(X_i)\sigma(X_i)\varepsilon_i \end{aligned} \tag{2.6.13}$$

for $x \in [a,b]$. Next define an $(N+2)$-vector $\mathbf{U}$

$$\mathbf{U} \equiv \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} B_{j,2}(X_i)\sigma(X_i)\varepsilon_i \right\}_{j=-1}^{N}, \tag{2.6.14}$$

and 2-vectors $\{\mathbf{\Lambda}_j\}_{j=0}^{N}$

$$\mathbf{\Lambda}_j = \begin{pmatrix} \Lambda_{j1} \\ \Lambda_{j2} \end{pmatrix} \equiv \tilde{\mathbf{S}}_j \mathbf{U} = \frac{1}{\sqrt{n}} \begin{pmatrix} \sum_{i=1}^{n} \sum_{j'=-1}^{N} s_{j-1,j'} B_{j',2}(X_i)\sigma(X_i)\varepsilon_i \\ \sum_{i=1}^{n} \sum_{j'=-1}^{N} s_{j,j'} B_{j',2}(X_i)\sigma(X_i)\varepsilon_i \end{pmatrix}, \tag{2.6.15}$$

in which the $(j-1)$-th and j-th rows of the matrix $\mathbf{S}$ is denoted as an $2 \times (N+2)$ matrix

$$\tilde{\mathbf{S}}_j = \begin{pmatrix} s_{j-1,-1} & s_{j-1,0} & \cdots & s_{j-1,N} \\ s_{j,-1} & s_{j,0} & \cdots & s_{j,N} \end{pmatrix}, \ 0 \le j \le N. \tag{2.6.16}$$

Then, one can write $\hat{\varepsilon}_2(x)$ in the following matrix form

$$\hat{\varepsilon}_2(x) = \mathbf{D}^T(x)\,\mathbf{\Lambda}_{j(x)},\ x \in [a,b], \tag{2.6.17}$$

in which the function $\mathbf{D}(x)$ is a 2-vector such that

$$\mathbf{D}(x) \equiv \left\{ D_{j(x)-1}(x), D_{j(x)}(x) \right\}^T, D_j(x) \equiv n^{-1/2} B_{j,2}(x), -1 \le j \le N. \tag{2.6.18}$$

The next lemma provides the pointwise variance of $\hat{\varepsilon}_2(x)$.

LEMMA 2.6.11. *The pointwise variance of $\hat{\varepsilon}_2(x)$ is the function $\sigma_{n,2}^2(x)$ defined in (2.2.10), which satisfies*

$$E\left\{\hat{\varepsilon}_2^2(x)\right\} \equiv \sigma_{n,2}^2(x) = \frac{3\sigma^2(x)}{2f(x)nh} \mathbf{\Delta}^T(x)\,\mathbf{S}_{j(x)} \mathbf{\Delta}(x) \left\{1 + r_{n,2}(x)\right\}, \tag{2.6.19}$$

with $\sup_{x\in[a,b]} |r_{n,2}(x)| \to 0$, $j(x)$ is as defined in (2.2.1), $\mathbf{\Delta}(x)$ as defined in (2.2.7) and matrix $\mathbf{S}_j$ in (2.2.5). Consequently, there exist positive constants c_σ and C_σ such that for large enough n

$$c_\sigma (nh)^{-1/2} \le \sigma_{n,2}(x) \le C_\sigma (nh)^{-1/2}, \forall x \in [a,b]. \tag{2.6.20}$$

PROOF. From (2.6.15) and (2.6.17), one has

$$E\left\{\hat{\varepsilon}_2^2(x)\right\} = \mathbf{D}^T(x)\operatorname{cov}\left(\mathbf{\Lambda}_{j(x)}\right)\mathbf{D}(x) = \mathbf{D}^T(x)\,\tilde{\mathbf{S}}_{j(x)}\operatorname{cov}(\mathbf{U})\,\tilde{\mathbf{S}}_{j(x)}^T\mathbf{D}(x).$$

Note that $E(\varepsilon_i|X_i) = 0$, $E\left[B_{j,2}(X_i)B_{l,2}(X_k)\sigma(X_i)\sigma(X_k)\varepsilon_i\varepsilon_k\right] = 0, \forall i \neq k$, the jl-th entry of the covariance matrix of $\mathbf{U}$ defined by (2.6.14) is

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{n} E\left\{B_{j,2}(X_i)B_{l,2}(X_k)\sigma(X_i)\sigma(X_k)\varepsilon_i\varepsilon_k\right\}\\ = \;&\frac{1}{n}\sum_{i=1}^{n} E\left\{B_{j,2}(X_i)B_{l,2}(X_i)\sigma^2(X_i)\right\} = \int \sigma^2(v)\,B_{j,2}(v)\,B_{l,2}(v)\,f(v)\,dv = \sigma_{jl},\end{aligned}$$

which is the jl-th entry of the matrix $\mathbf{\Sigma}$ defined in (2.2.6), i.e., $\operatorname{cov}(\mathbf{U}) = \mathbf{\Sigma}$. The rest of the proof is simple algebra. □

2.6.6 Proof of Theorem 2.2.1 with $k = 2$

Prior to the proof Theorem, we introduce some notation. First we define 2-vectors $\{\mathbf{Z}_j\}_{j=0}^N$

$$\mathbf{Z}_j \equiv (Z_{j1}, Z_{j2}) = \mathbf{\Lambda}_j^T\left\{\operatorname{cov}\left(\mathbf{\Lambda}_j\right)\right\}^{-1/2} = \begin{pmatrix} \beta_{11}^{(j)}\Lambda_{j1} + \beta_{12}^{(j)}\Lambda_{j2} \\ \beta_{12}^{(j)}\Lambda_{j1} + \beta_{22}^{(j)}\Lambda_{j2} \end{pmatrix}, \qquad (2.6.21)$$

where denote

$$\left\{\operatorname{cov}\left(\mathbf{\Lambda}_j\right)\right\}^{-1/2} \equiv \begin{pmatrix} \beta_{11}^{(j)} & \beta_{12}^{(j)} \\ \beta_{12}^{(j)} & \beta_{22}^{(j)} \end{pmatrix}. \qquad (2.6.22)$$

Then it is clear that $\operatorname{var}(\mathbf{Z}_j) = \mathbf{I}$, $\operatorname{var}(Z_{j\gamma}) = 1, \gamma = 1, 2$, for any $j = 0, ..., N$.

The covariance matrix of $\mathbf{\Lambda}_j$ approximates $\sigma^2(t_{j+1})\mathbf{S}_j$ defined in (2.2.5) uniformly.

LEMMA 2.6.12. *For* $\{\mathbf{\Lambda}_j\}_{j=0}^N$ *defined in (2.6.15) and matrix* $\mathbf{S}_j$ *defined in (2.2.5), one has*

$$\operatorname{cov}\left(\mathbf{\Lambda}_j\right) = \sigma^2(t_{j+1})\mathbf{S}_j + \tilde{\mathbf{R}}_j, 0 \leq j \leq N, \lim_{n\to\infty}\max_{0\leq j\leq N}\left|\tilde{\mathbf{R}}_j\right| = 0.$$

PROOF. Since $\mathbf{\Lambda}_j = \tilde{\mathbf{S}}_j\mathbf{U}$ with $\tilde{\mathbf{S}}_j$ defined in (2.6.16) and $\operatorname{cov}(\mathbf{U}) = \mathbf{\Sigma}$ as in the proof of Lemma 2.6.11. Thus the covariance matrix of $\mathbf{\Lambda}_j$ is

$$\operatorname{cov}\left(\mathbf{\Lambda}_j\right) = \tilde{\mathbf{S}}_j\mathbf{\Sigma}\tilde{\mathbf{S}}_j^T = \begin{pmatrix} \sum_{k,l=-1}^N s_{j-1,k}s_{j-1,l}\sigma_{kl} & \sum_{k,l=-1}^N s_{j,k}s_{j-1,l}\sigma_{kl} \\ \sum_{k,l=-1}^N s_{j-1,k}s_{j,l}\sigma_{kl} & \sum_{k,l=-1}^N s_{j,k}s_{j,l}\sigma_{kl} \end{pmatrix}.$$

By Assumption (A2), (2.6.16) and (2.2.6)

$$\sigma_{kl} = \int \sigma^2(v)\,B_{k,2}(v)\,B_{l,2}(v)\,f(v)\,dv = \sigma^2(t_{k+1})\,v_{kl} + cw\left(f\sigma^2, h\right).$$

Similarly, one also has $\sigma_{kl} = \sigma^2(t_{l+1}) v_{kl} + cw(f\sigma^2, h)$. Thus

$$\operatorname{cov}(\boldsymbol{\Lambda}_j) = \sum_{k,l=-1}^{N} \begin{pmatrix} s_{j-1,k}s_{j-1,l}v_{kl}\sigma^2(t_{l+1}) & s_{j,k}s_{j-1,l}v_{kl}\sigma^2(t_{l+1}) \\ s_{j-1,k}s_{j,l}v_{kl}\sigma^2(t_{k+1}) & s_{j,k}s_{j,l}v_{kl}\sigma^2(t_{k+1}) \end{pmatrix} + \tilde{\mathbf{R}}_j^*,$$

where

$$\tilde{\mathbf{R}}_j^* = cw\left(f\sigma^2, h\right) \begin{pmatrix} \sum_{k,l=-1}^{N} s_{j-1,k}s_{j-1,l} & \sum_{k,l=-1}^{N} s_{j,k}s_{j-1,l} \\ \sum_{k,l=-1}^{N} s_{j-1,k}s_{j,l} & \sum_{k,l=-1}^{N} s_{j,k}s_{j,l} \end{pmatrix}.$$

Note that $\sum_{k=-1}^{N} s_{j,k}v_{kl} = 0$ if $l \neq j$ and $\sum_{k=-1}^{N} s_{j,k}v_{kl} = 1$ if $l = j$, thus

$$\operatorname{cov}(\boldsymbol{\Lambda}_j) = \begin{pmatrix} s_{j-1,j}\sigma^2(t_j) & s_{j-1,j}\sigma^2(t_{j+1}) \\ s_{j-1,j}\sigma^2(t_{j+1}) & s_{j,j}\sigma^2(t_{j+1}) \end{pmatrix} + \tilde{\mathbf{R}}_j^* = \sigma^2(t_{j+1})\mathbf{S}_j + \tilde{\mathbf{R}}_j.$$

By (2.6.11), $\max_{0\le j\le N} \left|\tilde{\mathbf{R}}_j\right| \le Cw\left(f\sigma^2, h\right) \to 0$, as $n \to \infty$. □

LEMMA 2.6.13. *For the matrices $\Xi_j^{-1/2}$ defined in (2.2.7)*

$$\lim_{n\to\infty} \max_{0\le j\le N} \left|\Xi_j^{-1/2} - \sigma(t_{j+1})\left\{\operatorname{cov}(\boldsymbol{\Lambda}_j)\right\}^{-1/2}\right| = 0. \tag{2.6.23}$$

PROOF. Note that $\Xi_j^{-1/2}$, $\left\{\operatorname{cov}(\boldsymbol{\Lambda}_j)\right\}^{-1/2}$are symmetric matrices and using the following fact for symmetric matrices $\mathbf{A}$ and $\mathbf{B}$

$$\begin{aligned} c\left|\mathbf{A}^{-1/2} - \mathbf{B}^{-1/2}\right| &= c \max_{i=1,2}\left|\left(\mathbf{A}^{-1/2} - \mathbf{B}^{-1/2}\right)e_i\right| \\ &\le \max_{i=1,2}\left|\left(\mathbf{B}\mathbf{A}^{1/2} + \mathbf{A}\mathbf{B}^{1/2}\right)\left(\mathbf{A}^{-1/2} - \mathbf{B}^{-1/2}\right)e_i\right| = |\mathbf{B} - \mathbf{A}|, \end{aligned}$$

together with Lemma 2.6.12, one has

$$\begin{aligned} c\left|\Xi_j^{-1/2} - \sigma(t_{j+1})\left\{\operatorname{cov}(\boldsymbol{\Lambda}_j)\right\}^{-1/2}\right| &\le \left|\sigma^{-2}(t_{j+1})\operatorname{cov}(\boldsymbol{\Lambda}_j) - \Xi_j\right| \\ &\le \left|\mathbf{S}_j - \Xi_j\right| + \left|\sigma^{-2}(t_{j+1})\operatorname{cov}(\boldsymbol{\Lambda}_j) - \mathbf{S}_j\right| = \left|\mathbf{S}_j - \Xi_j\right| + \sigma^{-2}(t_{j+1})\tilde{\mathbf{R}}_j. \end{aligned}$$

The desired result follows from Lemma 2.6.9. □

LEMMA 2.6.14. *Under Assumptions (A1)-(A5), for the variables $Z_{j\gamma}$, $\gamma = 1, 2$, $0 \le j \le N$, defined in (2.6.21), one has*

$$\max_{\gamma=1,2} \limsup_{n\to\infty} P\left[\max_{0\le j\le N}\left\{Z_{j\gamma}^2\right\} > 2\left\{\log(N+1)\right\}\left\{d_n(\alpha/2)\right\}^2\right] \le \alpha/2. \tag{2.6.24}$$

PROOF. Without loss of generality, we prove (2.6.24) only for $\gamma = 1$.

$$
\begin{aligned}
& P\left[\max_{0\le j\le N}\left\{Z_{j1}^2\right\} > 2\left\{\log\left(N+1\right)\right\}\left\{d_n\left(\alpha/2\right)\right\}^2\right] \\
= \ & P\left[\max_{0\le j\le N}\left\{Z_{j1}\right\} > \left\{2\log\left(N+1\right)\right\}^{1/2} d_n\left(\alpha/2\right)\right],
\end{aligned}
$$

where, according to (2.6.21) and (2.6.22)

$$
Z_{j1} = \beta_{11}^{(j)}\Lambda_{j1} + \beta_{12}^{(j)}\Lambda_{j2} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{k=-1}^{N}\left(\beta_{11}^{(j)}s_{j-1,k} + \beta_{12}^{(j)}s_{j,k}\right)B_{k,2}(X_i)\sigma(X_i)\varepsilon_i.
$$

Let $\zeta_{i,j} = \sum_{k=-1}^{N}\left(\beta_{11}^{(j)}s_{j-1,k} + \beta_{12}^{(j)}s_{j,k}\right)B_{k,2}(X_i)\sigma(X_i)\varepsilon_i$, $j = 0, \ldots, N$, $i = 1, \ldots, n$, then

$$
\sqrt{n}Z_{j1} = S_n = \sum_{i=1}^{n}\zeta_{i,j},\ E\left(\sqrt{n}Z_{j1}\right)^2 = nEZ_{j1}^2 = n.
$$

So one only needs to find a bound for $E\left|\zeta_{1,j}\right|^3$ in order to apply Lemma 2.6.3 to S_n. By the boundedness of $\max_{0\le j\le N}\left|\Xi_j\right|$, (2.6.3) and (2.6.23)

$$
\begin{aligned}
E\left|\zeta_{1,j}\right|^3 &= E\left[\left|\sum_{k=-1}^{N}\left(\beta_{11}^{(j)}s_{j-1,k} + \beta_{12}^{(j)}s_{j,k}\right)B_{k,2}(X_1)\right|^3\sigma^3(X_1)\left|\varepsilon_1^3\right|\right] \\
&\le M_0E\left[\left|\sum_{k=-1}^{N}\left(\beta_{11}^{(j)}s_{j-1,k} + \beta_{12}^{(j)}s_{j,k}\right)B_{k,2}(X_1)\right|^3\sigma^3(X_1)\right] \le C(f,\sigma)h^{-1/2}.
\end{aligned}
$$

Lemma 2.6.3 entails that $\Delta_n = o\left(n^{-1/2}h^{-1/2}\right) = o\left(N^{-2}\right)$, in which Δ_n is

$$
\max_{0\le j\le N}\sup_z\left|P\left\{Z_{j1}\le z\right\} - \Phi\left(z\right)\right| = \max_{0\le j\le N}\sup_z\left|P\left\{n^{-1/2}\sum_{i=1}^{n}\zeta_{i,j}\le z\right\} - \Phi\left(z\right)\right|.
$$

By Lemma 2.6.4, one has uniformly in j

$$
P\left[\left|Z_{j1}\right| \le \left\{2\log\left(N+1\right)\right\}^{1/2}d_n\left(\alpha/2\right)\right] = 1 - \frac{\alpha}{2\left(N+1\right)} + o\left(N^{-1}\right).
$$

Therefore

$$
\begin{aligned}
& P\left[\max_{0\le j\le N}\left|Z_{j1}\right| > \left\{2\log\left(N+1\right)\right\}^{1/2}d_n\left(\alpha/2\right)\right] \\
\le \ & \sum_{j=0}^{N}P\left[\left|Z_{j1}\right| > \left\{2\log\left(N+1\right)\right\}^{1/2}d_n\left(\alpha/2\right)\right] \\
= \ & \sum_{j=0}^{N}\left[1 - \left\{1 - \frac{\alpha}{2\left(N+1\right)}\right\}\right] + o\left(1\right) = \alpha/2 + o\left(1\right).
\end{aligned}
$$

Hence

$$\limsup_{n\to\infty} P\left[\max_{0\le j\le N}\left\{Z_{j1}^2\right\} > 2\left\{\log(N+1)\right\}\left\{d_n(\alpha/2)\right\}^2\right] = \alpha/2. \qquad \square$$

LEMMA 2.6.15. *For a given $0 < \alpha < 1$, and $\sigma_{n,2}(x)$ as given in (2.2.10)*

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

PROOF. Note that $\hat{\varepsilon}_2(x) = \mathbf{D}^T(x)\,\mathbf{\Lambda}_{j(x)}$, where $\mathbf{D}(x)$ and $\mathbf{\Lambda}_{j(x)}$ are defined in (2.6.18) and (2.6.15). Thus, standardization leads to

$$\left\{\sigma_{n,2}^{-1}(x)\hat{\varepsilon}_2(x)\right\}^2 = \frac{\left\{\sigma_{n,2}^{-1}(x)\mathbf{D}(x)\right\}^T \mathbf{\Lambda}_{j(x)}\mathbf{\Lambda}_{j(x)}^T\left\{\sigma_{n,2}^{-1}(x)\mathbf{D}(x)\right\}}{\left\{\sigma_{n,2}^{-1}(x)\mathbf{D}(x)\right\}^T \operatorname{cov}\left(\mathbf{\Lambda}_{j(x)}\right)\left\{\sigma_{n,2}^{-1}(x)\mathbf{D}(x)\right\}}. \tag{2.6.25}$$

Define for any $j = 0, ..., N$, $\mathbf{Q}_j = \mathbf{\Lambda}_j^T\left\{\operatorname{cov}\left(\mathbf{\Lambda}_j\right)\right\}^{-1}\mathbf{\Lambda}_j = \mathbf{Z}_j\mathbf{Z}_j^T = \sum_{\gamma=1,2} Z_{j\gamma}^2$. The maximization lemma of Johnson and Wichern (1992), page 166, ensures that for any $x \in [a,b]$

$$\frac{\left\{\frac{\mathbf{D}(x)}{\sigma_{n,2}(x)}\right\}^T \mathbf{\Lambda}_{j(x)}\mathbf{\Lambda}_{j(x)}^T\left\{\frac{\mathbf{D}(x)}{\sigma_{n,2}(x)}\right\}}{\left\{\frac{\mathbf{D}(x)}{\sigma_{n,2}(x)}\right\}^T \operatorname{cov}\left(\mathbf{\Lambda}_{j(x)}\right)\left\{\frac{\mathbf{D}(x)}{\sigma_{n,2}(x)}\right\}} \le \mathbf{\Lambda}_{j(x)}^T\left\{\operatorname{cov}\left(\mathbf{\Lambda}_{j(x)}\right)\right\}^{-1}\mathbf{\Lambda}_{j(x)} = \mathbf{Q}_{j(x)},$$

which together with (2.6.25) entails that $\sup_{x\in[a,b]}\left|\sigma_{n,2}^{-1}(x)\hat{\varepsilon}_2(x)\right|^2 \le \max_{0\le j\le N}\mathbf{Q}_j$. Thus (2.6.24) implies

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

$$\liminf_{n\to\infty} P\left[\max_{0\le j\le N}\mathbf{Q}_j \le 4\left\{\log(N+1)\right\}\left\{d_n(\alpha/2)\right\}^2\right] \ge$$

$$1-\sum_{\gamma=1,2}\limsup_{n\to\infty} P\left[\max_{0\le j\le N}\left\{Z_{j\gamma}^2\right\} > 2\left\{\log(N+1)\right\}\left\{d_n(\alpha/2)\right\}^2\right] \ge 1-\alpha/2\times 2 = 1-\alpha. \qquad \square$$

The next lemma's proof follows from Lemma 2.6.8, (2.6.12), (2.6.13), (2.3.2) and (2.6.20).

LEMMA 2.6.16. *Under Assumptions (A3) and (A5), one has*

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

PROOF OF PROPOSITION 2.3.1 WITH $k = 2$. It follows from Lemmas 2.6.15 and 2.6.16 automatically. □

PROOF OF THEOREM 2.2.1 WITH $k = 2$. Note that equation (2.3.6) implies that $\|\tilde{m}_2(x) - m(x)\|_\infty = O_p(h^2)$, hence

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

which implies that the bias order is negligible compared to the noise order. Applying (2.3.7) with $k = 2$ in Proposition 2.3.1

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

CHAPTER 3

Spline-Backfitted Kernel Smoothing of NAAR Models

3.1 Introduction

For the past two decades, various non- and semiparametric regression techniques have been developed for the analysis of nonlinear time series; see, for example, Robinson (1983), Tjøstheim and Auestad (1994), Huang and Yang (2004), to name one article representative of each decade. Application to high dimensional time series data, however, has been hampered due to the scarcity of smoothing tools that are not only computationally expedient but also theoretically reliable. This has motivated the proposed procedures of this chapter.

For the NAAR model in (1.3.1), estimators of the unknown component functions $\{m_\alpha(\cdot)\}_{\alpha=1}^d$ are proposed based on a geometrically strong mixing sample $\{Y_i, X_{i,1}, ..., X_{i,d}\}_{i=1}^n$. If the data were actually i.i.d. observations instead of a time series realization, many methods would be available for estimating $\{m_\alpha(\cdot)\}_{\alpha=1}^d$. For instance, there are four types kernel-based estimators: the classic backfitting estimators (CBE) of Hastie and Tibshirani (1990), Opsomer and Ruppert (1997); marginal integration estimators (MIE) of Linton and Nielsen (1995), Linton and Härdle (1996), Fan, Härdle and Mammen (1998), Sperlich, Tjøstheim and Yang (2002), Yang, Sperlich and Härdle (2003) and a kernel based method of estimating rate to optimality of Hengartner and Sperlich (2005); the smoothing backfitting estimators (SBE) of Mammen, Linton and Nielsen (1999); and the two-stage estimators, such as one step backfitting of the integration estimators of Linton (1997), one step backfitting of the projection estimators of Horowitz, Klemmelä and Mammen (2006), and one Newton step from the nonlinear LSE estimators of Horowitz and Mammen (2004). For the spline estimators, see Stone (1985), (1994), Huang (1998), and Xue and Yang (2006 b).

In time series context, however, there are fewer theoretically justified methods due to the additional difficulty posed by dependence in data. Some of these are: the kernel estimators via marginal integration of Tjøstheim and Auestad (1994), Yang, Härdle and Nielsen (1999); and the spline estimators of Huang and Yang (2004). In addition, Xue and Yang (2006 a) have extended the marginal integration kernel estimator and spline estimator to additive coefficient models for weakly dependent data. All of these existing methods are unsatisfactory in regard to either the computational or the theoretical issue. The existing kernel methods are too computationally intensive for high dimension d, thus limiting their applicability to small number of predictors. Spline methods, on the other hand, provide only convergence rates but no asymptotic distributions, so no measures of confidence can be assigned to the estimators.

If the last $d-1$ component functions were known by "oracle", one could create $\{Y_{i,1}, X_{i,1}\}_{i=1}^{n}$ with $Y_{i,1} = Y_i - c - \sum_{\alpha=2}^{d} m_\alpha (X_{i,\alpha}) = m_1 (X_{i,1}) + \sigma (X_{i,1}, ..., X_{i,d}) \varepsilon_i$, from which one could compute an "oracle smoother" to estimate the only unknown function $m_1 (x_1)$, thus effectively bypassing the "curse of dimensionality". The idea of Linton (1997) was to obtain an approximation to the unobservable variables $Y_{i,1}$ by replacing $m_\alpha (X_{i,\alpha}), i = 1, ..., n, \alpha = 2, ..., d$ with marginal integration kernel estimates and arguing that the error incurred by this "cheating" is of smaller magnitude than the rate $O\left(n^{-2/5}\right)$ for estimating function $m_1 (x_1)$ from the unobservable data. The procedure of Linton (1997) is modified by substituting $m_\alpha (X_{i,\alpha}), i = 1, ..., n, \alpha = 2, ..., d$ with spline estimators, specifically, a two-stage estimation procedure is proposed: first one pre-estimates $\{m_\alpha (x_\alpha)\}_{\alpha=2}^{d}$ by its pilot estimator through an under smoothed centered standard spline procedure, next one constructs the pseudo response $\hat{Y}_{i,1}$ and approximates $m_1 (x_1)$ by its Nadaraya-Watson estimator as given in (3.2.12).

The above proposed spline-backfitted kernel (SPBK) estimation method has several advantages compared to most of the existing methods. Firstly, as Sperlich, Tjøstheim and Yang (2002) mentioned, Linton (1997) mixed up different projections, making it uninterpretable if the real data generating process deviates from additivity. While the projections in both steps here are with respect to the same measure. Secondly, since our pilot spline estimator is thousands of times faster than the pilot kernel estimators in Linton (1997), the proposed method is computationally expedient, see Table 4.4. Thirdly, the SPBK estimator can be shown as efficient as the "oracle smoother" uniformly over any compact range, whereas Linton (1997) proved such "oracle efficiency" only at a single point. Moreover, the regularity conditions considered here are natural and appealing and close to being the minimal compared to the papers mentioned above. In contrast, higher order smoothness is needed with growing dimensionality of the regressors in Linton and Nielsen (1995). Stronger and more obscure conditions are assumed for the two-stage estimation proposed by Horowitz and Mammen (2004).

The SPBK estimator achieves its seemingly surprising success by borrowing the strengths of both spline and kernel: spline does a quick initial estimation of all additive components and removes them all except the one of interest; kernel smoothing is then applied to the cleaned univariate data to estimate with asymptotic distribution. Propositions 3.4.1 and 3.5.1 are the keys in understanding the proposed estimators' uniform oracle efficiency. They accomplish the well-known "reducing bias by undersmoothing" in the first step using spline and "averaging out the variance" in the second step with kernel, both steps taking advantage of the joint asymptotics of kernel and spline functions, which is the new feature of the proofs here.

Fan and Jiang (2005) provides generalized likelihood ratio (GLR) tests for additive models using the backfitting estimator. Similar GLR test based on the SPBK estimator is feasible for future research.

The rest of the chapter is organized as follows. Section 3.2 introduces the SPBK estimator, and states its asymptotic "oracle efficiency" under appropriate assumptions. Section 3.3 provides some insights into the ideas behind the proofs of the main theoretical results, by decomposing the estimator's "cheating" error into a bias and a variance part. Section 3.4 shows the uniform order of the bias term. Section 3.5 shows the uniform order of the variance term. Section 3.6 presents Monte Carlo results to demonstrate that the SPBK estimator does indeed possess the claimed asymptotic properties. All technical proofs are contained in Section 3.7.

3.2 The SPBK estimator

In this section, a spline-backfitted kernel estimation procedure is proposed. For convenience, denote vectors as $\mathbf{x} = (x_1, ..., x_d)$ and take $\|\cdot\|$ as the usual Euclidian norm on R^d such that $\|\mathbf{x}\| = \sqrt{\sum_{\alpha=1}^{d} x_\alpha^2}$, and $\|\cdot\|_\infty$ the sup norm, $\|\mathbf{x}\|_\infty = \sup_{1 \leq \alpha \leq d} |x_\alpha|$. In what follows, let Y_i and $\mathbf{X}_i = (X_{i,1}, ..., X_{i,d})^T$ be the ith response and predictor vector. Denote $\mathbf{Y} = (Y_1, ..., Y_n)^T$ the response vector and $(\mathbf{X}_1, ..., \mathbf{X}_n)^T$ the design matrix.

Assume that the predictor X_α is distributed on a compact interval $[a_\alpha, b_\alpha], \alpha = 1, ..., d$. Without loss of generality, all intervals $[a_\alpha, b_\alpha] = [0, 1], \alpha = 1, ..., d$. We pre-select an integer $N = N_n \sim n^{2/5} \log n$, see Assumption (B6) below. For any $\alpha = 1, ..., d$, the constant B-spline function in (1.5.2) can be rewritten as the indicator function $I_{J,\alpha}(x_\alpha)$ of the $(N+1)$ equally-spaced subintervals of the finite interval $[0,1]$ with length $H = H_n = (N+1)^{-1}$, that is

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

Define the following centered spline basis

$$b_{J,\alpha}(x_\alpha) = I_{J+1,\alpha}(x_\alpha) - \frac{\|I_{J+1,\alpha}\|_2}{\|I_{J,\alpha}\|_2} I_{J,\alpha}(x_\alpha), \forall \alpha = 1, ..., d, J = 1, ..., N, \tag{3.2.2}$$

with the standardized version given for any $\alpha = 1, ..., d$,

$$B_{J,\alpha}(x_\alpha) = \frac{b_{J,\alpha}(x_\alpha)}{\|b_{J,\alpha}\|_2}, \forall J = 1, ..., N. \tag{3.2.3}$$

Define next the $(1 + dN)$-dimensional space $G = G[0,1]$ of additive spline functions as the linear space spanned by $\{1, B_{J,\alpha}(x_\alpha), \alpha = 1, ..., d, J = 1, ..., N\}$, while denote by $G_n \subset R^n$ spanned by $\left\{1, \{B_{J,\alpha}(X_{i,\alpha})\}_{i=1}^n, \alpha = 1, ..., d, J = 1, ..., N\right\}$. 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 G so that the vector $\{\hat{m}(\mathbf{X}_1), ..., \hat{m}(\mathbf{X}_n)\}^T$ best approximates the response vector $\mathbf{Y}$. To be precise

$$\hat{m}(\mathbf{x}) = \hat{\lambda}'_0 + \sum_{\alpha=1}^{d}\sum_{J=1}^{N} \hat{\lambda}'_{J,\alpha} I_{J,\alpha}(x_\alpha), \tag{3.2.4}$$

where the coefficients $\left(\hat{\lambda}'_0, \hat{\lambda}'_{1,1}, ..., \hat{\lambda}'_{N,d}\right)$ are solutions of the least squares problem

$$\left\{\hat{\lambda}'_0, \hat{\lambda}'_{1,1}, ..., \hat{\lambda}'_{N,d}\right\}^T = \text{argmin}_{R^{dN+1}} \sum_{i=1}^{n} \left\{ Y_i - \lambda_0 - \sum_{\alpha=1}^{d}\sum_{J=1}^{N} \lambda_{J,\alpha} I_{J,\alpha}(X_{i,\alpha}) \right\}^2 .$$

Simple linear algebra shows that

$$\hat{m}(\mathbf{x}) = \hat{\lambda}_0 + \sum_{\alpha=1}^{d}\sum_{J=1}^{N} \hat{\lambda}_{J,\alpha} B_{J,\alpha}(x_\alpha), \tag{3.2.5}$$

where $\left(\hat{\lambda}_0, \hat{\lambda}_{1,1}, ..., \hat{\lambda}_{N,d}\right)$ are solutions of the following least squares problem

$$\left\{\hat{\lambda}_0, \hat{\lambda}_{1,1}, ..., \hat{\lambda}_{N,d}\right\}^T = \text{argmin}_{R^{dN+1}} \sum_{i=1}^{n} \left\{ Y_i - \lambda_0 - \sum_{\alpha=1}^{d}\sum_{J=1}^{N} \lambda_{J,\alpha} B_{J,\alpha}(X_{i,\alpha}) \right\}^2 , \tag{3.2.6}$$

while (3.2.4) is used for data analytic implementation, the mathematically equivalent expression (3.2.5) is convenient for asymptotic analysis.

The pilot estimators of each component function and the constant are

$$\begin{aligned} \hat{m}_\alpha(x_\alpha) &= \sum_{J=1}^{N} \hat{\lambda}_{J,\alpha} B_{J,\alpha}(x_\alpha) - n^{-1}\sum_{i=1}^{n}\sum_{J=1}^{N} \hat{\lambda}_{J,\alpha} B_{J,\alpha}(X_{i,\alpha}), \\ \hat{m}_c &= \hat{\lambda}_0 + n^{-1}\sum_{\alpha=1}^{d}\sum_{i=1}^{n}\sum_{J=1}^{N} \hat{\lambda}_{J,\alpha} B_{J,\alpha}(X_{i,\alpha}). \end{aligned} \tag{3.2.7}$$

These pilot estimators are then used to define new pseudo-responses $\hat{Y}_{i,1}$, which are estimates of the unobservable "oracle" responses $Y_{i,1}$. Specifically,

$$\hat{Y}_{i,1} = Y_i - \hat{c} - \sum_{\alpha=2}^{d} \hat{m}_\alpha(X_{i,\alpha}), Y_{i,1} = Y_i - c - \sum_{\alpha=2}^{d} m_\alpha(X_{i,\alpha}), \tag{3.2.8}$$

where $\hat{c} = \overline{Y}_n = n^{-1}\sum_{i=1}^{n} Y_i$, which is a $\sqrt{n}$-consistent estimator of c by Central Limit Theorem. Next, define the spline-backfitted kernel (SPBK) estimator of $m_1(x_1)$ as $\hat{m}^*_1(x_1)$

based on $\left\{\hat{Y}_{i,1}, X_{i,1}\right\}_{i=1}^{n}$, which attempts to mimic the would-be Nadaraya-Watson estimator $\tilde{m}_1^*(x_1)$ of $m_1(x_1)$ based on $\{Y_{i,1}, X_{i,1}\}_{i=1}^{n}$ if the unobservable "oracle" responses $\{Y_{i,1}\}_{i=1}^{n}$ were available

$$\hat{m}_1^*(x_1) = \frac{\sum_{i=1}^{n} K_h(X_{i,1} - x_1)\hat{Y}_{i,1}}{\sum_{i=1}^{n} K_h(X_{i,1} - x_1)}, \tilde{m}_1^*(x_1) = \frac{\sum_{i=1}^{n} K_h(X_{i,1} - x_1) Y_{i,1}}{\sum_{i=1}^{n} K_h(X_{i,1} - x_1)}, \qquad (3.2.9)$$

where $\hat{Y}_{i,1}$ and $Y_{i,1}$ are defined in (3.2.8).

Throughout this chapter, on any fixed interval $[0,1]$, denote the class of Lipschitz continuous functions for any fixed constant $C > 0$ as

$$\text{Lip}([0,1], C) = \left\{m \middle| \left|m(x) - m(x')\right| \leq C\left|x - x'\right|, \forall x, x' \in [0,1]\right\}.$$

(B1) *The additive component function* $m_1(x_1) \in C^{(2)}[0,1]$ *defined in (1.5.1), while there is a constant* $0 < C_\infty < \infty$ *such that* $m_\beta \in Lip([0,1], C_\infty)$, $\forall \beta = 2, ..., d$.

(B2) *There exist positive constants* K_0 *and* λ_0 *such that* $\alpha(n) \leq K_0 e^{-\lambda_0 n}$ *holds for all* n, *with the* α*-mixing coefficients for* $\left\{\mathbf{Z}_i = \left(\mathbf{X}_i^T, \varepsilon_i\right)\right\}_{i=1}^{n}$ *defined as*

$$\alpha(k) = \sup_{B \in \sigma\{\mathbf{Z}_s, s \leq t\}, C \in \sigma\{\mathbf{Z}_s, s \geq t+k\}} \left|P(B \cap C) - P(B)P(C)\right|, \ k \geq 1. \qquad (3.2.10)$$

(B3) *The noise* ε_i *satisfies* $E(\varepsilon_i|\mathbf{X}_i) = 0, E(\varepsilon_i^2|\mathbf{X}_i) = 1$, $E\left(|\varepsilon_i|^{2+\delta}|\mathbf{X}_i\right) < M_\delta$ *for some* $\delta > 1/2$ *and a finite positive* M_δ. *The conditional standard deviation function* $\sigma(\mathbf{x})$ *is continuous on* $[0,1]^d$ and

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

(B4) *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* $[0,1]$ *as well as the uniform upper bound* C_f *and lower bound* c_f.

(B5) *The kernel function $K \in Lip\left([-1,1], C_k\right)$ for some constant $C_k > 0$,* and *is bounded, nonnegative, symmetric, and supported on $[-1,1]$. 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.*

(B6) *The number of interior knots $N \sim n^{2/5} \log n$, i.e., $c_N n^{2/5} \log n \leq N \leq C_N n^{2/5} \log n$ for some positive constants c_N,C_N, and the interval width $H = (N+1)^{-1}$.*

REMARK 3.2.1. The smoothness assumption of the true component functions is greatly relaxed and Assumption (B1) is closed to the minimal. By the result of Pham (1986), a geometrically ergodic time series is a strongly mixing sequence. Therefore, Assumption (B2) is suitable for (1.3.1) as a time series model under aforementioned assumptions. Assumption (B3)-(B5) are typical in the nonparametric smoothing literature, see for instance, Fan, and Gijbels (1996). For (B6), the proof of Theorem 3.2.1 in Section 3.7 will make it clear that the number of knots can be of the more general form $N \sim n^{2/5} N'$, where the sequence N' satisfies $N' \to \infty$, $n^{-\theta} N' \to 0$ for any $\theta > 0$. There is no optimal way to choose N' as in the literature. Here N is selected to be of barely larger order than $n^{2/5}$.

The asymptotic property of the kernel smoother $\tilde{m}_1^*(x_1)$ is well-developed. Under Assumptions (B1)-(B5), it is straightforward to verify (as in Bosq 1998) that

$$\sup_{x_1 \in [h, 1-h]} \left|\tilde{m}_1^*(x_1) - m_1(x_1)\right| = o_p\left(n^{-2/5} \log n\right)$$

$$\sqrt{nh}\left\{\tilde{m}_1^*(x_1) - m_1(x_1) - b_1(x_1) h^2\right\} \xrightarrow{D} N\left\{0, v_1^2(x_1)\right\},$$

where

$$\begin{aligned} b_1(x_1) &= \int u^2 K(u)\, du \left\{m_1''(x_1) f_1(x_1)/2 + m_1'(x_1) f_1'(x_1)\right\} f_1^{-1}(x_1), \\ v_1^2(x_1) &= \int K^2(u)\, du E\left[\sigma^2(X_1, ..., X_d) \left|X_1 = x_1\right.\right] f_1^{-1}(x_1). \end{aligned} \tag{3.2.11}$$

The following theorem states that the asymptotic uniform magnitude of difference between $\hat{m}_1^*(x_1)$ and $\tilde{m}_1^*(x_1)$ is of order $o_p\left(n^{-2/5}\right)$, which is dominated by the asymptotic uniform size of $\tilde{m}_1^*(x_1) - m_1(x_1)$. As a result, $\hat{m}_1^*(x_1)$ will have the same asymptotic distribution as $\tilde{m}_1^*(x_1)$.

THEOREM 3.2.1. *Under Assumptions (B1) to (B6), the SPBK estimator $\hat{m}_1^*(x_1)$ given in (3.2.9) satisfies*

$$\sup_{x_1\in[0,1]} \left|\hat{m}_1^*(x_1) - \tilde{m}_1^*(x_1)\right| = o_p\left(n^{-2/5}\right).$$

Hence with $b_1(x_1)$ and $v_1^2(x_1)$ as defined in (3.2.11), for any $x_1 \in [h, 1-h]$

$$\sqrt{nh}\left\{\hat{m}_1^*(x_1) - m_1(x_1) - b_1(x_1)h^2\right\} \xrightarrow{D} N\left\{0, v_1^2(x_1)\right\}.$$

REMARK 3.2.2. The above theorem holds for $\hat{m}_\alpha^*(x_\alpha)$ similarly constructed as $\hat{m}_1^*(x_1)$, for any $\alpha = 2, ..., d$, i.e.,

$$\hat{m}_\alpha^*(x_\alpha) = \frac{\sum_{i=1}^n K_h\left(X_{i,\alpha} - x_\alpha\right)\hat{Y}_{i,\alpha}}{\sum_{i=1}^n K_h\left(X_{i,1} - x_\alpha\right)}, \ \hat{Y}_{i,\alpha} = Y_i - \hat{c} - \sum_{1\leq\beta\leq d,\beta\neq\alpha} \hat{m}_\beta\left(X_{i,\beta}\right), \quad (3.2.12)$$

where $\hat{m}_\beta\left(X_{i,\beta}\right)$, $\beta = 1, ..., d$ are the pilot estimators of each component function given in (3.2.7). Similar constructions can be based on local polynomial instead of Nadaraya-Watson estimator. For more on the properties of local polynomial estimators, in particular, its minimax efficiency, see Fan and Gijbels (1996).

REMARK 3.2.3. Compared to the SBE in Mammen, Linton and Nielsen (1999), the variance term $v_1(x_1)$ is identical to that of SBE and the bias term $b_1(x_1)$ is much more explicit than that of SBE at least when Nadaraya-Watson smoother is used. Theorem 3.2.1 can be used to construct asymptotic confidence intervals. Under Assumptions (B1)-(B6), for any $\alpha \in (0,1)$, an asymptotic $100(1-\alpha)\%$ pointwise confidence intervals for $m(x)$ is

$$\hat{m}_1^*(x_1) - b_1(x_1)h^2 \pm z_{\alpha/2}\hat{\sigma}_1(x_1)\left\{\int K^2(u)\,du\right\}^{1/2} \Big/ \left\{nh\hat{f}_1(x_1)\right\}^{1/2}, \quad (3.2.13)$$

where $\hat{\sigma}_1(x_1)$ and $\hat{f}_1(x_1)$ are any constant estimators of $E\left[\sigma^2(\mathrm{X})\,|X_1 = x_1\right]$ and $f_1(x_1)$.

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

COROLLARY 3.2.1. *Under Assumptions (B1) to (B6) and the additional assumption that $m_\alpha(x_\alpha) \in C^{(2)}[0,1]$, $\alpha = 2, ..., d$, for any $\mathrm{x} \in [0,1]^d$, the SPBK estimator $\hat{m}_\alpha^*(\mathrm{x})$, $\alpha = 1, ..., d$, are defined as given in (3.2.12). Let*

$$\hat{m}^*(\mathrm{x}) = \hat{c} + \sum_{\alpha=1}^d \hat{m}_\alpha^*(x_\alpha), b(\mathrm{x}) = \sum_{\alpha=1}^d b_\alpha(x_\alpha), v^2(\mathrm{x}) = \sum_{\alpha=1}^d v_\alpha^2(x_\alpha),$$

then

$$\sqrt{nh}\left\{\hat{m}^*(\mathbf{x}) - m(\mathbf{x}) - b(\mathbf{x})h^2\right\} \xrightarrow{D} N\left\{0, v^2(\mathbf{x})\right\}.$$

3.3 Decomposition

In this section, some additional notations are introduced in order to shed some light on the ideas behind the proof of Theorem 3.2.1. Denote by $\|\phi\|_2$ the theoretical L^2 norm of a function ϕ on $[0,1]^d$, $\|\phi\|_2^2 = E\{\phi^2(\mathbf{X})\} = \int_{[0,1]^d} \phi^2(\mathbf{x}) f(\mathbf{x})\, d\mathbf{x}$, and the empirical L^2 norm as $\|\phi\|_{2,n}^2 = n^{-1}\sum_{i=1}^n \phi^2(\mathbf{X}_i)$. The corresponding inner products for L^2-integrable functions ϕ, φ on $[0,1]^d$ are

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

The evaluation of spline estimator $\hat{m}(\mathbf{x})$ at the n observations results in an n-dimensional vector, $\hat{m}(\mathbf{X}_1, \ldots, \mathbf{X}_n) = \{\hat{m}(\mathbf{X}_1), \ldots, \hat{m}(\mathbf{X}_n)\}^T$, which can be considered as the projection of $\mathbf{Y}$ on the space G_n with respect to the empirical inner product $\langle \cdot, \cdot \rangle_{2,n}$. In general, for any n-dimensional vector $\mathbf{\Lambda} = \{\Lambda_1, \ldots, \Lambda_n\}^T$, define $\mathbf{P}_n\mathbf{\Lambda}(\mathbf{x})$ as the spline function constructed from the projection of $\mathbf{\Lambda}$ on the inner product space $\left(G_n, \langle \cdot, \cdot \rangle_{2,n}\right)$

$$\mathbf{P}_n\mathbf{\Lambda}(\mathbf{x}) = \hat{\lambda}_0 + \sum_{\alpha=1}^d \sum_{J=1}^N \hat{\lambda}_{J,\alpha} B_{J,\alpha}(x_\alpha),$$

with the coefficients $\left(\hat{\lambda}_0, \hat{\lambda}_{1,1}, \ldots, \hat{\lambda}_{N,d}\right)$ given in (3.2.6). Next, the multivariate function $\mathbf{P}_n\mathbf{\Lambda}(\mathbf{x})$ is decomposed into empirically centered additive components $\mathbf{P}_{n,\alpha}\mathbf{\Lambda}(x_\alpha)$, $\alpha = 1, \ldots, d$ and the constant component $\mathbf{P}_{n,c}\mathbf{\Lambda}$

$$\mathbf{P}_{n,\alpha}\mathbf{\Lambda}(x_\alpha) = \mathbf{P}^*_{n,\alpha}\mathbf{\Lambda}(x_\alpha) - n^{-1}\sum_{i=1}^n \mathbf{P}^*_{n,\alpha}\mathbf{\Lambda}(X_{i,\alpha}), \tag{3.3.1}$$

$$\mathbf{P}_{n,c}\mathbf{\Lambda} = \hat{\lambda}_0 + n^{-1}\sum_{\alpha=1}^d \sum_{i=1}^n \mathbf{P}^*_{n,\alpha}\mathbf{\Lambda}(X_{i,\alpha}). \tag{3.3.2}$$

where $\mathbf{P}^*_{n,\alpha}\mathbf{\Lambda}(x_\alpha) = \sum_{J=1}^N \hat{\lambda}_{J,\alpha} B_{J,\alpha}(x_\alpha)$. With these new notations, one can rewrite the spline estimators $\hat{m}(\mathbf{x}), \hat{m}_\alpha(x_\alpha), \hat{m}_c$ defined in (3.2.5) and (3.2.7) 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 $Y = m(\mathbf{X}) + \sigma(\mathbf{X})\varepsilon = m(\mathbf{X}) + \mathbf{E}$ with noise vector $\mathbf{E} = \{\sigma(\mathbf{X}_i)\varepsilon_i\}_{i=1}^n$, one defines similarly the noiseless spline smoothers

$$\tilde{m}(\mathbf{x}) = \mathbf{P}_n\{m(\mathbf{X})\}(\mathbf{x}), \tilde{m}_\alpha(x_\alpha) = \mathbf{P}_{n,\alpha}\{m(\mathbf{X})\}(x_\alpha), \tilde{m}_c = \mathbf{P}_{n,c}\{m(\mathbf{X})\}, \tag{3.3.3}$$

and the variance spline components

$$\tilde{\varepsilon}(\mathbf{x}) = \mathbf{P}_n\mathbf{E}(\mathbf{x}), \tilde{\varepsilon}_\alpha(x_\alpha) = \mathbf{P}_{n,\alpha}\mathbf{E}(x_\alpha), \tilde{\varepsilon}_c = \mathbf{P}_{n,c}\mathbf{E}. \tag{3.3.4}$$

Due to the linearity of operators $\mathbf{P}_n, \mathbf{P}_{n,c}, \mathbf{P}_{n,\alpha}, \alpha = 1, ..., d$, one has the following crucial decomposition for proving Theorem 3.2.1,

$$\hat{m}(\mathbf{x}) = \tilde{m}(\mathbf{x}) + \tilde{\varepsilon}(\mathbf{x}), \ \hat{m}_c = \tilde{m}_c + \tilde{\varepsilon}_c, \ \hat{m}_\alpha(x_\alpha) = \tilde{m}_\alpha(x_\alpha) + \tilde{\varepsilon}_\alpha(x_\alpha) \tag{3.3.5}$$

for $\alpha = 1, ..., d$. As closer examination is needed later for $\tilde{\varepsilon}(\mathbf{x})$ and $\tilde{\varepsilon}_\alpha(x_\alpha)$, one defines in addition $\tilde{\mathbf{a}} = \{\tilde{a}_0, \tilde{a}_{1,1}, ..., \tilde{a}_{N,d}\}^T$ as the minimizer of the following

$$\sum_{i=1}^n \left\{\sigma(\mathbf{X}_i)\varepsilon_i - a_0 - \sum_{\alpha=1}^d \sum_{J=1}^N a_{J,\alpha} B_{J,\alpha}(X_{i,\alpha})\right\}^2. \tag{3.3.6}$$

Then $\tilde{\varepsilon}(\mathbf{x})$ in (3.3.4) 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 (3.3.6), and vector $\mathbf{B}(\mathbf{x})$ and matrix $\mathbf{B}$ are defined as

$$\mathbf{B}(\mathbf{x}) = \{1, B_{1,1}(x_1), ..., B_{N,d}(x_d)\}^T, \ \mathbf{B} = \{\mathbf{B}(\mathbf{X}_1), ..., \mathbf{B}(\mathbf{X}_n)\}^T. \tag{3.3.7}$$

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

$$\tilde{\mathbf{a}} = \left\{\begin{matrix} 1 & \mathbf{0}_{dN}^T \\ \mathbf{0}_{dN} & \left\langle B_{J,\alpha}, B_{J',\alpha'}\right\rangle_{2,n} \end{matrix}\right\}_{\substack{1\le\alpha,\alpha'\le d,\\ 1\le J,J'\le N}}^{-1} \left\{\begin{matrix} \frac{1}{n}\sum_{i=1}^n \sigma(\mathbf{X}_i)\varepsilon_i \\ \frac{1}{n}\sum_{i=1}^n B_{J,\alpha}(X_{i,\alpha})\sigma(\mathbf{X}_i)\varepsilon_i \end{matrix}\right\}_{\substack{1\le J\le N,\\ 1\le\alpha\le d}}, \tag{3.3.8}$$

where $\mathbf{0}_p$ is a p-vector with all elements 0. The main objective here is to study the difference between the smoothed backfitted estimator $\hat{m}_1^*(x_1)$ and the smoothed "oracle" estimator $\tilde{m}_1^*(x_1)$, both given in (3.2.9).

From now on, assume without loss of generality that $d = 2$ for notational brevity. Making use of the definition of $\hat{c}$ and the signal noise decomposition (3.3.5), the difference $\tilde{m}_1^*(x_1) - \hat{m}_1^*(x_1) - \hat{c} + c$ can be treated as the sum of two terms

$$\frac{\frac{1}{n}\sum_{i=1}^n K_h(X_{i,1} - x_1)\{\hat{m}_2(X_{i,2}) - m_2(X_{i,2})\}}{\frac{1}{n}\sum_{i=1}^n K_h(X_{i,1} - x_1)} = \frac{\Psi_b(x_1) + \Psi_v(x_1)}{\frac{1}{n}\sum_{i=1}^n K_h(X_{i,1} - x_1)}, \tag{3.3.9}$$

where

$$\Psi_b(x_1) = \frac{1}{n}\sum_{i=1}^{n} K_h(X_{i,1}-x_1)\{\tilde{m}_2(X_{i,2}) - m_2(X_{i,2})\}, \tag{3.3.10}$$

$$\Psi_v(x_1) = \frac{1}{n}\sum_{i=1}^{n} K_h(X_{i,1}-x_1)\,\tilde{\varepsilon}_2(X_{i,2}). \tag{3.3.11}$$

The term $\Psi_b(x_1)$ is induced by the bias term $\tilde{m}_2(X_{i,2}) - m_2(X_{i,2})$, while $\Psi_v(x_1)$ is related to the variance term $\tilde{\varepsilon}_2(X_{i,2})$. Both of these two terms have order $o_p(n^{-2/5})$ by Propositions 3.4.1 and 3.5.1 in the next two sections. Standard theory of kernel density estimation ensures that the denominator term in (3.3.9), $n^{-1}\sum_{i=1}^{n} K_h(X_{i,1}-x_1)$, has a positive lower bound for $x_1 \in [0,1]$. The additional nuisance term $\hat{c} - c$ is of clearly order $O_p\left(n^{-1/2}\right)$ and thus $o_p\left(n^{-2/5}\right)$, which needs no further arguments for the proofs. Theorem 3.2.1 then follows from Propositions 3.4.1 and 3.5.1.

3.4 Bias reduction

In this section, we show that the bias term $\Psi_b(x_1)$ of (3.3.10) is uniformly of order $o_p\left(n^{-2/5}\right)$ for $x_1 \in [0,1]$, which is given by Proposition 3.4.1 as below.

PROPOSITION 3.4.1. *Under Assumptions (B1) to (B2), and (B4) to (B6)*

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

One important result from page 149, de Boor (2001), is cited before the proof.

LEMMA 3.4.1. *There exists a constant $C_\infty > 0$ such that for any component function $m_\alpha \in Lip([0,1], C_\infty)$ and function $g_\alpha \in G$, $\alpha = 1, ..., d$, $\|g_\alpha - m_\alpha\|_\infty \leq C_\infty H$.*

LEMMA 3.4.2. *Under Assumption (B1), there exists function g_1, $g_2 \in G$, such that*

$$\left\| \tilde{m} - g + \sum_{\alpha=1}^{2} \langle 1, g_\alpha(X_\alpha)\rangle_{2,n} \right\|_{2,n} = O_p\left(n^{-1/2} + H\right),$$

where $g(\mathbf{x}) = c + \sum_{\alpha=1}^{2} g_\alpha(x_\alpha)$ and $\tilde{m}$ is defined in (3.3.3).

PROOF. By Lemma 3.4.1, there is a constant $C_\infty > 0$ such that for function $g_\alpha \in G$, $\|g_\alpha - m_\alpha\|_\infty \leq C_\infty H$, $\alpha = 1, 2$. Thus $\|g - m\|_\infty \leq \sum_{\alpha=1}^{2} \|g_\alpha - m_\alpha\|_\infty \leq 2C_\infty H$ and $\|\tilde{m} - m\|_{2,n} \leq \|g - m\|_{2,n} \leq 2C_\infty H$. The triangular inequality then implies that

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

$$\begin{aligned} \left|\langle g_\alpha(X_\alpha), 1\rangle_{2,n}\right| &\leq \left|\langle 1, g_\alpha(X_\alpha)\rangle_{2,n} - \langle 1, m_\alpha(X_\alpha)\rangle_{2,n}\right| + \left|\langle 1, m_\alpha(X_\alpha)\rangle_{2,n}\right| \\ &\leq C_\infty H + O_p\left(n^{-1/2}\right). \end{aligned} \tag{3.4.1}$$

Therefore

$$\begin{aligned} &\left\|\tilde{m} - g + \sum_{\alpha=1}^{2} \langle 1, g_\alpha(X_\alpha)\rangle_{2,n}\right\|_{2,n} \leq \|\tilde{m} - g\|_{2,n} + \sum_{\alpha=1}^{2} \left|\langle 1, g_\alpha(X_\alpha)\rangle_{2,n}\right| \\ \leq\ & 6C_\infty H + O_p\left(n^{-1/2}\right) = O_p\left(n^{-1/2} + H\right). \end{aligned}$$

□

PROOF OF PROPOSITION 3.4.1. Denote

$$\begin{aligned} R_1 &= \sup_{x_1 \in [0,1]} \left|\frac{\sum_{i=1}^{n} K_h(X_{i,1} - x_1)\{g_2(X_{i,2}) - m_2(X_{i,2})\}}{\sum_{i=1}^{n} K_h(X_{i,1} - x_1)}\right| \\ R_2 &= \sup_{x_1 \in [0,1]} \left|\frac{\sum_{i=1}^{n} K_h(X_{i,1} - x_1)\left\{\tilde{m}_2(X_{i,2}) - g_2(X_{i,2}) + \langle 1, g_2(X_2)\rangle_{2,n}\right\}}{\sum_{i=1}^{n} K_h(X_{i,1} - x_1)}\right|, \end{aligned}$$

then $\sup_{x_1 \in [0,1]} |\Psi_b(x_1)| \leq \left|\langle 1, g_2(X_2)\rangle_{2,n}\right| + R_1 + R_2$. For R_1, using Lemma 3.4.1

$$R_1 \leq C_\infty H. \tag{3.4.2}$$

To deal with R_2, let $B^*_{J,2}(x_\alpha) = B_{J,2}(x_\alpha) - \langle 1, B_{J,2}(X_\alpha)\rangle_{2,n}$, for $J = 1, \ldots, N$, $\alpha = 1, 2$, then one can write

$$\tilde{m}(\mathrm{x}) - g(\mathrm{x}) + \sum_{\alpha=1}^{2} \langle 1, g_\alpha(X_\alpha)\rangle_{2,n} = \tilde{a}^* + \sum_{\alpha=1}^{2} \sum_{J=1}^{N} \tilde{a}^*_{J,\alpha} B^*_{J,\alpha}(x_\alpha).$$

Thus, $n^{-1} \sum_{i=1}^{n} K_h(X_{i,1} - x_1)\left\{\tilde{m}_2(X_{i,2}) - g_2(X_{i,2}) + \langle 1, g_2(X_2)\rangle_{2,n}\right\}$ can be rewritten as $n^{-1} \sum_{i=1}^{n} K_h(X_{i,1} - x_1) \sum_{J=1}^{N} \tilde{a}^*_{J,2} B^*_{J,2}(X_{i,2})$, bounded by

$$\begin{aligned} &\sum_{J=1}^{N} \left|\tilde{a}^*_{J,2}\right| \sup_{1 \leq J \leq N} \left|n^{-1} \sum_{i=1}^{n} K_h(X_{i,1} - x_1) B^*_{J,2}(X_{i,2})\right| \\ \leq\ & \sum_{J=1}^{N} \left|\tilde{a}^*_{J,2}\right| \left\{\sup_{1 \leq J \leq N} \left|n^{-1} \sum_{i=1}^{n} \omega_J(\mathrm{X}_i, x_1)\right| + A_{n,1} \left|n^{-1} \sum_{i=1}^{n} K_h(X_{i,1} - x_1)\right|\right\}, \end{aligned}$$

where $A_{n,1} = O_p(\log n/\sqrt{n})$ as in (3.7.15), $\omega_J(\mathbf{X}_i, x_1)$ is as given in (3.5.5) with mean $\mu_{\omega_J}(x_1)$. By Lemma 3.7.2

$$\sup_{x_1\in[0,1]}\sup_{1\le J\le N}\left|n^{-1}\sum_{i=1}^{n}\omega_J(\mathbf{X}_i,x_1)\right| \le \sup_{x_1\in[0,1]}\sup_{1\le J\le N}\left|n^{-1}\sum_{i=1}^{n}\omega_J(\mathbf{X}_i,x_1)-\mu_{\omega_J}(x_1)\right|$$

$$+\sup_{x_1\in[0,1]}\sup_{1\le J\le N}\left|\mu_{\omega_J}(x_1)\right| = O_p\left(\log n/\sqrt{nh}\right)+O_p\left(H^{1/2}\right)=O_p\left(H^{1/2}\right).$$

Therefore, one has

$$\sup_{x_1\in[0,1]}\left|n^{-1}\sum_{i=1}^{n}K_h(X_{i,1}-x_1)\left\{\tilde{m}_2(X_{i,2})-g_2(X_{i,2})+\langle 1,g_2(X_2)\rangle_{2,n}\right\}\right|$$

$$\le\left\{N\sum_{J=1}^{N}\left(\tilde{a}_{J,2}^{*}\right)^2\right\}^{1/2}\left\{O_p\left(H^{1/2}\right)+O_p\left(\frac{\log n}{\sqrt{n}}\right)\right\}=O_p\left(\left\{\sum_{J=1}^{N}\left(\tilde{a}_{J,2}^{*}\right)^2\right\}^{1/2}\right)$$

$$=O_p\left(\left\|\tilde{m}-g+\sum_{\alpha=1}^{2}\langle 1,g_\alpha(X_\alpha)\rangle_{2,n}\right\|_2\right)=O_p\left(\left\|\tilde{m}-g+\sum_{\alpha=1}^{2}\langle 1,g_\alpha(X_\alpha)\rangle_{2,n}\right\|_{2,n}\right).$$

where the last step follows from Lemma 3.7.7. Thus, by lemma 3.4.2

$$R_2=O_p\left(n^{-1/2}+H\right). \tag{3.4.3}$$

Combining (3.4.1), (3.4.2) and (3.4.3), one establishes Proposition 3.4.1. □

3.5 Variance reduction

This section shows that the term $\Psi_v(x_1)$ given in (3.3.11) is uniformly of order $o_p\left(n^{-2/5}\right)$. This is the most challenging part to be proved, mostly done in Section 3.7. Define an auxiliary entity

$$\tilde{\varepsilon}_2^{*}=\sum_{J=1}^{N}\tilde{a}_{J,2}B_{J,2}(x_2), \tag{3.5.1}$$

where $\tilde{a}_{J,2}$ is given in (3.3.8). Definitions (3.3.1) and (3.3.2) imply that $\tilde{\varepsilon}_2(x_2)$ is simply the empirical centering of $\tilde{\varepsilon}_2^{*}(x_2)$, i.e.

$$\tilde{\varepsilon}_2(x_2)\equiv\tilde{\varepsilon}_2^{*}(x_2)-n^{-1}\sum_{i=1}^{n}\tilde{\varepsilon}_2^{*}(X_{i,2}). \tag{3.5.2}$$

PROPOSITION 3.5.1. *Under Assumptions (B2) to (B6),*

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

According to (3.5.2), one can write $\Psi_v(x_1) = \Psi_v^{(2)}(x_1) - \Psi_v^{(1)}(x_1)$, where

$$\Psi_v^{(1)}(x_1) = n^{-1}\sum_{l=1}^{n} K_h(X_{l,1} - x_1)\cdot n^{-1}\sum_{i=1}^{n} \tilde{\varepsilon}_2^*(X_{i,2}), \tag{3.5.3}$$

$$\Psi_v^{(2)}(x_1) = n^{-1}\sum_{l=1}^{n} K_h(X_{l,1} - x_1)\tilde{\varepsilon}_2^*(X_{l,2}), \tag{3.5.4}$$

in which $\tilde{\varepsilon}_2^*(X_{i,2})$ is given in (3.5.1). Further one denotes

$$\omega_J(\mathbf{X}_l, x_1) = K_h(X_{l,1} - x_1) B_{J,2}(X_{l,2}), \ \mu_{\omega_J}(x_1) = E\omega_J(\mathbf{X}_l, x_1), \tag{3.5.5}$$

by (3.3.8) and (3.5.1), $\Psi_v^{(2)}(x_1)$ can be rewritten as

$$\Psi_v^{(2)}(x_1) = n^{-1}\sum_{l=1}^{n}\sum_{J=1}^{N} \tilde{a}_{J,2}\omega_J(\mathbf{X}_l, x_1). \tag{3.5.6}$$

The uniform order of $\Psi_v^{(1)}(x_1)$ and $\Psi_v^{(2)}(x_1)$ are given in the following two lemmas.

LEMMA 3.5.1. *Under Assumptions (B2) to (B6), $\Psi_v^{(1)}(x_1)$ in (3.5.3) satisfies*

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

PROOF OF LEMMA 3.5.1. Based on (3.5.1)

$$\begin{aligned} n^{-1}\sum_{i=1}^{n} \tilde{\varepsilon}_2^*(X_{i,2}) &= \sum_{J=1}^{N} \tilde{a}_{J,2}\left\{n^{-1}\sum_{i=1}^{n} B_{J,2}(X_{i,2})\right\} \\ &\leq \left|\sum_{J=1}^{N} \tilde{a}_{J,2}\right| \cdot \sup_{1\leq J\leq N}\left|\frac{1}{n}\sum_{i=1}^{n} B_{J,2}(X_{i,2})\right|. \end{aligned}$$

Lemma 3.7.5 implies that

$$\left|\sum_{J=1}^{N} \tilde{a}_{J,2}\right| \leq \left\{N\cdot\sum_{J=1}^{N} \tilde{a}_{J,2}^2\right\}^{1/2} \leq \left\{N\cdot\tilde{\mathbf{a}}^T\tilde{\mathbf{a}}\right\}^{1/2} = O_p\left(Nn^{-1/2}\log n\right).$$

By (3.7.15),(3.7.18), $\sup_{1\leq J\leq N}\left|n^{-1}\sum_{i=1}^{n} B_{J,2}(X_{i,2})\right| \leq A_{n,1} = O_p\left(n^{-1/2}\log n\right)$, so

$$n^{-1}\sum_{i=1}^{n} \tilde{\varepsilon}_2^*(X_{i,2}) = O_p\left\{N(\log n)^2/n\right\}. \tag{3.5.7}$$

By Assumption (B5) on the kernel function K, standard theory on kernel density estimation entails that $\sup_{x_1\in[0,1]}\left|n^{-1}\sum_{l=1}^{n} K_h\left(X_{l,1}-x_1\right)\right| = O_p(1)$. Thus with (3.5.7) the lemma follows immediately. □

LEMMA 3.5.2. *Under Assumptions (B2) to (B6),* $\Psi_v^{(2)}(x_1)$ *in (3.5.4) satisfies*

$$\sup_{x_1\in[0,1]}\left|\Psi_v^{(2)}(x_1)\right| = O_p(H).$$

Lemma 3.5.2 follows from Lemmas 3.7.9 and 3.7.10. Proposition 3.5.1 follows from Lemmas 3.5.1 and 3.5.2.

3.6 Simulations

In this section two simulation experiments are carried out to illustrate the finite-sample behavior of the SPBK estimators $\hat{m}_\alpha^*(x_\alpha)$ for $\alpha = 1,...,d$. The programming codes are available both in R 2.2.1 and XploRe. For more information on XploRe, see Härdle, Hlávka and Klinke (2000) or visit the following website, http://www.xplore-stat.de.

The number of knots N for the spline estimation as in (3.2.6) will be determined by the sample size and a tuning constant c. To be precise

$$N = \min\left(\left[cn^{2/5}\log n\right] + 1, \left[(n/2-1)\,d^{-1}\right]\right),$$

in which $[a]$ denotes the integer part of a. In this simulation study, c is chosen to be 0.5 and 1.0. As seen in Table 4.3, the choice of c makes little difference, so we always recommend to use $c = 0.5$ to save computation for massive data set. The additional constraint that $N \leq (n/2-1)\,d^{-1}$ ensures that the number of terms in the linear least squares problem (3.2.6), $1+dN$, is no greater than $n/2$, which is necessary when the sample size n is moderate and dimension d is high.

We have obtained for comparison both the SPBK estimator $\hat{m}_\alpha^*(x_\alpha)$ and the "oracle" estimator $\tilde{m}_\alpha^*(x_\alpha)$ by Nadaraya-Watson regression estimation using quartic kernel and the rule-of-thumb bandwidth.

We consider first the accuracy of the estimation, measured in terms of mean average squared error. To see that the SPBK estimator $\hat{m}_\alpha^*(x_\alpha)$ is as efficient as the "oracle" smoother $\tilde{m}_{\alpha}^{*}(x_{\alpha})$, define the following empirical relative efficiency of $\hat{m}_{\alpha}^{*}(x_{\alpha})$ with respect to $\tilde{m}_{\alpha}^{*}(x_{\alpha})$ as

$$\text{eff}_{\alpha} = \left[\frac{\sum_{i=1}^{n}\left\{\tilde{m}_{\alpha}^{*}(X_{i,\alpha}) - m_{\alpha}(X_{i,\alpha})\right\}^{2}}{\sum_{i=1}^{n}\left\{\hat{m}_{\alpha}^{*}(X_{i,\alpha}) - m_{\alpha}(X_{i,\alpha})\right\}^{2}}\right]^{1/2}. \tag{3.6.1}$$

Theorem 3.2.1 indicates that the eff_{α} should be close to 1 for all $\alpha = 1, ..., d$. Figure 4.15 and 4.16 provide the kernel density estimations of the above empirical efficiencies to observe the convergence, where one sees that the center of the density plots is going toward the standard line 1.0 and the shape of those plots becomes narrower as well when sample size n is increasing.

3.6.1 Example 1

A time series $\{Y_t\}_{t=-1999}^{n+3}$ is generated according to the NAAR model with sine functions given in Chen and Tsay (1993),

$$Y_t = 1.5\sin\left(\frac{\pi}{2}Y_{t-2}\right) - 1.0\sin\left(\frac{\pi}{2}Y_{t-3}\right) + \sigma_0\varepsilon_t,\ \sigma_0 = 0.5, 1.0,$$

where $\{\varepsilon_t\}_{t=-1996}^{n+3}$ are i.i.d. standard normal errors. Let $\mathbf{X}_t^T = \{Y_{t-1}, Y_{t-2}, Y_{t-3}\}$. Theorem 3, page 91 of Doukhan (1994) establishes that $\left\{Y_t, \mathbf{X}_t^T\right\}_{t=-1996}^{n+3}$ is geometrically ergodic. The first 2000 observations are discarded to make the last $n+3$ observations $\{Y_t\}_{t=1}^{n+3}$ behave like a geometrically α-mixing and strictly stationary time series. The multivariate datum $\left\{Y_t, \mathbf{X}_t^T\right\}_{t=4}^{n+3}$ then satisfies Assumptions (B1) to (B6) except that instead of being $[0, 1]$, the range of $Y_{t-\alpha}, \alpha = 1, 2, 3$ needs to be recalibrated. Since there is no exact knowledge of the distribution of the Y_t, many realizations of size 50000 have been generated from which one sees that more than 95% of the observations fall in $[-2.58, 2.58]$ ($[-3.14, 3.14]$) with $\sigma_0 = 0.5$ ($\sigma_0 = 1$) . We will estimate the functions $\{m_{\alpha}(x_{\alpha})\}_{\alpha=1}^{3}$ for $x_{\alpha} \in [-2.58, 2.58]$ ($[-3.14, 3.14]$) with $\sigma_0 = 0.5$ ($\sigma_0 = 1.0$), where

$$\begin{aligned} m_1(x_1) &\equiv 0, \\ m_2(x_2) &\equiv 1.5\sin\left(\frac{\pi}{2}x_2\right) - E\left[1.5\sin\left(\frac{\pi}{2}Y_t\right)\right], \\ m_3(x_3) &\equiv -1.0\sin\left(\frac{\pi}{2}x_3\right) - E\left[-1.0\sin\left(\frac{\pi}{2}Y_t\right)\right]. \end{aligned}$$

Sample size n is chosen to be 100, 200, 500 and 1000. Table 4.3 lists the average squared error (ASE) of the SPBK estimators and the constant spline pilot estimators from 100 Monte Carlo replications. As expected, increases in sample size reduce ASE for both estimators and across all combination of c values and noise levels. Table 4.3 also shows that the SPBK estimators improve upon the spline pilot estimators immensely regardless of noise level and sample size, which implies that our second Nadaraya-Watson smoothing step is not redundant.

To have some impression of the actual function estimates, at noise level $\sigma_0 = 0.5$ with sample size $n = 200$, 500, the oracle estimators (thin dotted lines), SPBK estimators $\hat{m}_{\alpha}^{*}$ (thin solid lines) and their 95% pointwise confidence intervals (upper and lower dashed curves) for the true functions m_{α} (thick solid lines) have been plotted in Figure 4.12, 4.13 and 4.14. The visual impression of the SPBK estimators are rather satisfactory and their the performance improves with increasing n.

To see the convergence, Figure 4.15 plots the kernel density estimations of the 100 empirical efficiencies for sample sizes $n = 100, 200, 500$ and 1000 at the noise level $\sigma_0 = 0.5$. The vertical line at efficiency $= 1$ is the standard line for the comparison of $\hat{m}_{\alpha}^{*}(x_{\alpha})$ and $\tilde{m}_{\alpha}^{*}(x_{\alpha})$. One can clearly see from Figure 4.15 that as sample size n increases the efficiency distribution converges to 1, confirmative to the conclusions of Theorem 3.2.1.

Lastly, the computing time of Example 3.6.1 is provided based on 100 replications done on an ordinary PC with Intel Pentium IV 1.86 GHz processor and 1.0 GB RAM. The average time run by XploRe to generate one sample of size n and compute the SPBK estimator and marginal integration estimator (MIE) has been reported in Table 4.4. The MIEs have been obtained by directly recalling the "intest" in XploRe. As expected, the computing time for MIE is extremely sensitive to sample size due to the fact that it requires n^2 least squares in two steps. In contrast, at least for large sample data, the proposed SPBK is thousands of times faster than MIE. Thus our SPBK estimation is feasible and appealing to deal with massive data set.

3.6.2 Example 2

Consider the following nonlinear additive heteroscedastic model

$$Y_t = \sum_{\alpha=1}^{d} \sin\left(\frac{\pi}{2.5} X_{t-\alpha}\right) + \sigma(\mathbf{X})\varepsilon_t, \varepsilon_t \overset{i.i.d}{\sim} N(0,1),$$

in which $\mathbf{X}_t^T = \{X_{t-1}, ..., X_{t-d}\}$ is a sequence of i.i.d random variables with standard normal distribution truncated in the interval $[-2.5, 2.5]$ and the conditional standard deviation function is defined as

$$\sigma(\mathbf{X}) = \sigma_0 \frac{\sqrt{d}}{2} \cdot \frac{5 - \exp\left(\sum_{\alpha=1}^{d} |X_{t-\alpha}| \big/ d\right)}{5 + \exp\left(\sum_{\alpha=1}^{d} |X_{t-\alpha}| \big/ d\right)}, \ \sigma_0 = 0.1.$$

This choice of $\sigma(\mathbf{X})$ ensures that the 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.

For $d = 30$, 100 replications have been done for sample sizes $n = 500, 1000, 1500$ and 2000. The kernel density estimator of the 100 empirical efficiencies is graphically represented in Figures 4.16 and 4.17. Again one sees that with increasing sample sizes, the relative efficiency are becoming closer to the vertical standard line, with narrower spread out.

3.7 Proof of Theorems

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

3.7.1 Preliminaries

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

$$c_{J,\alpha} = \left\|I_{J,\alpha}\right\|_2^2 = \int I_{J,\alpha}^2(x_\alpha) f_\alpha(x_\alpha)\, dx_\alpha. \tag{3.7.1}$$

LEMMA 3.7.1. *Under Assumptions (B4) and (B6), one has:*

(i) there exist constants $C_0(f)$ and $C_1(f)$ depending on the marginal densities $f_\alpha(x_\alpha), \alpha = 1, 2$, such that $C_0(f) H \leq \left\|b_{J,\alpha}\right\|_2^2 \leq C_1(f) H$.

(ii)

$$E\left\{B_{J,\alpha}(X_{i,\alpha}) B_{J',\alpha}(X_{i,\alpha})\right\} = \begin{cases} 1 & J' = J \\ -c_{J+1,\alpha} \left\|b_{J,\alpha}\right\|_2^{-1} \left\|b_{J',\alpha}\right\|_2^{-1} & J' = J - 1 \\ -c_{J+2,\alpha} \left\|b_{J,\alpha}\right\|_2^{-1} \left\|b_{J',\alpha}\right\|_2^{-1} & J' = J + 1 \\ 0 & \left|J - J'\right| > 1 \end{cases}$$
$$\sim \begin{cases} 1 & \left|J' - J\right| \leq 1 \\ 0 & \left|J' - J\right| > 1 \end{cases}$$

and for $k \geq 1$,

$$E\left|B_{J,\alpha}(X_{i,\alpha}) B_{J',\alpha}(X_{i,\alpha})\right|^k = \begin{cases} \left\|b_{J,\alpha}\right\|_2^{-2k} \left(c_{J+1,\alpha} + c_{J+1,\alpha}^{2k} / c_{J,\alpha}^{2k}\right) & J = J' \\ c_{J+1,\alpha}^k \left\|b_{J,\alpha}\right\|_2^{-k} \left\|b_{J',\alpha}\right\|_2^{-k} / c_{J,\alpha}^{k-1} & J' = J - 1 \\ c_{J+2,\alpha}^k \left\|b_{J,\alpha}\right\|_2^{-k} \left\|b_{J',\alpha}\right\|_2^{-k} / c_{J+1,\alpha}^{k-1} & J' = J + 1 \\ 0 & \left|J - J'\right| > 1 \end{cases}$$
$$\sim \begin{cases} H^{1-k} & \left|J' - J\right| \leq 1 \\ 0 & \left|J' - J\right| > 1 \end{cases}$$

where $c_{J,\alpha}$, $\alpha = 1, 2, J = 1, ..., N + 1$ are given in (3.7.1).

Proof. Note that for any $\alpha = 1, 2$, $J = 1, ..., N$, $b_{J,\alpha}(x_\alpha)$ in (3.2.2) can be rewritten 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

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

In Assumption (B4), the two positive constants c_f, C_f are the upper and lower bounds of $f_\alpha(x_\alpha)$, then

$$c_f H \leq c_{J,\alpha} \leq C_f H,$$

$$C_0(f) H = c_f\left(1 + c_f/C_f\right) H \leq \left\|b_{J,\alpha}\right\|_2^2 \leq C_f\left(1 + C_f/c_f\right) H = C_1(f) H,$$

for all $J = 1, ..., N + 1, \alpha = 1, 2$. The proof of (ii) is trivial. □

Lemma 3.7.2. *Under Assumptions (B4) to (B6), for $\mu_{\omega_J}(x_1)$ given in (3.5.5)*

$$\sup_{x_1 \in [0,1]} \sup_{1 \leq J \leq N} \left|\mu_{\omega_J}(x_1)\right| = O\left(H^{1/2}\right).$$

PROOF. By definition, $\left|\mu_{\omega_J}(x_1)\right| = \left|E\left\{K_h\left(X_{l,1} - x_1\right) B_{J,2}\left(X_{l,2}\right)\right\}\right|$ is bounded by

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

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

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

the proof of the lemma is then completed, by (i) of Lemma 3.7.1. □

LEMMA 3.7.3. *Under Assumptions (B4) and (B6), there exist constants* $C_0 > c_0 > 0$ *such that for any* $\mathbf{a} = \left(a_0, a_{1,1}, ..., a_{N,1}, a_{1,2}, ..., a_{N,2}\right)$,

$$c_0\left(a_0^2 + \sum_{J,\alpha} a_{J,\alpha}^2\right) \le \left\|a_0 + \sum_{J,\alpha} a_{J,\alpha} B_{J,\alpha}\right\|_2^2 \le C_0\left(a_0^2 + \sum_{J,\alpha} a_{J,\alpha}^2\right). \tag{3.7.2}$$

PROOF. Lemma 1 of Stone (1985) provides a constant $c_0 > 0$ such that

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

then (3.7.2) follows if there exist constants $C_0' > c_0' > 0$ such that for $\alpha = 1, 2$

$$c_0' \sum_{J=1}^N a_{J,\alpha}^2 \le \left\|\sum_{J=1}^N a_{J,\alpha} B_{J,\alpha}\right\|_2^2 \le C_0' \sum_{J=1}^N a_{J,\alpha}^2. \tag{3.7.3}$$

To prove (3.7.3), the original B-Spline basis is employed. Without loss of generality, let $\alpha = 1$, and use the constant basis $\left\{I_{J,1}(x_1)\right\}_{J=1}^{N+1}$. Represent the term $\sum_{J=1}^N a_{J,1} B_{J,1}(x_1)$ as follows

$$\sum_{J=1}^N a_{J,1} B_{J,1}(x_1) = \sum_{J=1}^{N+1} d_{J,1} I_{J,1}(x_1). \tag{3.7.4}$$

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

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

The uniform boundedness of the joint density in Assumption (B4) implies that

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

Then Lemma 3.7.1 and (3.7.4) lead to

$$\sum_{J=1}^{N+1} d_{J,1}^2 = \sum_{J=1}^{N} \frac{a_{J,1}^2}{\|b_{J,1}\|_2^2}\left\{\left(\frac{c_{J+1,1}}{c_{J,1}}\right)^2 + 1\right\}.$$

Then

$$c_a\sum_{J=1}^{N} a_{J,1}^2 H^{-1} \le \sum_{J=1}^{N+1} d_{J,1}^2 \le C_a\sum_{J=1}^{N} a_{J,1}^2 H^{-1},$$

for positive constants c_a and C_a. Therefore,

$$c_f c_a\sum_{J=1}^{N} a_{J,1}^2 \le \left\|\sum_{J=1}^{N} a_{J,1}B_{J,1}\right\|_2^2 = \left\|\sum_{J=1}^{N+1} d_{J,1}I_{J,1}\right\|_2^2 \le C_f C_a\sum_{J=1}^{N} a_{J,1}^2,$$

i.e. (3.7.3) holds given $c_0' = c_f c_a$, $C_0' = C_f C_a$. □

Lemmas 2.6.2 and 3.7.2 entail the next Lemma 3.7.4, which shows the uniform supremum magnitude of $n^{-1}\sum_{l=1}^{n}\left\{\omega_J(\mathbf{X}_l, x_1) - \mu_{\omega_J}(x_1)\right\}$ and $n^{-1}\sum_{l=1}^{n}\omega_J(\mathbf{X}_l, x_1)$. The quantities $\omega_J(\mathbf{X}_l, x_1)$ and $\mu_{\omega_J}(x_1)$ are defined in (3.5.5).

LEMMA 3.7.4. *Under Assumptions (B2), (B4) to (B6)*

$$\sup_{x_1\in[0,1]}\sup_{1\le J\le N}\left|n^{-1}\sum_{l=1}^{n}\left\{\omega_J(\mathbf{X}_l, x_1) - \mu_{\omega_J}(x_1)\right\}\right| = O_p\left(\log n/\sqrt{nh}\right), \tag{3.7.5}$$

$$\sup_{x_1\in[0,1]}\sup_{1\le J\le N}\left|n^{-1}\sum_{l=1}^{n}\omega_J(\mathbf{X}_l, x_1)\right| = O_p\left(H^{1/2}\right). \tag{3.7.6}$$

PROOF. For simplicity, denote $\omega_J^*(\mathbf{X}_l, x_1) = \omega_J(\mathbf{X}_l, x_1) - \mu_{\omega_J}(x_1)$. Then

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

While $E\omega_J^2(\mathbf{X}_l, x_1)$ is equal to

$$h^{-1}\left\|b_{J,2}\right\|_2^{-2}\int\int K^2(v_1)\left\{I_{J+1,2}(u_2) + \frac{c_{J+1,2}}{c_{J,2}}I_{J,2}(u_2)\right\}f(hv_1 + x_1, u_2)\,dv_1du_2,$$

which implies that $E\omega_J^2(\mathbf{X}_l, x_1) \sim h^{-1}$ and $E\omega_J^2(\mathbf{X}_l, x_1) \gg \mu_{\omega_J}^2(x_1)$. Hence for n sufficiently large

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

for some positive constant c^*. When $r \geq 3$, the r-th moment $E\left|\omega_J(\mathbf{X}_l, x_1)\right|^r$ is

$$\left\{\left\|b_{J,2}\right\|_2\right\}^{-r}\int\int K_h^r(u_1 - x_1)\left\{I_{J+1,2}(u_2) + \left(\frac{c_{J+1,2}}{c_{J,2}}\right)^r I_{J,2}(u_2)\right\}f(u_1, u_2)\,du_1du_2.$$

It is clear that $E\left|\omega_J(\mathbf{X}_l, x_1)\right|^r \sim h^{(1-r)}H^{1-r/2}$ and $\left|E\omega_J(\mathbf{X}_l, x_1)\right|^r \leq CH^{r/2}$ by Lemma 3.7.2, thus $E\left|\omega_J(\mathbf{X}_l, x_1)\right|^r \gg \left|\mu_{\omega_J}(x_1)\right|^r$.

$$\begin{aligned}
& E\left|\omega_J^*(\mathbf{X}_l, x_1)\right|^r = E\left|\omega_J(\mathbf{X}_l, x_1) - \mu_{\omega_J}(x_1)\right|^r \\
\leq\ & 2^{r-1}\left(E\left|\omega_J(\mathbf{X}_l, x_1)\right|^r + \left|\mu_{\omega_J}(x_1)\right|^r\right) \leq \left\{ch^{-1}H^{-1/2}\right\}^{(r-2)} r!E\left|\omega_J^*(\mathbf{X}_l, x_1)\right|^2,
\end{aligned}$$

then there exists a constant $c_* = ch^{-1}H^{-1/2}$ such that

$$E\left|\omega_J^*(\mathbf{X}_l, x_1)\right|^r \leq c_*^{r-2}r!E\left|\omega_J^*(\mathbf{X}_l, x_1)\right|^2,$$

that means the sequence of random variables $\left\{\omega_J^*(\mathbf{X}_l, x_1)\right\}_{l=1}^n$ satisfies the Cramér's condition, hence by the Bernstein's inequality one has for $r = 3$

$$P\left\{\left|n^{-1}\sum_{l=1}^n \omega_J^*(\mathbf{X}_l, x_1)\right| \geq \rho_n\right\} \leq a_1\exp\left(-\frac{q\rho_n^2}{25m_2^2 + 5c_*\rho_n}\right) + a_2(3)\,\alpha\left(\left[\frac{n}{q+1}\right]\right)^{6/7},$$

where

$$\rho_n = \rho\frac{\log n}{\sqrt{nh}}, \text{ with } a_1 = 2\frac{n}{q} + 2\left(1 + \frac{\rho_n^2}{25m_2^2 + 5c_*\rho_n}\right), \text{ with } m_2^2 \sim h^{-1},$$

$$a_2(3) = 11n\left(1 + \frac{5m_3^{6/7}}{\rho_n}\right), \text{ with } m_3 = \max_{1\leq i\leq n}\left\|\omega_J^*(\mathbf{X}_l, x_1)\right\|_3 \leq \left\{C_0\left(2h^{-1}\right)^2\right\}^{1/3}.$$

Observe that $5c_*\rho_n = o(1)$, then by taking q such that $\left[\frac{n}{q+1}\right] \geqslant c_0 \log n$, $q \geqslant c_1 n/\log n$ for some constants c_0, c_1, one has $a_1 = O(n/q) = O(\log n)$, $a_2(3) = o(n^2)$. Assumption (B2) yields that

$$\alpha\left(\left[\frac{n}{q+1}\right]\right)^{6/7} \le \left\{K_0 \exp\left(-\lambda_0\left[\frac{n}{q+1}\right]\right)\right\}^{6/7} \le Cn^{-6\lambda_0 c_0/7}.$$

Thus, for n large enough,

$$P\left\{\frac{1}{n}\left|\sum_{l=1}^{n} \omega_J^*(\mathrm{X}_l, x_1)\right| > \frac{\rho\log n}{\sqrt{nh}}\right\} \le cn^{-c_2\rho^2}\log n + Cn^{2-6\lambda_0 c_0/7}. \tag{3.7.7}$$

We divide the interval $[0,1]$ into $M_n \sim n^6$ equally spaced intervals with disjoint endpoints $0 = x_{1,0} < x_{1,1} < \dots < x_{1,M_n} = 1$. Employing the discretization method, one has

$$\begin{aligned}\sup_{x_1\in[0,1]}\sup_{1\le J\le N}\left|n^{-1}\sum_{l=1}^{n}\omega_J^*(\mathrm{X}_l, x_1)\right| &= \sup_{0\le k\le M_n}\sup_{1\le J\le N}\left|n^{-1}\sum_{l=1}^{n}\omega_J^*(\mathrm{X}_l, x_{1,k})\right| \\ &+ \sup_{1\le k\le M_n}\sup_{1\le J\le N}\sup_{x_1\in[x_{1,k-1},x_{1,k}]}\left|n^{-1}\sum_{l=1}^{n}\left\{\omega_J^*(\mathrm{X}_l, x_1) - \omega_J^*(\mathrm{X}_l, x_{1,k})\right\}\right|.\end{aligned} \tag{3.7.8}$$

By (3.7.7), there exists large enough $\rho > 0$ such that for any $1 \le k \le M_n, 1 \le J \le N$

$$P\left\{\frac{1}{n}\left|\sum_{l=1}^{n}\omega_J^*(\mathrm{X}_l, x_{1,k})\right| > \rho(nh)^{-1/2}\log n\right\} \le n^{-10},$$

which implies that

$$\begin{aligned}&\sum_{n=1}^{\infty} P\left\{\sup_{0\le k\le M_n}\sup_{1\le J\le N}\left|n^{-1}\sum_{l=1}^{n}\omega_J^*(\mathrm{X}_l, x_{1,k})\right| \ge \rho\frac{\log n}{\sqrt{nh}}\right\} \\ \le\ &\sum_{n=1}^{\infty}\sum_{k=1}^{M_n}\sum_{J=1}^{N} P\left\{\left|n^{-1}\sum_{l=1}^{n}\omega_J^*(\mathrm{X}_l, x_{1,k})\right| \ge \rho\frac{\log n}{\sqrt{nh}}\right\} \le \sum_{n=1}^{\infty} NM_n n^{-10} < \infty.\end{aligned}$$

Thus, Borel-Cantelli Lemma entails that

$$\sup_{0\le k\le M_n}\sup_{1\le J\le N}\left|n^{-1}\sum_{l=1}^{n}\omega_J^*(\mathrm{X}_l, x_{1,k})\right| = O_p\left(\log n/\sqrt{nh}\right). \tag{3.7.9}$$

Employing Lipschitz continuity of kernel K, one has for $x_1 \in [x_{1,k-1}, x_{1,k}]$

$$\sup_{1\le k\le M_n}\left|K_h(X_{l,1} - x_1) - K_h(X_{l,1} - x_{1,k})\right| \le C_K M_n^{-1} h^{-2}.$$

Hence one has

$$\sup_{1\leq k\leq M_n}\sup_{1\leq J\leq N}\sup_{x_1\in[x_{1,k-1},x_{1,k}]}\left|n^{-1}\sum_{l=1}^{n}\left\{\omega_J^*\left(\mathbf{X}_l,x_1\right)-\omega_J^*\left(\mathbf{X}_l,x_{1,k}\right)\right\}\right|$$
$$\leq\ C_K M_n^{-1}h^{-2}\sup_{x_2\in[0,1]}\sup_{1\leq J\leq N}\left|B_{J,2}\left(x_2\right)\right|=O\left(M_n^{-1}h^{-2}H^{-1/2}\right).$$

Thus, one has

$$\sup_{1\leq k\leq M_n}\sup_{1\leq J\leq N}\sup_{x_1\in[x_{1,k-1},x_{1,k}]}\left|\frac{1}{n}\sum_{l=1}^{n}\left\{\omega_J^*\left(\mathbf{X}_l,x_1\right)-\omega_J^*\left(\mathbf{X}_l,x_{1,k}\right)\right\}\right|=o\left(\frac{1}{n}\right),\quad(3.7.10)$$

since $M_n\sim n^6$. (3.7.5) follows instantly from (3.7.8), (3.7.9) and (3.7.10). As a result of Lemma 3.7.2 and (3.7.5), (3.7.6) holds. □

The next lemma provides the size of $\tilde{\mathbf{a}}^T\tilde{\mathbf{a}}$.

LEMMA 3.7.5. *Under Assumptions (B2) to (B6), the least square solution* $\tilde{\mathbf{a}}$ *defined in (3.3.6) satisfies*

$$\tilde{\mathbf{a}}^T\tilde{\mathbf{a}}=\tilde{a}_0^2+\sum_{J=1}^{N}\sum_{\alpha=1}^{2}\tilde{a}_{J,\alpha}^2=O_p\left(N\left(\log n\right)^2/n\right).\quad(3.7.11)$$

PROOF. According to (3.3.8) and (3.3.7), $\tilde{\mathbf{a}}=\left(\mathbf{B}^T\mathbf{B}\right)^{-1}\mathbf{B}^T\mathbf{E}$, then

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

As the matrix $\mathbf{B}$ is given in (3.3.7), one has

$$\|\mathbf{B}\tilde{\mathbf{a}}\|_{2,n}^2=\tilde{\mathbf{a}}^T\begin{pmatrix}1 & \\ & \left\langle B_{J,\alpha},B_{J',\alpha'}\right\rangle_{2,n}\end{pmatrix}\tilde{\mathbf{a}}=\tilde{\mathbf{a}}^T\left(n^{-1}\mathbf{B}^T\mathbf{E}\right).\quad(3.7.12)$$

According to (3.7.21), $\|\mathbf{B}\tilde{\mathbf{a}}\|_{2,n}^2$ is bounded below in probability by $(1-A_n)\|\mathbf{B}\tilde{\mathbf{a}}\|_2^2$. By (3.7.2), one has

$$\|\mathbf{B}\tilde{\mathbf{a}}\|_2^2=\left\|\tilde{a}_0^2+\sum_{J=1}^{N}\sum_{\alpha=1}^{2}\tilde{a}_{J,\alpha}^2\right\|_2^2\geq c_0\left(\tilde{a}_0^2+\sum_{J,\alpha}\tilde{a}_{J,\alpha}^2\right).\quad(3.7.13)$$

Meanwhile one can show that $\tilde{\mathbf{a}}^T\left(n^{-1}\mathbf{B}^T\mathbf{E}\right)$ is bounded above by

$$\left(\tilde{a}_0^2+\sum_{J,\alpha}\tilde{a}_{J,\alpha}^2\right)^{1/2}\left[\left\{\frac{1}{n}\sum_{i=1}^{n}\sigma\left(\mathbf{X}_i\right)\varepsilon_i\right\}^2+\sum_{J,\alpha}\left\{\frac{1}{n}\sum_{i=1}^{n}B_{J,\alpha}\left(X_{i,\alpha}\right)\sigma\left(\mathbf{X}_i\right)\varepsilon_i\right\}^2\right]^{1/2}.\quad(3.7.14)$$

Combining (3.7.12), (3.7.13) and (3.7.14), the squared norm $\tilde{\mathbf{a}}^T\tilde{\mathbf{a}}$ is bounded by

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

Using the same truncation version of ε as in Lemma 3.7.10, Bernstein inequality entails that

$$\left|n^{-1}\sum_{i=1}^{n}\sigma\left(\mathbf{X}_i\right)\varepsilon_i\right|+\max_{1\le J\le N,\alpha=1,2}\left|n^{-1}\sum_{i=1}^{n}B_{J,\alpha}\left(X_{i,\alpha}\right)\sigma\left(\mathbf{X}_i\right)\varepsilon_i\right|=O_p\left(\log n/\sqrt{n}\right).$$

Therefore (3.7.11) holds since A_n is of order $o_p(1)$. □

3.7.2 Empirical approximation of the theoretical inner product

Let

$$A_{n,1} = \sup_{J,\alpha}\left|\left\langle 1,B_{J,\alpha}\right\rangle_{2,n}-\left\langle 1,B_{J,\alpha}\right\rangle_{2}\right| = \sup_{J,\alpha}\left|n^{-1}\sum_{i=1}^{n}B_{J,\alpha}\left(X_{i,\alpha}\right)\right|, \quad (3.7.15)$$

$$A_{n,2} = \sup_{J,J',\alpha}\left|\left\langle B_{J,\alpha},B_{J',\alpha}\right\rangle_{2,n}-\left\langle B_{J,\alpha},B_{J',\alpha}\right\rangle_{2}\right|, \quad (3.7.16)$$

$$A_{n,3} = \sup_{1\le J,J'\le N,\alpha\neq\alpha'}\left|\left\langle B_{J,\alpha},B_{J',\alpha'}\right\rangle_{2,n}-\left\langle B_{J,\alpha},B_{J',\alpha'}\right\rangle_{2}\right|. \quad (3.7.17)$$

LEMMA 3.7.6. *Under Assumptions (B2), (B4) and (B6), one has*

$$A_{n,1} = O_p\left(n^{-1/2}\log n\right), \quad (3.7.18)$$

$$A_{n,2} = O_p\left(n^{-1/2}H^{-1/2}\log n\right), \quad (3.7.19)$$

$$A_{n,3} = O_p\left(n^{-1/2}\log n\right). \quad (3.7.20)$$

PROOF. The proof of (3.7.18) follows from Bernstein's inequality immediately, thus is omitted. Here we only prove (3.7.19) and (3.7.20). We will discuss case by case with various α, α', J and J', via Bernstein's inequality. For brevity, set

$$\xi_i=\xi_{i,J,J',\alpha,\alpha'}=n^{-1}\left[B_{J,\alpha}\left(X_{i,\alpha}\right)B_{J',\alpha'}\left(X_{i,\alpha'}\right)-E\left\{B_{J,\alpha}\left(X_{i,\alpha}\right)B_{J,'\alpha'}\left(X_{i,\alpha'}\right)\right\}\right],$$

then

$$A_{n,2}=\sup_{1\le J\le N,\alpha=1,2}\left|\sum_{i=1}^{n}\xi_{i,J,J',\alpha,\alpha}\right|,\quad A_{n,3}=\sup_{1\le J,J'\le N,\alpha\neq\alpha'}\left|\sum_{i=1}^{n}\xi_{i,J,J',\alpha,\alpha'}\right|.$$

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

CASE 1.1 when $\left|J - J'\right| > 1$. The definition of $B_{J,1}$ in (3.2.3) will guarantee that $B_{J,1}\left(X_{i,1}\right) B_{J',1}\left(X_{i,1}\right) = 0$ if $\left|J - J'\right| > 1$.

CASE 1.2 when $J = J'$. By Lemma 3.7.1, the variable ξ_i and its second moment can be simplified as follows

$$\xi_i = n^{-1}\left\{B_{J,1}^2\left(X_{i,1}\right) - 1\right\}, E\xi_i^2 = \frac{1}{n^2}E\left\{B_{J,1}^2\left(X_{i,1}\right) - 1\right\}^2 = \frac{1}{n^2}\left\{EB_{J,1}^4\left(X_{i,1}\right) - 1\right\},$$

in which $EB_{J,1}^4\left(X_{i,1}\right) = \left\|b_{J,1}\right\|_2^{-4}\left(c_{J+1,1} + c_{J+1,1}^4/c_{J,1}^3\right)$.The selection of H will make $EB_{J,1}^4\left(X_{i,1}\right)$ the major term of $\left\{EB_{J,1}^4\left(X_{i,1}\right) - 1\right\}$, then there exist constants $c_{\xi,2}$ and $C'_{\xi,2} > 0$ such that

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

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

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

where $EB_{J,1}^{2k}\left(X_{i,1}\right) \sim 1$ according to Lemma 3.7.1. Hence there exists a constant $C_{\xi,2} > 0$ such that

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

Next step is to verify the Cramér's condition

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

in which $C_{\xi,2}^* = \left(2C_{\xi,2}n^{-1}H^{-1}\right)\max\left(1, 2C_{\xi,2}^2c_{\xi,2}^{-1}\right)$. Applying Lemma 2.6.2 and Borel - Cantelli lemma, when $J = J'$, $\alpha = \alpha' = 1$, one has (3.7.19).

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

$$E\xi_i^2 = n^{-2}\left[EB_{J,1}^2\left(X_{i,1}\right) B_{J+1,1}^2\left(X_{i,1}\right) - \left\{EB_{J,1}\left(X_{i,1}\right) B_{J+1,1}\left(X_{i,1}\right)\right\}^2\right],$$

where $\left\{EB_{J,1}\left(X_{i,1}\right) B_{J+1,1}\left(X_{i,1}\right)\right\}^2 \sim 1$, $EB_{J,1}^2\left(X_{i,1}\right) B_{J+1,1}^2\left(X_{i,1}\right) \sim H^{-1}$, according to Lemma 3.7.1. Hence, $E\xi_i^2 \sim H^{-1}$. The k-th moment is given by

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

where $\left|EB_{J,1}\left(X_{i,1}\right)B_{J+1,1}\left(X_{i,1}\right)\right|^k \sim 1$ and $E\left|B_{J,1}\left(X_{i,1}\right)B_{J+1,1}\left(X_{i,1}\right)\right|^k \sim H^{1-k}$, according to Lemma 3.7.1. Hence there exists a constant $C_{\xi,3} > 0$ such that

$$E\left|\xi_i\right|^k \leq C_{\xi,3}^k n^{-k}2^{k-1}H^{1-k}.$$

Similar as in Case 1.2, (3.7.19) follows by using Bernstein's inequality.

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

CASE 3 when $\alpha \neq \alpha'$. Without loss of generality, suppose $\alpha = 1, \alpha' = 2$. First we still need to calculate the order of second moment $E\xi_i^2$,

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

The boundedness of the density function $f\left(x_1, x_2\right)$ implies that

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

for some constant $C_{B,1} > 0$, where the last step is derived by Lemma 3.7.1. As a consequence, $\left|EB_{J,1}\left(X_{i,1}\right)B_{J',2}\left(X_{i,2}\right)\right|^k \leq C_{B,1}^k H^k$. Meanwhile, by Assumption (B4) and Lemma 3.7.1,

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

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

$$c_\xi n^{-2} \leq E\xi_i^2 \leq C'_\xi n^{-2}.$$

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

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

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

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

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

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

Employing the Bernstein's inequality and the fact that $E\xi_i^2 \sim n^{-2}$, one has

$$\sup_{1 \leq J, J' \leq N} \sup_{\alpha \neq \alpha'} \left|\sum_{i=1}^n \frac{1}{n}\left[B_{J,\alpha}\left(X_{i,\alpha}\right) B_{J',\alpha'}\left(X_{i,\alpha'}\right) - E\left\{B_{J,\alpha}\left(X_{i,\alpha}\right) B_{J,'\alpha'}\left(X_{i,\alpha'}\right)\right\}\right]\right|$$

is of order $O_p\left(n^{-1/2} \log n\right)$. So the proof of (3.7.19) and (3.7.20) is completed. □

LEMMA 3.7.7. *Under Assumptions (B2), (B4) and (B6), 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*

$$A_n = \sup_{g_1, g_2 \in G^{(-1)}} \frac{\left|\langle g_1, g_2\rangle_{2,n} - \langle g_1, g_2\rangle_2\right|}{\left\|g_1\right\|_2 \left\|g_2\right\|_2} = O_p\left(\frac{\log n}{n^{1/2} H^{1/2}}\right) = o_p(1). \qquad (3.7.21)$$

PROOF. For every $g_1, g_2 \in G^{(-1)}$, one can write

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

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

$$\left|\langle g_1, g_2\rangle_{2,n} - \langle g_1, g_2\rangle_2\right| \le \left|\sum_{J,\alpha} \langle a_0', a_{J,\alpha} B_{J,\alpha}\rangle_{2,n}\right| + \left|\sum_{J',\alpha'} \left\langle a_0, a'_{J',\alpha'} B_{J',\alpha'}\right\rangle_{2,n}\right|$$
$$+ \sum_{J,J',\alpha,\alpha'} \left|a_{J,\alpha}\right| \left|a'_{J',\alpha'}\right| \left|\left\langle B_{J,\alpha}, B_{J',\alpha'}\right\rangle_{2,n} - \left\langle B_{J,\alpha}, B_{J',\alpha'}\right\rangle_2\right| = L_1 + L_2 + L_3.$$

The equivalence of norms given in equation (3.7.2) and definition (3.7.15) lead to

$$\begin{aligned} L_1 &\le A_{n,1} \cdot \left|a_0'\right| \cdot \sum_{J,\alpha} \left|a_{J,\alpha}\right| \le C_0 A_{n,1} \left(a_0'^2 + \sum_{J,\alpha} a'^2_{J,\alpha}\right)^{1/2} \left(\sum_{J,\alpha} a^2_{J,\alpha}\right)^{1/2} N^{1/2} \\ &\le C_{A,1} A_{n,1} \|g_1\|_2 \|g_2\|_2 H^{-1/2} = O_p\left(n^{-1/2} H^{-1/2} \log n\right) \|g_1\|_2 \|g_2\|_2 . \end{aligned}$$

Similarly, one has

$$L_2 \le C'_{A,1} A_{n,1} \|g_1\|_2 \|g_2\|_2 H^{-1/2} = O_p\left(n^{-1/2} H^{-1/2} \log n\right) \|g_1\|_2 \|g_2\|_2 .$$

For the last term L_3, one has, by definitions (3.7.16) and (3.7.17)

$$\begin{aligned} L_3 &\le \sum_{J,J',\alpha,\alpha'} \left|a_{J,\alpha}\right| \left|a'_{J',\alpha'}\right| \max\left(A_{n,2}, A_{n,3}\right) \\ &\le C_{A,2} \max\left(A_{n,2}, A_{n,3}\right) \left\{\sum_{J,\alpha} a^2_{J,\alpha}\right\}^{1/2} \left\{\sum_{J',\alpha'} a^2_{J',\alpha'}\right\}^{1/2} \\ &\le C_{A,2} \max\left(A_{n,2}, A_{n,3}\right) \|g_1\|_2 \|g_2\|_2 = O_p\left(n^{-1/2} H^{-1/2} \log n\right) \|g_1\|_2 \|g_2\|_2 . \end{aligned}$$

Therefore, statement (3.7.21) is established. □

3.7.3 Proof of Lemma 3.5.2

In the following, denote $\mathbf{V}$ as the theoretical inner product of the B-spline basis $\left\{1, B_{J,\alpha}(x_\alpha), J = 1, ..., N, \alpha = 1, 2\right\}$, i.e.

$$\mathbf{V} = \begin{pmatrix} 1 & 0_{2N}^T \\ 0_{2N} & \left\langle B_{J,\alpha}, B_{J',\alpha'}\right\rangle_2 \end{pmatrix}_{\substack{1\le\alpha,\alpha'\le 2, \\ 1\le J,J'\le N}} = \begin{pmatrix} 1 & 0_N^T & 0_N^T \\ 0_N & \mathbf{V}_{11} & \mathbf{V}_{12} \\ 0_N & \mathbf{V}_{21} & \mathbf{V}_{22} \end{pmatrix}. \quad (3.7.22)$$

where $\mathbf{0}_p = \{0, ..., 0\}^T$. Let $\mathbf{S}$ be the inverse matrix of $\mathbf{V}$, i.e.

$$\mathbf{S} = \mathbf{V}^{-1} = \begin{pmatrix} 1 & \mathbf{0}_N^T & \mathbf{0}_N^T \\ \mathbf{0}_N & \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{0}_N & \mathbf{S}_{21} & \mathbf{S}_{22} \end{pmatrix}. \tag{3.7.23}$$

The next lemma on the positive definiteness of matrices $\mathbf{V}$ and $\mathbf{S}$ is a sufficient step to achieve Lemmas 3.7.9 and 3.7.10.

LEMMA 3.7.8. *Under Assumptions (B4) and (B6), for the matrices* $\mathbf{V}$ *and* $\mathbf{S}$ *defined in (3.7.22) and (3.7.23) respectively, there exist constants* $C_V > c_V > 0$ *and* $C_S > c_S > 0$ *such that*

$$c_V \mathbf{I}_{2N+1} \le \mathbf{V} \le C_V \mathbf{I}_{2N+1}, \ c_S \mathbf{I}_{2N+1} \le \mathbf{S} \le C_S \mathbf{I}_{2N+1}. \tag{3.7.24}$$

PROOF. Take a real vector $\beta = (\beta_0, \beta_{1,1}, ..., \beta_{N,1}, \beta_{1,2}, ..., \beta_{N,2})^T \in R^{2N+1}$, one has

$$\left\|\beta^T \mathbf{B}_2(\mathbf{x})\right\|_2^2 = \beta^T \begin{pmatrix} 1 & \mathbf{0}_{2N}^T \\ \mathbf{0}_{2N} & \left\langle B_{J,\alpha}, B_{J',\alpha'} \right\rangle_2 \end{pmatrix} \beta = \beta^{\mathbf{T}} \mathbf{V} \beta,$$

where denote $\mathbf{B}_2(\mathbf{x}) = \{1, B_{1,1}(X_1), ..., B_{N,2}(X_2)\}^T$. According to (3.7.2), there exist constants $C_V > c_V > 0$ such that

$$C_V \left(\beta_0^2 + \sum_{J,\alpha} \beta_{J,\alpha}^2\right) \ge \left\|\beta^T \mathbf{B}_2(\mathbf{x})\right\|_2^2 = \beta_0^2 + \left\|\sum_{J,\alpha} \beta_{J,\alpha} B_{J,\alpha}(x_\alpha)\right\|_2^2,$$

$$\left\|\beta^T \mathbf{B}_2(\mathbf{x})\right\|_2^2 = \beta_0^2 + \left\|\sum_{J,\alpha} \beta_{J,\alpha} B_{J,\alpha}(x_\alpha)\right\|_2^2 \ge c_V \left(\beta_0^2 + \sum_{J,\alpha} \beta_{J,\alpha}^2\right),$$

thus one concludes that

$$C_V \beta^{\mathbf{T}} \beta = C_V \left(\beta_0^2 + \sum_{J,\alpha} \beta_{J,\alpha}^2\right) \ge \beta^{\mathbf{T}} \mathbf{V} \beta \ge c_V \left(\beta_0^2 + \sum_{J,\alpha} \beta_{J,\alpha}^2\right) = c_V \beta^{\mathbf{T}} \beta,$$

which implies that $c_V \mathbf{I}_{2N+1} \le \mathbf{V} \le C_V \mathbf{I}_{2N+1}$. The second half of (3.7.24) follows by changing β by $\mathbf{V}^{-1/2}\beta$. □

As an application of the above Lemma, for any $(2N+1)$-vectors $\mathbf{x}$ and $\mathbf{y}$

$$\mathbf{x}^{\mathbf{T}} \mathbf{S} \mathbf{y} \le C_S(2N+1) \|\mathbf{x}\| \cdot \|\mathbf{y}\|, \tag{3.7.25}$$

where C_S is the same as in (3.7.24). Note that $\tilde{\mathbf{a}}$ given in (3.3.8) can be rewritten as

$$\tilde{\mathbf{a}} = \left(\mathbf{B}^T\mathbf{B}\right)^{-1}\mathbf{B}^T\mathbf{E} = \left(\frac{1}{n}\mathbf{B}^T\mathbf{B}\right)^{-1}\left(\frac{1}{n}\mathbf{B}^T\mathbf{E}\right) = \left(\mathbf{V}+\mathbf{V}^*\right)^{-1}\left(\frac{1}{n}\mathbf{B}^T\mathbf{E}\right), \qquad (3.7.26)$$

where $\mathbf{V}^*$ is the difference between empirical and theoretical inner product matrices, i.e.

$$\mathbf{V}^* = \begin{pmatrix} 0 & \mathbf{0}_{2N}^T \\ \mathbf{0}_{2N} & \left\langle B_{J,\alpha}, B_{J',\alpha'}\right\rangle_{2,n} - \left\langle B_{J,\alpha}, B_{J',\alpha'}\right\rangle_2 \end{pmatrix}_{\substack{1\le\alpha,\alpha'\le 2,\\ 1\le J,J'\le N}}.$$

Now define $\hat{\mathbf{a}} = \left\{\hat{a}_0, \hat{a}_{1,1}, ..., \hat{a}_{N,1}, \hat{a}_{1,2}, ..., \hat{a}_{N,2}\right\}^T$ by replacing $\left(\mathbf{V}+\mathbf{V}^*\right)^{-1}$ with $\mathbf{V}^{-1} = \mathbf{S}$ in the above formula, that is

$$\hat{\mathbf{a}} = \mathbf{V}^{-1}\left(n^{-1}\mathbf{B}^T\mathbf{E}\right) = \mathbf{S}\left(n^{-1}\mathbf{B}^T\mathbf{E}\right). \qquad (3.7.27)$$

and define

$$\hat{\Psi}_v^{(2)}(x_1) = n^{-1}\sum_{i=1}^{n}\sum_{J=1}^{N}\hat{a}_{J,2}\omega_J(\mathbf{X}_i, x_1). \qquad (3.7.28)$$

The next lemma shows that the difference between $\Psi_v^{(2)}(x_1)$ in (3.5.6) and $\hat{\Psi}_v^{(2)}(x_1)$ in (3.7.28) is negligible uniformly over $x_1 \in [0,1]$.

LEMMA 3.7.9. *Under Assumptions (B2) to (B6),*

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

PROOF. According to (3.7.26) and (3.7.27), one has $\mathbf{V}\,\hat{\mathbf{a}} = \left(\mathbf{V}+\mathbf{V}^*\right)\,\tilde{\mathbf{a}}$, which implies that $\mathbf{V}^*\tilde{\mathbf{a}} = \mathbf{V}\left(\hat{\mathbf{a}} - \tilde{\mathbf{a}}\right)$. Using (3.7.19) and (3.7.20), one obtains that

$$\left\|\mathbf{V}\left(\hat{\mathbf{a}} - \tilde{\mathbf{a}}\right)\right\| = \left\|\mathbf{V}^*\tilde{\mathbf{a}}\right\| \le O_p\left(n^{-1/2}H^{-1}\log n\right)\left\|\tilde{\mathbf{a}}\right\|.$$

By Lemma 3.7.5, $\left\|\tilde{\mathbf{a}}\right\| = O_p\left(n^{-1/2}N^{1/2}\log n\right)$, so one has

$$\left\|\mathbf{V}\left(\hat{\mathbf{a}} - \tilde{\mathbf{a}}\right)\right\| \le O_p\left\{(\log n)^2 n^{-1}N^{3/2}\right\}.$$

Thus according to Lemma 3.7.8, one has

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

Using Lemma 3.7.5 again, one has

$$\|\hat{\mathbf{a}}\| \le \|(\hat{\mathbf{a}}-\tilde{\mathbf{a}})\| + \|\tilde{\mathbf{a}}\| = O_p\left(\log n\sqrt{N/n}\right). \tag{3.7.29}$$

Hence

$$\left|\Psi_v^{(2)}(x_1) - \hat{\Psi}_v^{(2)}(x_1)\right| = \left|\sum_{J=1}^{N}\left(\tilde{a}_{J,2}-\hat{a}_{J,2}\right)\frac{1}{n}\sum_{l=1}^{n}\omega_J(\mathbf{X}_l, x_1)\right|.$$

Cauchy-Schwartz inequality implies that

$$\sup_{x\in[0,1]}\left|\Psi_v^{(2)}(x_1) - \hat{\Psi}_v^{(2)}(x_1)\right| \le \sqrt{N}O_p\left(\frac{(\log n)^2}{nH}\right)O_p\left(H^{1/2}\right) = O_p\left(\frac{(\log n)^2}{nH}\right).$$

Therefore the lemma follows. □

Lemma 3.7.10. *Under Assumptions (B2) to (B6), for* $\hat{\Psi}_v^{(2)}(x_1)$ *as defined in (3.7.28)*

$$\sup_{x_1\in[0,1]}\left|\hat{\Psi}_v^{(2)}(x_1)\right| = \sup_{x_1\in[0,1]}\left|n^{-1}\sum_{i=1}^{n}K_h\left(X_{i,1}-x_1\right)\sum_{J=1}^{N}\hat{a}_{J,2}B_{J,2}\left(X_{i,2}\right)\right| = O_p(H).$$

Proof. Note that

$$\begin{aligned}\left|\hat{\Psi}_v^{(2)}(x_1)\right| &\le \left|\sum_{J=1}^{N}\hat{a}_{J,2}\mu_{\omega_J}(x_1)\right| + \left|\sum_{J=1}^{N}\hat{a}_{J,2}n^{-1}\sum_{i=1}^{n}\left\{\omega_J(\mathbf{X}_i, x_1) - \mu_{\omega_J}(x_1)\right\}\right| \\ &= Q_1(x_1) + Q_2(x_1).\end{aligned} \tag{3.7.30}$$

By Cauchy-Schwartz inequality, one has

$$Q_2^2(x_1) \le \sum_{J=1}^{N}\hat{a}_{J,2}^2\sum_{J=1}^{N}\left\{\sup_{x_1\in[0,1]}\left|n^{-1}\sum_{i=1}^{n}\left\{\omega_J(\mathbf{X}_i, x_1) - \mu_{\omega_J}(x_1)\right\}\right|\right\}^2.$$

Observe that $\|\hat{\mathbf{a}}\| = O_p\left(\log n\sqrt{N/n}\right)$ as given in (3.7.29) and

$$\sup_{x_1\in[0,1]}\left|n^{-1}\sum_{i=1}^{n}\left\{\omega_J(\mathbf{X}_i, x_1) - \mu_{\omega_J}(x_1)\right\}\right| = O_p\left(\log n/\sqrt{nh}\right),$$

given in Lemma 3.7.4, so by Assumptions (B5) and (B6)

$$\begin{aligned}\sup_{x_1\in[0,1]}Q_2(x_1) &= O_p\left(\log n\sqrt{N/n}\right)\sqrt{N}O_p\left(\frac{\log n}{\sqrt{nh}}\right) = O_p\left\{\frac{N(\log n)^2}{n\sqrt{h}}\right\} \\ &= O_p\left\{(\log n)^3/\sqrt{n}\right\}.\end{aligned} \tag{3.7.31}$$

Using the discretization idea again as in the proof of Lemma 3.7.4, one has

$$\sup_{x_1\in[0,1]} Q_1(x_1) \le \max_{1\le k\le M_n}\left|\sum_{J=1}^{N}\hat{a}_{J,2}\mu_{\omega_J}(x_{1,k})\right| + \tag{3.7.32}$$

$$\max_{1\le k\le M_n}\sup_{x_1\in[x_{1,k-1},x_{1,k}]}\left|\sum_{J=1}^{N}\hat{a}_{J,2}\mu_{\omega_J}(x_1)-\sum_{J=1}^{N}\hat{a}_{J,2}\mu_{\omega_J}(x_{1,k})\right| = T_1+T_2,$$

where $M_n \sim n$. Define next

$$W_1 = \max_{1\le k\le M_n}\left|n^{-1}\sum_{1\le i\le n}\sum_{1\le J,J'\le N}\mu_{\omega_J}(x_{1,k})\, s_{J+N+1,J'+1}B_{J',1}(X_{i,1})\sigma(\mathbf{X}_i)\,\varepsilon_i\right|,$$

$$W_2 = \max_{1\le k\le M_n}\left|n^{-1}\sum_{1\le i\le n}\sum_{1\le J,J'\le N}\mu_{\omega_J}(x_{1,k})\, s_{J+N+1,J'+N+1}B_{J',2}(X_{i,2})\sigma(\mathbf{X}_i)\,\varepsilon_i\right|,$$

then it is clear that $T_1 \le W_1 + W_2$. To show that both of the two terms W_1 and W_2 have order $O_p(H)$, we truncate the random variable ε_i at the level of

$$D_n = n^{\theta_0}\left(\frac{1}{2+\delta} < \theta_0 < \frac{2}{5}\right). \tag{3.7.33}$$

where δ is the same as in Assumption (B3). Without loss of generality, we only give the proof of $W_1 = O_p(H)$. Let

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

$$U_{i,k} = \sum_{1\le J,J'\le N}\mu_{\omega_J}(x_{1,k})\, s_{J+N+1,J'+1}B_{J',1}(X_{i,1})\sigma(\mathbf{X}_i)\,\varepsilon_{i,D}^{*},$$

and denote W_1^D as the truncated centered version of W_1, i.e.,

$$W_1^D = \max_{1\le k\le M_n}\left|n^{-1}\sum_{i=1}^{n}U_{i,k}\right|. \tag{3.7.34}$$

Next we show that $\left|W_1 - W_1^D\right| = O_p(H)$. Note that $\left|W_1 - W_1^D\right| \le \Lambda_1 + \Lambda_2$, where

$$\Lambda_1 = \max_{1\le k\le M_n}\left|\frac{1}{n}\sum_{i=1}^{n}\sum_{1\le J,J'\le N}\mu_{\omega_J}(x_{1,k})\, s_{J+N+1,J'+1}B_{J',1}(X_{i,1})\sigma(\mathbf{X}_i)\,E\left(\varepsilon_{i,D}^{-}\,|\mathbf{X}_i\right)\right|,$$

$$\Lambda_2 = \max_{1\le k\le M_n}\left|\frac{1}{n}\sum_{i=1}^{n}\sum_{1\le J,J'\le N}\mu_{\omega_J}(x_{1,k})\, s_{J+N+1,J'+1}B_{J',1}(X_{i,1})\sigma(\mathbf{X}_i)\,\varepsilon_{i,D}^{+}\right|.$$

Let $\boldsymbol{\mu}_\omega(x_{1,k}) = \left\{\mu_{\omega_1}(x_{1,k}), \cdots, \mu_{\omega_N}(x_{1,k})\right\}^T$, then

$$\Lambda_1 = \max_{1\le k\le M_n} \left| \boldsymbol{\mu}_\omega(x_{1,k})^T \mathrm{S}_{21} \left\{ n^{-1}\sum_{i=1}^n B_{J',1}(X_{i,1})\sigma(\mathbf{X}_i) E\left(\varepsilon_{i,D}^- |\mathbf{X}_i\right)\right\}_{J'=1}^N \right|$$

$$\le C_S \max_{1\le k\le M_n} \left\{ \sum_{J=1}^N \mu_{\omega_J}^2(x_{1,k}) \sum_{J=1}^N \left\{ \frac{1}{n}\sum_{i=1}^n B_{J,1}(X_{i,1})\sigma\left(\bar{\mathbf{X}}_i\right) E\left(\varepsilon_{i,D}^- |\mathbf{X}_i\right)\right\}^2 \right\}^{1/2},$$

according to (3.7.25). By Assumption (B3),

$$\left|E\left(\varepsilon_{i,D}^- |\mathbf{X}_i\right)\right| = \left|E\left(\varepsilon_{i,D}^+ |\mathbf{X}_i\right)\right| \le \frac{E\left(|\varepsilon_i|^{2+\delta} |\mathbf{X}_i\right)}{D_n^{1+\delta}} \le M_\delta D_n^{-(1+\delta)},$$

and $\sup_{J,\alpha} \left|\frac{1}{n}\sum_{i=1}^n B_{J,1}(X_{i,1})\sigma(\mathbf{X}_i)\right| = O_p(\log n/\sqrt{n})$ by Bernstein inequality given in Lemma 2.6.2. Therefore

$$\begin{aligned} \Lambda_1 &\le M_\delta D_n^{-(1+\delta)} \max_{1\le k\le M_n} \left[\sum_{J=1}^N \mu_{\omega_J}^2(x_{1,k}) \sum_{J=1}^N \left\{ \frac{1}{n}\sum_{i=1}^n B_{J,1}(X_{i,1})\sigma(\mathbf{X}_i)\right\}^2 \right]^{1/2} \\ &= O_p\left\{ N D_n^{-(1+\delta)} \log^2 n/n \right\} = O_p(H), \end{aligned}$$

where the last step follows from the choice of D_n in (3.7.33). Meanwhile

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

since $\delta > 1/2$. By Borel-Cantelli Lemma, one has with probability 1,

$$n^{-1}\sum_{i=1}^n \sum_{1\le J,J'\le N} \mu_{\omega_J}(x_{1,k})\, s_{J+N+1,J'+1} B_{J',1}(X_{i,1})\sigma(\mathbf{X}_i)\,\varepsilon_{i,D}^+ = 0$$

for large n. Therefore, one has $\left|W_1 - W_1^D\right| \le \Lambda_1 + \Lambda_2 = O_p(H)$. Next we want to show that $W_1^D = O_p(H)$, with W_1^D defined in (3.7.34). Since

$$U_{i,k} = \boldsymbol{\mu}_\omega(x_{1,k})^T \mathrm{S}_{21} \left\{B_{1,1}(X_{i,1}), \cdots, B_{1,N}(X_{i,1})\right\}^T \sigma(\mathbf{X}_i)\,\varepsilon_{i,D}^*,$$

so the variance of $U_{i,k}$ is

$$\boldsymbol{\mu}_\omega(x_{1,k})^T \mathrm{S}_{21}\mathrm{var}\left(\left\{B_{1,1}(X_{i,1}), \cdots, B_{N,1}(X_{i,1})\right\}^T \sigma(\mathbf{X}_i)\,\varepsilon_{i,D}^*\right) \mathrm{S}_{21}\boldsymbol{\mu}_\omega(x_{1,k}).$$

According to Assumption (B3), $\sigma(\mathbf{x})$ is continuous on a compact set $[0,1]^d$, so it is clear that $c_\sigma^2 \mathbf{V}_{11} \le \operatorname{var}\left(\{B_{1,1}(X_{i,1}), \cdots, B_{N,1}(X_{i,1})\}^T \sigma(\mathbf{X}_i)\right) \le C_\sigma^2 \mathbf{V}_{11}$. Thus

$$\begin{aligned}
\operatorname{var}(U_{i,k}) &\sim \boldsymbol{\mu}_\omega(x_{1,k})^T \mathbf{S}_{21}\mathbf{V}_{11}\mathbf{S}_{21}\boldsymbol{\mu}_\omega(x_{1,k}) V_{\varepsilon,D} \\
&= \boldsymbol{\mu}_\omega(x_{1,k})^T \mathbf{S}_{21}\boldsymbol{\mu}_\omega(x_{1,k}) V_{\varepsilon,D},
\end{aligned}$$

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

$$c_S c_\sigma^2 \{\kappa(x_{1,k})\}^2 V_{\varepsilon,D} \le \operatorname{var}(U_{i,k}) \le C_S C_\sigma^2 \{\kappa(x_{1,k})\}^2 V_{\varepsilon,D}.$$

When $r \ge 3$, the r-th moment $EU^r_{i,k}$ is

$$\begin{aligned}
E|U_{i,k}|^r &= E\left|\sum_{1\le J,J'\le N} \mu_{\omega_J}(x_{1,k})\, s_{J+N+1,J'+1} B_{J',1}(X_{i,1})\sigma(\mathbf{X}_i)\,\varepsilon^*_{i,D}\right|^r \\
&\le E\left|\sum_{1\le J,J'\le N} \mu_{\omega_J}(x_{1,k})\, s_{J+N+1,J'+1} B_{J',1}(X_{i,1})\sigma(\mathbf{X}_i)\right|^r E\left(\left|\varepsilon^*_{i,D}\right|^r |\mathbf{X}_i\right) \\
&\le E\left|\sum_{1\le J,J'\le N} \mu_{\omega_J}(x_{1,k})\, s_{J+N+1,J'+1} B_{J',1}(X_{i,1})\sigma(\mathbf{X}_i)\right|^r D_n^{r-2} V_{\varepsilon,D},
\end{aligned}$$

while

$$\begin{aligned}
&E\left|\sum_{1\le J,J'\le N} \mu_{\omega_J}(x_{1,k})\, s_{J+N+1,J'+1} B_{J',1}(X_{i,1})\sigma(\mathbf{X}_i)\right|^r \\
&= E\left|\boldsymbol{\mu}_\omega(x_{1,k})^T \mathbf{S}_{21}\{B_{1,1}(X_{i,1}), \cdots, B_{1,N}(X_{i,1})\}^T \sigma(\mathbf{X}_i)\right|^r \\
&\le C_S^r C_\sigma^r E\left|\boldsymbol{\mu}_\omega(x_{1,k})^T \{B_{1,1}(X_{i,1}), \cdots, B_{1,N}(X_{i,1})\}^T\right|^r \\
&\le C_S^r C_\sigma^r \{\kappa(x_{1,k})\}^r E\left\{\sum_{J=1}^N B^2_{J,1}(X_{i,1})\right\}^{r/2} \\
&\le C_S^r C_\sigma^r \{\kappa(x_{1,k})\}^r O\left(H^{1-r/2}\right).
\end{aligned}$$

Therefore

$$\begin{aligned}
.E|U_{i,k}|^r &\le C_S^r C_\sigma^r \{\kappa(x_{1,k})\}^r O\left(H^{1-r/2}\right) D_n^{r-2} V_{\varepsilon,D} \\
&\le \left\{c_0 \kappa(x_{1,k}) D_n H^{-1/2}\right\}^{r-2} r! E|U_{i,k}|^2 < +\infty,
\end{aligned}$$

which means the sequence of random variables $\{U_{i,k}\}_{i=1}^{n}$ satisfies the Cramér's condition with Cramér's constant equal to $c_* = c_0 \kappa\left(x_{1,k}\right) D_n H^{-1/2}$, hence by the Bernstein's inequality we have for $r = 3$

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

where

$$\rho_n = \rho H,\ a_1 = 2\frac{n}{q} + 2\left(1 + \frac{\rho_n^2}{25m_2^2 + 5c_*\rho_n}\right),\ a_2(3) = 11n\left(1 + \frac{5m_3^{6/7}}{\rho_n}\right),$$

$$m_2^2 \sim \left\{\kappa\left(x_{1,k}\right)\right\}^2 V_{\varepsilon,D},\ m_3 \leq \left\{c\left\{\kappa\left(x_{1,k}\right)\right\}^3 H^{-1/2} D_n V_{\varepsilon,D}\right\}^{1/3}.$$

Observe that $5c_*\rho_n = o(1)$, then by taking q such that $\left[\frac{n}{q+1}\right] \geqslant c_0 \log n$, $q \geqslant c_1 n/\log n$ for some constants c_0, c_1, one has $a_1 = O(n/q) = O(\log n)$, $a_2(3) = o\left(n^2\right)$. Assumption (B2) yields that

$$\alpha\left(\left[\frac{n}{q+1}\right]\right)^{6/7} \leq \left\{K_0 \exp\left(-\lambda_0\left[\frac{n}{q+1}\right]\right)\right\}^{6/7} \leq Cn^{-6\lambda_0 c_0/7},$$

and as $n \to \infty$, one has

$$\frac{q\rho_n^2}{25m_2^2 + 5c_*\rho_n} \sim \frac{q\rho_n}{c_*} = \frac{\rho n^{2/5}}{c_0(\log n)^{5/2} D_n} \to +\infty.$$

Thus, for n large enough,

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

Taking c_0, ρ large enough, $P\left\{\left|\frac{1}{n}\sum_{i=1}^{n} U_{i,k}\right| > \rho H\right\} \leq n^{-3}$, for large n. Hence

$$\sum_{n=1}^{\infty} P\left(\left|W_1^D\right| \geq \rho H\right) = \sum_{n=1}^{\infty}\sum_{k=1}^{M_n} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} U_{i,k}\right| \geq \rho H\right) \leq \sum_{n=1}^{\infty} M_n n^{-3} < \infty.$$

Thus, Borel-Cantelli Lemma entails that $W_1^D = O_p(H)$. Noting that $\left|W_1 - W_1^D\right| = O_p(H)$, one obtains that $W_1 = O_p(H)$. Similarly one can show that $W_2 = O_p(H)$. Hence

$$T_1 \leq W_1 + W_2 = O_p(H). \tag{3.7.35}$$

Employing Lipschitz continuity of kernel K, the term T_2^2 is bounded by

$$\|\hat{a}\|^2 \max_{1\le k\le M_n} \sup_{x_1\in[x_{1,k-1},x_{1,k}]} \sum_{J=1}^{N} \left\{\mu_{\omega_J}(x_1) - \mu_{\omega_J}(x_{1,k})\right\}^2 \le \|\hat{a}\|^2 \times$$

$$\max_{1\le k\le M_n} \sup_{x_1\in[x_{1,k-1},x_{1,k}]} \sum_{J=1}^{N} E\left[\left\{K_h(X_{11}-x_1) - K_h(X_{11}-x_{1,k})\right\}^2 \left\{B_{J,2}(X_{12})\right\}^2\right]$$

Therefore, according to Assumption (B5), Lemma 3.7.1 (ii) and (3.7.29),

$$T_2 \le O_p\left(\frac{N^{1/2}\log n}{n^{1/2}}\right) \frac{\left\{\sum_{J=1}^{N} EB_{J,2}^2(X_{12})\right\}^{1/2}}{h^2 M_n} = O\left(\frac{N^{1/2}\log n}{n^{1/2}h^2 M_n}\right) = o_p\left(n^{-1/2}\right). \tag{3.7.36}$$

Combining (3.7.32), (3.7.35) and (3.7.36) one has $\sup_{x_1\in[0,1]} Q_1(x_1) = O_p(H)$. The desired result follows from (3.7.30) and (3.7.31). □

CHAPTER 4

Spline Single-Index Prediction Model

4.1 Introduction

Consider the stochastic heteroscedastic regression model given in (1.1.1), an attractive dimension reduction method to deal with the "curse of dimensionality" is the single-index model, similar to the first step of projection pursuit regression, see Friedman and Stuetzle (1981), Hall (1989), Huber (1985), Chen (1991). The basic appeal of single-index model is its simplicity: the d-variate function $m(\mathbf{x}) = m(x_1, ..., x_d)$ is expressed as a univariate function of $\mathbf{x}^T\boldsymbol{\theta}_0 = \sum_{p=1}^{d} x_p\theta_{0,p}$. Over the last two decades, many authors had devised various intelligent estimators of the single-index coefficient vector $\boldsymbol{\theta}_0 = \left(\theta_{0,1}, ..., \theta_{0,d}\right)^T$, for instance, Powell, Stock and Stoker (1989), Härdle and Stoker (1989), Ichimura (1993), Klein and Spady (1993), Härdle, Hall and Ichimura (1993), Horowitz and Härdle (1996), Carroll, Fan, Gijbels and Wand (1997), Xia and Li (1999), Hristache, Juditski and Spokoiny (2001). More recently, Xia, Tong, Li and Zhu (2002) proposed the minimum average variance estimation (MAVE) for several index vectors.

All the aforementioned methods assume that the d-variate regression function $m(\mathbf{x})$ is exactly a univariate function of some $\mathbf{x}^T\boldsymbol{\theta}_0$ and obtain a root-n consistent estimator of $\boldsymbol{\theta}_0$. If this model is misspecified (m is not a genuine single-index function), however, a goodness-of-fit test then becomes necessary and the estimation of $\boldsymbol{\theta}_0$ must be redefined, see Xia, Li, Tong and Zhang (2004). Here instead of presuming that underlying true function m is a single-index function, a univariate function g is estimated that optimally approximates the multivariate function m in the sense of

$$g(\nu) = E\left[m(\mathbf{X}) | \mathbf{X}^T\boldsymbol{\theta}_0 = \nu\right], \tag{4.1.1}$$

where the unknown parameter $\boldsymbol{\theta}_0$ is called the SIP coefficient, used for simple interpretation once estimated; $\mathbf{X}^T\boldsymbol{\theta}_0$ is the latent SIP variable; and g is a smooth but unknown function used for further data summary, called the link prediction function. Our method therefore is clearly interpretable regardless of the goodness-of-fit of the single-index model, making it much more relevant in applications.

Estimators of $\boldsymbol{\theta}_0$ and g are proposed in this chapter based on weakly dependent sample, which includes many existing nonparametric time series models, that are (i) computationally expedient and (ii) theoretically reliable. Estimation of both $\boldsymbol{\theta}_0$ and g has been done via the kernel smoothing techniques in existing literature, while polynomial spline smoothing is used here. The greatest advantages of spline smoothing, as pointed out in Huang and Yang (2004), Xue and Yang (2006 b) are its simplicity and fast computation. The proposed procedure involves two stages: estimation of $\boldsymbol{\theta}_0$ by some $\sqrt{n}$-consistent $\hat{\boldsymbol{\theta}}$, minimizing an empirical version of the mean squared error, $R(\boldsymbol{\theta}) = E\{Y - E(Y | \mathbf{X}^T\boldsymbol{\theta})\}^2$; spline smoothing of Y on $\mathbf{X}^T\hat{\boldsymbol{\theta}}$ to obtain a cubic spline estimator $\hat{g}$ of g. The best single-index approximation to $m(\mathbf{x})$ is then $\hat{m}(\mathbf{x}) = \hat{g}\left(\mathbf{x}^T\hat{\boldsymbol{\theta}}\right)$.

Under geometrically strong mixing condition, strong consistency and $\sqrt{n}$-rate asymptotic normality of the estimator $\hat{\boldsymbol{\theta}}$ of the SIP coefficient $\boldsymbol{\theta}_0$ in (4.1.1) are obtained. Proposition 4.2.2 is the key in understanding the efficiency of the proposed estimator. It shows that the derivatives of the risk function up to order 2 are uniformly almost surely approximated by their empirical versions.

Practical performance of the SIP estimators is examined via Monte Carlo examples. The estimator of the SIP coefficient performs very well for data of both moderate and high dimension d, of sample size n from small to large, see Tables 4.5 and 4.6, Figures 4.19, 4.20 and 4.21. By taking advantages of the spline smoothing and the iterative optimization routines, one reduces the computation burden immensely for massive data sets. Table 4.6 reports the computing time of one simulation example on an ordinary PC, which shows that for massive data sets, the SIP method is much faster than the MAVE method. For instance, the SIP estimation of a 200-dimensional $\boldsymbol{\theta}_0$ from a data of size 1000 takes on average mere 2.84 seconds, while the MAVE method needs to spend 2432.56 seconds on average to obtain a comparable estimates. Hence on account of criteria (i) and (ii), our method is indeed appealing. Applying the proposed SIP procedure to the rive flow data of Iceland, we have obtained superior forecasts, based on a 9-dimensional index selected by BIC, see Figure 4.25.

The rest of this chapter is organized as follows. Section 4.2 gives details of the model specification, proposed methods of estimation and main results. Section 4.3 describes the actual procedure to implement the estimation method. Section 4.4 reports the main findings in an extensive simulation study. The proposed SIP model and the estimation procedure are applied in Section 4.5 to the river flow data of Iceland. Most of the technical proofs are contained in Section 4.6.

4.2 The Method and Main Results

4.2.1 Identifiability and definition of the index coefficient

It is obvious that without constraints, the SIP coefficient vector $\boldsymbol{\theta}_0 = \left(\theta_{0,1}, ..., \theta_{0,d}\right)^T$ is identified only up to a constant factor. Typically, one requires that $\|\boldsymbol{\theta}_0\| = 1$ which entails that at least one of the coordinates $\theta_{0,1}, ..., \theta_{0,d}$ is nonzero. One could assume without loss of generality that $\theta_{0,d} > 0$, and the candidate $\boldsymbol{\theta}_0$ would then belong to the upper unit hemisphere $S_+^{d-1} = \left\{(\theta_1, ..., \theta_d) \,|\, \sum_{p=1}^d \theta_p^2 = 1, \theta_d > 0\right\}$.

For a fixed $\boldsymbol{\theta} = (\theta_1, ..., \theta_d)^T$, denote $X_{\boldsymbol{\theta}} = \mathbf{X}^T\boldsymbol{\theta}$, $X_{\boldsymbol{\theta},i} = \mathbf{X}_i^T\boldsymbol{\theta}$, $1 \leq i \leq n$. Let

$$m_{\boldsymbol{\theta}}\left(X_{\boldsymbol{\theta}}\right) = E\left(Y|X_{\boldsymbol{\theta}}\right) = E\left\{m\left(\mathbf{X}\right)|X_{\boldsymbol{\theta}}\right\}. \tag{4.2.1}$$

Define the risk function of $\boldsymbol{\theta}$ as

$$R\left(\boldsymbol{\theta}\right) = E\left[\left\{Y - m_{\boldsymbol{\theta}}\left(X_{\boldsymbol{\theta}}\right)\right\}^2\right] = E\left\{m\left(\mathbf{X}\right) - m_{\boldsymbol{\theta}}\left(X_{\boldsymbol{\theta}}\right)\right\}^2 + E\sigma^2\left(\mathbf{X}\right), \tag{4.2.2}$$

which is uniquely minimized at $\boldsymbol{\theta}_0 \in S_+^{d-1}$, i.e.

$$\boldsymbol{\theta}_0 = \arg \min_{\boldsymbol{\theta} \in S_+^{d-1}} R\left(\boldsymbol{\theta}\right).$$

REMARK 4.2.1. Note that S_{+}^{d-1} is not a compact set, so a cap shape subset of S_{+}^{d-1} is introduced

$$S_c^{d-1} = \left\{ (\theta_1, ..., \theta_d) \mid \sum_{p=1}^{d} \theta_p^2 = 1, \theta_d \geq \sqrt{1-c^2} \right\}, c \in (0,1)$$

Clearly, for an appropriate choice of c, $\boldsymbol{\theta}_0 \in S_c^{d-1}$, which is assumed in the rest of the chapter.

Denote $\boldsymbol{\theta}_{-d} = (\theta_1, ..., \theta_{d-1})^T$, since for fixed $\boldsymbol{\theta} \in S_{+}^{d-1}$, the risk function $R(\boldsymbol{\theta})$ depends only on the first $d-1$ values in $\boldsymbol{\theta}$, so $R(\boldsymbol{\theta})$ is a function of $\boldsymbol{\theta}_{-d}$

$$R^*(\boldsymbol{\theta}_{-d}) = R\left(\theta_1, \theta_2, ..., \theta_{d-1}, \sqrt{1 - \|\boldsymbol{\theta}_{-d}\|_2^2}\right),$$

with well-defined score and Hessian matrices

$$S^*(\boldsymbol{\theta}_{-d}) = \frac{\partial}{\partial \boldsymbol{\theta}_{-d}} R^*(\boldsymbol{\theta}_{-d}), H^*(\boldsymbol{\theta}_{-d}) = \frac{\partial^2}{\partial \boldsymbol{\theta}_{-d} \partial \boldsymbol{\theta}_{-d}^T} R^*(\boldsymbol{\theta}_{-d}). \tag{4.2.3}$$

ASSUMPTION (C1): *The Hessian matrix $H^*(\boldsymbol{\theta}_{0,-d})$ is positive definite and the risk function R^* is locally convex at $\boldsymbol{\theta}_{0,-d}$, i.e., for any $\varepsilon > 0$, there exists $\delta > 0$ such that $R^*(\boldsymbol{\theta}_{-d}) - R^*(\boldsymbol{\theta}_{0,-d}) < \delta$ implies $\|\boldsymbol{\theta}_{-d} - \boldsymbol{\theta}_{0,-d}\|_2 < \varepsilon$.*

4.2.2 Variable transformation

Throughout this chapter, denote by $B_a^d = \left\{ \mathbf{x} \in R^d \mid \|\mathbf{x}\| \leq a \right\}$ the d-dimensional ball with radius a and center $\mathbf{0}$ and

$$C^{(k)}\left(B_a^d\right) = \left\{ m \,\middle|\, \text{the } k\text{th order partial derivatives of } m \text{ are continuous on } B_a^d \right\}$$

the space of k-th order smooth functions.

ASSUMPTION (C2): *The density function of* $\mathbf{X}$, $f(\mathbf{x}) \in C^{(4)}\left(B_a^d\right)$, *and there are constants* $0 < c_f \leq C_f$ *such that*

$$\begin{cases} c_f / \mathrm{Vol}_d\left(B_a^d\right) \leq f(\mathbf{x}) \leq C_f / \mathrm{Vol}_d\left(B_a^d\right), & \mathbf{x} \in B_a^d \\ f(\mathbf{x}) \equiv 0, & \mathbf{x} \notin B_a^d \end{cases}.$$

For a fixed $\boldsymbol{\theta}$, define the transformed variables of the SIP variable $X_{\boldsymbol{\theta}}$

$$U_{\boldsymbol{\theta}} = F_d(X_{\boldsymbol{\theta}}), U_{\boldsymbol{\theta},i} = F_d(X_{\boldsymbol{\theta},i}), 1 \leq i \leq n, \tag{4.2.4}$$

in which F_d is the a rescaled centered Beta$\{(d+1)/2, (d+1)/2\}$ cumulative distribution function, i.e.

$$F_d(\nu) = \int_{-1}^{\nu/a} \frac{\Gamma(d+1)}{\Gamma\{(d+1)/2\}^2 2^d}\left(1-t^2\right)^{(d-1)/2} dt, \nu \in [-a, a]. \qquad (4.2.5)$$

REMARK 4.2.2. For any fixed $\boldsymbol{\theta}$, the transformed variable $U_{\boldsymbol{\theta}}$ in (4.2.4) has a quasi-uniform $[0,1]$ distribution. Let $f_{\boldsymbol{\theta}}(u)$ be the probability density function of $U_{\boldsymbol{\theta}}$, then for any $u \in [0,1]$

$$f_{\boldsymbol{\theta}}(u) = \left\{F_d'(v)\right\} f_{X_{\boldsymbol{\theta}}}(v), \; v = F_d^{-1}(u),$$

in which $f_{X_{\boldsymbol{\theta}}}(v) = \lim_{\Delta\nu \to 0} P(\nu \leq X_{\boldsymbol{\theta}} \leq \nu + \Delta\nu)$. Noting that $x_{\boldsymbol{\theta}}$ is exactly the projection of x on $\boldsymbol{\theta}$, let $\mathcal{D}_\nu = \{\mathrm{x} | \nu \leq x_{\boldsymbol{\theta}} \leq \nu + \Delta\nu\} \cap B_{\mathrm{a}}^d$, then one has

$$P(\nu \leq X_{\boldsymbol{\theta}} \leq \nu + \Delta\nu) = P(\mathbf{X} \in \mathcal{D}_\nu) = \int_{\mathcal{D}_\nu} f(\mathbf{x})\, d\mathbf{x}.$$

According to Assumption (C2)

$$\frac{c_f \mathrm{Vol}_d(\mathcal{D}_\nu)}{\mathrm{Vol}_d\left(B_{\mathrm{a}}^d\right)} \leq P(\nu \leq X_{\boldsymbol{\theta}} \leq \nu + \Delta\nu) \leq \frac{C_f \mathrm{Vol}_d(\mathcal{D}_\nu)}{\mathrm{Vol}_d\left(B_{\mathrm{a}}^d\right)}.$$

On the other hand

$$\mathrm{Vol}_d(\mathcal{D}_\nu) = \mathrm{Vol}_{d-1}(\mathcal{J}_\nu)\Delta\nu + o(\Delta\nu),$$

where $\mathcal{J}_\nu = \{\mathrm{x} | x_{\boldsymbol{\theta}} = v\} \cap B_{\mathrm{a}}^d$. Note that the volume of B_{a}^d is $\pi^{d/2} a^d / \Gamma(d/2+1)$ and

$$\mathrm{Vol}_{d-1}(\mathcal{J}_\nu) = \pi^{(d-1)/2}\left(a^2 - \nu^2\right)^{(d-1)/2} \Big/ \Gamma\{(d+1)/2\},$$

thus

$$\frac{\mathrm{Vol}_{d-1}(\mathcal{J}_\nu)}{\mathrm{Vol}_d\left(B_{\mathrm{a}}^d\right)} = \frac{1}{a\sqrt{\pi}} \frac{\Gamma(d+1)}{\left\{\Gamma\left(\frac{d+1}{2}\right)\right\}^2 2^d} \left\{1 - \left(\frac{\nu}{a}\right)^2\right\}^{(d-1)/2}.$$

Therefore $0 < c_f \leq f_{\boldsymbol{\theta}}(u) \leq C_f < \infty$, for any fixed $\boldsymbol{\theta}$ and $u \in [0,1]$.

In terms of the transformed SIP variable $U_{\boldsymbol{\theta}}$ in (4.2.4), one can rewrite the regression function $m_{\boldsymbol{\theta}}$ in (4.2.1) for fixed $\boldsymbol{\theta}$

$$\gamma_{\boldsymbol{\theta}}(U_{\boldsymbol{\theta}}) = E\{m(\mathbf{X}) | U_{\boldsymbol{\theta}}\} = E\{m(\mathbf{X}) | X_{\boldsymbol{\theta}}\} = m_{\boldsymbol{\theta}}(X_{\boldsymbol{\theta}}), \qquad (4.2.6)$$

then the risk function $R(\boldsymbol{\theta})$ in (4.2.2) can be expressed as

$$R(\boldsymbol{\theta}) = E\left[\{Y - \gamma_{\boldsymbol{\theta}}(U_{\boldsymbol{\theta}})\}^2\right] = E\{m(\mathbf{X}) - \gamma_{\boldsymbol{\theta}}(U_{\boldsymbol{\theta}})\}^2 + E\sigma^2(\mathbf{X}). \qquad (4.2.7)$$

4.2.3 Estimation Method

Estimation of both $\boldsymbol{\theta}_0$ and g requires a degree of statistical smoothing, and all estimation here is carried out via cubic spline. In the following, define the estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}_0$ and the estimator $\hat{g}$ of g.

According to the definition of B-spline in Section 1.5 of Chapter 1, for fixed $\boldsymbol{\theta}$, the cubic spline estimator $\hat{\gamma}_{\boldsymbol{\theta}}$ of $\gamma_{\boldsymbol{\theta}}$ and the related estimator $\hat{m}_{\boldsymbol{\theta}}$ of $m_{\boldsymbol{\theta}}$ are defined as

$$\hat{\gamma}_{\boldsymbol{\theta}}(\cdot) = \arg \min_{\gamma(\cdot)\in G^{(2)}[0,1]} \sum_{i=1}^{n} \{Y_i - \gamma(U_{\boldsymbol{\theta},i})\}^2, \ \hat{m}_{\boldsymbol{\theta}}(\nu) = \hat{\gamma}_{\boldsymbol{\theta}}\{F_d(\nu)\}. \tag{4.2.8}$$

Define the empirical risk function of $\boldsymbol{\theta}$

$$\hat{R}(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^{n}\{Y_i - \hat{\gamma}_{\boldsymbol{\theta}}(U_{\boldsymbol{\theta},i})\}^2 = n^{-1}\sum_{i=1}^{n}\{Y_i - \hat{m}_{\boldsymbol{\theta}}(X_{\boldsymbol{\theta},i})\}^2, \tag{4.2.9}$$

then the spline estimator of the SIP coefficient $\boldsymbol{\theta}_0$ is defined as

$$\hat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta}\in S_c^{d-1}} \hat{R}(\boldsymbol{\theta}),$$

and the cubic spline estimator of g is $\hat{m}_{\boldsymbol{\theta}}$ with $\boldsymbol{\theta}$ replaced by $\hat{\boldsymbol{\theta}}$, i.e.

$$\hat{g}(\nu) = \left\{\arg \min_{\gamma(\cdot)\in G^{(2)}[0,1]} \sum_{i=1}^{n} \left\{Y_i - \gamma\left(U_{\hat{\boldsymbol{\theta}},i}\right)\right\}^2\right\}\{F_d(\nu)\}. \tag{4.2.10}$$

4.2.4 Asymptotic results

The following are some other assumptions to achieve the main theorems.

Assumption (C3): *The regression function $m \in C^{(4)}\left(B_a^d\right)$ for some $a > 0$.*

Assumption (C4): *The noise ε satisfies $E(\varepsilon\,|\mathbf{X}) = 0$, $E\left(\varepsilon^2\,|\mathbf{X}\right) = 1$ and there exists a positive constant M such that $\sup_{\mathbf{x}\in B^d} E\left(|\varepsilon|^3\,|\mathbf{X} = \mathbf{x}\right) < M$. The standard deviation function $\sigma(\mathbf{x})$ is continuous on B_a^d,*

$$0 < c_\sigma \leq \inf_{\mathbf{x}\in B_a^d} \sigma(\mathbf{x}) \leq \sup_{\mathbf{x}\in B_a^d} \sigma(\mathbf{x}) \leq C_\sigma < \infty.$$

Assumption (C5): *There exist positive constants K_0 and λ_0 such that $\alpha(n) \leq K_0 e^{-\lambda_0 n}$ holds for all n, with the α-mixing coefficient for $\left\{\mathbf{Z}_i = \left(\mathbf{X}_i^T, \varepsilon_i\right)\right\}_{i=1}^{n}$ defined as*

$$\alpha(k) = \sup_{B\in\sigma\{\mathbf{Z}_s, s\leq t\}, C\in\sigma\{\mathbf{Z}_s, s\geq t+k\}} |P(B\cap C) - P(B)P(C)|, \ k \geq 1.$$

ASSUMPTION (C6): *The number of interior knots N satisfies:* $n^{1/6} \ll N \ll n^{1/5}(\log n)^{-2/5}$.

REMARK 4.2.3. Assumptions (C3) and (C4) are typical in the nonparametric smoothing literature, see for instance, Härdle (1990), Fan and Gijbels (1996), Xia, Tong Li and Zhu (2002). By the result of Pham (1986), a geometrically ergodic time series is a strongly mixing sequence. Therefore, Assumption (C5) is suitable for (1.1.1) as a time series model under aforementioned assumptions.

THEOREM 4.2.1. *Under Assumptions (C1)-(C6), one has*

$$\hat{\boldsymbol{\theta}}_{-d} \longrightarrow \boldsymbol{\theta}_{0,-d}, a.s.. \tag{4.2.11}$$

PROOF. Denote by $(\Omega, \mathcal{F}, \mathcal{P})$ the probability space on which all $\left\{\left(\mathbf{X}_i^T, Y_i\right)\right\}_{i=1}^{\infty}$ are defined. By Proposition 4.2.2, given at the end of this section

$$\sup_{\left\|\boldsymbol{\theta}_{-d}\right\|_2 \leq \sqrt{1-c^2}} \left|\hat{R}^*\left(\boldsymbol{\theta}_{-d}\right) - R^*\left(\boldsymbol{\theta}_{-d}\right)\right| \longrightarrow 0, a.s.. \tag{4.2.12}$$

So for any $\delta > 0$ and $\omega \in \Omega$, there exists an integer $n_0(\omega)$, such that when $n > n_0(\omega)$, $\hat{R}^*\left(\boldsymbol{\theta}_{0,-d}, \omega\right) - R^*\left(\boldsymbol{\theta}_{0,-d}\right) < \delta/2$. Note that $\hat{\boldsymbol{\theta}}_{-d} = \hat{\boldsymbol{\theta}}_{-d}(\omega)$ is the minimizer of $\hat{R}^*\left(\boldsymbol{\theta}_{-d}, \omega\right)$, so $\hat{R}^*\left(\hat{\boldsymbol{\theta}}_{-d}(\omega), \omega\right) - R^*\left(\boldsymbol{\theta}_{0,-d}\right) < \delta/2$. Using (4.2.12), there exists $n_1(\omega)$, such that when $n > n_1(\omega)$, $R^*\left(\hat{\boldsymbol{\theta}}_{-d}(\omega), \omega\right) - \hat{R}^*\left(\hat{\boldsymbol{\theta}}_{-d}(\omega), \omega\right) < \delta/2$. Thus, when $n > \max\left(n_0(\omega), n_1(\omega)\right)$,

$$R^*\left(\hat{\boldsymbol{\theta}}_{-d}(\omega), \omega\right) - R^*\left(\boldsymbol{\theta}_{0,-d}\right) < \delta/2 + \hat{R}^*\left(\hat{\boldsymbol{\theta}}_{-d}(\omega), \omega\right) - R^*\left(\boldsymbol{\theta}_{0,-d}\right) < \delta/2 + \delta/2 = \delta.$$

According to Assumption (C1), R^* is locally convex at $\boldsymbol{\theta}_{0,-d}$, so for any $\varepsilon > 0$ and any ω, if $R^*\left(\hat{\boldsymbol{\theta}}_{-d}(\omega), \omega\right) - R^*\left(\boldsymbol{\theta}_{0,-d}\right) < \delta$, then $\left\|\hat{\boldsymbol{\theta}}_{-d}(\omega) - \boldsymbol{\theta}_{0,-d}\right\| < \varepsilon$ for n large enough , which implies the strong consistency. □

THEOREM 4.2.2. *Under Assumptions (C1)-(C6), one has*

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_{-d} - \boldsymbol{\theta}_{0,-d}\right) \xrightarrow{d} N\left\{0, \Sigma\left(\boldsymbol{\theta}_0\right)\right\},$$

where $\Sigma(\boldsymbol{\theta}_0) = \left\{H^*\left(\boldsymbol{\theta}_{0,-d}\right)\right\}^{-1}\Psi(\boldsymbol{\theta}_0)\left\{H^*\left(\boldsymbol{\theta}_{0,-d}\right)\right\}^{-1}$, $H^*\left(\boldsymbol{\theta}_{0,-d}\right) = \{l_{pq}\}_{p,q=1}^{d-1}$ *and* $\Psi(\boldsymbol{\theta}_0) = \{\psi_{pq}\}_{p,q=1}^{d-1}$ *with*

$$\begin{aligned} l_{p,q} = -2E\left[\left\{\dot{\gamma}_p\dot{\gamma}_q + \gamma_{\boldsymbol{\theta}_0}\ddot{\gamma}_{p,q}\right\}\left(U_{\boldsymbol{\theta}_0}\right)\right] + 2\theta_{0,q}\theta_{0,d}^{-1}E\left[\left\{\dot{\gamma}_p\dot{\gamma}_d\left(U_{\boldsymbol{\theta}_0}\right) + \gamma_{\boldsymbol{\theta}_0}\ddot{\gamma}_{p,d}\right\}\left(U_{\boldsymbol{\theta}_0}\right)\right] \\ +2\theta_{0,d}^{-3}E\left[\left(\gamma_{\boldsymbol{\theta}_0}\dot{\gamma}_d\right)\left(U_{\boldsymbol{\theta}_0}\right)\right]\left\{\left(\theta_{0,d}^2 + \theta_{0,p}^2\right)I_{\{p=q\}} + \theta_{0,p}\theta_{0,q}I_{\{p\neq q\}}\right\} \\ +2\theta_{0,p}\theta_{0,d}^{-1}E\left[\left\{\dot{\gamma}_p\dot{\gamma}_q + \gamma_{\boldsymbol{\theta}_0}\ddot{\gamma}_{p,q}\right\}\left(U_{\boldsymbol{\theta}_0}\right)\right] - 2\theta_{0,p}\theta_{0,q}\theta_{0,d}^{-2}E\left[\left\{\dot{\gamma}_d^2 + \gamma_{\boldsymbol{\theta}_0}\ddot{\gamma}_{d,d}\right\}\left(U_{\boldsymbol{\theta}_0}\right)\right], \\ \psi_{pq} = 4E\left[\left\{\left(\dot{\gamma}_p - \theta_{0,p}\theta_{0,d}^{-1}\dot{\gamma}_d\right)\left(\dot{\gamma}_q - \theta_{0,q}\theta_{0,d}^{-1}\dot{\gamma}_d\right)\right\}\left(U_{\boldsymbol{\theta}_0}\right)\left\{\gamma_{\boldsymbol{\theta}_0}\left(U_{\boldsymbol{\theta}_0}\right) - Y\right\}^2\right], \end{aligned}$$

in which $\dot{\gamma}_p$ *and* $\ddot{\gamma}_{p,q}$ *are the values of* $\frac{\partial}{\partial\theta_p}\gamma_{\boldsymbol{\theta}}$, $\frac{\partial^2}{\partial\theta_p\partial\theta_q}\gamma_{\boldsymbol{\theta}}$ *taking at* $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, *for any* $p, q = 1, 2, ..., d-1$ *and* $\gamma_{\boldsymbol{\theta}}$ *is given in (4.2.6).*

REMARK 4.2.4. Consider the Generalized Linear Model (GLM): $Y = g\left(\mathbf{X}^T\boldsymbol{\theta}_0\right) + \sigma(\mathbf{X})\varepsilon$, where g is a known link function. Let $\tilde{\boldsymbol{\theta}}$ be the nonlinear least squared estimator of $\boldsymbol{\theta}_0$ in GLM. Theorem 4.2.2 shows that under the assumptions (C1)-(C6), the asymptotic distribution of the $\hat{\boldsymbol{\theta}}_{-d}$ is the same as that of $\tilde{\boldsymbol{\theta}}$. This implies that the proposed SIP estimator $\hat{\boldsymbol{\theta}}_{-d}$ is as efficient as if the true link function g is known.

The next two propositions play an important role in the proof of the main results. Proposition 4.2.1 establishes the uniform convergence rate of the derivatives of $\hat{\gamma}_{\boldsymbol{\theta}}$ up to order 2 to those of $\gamma_{\boldsymbol{\theta}}$ in $\boldsymbol{\theta}$. Proposition 4.2.2 shows that the derivatives of the risk function up to order 2 are uniformly almost surely approximated by their empirical versions.

PROPOSITION 4.2.1. *Under Assumptions (C2)-(C6), with probability* 1

$$\sup_{\boldsymbol{\theta}\in S_c^{d-1}}\sup_{u\in[0,1]}\left|\hat{\gamma}_{\boldsymbol{\theta}}(u) - \gamma_{\boldsymbol{\theta}}(u)\right| = O\left\{(nh)^{-1/2}\log n + h^4\right\}, \tag{4.2.13}$$

$$\sup_{1\leq p\leq d}\sup_{\boldsymbol{\theta}\in S_c^{d-1}}\max_{1\leq i\leq n}\left|\frac{\partial}{\partial\theta_p}\left\{\hat{\gamma}_{\boldsymbol{\theta}}\left(U_{\boldsymbol{\theta},i}\right) - \gamma_{\boldsymbol{\theta}}\left(U_{\boldsymbol{\theta},i}\right)\right\}\right| = O\left(\frac{\log n}{\sqrt{nh^3}} + h^3\right), \tag{4.2.14}$$

$$\sup_{1\leq p,q\leq d}\sup_{\boldsymbol{\theta}\in S_c^{d-1}}\max_{1\leq i\leq n}\left|\frac{\partial^2}{\partial\theta_p\partial\theta_q}\left\{\hat{\gamma}_{\boldsymbol{\theta}}\left(U_{\boldsymbol{\theta},i}\right) - \gamma_{\boldsymbol{\theta}}\left(U_{\boldsymbol{\theta},i}\right)\right\}\right| = O\left(\frac{\log n}{\sqrt{nh^5}} + h^2\right). \tag{4.2.15}$$

PROPOSITION 4.2.2. *Under Assumptions (C2)-(C6), one has for* $k = 0, 1, 2$

$$\sup_{\left\|\boldsymbol{\theta}_{-d}\right\|\leq\sqrt{1-c^2}}\left|\frac{\partial^k}{\partial^k\boldsymbol{\theta}_{-d}}\left\{\hat{R}^*\left(\boldsymbol{\theta}_{-d}\right) - R^*\left(\boldsymbol{\theta}_{-d}\right)\right\}\right| = o(1), a.s..$$

Proofs of Theorem 4.2.2, Propositions 4.2.1 and 4.2.2 are given in Section 4.6.

4.3 Implementation

This section describes the actual procedure to implement the estimation of $\boldsymbol{\theta}_0$ and g. We first introduce some new notation. For fixed $\boldsymbol{\theta}$, denote the B-spline matrix as $\mathbf{B}_{\boldsymbol{\theta}} = \left\{B_{j,4}\left(U_{\boldsymbol{\theta},i}\right)\right\}_{i=1,j=-3}^{n,\,N}$ and

$$\mathbf{P}_{\boldsymbol{\theta}} = \mathbf{B}_{\boldsymbol{\theta}}\left(\mathbf{B}_{\boldsymbol{\theta}}^T\mathbf{B}_{\boldsymbol{\theta}}\right)^{-1}\mathbf{B}_{\boldsymbol{\theta}}^T \tag{4.3.1}$$

as the projection matrix onto the cubic spline space $G_{n,\boldsymbol{\theta}}^{(2)}$. For any $p = 1, ..., d$, denote

$$\dot{\mathbf{B}}_p = \frac{\partial}{\partial\theta_p}\mathbf{B}_{\boldsymbol{\theta}},\ \dot{\mathbf{P}}_p = \frac{\partial}{\partial\theta_p}\mathbf{P}_{\boldsymbol{\theta}}.$$

as the first order partial derivatives of $\mathbf{B}_{\boldsymbol{\theta}}$ and $\mathbf{P}_{\boldsymbol{\theta}}$ with respect to $\boldsymbol{\theta}$.

It is easy to see that the distribution function F_d in (4.2.5) satisfies

$$\dot{F}_d(x) = \frac{d}{dx}F_d = \frac{\Gamma(d+1)}{a\Gamma\{(d+1)/2\}^2 2^d}\left(1-\frac{x^2}{a^2}\right)^{\frac{d-1}{2}} I\left(|x| \le a\right). \tag{4.3.2}$$

Let $\hat{S}^*(\boldsymbol{\theta}_{-d})$ be the score vector of $\hat{R}^*\left(\boldsymbol{\theta}_{-d}\right)$, i.e.

$$\hat{S}^*(\boldsymbol{\theta}_{-d}) = \frac{\partial}{\partial\boldsymbol{\theta}_{-d}}\hat{R}^*\left(\boldsymbol{\theta}_{-d}\right). \tag{4.3.3}$$

The next lemma provides the exact forms of $\hat{S}^*(\boldsymbol{\theta}_{-d})$.

Lemma 4.3.1. *For the score vector of $\hat{R}^*\left(\boldsymbol{\theta}_{-d}\right)$ defined in (4.3.3), one has*

$$\hat{S}^*\left(\boldsymbol{\theta}_{-d}\right) = -n^{-1}\left\{\mathbf{Y}^T\dot{\mathbf{P}}_p\mathbf{Y} - \theta_p\theta_d^{-1}\mathbf{Y}^T\dot{\mathbf{P}}_d\mathbf{Y}\right\}_{p=1}^{d-1}, \tag{4.3.4}$$

where for any $p = 1, 2, ..., d$

$$\mathbf{Y}^T\dot{\mathbf{P}}_p\mathbf{Y} = 2\mathbf{Y}^T\left(\mathbf{I} - \mathbf{P}_{\boldsymbol{\theta}}\right)\dot{\mathbf{B}}_p\left(\mathbf{B}_{\boldsymbol{\theta}}^T\mathbf{B}_{\boldsymbol{\theta}}\right)^{-1}\mathbf{B}_{\boldsymbol{\theta}}^T\mathbf{Y}, \tag{4.3.5}$$

where $\dot{\mathbf{B}}_p = \left\{\left\{B_{j,3}\left(U_{\boldsymbol{\theta},i}\right) - B_{j+1,3}\left(U_{\boldsymbol{\theta},i}\right)\right\}\dot{F}_d\left(\mathbf{X}_{\boldsymbol{\theta},i}\right)h^{-1}X_{i,p}\right\}_{i=1,j=-3}^{n,\,N}$ with $\dot{F}_d$ in (4.3.2).

Proof. For any $p = 1, 2, ..., d$, the derivatives of B-splines in de Boor (2001) implies

$$\begin{aligned}
\dot{\mathbf{B}}_p &= \left\{\frac{\partial}{\partial\theta_p}B_{j,4}\left(U_{\boldsymbol{\theta},i}\right)\right\}_{i=1,j=-3}^{n,\,N} = \left\{\frac{d}{du}B_{j,4}\left(U_{\boldsymbol{\theta},i}\right)\frac{d}{d\theta_p}U_{\boldsymbol{\theta},i}\right\}_{i=1,j=-3}^{n,\,N} \\
&= 3\left\{\left\{\frac{B_{j,3}\left(U_{\boldsymbol{\theta},i}\right)}{t_{j+3}-t_j} - \frac{B_{j+1,3}\left(U_{\boldsymbol{\theta},i}\right)}{t_{j+4}-t_{j+1}}\right\}\dot{F}_d\left(\mathbf{X}_{\boldsymbol{\theta},i}\right)X_{i,p}\right\}_{i=1,j=-3}^{n,\,N} \\
&= \left\{\left\{B_{j,3}\left(U_{\boldsymbol{\theta},i}\right) - B_{j+1,3}\left(U_{\boldsymbol{\theta},i}\right)\right\}\dot{F}_d\left(\mathbf{X}_{\boldsymbol{\theta},i}\right)h^{-1}X_{i,p}\right\}_{i=1,j=-3}^{n,\,N}.
\end{aligned}$$

Next, note that

$$\begin{aligned}\dot{\mathbf{P}}_p &= \dot{\mathbf{B}}_p\left(\mathbf{B}_\theta^T\mathbf{B}_\theta\right)^{-1}\mathbf{B}_\theta^T + \mathbf{B}_\theta\left[\frac{\partial}{\partial\theta_p}\left\{\left(\mathbf{B}_\theta^T\mathbf{B}_\theta\right)^{-1}\mathbf{B}_\theta^T\right\}\right] \\ &= \dot{\mathbf{B}}_p\left(\mathbf{B}_\theta^T\mathbf{B}_\theta\right)^{-1}\mathbf{B}_\theta^T + \mathbf{B}_\theta\left\{\frac{\partial}{\partial\theta_p}\left(\mathbf{B}_\theta^T\mathbf{B}_\theta\right)^{-1}\right\}\mathbf{B}_\theta^T + \mathbf{B}_\theta\left(\mathbf{B}_\theta^T\mathbf{B}_\theta\right)^{-1}\dot{\mathbf{B}}_p^T.\end{aligned}$$

Since

$$0 \equiv \frac{\partial\left\{\left(\mathbf{B}_\theta^T\mathbf{B}_\theta\right)^{-1}\mathbf{B}_\theta^T\mathbf{B}_\theta\right\}}{\partial\theta_p} = \frac{\partial\left(\mathbf{B}_\theta^T\mathbf{B}_\theta\right)^{-1}}{\partial\theta_p}\mathbf{B}_\theta^T\mathbf{B}_\theta + \left(\mathbf{B}_\theta^T\mathbf{B}_\theta\right)^{-1}\frac{\partial\left(\mathbf{B}_\theta^T\mathbf{B}_\theta\right)}{\partial\theta_p},$$

and $\frac{\partial}{\partial\theta_p}\left(\mathbf{B}_\theta^T\mathbf{B}_\theta\right) = \dot{\mathbf{B}}_p^T\mathbf{B}_\theta + \mathbf{B}_\theta^T\dot{\mathbf{B}}_p$, thus

$$\frac{\partial}{\partial\theta_p}\left(\mathbf{B}_\theta^T\mathbf{B}_\theta\right)^{-1} = -\left(\mathbf{B}_\theta^T\mathbf{B}_\theta\right)^{-1}\left(\dot{\mathbf{B}}_p^T\mathbf{B}_\theta + \mathbf{B}_\theta^T\dot{\mathbf{B}}_p\right)\left(\mathbf{B}_\theta^T\mathbf{B}_\theta\right)^{-1}.$$

Hence

$$\dot{\mathbf{P}}_p = \left(\mathbf{I} - \mathbf{P}_\theta\right)\dot{\mathbf{B}}_p\left(\mathbf{B}_\theta^T\mathbf{B}_\theta\right)^{-1}\mathbf{B}_\theta^T + \mathbf{B}_\theta\left(\mathbf{B}_\theta^T\mathbf{B}_\theta\right)^{-1}\dot{\mathbf{B}}_p^T\left(\mathbf{I} - \mathbf{P}_\theta\right).$$

Thus, (4.3.5) follows immediately. □

In practice, the estimation is implemented via the following procedure.

Step 1. *Standardize the predictor vectors* $\{\mathbf{X}_i\}_{i=1}^n$ *and for each fixed* $\boldsymbol{\theta} \in S_c^{d-1}$ *obtain the CDF transformed variables* $\left\{U_{\theta,i}\right\}_{i=1}^n$ *of the SIP variable* $\left\{X_{\theta,i}\right\}_{i=1}^n$ *through formula (4.2.5), where the radius a is taken to be the 95% percentile of* $\{\|\mathbf{X}_i\|\}_{i=1}^n$.

Step 2. *Compute quadratic and cubic B-spline basis at each value* $U_{\theta,i}$, *where the number of interior knots* N *is*

$$N = \min\left\{c_1\left[n^{1/5.5}\right], c_2\right\}, \tag{4.3.6}$$

Step 3. *Find the estimator* $\hat{\boldsymbol{\theta}}$ *of* $\boldsymbol{\theta}_0$ *by minimizing* $\hat{R}^*$ *through the port optimization routine with* $(0, 0, ..., 1)^T$ *as the initial value and the empirical score vector* $\hat{S}^*$ *in (4.3.4). If* $d < n$, *one can take the simple LSE (without the intercept) for data* $\{Y_i, \mathbf{X}_i\}_{i=1}^n$ *with its last coordinate set positive.*

Step 4. *Obtain the spline estimator* $\hat{g}$ *of* g *by plugging* $\hat{\boldsymbol{\theta}}$ *obtained in Step 3 into (4.2.10).*

REMARK 4.3.1. In (4.3.6), c_1 and c_2 are positive integers and $[\nu]$ denotes the integer part of ν. The choice of the tuning parameter c_1 makes little difference for a large sample and according to the asymptotic theory there is no optimal way to set these constants. We recommend using $c_1 = 1$ to save computing for massive data sets. The first term ensures Assumption (C6). The addition constrain c_2 can be taken from 5 to 10 for smooth monotonic or smooth unimodel regression and $c_2 > 10$ if has many local minima and maxima, which is very unlikely in application.

4.4 Simulations

In this section, two simulations are carried out to illustrate the finite-sample behavior of the SIP estimation method. The number of interior knots N is computed according to (4.3.6) with $c_1 = 1, c_2 = 5$. All of the codes have been written in R.

4.4.1 Example 1

Consider the model in Xia, Li, Tong and Zhang (2004)

$$Y = m(\mathbf{X}) + \sigma_0 \varepsilon, \ \sigma_0 = 0.3, 0.5, \ \varepsilon \overset{i.i.d}{\sim} N(0,1)$$

where $\mathbf{X} = (X_1, X_2)^T \sim N(0, I_2)$, truncated by $[-2.5, 2.5]^2$ and

$$m(\mathrm{x}) = x_1 + x_2 + 4\exp\left\{-(x_1 + x_2)^2\right\} + \delta\left(x_1^2 + x_2^2\right)^{1/2}. \tag{4.4.1}$$

If $\delta = 0$, then the underlying true function m is exactly a single-index function, i.e., $m(\mathbf{X}) = \sqrt{2}\mathbf{X}^T\boldsymbol{\theta}_0 + 4\exp\left\{-2\left(\mathbf{X}^T\boldsymbol{\theta}_0\right)^2\right\}$, where $\boldsymbol{\theta}_0^T = (1,1)/\sqrt{2}$. While $\delta \neq 0$, then m is not a genuine single-index function. An impression of the bivariate function m for $\delta = 0$ and $\delta = 1$ can be gained in Figure 4.18.

For $\delta = 0, 1$, one hundred random realizations of each sample size $n = 50, 100, 300$ are drawn respectively. To demonstrate how close the SIP estimator is to the true index parameter $\boldsymbol{\theta}_0$, Table 4.5 lists the sample mean (MEAN), bias (BIAS), standard deviation (SD), the mean squared error (MSE) of the estimates of $\boldsymbol{\theta}_0$ and the average MSE of both directions. From this table, one sees that the SIP estimators are very accurate for both cases $\delta = 0$ and $\delta = 1$, which shows that the proposed method is robust against the deviation from single-index model. As we expected, when the sample size increases, the SIP coefficient is more accurately estimated. Moreover, for $n = 100, 300$, the total average is inversely proportional to n.

4.4.2 Example 2

Consider the heteroscedastic regression model (1.1.1) with

$$m(\mathrm{X}) = \sin\left(\frac{\pi}{4}\mathrm{X}^T\theta_0\right), \ \sigma(\mathrm{X}) = \sigma_0\frac{\left\{5 - \exp\left(\|\mathrm{X}\|/\sqrt{d}\right)\right\}}{5 + \exp\left(\|\mathrm{X}\|/\sqrt{d}\right)}, \tag{4.4.2}$$

in which $\mathbf{X}_i = \{X_{i,1}, ..., X_{i,d}\}^T$ and ε_i, $i = 1, ..., n$, are $\overset{i.i.d}{\sim} N(0,1)$, $\sigma_0 = 0.2$. In this simulation, the true parameter $\theta_0^T = (1,1,0,...,0,1)/\sqrt{3}$ for different sample size n and dimension d. The superior performance of SIP estimators is borne out in comparison with MAVE of Xia, Tong, Li and Zhu (2002). We also investigate the behavior of SIP estimators in the previously unexplored cases that n is smaller than or equal to d, for instance, $n = 100, d = 100, 200$ and $n = 200, d = 200, 400$. The average MSEs of the d dimensions are listed in Table 4.6, from which one sees that the performance of the SIP estimators are quite reasonable and in most of the scenarios $n \leq d$, the SIP estimators still work astonishingly well where the MAVEs become unreliable. For $n = 100$, $d = 10, 50, 100, 200$, the estimates of the link prediction function from model (4.4.2) are plotted in Figures 4.20 and 4.21, which are rather satisfactory even when dimension exceeds the sample size.

Theorem 4.2.1 indicates that $\hat{\theta}_{-d}$ is strongly consistent of $\theta_{0,-d}$. To see the convergence, we run 100 replications and in each replication, the value of $\|\hat{\theta} - \theta_0\|/\sqrt{d}$ is computed. Figures 4.22 and 4.23 plot the kernel density estimations of the 100 $\|\hat{\theta} - \theta_0\|$ in Example 2, in which dimension $d = 10, 50, 100, 200$. There are four types of line characteristics: the dotted-dashed line ($n = 100$), dotted line ($n = 200$), dashed line (500) and solid line ($n = 1000$). As sample sizes increasing, the squared errors are becoming closer to 0, with narrower spread out, confirmative to the conclusions of Theorem 4.2.1.

Lastly, Table 4.6 reports the average computing time of Example 2 to generate one sample of size n and perform the SIP or MAVE procedure done on the same ordinary Pentium IV PC. From Table 4.6, one sees that the proposed SIP estimator is much faster than the MAVE. The computing time for MAVE is extremely sensitive to sample size as we expected. For very large d, MAVE becomes unstable to the point of the breaking down in four cases.

4.5 Application

In this section the proposed SIP model is demonstrated through the river flow data of Jökulsá Eystri River of Iceland, from January 1, 1972 to December 31, 1974. There are 1096 observations, see Tong (1990). The response variables are the daily river flow (Y_t), measured in meter cubed per second of Jökulsá Eystri River. The exogenous variables are temperature (X_t) in degrees Celsius and daily precipitation (Z_t) in millimeters collected at the meteorological station at Hveravellir.

This data set was analyzed earlier through threshold autoregressive (TAR) models by Tong, Thanoon and Gudmundsson (1985), Tong (1990), and nonlinear additive autoregressive (NAARX) models by Chen and Tsay (1993). Figure 4.24 shows the plots of the three time series, from which some nonlinear and non-stationary features of the river flow series are evident. To make these series stationary, the trends are removed by a simple quadratic spline regression and these trends (dashed lines) are shown in Figure 4.24. By an abuse of notation, we shall continue to use X_t, Y_t, Z_t to denote the detrended series.

In the analysis, we pre-select all the lagged values in the last 7 days (1 week), i.e., the predictor pool is $\{Y_{t-1}, ..., Y_{t-7}, X_t, X_{t-1}, ..., X_{t-7}, Z_t, Z_{t-1}, ..., Z_{t-7}, \}$. Using BIC similar to Huang and Yang (2004) for the proposed spline SIP model with 3 interior knots, the following 9 explanatory variables are selected from the above set $\{Y_{t-1}, ..., Y_{t-4}, X_t, X_{t-1}, X_{t-2}, Z_t, Z_{t-1}\}$. Based on this selection, we fit the SIP model again and obtain the estimate of the SIP coefficient

$$\hat{\theta} = \{-0.877, 0.382, -0.208, 0.125, -0.046, -0.034, 0.004, -0.126, 0.079\}^T.$$

The first two plots of Figure 4.25 display the fitted river flow series and the residuals against time.

Next we examine the forecasting performance of the SIP method. We start with estimating the SIP estimator using only observations of the first two years, then we perform the out-of-sample rolling forecast of the entire third year. The observed values of the exogenous variables are used in the forecast. The last plot of Figure 4.25 shows the SIP out-of-sample forecasts. For the purpose of comparison, the MAVE method is also used, in which the same predictor vector is selected by using BIC. The mean squared prediction error is 60.52 for the SIP model, 61.25 for MAVE, 65.62 for NAARX, 66.67 for TAR and 81.99 for the linear regression model, see Chen and Tsay (1993). Among the above five models, the SIP model produces the best forecasts.

4.6 Proof of The Theorems

4.6.1 Preliminaries

In this section, some properties of the B-spline are introduced.

LEMMA 4.6.1. *For each $0 < r \leq \infty$, there exist constants $c > 0$ such that for each spline combination $\sum_{j=-k+1}^{N} \alpha_{j,k} B_{j,k}$ up to order $k = 4$, one has*

$$\begin{cases} ch^{1/r} \|\alpha\|_r \leq \left\|\sum_{k=2}^{4} \sum_{j=-k+1}^{N} \alpha_{j,k} B_{j,k}\right\|_r \leq \left(3^{r-1} h\right)^{1/r} \|\alpha\|_r, & 1 \leq r \leq \infty \\ ch^{1/r} \|\alpha\|_r \leq \left\|\sum_{k=2}^{4} \sum_{j=-k+1}^{N} \alpha_{j,k} B_{j,k}\right\|_r \leq (3h)^{1/r} \|\alpha\|_r, & 0 < r < 1 \end{cases},$$

where $\boldsymbol{\alpha} := \left(\alpha_{-1,2}, \alpha_{0,2}, ..., \alpha_{N,2}, ..., \alpha_{N,4}\right)$. In particular, under Assumption A2, for any fixed $\boldsymbol{\theta}$, one has

$$ch^{1/2} \|\alpha\|_2 \leq \left\|\sum_{k=2}^{4} \sum_{j=-k+1}^{N} \alpha_{j,k} B_{j,k}\right\|_{2,\theta} \leq Ch^{1/2} \|\alpha\|_2$$

PROOF. It follows from the B-spline Property on page 96 of de Boor (2001), $\sum_{k=2}^{4} \sum_{j=-k+1}^{N} B_{j,k} \equiv 3$ on $[0, 1]$. So the right inequality is immediate for $r = \infty$. When $1 \leq r < \infty$, Hölder's inequality implies that

$$\begin{aligned} \left|\sum_{k=2}^{4} \sum_{j=-k+1}^{N} \alpha_{j,k} B_{j,k}\right| &\leq \left(\sum_{k=2}^{4} \sum_{j=-k+1}^{N} \left|\alpha_{j,k}\right|^r B_{j,k}\right)^{1/r} \left(\sum_{k=2}^{4} \sum_{j=-k+1}^{N} B_{j,k}\right)^{1-1/r} \\ &= 3^{1-1/r} \left(\sum_{k=2}^{4} \sum_{j=-k+1}^{N} \left|\alpha_{j,k}\right|^r B_{j,k}\right)^{1/r}. \end{aligned}$$

Since all the knots are equally spaced, $\int_{-\infty}^{\infty} B_{j,k}(u)\, du \leq h$, the right inequality follows from

$$\int_0^1 \left|\sum_{k=2}^{4} \sum_{j=-k+1}^{N} \alpha_{j,k} B_{j,k}(u)\right|^r du \leq 3^{r-1} h \|\alpha\|_r^r.$$

When $r < 1$, one has

$$\left|\sum_{k=2}^{4}\sum_{j=-k+1}^{N}\alpha_{j,k}B_{j,k}\right|^{r} \leq \sum_{k=2}^{4}\sum_{j=-k+1}^{N}\left|\alpha_{j,k}\right|^{r}B_{j,k}^{r}.$$

Since $\int_{-\infty}^{\infty}B_{j,k}^{r}(u)\,du \leq t_{j+k}-t_{j} = kh$ and

$$\int_{0}^{1}\left|\sum_{k=2}^{4}\sum_{j=-k+1}^{N}\alpha_{j,k}B_{j,k}(u)\right|^{r}du \leq \|\alpha\|_{r}^{r}\int_{-\infty}^{\infty}B_{j,k}^{r}(u)\,du \leq 3h\,\|\alpha\|_{r}^{r},$$

the right inequality follows in this case as well. For the left inequalities, Theorem 5.4.2, DeVore and Lorentz (1993) implies that

$$\left|\alpha_{j,k}\right| \leq C_{1}h^{-1/r}\int_{t_{j}}^{t_{j+1}}\left|\sum_{j=-k+1}^{N}\alpha_{j,k}B_{j,k}(u)\right|^{r}du$$

for any $0 < r \leq \infty$, so

$$\left|\alpha_{j,k}\right|^{r} \leq C_{1}^{r}h^{-1}\int_{t_{j}}^{t_{j+1}}\left|\sum_{j=-k+1}^{N}\alpha_{j,k}B_{j,k}(u)\right|^{r}du.$$

Since each $u \in [0,1]$ appears in at most k intervals $\left(t_{j},t_{j+k}\right)$, adding up these inequalities, one obtains that

$$\|\alpha\|_{r}^{r} \leq C_{1}h^{-1}\sum_{k=1}^{4}\int_{t_{j}}^{t_{j+k}}\left|\sum_{j=-k+1}^{N}\alpha_{j,k}B_{j,k}(u)\right|^{r}du \leq 3Ch^{-1}\left\|\sum_{j=-k+1}^{N}\alpha_{j,k}B_{j,k}\right\|_{r}^{r}.$$

The left inequality follows. □

For any functions ϕ and φ, define the empirical inner product and the empirical norm as

$$\langle\phi,\varphi\rangle_{\theta} = \int_{0}^{1}\phi(u)\,\varphi(u)\,f_{\theta}(u)\,du,\quad \|\phi\|_{2,n,\theta}^{2} = n^{-1}\sum_{i=1}^{n}\phi^{2}\left(U_{\theta,i}\right).$$

In addition, if functions ϕ,φ are $L_2\,[0,1]$-integrable, define the theoretical inner product and its corresponding theoretical L_2 norm as

$$\|\phi\|_{2,\theta}^{2} = \int_{0}^{1}\phi^{2}(u)\,f_{\theta}(u)\,du,\quad \langle\phi,\varphi\rangle_{n,\theta} = n^{-1}\sum_{i=1}^{n}\phi\left(U_{\theta,i}\right)\varphi\left(U_{\theta,i}\right).$$

LEMMA 4.6.2. *Under Assumptions (C2), (C5) and (C6), with probability* 1,

$$\sup_{\theta\in S_{c}^{d-1}}\max_{\substack{k,k'=2,3,4\\ 1\leq j,j'\leq N}}\left|\left\langle B_{j,k},B_{j',k'}\right\rangle_{n,\theta}-\left\langle B_{j,k},B_{j',k'}\right\rangle_{\theta}\right| = O\left\{(nN)^{-1/2}\log n\right\}.$$

PROOF. We only prove the case $k = k' = 4$, all other cases are similar. Let

$$\zeta_{\theta,j,j',i} = B_{j,4}\left(U_{\theta,i}\right) B_{j',4}\left(U_{\theta,i}\right) - EB_{j,4}\left(U_{\theta,i}\right) B_{j',4}\left(U_{\theta,i}\right),$$

with the second moment

$$E\zeta^2_{\theta,j,j',i} = E\left[B^2_{j,4}\left(U_{\theta,i}\right) B^2_{j',4}\left(U_{\theta,i}\right)\right] - \left\{EB_{j,4}\left(U_{\theta,i}\right) B_{j',4}\left(U_{\theta,i}\right)\right\}^2,$$

where $\left\{EB_{j,4}\left(U_{\theta,i}\right) B_{j',4}\left(U_{\theta,i}\right)\right\}^2 \sim N^{-2}$ and $E\left[B^2_{j,4}\left(U_{\theta,i}\right) B^2_{j',4}\left(U_{\theta,i}\right)\right] \sim N^{-1}$ by Assumption (C2). Hence, $E\zeta^2_{\theta,j,j',i} \sim N^{-1}$. The k-th moment is given by

$$\begin{aligned} E\left|\zeta_{\theta,j,j',i}\right|^k &= E\left|B_{j,4}\left(U_{\theta,i}\right) B_{j',4}\left(U_{\theta,i}\right) - EB_{j,4}\left(U_{\theta,i}\right) B_{j',4}\left(U_{\theta,i}\right)\right|^k \\ &\le 2^{k-1}\left\{E\left|B_{j,4}\left(U_{\theta,i}\right) B_{j',4}\left(U_{\theta,i}\right)\right|^k + \left|EB_{j,4}\left(U_{\theta,i}\right) B_{j',4}\left(U_{\theta,i}\right)\right|^k\right\}, \end{aligned}$$

where $\left|EB_{j,4}\left(U_{\theta,i}\right) B_{j',4}\left(U_{\theta,i}\right)\right|^k \sim N^{-k}$, $E\left|EB_{j,4}\left(U_{\theta,i}\right) B_{j',4}\left(U_{\theta,i}\right)\right|^k \sim N^{-1}$ by Assumption (C2). Thus, there exists a positive constant C such that $E\left|\zeta_{\theta,j,j',i}\right|^k \le C^{k-1}k!E\zeta^2_{j,j',i}$. So the Cramér's condition is satisfied with Cramér's constant c^*. By Lemma 2.6.2, one has for $k = 3$

$$P\left\{\left|n^{-1}\sum_{i=1}^{n}\zeta_{\theta,j,j',i}\right| \ge \delta_n\right\} \le a_1 \exp\left(-\frac{q\delta_n^2}{25m_2^2 + 5c^*\delta_n}\right) + a_2(k)\,\alpha\left(\left[\frac{n}{q+1}\right]\right)^{6/7},$$

where

$$\delta_n = \delta\frac{\log n}{\sqrt{nN}},\quad a_1 = 2\frac{n}{q} + 2\left(1 + \frac{\delta^2 (nN)^{-1}\log^2 n}{25m_2^2 + 5c^*\delta_n}\right),\quad m_2^2 \sim N^{-1},$$

$$a_2(3) = 11n\left(1 + \frac{5m_3^{6/7}}{\delta_n}\right),\quad m_3 = \max_{1\le i\le n}\left\|\zeta_{\theta,j,j',i}\right\|_3 \le cN^{1/3}.$$

Observe that $5c\delta_n = o(1)$ by Assumption (C6), then by taking q such that $\left[\frac{n}{q+1}\right] \ge c_0 \log n$, $q \ge c_1 n/\log n$ for some constants c_0, c_1, one has $a_1 = O(n/q) = O(\log n)$, $a_2(3) = o\left(n^2\right)$ via Assumption (C6) again. Assumption A5 yields that

$$\alpha\left(\left[\frac{n}{q+1}\right]\right)^{6/7} \le \left\{K_0 \exp\left(-\lambda_0\left[\frac{n}{q+1}\right]\right)\right\}^{6/7} \le Cn^{-6\lambda_0 c_0/7}.$$

Thus, for fixed $\theta \in S_c^{d-1}$, when n large enough

$$P\left\{\frac{1}{n}\left|\sum_{i=1}^{n}\zeta_{\theta,j,j',i}\right| > \delta_n\right\} \le c\log n \exp\left\{-c_2\delta^2\log n\right\} + Cn^{2-6\lambda_0 c_0/7}. \tag{4.6.1}$$

We divide the $d-1$ intervals into $n^{6/(d-1)}$ equally spaced intervals with disjoint endpoints $-1 = \theta_{p,0} < \theta_{p,1} < \dots < \theta_{p,M_n} = 1$, for $p = 1, \dots, d-1$. Projecting these small cylinders onto S_c^{d-1}, the radius of each patch Λ_r, $r = 1, \dots, M_n$ is bounded by cM_n^{-1}. Denote the projection of the M_n points as $\boldsymbol{\theta}_r = \left(\boldsymbol{\theta}_{r,-d}, \sqrt{1 - \|\boldsymbol{\theta}_{r,-d}\|_2^2}\right)$, $r = 0, 1, \dots, M_n$. Employing the discretization method, $\sup_{\boldsymbol{\theta} \in S_c^{d-1}} \max_{1 \leq j,j' \leq N} \left|\zeta_{\boldsymbol{\theta},j,j',i}\right|$ is bounded by

$$\sup_{0 \leq r \leq M_n} \max_{1 \leq j,j' \leq N} \left|\zeta_{\boldsymbol{\theta}_r,j,j',i}\right| + \sup_{0 \leq r \leq M_n} \max_{1 \leq j,j' \leq N} \sup_{\boldsymbol{\theta} \in \Lambda_r} \left|\zeta_{\boldsymbol{\theta},j,j',i} - \zeta_{\boldsymbol{\theta}_r,j,j',i}\right|. \qquad (4.6.2)$$

By (4.6.1) and Assumption (C6), there exists large enough value $\delta > 0$ such that

$$P\left\{\frac{1}{n}\left|\sum_{i=1}^{n} \zeta_{\boldsymbol{\theta}_r,j,j',i}\right| > \delta_n\right\} \leq n^{-10},$$

which implies that

$$\sum_{n=1}^{\infty} P\left\{\max_{1 \leq j,j' \leq N}\left|n^{-1}\sum_{l=1}^{n} \zeta_{\boldsymbol{\theta}_r,j,j',i}\right| \geq \delta_n\right\} \leq 2\sum_{n=1}^{\infty} N^2 M_n n^{-10} \leq C\sum_{n=1}^{\infty} n^{-3} < \infty.$$

Thus, Borel-Cantelli Lemma entails that

$$\sup_{0 \leq r \leq M_n} \max_{1 \leq j,j' \leq N}\left|n^{-1}\sum_{l=1}^{n} \zeta_{\boldsymbol{\theta}_r,j,j',i}\right| = O\left(\frac{\log n}{\sqrt{nN}}\right), a.s.. \qquad (4.6.3)$$

Employing Lipschitz continuity of the cubic B-spline, one has with probability 1

$$\sup_{0 \leq r \leq M_n} \max_{1 \leq j,j' \leq N} \sup_{\boldsymbol{\theta} \in \Lambda_r}\left|n^{-1}\sum_{i=1}^{n}\left\{\zeta_{\boldsymbol{\theta},j,j',i} - \zeta_{\boldsymbol{\theta}_r,j,j',i}\right\}\right| = O\left(M_n^{-1}h^{-6}\right). \qquad (4.6.4)$$

Therefore Assumption (C2), (4.6.2), (4.6.3) and (4.6.4) lead to the desired result. □

Denote by $G = G^{(0)} \cup G^{(1)} \cup G^{(2)}$ the space of all linear, quadratic and cubic spline functions on $[0,1]$. We establish the uniform rate at which the empirical inner product approximates the theoretical inner product for all B-splines $B_{j,k}$ with $k = 2, 3, 4$.

LEMMA 4.6.3. *Under Assumptions (C2), (C5) and (C6), one has*

$$A_n = \sup_{\boldsymbol{\theta} \in S_c^{d-1}} \sup_{\gamma_1,\gamma_2 \in G}\left|\frac{\langle\gamma_1,\gamma_2\rangle_{n,\boldsymbol{\theta}} - \langle\gamma_1,\gamma_2\rangle_{\boldsymbol{\theta}}}{\|\gamma_1\|_{2,\boldsymbol{\theta}}\|\gamma_2\|_{2,\boldsymbol{\theta}}}\right| = O\left\{(nh)^{-1/2}\log n\right\}, a.s.. \qquad (4.6.5)$$

PROOF. Denote without loss of generality,

$$\gamma_1 = \sum_{k=2}^{4}\sum_{j=-k+1}^{N} \alpha_{jk}B_{j,k}, \quad \gamma_2 = \sum_{k=2}^{4}\sum_{j=-k+1}^{N} \beta_{jk}B_{j,k},$$

for any two $3\,(N+3)$-vectors

$$\boldsymbol{\alpha} = \left(\alpha_{-1,2}, \alpha_{0,2}, ..., \alpha_{N,2}, ..., \alpha_{N,4}\right), \boldsymbol{\beta} = \left(\beta_{-1,2}, \beta_{0,2}, ..., \beta_{N,2}, ..., \beta_{N,4}\right).$$

Then for fixed $\boldsymbol{\theta}$

$$\begin{aligned}\langle\gamma_1, \gamma_2\rangle_{n,\boldsymbol{\theta}} &= \frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{k=2}^{4}\sum_{j=-k+1}^{N}\alpha_{j,k}B_{j,k}\left(U_{\boldsymbol{\theta},i}\right)\right\}\left\{\sum_{k=2}^{4}\sum_{j=-k+1}^{N}\beta_{j,k}B_{j,k}\left(U_{\boldsymbol{\theta},i}\right)\right\}\\ &= \sum_{k=2}^{4}\sum_{j=-k+1}^{N}\sum_{k'=2}^{4}\sum_{j'=-k+1}^{N}\alpha_{j,k}\beta_{j',k'}\left\langle B_{j,k}, B_{j',k'}\right\rangle_{n,\boldsymbol{\theta}},\end{aligned}$$

$$\begin{aligned}\|\gamma_1\|_{2,\boldsymbol{\theta}}^2 &= \sum_{k=2}^{4}\sum_{j=-k+1}^{N}\sum_{k'=2}^{4}\sum_{j'=-k+1}^{N}\alpha_{j,k}\alpha_{j',k'}\left\langle B_{j,k}, B_{j',k'}\right\rangle_{\boldsymbol{\theta}},\\ \|\gamma_2\|_{2,\boldsymbol{\theta}}^2 &= \sum_{k=2}^{4}\sum_{j=-k+1}^{N}\sum_{k'=2}^{4}\sum_{j'=-k+1}^{N}\beta_{j,k}\beta_{j',k'}\left\langle B_{j,k}, B_{j',k'}\right\rangle_{\boldsymbol{\theta}}.\end{aligned}$$

According to Lemma 4.6.1, one has for any $\boldsymbol{\theta} \in S_c^{d-1}$,

$$c_1 h\,\|\boldsymbol{\alpha}\|_2^2 \le \|\gamma_1\|_{2,\boldsymbol{\theta}}^2 \le c_2 h\,\|\boldsymbol{\alpha}\|_2^2, c_1 h\,\|\boldsymbol{\beta}\|_2^2 \le \|\gamma_2\|_{2,\boldsymbol{\theta}}^2 \le c_2 h\,\|\boldsymbol{\beta}\|_2^2,$$

$$c_1 h\,\|\boldsymbol{\alpha}\|_2\,\|\boldsymbol{\beta}\|_2 \le \|\gamma_1\|_{2,\boldsymbol{\theta}}\,\|\gamma_2\|_{2,\boldsymbol{\theta}} \le c_2 h\,\|\boldsymbol{\alpha}\|_2\,\|\boldsymbol{\beta}\|_2.$$

Hence

$$\begin{aligned}A_n &= \sup_{\boldsymbol{\theta}\in S_c^{d-1}}\ \sup_{\gamma_1\in\gamma,\gamma_2\in\Gamma}\left|\frac{\langle\gamma_1,\gamma_2\rangle_{n,\boldsymbol{\theta}} - \langle\gamma_1,\gamma_2\rangle_{\boldsymbol{\theta}}}{\|\gamma_1\|_{2,\boldsymbol{\theta}}\,\|\gamma_2\|_{2,\boldsymbol{\theta}}}\right| \le \frac{\|\boldsymbol{\alpha}\|_\infty\,\|\boldsymbol{\beta}\|_\infty}{c_1 h\,\|\boldsymbol{\alpha}\|_2\,\|\boldsymbol{\beta}\|_2}\\ &\times \sup_{\boldsymbol{\theta}\in S_c^{d-1}}\ \max_{\substack{k,k'=2,3,4\\ 1\le j,j'\le N}}\left|\frac{1}{n}\sum_{i=1}^{n}\left\{\left\langle B_{j,k}, B_{j',k'}\right\rangle_{n,\boldsymbol{\theta}} - \left\langle B_{j,k}, B_{j',k'}\right\rangle_{\boldsymbol{\theta}}\right\}\right|,\end{aligned}$$

$$A_n \le c_0 h^{-1}\sup_{\boldsymbol{\theta}\in S_c^{d-1}}\ \max_{\substack{k,k'=2,3,4\\ 1\le j,j'\le N}}\left|\frac{1}{n}\sum_{i=1}^{n}\left\{\left\langle B_{j,k}, B_{j',k'}\right\rangle_{n,\boldsymbol{\theta}} - \left\langle B_{j,k}, B_{j',k'}\right\rangle_{\boldsymbol{\theta}}\right\}\right|,$$

which, together with Lemma 4.6.2, imply (4.6.5). □

4.6.2 Proof of Proposition 4.2.1

For any fixed $\boldsymbol{\theta}$, write the response $\mathbf{Y}^T = (Y_1, ..., Y_n)$ as the sum of a signal vector $\boldsymbol{\gamma}_{\boldsymbol{\theta}}$, a parametric noise vector $\mathbf{E}_{\boldsymbol{\theta}}$ and a systematic noise vector $\mathbf{E}$, i.e.,

$$\mathbf{Y} = \boldsymbol{\gamma}_{\boldsymbol{\theta}} + \mathbf{E}_{\boldsymbol{\theta}} + \mathbf{E},$$

in which the vectors $\gamma_{\theta}^{T} = \left\{\gamma_{\theta}\left(U_{\theta,1}\right), ..., \gamma_{\theta}\left(U_{\theta,n}\right)\right\}$, $\mathbf{E}^{T} = \left\{\sigma\left(\mathbf{X}_1\right)\varepsilon_1, ..., \sigma\left(\mathbf{X}_n\right)\varepsilon_n\right\}$ and $\mathbf{E}_{\theta}^{T} = \left\{m\left(\mathbf{X}_1\right) - \gamma_{\theta}\left(U_{\theta,1}\right), ..., m\left(\mathbf{X}_n\right) - \gamma_{\theta}\left(U_{\theta,n}\right)\right\}$.

REMARK 4.A.1: If m is a genuine single-index function, then $\mathbf{E}_{\theta_0} \equiv 0$, thus the proposed SIP model is exactly the single-index model.

Let $G_{n,\,\theta}^{(2)}$ be the cubic spline space spanned by $\left\{\mathbf{B}_{j,4}\left(U_{\theta,i}\right)\right\}_{i=1}^{n}$, $-3 \leq j \leq N$ for fixed θ. Projecting $\mathbf{Y}$ onto $G_{n,\,\theta}^{(2)}$ yields that

$$\hat{\gamma}_{\theta} = \left\{\hat{\gamma}_{\theta}\left(U_{\theta,1}\right), ..., \hat{\gamma}_{\theta}\left(U_{\theta,n}\right)\right\}^{T} = \mathrm{Proj}_{G_{n,\theta}^{(2)}}\gamma_{\theta} + \mathrm{Proj}_{G_{n,\theta}^{(2)}}\mathbf{E}_{\theta} + \mathrm{Proj}_{G_{n,\theta}^{(2)}}\mathbf{E},$$

where $\hat{\gamma}_{\theta}$ is given in (4.2.8). We break the spline estimation error $\hat{\gamma}_{\theta}\left(u_{\theta}\right) - \gamma_{\theta}\left(u_{\theta}\right)$ into a bias term $\tilde{\gamma}_{\theta}\left(u_{\theta}\right) - \gamma_{\theta}\left(u_{\theta}\right)$ and two noise terms $\bar{\varepsilon}_{\theta}\left(u_{\theta}\right)$ and $\hat{\varepsilon}_{\theta}\left(u_{\theta}\right)$

$$\hat{\gamma}_{\theta}\left(u_{\theta}\right) - \gamma_{\theta}\left(u_{\theta}\right) = \left\{\tilde{\gamma}_{\theta}\left(u_{\theta}\right) - \gamma_{\theta}\left(u_{\theta}\right)\right\} + \bar{\varepsilon}_{\theta}\left(u_{\theta}\right) + \hat{\varepsilon}_{\theta}\left(u_{\theta}\right), \tag{4.6.6}$$

where

$$\tilde{\gamma}_{\theta}\left(u\right) = \left\{B_{j,4}\left(u\right)\right\}_{-3\leq j\leq N}^{T}\mathbf{V}_{n,\theta}^{-1}\left\{\left\langle\gamma_{\theta}, B_{j,4}\right\rangle_{n,\theta}\right\}_{j=-3}^{N}, \tag{4.6.7}$$

$$\bar{\varepsilon}_{\theta}\left(u\right) = \left\{B_{j,4}\left(u\right)\right\}_{-3\leq j\leq N}^{T}\mathbf{V}_{n,\theta}^{-1}\left\{\left\langle\mathbf{E}_{\theta}, B_{j,4}\right\rangle_{n,\theta}\right\}_{j=-3}^{N}, \tag{4.6.8}$$

$$\hat{\varepsilon}_{\theta}\left(u\right) = \left\{B_{j,4}\left(u\right)\right\}_{-3\leq j\leq N}^{T}\mathbf{V}_{n,\theta}^{-1}\left\{\left\langle\mathbf{E}, B_{j,4}\right\rangle_{n,\theta}\right\}_{j=-3}^{N}. \tag{4.6.9}$$

In the above, denote by $\mathbf{V}_{n,\theta}$ the empirical inner product matrix of the cubic B-spline basis and similarly, the theoretical inner product matrix as $\mathbf{V}_{\theta}$

$$\mathbf{V}_{n,\theta} = \frac{1}{n}\mathbf{B}_{\theta}^{T}\mathbf{B}_{\theta} = \left\{\left\langle B_{j',4}, B_{j,4}\right\rangle_{n,\theta}\right\}_{j,j'=-3}^{N}, \mathbf{V}_{\theta} = \left\{\left\langle B_{j',4}, B_{j,4}\right\rangle_{\theta}\right\}_{j,j'=-3}^{N}. \tag{4.6.10}$$

The next lemma is a special case of Theorem 13.4.3 in DeVore and Lorentz (1993).

LEMMA 4.6.4. *If a bi-infinite matrix with bandwidth r has a bounded inverse $\mathbf{A}^{-1}$ on l_2 and $\kappa = \kappa\left(\mathbf{A}\right) := \left\|\mathbf{A}\right\|_2\left\|\mathbf{A}^{-1}\right\|_2$ is the condition number of $\mathbf{A}$, then $\left\|\mathbf{A}^{-1}\right\|_{\infty} \leq 2c_0\left(1-\nu\right)^{-1}$, with $c_0 = \nu^{-2r}\left\|\mathbf{A}^{-1}\right\|_2$, $\nu = \left(\kappa^2 - 1\right)^{1/4r}\left(\kappa^2 + 1\right)^{-1/4r}$.*

LEMMA 4.6.5. *Under Assumptions (C2), (C5) and (C6), there exist constants $0 < c_V < C_V$ such that $c_V N^{-1}\left\|\mathbf{w}\right\|_2^2 \leq \mathbf{w}^T\mathbf{V}_{\theta}\mathbf{w} \leq C_V N^{-1}\left\|\mathbf{w}\right\|_2^2$ and*

$$c_V N^{-1}\left\|\mathbf{w}\right\|_2^2 \leq \mathbf{w}^T\mathbf{V}_{n,\theta}\mathbf{w} \leq C_V N^{-1}\left\|\mathbf{w}\right\|_2^2, a.s., \tag{4.6.11}$$

with matrices $\mathbf{V}_{\theta}$ *and* $\mathbf{V}_{n,\theta}$ *defined in (4.6.10). In addition, there exists a constant* $C > 0$ *such that*

$$\sup_{\theta \in S_c^{d-1}} \left\| \mathbf{V}_{n,\theta}^{-1} \right\|_{\infty} \leq CN, a.s., \sup_{\theta \in S_c^{d-1}} \left\| \mathbf{V}_{\theta}^{-1} \right\|_{\infty} \leq CN. \tag{4.6.12}$$

PROOF. First we compute the lower and upper bounds for the eigenvalues of $\mathbf{V}_{n,\theta}$. Let $\mathbf{w}$ be any $(N+4)$-vector and denote $\gamma_{\mathbf{w}}(u) = \sum_{j=-3}^{N} w_j B_{j,4}(u)$, then $\mathbf{B}_{\theta}\mathbf{w} = \{\gamma_{\mathbf{w}}(U_{\theta,1}), ..., \gamma_{\mathbf{w}}(U_{\theta,n})\}^T$ and the definition of A_n in (4.6.5) from Lemma 4.6.3 entails that

$$\|\gamma_{\mathbf{w}}\|_{2,\theta}^2 (1 - A_n) \leq \mathbf{w}^T \mathbf{V}_{n,\theta} \mathbf{w} = \|\gamma_{\mathbf{w}}\|_{2,n,\theta}^2 \leq \|\gamma_{\mathbf{w}}\|_{2,\theta}^2 (1 + A_n). \tag{4.6.13}$$

Using Theorem 5.4.2 of DeVore and Lorentz (1993) and Assumption (C2), one obtains that

$$c_f \frac{C}{N} \|\mathbf{w}\|_2^2 \leq \|\gamma_{\mathbf{w}}\|_{2,\theta}^2 = \mathbf{w}^T \mathbf{V}_{\theta} \mathbf{w} = \left\| \sum_{j=-3}^{N} w_j B_{j,4} \right\|_{2,\theta}^2 \leq C_f \frac{C}{N} \|\mathbf{w}\|_2^2, \tag{4.6.14}$$

which, together with (4.6.13), yield

$$c_f C N^{-1} \|\mathbf{w}\|_2^2 (1 - A_n) \leq \mathbf{w}^T \mathbf{V}_{n,\theta} \mathbf{w} \leq C_f C N^{-1} \|\mathbf{w}\|_2^2 (1 + A_n). \tag{4.6.15}$$

Now the order of A_n in (4.6.5), together with (4.6.14) and (4.6.15) implies (4.6.11), in which $c_V = c_f C, C_V = C_f C$. Next, denote by $\lambda_{\max}(\mathbf{V}_{n,\theta})$ and $\lambda_{\min}(\mathbf{V}_{n,\theta})$ the maximum and minimum eigenvalue of $\mathbf{V}_{n,\theta}$, simple algebra and (4.6.11) entail that

$$C_V N^{-1} \geq \|\mathbf{V}_{n,\theta}\|_2 = \lambda_{\max}(\mathbf{V}_{n,\theta}), \left\| \mathbf{V}_{n,\theta}^{-1} \right\|_2 = \lambda_{\min}^{-1}(\mathbf{V}_{n,\theta}) \leq c_V^{-1} N, a.s.,$$

thus

$$\kappa := \|\mathbf{V}_{n,\theta}\|_2 \left\| \mathbf{V}_{n,\theta}^{-1} \right\|_2 = \lambda_{\max}(\mathbf{V}_{n,\theta}) \lambda_{\min}^{-1}(\mathbf{V}_{n,\theta}) \leq C_V c_V^{-1} < \infty, a.s..$$

Meanwhile, let $\mathbf{w}_j =$ the $(N+4)$-vector with all zeros except the j-th element being $1, j = -3, ..., N$. Then clearly

$$\mathbf{w}_j^T \mathbf{V}_{n,\theta} \mathbf{w}_j = \frac{1}{n} \sum_{i=1}^{n} B_{j,4}^2 (U_{\theta,i}) = \|B_{j,4}\|_{n,\theta}^2, \|\mathbf{w}_j\|_2 = 1, -3 \leq j \leq N$$

and in particular

$$\begin{aligned} \mathbf{w}_0^T \mathbf{V}_{n,\theta} \mathbf{w}_0 &\leq \lambda_{\max}(\mathbf{V}_{n,\theta}) \|\mathbf{w}_0\|_2 = \lambda_{\max}(\mathbf{V}_{n,\theta}), \\ \mathbf{w}_{-3}^T \mathbf{V}_{n,\theta} \mathbf{w}_{-3} &\geq \lambda_{\min}(\mathbf{V}_{n,\theta}) \|\mathbf{w}_{-3}\|_2 = \lambda_{\min}(\mathbf{V}_{n,\theta}). \end{aligned}$$

This, together with (4.6.5) yields that

$$\kappa = \lambda_{\max}\left(\mathrm{V}_{n,\boldsymbol{\theta}}\right)\lambda_{\min}^{-1}\left(\mathrm{V}_{n,\boldsymbol{\theta}}\right) \geq \frac{\mathrm{w}_0^T \mathrm{V}_{n,\boldsymbol{\theta}} \mathrm{w}_0}{\mathrm{w}_{-3}^T \mathrm{V}_{n,\boldsymbol{\theta}} \mathrm{w}_{-3}} = \frac{\left\|B_{0,4}\right\|_{n,\boldsymbol{\theta}}^2}{\left\|B_{-3,4}\right\|_{n,\boldsymbol{\theta}}^2} \geq \frac{\left\|B_{0,4}\right\|_{\boldsymbol{\theta}}^2}{\left\|B_{-3,4}\right\|_{\boldsymbol{\theta}}^2}\frac{1-A_n}{1+A_n},$$

which leads to $\kappa \geq C > 1, a.s.$ because the definition of B-spline and Assumption (C2) ensure that $\left\|B_{0,4}\right\|_{\boldsymbol{\theta}}^2 \geq C_0 \left\|B_{-3,4}\right\|_{\boldsymbol{\theta}}^2$ for some constant $C_0 > 1$. Next applying Lemma 4.6.4 with $\nu = \left(\kappa^2-1\right)^{1/16}\left(\kappa^2+1\right)^{-1/16}$ and $c_0 = \nu^{-8}\left\|\mathrm{V}_{n,\boldsymbol{\theta}}^{-1}\right\|_2$, one gets $\left\|\mathrm{V}_{n,\boldsymbol{\theta}}^{-1}\right\|_\infty \leq 2\nu^{-8}N\left(1-\nu\right)^{-1} = CN, a.s.$. Hence part one of (4.6.12) follows. Part two of (4.6.12) is proved in the same fashion. □

In the following, denote by $Q_T\left(m\right)$ the 4-th order quasi-interpolant of m corresponding to the knots T, see equation (4.12), page 146 of DeVore and Lorentz (1993). According to Theorem 7.7.4, DeVore and Lorentz (1993), the following lemma holds.

LEMMA 4.6.6. *There exists a constant $C > 0$, such that for $0 \leq k \leq 2$ and $\gamma \in C^{(4)}\left[0,1\right]$*

$$\left\|\left(\gamma - Q_T\left(\gamma\right)\right)^{(k)}\right\|_\infty \leq C\left\|\gamma^{(4)}\right\|_\infty h^{4-k},$$

LEMMA 4.6.7. *Under Assumptions (C2), (C3), (C5) and (C6), there exists an absolute constant $C > 0$, such that for function $\tilde{\gamma}_{\boldsymbol{\theta}}\left(u\right)$ in (4.6.7)*

$$\sup_{\boldsymbol{\theta}\in S_c^{d-1}}\left\|\frac{d^k}{du^k}\left(\tilde{\gamma}_{\boldsymbol{\theta}} - \gamma_{\boldsymbol{\theta}}\right)\right\|_\infty \leq C\left\|m^{(4)}\right\|_\infty h^{4-k}, a.s., 0 \leq k \leq 2, \qquad (4.6.16)$$

PROOF. According to Lemma 2.3.3, there exists an absolute constant $C > 0$, such that

$$\sup_{\boldsymbol{\theta}\in S_c^{d-1}}\left\|\tilde{\gamma}_{\boldsymbol{\theta}} - \gamma_{\boldsymbol{\theta}}\right\|_\infty \leq C \sup_{\boldsymbol{\theta}\in S_c^{d-1}}\inf_{\gamma\in G^{(2)}}\left\|\gamma - \gamma_{\boldsymbol{\theta}}\right\|_\infty \leq C\left\|m^{(4)}\right\|_\infty h^4, a.s., \qquad (4.6.17)$$

which proves (4.6.16) for the case $k = 0$. Applying Lemma 4.6.6, one has for $0 \leq k \leq 2$

$$\sup_{\boldsymbol{\theta}\in S_c^{d-1}}\left\|\frac{d^k}{du^k}\left\{Q_T\left(\gamma_{\boldsymbol{\theta}}\right) - \gamma_{\boldsymbol{\theta}}\right\}\right\|_\infty \leq C \sup_{\boldsymbol{\theta}\in S_c^{d-1}}\left\|\gamma_{\boldsymbol{\theta}}^{(4)}\right\|_\infty h^{4-k} \leq C\left\|m^{(4)}\right\|_\infty h^{4-k}, \qquad (4.6.18)$$

As a consequence of (4.6.17) and (4.6.18) for the case $k = 0$, one has

$$\sup_{\boldsymbol{\theta}\in S_c^{d-1}}\left\|Q_T\left(\gamma_{\boldsymbol{\theta}}\right) - \tilde{\gamma}_{\boldsymbol{\theta}}\right\|_\infty \leq C\left\|m^{(4)}\right\|_\infty h^4, a.s.,$$

which, according to the differentiation of B-spline given in de Boor (2001), entails that

$$\sup_{\boldsymbol{\theta}\in S_c^{d-1}}\left\|\frac{d^k}{du^k}\left\{Q_T\left(\gamma_{\boldsymbol{\theta}}\right) - \tilde{\gamma}_{\boldsymbol{\theta}}\right\}\right\|_\infty \leq C\left\|m^{(4)}\right\|_\infty h^{4-k}, a.s., \ 0 \leq k \leq 2. \qquad (4.6.19)$$

Combining (4.6.18) and (4.6.19) proves (4.6.16) for $k = 1, 2$. □

Lemma 4.6.8. *Under Assumptions (C1), (C2), (C4) and (C5), there exists an absolute constant $C > 0$, such that*

$$\sup_{1\leq p\leq d}\sup_{\theta\in S_c^{d-1}}\left\|\frac{\partial}{\partial\theta_p}\left\{\tilde{\gamma}_{\theta}\left(U_{\theta,i}\right)-\gamma_{\theta}\left(U_{\theta,i}\right)\right\}_{i=1}^{n}\right\|_{\infty}\leq C\left\|m^{(4)}\right\|_{\infty}h^{3}, a.s., \tag{4.6.20}$$

$$\sup_{1\leq p,q\leq d}\sup_{\theta\in S_c^{d-1}}\left\|\frac{\partial^2}{\partial\theta_p\partial\theta_q}\left\{\tilde{\gamma}_{\theta}\left(U_{\theta,i}\right)-\gamma_{\theta}\left(U_{\theta,i}\right)\right\}_{i=1}^{n}\right\|_{\infty}\leq C\left\|m^{(4)}\right\|_{\infty}h^{2}, a.s.. \tag{4.6.21}$$

Proof. According to the definition of $\tilde{\gamma}_{\theta}$ in (4.6.7), and the fact that $Q_T(\gamma_{\theta})$ is a cubic spline on the knots T

$$\left\{\left\{Q_T(\gamma_{\theta})-\tilde{\gamma}_{\theta}\right\}\left(U_{\theta,i}\right)\right\}_{i=1}^{n}=\mathbf{P}_{\theta}\left\{\left\{Q_T(\gamma_{\theta})-\gamma_{\theta}\right\}\left(U_{\theta,i}\right)\right\}_{i=1}^{n},$$

which entails that

$$\frac{\partial}{\partial\theta_p}\left\{\left\{Q_T(\gamma_{\theta})-\tilde{\gamma}_{\theta}\right\}\left(U_{\theta,i}\right)\right\}_{i=1}^{n}=\frac{\partial}{\partial\theta_p}\mathbf{P}_{\theta}\left\{\left\{Q_T(\gamma_{\theta})-\gamma_{\theta}\right\}\left(U_{\theta,i}\right)\right\}_{i=1}^{n}$$
$$=\dot{\mathbf{P}}_p\left\{\left\{Q_T(\gamma_{\theta})-\gamma_{\theta}\right\}\left(U_{\theta,i}\right)\right\}_{i=1}^{n}+\mathbf{P}_{\theta}\frac{\partial}{\partial\theta_p}\left\{\left\{Q_T(\gamma_{\theta})-\gamma_{\theta}\right\}\left(U_{\theta,i}\right)\right\}_{i=1}^{n}.$$

Since

$$\frac{\partial}{\partial\theta_p}\left\{\left\{Q_T(\gamma_{\theta})-\gamma_{\theta}\right\}\left(U_{\theta,i}\right)\right\}_{i=1}^{n}=\left\{\left\{Q_T\left(\frac{\partial}{\partial\theta_p}\gamma_{\theta}\right)-\frac{\partial}{\partial\theta_p}\gamma_{\theta}\right\}\left(U_{\theta,i}\right)\right\}_{i=1}^{n}$$
$$+\left\{\frac{d}{du}\left\{Q_T(\gamma_{\theta})-\gamma_{\theta}\right\}\left(U_{\theta,i}\right)X_{ip}\right\}_{i=1}^{n},$$

applying (4.6.19) to the decomposition above produces (4.6.20). The proof of (4.6.21) is similar. □

Lemma 4.6.9. *Under Assumptions (C2), (C5) and (C6), there exists a constant $C > 0$ such that*

$$\sup_{\theta\in S_c^{d-1}}\left\|n^{-1}\mathbf{B}_{\theta}^{T}\right\|_{\infty}\leq Ch, a.s., \sup_{1\leq p\leq d}\sup_{\theta\in S_c^{d-1}}\left\|n^{-1}\dot{\mathbf{B}}_{p}^{T}\right\|_{\infty}\leq C, a.s., \tag{4.6.22}$$

$$\sup_{\theta\in S_c^{d-1}}\left\|\mathbf{P}_{\theta}\right\|_{\infty}\leq C, a.s., \sup_{1\leq p\leq d}\sup_{\theta\in S_c^{d-1}}\left\|\dot{\mathbf{P}}_{p}\right\|_{\infty}\leq Ch^{-1}, a.s.. \tag{4.6.23}$$

Proof. To prove (4.6.22), observe that for any vector $\mathbf{a}\in R^n$, with probability 1

$$\left\|n^{-1}\mathbf{B}_{\theta}^{T}\mathbf{a}\right\|_{\infty}\leq\left\|\mathbf{a}\right\|_{\infty}\max_{-3\leq j\leq N}\left|n^{-1}\sum_{i=1}^{n}B_{j,4}\left(U_{\theta,i}\right)\right|\leq Ch\left\|\mathbf{a}\right\|_{\infty}, \left\|n^{-1}\dot{\mathbf{B}}_{p}^{T}\mathbf{a}\right\|_{\infty}$$

$$\leq \|\mathbf{a}\|_\infty \max_{-3\leq j\leq N}\left|\frac{1}{nh}\sum_{i=1}^{n}\left\{\left(B_{j,3}-B_{j+1,3}\right)\left(U_{\theta,i}\right)\right\}\dot{F}_d\left(\mathbf{X}_{\theta,i}\right)X_{i,p}\right| \leq C\|\mathbf{a}\|_\infty .$$

To prove (4.6.23), one only needs to use (4.6.12), (4.6.22) and (4.3.1). □

LEMMA 4.6.10. *Under Assumptions (C2) and (C4)-(C6), one has with probability* 1

$$\sup_{\theta\in S_c^{d-1}}\left\|\frac{\mathbf{B}_\theta^T\mathbf{E}}{n}\right\|_\infty = \max_{-3\leq j\leq N}\left|n^{-1}\sum_{i=1}^{n}B_{j,4}\left(U_{\theta,i}\right)\sigma\left(\mathbf{X}_i\right)\varepsilon_i\right| = O\left(\frac{\log n}{\sqrt{nN}}\right), \quad (4.6.24)$$

$$\sup_{1\leq p\leq d}\sup_{\theta\in S_c^{d-1}}\left\|\frac{\partial}{\partial\theta_p}\left(\frac{\mathbf{B}_\theta^T\mathbf{E}}{n}\right)\right\|_\infty = \sup_{1\leq p\leq d}\sup_{\theta\in S_c^{d-1}}\left\|\frac{\dot{\mathbf{B}}_p^T\mathbf{E}}{n}\right\|_\infty = O\left(\frac{\log n}{\sqrt{nh}}\right). \quad (4.6.25)$$

Similarly, under Assumptions (C2), (C4)-(C6), with probability 1

$$\begin{aligned}\sup_{\theta\in S_c^{d-1}}\left\|\frac{\mathbf{B}_\theta^T\mathbf{E}_\theta}{n}\right\|_\infty &= \sup_{\theta\in S_c^{d-1}}\max_{-3\leq j\leq N}\left|\frac{1}{n}\sum_{i=1}^{n}B_{j,4}\left(U_{\theta,i}\right)\left\{m\left(\mathbf{X}_i\right)-\gamma_\theta\left(U_{\theta,i}\right)\right\}\right| \\ &= O\left(\frac{\log n}{\sqrt{nN}}\right), \end{aligned} \quad (4.6.26)$$

$$\sup_{1\leq p\leq d}\sup_{\theta\in S_c^{d-1}}\left\|\frac{\partial}{\partial\theta_p}\left(\frac{\mathbf{B}_\theta^T\mathbf{E}_\theta}{n}\right)\right\|_\infty = O\left(\frac{\log n}{\sqrt{nh}}\right), a.s.. \quad (4.6.27)$$

PROOF. We decompose the noise variable ε_i into a truncated part and a tail part $\varepsilon_i = \varepsilon_{i,1}^{D_n} + \varepsilon_{i,2}^{D_n} + m_i^{D_n}$, where $D_n = n^\eta\,(1/3 < \eta < 2/5)$, $\varepsilon_{i,1}^{D_n} = \varepsilon_i I\left\{|\varepsilon_i| > D_n\right\}$,

$$\varepsilon_{i,2}^{D_n} = \varepsilon_i I\left\{|\varepsilon_i| \leq D_n\right\} - m_i^{D_n}, m_i^{D_n} = E\left[\varepsilon_i I\left\{|\varepsilon_i| \leq D_n\right\}|\mathbf{X}_i\right].$$

It is straightforward to verify that the mean of the truncated part is uniformly bounded by D_n^{-2}, so the boundedness of B-spline basis and of the function σ^2 entail that

$$\sup_{\theta\in S_c^{d-1}}\left|\frac{1}{n}\sum_{i=1}^{n}B_{j,4}\left(U_{\theta,i}\right)\sigma\left(\mathbf{X}_i\right)m_i^{D_n}\right| = O\left(D_n^{-2}\right) = o\left(n^{-2/3}\right).$$

The tail part vanishes almost surely

$$\sum_{n=1}^{\infty}P\left\{|\varepsilon_n| > D_n\right\} \leq \sum_{n=1}^{\infty}D_n^{-3} < \infty.$$

Borel-Cantelli Lemma implies that

$$\left|\frac{1}{n}\sum_{i=1}^{n}B_{j,4}\left(U_{\theta,i}\right)\sigma\left(\mathbf{X}_i\right)\varepsilon_{i,1}^{D_n}\right| = O\left(n^{-k}\right), \text{ for any } k > 0.$$

For the truncated part, using Bernstein's inequality and discretization as in Lemma 4.6.2

$$\sup_{\theta \in S_c^{d-1}} \sup_{1 \leq j \leq N} \left| n^{-1} \sum_{i=1}^{n} B_{j,4}\left(U_{\theta,i}\right) \sigma\left(\mathbf{X}_i\right) \varepsilon_{i,2}^{D_n} \right| = O\left(\log n / \sqrt{nN}\right), a.s..$$

Therefore (4.6.24) is established as with probability 1

$$\sup_{\theta \in S_c^{d-1}} \left\| \frac{1}{n} \mathbf{B}_{\theta}^{\mathbf{T}} \mathbf{E} \right\|_{\infty} = o\left(n^{-2/3}\right) + O\left(n^{-k}\right) + O\left(\log n / \sqrt{nN}\right) = O\left(\log n / \sqrt{nN}\right).$$

The proofs of (4.6.25), (4.6.26) are similar as $E\left\{m\left(\mathbf{X}_i\right) - \gamma_{\theta}\left(U_{\theta,i}\right) \left| U_{\theta,i}\right.\right\} \equiv 0$, but no truncation is needed for (4.6.26) as $\sup_{\theta \in S_c^{d-1}} \max_{1 \leq i \leq n} \left| m\left(\mathbf{X}_i\right) - \gamma_{\theta}\left(U_{\theta,i}\right) \right| \leq C < \infty$. Meanwhile, to prove (4.6.27), we note that for any $p = 1, ..., d$

$$\frac{\partial}{\partial \theta_p}\left(\mathbf{B}_{\theta}^{T} \mathbf{E}_{\theta}\right) = \left\{ \sum_{i=1}^{n} \frac{\partial}{\partial \theta_p}\left[B_{j,4}\left(U_{\theta,i}\right)\left\{m\left(\mathbf{X}_i\right) - \gamma_{\theta}\left(U_{\theta,i}\right)\right\}\right] \right\}_{j=-3}^{N}.$$

According to (4.2.6), one has $\gamma_{\theta}\left(U_{\theta}\right) \equiv E\left\{m\left(\mathbf{X}\right) \left| U_{\theta}\right.\right\}$, hence

$$E\left[B_{j,4}\left(U_{\theta}\right)\left\{m\left(\mathbf{X}\right) - \gamma_{\theta}\left(U_{\theta}\right)\right\}\right] \equiv 0, -3 \leq j \leq N, \theta \in S_c^{d-1}.$$

Applying Assumptions (C2) and (C3), one can differentiate through the expectation, thus

$$E\left\{\frac{\partial}{\partial \theta_p}\left[B_{j,4}\left(U_{\theta}\right)\left\{m\left(\mathbf{X}\right) - \gamma_{\theta}\left(U_{\theta}\right)\right\}\right]\right\} \equiv 0, 1 \leq p \leq d, -3 \leq j \leq N, \theta \in S_c^{d-1},$$

which allows one to apply the Bernstein's inequality to obtain that with probability 1

$$\left\| \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \theta_p}\left[B_{j,4}\left(U_{\theta,i}\right)\left\{m\left(\mathbf{X}_i\right) - \gamma_{\theta}\left(U_{\theta,i}\right)\right\}\right] \right\}_{j=-3}^{N} \right\|_{\infty} = O\left\{(nh)^{-1/2} \log n\right\},$$

which is (4.6.27). □

Lemma 4.6.11. *Under Assumptions (C2) and (C4)-(C6), for $\hat{\varepsilon}_{\theta}(u)$ in (4.6.9), one has*

$$\sup_{\theta \in S_c^{d-1}} \sup_{u \in [0,1]} \left|\hat{\varepsilon}_{\theta}(u)\right| = O\left\{(nh)^{-1/2} \log n\right\}, a.s.. \tag{4.6.28}$$

Proof. Denote $\hat{\mathbf{a}} \equiv \left(\hat{a}_{-3}, \cdots, \hat{a}_N\right)^T = \left(\mathbf{B}_{\theta}^T \mathbf{B}_{\theta}\right)^{-1} \mathbf{B}_{\theta}^T \mathbf{E} = \mathbf{V}_{n,\theta}^{-1}\left(n^{-1} \mathbf{B}_{\theta}^{\mathbf{T}} \mathbf{E}\right)$, then $\hat{\varepsilon}_{\theta}(u) = \sum_{j=-3}^{N} \hat{a}_j B_{j,4}(u)$, so the order of $\hat{\varepsilon}_{\theta}(u)$ is related to that of $\hat{\mathbf{a}}$. In fact, by Theorem 5.4.2 in DeVore and Lorentz (1993)

$$\begin{aligned} \sup_{\theta \in S_c^{d-1}} \sup_{u \in [0,1]} \left|\hat{\varepsilon}_{\theta}(u)\right| &\leq \sup_{\theta \in S_c^{d-1}} \|\hat{\mathbf{a}}\|_{\infty} = \\ \sup_{\theta \in S_c^{d-1}} \left\| \mathbf{V}_{n,\theta}^{-1}\left(n^{-1} \mathbf{B}_{\theta}^{\mathbf{T}} \mathbf{E}\right) \right\|_{\infty} &\leq CN \sup_{\theta \in S_c^{d-1}} \left\| n^{-1} \mathbf{B}_{\theta}^{\mathbf{T}} \mathbf{E} \right\|_{\infty}, a.s., \end{aligned}$$

where the last inequality follows from (4.6.12) of Lemma 4.6.5. Applying (4.6.24) of Lemma 4.6.10, one has established (4.6.28). □

LEMMA 4.6.12. *Under Assumptions (C2) and (C4)-(C6), for $\tilde{\varepsilon}_{\boldsymbol{\theta}}(u)$ in (4.6.8), one has*

$$\sup_{\boldsymbol{\theta}\in S_c^{d-1}} \sup_{u\in[0,1]} \left|\tilde{\varepsilon}_{\boldsymbol{\theta}}(u)\right| = O\left\{(nh)^{-1/2}\log n\right\}, a.s.. \tag{4.6.29}$$

The proof is similar to Lemma 4.6.11, thus omitted. □

The next result evaluates the uniform size of the noise derivatives.

LEMMA 4.6.13. *Under Assumptions (C2)-(C6), one has with probability* 1

$$\sup_{1\le p\le d} \sup_{\boldsymbol{\theta}\in S_c^{d-1}} \max_{1\le i\le n} \left|\frac{\partial}{\partial\theta_p}\hat{\varepsilon}_{\boldsymbol{\theta}}\left(U_{\boldsymbol{\theta},i}\right)\right| = O\left\{(nh^3)^{-1/2}\log n\right\}, \tag{4.6.30}$$

$$\sup_{1\le p\le d} \sup_{\boldsymbol{\theta}\in S_c^{d-1}} \max_{1\le i\le n} \left|\frac{\partial}{\partial\theta_p}\tilde{\varepsilon}_{\boldsymbol{\theta}}\left(U_{\boldsymbol{\theta},i}\right)\right| = O\left\{(nh^3)^{-1/2}\log n\right\}, \tag{4.6.31}$$

$$\sup_{1\le p,q\le d} \sup_{\boldsymbol{\theta}\in S_c^{d-1}} \max_{1\le i\le n} \left|\frac{\partial^2}{\partial\theta_p\partial\theta_q}\hat{\varepsilon}_{\boldsymbol{\theta}}\left(U_{\boldsymbol{\theta},i}\right)\right| = O\left\{(nh^5)^{-1/2}\log n\right\}, \tag{4.6.32}$$

$$\sup_{1\le p,q\le d} \sup_{\boldsymbol{\theta}\in S_c^{d-1}} \max_{1\le i\le n} \left|\frac{\partial^2}{\partial\theta_p\partial\theta_q}\tilde{\varepsilon}_{\boldsymbol{\theta}}\left(U_{\boldsymbol{\theta},i}\right)\right| = O\left\{(nh^5)^{-1/2}\log n\right\}. \tag{4.6.33}$$

PROOF. Note that

$$\left\{\frac{\partial}{\partial\theta_p}\hat{\varepsilon}_{\boldsymbol{\theta}}\left(U_{\boldsymbol{\theta},i}\right)\right\}_{i=1}^n = (\mathbf{I}-\mathbf{P}_{\boldsymbol{\theta}})\dot{\mathbf{B}}_p\left(\mathbf{B}_{\boldsymbol{\theta}}^T\mathbf{B}_{\boldsymbol{\theta}}\right)^{-1}\mathbf{B}_{\boldsymbol{\theta}}^T\mathbf{E} + \mathbf{B}_{\boldsymbol{\theta}}\left(\mathbf{B}_{\boldsymbol{\theta}}^T\mathbf{B}_{\boldsymbol{\theta}}\right)^{-1}\dot{\mathbf{B}}_p^T(\mathbf{I}-\mathbf{P}_{\boldsymbol{\theta}})\mathbf{E}.$$

Applying (4.6.24) and (4.6.25) of Lemma 4.6.10, (4.6.12) of Lemma 4.6.5, (4.6.22) and (4.6.23) of Lemma 4.6.9, one derives (4.6.30). To prove (4.6.31), note that

$$\left\{\frac{\partial}{\partial\theta_p}\tilde{\varepsilon}_{\boldsymbol{\theta}}\left(U_{\boldsymbol{\theta},i}\right)\right\}_{i=1}^n = \frac{\partial}{\partial\theta_p}\left\{\mathbf{P}_{\boldsymbol{\theta}}\mathbf{E}_{\boldsymbol{\theta}}\right\} = \dot{\mathbf{P}}_p\mathbf{E}_{\boldsymbol{\theta}} + \mathbf{P}_{\boldsymbol{\theta}}\frac{\partial}{\partial\theta_p}\mathbf{E}_{\boldsymbol{\theta}} = T_1 + T_2, \tag{4.6.34}$$

in which

$$\begin{aligned} T_1 &= \left\{(\mathbf{I}-\mathbf{P}_{\boldsymbol{\theta}})\dot{\mathbf{B}}_p - \mathbf{B}_{\boldsymbol{\theta}}\left(\mathbf{B}_{\boldsymbol{\theta}}^T\mathbf{B}_{\boldsymbol{\theta}}\right)^{-1}\dot{\mathbf{B}}_p^T\mathbf{B}_{\boldsymbol{\theta}}\right\}\left(\mathbf{B}_{\boldsymbol{\theta}}^T\mathbf{B}_{\boldsymbol{\theta}}\right)^{-1}\mathbf{B}_{\boldsymbol{\theta}}^T\mathbf{E}_{\boldsymbol{\theta}} \\ &= \left\{(\mathbf{I}-\mathbf{P}_{\boldsymbol{\theta}})\dot{\mathbf{B}}_p - \mathbf{B}_{\boldsymbol{\theta}}\left(\frac{\mathbf{B}_{\boldsymbol{\theta}}^T\mathbf{B}_{\boldsymbol{\theta}}}{n}\right)^{-1}\frac{\dot{\mathbf{B}}_p^T\mathbf{B}_{\boldsymbol{\theta}}}{n}\right\}\left(\frac{\mathbf{B}_{\boldsymbol{\theta}}^T\mathbf{B}_{\boldsymbol{\theta}}}{n}\right)^{-1}\frac{\mathbf{B}_{\boldsymbol{\theta}}^T\mathbf{E}_{\boldsymbol{\theta}}}{n}, \end{aligned}$$

$$T_2 = \mathbf{B}_{\boldsymbol{\theta}}\left(\frac{\mathbf{B}_{\boldsymbol{\theta}}^T\mathbf{B}_{\boldsymbol{\theta}}}{n}\right)^{-1}\frac{\partial}{\partial\theta_p}\left(\frac{\mathbf{B}_{\boldsymbol{\theta}}^T\mathbf{E}_{\boldsymbol{\theta}}}{n}\right).$$

By (4.6.24), (4.6.12), (4.6.22) and (4.6.23), one derives

$$\sup_{\theta \in S_c^{d-1}} \|T_1\|_\infty = O\left(n^{-1/2} N^{3/2} \log n\right), a.s., \tag{4.6.35}$$

while (4.6.27) of Lemma 4.6.10, (4.6.12) of Lemma 4.6.5

$$\sup_{\theta \in S_c^{d-1}} \|T_2\|_\infty = N \times O\left(n^{-1/2} h^{-1/2} \log n\right) = O\left(n^{-1/2} h^{-3/2} \log n\right), a.s.. \tag{4.6.36}$$

Now, putting together (4.6.34), (4.6.35) and (4.6.36), one can establish (4.6.31). The proof for (4.6.32) and (4.6.33) are similar. □

Proof of Proposition 4.2.1. According to the decomposition (4.6.6)

$$|\hat{\gamma}_\theta(u) - \gamma_\theta(u)| = |\{\tilde{\gamma}_\theta(u) - \gamma_\theta(u)\} + \tilde{\varepsilon}_\theta(u) + \hat{\varepsilon}_\theta(u)|.$$

Then (4.2.13) follows directly from (4.6.16) of Lemma 4.6.7, (4.6.28) of Lemma 4.6.11 and (4.6.29) of Lemma 4.6.12. Again by definitions (4.6.8) and (4.6.9), we write

$$\frac{\partial}{\partial \theta_p}\{(\hat{\gamma}_\theta - \gamma_\theta)(U_{\theta,i})\} = \frac{\partial}{\partial \theta_p}(\tilde{\gamma}_\theta - \gamma_\theta)(U_{\theta,i}) + \frac{\partial}{\partial \theta_p}\tilde{\gamma}_\theta(U_{\theta,i}) + \frac{\partial}{\partial \theta_p}\hat{\varepsilon}_\theta(U_{\theta,i}).$$

It is clear from (4.6.20), (4.6.30) and (4.6.31) that with probability 1

$$\sup_{1 \le p \le d} \sup_{\theta \in S_c^{d-1}} \max_{1 \le i \le n} \left|\frac{\partial}{\partial \theta_p}(\tilde{\gamma}_\theta - \gamma_\theta)(U_{\theta,i})\right| = O\left(h^3\right),$$

$$\sup_{1 \le p \le d} \sup_{\theta \in S_c^{d-1}} \max_{1 \le i \le n} \left\{\left|\frac{\partial}{\partial \theta_p}\tilde{\varepsilon}_\theta(U_{\theta,i})\right| + \left|\frac{\partial}{\partial \theta_p}\hat{\varepsilon}_\theta(U_{\theta,i})\right|\right\} = O\left\{\left(nh^3\right)^{-1/2} \log n\right\}.$$

Putting together all the above yields (4.2.14). The proof of (4.2.15) is similar. □

4.6.3 Proof of Proposition 4.2.2

Lemma 4.6.14.. *Under Assumptions (C2)-(C6), one has*

$$\sup_{\theta \in S_c^{d-1}} \left|\hat{R}(\theta) - R(\theta)\right| = o(1), a.s..$$

Proof. For the empirical risk function $\hat{R}(\theta)$ in (4.2.9), one has

$$\hat{R}(\theta) = n^{-1} \sum_{i=1}^{n} \{\hat{\gamma}_\theta(U_{\theta,i}) - m(\mathbf{X}_i) - \sigma(\mathbf{X}_i)\varepsilon_i\}^2$$

$$= n^{-1}\sum_{i=1}^{n}\left\{\hat{\gamma}_{\theta}\left(U_{\theta,i}\right) - \gamma_{\theta}\left(U_{\theta,i}\right) + \gamma_{\theta}\left(U_{\theta,i}\right) - m\left(\mathbf{X}_i\right) - \sigma\left(\mathbf{X}_i\right)\varepsilon_i\right\}^2,$$

hence

$$\hat{R}\left(\theta\right) = n^{-1}\sum_{i=1}^{n}\left\{\hat{\gamma}_{\theta}\left(U_{\theta,i}\right) - \gamma_{\theta}\left(U_{\theta,i}\right)\right\}^2 + n^{-1}\sum_{i=1}^{n}\sigma^2\left(\mathbf{X}_i\right)\varepsilon_i^2$$

$$+2n^{-1}\sum_{i=1}^{n}\left\{\hat{\gamma}_{\theta}\left(U_{\theta,i}\right) - \gamma_{\theta}\left(U_{\theta,i}\right)\right\}\left\{\gamma_{\theta}\left(U_{\theta,i}\right) - m\left(\mathbf{X}_i\right) - \sigma\left(\mathbf{X}_i\right)\varepsilon_i\right\}$$

$$+n^{-1}\sum_{i=1}^{n}\left\{\gamma_{\theta}\left(U_{\theta,i}\right) - m\left(\mathbf{X}_i\right)\right\}^2 + 2n^{-1}\sum_{i=1}^{n}\left\{\gamma_{\theta}\left(U_{\theta,i}\right) - m\left(\mathbf{X}_i\right)\right\}\sigma\left(\mathbf{X}_i\right)\varepsilon_i,$$

where $\hat{\gamma}_{\theta}\left(x\right)$ is defined in (4.2.8). Using the expression of $R\left(\theta\right)$ in (4.2.7), one has

$$\sup_{\theta\in S_c^{d-1}}\left|\hat{R}\left(\theta\right) - R\left(\theta\right)\right| \leq I_1 + I_2 + I_3 + I_4,$$

with

$$I_1 = \sup_{\theta\in S_c^{d-1}}\left|n^{-1}\sum_{i=1}^{n}\left\{\hat{\gamma}_{\theta}\left(U_{\theta,i}\right) - \gamma_{\theta}\left(U_{\theta,i}\right)\right\}^2\right|,$$

$$I_2 = \sup_{\theta\in S_c^{d-1}}\left|2n^{-1}\sum_{i=1}^{n}\left\{\hat{\gamma}_{\theta}\left(U_{\theta,i}\right) - \gamma_{\theta}\left(U_{\theta,i}\right)\right\}\left\{\gamma_{\theta}\left(U_{\theta,i}\right) - m\left(\mathbf{X}_i\right) - \sigma\left(\mathbf{X}_i\right)\varepsilon_i\right\}\right|,$$

$$I_3 = \sup_{\theta\in S_c^{d-1}}\left|n^{-1}\sum_{i=1}^{n}\left\{\gamma_{\theta}\left(U_{\theta,i}\right) - m\left(\mathbf{X}_i\right)\right\}^2 - E\left\{\gamma_{\theta}\left(U_{\theta}\right) - m\left(\mathbf{X}\right)\right\}^2\right|,$$

$$I_4 = \sup_{\theta\in S_c^{d-1}}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\sigma^2\left(\mathbf{X}_i\right)\varepsilon_i^2 - E\sigma^2\left(\mathbf{X}\right)\right| + \left|\frac{2}{n}\sum_{i=1}^{n}\left\{\gamma_{\theta}\left(U_{\theta,i}\right) - m\left(\mathbf{X}_i\right)\right\}\sigma\left(\mathbf{X}_i\right)\varepsilon_i\right|\right\}.$$

Bernstein inequality and strong law of large number for α mixing sequence imply that

$$I_3 + I_4 = o(1), a.s.. \tag{4.6.37}$$

Now (4.2.13) of Proposition 4.2.1 provides that

$$\sup_{\theta\in S_c^{d-1}}\sup_{u\in[0,1]}\left|\hat{\gamma}_{\theta}\left(u\right) - \gamma_{\theta}\left(u\right)\right| = O\left(n^{-1/2}h^{-1/2}\log n + h^4\right), a.s.,$$

which entail that

$$I_1 = O\left\{\left(n^{-1/2}h^{-1/2}\log n\right)^2 + \left(h^4\right)^2\right\}, a.s., \tag{4.6.38}$$

$$I_2 \leq O\left\{\left(nh\right)^{-1/2}\log n + h^4\right\} \times \sup_{\theta\in S_c^{d-1}} 2n^{-1}\sum_{i=1}^{n}\left|\gamma_{\theta}\left(U_{\theta,i}\right) - m\left(\mathbf{X}_i\right) - \sigma\left(\mathbf{X}_i\right)\varepsilon_i\right|.$$

Hence

$$I_2 \leq O\left(n^{-1/2}h^{-1/2}\log n + h^4\right), a.s.. \tag{4.6.39}$$

The lemma now follows from (4.6.37), (4.6.38) and (4.6.39) and Assumption (C6). □

LEMMA 4.6.15. *Under Assumptions (C2) - (C6), one has*

$$\sup_{\theta \in S_c^{d-1}} \sup_{1 \leq p \leq d} \left| \frac{\partial}{\partial \theta_p} \left\{ \hat{R}(\theta) - R(\theta) \right\} - n^{-1} \sum_{i=1}^{n} \xi_{\theta,i,p} \right| = o\left(n^{-1/2}\right), \ a.s., \tag{4.6.40}$$

in which

$$\xi_{\theta,i,p} = 2\left\{\gamma_\theta\left(U_{\theta,i}\right) - Y_i\right\} \frac{\partial}{\partial \theta_p} \gamma_\theta\left(U_{\theta,i}\right) - \frac{\partial}{\partial \theta_p} R(\theta), \ E\left(\xi_{\theta,i,p}\right) = 0. \tag{4.6.41}$$

Furthermore for $k = 1, 2$

$$\sup_{\theta \in S_c^{d-1}} \left| \frac{\partial^k}{\partial \theta^k} \left\{ \hat{R}(\theta) - R(\theta) \right\} \right| = O\left(n^{-1/2}h^{-1/2-k}\log n + h^{4-k}\right), a.s.. \tag{4.6.42}$$

PROOF. Note that for any $p = 1, 2, ..., d$

$$\frac{1}{2}\frac{\partial}{\partial \theta_p}\hat{R}(\theta) = n^{-1}\sum_{i=1}^{n}\left\{\hat{\gamma}_\theta\left(U_{\theta,i}\right) - Y_i\right\}\frac{\partial}{\partial \theta_p}\hat{\gamma}_\theta\left(U_{\theta,i}\right),$$

$$\begin{aligned}\frac{1}{2}\frac{\partial}{\partial \theta_p}R(\theta) &= E\left[\left\{\gamma_\theta\left(U_\theta\right) - m(\mathbf{X})\right\}\frac{\partial}{\partial \theta_p}\gamma_\theta\left(U_\theta\right)\right] \\ &= E\left[\left\{\gamma_\theta\left(U_\theta\right) - m(\mathbf{X}) - \sigma(\mathbf{X})\varepsilon\right\}\frac{\partial}{\partial \theta_p}\gamma_\theta\left(U_\theta\right)\right].\end{aligned}$$

Thus $E\left(\xi_{\theta,i,p}\right) = 2E\left[\left\{\gamma_\theta\left(U_{\theta,i}\right) - Y_i\right\}\frac{\partial}{\partial \theta_p}\gamma_\theta\left(U_{\theta,i}\right)\right] - \frac{\partial}{\partial \theta_p}R(\theta) = 0$ and

$$\frac{1}{2}\frac{\partial}{\partial \theta_p}\left\{\hat{R}(\theta) - R(\theta)\right\} = (2n)^{-1}\sum_{i=1}^{n}\xi_{\theta,i,p} + J_{1,\theta,p} + J_{2,\theta,p} + J_{3,\theta,p}, \tag{4.6.43}$$

with

$$\begin{aligned}J_{1,\theta,p} &= n^{-1}\sum_{i=1}^{n}\left\{\hat{\gamma}_\theta\left(U_{\theta,i}\right) - \gamma_\theta\left(U_{\theta,i}\right)\right\}\frac{\partial}{\partial \theta_p}\left(\hat{\gamma}_\theta - \gamma_\theta\right)\left(U_{\theta,i}\right), \\ J_{2,\theta,p} &= n^{-1}\sum_{i=1}^{n}\left\{\gamma_\theta\left(U_{\theta,i}\right) - m(\mathbf{X}_i) - \sigma(\mathbf{X}_i)\varepsilon_i\right\}\frac{\partial}{\partial \theta_p}\left(\hat{\gamma}_\theta - \gamma_\theta\right)\left(U_{\theta,i}\right), \\ J_{3,\theta,p} &= n^{-1}\sum_{i=1}^{n}\left\{\hat{\gamma}_\theta\left(U_{\theta,i}\right) - \gamma_\theta\left(U_{\theta,i}\right)\right\}\frac{\partial}{\partial \theta_p}\gamma_\theta\left(U_{\theta,i}\right).\end{aligned}$$

Bernstein inequality implies that

$$\sup_{\theta\in S_c^{d-1}}\sup_{1\le p\le d}\left|n^{-1}\sum_{i=1}^{n}\xi_{\theta,i,p}\right| = O\left(n^{-1/2}\log n\right), a.s.. \tag{4.6.44}$$

Meanwhile, applying (4.2.13) and (4.2.14) of Proposition 4.2.1, one obtains that

$$\begin{aligned}\sup_{\theta\in S_c^{d-1}}\sup_{1\le p\le d}\left|J_{1,\theta,p}\right| &= O\left\{(nh)^{-1/2}\log n + h^4\right\}\times O\left\{\left(nh^3\right)^{-1/2}\log n + h^3\right\}\\ &= O\left(n^{-1}h^{-2}\log^2 n + h^7\right), a.s.. \end{aligned}\tag{4.6.45}$$

Note that

$$\begin{aligned} J_{2,\theta,p} &= n^{-1}\sum_{i=1}^{n}\left\{\gamma_\theta\left(U_{\theta,i}\right) - m\left(\mathbf{X}_i\right) - \sigma\left(\mathbf{X}_i\right)\varepsilon_i\right\}\frac{\partial}{\partial\theta_p}\left(\tilde{\gamma}_\theta - \gamma_\theta\right)\left(U_{\theta,i}\right)\\ &\quad -n^{-1}\left(\mathbf{E}+\mathbf{E}_\theta\right)^T\frac{\partial}{\partial\theta_p}\left\{\mathbf{P}_\theta\left(\mathbf{E}+\mathbf{E}_\theta\right)\right\}.\end{aligned}$$

Applying (4.2.13), one gets

$$\sup_{\theta\in S_c^{d-1}}\sup_{1\le p\le d}\left|J_{2,\theta,p} + n^{-1}\left(\mathbf{E}+\mathbf{E}_\theta\right)^T\frac{\partial}{\partial\theta_p}\left\{\mathbf{P}_\theta\left(\mathbf{E}+\mathbf{E}_\theta\right)\right\}\right| = O\left(h^3\right), a.s.,$$

while (4.6.24), (4.6.26) and (4.6.12) entail that with probability 1

$$\sup_{\theta\in S_c^{d-1}}\sup_{1\le p\le d}\left|n^{-1}\left(\mathbf{E}+\mathbf{E}_\theta\right)^T\frac{\partial}{\partial\theta_p}\left\{\mathbf{P}_\theta\left(\mathbf{E}+\mathbf{E}_\theta\right)\right\}\right|$$

$$= O\left\{(nN)^{-1/2}\log n\right\}\times N\times N\times O\left\{(nN)^{-1/2}\log n\right\} = O\left\{n^{-1}N\log^2 n\right\},$$

thus

$$\sup_{\theta\in S_c^{d-1}}\sup_{1\le p\le d}\left|J_{2,\theta,p}\right| = O\left(h^3 + n^{-1}N\log^2 n\right), a.s.. \tag{4.6.46}$$

Lastly

$$J_{3,\theta,p} - n^{-1}\sum_{i=1}^{n}\left(\tilde{\gamma}_\theta - \gamma_\theta\right)\frac{\partial}{\partial\theta_p}\gamma_\theta\left(U_{\theta,i}\right)$$

$$= n^{-1}\left(\mathbf{E}+\mathbf{E}_\theta\right)^T\mathbf{B}_\theta\left(\frac{\mathbf{B}_\theta^T\mathbf{B}_\theta}{n}\right)^{-1}\frac{\mathbf{B}_\theta^T}{n}\frac{\partial}{\partial\theta_p}\gamma_\theta.$$

By applying (4.6.24), (4.6.26), and (4.6.12), it is clear that with probability 1

$$\begin{aligned}&\sup_{\theta\in S_c^{d-1}}\sup_{1\le p\le d}\left|\left(n^{-1}\mathbf{B}_\theta^T\mathbf{E}+n^{-1}\mathbf{B}_\theta^{\mathbf{T}}\mathbf{E}_\theta\right)^T\left(\frac{\mathbf{B}_\theta^T\mathbf{B}_\theta}{n}\right)^{-1}\frac{\mathbf{B}_\theta^T}{n}\frac{\partial}{\partial\theta_p}\gamma_\theta\right|\\ &= O\left\{(nN)^{-1/2}\log n\right\}\times N\times O\left\{h + (nN)^{-1/2}\log n\right\}\\ &= O\left\{n^{-1}\log^2 n + (nN)^{-1/2}\log n\right\},\end{aligned}$$

while by applying (4.6.16) of Lemma 4.6.7, one has

$$\sup_{\boldsymbol{\theta}\in S_c^{d-1}} \sup_{1\le p\le d} \left| n^{-1} \sum_{i=1}^{n} \left(\tilde{\gamma}_{\boldsymbol{\theta}} - \gamma_{\boldsymbol{\theta}}\right) \frac{\partial}{\partial \theta_p} \gamma_{\boldsymbol{\theta}} \left(U_{\boldsymbol{\theta},i}\right) \right| = O\left(h^4\right), a.s.,$$

together, the above entail that

$$\sup_{\boldsymbol{\theta}\in S_c^{d-1}} \sup_{1\le p\le d} \left|J_{3,\boldsymbol{\theta},p}\right| = O\left\{h^4 + n^{-1}\log^2 n + (nN)^{-1/2}\log n\right\}, a.s.. \qquad (4.6.47)$$

Therefore, (4.6.43), (4.6.45), (4.6.46), (4.6.47) and Assumption A6 lead to (4.6.40), which, together with (4.6.44), establish (4.6.42) for $k = 1$.

Note that the second order derivative of $\hat{R}(\boldsymbol{\theta})$ and $R(\boldsymbol{\theta})$ with respect to θ_p, θ_q are

$$2n^{-1}\left[\sum_{i=1}^{n} \left\{\hat{\gamma}_{\boldsymbol{\theta}}\left(U_{\boldsymbol{\theta},i}\right) - Y_i\right\} \frac{\partial^2}{\partial\theta_p\partial\theta_q}\hat{\gamma}_{\boldsymbol{\theta}}\left(U_{\boldsymbol{\theta},i}\right) + \sum_{i=1}^{n} \frac{\partial}{\partial\theta_q}\hat{\gamma}_{\boldsymbol{\theta}}\left(U_{\boldsymbol{\theta},i}\right) \frac{\partial}{\partial\theta_p}\hat{\gamma}_{\boldsymbol{\theta}}\left(U_{\boldsymbol{\theta},i}\right)\right],$$

$$2\left[E\left\{\gamma_{\boldsymbol{\theta}}\left(U_{\boldsymbol{\theta}}\right) - m\left(\mathbf{X}\right)\right\} \frac{\partial^2}{\partial\theta_p\partial\theta_q}\gamma_{\boldsymbol{\theta}}\left(U_{\boldsymbol{\theta}}\right) + E\left\{\frac{\partial}{\partial\theta_q}\gamma_{\boldsymbol{\theta}}\left(U_{\boldsymbol{\theta}}\right)\frac{\partial}{\partial\theta_p}\gamma_{\boldsymbol{\theta}}\left(U_{\boldsymbol{\theta}}\right)\right\}\right].$$

The proof of (4.6.42) for $k = 2$ follows from (4.2.13), (4.2.14) and (4.2.15). □

PROOF OF PROPOSITION 4.2.2. The result follows from Lemma 4.6.14, Lemma 4.6.15, equations (4.6.50) and (4.6.51). □

4.6.4 Proof of the Theorem 4.2.2

Let $\hat{S}_p^*\left(\boldsymbol{\theta}_{-d}\right)$ be the p-th element of $\hat{S}^*\left(\boldsymbol{\theta}_{-d}\right)$ and for $\gamma_{\boldsymbol{\theta}}$ in (4.2.6), denote

$$\eta_{i,p} := 2\left\{\dot{\gamma}_p - \theta_{0,p}\theta_{0,d}^{-1}\dot{\gamma}_d\right\}\left(U_{\boldsymbol{\theta}_0,i}\right)\left\{\gamma_{\boldsymbol{\theta}_0}\left(U_{\boldsymbol{\theta}_0,i}\right) - Y_i\right\}, \qquad (4.6.48)$$

where $\dot{\gamma}_p$ is value of $\frac{\partial}{\partial\theta_p}\gamma_{\boldsymbol{\theta}}$ taking at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, for any $p, q = 1, 2, ..., d-1$.

LEMMA 4.6.16. *Under Assumptions (C2)-(C6), one has*

$$\sup_{1\le p\le d-1}\left|\hat{S}_p^*\left(\boldsymbol{\theta}_{0,-d}\right) - n^{-1}\sum_{i=1}^{n}\eta_{i,p}\right| = o\left(n^{-1/2}\right), a.s.. \qquad (4.6.49)$$

PROOF. For any $p = 1, ..., d-1$

$$\hat{S}_p^*\left(\boldsymbol{\theta}_{-d}\right) - S_p^*\left(\boldsymbol{\theta}_{-d}\right) = \left(\frac{\partial}{\partial\theta_p} - \theta_p\theta_d^{-1}\frac{\partial}{\partial\theta_d}\right)\left\{\hat{R}\left(\boldsymbol{\theta}\right) - R\left(\boldsymbol{\theta}\right)\right\}.$$

Therefore, according to (4.6.40), (4.6.41) and (4.6.48)

$$\eta_{i,p} = n^{-1}\sum_{i=1}^{n} \xi_{\boldsymbol{\theta}_0,i,p} - \theta_{0,p}\theta_{0,d}^{-1} n^{-1}\sum_{i=1}^{n} \xi_{\boldsymbol{\theta}_0,i,d}, \; E\left(\eta_{i,p}\right) = 0,$$

$$\sup_{1\leq p\leq d-1}\left|\hat{S}_p^*\left(\boldsymbol{\theta}_{0,-d}\right) - S_p^*\left(\boldsymbol{\theta}_{0,-d}\right) - n^{-1}\sum_{i=1}^{n}\eta_{i,p}\right| = o\left(n^{-1/2}\right), a.s..$$

Since $S^*\left(\boldsymbol{\theta}_{-d}\right)$ attains its minimum at $\boldsymbol{\theta}_{0,-d}$, for $p = 1, ..., d-1$

$$S_p^*\left(\boldsymbol{\theta}_{0,-d}\right) \equiv \left(\frac{\partial}{\partial\theta_p} - \theta_p\theta_d^{-1}\frac{\partial}{\partial\theta_d}\right) R(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \equiv 0,$$

which yields (4.6.49). □

LEMMA 4.6.17. *The (p,q)-th entry of the Hessian matrix $H^*\left(\boldsymbol{\theta}_{0,-d}\right)$ equals $l_{p,q}$ given in Theorem 4.2.2.*

PROOF. It is easy to show that for any $p, q = 1, 2, ..., d$,

$$\frac{\partial}{\partial\theta_p}R(\boldsymbol{\theta}) = \frac{\partial}{\partial\theta_p}E\left\{m(\mathbf{X}) - \gamma_{\boldsymbol{\theta}}\left(U_{\boldsymbol{\theta}}\right)\right\}^2 = -2E\left[\gamma_{\boldsymbol{\theta}}\left(U_{\boldsymbol{\theta}}\right)\frac{\partial}{\partial\theta_p}\gamma_{\boldsymbol{\theta}}\left(U_{\boldsymbol{\theta}}\right)\right],$$

$$\frac{\partial^2}{\partial\theta_p\partial\theta_q}R(\boldsymbol{\theta}) = -2E\left[\frac{\partial}{\partial\theta_p}\gamma_{\boldsymbol{\theta}}\left(U_{\boldsymbol{\theta}}\right)\frac{\partial}{\partial\theta_q}\gamma_{\boldsymbol{\theta}}\left(U_{\boldsymbol{\theta}}\right) + \gamma_{\boldsymbol{\theta}}\left(U_{\boldsymbol{\theta}}\right)\frac{\partial^2}{\partial\theta_p\partial\theta_q}\gamma_{\boldsymbol{\theta}}\left(U_{\boldsymbol{\theta}}\right)\right].$$

Note that

$$\frac{\partial}{\partial\theta_p}R^*\left(\boldsymbol{\theta}_{-d}\right) = \frac{\partial}{\partial\theta_p}R(\boldsymbol{\theta}) - \frac{\theta_p}{\theta_d}\frac{\partial}{\partial\theta_d}R(\boldsymbol{\theta}), \tag{4.6.50}$$

$$\begin{aligned}\frac{\partial^2}{\partial\theta_p\partial\theta_q}R^*\left(\boldsymbol{\theta}_{-d}\right) &= \frac{\partial^2}{\partial\theta_p\partial\theta_q}R(\boldsymbol{\theta}) - \frac{\theta_q}{\theta_d}\frac{\partial^2}{\partial\theta_p\partial\theta_d}R(\boldsymbol{\theta}) - \frac{\theta_p}{\theta_d}\frac{\partial^2}{\partial\theta_d\partial\theta_q}R(\boldsymbol{\theta}) \\ &-\frac{\partial}{\partial\theta_q}\left(\frac{\theta_p}{\sqrt{1-\|\boldsymbol{\theta}_{-d}\|_2^2}}\right)\frac{\partial}{\partial\theta_d}R(\boldsymbol{\theta}) + \frac{\theta_p\theta_q}{\theta_d^2}\frac{\partial^2}{\partial\theta_d\partial\theta_d}R(\boldsymbol{\theta}).\end{aligned} \tag{4.6.51}$$

Thus

$$\frac{\partial}{\partial\theta_p}R^*\left(\boldsymbol{\theta}_{-d}\right) = -2E\left[\gamma_{\boldsymbol{\theta}}\left(U_{\boldsymbol{\theta}}\right)\frac{\partial}{\partial\theta_p}\gamma_{\boldsymbol{\theta}}\left(U_{\boldsymbol{\theta}}\right)\right] + 2\theta_d^{-1}\theta_p E\left[\gamma_{\boldsymbol{\theta}}\left(U_{\boldsymbol{\theta}}\right)\frac{\partial}{\partial\theta_d}\gamma_{\boldsymbol{\theta}}\left(U_{\boldsymbol{\theta}}\right)\right],$$

$$\frac{\partial^2}{\partial\theta_p\partial\theta_q}R^*\left(\boldsymbol{\theta}_{-d}\right) = -2E\left\{\frac{\partial}{\partial\theta_p}\gamma_{\boldsymbol{\theta}}\left(U_{\boldsymbol{\theta}}\right)\frac{\partial}{\partial\theta_q}\gamma_{\boldsymbol{\theta}}\left(U_{\boldsymbol{\theta}}\right) + \gamma_{\boldsymbol{\theta}}\left(U_{\boldsymbol{\theta}}\right)\frac{\partial^2}{\partial\theta_p\partial\theta_q}\gamma_{\boldsymbol{\theta}}\left(U_{\boldsymbol{\theta}}\right)\right\}$$

$$
\begin{aligned}
&+2\theta_q\theta_d^{-1}E\left\{\frac{\partial}{\partial\theta_d}\gamma_{\boldsymbol{\theta}}\left(U_{\boldsymbol{\theta}}\right)\frac{\partial}{\partial\theta_p}\gamma_{\boldsymbol{\theta}}\left(U_{\boldsymbol{\theta}}\right)+\gamma_{\boldsymbol{\theta}}\left(U_{\boldsymbol{\theta}}\right)\frac{\partial^2}{\partial\theta_p\partial\theta_d}\gamma_{\boldsymbol{\theta}}\left(U_{\boldsymbol{\theta}}\right)\right\}\\
&+2\frac{\partial}{\partial\theta_q}\left(\frac{\theta_p}{\sqrt{1-\|\boldsymbol{\theta}_{-d}\|_2^2}}\right)E\left\{\gamma_{\boldsymbol{\theta}}\left(U_{\boldsymbol{\theta}}\right)\frac{\partial}{\partial\theta_d}\gamma_{\boldsymbol{\theta}}\left(U_{\boldsymbol{\theta}}\right)\right\}\\
&+2\theta_p\theta_d^{-1}E\left\{\frac{\partial}{\partial\theta_p}\gamma_{\boldsymbol{\theta}}\left(U_{\boldsymbol{\theta}}\right)\frac{\partial}{\partial\theta_q}\gamma_{\boldsymbol{\theta}}\left(U_{\boldsymbol{\theta}}\right)+\gamma_{\boldsymbol{\theta}}\left(U_{\boldsymbol{\theta}}\right)\frac{\partial^2}{\partial\theta_p\partial\theta_q}\gamma_{\boldsymbol{\theta}}\left(U_{\boldsymbol{\theta}}\right)\right\}\\
&-2\theta_p\theta_q\theta_d^{-2}E\left[\left\{\frac{\partial}{\partial\theta_d}\gamma_{\boldsymbol{\theta}}\left(U_{\boldsymbol{\theta}}\right)\right\}^2+\gamma_{\boldsymbol{\theta}}\left(U_{\boldsymbol{\theta}}\right)\frac{\partial^2}{\partial\theta_d\partial\theta_d}\gamma_{\boldsymbol{\theta}}\left(U_{\boldsymbol{\theta}}\right)\right].
\end{aligned}
$$

Therefore we obtained the desired result. □

PROOF OF THEOREM 4.2.2. For any $p = 1, 2, ..., d-1$, let

$$f_p(t) = \hat{S}_p^*\left(t\hat{\boldsymbol{\theta}}_{-d} + (1-t)\boldsymbol{\theta}_{0,-d}\right), t \in [0,1],$$

then

$$\frac{d}{dt}f_p(t) = \sum_{q=1}^{d-1}\frac{\partial}{\partial\theta_q}\hat{S}_p^*\left(t\hat{\boldsymbol{\theta}}_{-d}+(1-t)\boldsymbol{\theta}_{0,-d}\right)\left(\hat{\theta}_q-\theta_{0,q}\right).$$

Note that $\hat{S}^*(\boldsymbol{\theta}_{-d})$ attains its minimum at $\hat{\boldsymbol{\theta}}_{-d}$, i.e., $\hat{S}_p^*\left(\hat{\boldsymbol{\theta}}_{-d}\right) \equiv 0$. Thus, for any $p = 1, 2, ..., d-1$, $t_p \in [0,1]$, one has

$$
\begin{aligned}
-\hat{S}_p^*\left(\boldsymbol{\theta}_{0,-d}\right) &= \hat{S}_p^*\left(\hat{\boldsymbol{\theta}}_{-d}\right) - \hat{S}_p^*\left(\boldsymbol{\theta}_{0,-d}\right) = f_p(1) - f_p(0)\\
&= \left\{\frac{\partial^2}{\partial\theta_q\theta_p}\hat{R}^*\left(t_p\hat{\boldsymbol{\theta}}_{-d}+(1-t_p)\boldsymbol{\theta}_{0,-d}\right)\right\}^T_{q=1,...,d-1}\left(\hat{\boldsymbol{\theta}}_{-d}-\boldsymbol{\theta}_{0,-d}\right),
\end{aligned}
$$

then

$$-\hat{S}^*\left(\boldsymbol{\theta}_{0,-d}\right) = \left\{\frac{\partial^2}{\partial\theta_q\partial\theta_p}\hat{R}^*\left(t_p\hat{\boldsymbol{\theta}}_{-d}+(1-t_p)\boldsymbol{\theta}_{0,-d}\right)\right\}_{p,q=1,...,d-1}\left(\hat{\boldsymbol{\theta}}_{-d}-\boldsymbol{\theta}_{0,-d}\right).$$

Now (4.2.11) of Theorem 4.2.1 and Proposition 4.2.2 with $k = 2$ imply that uniformly in $p, q = 1, 2, ..., d-1$

$$\frac{\partial^2}{\partial\theta_q\partial\theta_p}\hat{R}^*\left(t_p\hat{\boldsymbol{\theta}}_{-d}+(1-t_p)\boldsymbol{\theta}_{0,-d}\right) \longrightarrow l_{q,p}, a.s., \tag{4.6.52}$$

where $l_{p,q}$ is given in Theorem 4.2.2. Noting that $\sqrt{n}\left(\hat{\boldsymbol{\theta}}_{-d}-\boldsymbol{\theta}_{0,-d}\right)$ is represented as

$$-\left[\left\{\frac{\partial^2}{\partial\theta_q\partial\theta_p}\hat{R}^*\left(t_p\hat{\boldsymbol{\theta}}_{-d}+(1-t_p)\boldsymbol{\theta}_{0,-d}\right)\right\}_{p,q=1,...,d-1}\right]^{-1}\sqrt{n}\hat{S}^*\left(\boldsymbol{\theta}_{0,-d}\right),$$

where $\hat{S}^*\left(\boldsymbol{\theta}_{0,-d}\right)=\left\{\hat{S}_p^*\left(\boldsymbol{\theta}_{0,-d}\right)\right\}_{p=1}^{d-1}$ and according to (4.6.48) and Lemma 4.6.16

$$\hat{S}_p^*\left(\boldsymbol{\theta}_{0,-d}\right)=n^{-1}\sum_{i=1}^{n}\eta_{p,i}+o\left(n^{-1/2}\right), a.s., \ E\left(\eta_{p,i}\right)=0.$$

Let $\Psi\left(\boldsymbol{\theta}_0\right)=\left(\psi_{pq}\right)_{p,q=1}^{d-1}$ be the covariance matrix of $\sqrt{n}\left\{\hat{S}_p^*\left(\boldsymbol{\theta}_{0,-d}\right)\right\}_{p=1}^{d-1}$ with ψ_{pq} given in Theorem 4.2.2. Cramér-Wold device and central limit theorem for α mixing sequences entail that

$$\sqrt{n}\hat{S}^*\left(\boldsymbol{\theta}_{0,-d}\right)\xrightarrow{d}N\left\{\mathbf{0},\Psi\left(\boldsymbol{\theta}_0\right)\right\}.$$

Let $\Sigma\left(\boldsymbol{\theta}_0\right)=\left\{H^*\left(\boldsymbol{\theta}_{0,-d}\right)\right\}^{-1}\Psi\left(\boldsymbol{\theta}_0\right)\left[\left\{H^*\left(\boldsymbol{\theta}_{0,-d}\right)\right\}^T\right]^{-1}$, with $H^*\left(\boldsymbol{\theta}_{0,-d}\right)$ being the Hessian matrix defined in (4.2.3). The above limiting distribution of $\sqrt{n}\hat{S}^*\left(\boldsymbol{\theta}_{0,-d}\right)$, (4.6.52) and Slutsky's theorem imply that

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_{-d}-\boldsymbol{\theta}_{0,-d}\right)\xrightarrow{d}N\left\{\mathbf{0},\Sigma\left(\boldsymbol{\theta}_0\right)\right\}. \qquad \square$$

Table 4.1. Example 2.5.1: Piecewise constant spline bands coverage probabilities

noise level	sample size	confidence level	estimated bands	oracle bands
0.2	100	1 − 0.01 1 − 0.05	0.588 (0.588) 0.320 (0.288)	0.590 (0.582) 0.278 (0.276)
	200	1 − 0.01 1 − 0.05	0.660 (0.716) 0.410 (0.428)	0.772 (0.766) 0.522 (0.512)
	500	1 − 0.01 1 − 0.05	0.858 (0.856) 0.548 (0.556)	0.858 (0.856) 0.564 (0.554)
0.5	100	1 − 0.01 1 − 0.05	0.704 (0.792) 0.482 (0.542)	0.870 (0.864) 0.682 (0.666)
	200	1 − 0.01 1 − 0.05	0.762 (0.812) 0.568 (0.570)	0.880 (0.876) 0.690 (0.676)
	500	1 − 0.01 1 − 0.05	0.922 (0.924) 0.732 (0.744)	0.930 (0.926) 0.782 (0.776)

Table 4.2. Example 2.5.1: Piecewise linear spline bands coverage probabilities

noise level	sample size	confidence level 0.99	confidence level 0.95
0.2	100	0.980 (0.990)	0.948 (0.962)
	200	0.994 (0.996)	0.956 (0.978)
	500	0.994 (1.000)	0.950 (1.000)
0.5	100	0.984 (0.992)	0.956 (0.974)
	200	0.994 (1.000)	0.972 (0.988)
	500	0.996 (1.000)	0.978 (1.000)

Table 4.3. Report of Example 3.6.1

σ_0	n	c	component #1		component #2		component #3	
			1st stage	2nd stage	1st stage	2nd stage	1st stage	2nd stage
0.5	100	0.5	0.1231	0.0461	0.1476	0.0645	0.1254	0.0681
		1.0	0.1278	0.0520	0.1404	0.0690	0.1318	0.0726
	200	0.5	0.0539	0.0125	0.0616	0.0275	0.0577	0.0252
		1.0	0.0841	0.0144	0.0839	0.0290	0.0848	0.0285
	500	0.5	0.0263	0.0031	0.0306	0.0107	0.0278	0.0102
		1.0	0.0595	0.0044	0.0578	0.0115	0.0605	0.0119
	1000	0.5	0.0169	0.0015	0.0210	0.0053	0.0178	0.0054
		1.0	0.0364	0.0018	0.0367	0.0054	0.0375	0.0059
1.0	100	0.5	0.3008	0.0587	0.3298	0.1427	0.3236	0.1393
		1.0	0.3088	0.0586	0.3369	0.1364	0.3062	0.1316
	200	0.5	0.1742	0.0256	0.1783	0.0802	0.1892	0.0701
		1.0	0.2899	0.0328	0.2830	0.0824	0.3043	0.0721
	500	0.5	0.0924	0.0065	0.1124	0.0421	0.1004	0.0345
		1.0	0.2299	0.0078	0.2305	0.0458	0.2314	0.0362
	1000	0.5	0.0616	0.0033	0.0637	0.0270	0.0646	0.0224
		1.0	0.1460	0.0034	0.1433	0.0275	0.1429	0.0219

Table 4.4. The computing time of Example 3.6.1

Method	**n = 100**	**n = 200**	**n = 400**	**n = 1000**
MIE	10	76	628	10728
SPBK	0.7	0.9	1.2	4.5

Table 4.5. Report of Example 4.4.1

σ_0	n	θ_0	BIAS	SD	MSE	Average MSE
0.3	100	$\theta_{0,1}$	5e − 04 (−0.00236)	0.00825 (0.02093)	7e − 05 (0.00044)	7e − 05 (0.00043)
		$\theta_{0,2}$	−6e − 04 (0.00174)	0.00826 (0.02083)	7e − 05 (0.00043)	
	300	$\theta_{0,1}$	−0.00124 (−0.00129)	0.00383 (0.01172)	2e − 05 (0.00014)	2e − 05 (0.00014)
		$\theta_{0,2}$	−0.00124 (0.00110)	0.00383 (0.01160)	2e − 05 (0.00013)	
0.5	100	$\theta_{0,1}$	0.00121 (−0.00137)	0.01346 (0.02257)	0.00018 (0.00051)	0.00018 (0.00051)
		$\theta_{0,2}$	−0.00147 (0.00062)	0.01349 (0.02309)	0.00018 (0.00052)	
	300	$\theta_{0,1}$	−0.00204 (−0.00229)	0.00639 (0.01205)	4e − 05 (0.00015)	4e − 05 (0.00015)
		$\theta_{0,2}$	0.00197 (0.00208)	0.00637 (0.01190)	4e − 05 (0.00014)	

Table 4.6. Report of Example 4.4.2

Sample Size n	Dimension d	Average MSE		Time	
		MAVE	SIP	MAVE	SIP
50	4	0.00020	0.00018	1.91	0.19
	10	0.00031	0.00043	2.17	0.10
	50	0.00031	0.00043	3.29	0.10
	100	0.00681	0.00620	5.94	0.31
	200	0.00529	0.00407	27.90	0.49
100	4	0.00008	0.00008	3.28	0.09
	10	0.00012	0.00017	3.93	0.13
	50	0.00032	0.00127	8.48	0.16
	100	—	0.00395	—	0.44
	200	—	0.00324	—	0.73
200	4	0.00004	0.00003	5.32	0.17
	10	0.00005	0.00007	7.49	0.24
	50	0.00007	0.00030	15.42	0.24
	100	0.00015	0.00061	40.81	0.54
	200	—	0.00197	—	1.44
500	4	0.00002	0.00001	14.44	0.76
	10	0.00002	0.00003	24.54	0.79
	50	0.00002	0.00010	52.93	0.89
	100	0.00003	0.00012	143.07	0.99
	200	0.00004	0.00020	386.80	1.96
	400	—	0.00054	—	4.98
1000	4	0.00001	0.00001	33.57	1.95
	10	0.00001	0.00001	62.54	3.64
	50	0.00001	0.00003	155.38	2.72
	100	0.00001	0.00005	275.73	1.81
	200	0.00008	0.00006	2432.56	2.84
	400	—	0.00010	—	9.35

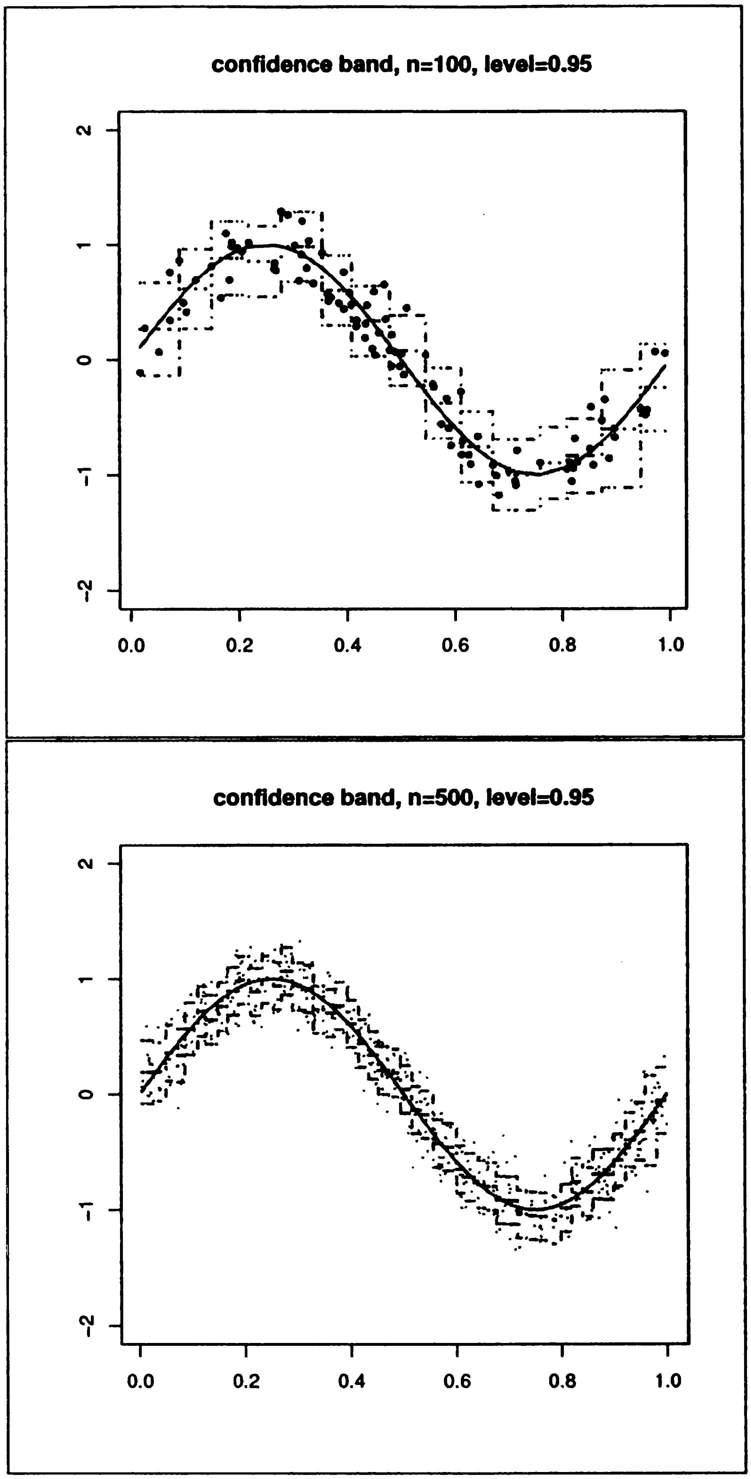


Figure 4.1. Example 2.5.1: 95% constant spline confidence bands with opt = 1

Note: confidence bands (upper and lower dashed curves) computed from (2.4.5) with $k = 1$, opt $= 1$, $\hat{m}_1(x)$ (center dotted curve), $m(x) = \sin(2\pi x)$ (center smooth solid curve).

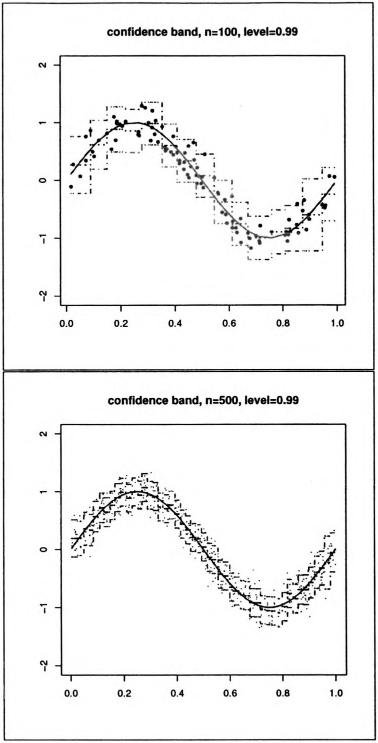


Figure 4.2. Example 2.5.1: 99% constant spline confidence bands with opt = 1
Note: confidence bands (upper and lower dashed curves) computed from (2.4.5) with
$k = 1$, opt $= 1$, $\hat{m}_1(x)$ (center dotted curve), $m(x) = \sin(2\pi x)$ (center smooth solid curve).

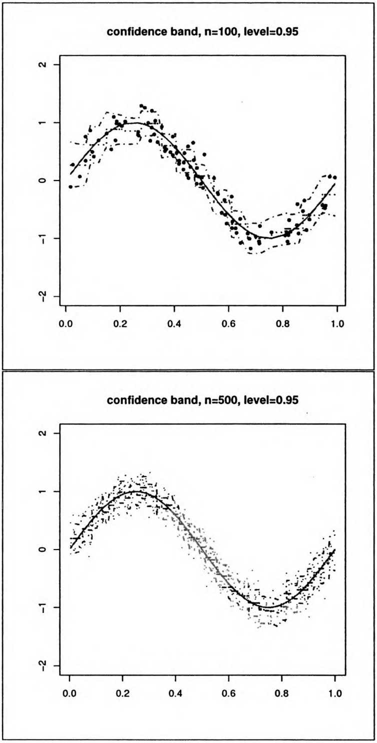


Figure 4.3. Example 2.5.1: 95% constant spline confidence bands with opt $= 2$

Note: confidence bands (upper and lower dashed curves) computed from (2.4.5) with $k = 1$, opt $= 2$, $\hat{m}_1(x)$ (center dotted curve), $m(x) = \sin(2\pi x)$ (center smooth solid curve).

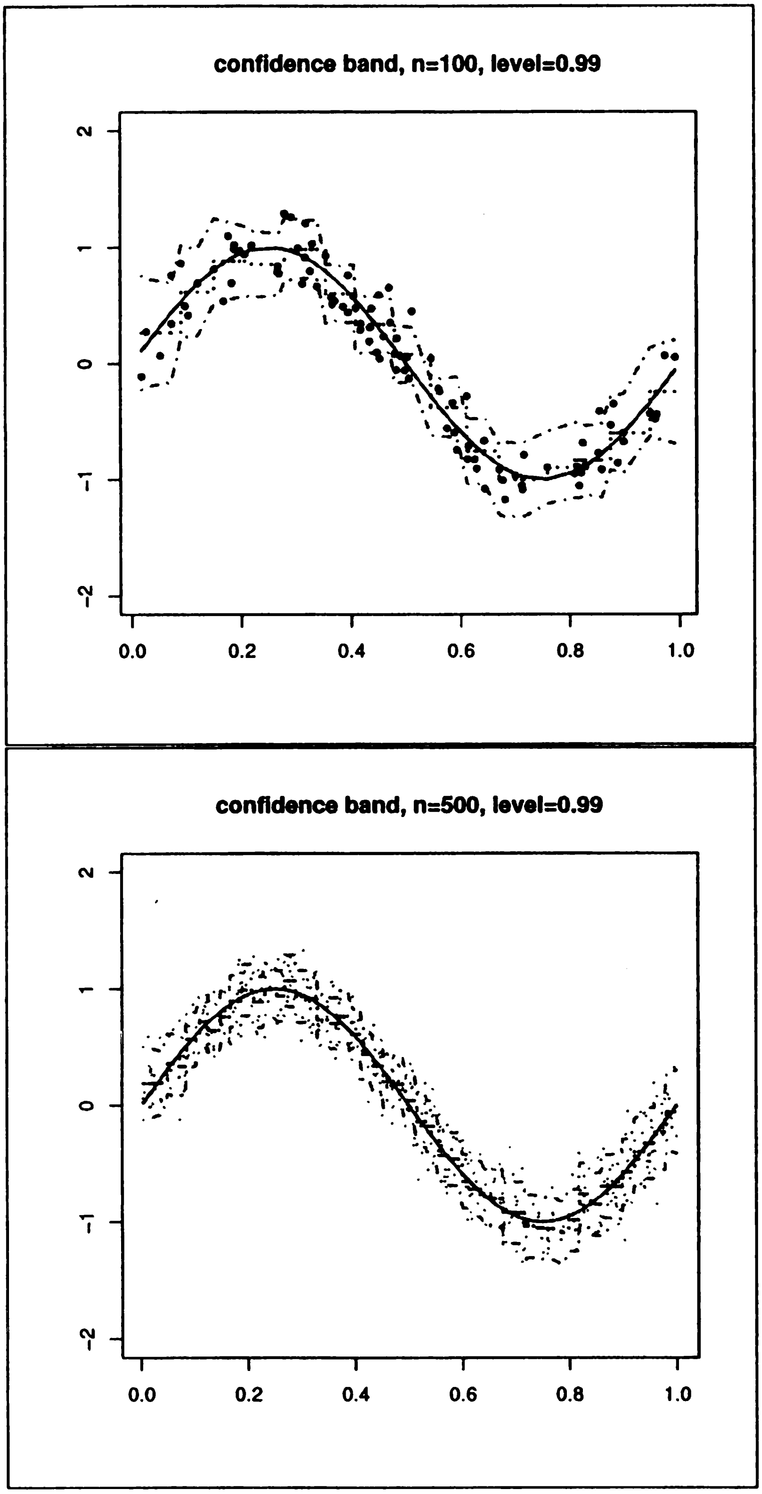
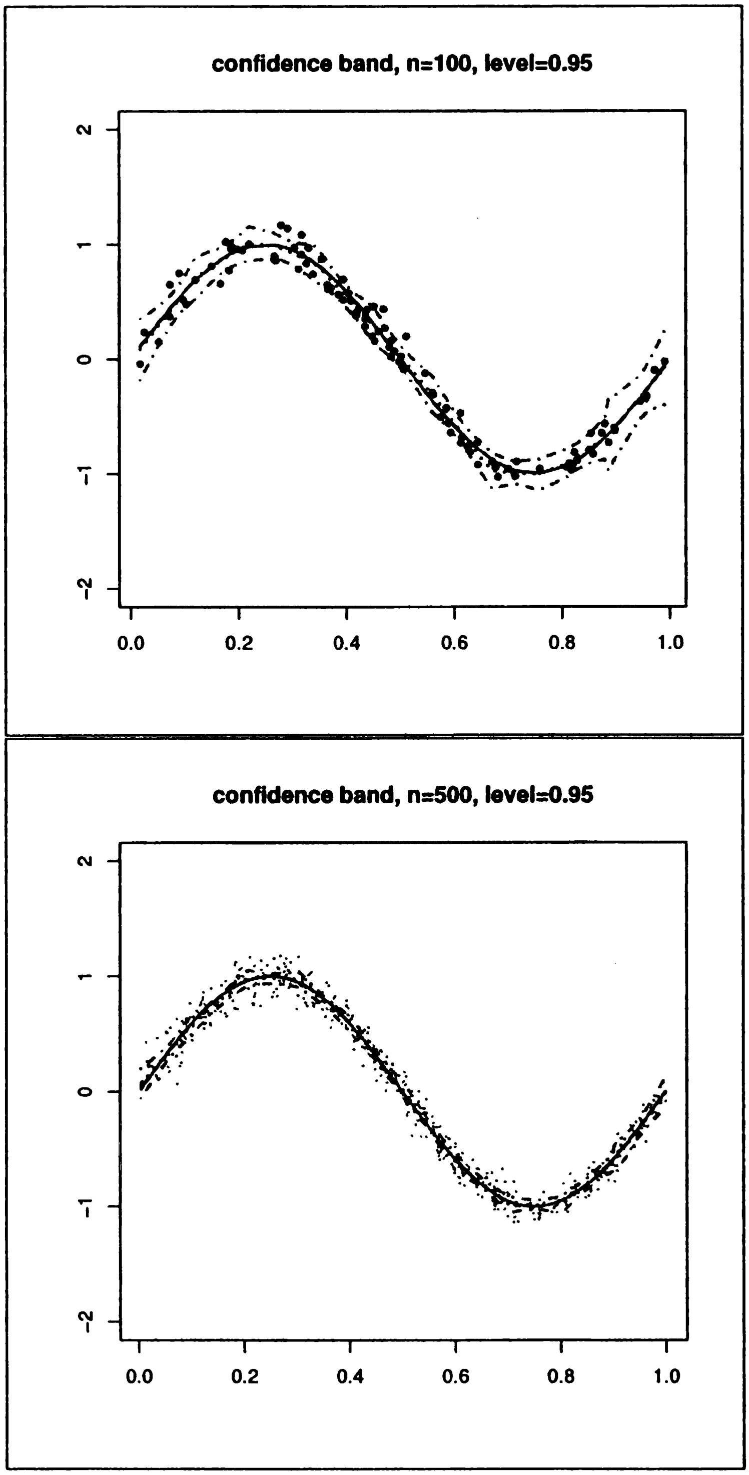


Figure 4.4. Example 2.5.1: 99% constant spline confidence bands with opt = 2

Note: confidence bands (upper and lower dashed curves) computed from (2.4.5) with $k = 1$, opt $= 2$, $\hat{m}_1(x)$ (center dotted curve), $m(x) = \sin(2\pi x)$ (center smooth solid curve).


confidence band, n=100, level=0.95

confidence band, n=500, level=0.95

Figure 4.5. Example 2.5.1: 95% linear spline confidence bands with opt = 1

Note: confidence bands (upper and lower dashed curves) computed from (2.4.5) with $k = 2$, opt = 1, $\hat{m}_2(x)$ (center dotted curve), $m(x) = \sin(2\pi x)$ (center smooth solid curve).

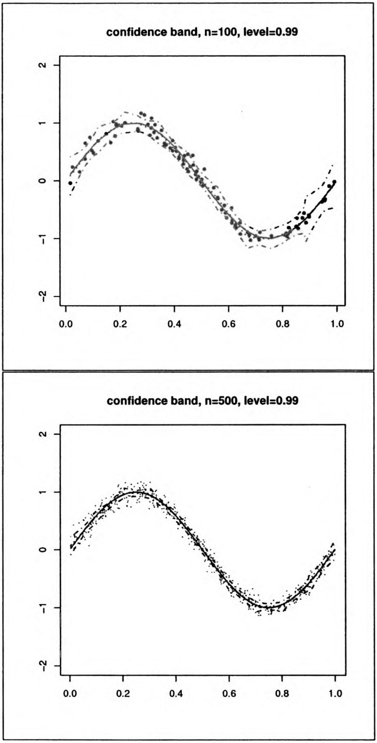


Figure 4.6. Example 2.5.1: 99% linear spline confidence bands with opt = 1

Note: confidence bands (upper and lower dashed curves) computed from (2.4.5) with $k = 2$, opt $= 1$, $\hat{m}_2(x)$ (center dotted curve), $m(x) = \sin(2\pi x)$ (center smooth solid curve).

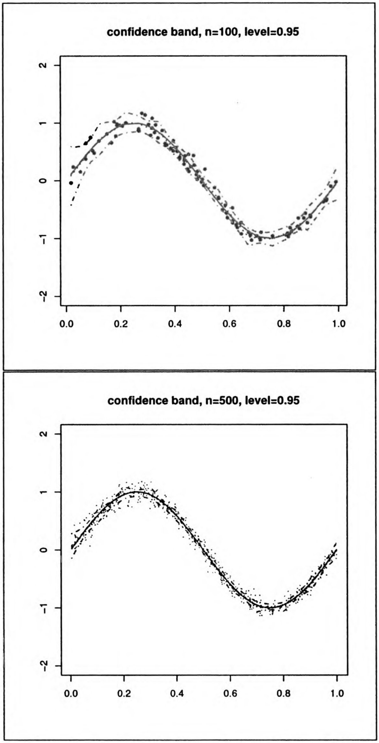


Figure 4.7. Example 2.5.1: 95% linear spline confidence bands with opt $= 2$

Note: confidence bands (upper and lower dashed curves) computed from (2.4.5) with $k = 2$, opt $= 2$, $\hat{m}_2(x)$ (center dotted curve), $m(x) = \sin(2\pi x)$ (center smooth solid curve).

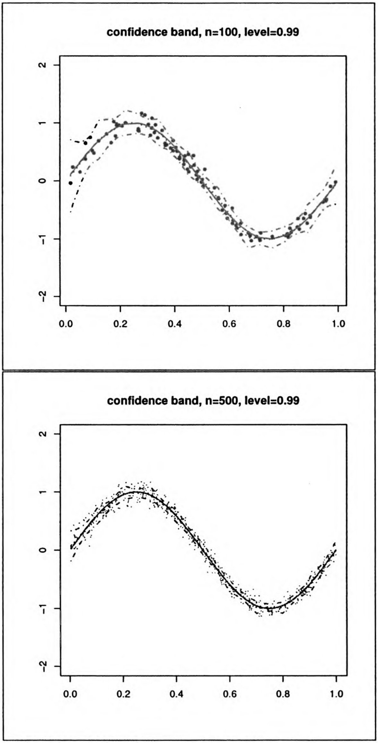


Figure 4.8. Example 2.5.1: 99% linear spline confidence bands with opt = 2

Note: confidence bands (upper and lower dashed curves) computed from (2.4.5) with $k = 2$, opt $= 2$, $\hat{m}_2(x)$ (center dotted curve), $m(x) = \sin(2\pi x)$ (center smooth solid curve).

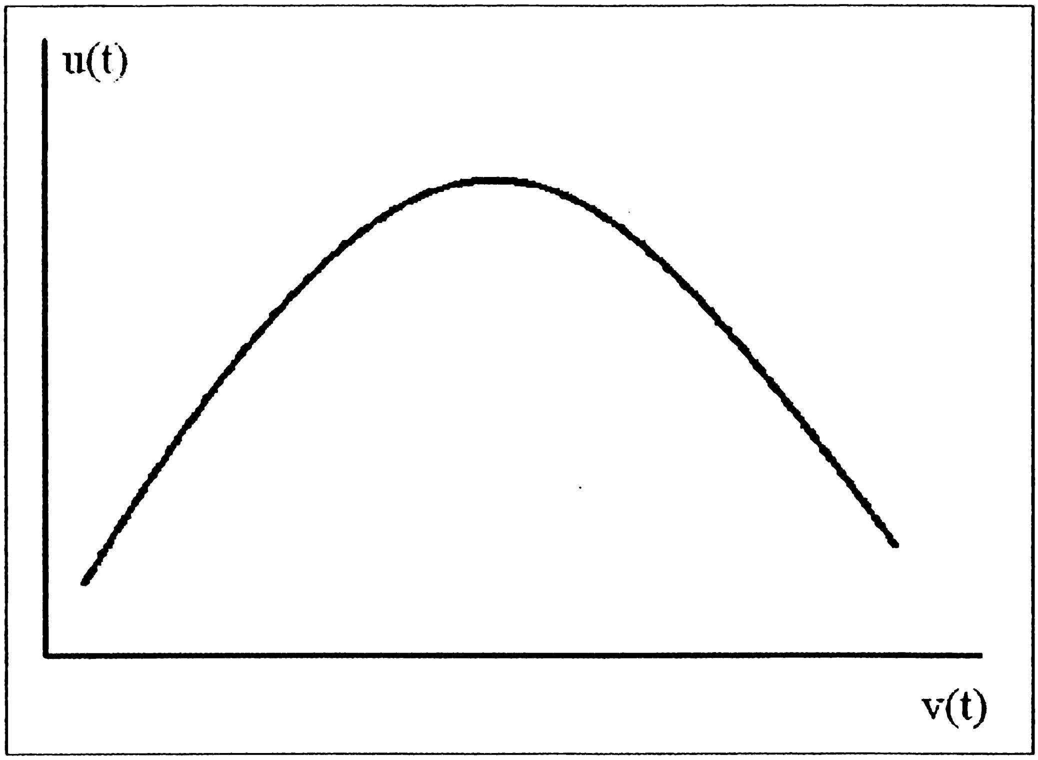


Figure 4.9. Example 2.5.2: Plot of the EKC in terms of $u(t)$ and $v(t)$

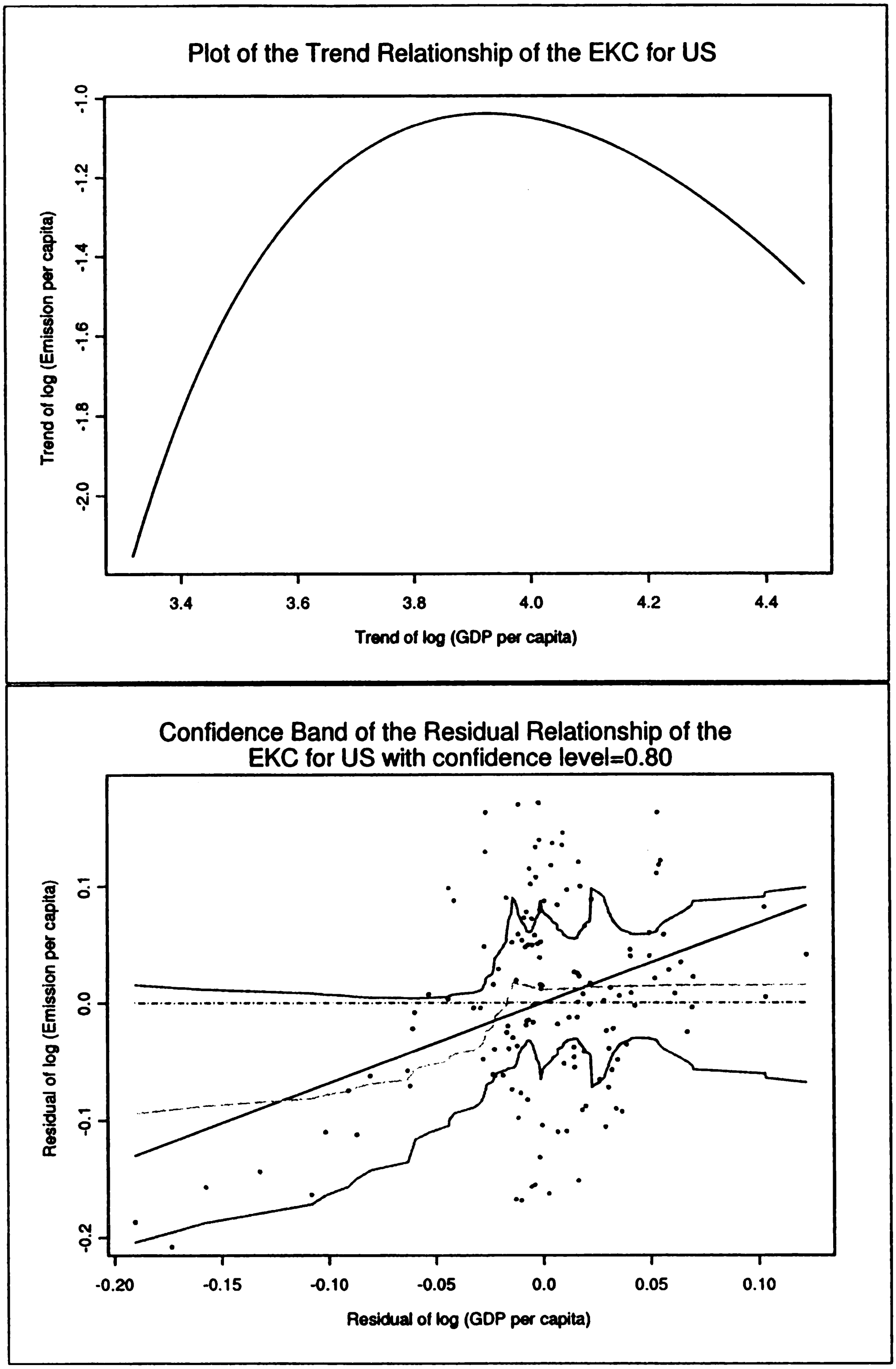
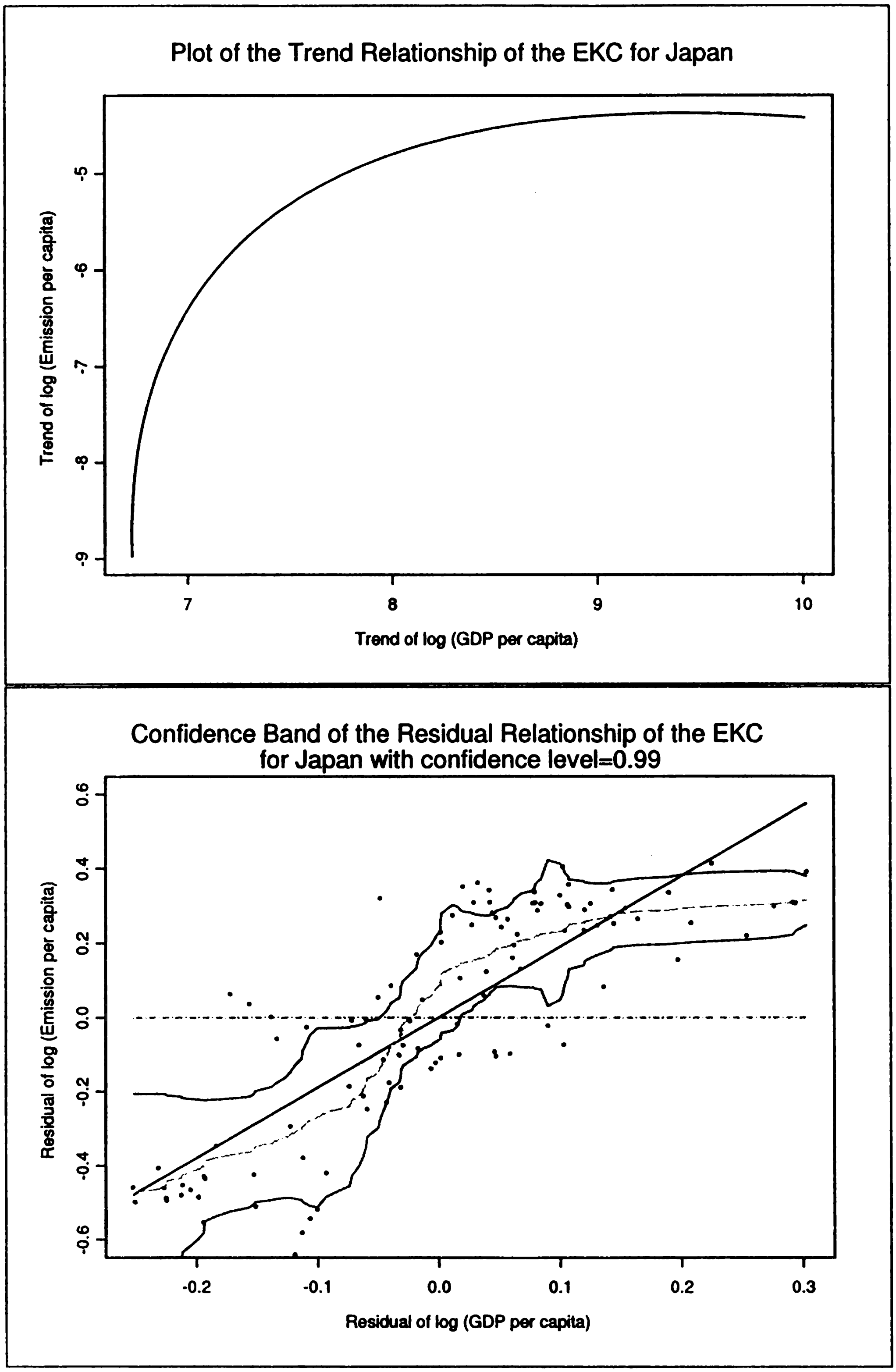


Figure 4.10. Example 2.5.2: Trend and noise analysis of US

Note: linear fit (solid), zero fit (dotted dashed) and spline fit (dashed) with 80% bands (upper and lower solid).



Plot of the Trend Relationship of the EKC for Japan

Confidence Band of the Residual Relationship of the EKC for Japan with confidence level=0.99

Figure 4.11. Example 2.5.2: Trend and noise analysis of Japan

Note: linear fit (solid), zero fit (dotted dashed) and spline fit (dashed) with 99% bands (upper and lower solid).

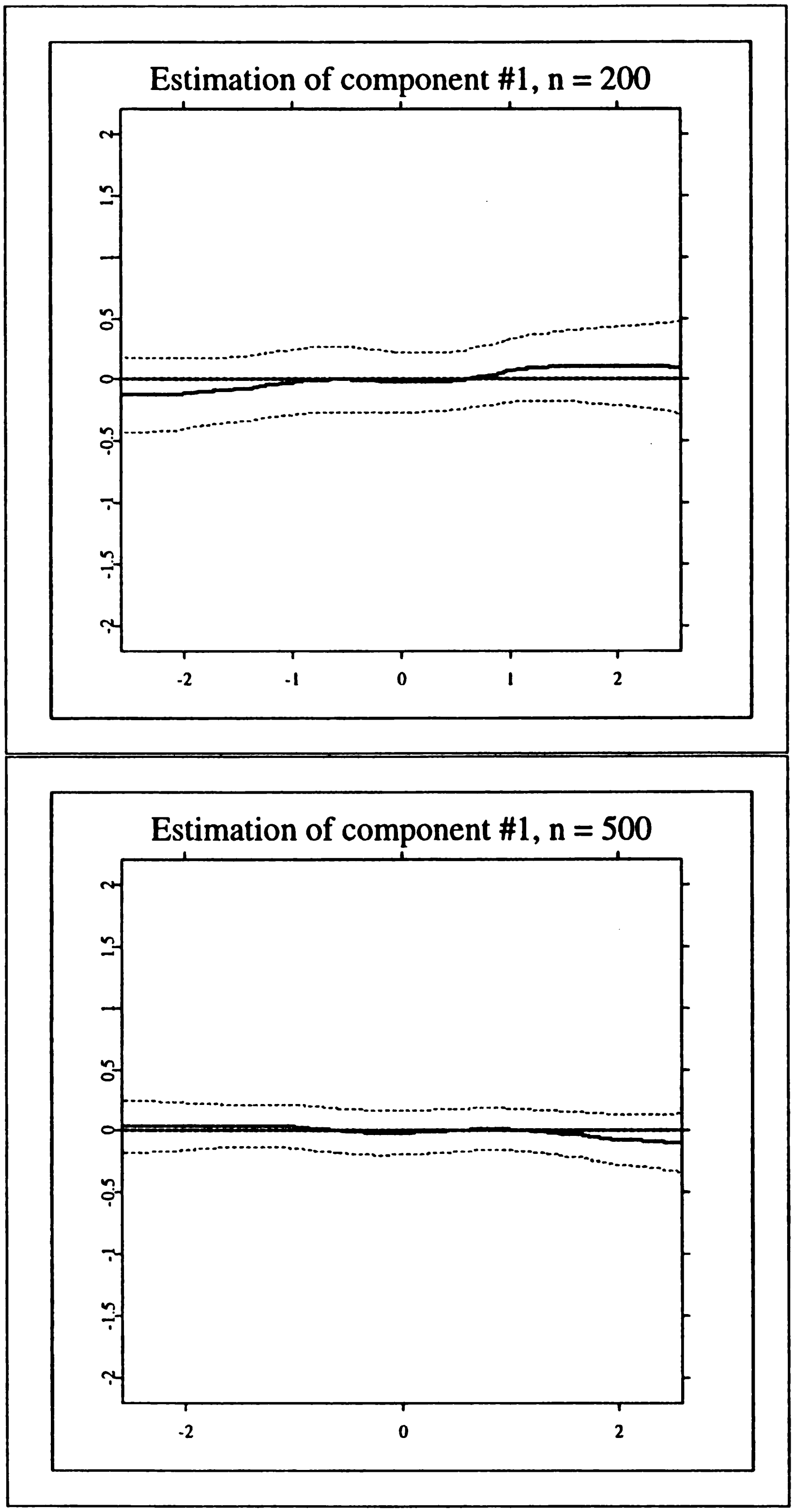
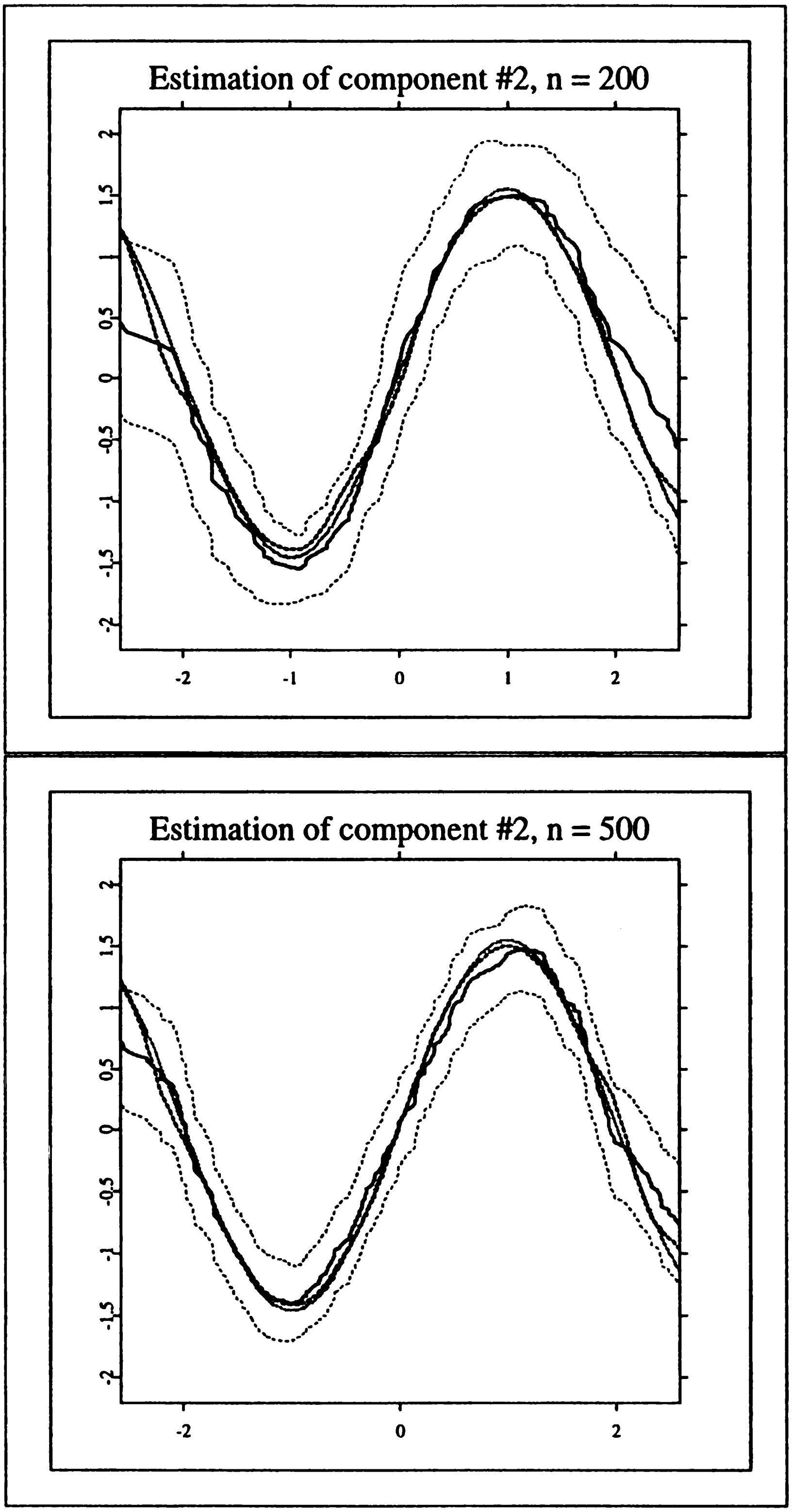
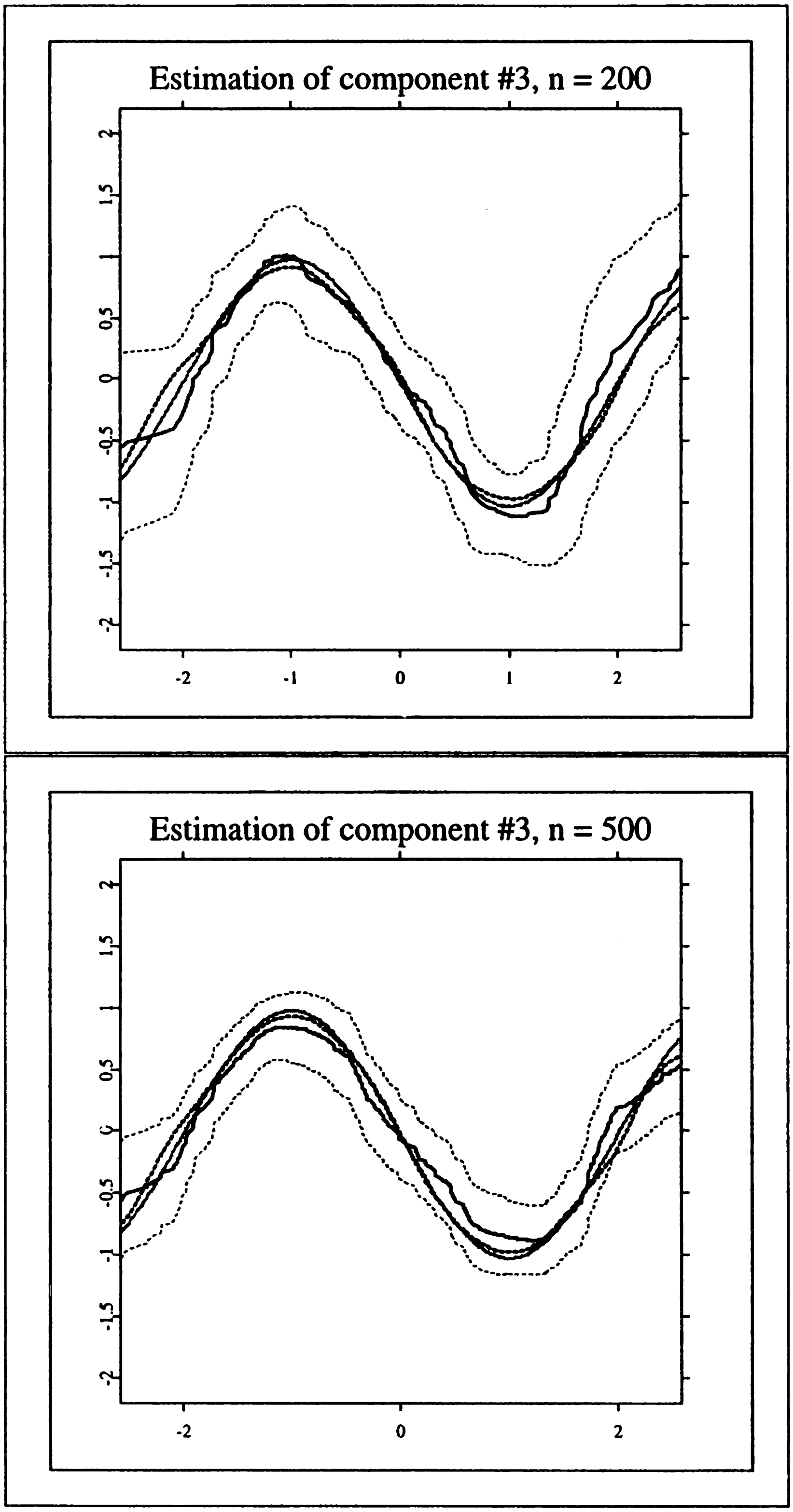


Figure 4.12. Example 3.6.1: SPBK estimator with confidence intervals for the first component Note: oracle estimator (dotted), SPBK estimator (solid) and 95% pointwise confidence intervals constructed by (3.2.13) (thin dashed) of the first component (smooth solid curve).


Estimation of component #2, n = 200

Estimation of component #2, n = 500

Figure 4.13. Example 3.6.1: SPBK estimator with confidence intervals for the second component Note: oracle estimator (dotted), SPBK estimator (solid) and 95% pointwise confidence intervals constructed by (3.2.13) (thin dashed) of the second component (smooth solid).



Estimation of component #3, n = 200

Estimation of component #3, n = 500

Figure 4.14. Example 3.6.1: SPBK estimator with confidence intervals for the third component

Note: oracle estimator (dotted), SPBK estimator (solid) and 95% pointwise confidence intervals constructed by (3.2.13) (thin dashed) of the third component (smooth solid).

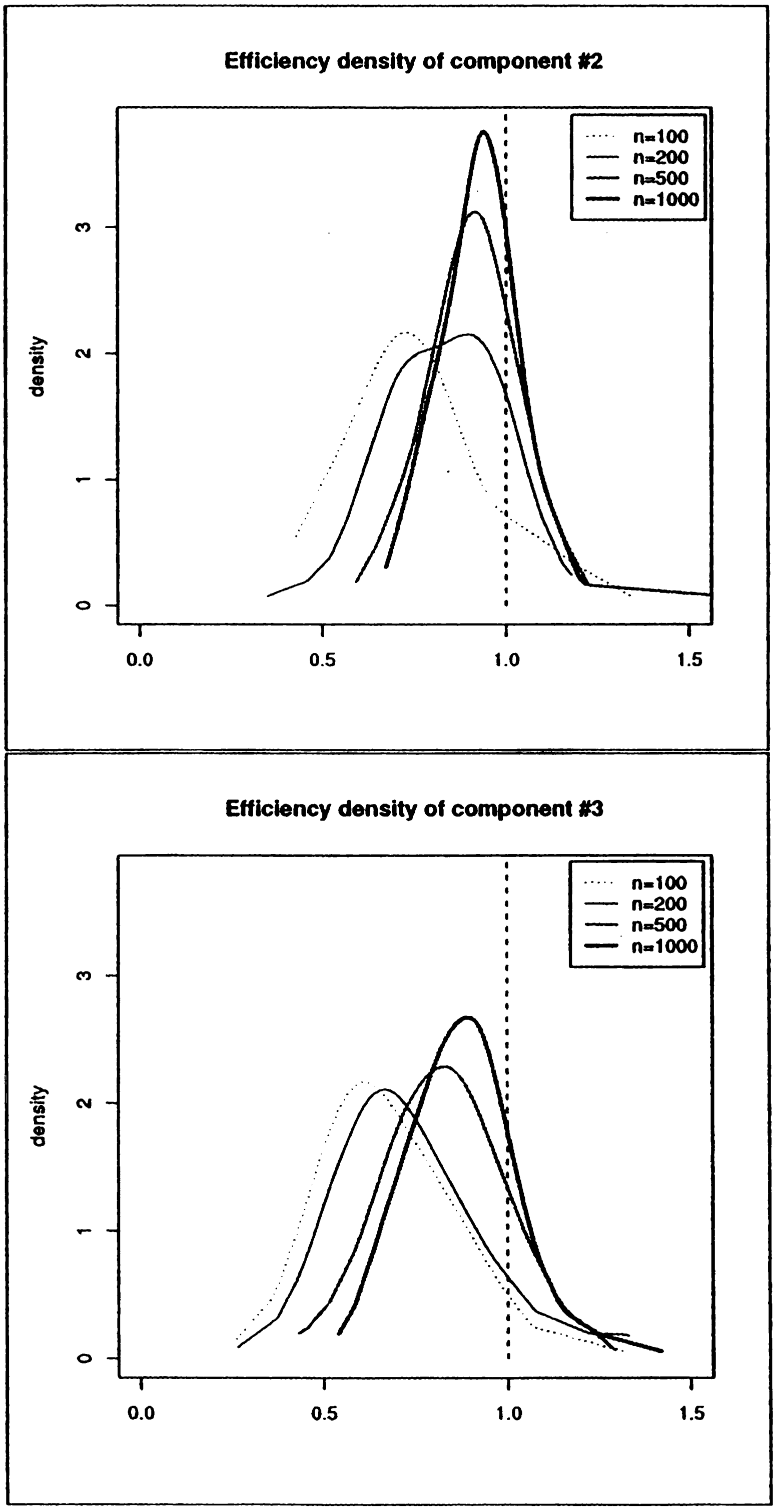


Figure 4.15. Example 3.6.1: Plot of the relative efficiencies of components 2 and 3

Note: the empirical efficiencies of $\hat{m}^*_\alpha(x_\alpha)$ to $\bar{m}^*_\alpha(x_\alpha)$ computed by (3.6.1) based on 100 replications, $\alpha = 2, 3$.

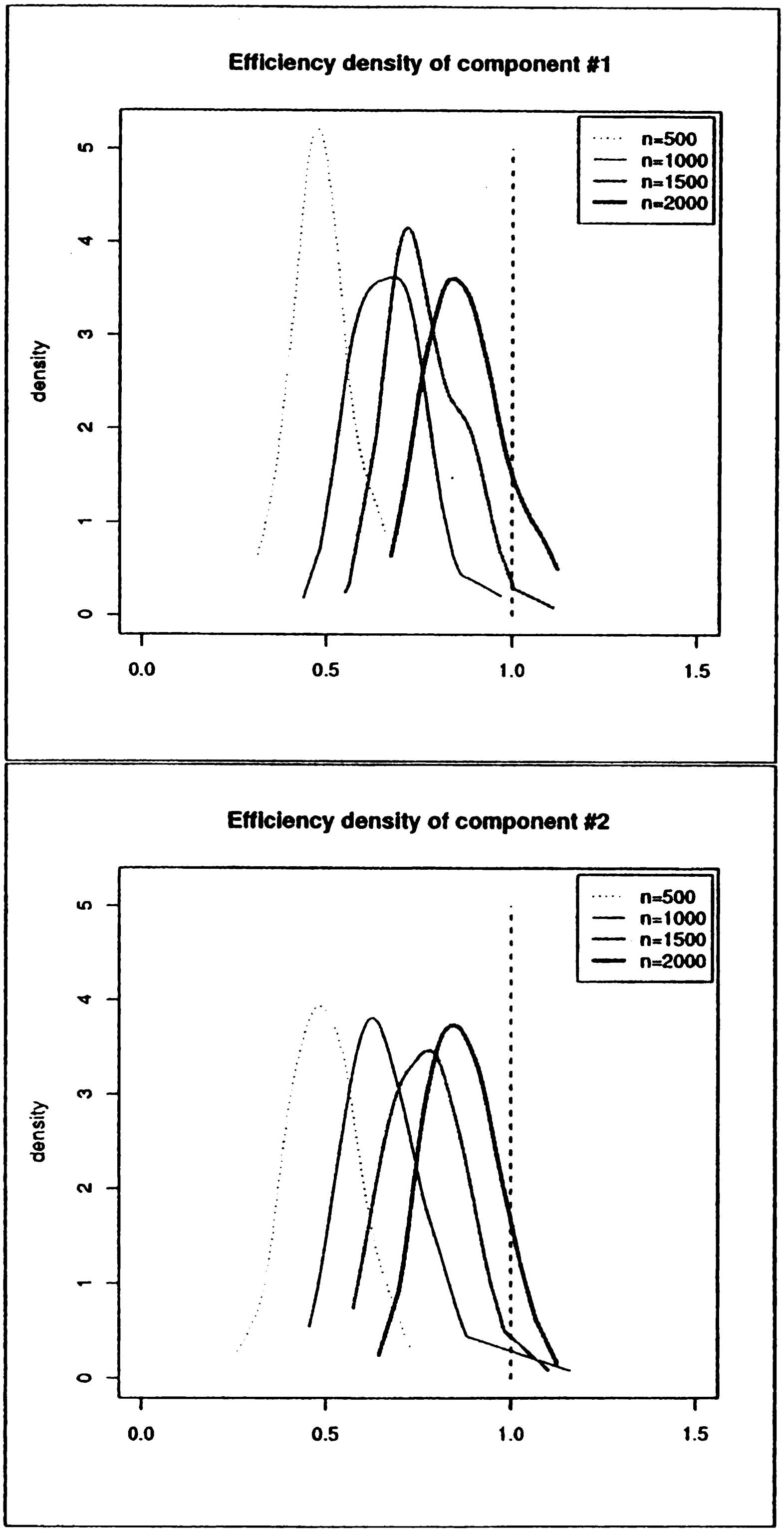
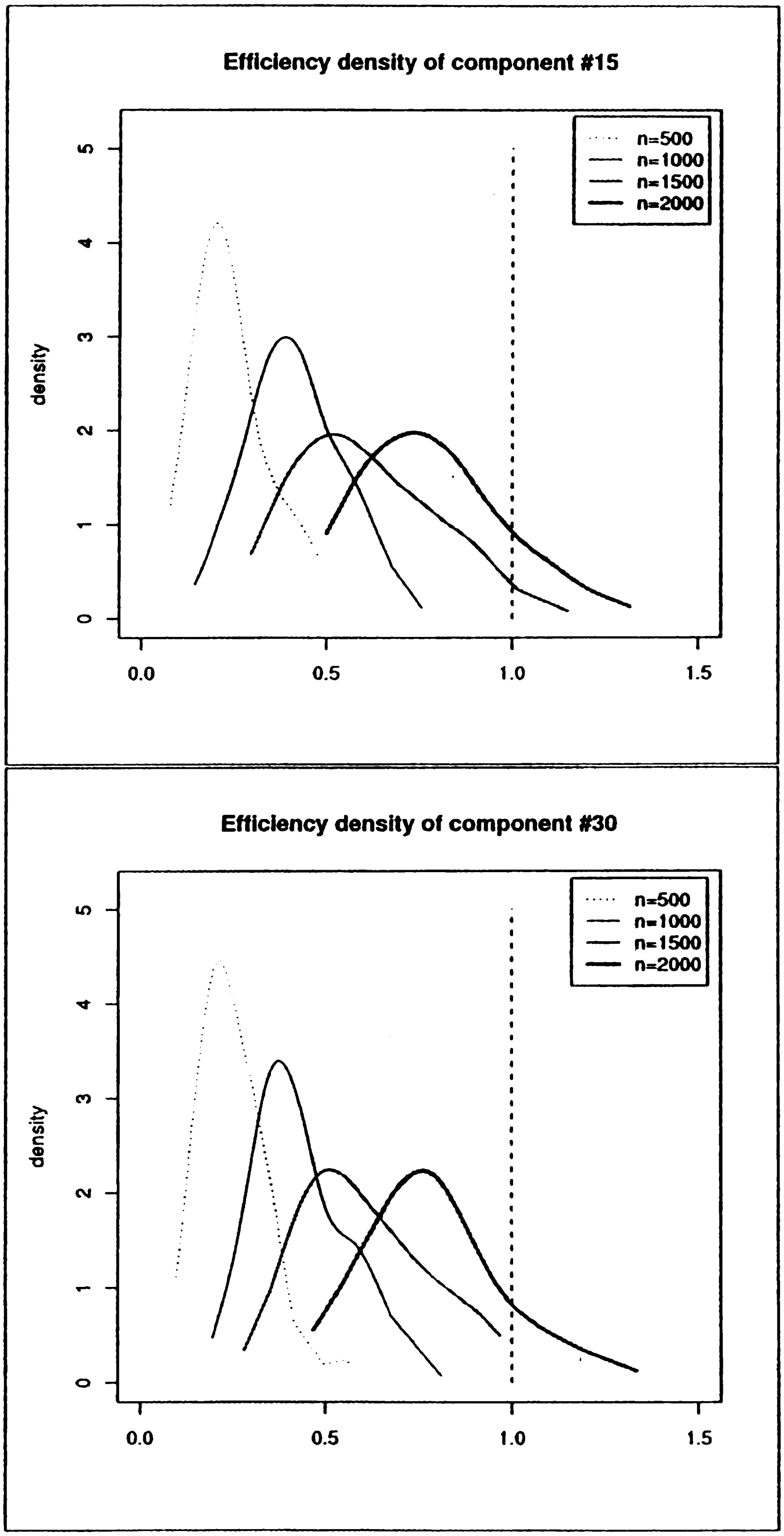


Figure 4.16. Example 3.6.2: Plot of the relative efficiencies of components 1 and 2

Note: the empirical efficiencies of $\hat{m}_{\alpha}^{*}(x_{\alpha})$ to $\tilde{m}_{\alpha}^{*}(x_{\alpha})$ computed by (3.6.1) based on 100 replications, $\alpha = 1, 2$.

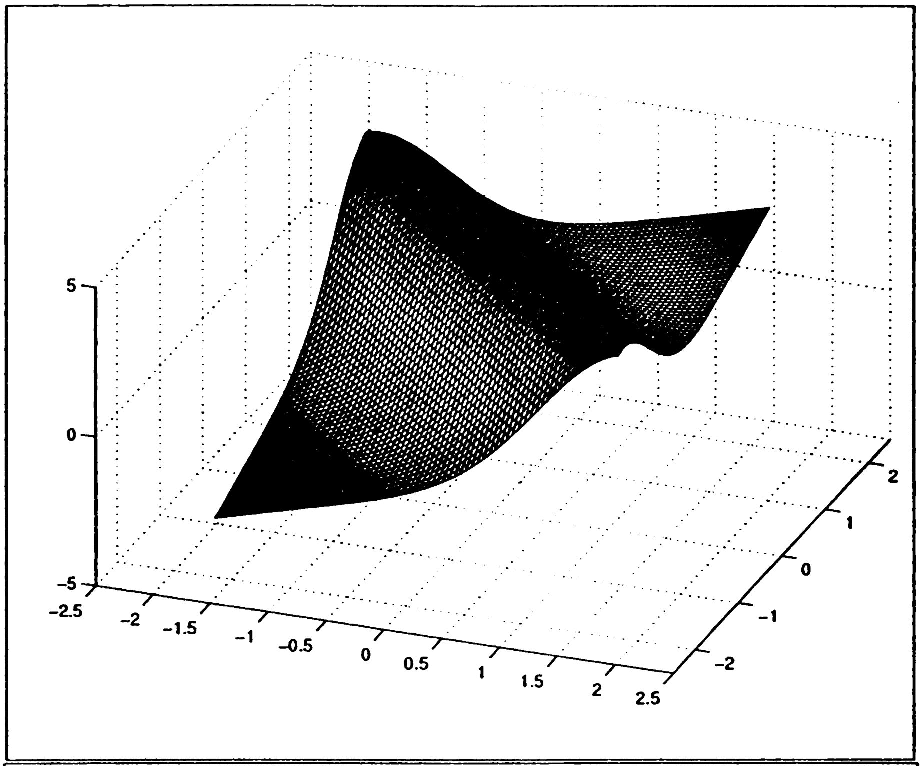


Efficiency density of component #15

Efficiency density of component #30

Figure 4.17. Example 3.6.2: Plot of the relative efficiencies of components 15 and 30

Note: the empirical efficiencies of $\hat{m}^*_\alpha(x_\alpha)$ to $\tilde{m}^*_\alpha(x_\alpha)$ computed by (3.6.1) based on 100 replications, $\alpha = 15, 30$.



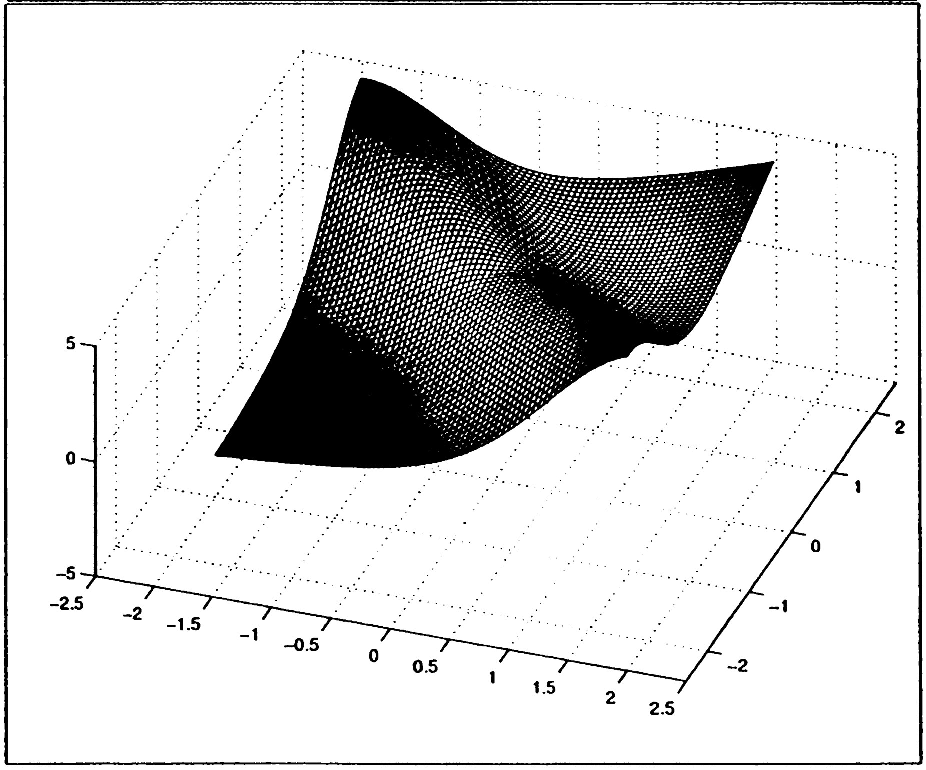


Figure 4.18. Example 4.4.1: The actual bivariate surface

Note: the actual surface m in model (4.4.1) with respect to $\delta = 0, 1$.

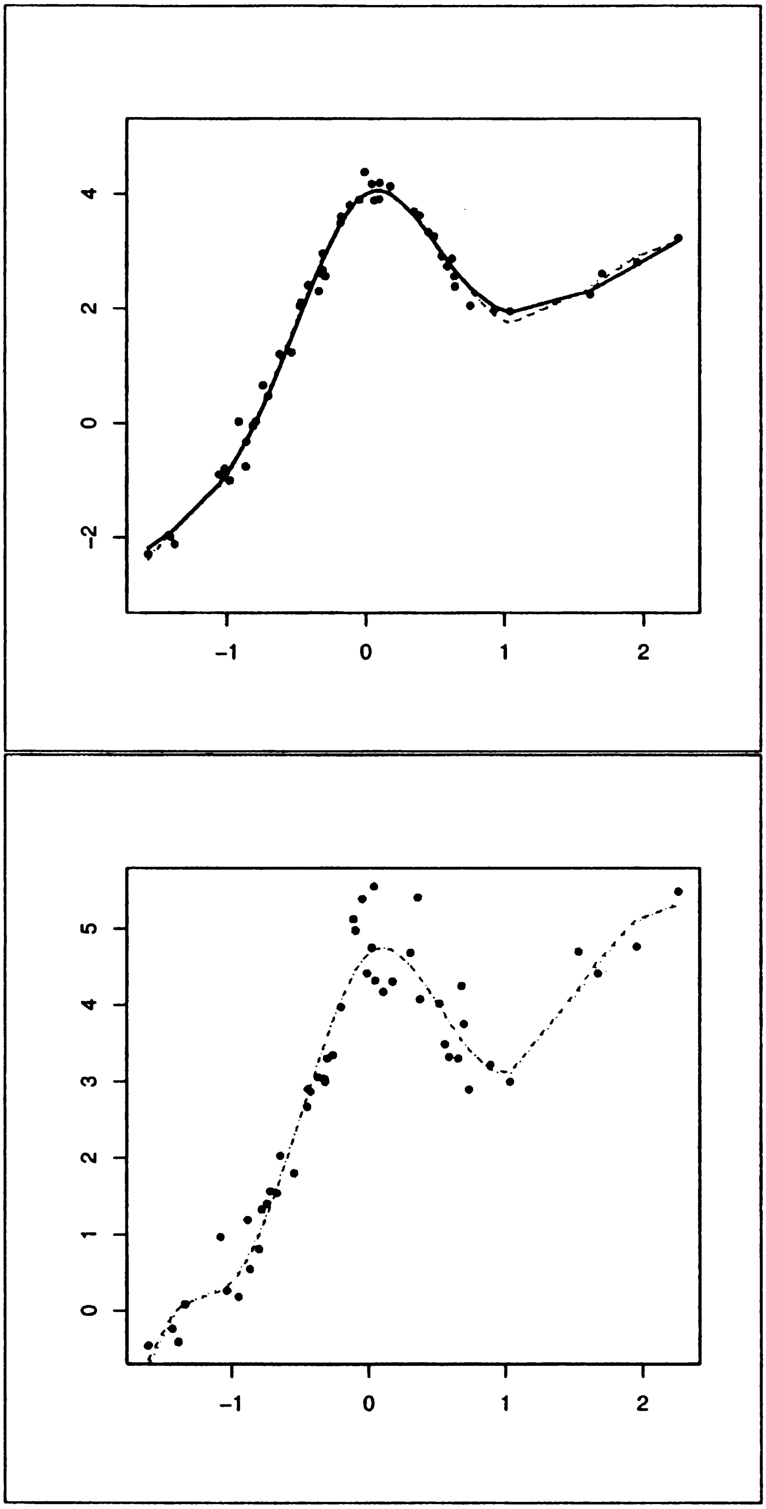


Figure 4.19. Example 4.4.1: The univariate approximation to the bivariate surface

Note: function g (solid curve); estimate of g (dotted curve) by $\boldsymbol{\theta}_0$; estimate of g (dashed curve) by $\hat{\boldsymbol{\theta}} = (0.69016, 0.72365)^T$ for $\delta = 0$ and $(0.72186, 0.69204)^T$ for $\delta = 1$.

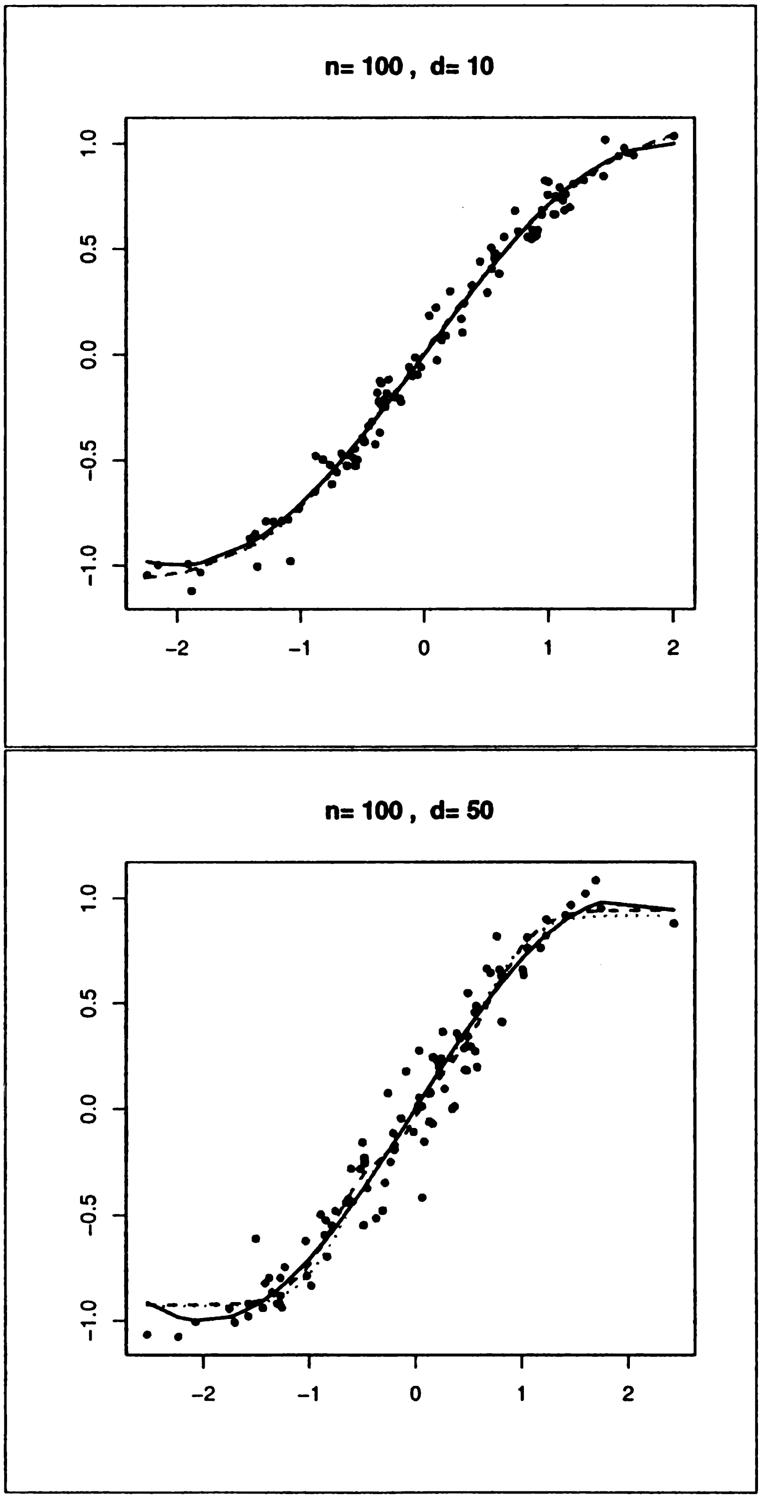


Figure 4.20. Example 4.4.2: The univariate approximation (d = 10, 50)

Note: estimate of g with $\hat{\theta}$ (dotted curve), estimate of g with θ_0 (dashed curve), true function $m(\mathrm{x})$ in (4.4.2) (solid curve).

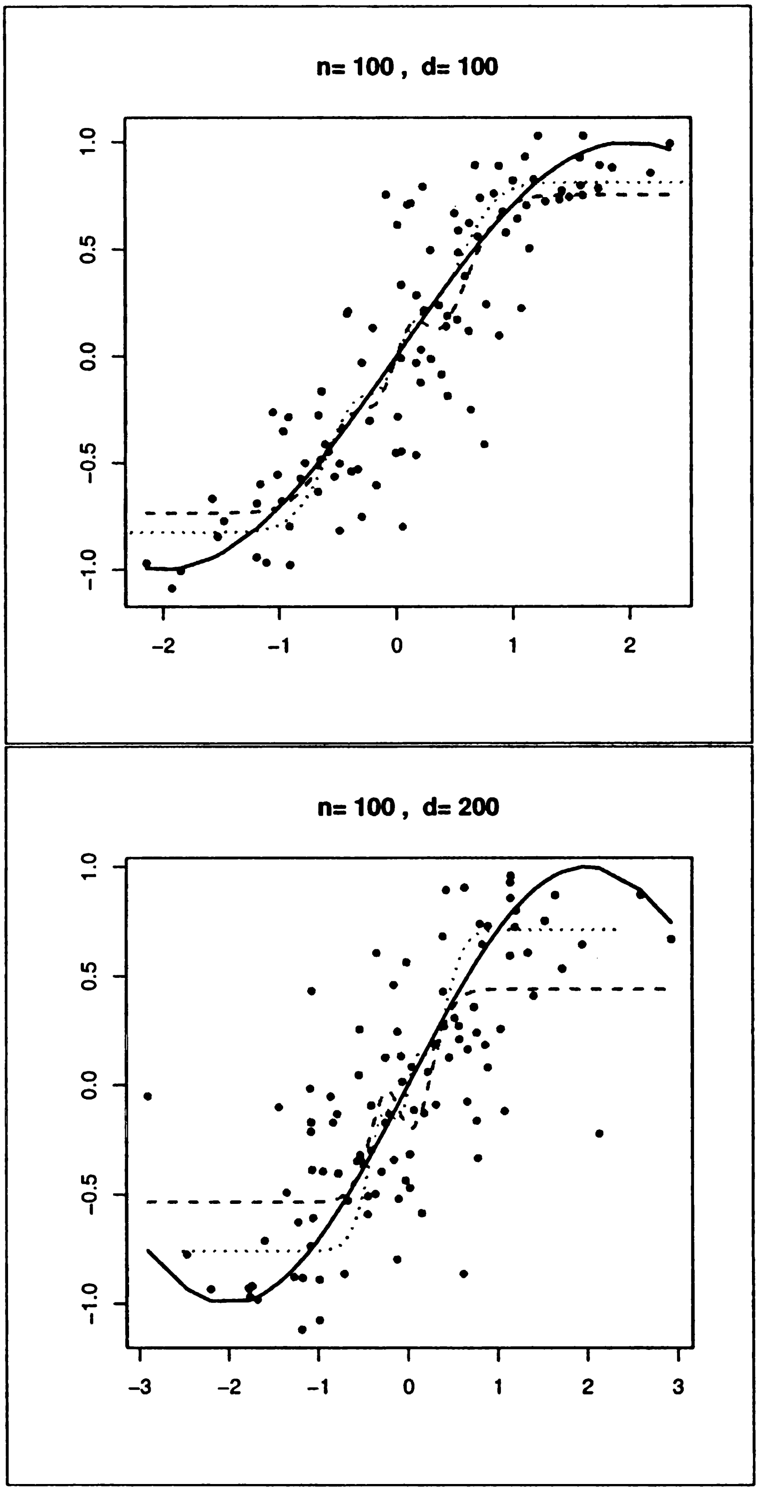


Figure 4.21. Example 4.4.2: The univariate approximation (d = 100, 200)

Note: estimate of g with $\hat{\boldsymbol{\theta}}$ (dotted curve), estimator of g with $\boldsymbol{\theta}_0$ (dashed curve), the true function $m(\mathbf{x})$ in (4.4.2) (solid curve).

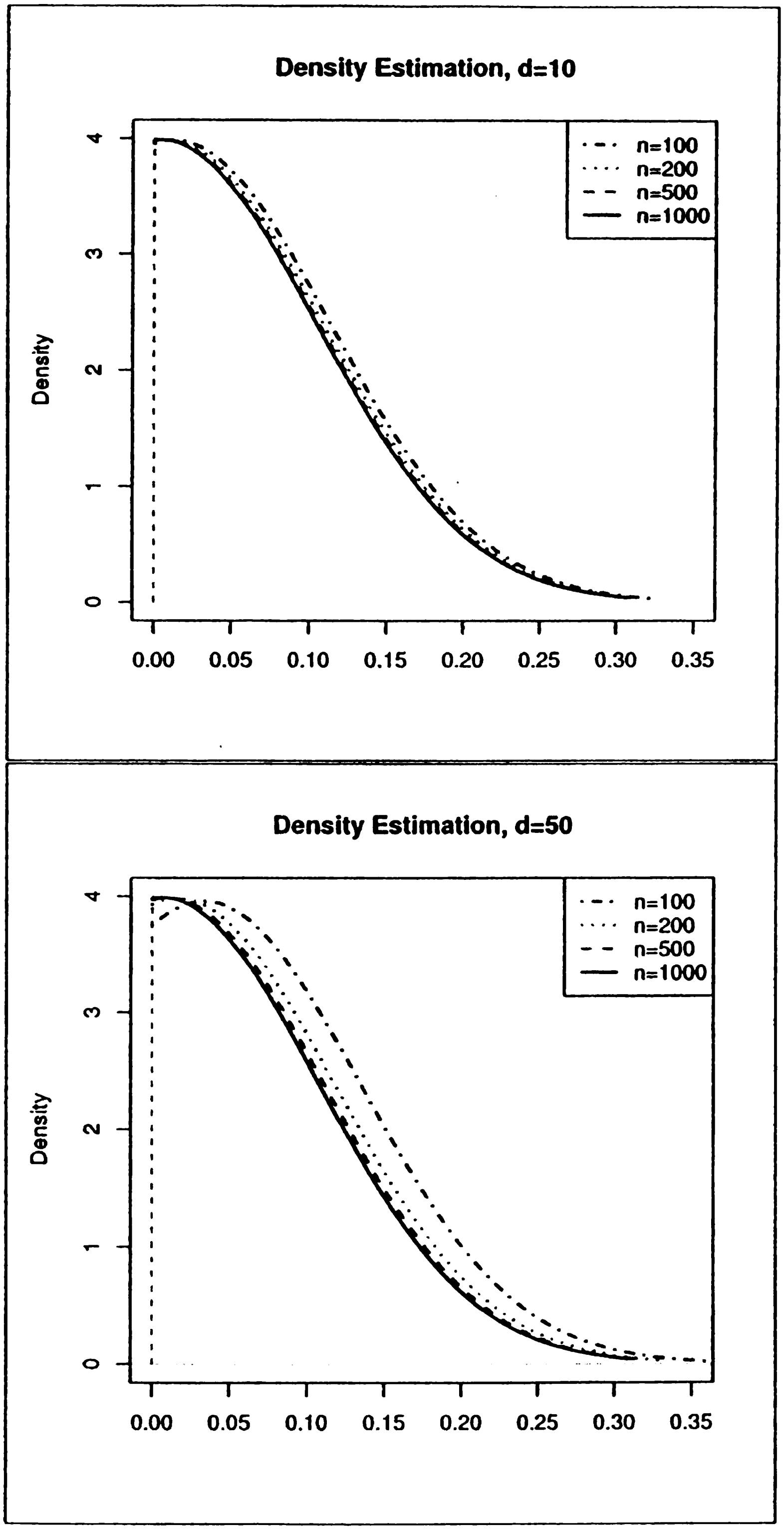


Figure 4.22. Example 4.4.2: Kernel density plots of the error norms

Note: the kernel density estimators of $\|\hat{\theta} - \theta_0\|/\sqrt{d}$ are based on 100 replications.

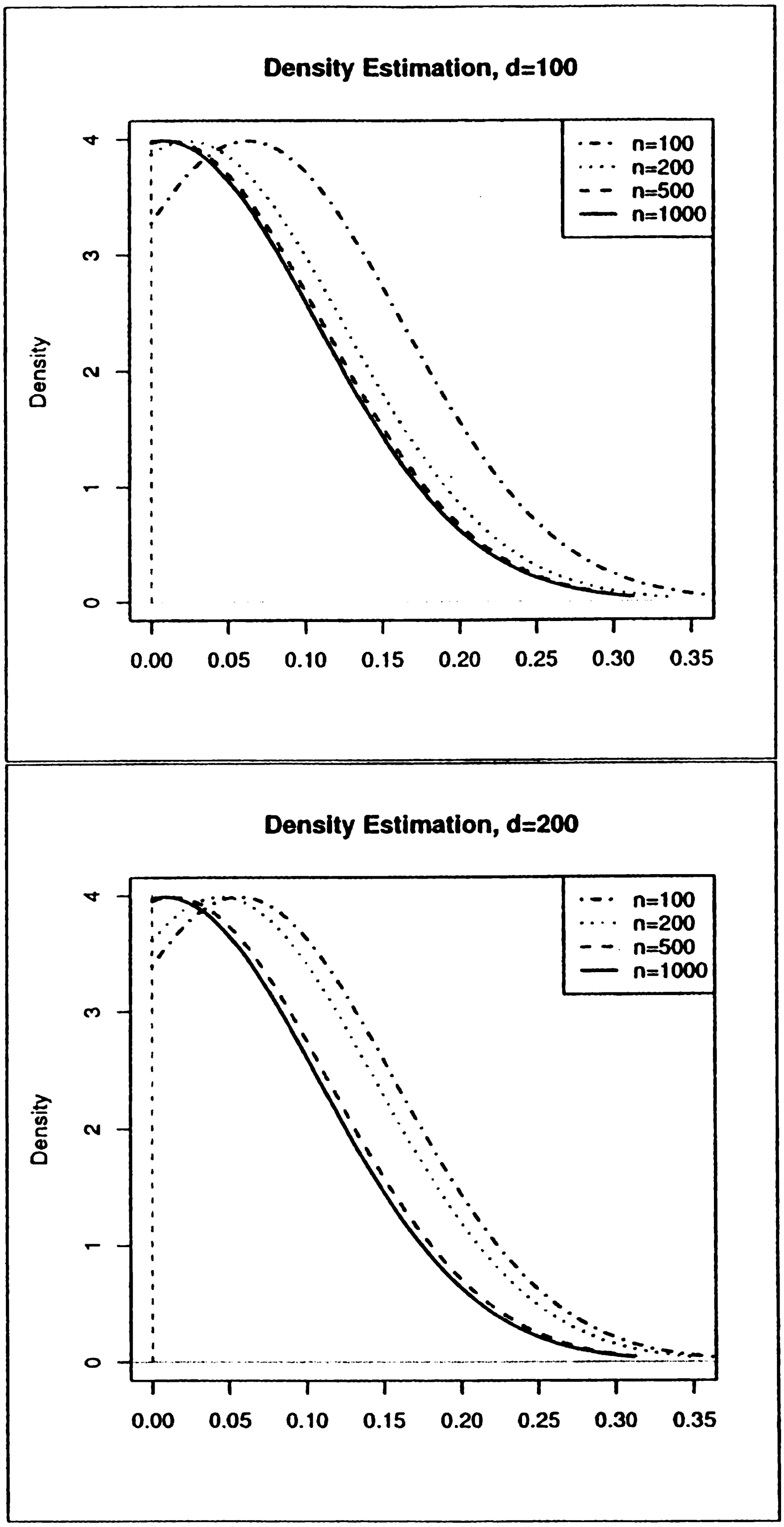


Figure 4.23. Example 4.4.2: Kernel density plots of the error norms

Note: the kernel density estimators of $\|\hat{\theta} - \theta_0\|/\sqrt{d}$ are based on 100 replications.

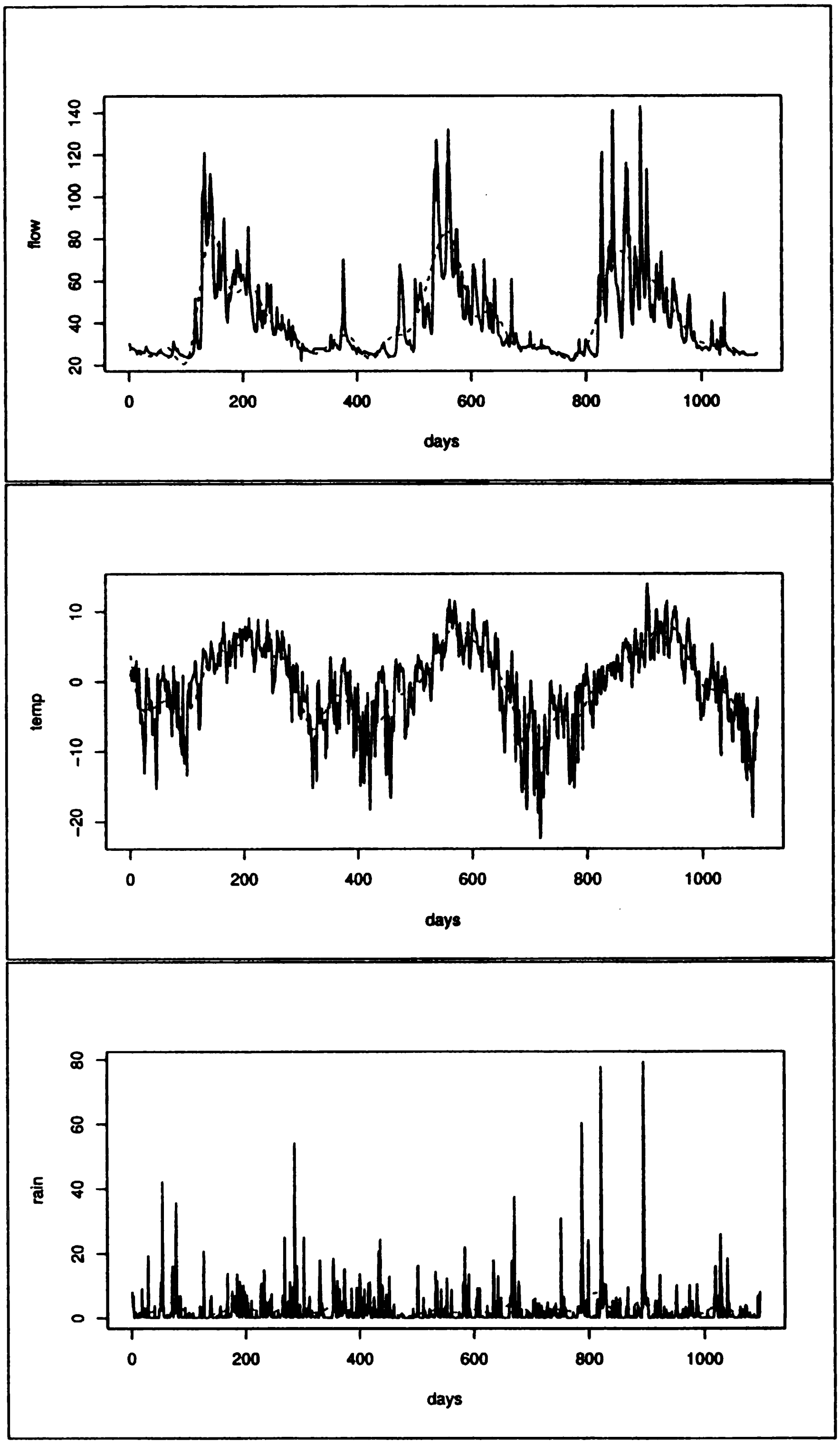


Figure 4.24. Time plots of the daily river flow data

Note: the first, second and third are flow (solid) with trend (dashed), temperature (solid) with trend (dashed line) and precipitation(solid) with trend (dashed) respectively.

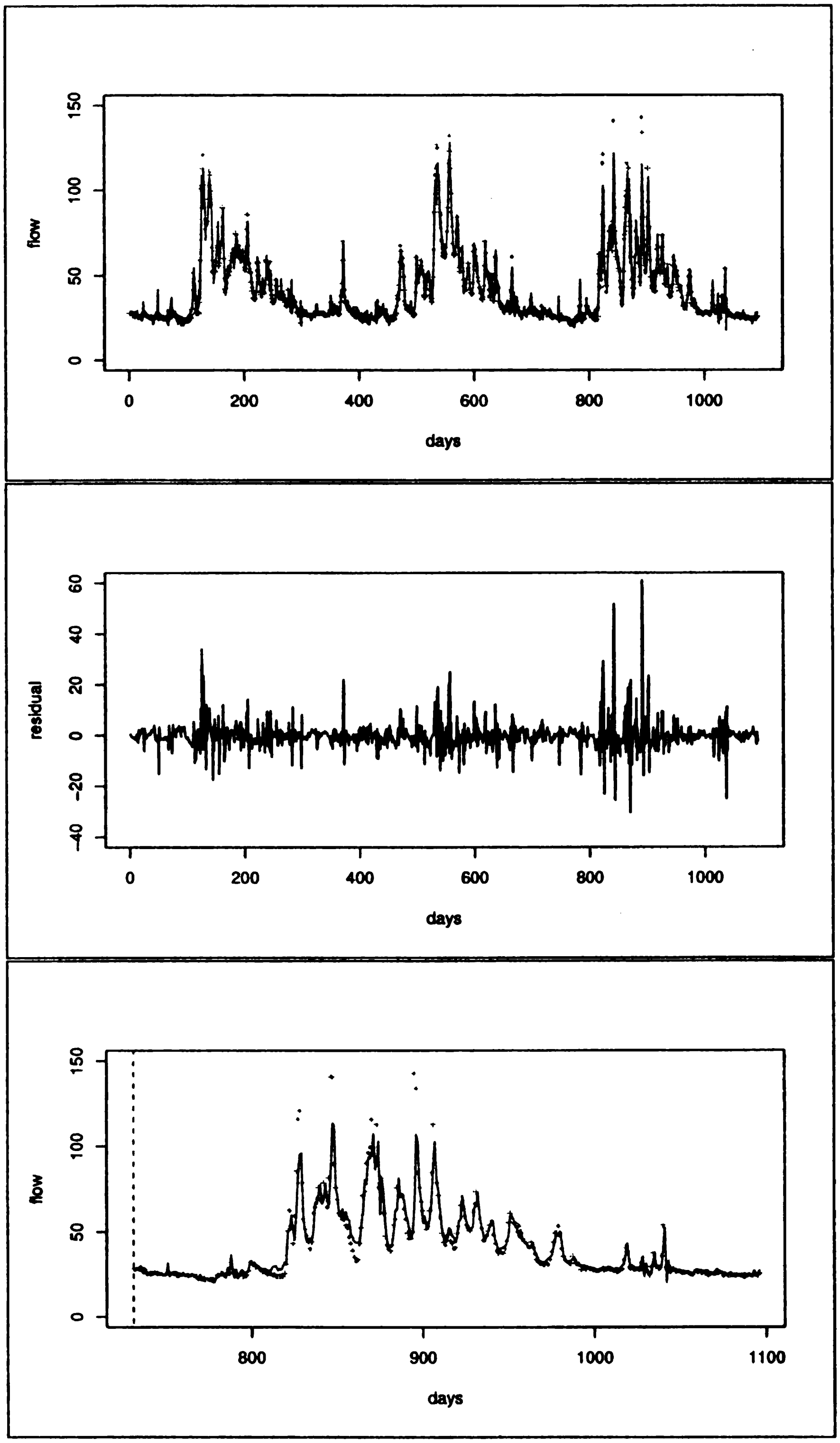


Figure 4.25. The fitted, residual and forecast plots of the river flow data

Note: the first is the river flow data ("+") with the SIP fitted values (line); the second is the residual plot; the third is the out-of-sample rolling forecasts (line) for the third year.

BIBLIOGRAPHY

[1] Bickel, P. J. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. *Ann. Statist.* **1** 1071-1095.

[2] Bosq, D. (1998). *Nonparametric Statistics for Stochastic Processes.* New York: Springer.

[3] Carroll, R., Fan, J., Gijbles, I. and Wand, M. P. (1997). Generalized partially linear single-index models. *J. Amer. Statist. Assoc.* **92** 477-489.

[4] Chen, H. (1991). Estimation of a projection -persuit type regression model. *Ann. Statist.* **19** 142-157.

[5] Chen, R. and Tsay, R. S. (1993). Nonlinear additive ARX models. *J. Amer. Statist. Assoc.* **88** 956–967.

[6] Chen, R., Yang, L. and Hafner, C. (2004). Nonparametric multi-step ahead prediction in time series analysis. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **66** 669-686.

[7] Claeskens, G. and Van Keilegom, I. (2003). Bootstrap confidence bands for regression curves and their derivatives. *Ann. Statist.* **31** 1852-1884.

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

[9] DeVore, R. A. and Lorentz, G. G. (1993). *Constructive Approximation: Polynomials and Splines Approximation.* Springer-Verlag, Berlin.

[10] Doukhan, P. (1994). *Mixing: Properties and Examples.* Springer-Verlag, New York.

[11] Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*, London: Chapman and Hall.

[12] Fan, J. and Jiang, J. (2005). Nonparametric inference for additive models. *J. Amer. Statist. Assoc.* **100** 890-907.

[13] Fan, J., Härdle, W. and Mammen, E. (1998). Direct estimation of low-dimensional components in additive models. *Ann. Statist.* **26** 943-971.

[14] Fan, J. and Yao, Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods.* New York: Springer.

[15] Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* **76** 817-823.

[16] Hall, P. (1989). On projection pursuit regression. *Ann. Statist.* **17** 573-588.

[17] Hall, P. and Titterington, D. M. (1988). On confidence bands in nonparametric density estimation and regression. *J. Multivariate Anal.* **27** 228-254.

[18] Härdle, W. (1989). Asymptotic maximal deviation of M-smoothers. *J. Multivariate Anal.* **29** 163-179.

[19] Härdle, W. (1990). *Applied Nonparametric Regression.* Cambridge University Press, Cambridge.

[20] Härdle, W. and Hall, P. and Ichimura, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.* **21** 157-178.

[21] Härdle, W. , Hlávka, Z. and Klinke, S. (2000). *XploRe Application Guide.* Springer-Verlag, Berlin.

[22] 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. *Ann. Statist.* **25** 1443-1450.

[23] Härdle, W. and Stoker, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *J. Amer. Statist. Assoc.* **84** 986-995.

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

[25] Hengartner, N. W. and Sperlich, S. (2005). Rate optimal estimation with the integration method in the presence of many covariates. *J. Multivariate Anal.* **95** 246-272.

[26] Horowitz, J. L. and Härdle, W. (1996). Direct semiparametric estimation of single-index models with discrete covariates. *J. Amer. Statist. Assoc.* **91** 1632-1640.

[27] Horowitz, J. and Mammen, E. (2004). Nonparametric estimation of an additive model with a link function. *Ann. Statist.* **32** 2412-2443.

[28] Horowitz, J. Klemelä, J. and Mammen, E. (2006). Optimal estimation in additive regression. *Bernoulli* **12** 271-298.

[29] Hristache, M., Juditski, A. and Spokoiny, V. (2001). Direct estimation of the index coeffcients in a single-index model. *Ann. Statist.* **29** 595-623.

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

[31] Huang, J. Z. (2003). Local asymptotics for polynomial spline regression. *Ann. Statist.* **31** 1600-1635.

[32] Huang, J. and Yang, L. (2004). Identification of nonlinear additive autoregressive models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **66**, 463-477.

[33] Huber, P. J. (1985). Projection pursuit (with discussion). *Ann. Statist.* **13** 435-525.

[34] Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models *Journal of Econometrics* **58** 71-120.

[35] Johnson, R. A. and Wichern, D. W. (1992). *Applied Multivariate Statistical Analysis.* New Jersey: Prentice Hall.

[36] Klein, R. W. and Spady. R. H. (1993). An efficient semiparametric estimator for binary response models. *Econometrica* **61** 387-421.

[37] Leadbetter, M. R., Lindgren, G. and Rootzén, H. (1983). *Extremes and Related Properties of Random Sequences and Processes.* New York: Springer.

[38] Lefohn, A. S., Husar, J. D. and Husar, R. B. (1999). Estimating historical anthropogenic global sulfur emission patterns for the period 1850-1990. *Atmospheric Environment.* **33** 3435-3444.

[39] Linton, O. B. and Nielsen, J. P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika* **82** 93-101.

[40] Linton, O. B. and Härdle, W. (1996). Estimating additive regression models with known links. *Biometrika* **83** 529-540.

[41] Linton, O. B. (1997). Efficient estimation of additive nonparametric regression models. *Biometrika* **84** 469-473.

[42] Maddison, A. (2003). *The World Economy: Historical Statistics.* Paris: OECD.

[43] Mammen, E., Linton, O. and Nielsen, J. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Ann. Statist.* **27** 1443-1490.

[44] Müller, H. G., Stadtmüller, U. and Schmitt, T. (1987). Bandwidth choice and confidence intervals for derivatives of noisy data. *Biometrika* **74** 743-749.

[45] Neumann, M. H. (1995). Automatic bandwidth choice and confidence intervals in nonparametric regression. *Ann. Statist.* **23** 1937-1959.

[46] Neumann, M. H. (1997). Pointwise confidence intervals in nonparametric regression with heteroscedastic error structure. *Statistics* **29** 1-36.

[47] Opsomer, J. D. and Ruppert, D. (1997). Fitting a bivariate additive model by local polynomial regression. *Ann. Statist.* **25** 186-211.

[48] Pham, D. T. (1986). The mixing properties of bilinear and generalized random coefficient autoregressive models. *Stochastic Anal. Appl.* **23** 291-300.

[49] Powell, J. L., Stock, J. H. and Stoker, T. M. (1989). Semiparametric estimation of index coefficients. *Econometrica.* **57** 1403-1430.

[50] Robinson, P. M. (1983). Nonparametric estimators for time series. *J. Time Ser. Anal.* **4** 185-207.

[51] Rosenblatt, M. (1976). On the maximal deviation of k-dimensional density estimates. *Ann. Probab.* **4** 1009-1015.

[52] Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* London: Chapman and Hall.

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

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

[55] Stone, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. *Ann. Statist.* **22** 118-184.

[56] Sunklodas, J. (1984). On the rate of convergence in the central limit theorem for strongly mixing random variables. *Lithuanian Math. J.* **24** 182-190.

[57] Tjøstheim, D. and Auestad, B. (1994). Nonparametric identification of nonlinear time series: projections. *Amer. Statist.* **89** 1398-1409.

[58] Tong, H. (1990). *Nonlinear Time Series: A Dynamical System Approach.* Oxford, U.K.: Oxford University Press.

[59] Tong, H., Thanoon, B. and Gudmundsson, G. (1985). Threshold time series modeling of two icelandic riverflow systems. *Time Series Analysis in Water Resources.* ed. K. W. Hipel, American Water Research Association.

[60] Tusnády, G. (1977). A remark on the approximation of the sample df in the multidimensional case. *Period. Math. Hungar.* **8** 53-55.

[61] Wang, L. and Yang, L. (2007). Spline-backfitted kernel smoothing of nonlinear additive autoregression model. *Ann. Statist.* Forthcoming.

[62] Xia, Y. (1998). Bias-corrected confidence bands in nonparametric regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **60** 797-811.

[63] Xia, Y. and Li, W. K. (1999). On single-index coefficient regression models. *J. Amer. Statist. Assoc.* **94** 1275-1285.

[64] Xia, Y., Li, W. K., Tong, H. and Zhang, D. (2004). A goodness-of-fit test for single-index models. *Statist. Sinica.* **14** 1-39.

[65] Xia, Y., Tong, H., Li, W. K. and Zhu, L. (2002). An adaptive estimation of dimension reduction space. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 363-410.

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

[67] Xue, L. and Yang, L. (2006 b). Additive coefficient modeling via polynomial spline. *Statistica Sinica* **16** 1423-1446.

[68] Yang, L., Härdle, W. and Nielsen, J. P. (1999). Nonparametric autoregression with multiplicative volatility and additive mean. *J. Time Ser. Anal.* **20** 579-604.

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

[70] Zhang, F. (1999). *Matrix Theory: Basic Results and Techniques.* New York: Springer.

[71] Zhou, S., Shen, X. and Wolfe, D. A. (1998). Local asymptotics of regression splines and confidence regions. *Ann. Statist.* **26** 1760-1782.