2
2006

This is to certify that the
dissertation entitled

ADDITIVE COEFFICIENT MODELING VIA MARGINAL INTEGRATION AND POLYNOMIAL SPLINE SMOOTHING

presented by

LAN XUE

has been accepted towards fulfillment
of the requirements for the

Ph.D. degree in Statistics

Major Professor's Signature

July 27, 2005

Date

MSU is an Affirmative Action/Equal Opportunity Institution

ADDITIVE COEFFICIENT MODELLING VIA MARGINAL INTEGRATION AND POLYNOMIAL SPLINE SMOOTHING

By

Lan Xue

A DISSERTATION

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

DOCTOR OF PHILOSOPHY

Department of Statistics and Probability

2005

ABSTRACT

ADDITIVE COEFFICIENT MODELLING VIA MARGINAL INTEGRATION AND POLYNOMIAL SPLINE SMOOTHING

By

Lan Xue

In this dissertation, we propose a flexible semi-parametric model called additive coefficient model (ACM). In the ACM, one assumes that the response depends linearly on some covariates, whose regression coefficients, however, are additive functions of another set of covariates. The ACM can be viewed as a generalization of the classic linear models in the sense that instead of assuming the coefficients to be constants like the linear model does, it allows the regression coefficients to vary with another set of covariates through an additive function form.

This dissertation focuses on the estimation of the ACM. Two different approaches are considered. One is the local polynomial based marginal integration method, and the other one is the polynomial spline estimation. The local polynomial smoothing is *local* in nature, whereas the polynomial spline is a *global* smoothing method. This difference, in turn, leads to the difference in the asymptotic behavior of the two types of estimators.

Under weak dependence, the point-wise asymptotic normality is established for the marginal integration estimators. It is found that the estimators of the parameters in the regression coefficients have rate of convergence $1/\sqrt{n}$, and the nonparametric additive components are estimated at the same rate of convergence as in univariate

smoothing. In contrast, only mean square convergence is established for the polynomial spline estimators. However, the polynomial spline method is much simpler in both computation and inference. The nonparametric versions of AIC and BIC are adopted easily based on polynomial spline estimation, for the model selection purpose.

Monte Carlo studies are conducted to compare the numerical performances of the two estimation methods, as well as the model selection procedures. The simulation studies show that besides being highly efficient in terms of computing, the polynomial spline estimators are also more accurate than or at least as good as the local polynomial based estimators. The ACM is also successfully applied to several interesting empirical examples: West German GNP, Housing price, and Sunspot data, where the semi-parametric additive coefficient model demonstrates superior performance in terms of out-of-sample forecasts.

To my grandfather, my parents, my sisters, and Li

ACKNOWLEDGEMENTS

I would like to thank my advisor, Professor Lijian Yang, for his unwavering support in the past five years of my doctoral program. I am very fortunate to have him as advisor. As an advisor, he always made himself available for all types of responsibilities. He provided continued guidance, endless supply of patience, and tremendous insight for my research. I owe every single piece of my achievement to him.

I'd like to thank Professors Dennis Gilliland, V. S. Mandrekar, Shlomo Levental and Jiaguo Qi for severing on my thesis committee. Professors Gilliland and Mandrekar provided me with tremendous help and support during my last year in the program, specially during the critical time of my job searching. Also I am very grateful to Professor Qi for supporting me in the CLIP program. Working in CLIP has been an invaluable experience for me.

I would also like to express my gratitude to Professor Connie Page, who taught me how to be a good consultant and provided me with a great deal of encouragement, and guidance when I worked at the statistical consulting service center. I also want to thank Professor Yijun Zuo for teaching me a wonderful course in Robust Statistics. My special thanks go to Professor James Stapleton for his continued support and friendship throughout the years. Also I want to thank Professor Habib Salehi and Cathy for assistance in my simulation.

Last, but not least, I'd like to thank all the professors and friends who ever helped me during my stay at Michigan State University. It was so nice being with you.

Contents

List of Tables

List of Figures

Chapter 1

The model

1.1 Introduction

An important task in statistical analysis is to quantify the association between two sets of variables, say a univariate variable Y and a d-dimensional vector $\mathbf{X}$. In regression analysis, one focuses on the averaged (or expected) response of Y given $\mathbf{X}$, i.e., $m(\mathbf{X}) = E(Y|\mathbf{X})$, which is also known as *regression function*. To estimate the unknown regression function $m(\mathbf{X})$, the parametric regression analysis begins with assuming $m(\mathbf{X})$ takes a pre-determined function form with only finitely many unknown parameters, i.e.,

$$m(\mathbf{X}) = m(\boldsymbol{\beta}, \mathbf{X}),$$

where $\boldsymbol{\beta}$ is a set of unknown coefficients, and the function $m(\boldsymbol{\beta}, \mathbf{x})$ is specified in advance. As a special case, the linear regression assumes $m(\boldsymbol{\beta}, \mathbf{x})$ is a linear function in $\boldsymbol{\beta}$. The unknown coefficients $\boldsymbol{\beta}$ can be estimated using e.g. least squares method. However, the restricted parametric form often can't explain (or approximate) well

the complicated data structure. Furthermore, the parametric regression can lead to excessive estimation biases and erroneous inferences, if a wrong model function $m(\boldsymbol{\beta}, \mathbf{x})$ is used.

On the other hand, nonparametric regression makes minimal assumptions about the regression function m. Without assuming $m(\boldsymbol{\beta}, \mathbf{x})$ take any particular form, it allows the data to speak for themselves , thus they uncover the data structure that linear and parametric regression are unable to detect. To estimate the nonparametric regression function m, several smoothing methods were developed, for example, kernel smoothing (Nadaraya 1964, Watson 1964, Gasser & Müller 1984), local polynomial smoothing (Cleveland 1979, Wand & Jones 1995, Fan & Gijbels 1996), polynomial spline (Stone 1985), smoothing spline (Eubank 1988, Wahba 1990), penalized spline (Eilers & Marx 1996, Ruppert, Wand & Carroll 2003) and Wavelet thresholding (Chiu 1992, Donoho & Johnstone 1995, Härdle, et al. 1998). In this dissertation , we focus on two of them: the local polynomial smoothing and the polynomial spline.

A serious limitation of the general nonparametric model is the "curse of dimensionality" phenomenon. This term refers to the fact that the convergence rate of nonparametric smoothing estimators becomes rather slow when the estimation target is a general function of a large number of variables without additional structures. Many efforts have been made to impose structures on the regression function to partly alleviate the "curse of dimensionality", which is broadly described as dimension reduction. Some well-known dimension reduction approaches are: (generalized) additive models (Chen & Tsay 1993a, Hastie & Tibshirani 1990, Sperlich, Tjøstheim & Yang 2002, Stone 1985), partially linear models (Härdle, Liang & Gao 2000) and varying

coefficient models (Hastie & Tibshirani 1993).

The idea of the varying coefficient model is especially appealing. It allows a response variable to depend linearly on some regressors, with coefficients as smooth functions of some other predictor variables. The additive-linear structure enables simple interpretation and avoids the curse of dimensionality problem in high dimensional cases. Specifically, consider a multivariate regression model in which a sample $\{(Y_i, \mathbf{X}_i, \mathbf{T}_i)\}_{i=1}^n$ is drawn that satisfies

$$Y_i = m(\mathbf{X}_i, \mathbf{T}_i) + \sigma(\mathbf{X}_i, \mathbf{T}_i)\varepsilon_i, \tag{1.1}$$

where for the response variables Y_i and predictor vectors $\mathbf{X}_i$ and $\mathbf{T}_i$, m and σ^2 are the conditional mean and variance functions

$$m(\mathbf{X}_i, \mathbf{T}_i) = E(Y_i|\mathbf{X}_i, \mathbf{T}_i), \quad \sigma^2(\mathbf{X}_i, \mathbf{T}_i) = \text{var}(Y_i|\mathbf{X}_i, \mathbf{T}_i) \tag{1.2}$$

and $E(\varepsilon_i|\mathbf{X}_i, \mathbf{T}_i) = 0$, $\text{var}(\varepsilon_i|\mathbf{X}_i, \mathbf{T}_i) = 1$. For the varying coefficient model, the conditional mean takes the following form

$$m(\mathbf{X}, \mathbf{T}) = \sum_{l=1}^{d} \alpha_l(X_l) T_l \tag{1.3}$$

in which all tuning variables $X_l, l = 1, ..., d$ make up the vector $\mathbf{X}$, and all linear predictor variables $T_l, l = 1, ..., d$ are univariate and distinct.

Hastie & Tibshirani (1993) proposed a backfitting algorithm to estimate the varying coefficient functions $\{\alpha_l(x_l)\}_{1 \le l \le d}$, but gave no asymptotic justification of the algorithm. A somewhat restricted model, the functional coefficient model, was proposed in the time series context by Chen & Tsay (1993b) and later in the context of longitudinal data by Hoover, Rice, Wu & Yang (1998), in which all the tuning

variables $X_l, l = 1, ..., d$ are the same and univariate. For more recent developments of the functional coefficient model, see Cai, Fan & Yao (2000). In a different direction, Yang, Härdle, Park & Xue (2004) studied inference for model (1.3) when all the tuning variables $\{X_{il}\}_{1 \le l \le d}$ are univariate but have a joint d-dimensional density. This model breaks the restrictive nature of the functional coefficient model that all the tuning variables $X_l, l = 1, ..., d$ have to be equal. On the other hand, it requires that none of the tuning variables $X_l, l = 1, ..., d$ are equal. In this dissertation, we propose a more flexible additive coefficient model, which includes functional/varying coefficient models as special cases.

1.2 The Additive Coefficient Model

We propose the following additive coefficient model which has a more flexible form, namely

$$m(\mathbf{X}, \mathbf{T}) = \sum_{l=1}^{d_1} \alpha_l(\mathbf{X}) T_l, \quad \alpha_l(\mathbf{X}) = \sum_{s=1}^{d_2} \alpha_{ls}(X_s), \forall 1 \le l \le d_1, \tag{1.4}$$

in which the coefficient functions $\{\alpha_l(\mathbf{X})\}_{l=1}^{d_1}$ are additive functions of the tuning variables $\mathbf{X} = \left(X_1, \ldots, X_{d_2}\right)^T$. Note that without the additivity restriction on the coefficient functions $\{\alpha_l(\mathbf{X})\}_{l=1}^{d_1}$, model (1.4) would be a kind of functional coefficient model with a multivariate tuning variable $\mathbf{X}$ instead of a univariate one as in the existing literature. The additive structure is imposed on the coefficient functions $\{\alpha_l(\mathbf{X})\}_{l=1}^{d_1}$, so that inference can be made on them without the "curse of dimensionality".

To understand the flexibility of this model, we look at some of the models that are included as special cases:

1. When the dimension of $\mathbf{X}$ is 1 ($d_2 = 1$), (1.4) reduces to the functional coefficient model of Chen & Tsay (1993b).

2. When the linear regressor vector $\mathbf{T}$ is constant ($d_1 = 1$, and $T_1 \equiv 1$), (1.4) reduces to the additive model of Chen & Tsay (1993a), Hastie & Tibshirani (1990).

3. When for any fixed $l = 1, ..., d_1$, $\alpha_{ls}(x_s) \equiv 0$ for all but one $s = 1, ..., d_2$, (1.4) reduces to the varying coefficient model (1.3) of Hastie & Tibshirani (1993).

4. When $d_1 = d_2 = d$, and $\alpha_{ls}(x_s) \equiv 0$ for $l \neq s$, (1.4) reduces to the varying-coefficient model of Yang, Härdle, Park & Xue (2004).

The additive coefficient model is a useful nonparametric alternative to the parametric models. To gain some insight into it, consider the application of our estimation procedure to the quarterly West German real GNP data from January 1960 to December 1990. Denote this time series by $\{G_t\}_{t=1}^{124}$, where G_t is the real GNP in the t-th quarter (the first quarter being from January 1, 1960 to April 1, 1960, the 124-th quarter being from September 1, 1990 to December 1, 1990). Yang & Tschernig (2002) deseasonalized this series by removing the four seasonal means from the series $\log(G_{t+4}/G_{t+3}), t = 1, ..., 120$. Denote the transformed time series as $\{Y_t\}_{t=1}^{120}$. As the nonparametric alternative to the optimal linear autoregressive model selected by the

Bayesian Information Criterion (BIC),

$$Y_t = c_1 Y_{t-2} + c_2 Y_{t-4} + \varepsilon_t, \tag{1.5}$$

we have fitted the following additive coefficient model (details in subsection 4.2.1),

$$\begin{aligned} Y_t &= \{c_1 + \alpha_{11}(Y_{t-1}) + \alpha_{12}(Y_{t-8})\} Y_{t-2} \\ &\quad + \{c_2 + \alpha_{21}(Y_{t-1}) + \alpha_{22}(Y_{t-8})\} Y_{t-4} + \sigma\varepsilon_t. \end{aligned} \tag{1.6}$$

Using this model, we can efficiently take into account the phenomenon that the effect of Y_{t-2}, Y_{t-4} on Y_t vary with Y_{t-1}, Y_{t-8}. The efficiency is evidenced by its superior out-of-sample one-step prediction at each of the last ten quarters. The averaged squared prediction error (ASPE) is 0.000112 for the linear autoregressive fit in (1.5), and 0.000077 to 0.000085 for fits of the additive coefficient model (1.6). Hence the reduction in ASPE is between 31% and 46%, see Table 4.2.3. Figure 4.1 clearly illustrates this improvement in prediction power, in which circle denotes the observed value, and cross (triangle) denotes the predictions by linear autoregressive model (1.5), and additive coefficient model (1.6) respectively. One can see that the additive coefficient model out-performs the linear autoregressive model in prediction for 8 of the 10 quarters.

1.3 Model Identification

For the additive coefficient model, the regression function $m(\mathbf{X}, \mathbf{T})$ in (1.4) needs to be identified. One practical solution is to rewrite it as

$$m(\mathbf{X}, \mathbf{T}) = \sum_{l=1}^{d_1} \alpha_l(\mathbf{X}) T_l, \quad \alpha_l(\mathbf{X}) = \alpha_{l0} + \sum_{s=1}^{d_2} \alpha_{ls}(X_s), \quad \forall 1 \le l \le d_1, \tag{1.7}$$

with the identification conditions

$$E\{w(\mathbf{X})\alpha_{ls}(X_s)\} \equiv 0, \quad l = 1, ..., d_1, s = 1, ..., d_2, \tag{1.8}$$

for some nonnegative weight function w, with $E\{w(\mathbf{X})\} = 1$. The weight function w is introduced so that estimation of the unknown functions $\{\alpha_l(\mathbf{X})\}_{1 \leq l \leq d_1}$ will be carried out only on the support of w, supp (w), which is compact according to assumption (A7). This is important as most of the asymptotic results for nonparametric estimators are developed only for values over compact sets. By having this weight function, the support of the distribution of $\mathbf{X}$ is not required to be compact. This relaxation is very desirable since most time series distributions are not compactly supported. See Yang & Tschernig (2002), p.1414 for similar use of the weight function.

Note that (1.8) does not impose any restriction on the model, since any regression function $m(\mathbf{X}, \mathbf{T}) = \sum_{l=1}^{d_1} \sum_{s=1}^{d_2} \alpha_{ls}^*(X_s)T_l$ can be reorganized to satisfy (1.8), by writing

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

with $\alpha_{l0} = E\left\{w(\mathbf{X}) \sum_{s=1}^{d_2} \alpha_{ls}^*(X_s)\right\}$, $\alpha_{ls}(X_s) = \alpha_{ls}^*(X_s) - E\{w(\mathbf{X})\alpha_{ls}^*(X_s)\}$.

In addition, for the functions $\{\alpha_{ls}(X_s)\}_{1 \leq l \leq d_1}^{1 \leq s \leq d_2}$ and parameters $\{\alpha_{l0}\}_{1 \leq l \leq d_1}$ to be uniquely determined, one imposes an additional assumption.

(A0) There exists a constant $C > 0$ such that for any set of measurable functions $\{b_{ls}(X_s)\}_{1 \leq l \leq d_1}^{1 \leq s \leq d_2}$ that satisfy (1.8) and any set of constants $\{a_l\}_{1 \leq l \leq d_1}$,

the following holds

$$E\left[\sum_{l=1}^{d_1}\left\{a_l+\sum_{s=1}^{d_2}b_{ls}(X_s)\right\}T_l\right]^2 \tag{1.9}$$

$$\geq C\left[\sum_{l=1}^{d_1}a_l^2+\sum_{l=1}^{d_1}\sum_{s=1}^{d_2}E\left\{b_{ls}^2(X_s)\right\}\right]. \tag{1.10}$$

Lemma 1.3.1. *Under assumptions (A0) and (A5) in the subsection 2.5.1, the representation in (1.7) subject to (1.8) is unique.*

Proof. Suppose that

$$m(\mathbf{X},\mathbf{T})=\sum_{l=1}^{d_1}\left\{\alpha_{l0}+\sum_{s=1}^{d_2}\alpha_{ls}(X_s)\right\}T_l=\sum_{l=1}^{d_1}\left\{\widetilde{\alpha}_{l0}+\sum_{s=1}^{d_2}\widetilde{\alpha}_{ls}(X_s)\right\}T_l$$

with both the set $\{\alpha_{ls}(X_s)\}_{1=1,s=1}^{d_1,d_2}$, $\{\alpha_{l0}\}_{1=1}^{d_1}$ and the set $\{\widetilde{\alpha}_{ls}(X_s)\}_{l=1,s=1}^{d_1,d_2}$, $\{\widetilde{\alpha}_{l0}\}_{1=1}^{d_1}$ satisfying (1.8). Then upon defining for all s,l

$$b_{ls}(X_s)\equiv\widetilde{\alpha}_{ls}(X_s)-\alpha_{ls}(X_s),\quad a_l\equiv\widetilde{\alpha}_{l0}-\alpha_{l0}$$

one has $\sum_{l=1}^{d_1}\left\{a_l+\sum_{s=1}^{d_2}b_{ls}(X_s)\right\}T_l\equiv 0$. Hence by assumption (A0)

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

entailing that for all s,l, $a_l\equiv 0$ and $b_{ls}^2(X_s)\equiv 0$ almost surely. Since assumption (A5) requires that all X_s are continuous random variables, one has $b_{ls}(x)\equiv 0$ for all s,l. ■

1.4 Data Generating Process

In this dissertation, we consider $\{(Y_i,\mathbf{X}_i,\mathbf{T}_i)\}_{i=1}^n$, a sample generated from the regression model (1.1) and (1.2) with its conditional mean function described by (1.7)

and the identifiability conditions (1.8), (1.10). Furthermore its error terms $\{\varepsilon_i\}_{i=1}^n$ are assumed to be i.i.d with $E\varepsilon_i = 0, E\varepsilon_i^2 = 1$, and with the additional property that ε_i is independent of $\{(\mathbf{X}_j, \mathbf{T}_j), j \leq i\}, i = 1, ..., n$. With this error structure, the explanatory variable vector $(\mathbf{X}_i, \mathbf{T}_i)$ can contain exogenous variables and/or lag variables of Y_i. If $(\mathbf{X}_i, \mathbf{T}_i)$ contains only the lags of Y_i, it is a semi-parametric autoregressive time series model, which is a useful extension of many existing nonlinear time series models such as exponential autoregressive model (EXPAR), threshold autoregressive model (TAR), and functional autoregressive model (FAR), as well as the linear autoregressive model.

To obtain the asymptotics of the estimators proposed in this dissertation , we need some additional properties on the data generating process $\{\zeta_i\}_{i=1}^\infty = \{(Y_i, \mathbf{X}_i, \mathbf{T}_i)\}_{i=1}^\infty$. First we assume $\{\zeta_i\}_{i=1}^\infty$ is strictly stationary. The following definition of strict stationarity is from Brockwell & Davis (1991).

Definition 1.4.1. (Strict Stationarity) The series $\{\zeta_i\}_{i=1}^\infty$ is said to be strictly stationary if the joint distributions of $\left(\zeta_{t_1}, \ldots, \zeta_{t_k}\right)'$ and $\left(\zeta_{t_1+h}, \ldots, \zeta_{t_k+h}\right)'$ are the same for all positive integers h and $t_1, \ldots, t_k \in \mathcal{Z}^+$.

Second, we assume $\{\zeta_i\}_{i=1}^\infty$ is weakly dependent. Generally speaking, weak dependence allows the observation at time t to be dependent with the observations at the other times, say, $t + k$, but requires this dependence diminishes to zero as the observations are far apart, i.e. $|k| \to \infty$. There are several definitions of weak dependence (or mixing) when the dependence is measured by different mixing coefficients. Here we quote the definitions of two commonly used weak dependence, the so-called

α-mixing and β-mixing from Bosq (1998).

Definition 1.4.2. (α-mixing) Let $\{\zeta_i\}_{i=1}^{\infty} = \{(Y_i, \mathbf{X}_i, \mathbf{T}_i)\}_{i=1}^{\infty}$ be a strictly stationary vector process. Let $\mathcal{F}_{n+k}^{\infty}$ and $\mathcal{F}_0^n$ denote the σ-algebras generated by $\{\zeta_i, i \geq n+k\}$ and $\{\zeta_0, \ldots, \zeta_n\}$ separately. Then the α- coefficient which measures the correlation between $\mathcal{F}_{n+k}^{\infty}$ and $\mathcal{F}_0^n$ is given as

$$\alpha(k) = \sup_{A \in \mathcal{F}_0^n, B \in \mathcal{F}_{n+k}^{\infty}} |P(A)P(B) - P(AB)| .$$

The vector process $\{\zeta_i\}_{i=1}^{\infty}$ is α-mixing (or strongly mixing) if its α-coefficient $\alpha(k) \to 0$, as $|k| \to \infty$. Specially, the vector process $\{\zeta_i\}_{i=1}^{\infty}$ is geometrically α-mixing if its α-coefficient goes to 0 geometrically fast, i.e. $\alpha(k) \leq c\rho^k$, for some constants $c > 0$, $0 < \rho < 1$.

Definition 1.4.3. (β-mixing) Let $\{\zeta_i\}_{i=1}^{\infty} = \{(Y_i, \mathbf{X}_i, \mathbf{T}_i)\}_{i=1}^{\infty}$ be a strictly stationary vector process. Let $\mathcal{F}_{n+k}^{\infty}$ and $\mathcal{F}_0^n$ denote the σ-algebras generated by $\{\zeta_i, i \geq n+k\}$ and $\{\zeta_0, \ldots, \zeta_n\}$ separately. Then the β-coefficient which measures the correlation between $\mathcal{F}_{n+k}^{\infty}$ and $\mathcal{F}_0^n$ is given as

$$\beta(k) = \sup_{n \geq 1} E \sup_{A \in \mathcal{F}_{n+k}^{\infty}} |P(A|\mathcal{F}_0^n) - P(A)|$$

The vector process $\{\zeta_i\}_{i=1}^{\infty}$ is β-mixing if its β-coefficient $\beta \to 0$ as $|k| \to \infty$. Similarly, the vector process $\{\zeta_i\}_{i=1}^{\infty}$ is geometrically β-mixing if its β-coefficient goes to 0 geometrically fast, i.e. $\beta(k) \leq c\rho^k$, for some constants $c > 0$, $0 < \rho < 1$.

The $\beta-$mixing is stronger than $\alpha-$mixing, because the coefficients satisfy the inequality that

$$\alpha(k) \leq \beta(k)/2.$$

Both α- and β- mixing are weaker than the *m-dependence*, i.e., $\sigma\{\zeta_t, t \leq T\}$ and $\sigma\{\zeta_t, t \geq T+k\}$ are independent for all $k > m$. Most importantly, the α- and β-mixing contain the usual linear autoregressive and moving average (ARMA) models. For more discussions about mixing, see Bosq (1996).

The rest of the dissertation is organized as follows. In chapter 2, a local polynomial based marginal integration method is proposed to estimate the coefficient functions. The asymptotic normality is developed. In chapter 3, a fast polynomial spline estimation is developed for the estimation. Also a model selection procedure based on nonparametric Bayesian Information Criterion (BIC) is proposed for inference purpose. In chapter 4, two simulation studies are given to compare the numerical performances of two proposed estimation methods, and the model selection procedure. Also the proposed methods are successfully applied to three empirical examples.

Chapter 2

Marginal Integration Estimation

2.1 Introduction

The main focus of this dissertation is to estimate the additive coefficient model (1.7), in which for every $l = 1, ..., d_2$, the coefficient of T_{il} consists of two parts, the unknown parameter α_{l0}, and the unknown univariate functions $\{\alpha_{ls}\}_{1 \leq s \leq d_2}$. The first approach we propose is the local polynomial based marginal integration method.

The marginal integration method was first discussed in Linton & Nielsen (1995) in the context of additive models, see also the marginal integration method for generalized additive models in Linton & Härdle (1996). To see how the marginal integration method works in our context, observe that according to the identification condition (1.8), for every $l = 1, ..., d_1$ one has,

$$\alpha_{l0} = E\{w(\mathbf{X})\alpha_l(\mathbf{X})\} = \int w(\mathbf{x})\alpha_l(\mathbf{x})\varphi(\mathbf{x})d\mathbf{x} \tag{2.1}$$

and for every point $\mathbf{x} = (x_1, ..., x_{d_2})^T$, and every $l = 1, ..., d_1, s = 1, ..., d_2$, one has,

$$\begin{aligned} \alpha_{l0} + \alpha_{ls}(x_s) &= E\left\{w_{-s}(\mathbf{X}_{-s})\alpha_l(x_s, \mathbf{X}_{-s})\right\} \\ &= \int w_{-s}(\mathbf{u}_{-s})\alpha_l(x_s, \mathbf{u}_{-s})\varphi_{-s}(\mathbf{u}_{-s})d\mathbf{u}_{-s} \end{aligned} \tag{2.2}$$

where

$$\begin{aligned} \mathbf{u}_{-s} &= (u_1, \ldots, u_s - 1, u_s + 1, \ldots, u_{d_2})^T, \\ (x_s, \mathbf{u}_{-s}) &= \left(u_1, \ldots, u_s - 1, x_s, u_s + 1, \ldots, u_{d_2}\right)^T, \end{aligned}$$

the density of $\mathbf{X}$ is φ, and the marginal density of

$$\mathbf{X}_{-s} = (X_1, \ldots, X_s - 1, X_s + 1, \ldots, X_{d_2})^T$$

is φ_{-s}, and $w_{-s}(\mathbf{x}_{-s}) = E\left\{w(X_s, \mathbf{x}_{-s})\right\} = \int w(u, \mathbf{x}_{-s})d\varphi_s(u)$. In addition, the marginal density of X_s is denoted by φ_s. Intuitively, one has

$$\alpha_{l0} \approx \frac{\sum_{i=1}^{n} w(\mathbf{X}_i)\alpha_l(\mathbf{X}_i)}{\sum_{i=1}^{n} w(\mathbf{X}_i)}, \qquad \alpha_{ls}(x_s) \approx \frac{\sum_{i=1}^{n} w_s(\mathbf{X}_{i,-s})\alpha_l(x_s, \mathbf{X}_{i,-s})}{\sum_{i=1}^{n} w_s(\mathbf{X}_{i,-s})} - \alpha_{l0} \tag{2.3}$$

and the d_2-dimensional functions $\{\alpha_l(\mathbf{x})\}_{l=1}^{d_1}$ in the above equations (2.3) can be replaced by the usual local polynomial estimators. This is the essential idea behind the marginal integration method. To gain more insight of it, we consider the following simple example.

Example: Suppose we have the data generating from the simple additive coefficient model

$$Y = \{2 + \sin(X_1) + X_2\}T_1 + \{1 + \sin(X_2)\}T_2 + \varepsilon,$$

where independent of each other, X_1 and X_2, follow $U[-\pi, \pi]$, and T_1, T_2 follow $N(0, 1)$ and ε is the normal noise term. In this case, we take w to be the identity

function. Denote $\alpha_1(\mathbf{X}) = 2 + \sin(X_1) + X_2$, and $\alpha_2(\mathbf{X}) = 1 + \sin(X_1)$. Then simple calculation shows that

$$E(\alpha_1(\mathbf{X})) = 2; \quad E(\alpha_2(\mathbf{X})) = 1$$

$$E(\alpha_1(x_1, X_2)) = 2 + \sin(x_1); \quad E(\alpha_2(x_1, X_2)) = 1 + \sin(x_1)$$

$$E(\alpha_2(X_1, x_2)) = 2 + x_2; \quad E(\alpha_2(X_1, x_2)) = 1.$$

We will discuss the same example in the simulation study.

2.2 Estimators of constants

According to the first approximation equation in (2.3), to estimate the constants $\{\alpha_{l0}\}_{l=1}^{d_1}$, we first estimate the unknown functions $\{\alpha_l(\mathbf{x})\}_{l=1}^{d_1}$ at those data points $\mathbf{X}_i$ that are in the support of the weight function w. More generally, for any fixed $\mathbf{x} \in$ supp (w), we approximate $\alpha_l(\mathbf{x})$ locally by a constant α_l, and estimate $\{\alpha_l(\mathbf{x})\}_{l=1}^{d_1}$ by minimizing the following weighted sum of squares with respect to $\boldsymbol{\alpha} = \left(\alpha_1, \ldots, \alpha_{d_1}\right)^T$,

$$\sum_{i=1}^{n} \left\{ Y_i - \sum_{l=1}^{d_1} \alpha_l T_{il} \right\}^2 K_H(\mathbf{X}_i - \mathbf{x}), \tag{2.4}$$

where K is a d_2-variate kernel function of order q_1, see assumption (A1), $H = \text{diag}\left\{h_{0,1}, \ldots, h_{0,d_2}\right\}$ is a diagonal matrix of positive numbers $h_{0,1}, \ldots, h_{0,d_2}$, called bandwidths, and

$$K_H(\mathbf{x}) = \frac{1}{\prod_{s=1}^{d_2} h_{0,s}} K\left(\frac{x_1}{h_{0,1}}, \ldots, \frac{x_{d_2}}{h_{0,d_2}}\right).$$

Let $\hat{\boldsymbol{\alpha}} = \left(\hat{\alpha}_1, \ldots, \hat{\alpha}_{d_1}\right)^T$ be the solution to the least squares problem in (2.4).

Note that $\hat{\alpha}$ is dependent on $\mathbf{x}$, as is (2.4), and the components in $\hat{\alpha}$ give the estimators for $\{\alpha_l(\mathbf{x})\}_{l=1}^{d_1}$. To emphasize the dependence on $\mathbf{x}$, we write $\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\alpha}}(\mathbf{x}) = \left(\hat{\alpha}_1(\mathbf{x}), \ldots, \hat{\alpha}_{d_1}(\mathbf{x})\right)^T$. More precisely, let

$$\mathbf{W}(\mathbf{x}) = \operatorname{diag}\left\{K_H(\mathbf{X}_i - \mathbf{x})/n\right\}_{1 \le i \le n}, \quad \mathbf{Z} = \begin{pmatrix} T_{11} & \cdots & T_{1d_1} \\ \vdots & \ddots & \vdots \\ T_{n1} & \cdots & T_{nd_1} \end{pmatrix}, \quad \mathbf{Y} = (Y_1, \ldots, Y_n)^T$$

and e_l be a d_1-dimensional vector with all entries 0 except the l-th entry being 1. Then $\{\hat{\alpha}_l(\mathbf{x})\}_1 \le l \le d_1$ is given by

$$\hat{\alpha}_l(\mathbf{x}) = e_l^T \left\{\mathbf{Z}^T\mathbf{W}(\mathbf{x})\mathbf{Z}\right\}^{-1} \mathbf{Z}^T\mathbf{W}(\mathbf{x})\mathbf{Y}. \tag{2.5}$$

By (2.3), the parameter α_{l0} can be estimated as a weighted average of $\hat{\alpha}_l(\mathbf{X}_i)$'s, i.e.,

$$\hat{\alpha}_{l0} = \frac{\sum_{i=1}^n w(\mathbf{X}_i)\hat{\alpha}_l(\mathbf{X}_i)}{\sum_{i=1}^n w(\mathbf{X}_i)}, \quad l = 1, \ldots, d_1. \tag{2.6}$$

Theorem 2.2.1. *Under assumptions (A1)-(A7) in subsection 2.5.1, for any* $l = 1, ..., d_1$,

$$\sqrt{n}(\hat{\alpha}_{l0} - \alpha_{l0}) \xrightarrow{\mathcal{L}} N\left\{0, \sigma_l^2\right\},$$

where the asymptotic variance σ_l^2 *is defined in (2.20).*

The rate of $1/\sqrt{n}$ at which $\hat{\alpha}_{l0}$ converges to α_{l0} is due to two special features of $\hat{\alpha}_l(\mathbf{x})$. First, the bias of $\hat{\alpha}_l(\mathbf{x})$ in estimating $\alpha_l(\mathbf{x})$ consists of terms of order $h_{0,1}^{q_1}, \ldots, h_{0,d_2}^{q_1}$, bounded by $1/\sqrt{n}$ according to assumption (A6) (a), see the derivation of Lemma 2.5.5 about the term P_{c1}. Second, the usual variance of $\hat{\alpha}_l(\mathbf{x})$ in estimating $\alpha_l(\mathbf{x})$ is proportional to $n^{-1}h_{0,1}^{-1} \cdots h_{0,d_2}^{-1}$, which gets reduced to $1/n$ due

to the effect of averaging in (2.6), see the derivation of the term P_{c2} in (2.18) and (2.19). This technique of simultaneously reducing the bias by the use of a higher order kernel and "integrating out the variance "is the common feature of all marginal integration procedures.

2.3 Estimators of function components

In the following, we illustrate a procedure for estimating the functions $\{\alpha_{ls}(x_s)\}_{l=1}^{d_1}$, for any fixed $s = 1, \ldots, d_1$. Let x_s be a point at which we want to evaluate the functions $\{\alpha_{ls}(x_s)\}_1 \leq l \leq d_1$. According to (2.3), we need to estimate $\{\alpha_l(\mathbf{x})\}_{l=1}^{d_1}$ at those points $(x_s, \mathbf{X}_{i,-s})$ that lie in the support of w. For any $\mathbf{x} \in \text{supp}(w)$, differently from estimating the constants, we approximate the function $\alpha_l(\mathbf{u})$ locally at $\mathbf{x}$ by $\alpha_l(\mathbf{u}) \approx \alpha_l + \sum_{j=1}^{p} \beta_{lj}(u_s - x_s)^j$, and estimate $\{\alpha_l(\mathbf{x})\}_{l=1}^{d_1}$ by minimizing the following weighted sum of squares with respect to $\boldsymbol{\alpha} = \left(\alpha_1, \ldots, \alpha_{d_1}\right)^T$, $\boldsymbol{\beta} = \left(\beta_{11}, \ldots, \beta_{1p}, \ldots, \beta_{d_1 1}, \ldots, \beta_{d_1 p}\right)^T$

$$\sum_{i=1}^{n} \left[Y_i - \sum_{l=1}^{d_1} \left\{ \alpha_l + \sum_{j=1}^{p} \beta_{lj}(X_{is} - x_s)^j \right\} T_{il} \right]^2 k_{h_s}(X_{is} - x_s) L_{G_s}(\mathbf{X}_{i,-s} - \mathbf{x}_{-s})$$

in which k is a univariate kernel, L is a $(d_2 - 1)$-variate kernel of order q_2, as in assumption (A1) in the subsection 2.5.1, the bandwidth matrix

$$G_s = \text{diag}\left\{ g_1, \ldots, g_S - 1, g_S + 1, \ldots, g_{d_2} \right\},$$

and

$$k_{h_s}(u_s) = \frac{1}{h_s} k\left(\frac{u_s}{h_s}\right),$$

$$L_{G_s}(\mathbf{u}_{-s}) = \frac{1}{\prod_{1 \le s' \le d_2, s' \ne s} g_{s'}} L\left(\frac{u_1}{g_1}, ..., \frac{u_{s-1}}{g_{s-1}}, \frac{u_{s+1}}{g_{s+1}}, ..., \frac{u_{d_2}}{g_{d_2}}\right)$$

for $\mathbf{u}_{-s} = \left(u_1, ..., u_S - 1, u_S + 1, ..., u_{d_2}\right)$. Let $\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}$ be the solution of the above least squares problem. Then the components in $\hat{\boldsymbol{\alpha}}$ give the estimators for $\{\alpha_l(\mathbf{x})\}_{l=1}^{d_1}$, which is given by

$$\hat{\alpha}_l(\mathbf{x}) = e_l^T \left\{\mathbf{Z}_s^T \mathbf{W}_s(\mathbf{x}) \mathbf{Z}_s\right\}^{-1} \mathbf{Z}_s^T \mathbf{W}_s(\mathbf{x}) \mathbf{Y}, \tag{2.7}$$

where e_l is a $(p+1)d_1$-dimensional vector with all entries 0 except the l-th entry being 1,

$$\mathbf{W}_s(\mathbf{x}) \equiv \operatorname{diag}\left\{n^{-1} k_{h_s}(X_{is} - x_s) L_{G_s}(\mathbf{X}_{i,-s} - \mathbf{x}_{-s})\right\}_{1 \le i \le n}$$

and

$$\begin{aligned} \mathbf{Z}_s &= \begin{bmatrix} \mathbf{T}_1^T, \{(X_{1s} - x_s)/h_s\} \mathbf{T}_1^T, ..., \{(X_{1s} - x_s)/h_s\}^p \mathbf{T}_1^T \\ \vdots \\ \mathbf{T}_n^T, \{(X_{ns} - x_s)/h_s\} \mathbf{T}_n^T, ..., \{(X_{ns} - x_s)/h_s\}^p \mathbf{T}_n^T \end{bmatrix} \\ &= \begin{bmatrix} [p\{(X_{1s} - x_s)/h_s\}]^T \otimes \mathbf{T}_1^T \\ \vdots \\ [p\{(X_{ns} - x_s)/h_s\}]^T \otimes \mathbf{T}_n^T \end{bmatrix} \end{aligned} \tag{2.8}$$

in which $p(u) = (1, u, \ldots, u^p)^T$ and $\otimes$ denotes the Kronecker product of matrices. Then for each s, we can construct the marginal integration estimators of α_{ls} for $l = 1, \ldots, d_1$ simultaneously, which are given by

$$\hat{\alpha}_{ls}(x_s) = \frac{\sum_{i=1}^n w_{-s}(\mathbf{X}_{i,-s}) \hat{\alpha}_l(x_s, \mathbf{X}_{i,-s})}{\sum_{i=1}^n w_{-s}(\mathbf{X}_{i,-s})} - \hat{\alpha}_{l0}. \tag{2.9}$$

where the term $\hat{\alpha}_{l0}$ is the $\sqrt{n}$-consistent estimator of α_{l0} in Theorem 2.2.1. The estimator $\hat{\alpha}_{ls}(x_s)$ is referred to as the p-th order local polynomial estimator, where p is the highest polynomial degree of variables $X_{is} - x_s$, $i = 1, ..., n$, in the definition of the design matrix $\mathbf{Z}_s$ in (2.8). In particular, the local linear ($p = 1$) and the local cubic estimators ($p = 3$) are the most commonly used.

Theorem 2.3.1. *Under assumptions A1-A7 in the subsection 2.5.1, one has, for any* $\mathbf{x} = (x_1, ..., x_{d_2})^T \in \text{supp}(w)$, *and* $l = 1, ..., d_1$, $s = 1, ..., d_2$,

$$\sqrt{nh_s}\left\{\hat{\alpha}_{ls}(x_s) - \alpha_{ls}(x_s) - h_s^{p+1}\eta_{ls}(x_s)\right\} \xrightarrow{\mathcal{L}} N\left\{0, \sigma_{ls}^2(x_s)\right\}, \tag{2.10}$$

where $\eta_{ls}(x_s)$ *and* $\sigma_{ls}^2(x_s)$ *are defined in (2.28) and (2.30), respectively.*

Finally, based on (2.6) and (2.9), one can predict Y given any realization $(\mathbf{x}, \mathbf{t})$ of $(\mathbf{X}, \mathbf{T})$ by the predictor

$$\hat{m}(\mathbf{x}, \mathbf{t}) = \sum_{l=1}^{d_1}\left\{\hat{\alpha}_{l0} + \sum_{s=1}^{d_2}\hat{\alpha}_{ls}(x_s)\right\}t_l. \tag{2.11}$$

To appreciate why α_{ls} can be estimated by $\hat{\alpha}_{ls}$ at the rate of $1/\sqrt{nh_s}$, which is the same as the rate of estimating a nonparametric function in the univariate case, we discuss two special features of $\hat{\alpha}_l(\mathbf{x})$ given in (2.7), which are similar to those discussed in subsection (2.2). First the bias of $\hat{\alpha}_l(\mathbf{x})$ in estimating $\alpha_l(\mathbf{x})$ is of order $h_s^{p+1} + g_{\max}^{q_2}$, where the first term can be understood as the approximation bias caused by locally approximating α_{ls} using a p-th degree polynomial, see the derivation of P_{s2} in Lemma 2.5.9, and the second term can be considered as the approximation bias by locally approximating functions $\{\alpha_{ls'}\}_{s' \neq s}$ using a constant, which is bounded by $g_{\max}^{q_2}$ since the kernel L is of order q_2, see P_{s3} in Lemma 2.5.9. The order $g_{\max}^{q_2}$ of the second

bias term is negligible compared to the rescaling factor of order $1/\sqrt{nh_s}$, according to (A6) (b). Hence, only the first bias term appears in the asymptotic distribution formula (2.10). As for the variance of $\hat{\alpha}_l(\mathbf{x})$ in estimating $\alpha_l(\mathbf{x})$, it is proportional to $n^{-1}h_s^{-1}g_1^{-1}\cdots g_{s-1}^{-1}g_{s+1}^{-1}\cdots g_{d_2}^{-1}$, but due to marginal averaging of variables $\mathbf{X}_{i,-s}$, the bandwidths $g_1,...,g_S-1,g_S+1,...,g_{d_2}$ related to $\mathbf{X}_{i,-s}$ are integrated out, see P_{s1} in Lemma 2.5.9. Then the variance of $\hat{\alpha}_{ls}$ is reduced to the order $n^{-1}h_s^{-1}$. If the same bandwidth h_s is used for all variable directions in $\mathbf{X}$, then Assumption (A6) (b) would imply that $nn^{-d_2/(2p+3)} \to \infty$ and hence restricting d_2 to be less than $2p+3$, for the asymptotic results of Theorem 2.3.1 to be true. That is why we prefer the flexibility of using a set of bandwidths $g_1,...,g_S-1,g_S+1,...,g_{d_2}$ different from h_s.

2.4 Implementation

Practical implementation of the estimators defined in (2.6) and (2.9) requires a rather intelligent choice of bandwidths $H = \text{diag}\left\{h_{0,1},...,h_{0,d_2}\right\}$, $\{h_s\}1 \leq s \leq d_2$, and $G_s = \text{diag}\left\{g_1,...,g_S-1,g_S+1,...,g_{d_2}\right\}$. In the following, we discuss the choices of such bandwidths.

- Note from Theorem 2.2.1 that the asymptotic distributions of the estimators $\{\hat{\alpha}_{l0}\}_{l=1}^{d_1}$ depend only on the quantity σ_l^2, not on the bandwidths in H. Hence we have only specified that H satisfy the order assumptions in (A6) (a) by taking $\hat{h}_{01} = \cdots = \hat{h}_{0d_2} = \sqrt{\hat{\text{var}}(\mathbf{X})}\log(n)\,n^{-1/(2q_1-1)}$, where q_1 is the order of the kernel K, required to be greater than $(d_2+1)/2$, and $\hat{\text{var}}(\mathbf{X}) =$

$\left\{ \prod_{s=1}^{d_2} \hat{\text{var}}\,(X_s) \right\}^{1/d_2}$, in which $\hat{\text{var}}\,(X_s)$ denotes the sample variance of $X_s, s = 1, ..., d_2$.

- The asymptotic distributions of the estimators $\{\hat{\alpha}_{ls}\}_{1 \le l \le d_1}^{1 \le s \le d_2}$ depend not only on the functions $\eta_{ls}(x_s)$ and $\sigma_{ls}^2(x_s)$ but also crucially on the choice of bandwidths h_s. Moreover, for each $s = 1, ..., d_2$, the coefficient functions $\{\alpha_{ls}(x_s)\}_{l=1}^{d_1}$ are estimated simultaneously. So we define the optimal bandwidth of h_s, denoted by $h_{s,\text{opt}}$, as the minimizer of the total asymptotic mean integrated squared errors of $\{\hat{\alpha}_{ls}(x_s), l = 1, \ldots, d_1\}$, which is defined as

$$\sum_{l=1}^{d_1} \text{AMISE}\,\{\hat{\alpha}_{ls}\} = h_s^{2(p+1)} \sum_{l=1}^{d_1} \int \eta_{ls}^2(x_s) dx_s + \frac{1}{nh_s} \sum_{l=1}^{d_1} \int \sigma_{ls}^2(x_s) dx_s.$$

Then $h_{s,\text{opt}}$ is found to be

$$h_{s,\text{opt}} = \left\{ \frac{\sum_{l=1}^{d_1} \int \sigma_{ls}^2\,(x_s)\, dx_s}{2n(p+1) \sum_{l=1}^{d_1} \int \eta_{ls}^2\,(x_s)\, dx_s} \right\}^{1/(2p+3)}$$

in which $\eta_{ls}\,(x_s)$ and $\sigma_{ls}^2\,(x_s)$ are the asymptotic bias and variance of $\hat{\alpha}_{ls}$ as in (2.28) and (2.30). According to the definitions of $\eta_{ls}\,(x_s)$, $\sigma_{ls}^2\,(x_s)$, $\int \eta_{ls}^2\,(x_s)\, dx_s$ and $\int \sigma_{ls}^2\,(x_s)\, dx_s$ can be approximated respectively by

$$\int \left[\frac{1}{(p+1)!} \sum_{l'=1}^{d_1} \alpha_{l's}^{(p+1)}\,(x_s) \int u^{p+1} \frac{1}{n} \sum_{i=1}^{n} \left\{ w_{-s}\,(\mathbf{X}_{i,-s})\, T_{il'} K_{ls}^*\,(u, x_s, \mathbf{X}_{i,-s}, \mathbf{T}_i) \right\} du \right]^2 dx_s,$$

$$\frac{1}{n} \sum_{i=1}^{n} \frac{w_{-s}^2\,(X_{i,-s})\, \varphi_{-s}^2\,(\mathbf{X}_{i,-s})\, \sigma^2\,(\mathbf{X}_i, \mathbf{T}_i)}{\varphi^2\,(\mathbf{X}_i)} \int K_{ls}^{*2}\,(u, \mathbf{X}_i, \mathbf{T}_i)\, du,$$

where the functions K_{ls}^* are defined in (2.29).

To implement this, one needs to evaluate terms such as $\alpha_{l's}^{(p+1)}\,(x_s)$, $\sigma^2(\mathbf{x}, \mathbf{t})$, $\varphi(\mathbf{x})$, $\varphi(\mathbf{x}_{-s})$ and K_{ls}^*. We propose the following simple estimation methods for those quantities. The resulting bandwidth is denoted as $\hat{h}_{s,\text{opt}}$.

1. The derivative functions $\alpha_{l's}^{(p+1)}(x_s)$ are estimated by fitting a polynomial regression model of degree $p+2$

$$E(Y|\mathbf{X},\mathbf{T}) = \sum_{l=1}^{d_1}\sum_{s=1}^{d_2}\sum_{k=0}^{p+2} a_{ls,k} X_s^k T_l.$$

Then $\alpha_{l's}^{(p+1)}(x_s)$ is estimated as $(p+1)!a_{l's,p+1} + (p+2)!a_{l's,p+2}x_s$. As a by-product, the mean squared error of this model, is used as an estimate of $\sigma^2(\mathbf{x})$.

2. Density functions $\varphi(\mathbf{x})$ and $\varphi(\mathbf{x}_{-s})$, are estimated as

$$\hat{\varphi}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\Pi_{s=1}^{d_2}\frac{1}{h(\mathbf{X},d_2)}\phi\left\{\frac{X_{is}-x_s}{h(\mathbf{X},d_2)}\right\},$$
$$\hat{\varphi}_{-s}(\mathbf{x}_{-s}) = \frac{1}{n}\sum_{i=1}^{n}\Pi_{s'\neq s}\frac{1}{h(\mathbf{X}_{-s},d_2-1)}\phi\left\{\frac{X_{is'}-x_{s'}}{h(\mathbf{X}_{-s},d_2-1)}\right\}$$

with the standard normal density ϕ and the rule-of-the-thumb bandwidth

$$h(\mathbf{X},m) = \sqrt{\hat{\text{var}}(\mathbf{X})}\{4/(m+2)\}^{1/(m+4)}n^{-1/(m+4)}.$$

3. According to the definition in (2.29), the dependence of the functions $K_{ls}^*(u,\mathbf{x},\mathbf{t})$ on u and $\mathbf{t}$ is explicitly known. The only unknown term $E(\mathbf{T}\mathbf{T}^T|\mathbf{X}=\mathbf{x})$ contained in $S_\alpha^{-1}(\mathbf{x})$ is estimated by fitting matrix polynomial regression

$$E(\mathbf{T}\mathbf{T}^T|\mathbf{X}=\mathbf{x}) = \mathbf{c} + \sum_{s=1}^{d_2}\sum_{k=1}^{p}\mathbf{c}_{s,k}x_s^k$$

in which the coefficients $\mathbf{c},\mathbf{c}_{s,k}$ are $d_1\times d_1$ matrices.

In this procedure, one simply uses polynomial regression to estimate some of the unknown quantities, which is easy to implement, but may lead the estimated

optimal bandwidths to be biased relative to the true optimal bandwidths. The development of a more sophisticated bandwidth selection method requires further investigation.

- Since Theorem 2.3.1 implies that the asymptotic distributions of the estimators $\{\hat{\alpha}_{ls}\}_{1 \leq l \leq d_1}^{1 \leq s \leq d_2}$ do not depend on $\{G_s\}_{s=1}^{d_2}$, we only specify that the G_s satisfies the order assumption in (A6) (b) $g_1 = \cdots = g_{s-1} = g_{s+1} = \cdots = g_{d_2} = \hat{h}_{s,\text{opt}}^{(p+1)/q_2} / \log(n)$, in which q_2, the order of the kernel function L, is required to be greater than $(d_2 - 1)/2$, and $\hat{h}_{s,\text{opt}}$ is the optimal bandwidth obtained using the above procedure.

Following the above discussion, the order of the kernels K and L are required to be greater than $(d_2 + 1)/2$ and $(d_2 - 1)/2$ respectively. If the dimension of X equals to 2, kernels K and L can have order 2. We have used the quadratic kernel $k(u) = \frac{15}{16}(1 - u^2)^2 1_{\{|u| \leq 1\}}$, where $1_{\{|u| \leq 1\}}$ is the indicator function of $[-1, 1]$ and the kernels K, L are product kernels.

Lastly, the matrix $\mathbf{Z}^T\mathbf{W}(\mathbf{x})\mathbf{Z}$ in (2.4) is computed as $\mathbf{Z}^T\mathbf{W}(\mathbf{x})\mathbf{Z} + n^{-1}\mathbf{T}\mathbf{T}^T$, and the matrix $\mathbf{Z}_s^T\mathbf{W}_s(\mathbf{x})\mathbf{Z}_s$ in (2.7) as

$$\mathbf{Z}_s^T\mathbf{W}_s(\mathbf{x})\mathbf{Z}_s + \left(n\hat{h}_{s,\text{opt}}\right)^{-1} \sqrt{\hat{\text{var}}(\mathbf{X})} \left\{ \int k(u)p(u)p(u)^T du \right\} \bigotimes \mathbf{T}\mathbf{T}^T,$$

following the ridge regression idea of Seifert & Gasser (1996).

2.5 Assumption and proofs

2.5.1 Assumptions

We have listed below some assumptions necessary for proving Theorems 2.2.1 and 2.3.1. Throughout this subsection, we denote by the same letters c, C etc., any positive constants, without distinction in each case.

(A1) *The kernel functions k, K and L are symmetric, Lipschitz continuous and compactly supported. The function k is a univariate probability density function, while K is d_2 variate, and of order q_1,* i.e. $\int K(\mathbf{u})d\mathbf{u} = 1$ *while*

$$\int K(\mathbf{u})u_1^{r_1}\cdots u_{d_2}^{r_{d_2}}d\mathbf{u} = 0,$$

for $1 \leq r_1 + \cdots + r_{d_2} \leq q_1 - 1$. *Kernel L is $(d_2 - 1)$ variate and of order q_2.*

Denote $p^* = \max(p+1, q_1, q_2)$. Then we assume further that

(A2) *The functions $\alpha_{ls}(x_s)$ have bounded continuous p^*-th derivatives for $1 \leq l \leq d_1, 1 \leq s \leq d_2$.*

(A3) *The vector process $\{\zeta_i\}_{i=1}^{\infty} = \{(Y_i, \mathbf{X}_i, \mathbf{T}_i)\}_{i=1}^{\infty}$ is strictly stationary and* geometrically *β-mixing.*

According to (1.1) of Bosq (1998), the strong mixing coefficient $\alpha(k) \leq \beta(k)/2$, hence

$$\alpha(k) \leq c\rho^k/2. \tag{2.12}$$

(A4) *The error term satisfies:*

(a) The innovations $\{\varepsilon_i\}_{i=1}^{\infty}$ are i.i.d with $E\varepsilon_i = 0, E\varepsilon_i^2 = 1$ and $E|\varepsilon_i|^{2+\delta} < +\infty$ for some $\delta > 0$. Also, the term ε_i is independent of $\{(\mathbf{X}_j, \mathbf{T}_j), j \leq i\}$ for all $i > 1$.

(b) The conditional standard deviation function $\sigma(\mathbf{x}, \mathbf{t})$ is bounded and Lipschitz continuous.

(A5) *The vector* $(\mathbf{X}, \mathbf{T})$ *has a joint probability density* $\psi(\mathbf{x}, \mathbf{t})$. *The marginal densities of* $\mathbf{X}$, X_s *and* $\mathbf{X}_{-s}$ *are denoted by* φ, φ_s *and* φ_{-s} *respectively.*

(a) Letting $q^* = \max(q_1, q_2) - 1$, we assume that $\psi(\mathbf{x}, \mathbf{t})$ has bounded continuous q^*-th partial derivatives with respect to $\mathbf{x}$. And the marginal density φ is bounded away from zero on the support of the weight function w.

(b) Let $S(\mathbf{x}) = E\left(\mathbf{T}\mathbf{T}^T | \mathbf{X} = \mathbf{x}\right)$. We assume there exists a $c > 0$, such that $S(\mathbf{x}) \geq c\mathbf{I}_{d_2}$ uniformly for $\mathbf{x} \in \text{supp}(w)$. Here $\mathbf{I}_{d_2}$ is the $d_2 \times d_2$ identity matrix.

(c) The random matrix $\mathbf{T}\mathbf{T}^T$ satisfies the Cramer's moment condition, i.e, there exists a positive constant c, such that $E|T_l T_{l'}|^k \leq c^{k-2} k! E|T_l T_{l'}|^2$, and $E|T_l T_{l'}|^2 \leq c$ holds uniformly for $k = 3, 4, \ldots$, and $1 \leq l, l' \leq d_1$.

(A6) *The bandwidths satisfy*:

(a) For $H = \text{diag}\{h_{01}, \ldots, h_{0d_2}\}$ in Theorem 2.2.1, $\sqrt{n} h_{\max}^{q_1} \to 0$ and $n h_{\text{prod}} \propto n^{\alpha}$ for some $\alpha > 0$, where $h_{\max} = \max\{h_{01}, \ldots, h_{0d_2}\}$, $h_{\text{prod}} = \prod_{i=1}^{d_2} h_{0i}$, and $\propto$ means proportional to.

(b) For the bandwidths h_s, and $G_s = \text{diag}\left\{g_1, \ldots, g_S - 1, g_S + 1, \ldots, g_{d_2} - 1\right\}$ of Theorem 2.3.1, $h_s = O\left\{n^{-1/(2p+3)}\right\}$, $nh_s g_{\text{prod}} \propto n^{\alpha}$ for some $\alpha > 0$ and $(nh_s \ln n)^{1/2} g_{\max}^{q_2} \to 0$, where $g_{\max} = \max\left\{g_1, \ldots, g_{s-1}, g_{s+1}, \ldots, g_{d_2} - 1\right\}$, $g_{\text{prod}} = \prod_{s' \neq s} g_{s'}$.

(A7) *The weight function w is nonnegative, has compact support with nonempty interior, and is Lipschitz continuous on its support.*

2.5.2 Technical lemmas

The proof of many results in this dissertation makes use of some inequalities about U-statistics and von Mises' statistics of dependent variables derived from Yoshihara (1976). In general, let $\xi_i, 1 \leq i \leq n$ denote a strictly stationary sequence of random variables with values in R^d and β-mixing coefficients $\beta(k), k = 1, 2, ...$, and r a fixed positive integer. Let $\{\theta_n(F)\}$ denote the functionals of the distribution function F of ξ_i

$$\theta_n(F) = \int g_n(x_1, ..., x_m)\, dF(x_1) \cdots dF(x_m),$$

where $\{g_n\}$ are measurable functions symmetric in their m arguments such that

$$\int |g_n(x_1, ..., x_m)|^{2+\delta}\, dF(x_1) \cdots dF(x_m) \leq M_n < +\infty.$$

$$\sup_{(i_1, ..., i_m) \in S_c} \int |g_n(x_1, ..., x_m)|^{2+\delta}\, dF_{\xi_{i_1}, ..., \xi_{i_m}}(x_1, ..., x_m) \leq M_{n,c} < +\infty,$$

for some $\delta > 0$, where $S_c = \{(i_1, ..., i_m) | \#_r(i_1, ..., i_m) = c\}, c = 0, ..., m - 1$ and for every $(i_1, ..., i_m), 1 \leq i_1 \leq \cdots \leq i_m \leq n$, $\#_r(i_1, ..., i_m) =$ the number of $j =$

$1, ..., m-1$ satisfying $i_{j+1} - i_j < r$. Clearly, the cardinality of each set S_c is less than n^{m-c}.

The von Mises' differentiable statistic and the U-statistic

$$\begin{aligned} \theta_n(F_n) &= \int g_n(x_1, ..., x_m)\, dF_n(x_1) \cdots dF_n(x_m) \\ &= \frac{1}{n^m} \sum_{i_1=1}^{n} \cdots \sum_{i_m=1}^{n} g_n(\xi_{i_1}, ..., \xi_{i_m}), \\ U_n &= \frac{1}{\binom{n}{m}} \sum_{1 \le i_1 < \cdots < i_m \le n} g_n(\xi_{i_1}, ..., \xi_{i_m}) \end{aligned}$$

allow decompositions as

$$\theta_n(F_n) = \theta_n(F) + \sum_{c=1}^{m} \binom{m}{c} V_n^{(c)}, V_n^{(c)} = \int g_{n,c}(x_1, ..., x_c) \prod_{j=1}^{c} [dF_n(x_j) - dF(x_j)],$$

$$\begin{aligned} U_n &= \theta_n(F) + \sum_{c=1}^{m} \binom{m}{c} U_n^{(c)}, \\ U_n^{(c)} &= \frac{(n-c)!}{n!} \sum_{1 \le i_1 < \cdots < i_c \le n} \int g_{n,c}\left(x_{i_1}, ..., x_{i_c}\right) \times \\ &\quad \prod_{j=1}^{c} \left[dI_{R_+^d}(x_j - \xi_{i_j}) - dF(x_j) \right], \end{aligned}$$

where $g_{n,c}$ are the projections of g_n

$$g_{n,c}(x_1, ..., x_c) = \int g_n(x_1, ..., x_m)\, dF(x_{c+1}) \cdots dF(x_m), c = 0, 1, ..., m,$$

so that $g_{n,0} = \theta_n(F), g_n = g_{n,m}$ and $I_{R_+^d}$ is the indicator function of the nonnegative part of R^d, $R_+^d = \{(y_1, ..., y_d) \in R^d | y_j \ge 0, j = 1, ..., d\}$.

Lemma 2.5.1. *If* $\beta(k) \le C_1 k^{-(2+\delta')/\delta'}, \delta > \delta' > 0$, *then*

$$\begin{aligned} &EV_n^{(c)2} + EU_n^{(c)2} \le C(m, \delta, r)\, n^{-c} \times \\ &\left\{ M_n^{2/(2+\delta)} \sum_{k=r+1}^{n} k\beta^{\delta/(2+\delta)}(k) + \sum_{c'=0}^{m-1} n^{-c'} M_{n,c'}^{\prime 2/(2+\delta)} \sum_{k=1}^{r} k\beta^{\delta/(2+\delta)}(k) \right\} \quad (2.13) \end{aligned}$$

for some constant $C(m,\delta,r) > 0$. *In particular, if one has* $\beta(k) \leq C_2\rho^k, 0 < \rho < 1$ *then*

$$EV_n^{(c)2} + EU_n^{(c)2} \leq C(m,\delta,r)\, C_2 C(\rho) n^{-c} \left\{ M_n^{2/(2+\delta)} + \sum_{c'=0}^{m-1} n^{-c'} M_{n,c'}^{\prime 2/(2+\delta)} \right\}. \quad (2.14)$$

Proof. The proof of Lemma 2 in Yoshihara (1976), which dealt with the special case of $g_n \equiv g, r = 1, M_n = M_n'$ and yielded (2.13), provides an obvious venue of extension to the more general setup. Elementary arguments then establish (2.14) under geometric mixing conditions. ■

For any $\mathbf{x} \in \text{supp}(w)$, we can write

$$\mathbf{Z}^T\mathbf{W}(\mathbf{x})\mathbf{Z} = \frac{1}{n}\sum_{i=1}^{n} K_H(\mathbf{X}_i - \mathbf{x})\mathbf{T}_i\mathbf{T}_i^T,$$

$$\begin{aligned} \mathbf{Z}_s^T\mathbf{W}_s(\mathbf{x})\mathbf{Z}_s &= \frac{1}{n}\sum_{i=1}^{n} k_{h_s}(X_{is} - x_s)\, L_{G_s}(\mathbf{X}_{i,-s} - \mathbf{x}_{-s}) \times \\ &\quad \left[\left\{ p\left(\frac{X_{is} - x_s}{h_s}\right) p^T\left(\frac{X_{is} - x_s}{h_s}\right) \right\} \bigotimes (\mathbf{T}_i\mathbf{T}_i^T) \right] \end{aligned}$$

in which, as before, $\bigotimes$ denotes the Kronecker product of matrices. Define also the following matrix

$$S_a(\mathbf{x}) = \left\{ \int k(u)p(u)p(u)^T du \right\} \bigotimes S(\mathbf{x}) \quad (2.15)$$

where $S(\mathbf{x}) = E(\mathbf{T}\mathbf{T}^T|\mathbf{X} = \mathbf{x})$ as defined in (A5) (b). For any matrix A, $|A|$ denotes the maximum absolute value of all elements in A.

Lemma 2.5.2. *Let*

$$b_1 = \ln n \left(h_{\max}^{q_1} + 1/\sqrt{nh_{\text{prod}}} \right), \; b_2 = \ln n \left(h_s + g_{\max}^{q_2} + 1/\sqrt{nh_s g_{\text{prod}}} \right),$$

and define the compact set $B = \text{supp}(w) \subset R^{d_2}$. *Under assumptions (A1)-(A6), as* $n \to \infty$, *with probability one*

$$\sup_{\mathbf{x}\in B} \left|\mathbf{Z}^T\mathbf{W}(\mathbf{x})\mathbf{Z} - \varphi(\mathbf{x})S(\mathbf{x})\right| = o(b_1),$$

$$\sup_{\mathbf{x}\in B} \left|\mathbf{Z}_s^T\mathbf{W}_s(\mathbf{x})\mathbf{Z}_s - \varphi(\mathbf{x})S_\alpha(\mathbf{x})\right| = o(b_2).$$

Proof: We only give the proof of the second part. Without loss of generality, one may assume B is bounded by the unit hypercube in R^{d_2}. Observe that

$$\begin{aligned} & \sup_{\mathbf{x}\in B} \left|\mathbf{Z}_s^T\mathbf{W}_s(\mathbf{x})\mathbf{Z}_s - \varphi(\mathbf{x})S_\alpha(\mathbf{x})\right| \\ \leq & \sup_{\mathbf{x}\in B} \left|E\mathbf{Z}_s^T\mathbf{W}_s(\mathbf{x})\mathbf{Z}_s - \varphi(\mathbf{x})S_\alpha(\mathbf{x})\right| + \sup_{\mathbf{x}\in B} \left|\mathbf{Z}_s^T\mathbf{W}_s(\mathbf{x})\mathbf{Z}_s - E(\mathbf{Z}_s^T\mathbf{W}_s(\mathbf{x})\mathbf{Z}_s)\right|. \end{aligned}$$

By a Taylor expansion and the fact that the kernel function L is of order q_2, we can show that

$$b_2^{-1} \sup_{\mathbf{x}\in B} \left|\mathbf{E}\left\{\mathbf{Z}_s^T\mathbf{W}_s(\mathbf{x})\mathbf{Z}_s\right\} - \varphi(\mathbf{x})S_\alpha(\mathbf{x})\right| \to 0.$$

For the second term, consider a covering of B by $v_n^{d_2}$ closed hypercubes $B_{jn} = \{\mathbf{x} : \|\mathbf{x} - \mathbf{x}_j\| \leq v_n^{-1}\}$, where $\{\mathbf{x}_j\}_{j=1}^{v_n^{d_2}}$ denote the center points of the $v_n^{d_2}$ closed hypercubes, and $\|\cdot\|$ denotes the supremum norm. Then

$$\begin{aligned} & b_2^{-1} \sup_{\mathbf{x}\in B} \left|\mathbf{Z}_s^T\mathbf{W}_s(\mathbf{x})\mathbf{Z}_s - E\left\{\mathbf{Z}_s^T\mathbf{W}_s(\mathbf{x})\mathbf{Z}_s\right\}\right| \\ \leq & b_2^{-1} \sup_j \sup_{\mathbf{x}\in B_{jn}} \left|\mathbf{Z}_s^T\mathbf{W}_s(\mathbf{x})\mathbf{Z}_s - \mathbf{Z}_s^T\mathbf{W}_s(\mathbf{x}_j)\mathbf{Z}_s\right| \\ & + b_2^{-1} \sup_j \sup_{\mathbf{x}\in B_{jn}} \left|E\mathbf{Z}_s^T\mathbf{W}_s(\mathbf{x})\mathbf{Z}_s - E\left\{\mathbf{Z}_s^T\mathbf{W}_s(\mathbf{x}_j)\mathbf{Z}_s\right\}\right| \\ & + b_2^{-1} \sup_j \left|\mathbf{Z}_s^T\mathbf{W}_s(\mathbf{x}_j)\mathbf{Z}_s - E\left\{\mathbf{Z}_s^T\mathbf{W}_s(\mathbf{x}_j)\mathbf{Z}_s\right\}\right|. \end{aligned} \tag{2.16}$$

Note that the elements in $\mathbf{Z}_s^T\mathbf{W}_s(\mathbf{x})\mathbf{Z}_s$ are of the form

$$\frac{1}{n}\sum_{i=1}^{n} k_{h_s}(X_{is} - x_s)\, L_{G_s}(\mathbf{X}_{i,-s} - \mathbf{x}_{-s}) \left(\frac{X_{is} - x_s}{h_s}\right)^k T_{il}T_{il'}$$

for $k = 0, \ldots, 2p$, $1 \leq l, l' \leq d_1$, which is denoted as $U_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} U_{n,i}(\mathbf{x})$. Index k, l, l' are suppressed for notation convenience. Then the elements in

$\left|\mathbf{Z}_s^T \mathbf{W}_s(\mathbf{x})\mathbf{Z}_s - \mathbf{Z}_s^T \mathbf{W}_s(\mathbf{x}_j)\mathbf{Z}_s\right|$ are

$$
\begin{aligned}
& |U_n(\mathbf{x}) - U_n(\mathbf{x}_j)| \leq \frac{1}{n}\sum_{i=1}^{n} |U_{n,i}(\mathbf{x}) - U_{n,i}(\mathbf{x}_j)| \\
\leq\ & \frac{1}{n}\sum_{i=1}^{n} \left| k_{h_s}(X_{is} - x_s) L_{G_s}(\mathbf{X}_{i,-s} - \mathbf{x}_{-s}) \left(\frac{X_{is} - x_s}{h_s}\right)^k \right. \\
& \left. - k_{h_s}(X_{is} - x_{js}) L_{G_s}(\mathbf{X}_{i,-s} - \mathbf{x}_{j,-s}) \left(\frac{X_{is} - x_{js}}{h_s}\right)^k \right| |T_{il} T_{il'}| .
\end{aligned}
$$

Under the assumption (A1), there exists a positive constant c, such that

$$
|U_n(\mathbf{x}) - U_n(\mathbf{x}_j)| \leq \frac{c}{(h_s g_{\text{prod}})^2 v_n} \sum_{i=1}^{n} |T_{il} T_{il'}| / n \leq \frac{c}{(h_s g_{\text{prod}})^2 v_n}
$$

almost surely, as a result of assumption (A5) (c) entails that $E\left(\mathbf{T}\mathbf{T}^T\right) < \infty$. Choosing $v_n = \left[(h_s g_{\text{prod}})^{-3}\right]$ (note $v_n \to \infty$), we have

$$
b_2^{-1} \sup_j \sup_{\mathbf{x} \in B_{jn}} \left|\mathbf{Z}_s^T \mathbf{W}_s(\mathbf{x})\mathbf{Z}_s - \mathbf{Z}_s^T \mathbf{W}_s(\mathbf{x}_j)\mathbf{Z}_s\right| = o(1)
$$

almost surely. Similarly, one can show that

$$
b_2^{-1} \sup_j \sup_{\mathbf{x} \in B_{jn}} \left|E\left\{\mathbf{Z}_s^T \mathbf{W}_s(\mathbf{x})\mathbf{Z}_s\right\} - E\left\{\mathbf{Z}_s^T \mathbf{W}_s(\mathbf{x}_j)\mathbf{Z}_s\right\}\right| = o(1).
$$

For the last term in (2.16), note that the elements in

$\mathbf{Z}_s^T \mathbf{W}_s(\mathbf{x}_j)\mathbf{Z}_s - E\left\{\mathbf{Z}_s^T \mathbf{W}_s(\mathbf{x}_j)\mathbf{Z}_s\right\}$ are of the form

$$
S_n(\mathbf{x}_j) = U_n(\mathbf{x}_j) - E\left\{U_n(\mathbf{x}_j)\right\} = \frac{1}{n}\sum_{i=1}^{n} \left[U_{n,i}(\mathbf{x}_j) - E\left\{U_{n,i}(\mathbf{x}_j)\right\}\right] = \frac{1}{n}\sum_{i=1}^{n} U_{n,i}^*(\mathbf{x}_j).
$$

By assumptions (A1) and (A5) (c) that $\mathbf{T}\mathbf{T}^T$ satisfies the Cramer's moment conditions, we have, for $d = 3, 4, \ldots$

$$
\begin{aligned}
E\left|U_{n,i}(\mathbf{x}_j)\right|^d &= E\left| k_{h_s}(X_{is} - x_{js}) L_{G_s}(\mathbf{X}_{i,-s} - \mathbf{x}_{j,-s}) \left(\frac{X_{is} - x_{js}}{h_s}\right)^k T_{il} T_{il'} \right|^d \\
&\leq c_n^d E\left|T_{il} T_{il'}\right|^d \leq c_n^{d-2} d! E\left|T_{il} T_{il'}\right|^2,
\end{aligned}
$$

where $c_n = C_0 (h_s g_{\text{prod}})^{-1}$ for some $C_0 > 0$. Meanwhile

$$\begin{aligned} E\left|U_{n,i}^*(\mathbf{x}_j)\right|^d &= E\left|U_{n,i}(\mathbf{x}_j) - E\left\{U_{n,i}(\mathbf{x}_j)\right\}\right|^d \\ &\leq \sum_{r=0}^{d} \left|E\left\{U_{n,i}(\mathbf{x}_j)\right\}\right|^{d-r} \binom{d}{r} E\left|U_{n,i}(\mathbf{x}_j)\right|^r \leq c_n^{d-2} d! E\left|T_{il} T_{il'}\right|^2 \end{aligned}$$

as long as the constant C_0 is sufficiently large. Applying Theorem 1.4 (Bosq 1998) and inequality (2.12), we have, for any integer $q \in \left[1, \frac{n}{2}\right]$, $\varepsilon > 0$ and each $k \geq 3$

$$P\left\{\left|S_n(\mathbf{x}_j)\right| > b_2\varepsilon\right\} \leq a_1 \exp\left(-\frac{q\varepsilon^2 b_2^2}{25m_2^2 + 5c_n b_2 \varepsilon}\right) + a_2(k) \frac{c}{2} \rho^{\left[\frac{n}{q+1}\right]^{2k/(2k+1)}},$$

where

$$\begin{aligned} a_1 &= \frac{2n}{q} + 2\left(1 + \frac{\varepsilon^2}{25m_2^2 + 5c_n b_2 \varepsilon}\right) \quad with \quad m_2^2 = E\left\{U^*(\mathbf{x}_j)\right\}^2, \\ a_2(k) &= 11n\left(1 + \frac{5m_p^{k/(2k+2)}}{b_2\varepsilon}\right) \quad with \quad m_p = \left\|U^*(\mathbf{x}_j)\right\|_p. \end{aligned}$$

By taking $q = \left[n/(\ln n)^2\right]$, the first term

$$a_1 \exp\left(-\frac{q\varepsilon^2 b_2^2}{25m_2^2 + 5c_n b_2 \varepsilon}\right) \leq c_1 \exp\left\{-c_2 (\ln n)^2\right\}$$

and the second term

$$a_2(k) \frac{c}{2} \rho^{\left[\frac{n}{q+1}\right]^{2k/(2k+1)}} \leq c_3 \exp\left\{-c_4 (\ln n)^2\right\},$$

where the $c_i's$ are strictly positive constants. So, for any integer $1 \leq j \leq v_n^d$, we have

$$P\left\{\left|S_n(\mathbf{x}_j)\right| > b_2\varepsilon\right\} \leq c_1 \exp\left\{-c_2 (\ln n)^2\right\} + c_3 \exp\left(-c_4 (\ln n)^2\right).$$

Then for any $\varepsilon > 0$

$$\begin{aligned} &P\left\{b_2^{-1} \sup_j \left|S_n(\mathbf{x}_j)\right| > \varepsilon\right\} \leq \sum_{j=1}^{v_n^d} P\left\{b_2^{-1}\left|S_n(\mathbf{x}_j)\right| > \varepsilon\right\} \\ \leq\ & v_n^d \left[c_1 \exp\left\{-c_2 (\ln n)^2\right\} + c_3 \exp\left(-c_4 (\ln n)^2\right)\right]. \end{aligned}$$

Since we have taken $v_n = \left[(h_s g_{\text{prod}})^{-3}\right]$,

$$\sum_n P\left\{b_2^{-1} \sup_j |S_n(\mathbf{x}_j)| > \varepsilon\right\}$$
$$\leq \sum_n v_n^d \left[c_1 \exp\left\{-c_2 (\ln n)^2\right\} + c_3 \exp\left(-c_4 (\ln n)^2\right)\right] < +\infty.$$

By the Borel-Cantelli lemma, we have, $b_2^{-1} \sup_j |S_n(\mathbf{x}_j)| \to 0$ almost surely. The rest of the lemma follows immediately. ■

2.5.3 Proof of Theorem 2.2.1

By observing that, $e_l^T \left\{\mathbf{Z}^T\mathbf{W}(\mathbf{X}_i)\mathbf{Z}\right\}^{-1} \mathbf{Z}^T\mathbf{W}(\mathbf{X}_i)\mathbf{Z}e_{l'} = \delta_{ll'}$, where $\delta_{ll'}$ equals to 1 if $l = l'$ and equals to 0 otherwise, we have

$$\frac{1}{n}\sum_{i=1}^n w(\mathbf{X}_i)\left\{\hat{\alpha}_{l0} - \alpha_{l0}\right\} = I + II + III \tag{2.17}$$

in which

$$I = \frac{1}{n}\sum_{i=1}^n w(\mathbf{X}_i)\, e_l^T \left\{\mathbf{Z}^T\mathbf{W}(\mathbf{X}_i)\mathbf{Z}\right\}^{-1} \mathbf{Z}^T\mathbf{W}(\mathbf{X}_i)\mathbf{E},$$
$$II = \frac{1}{n}\sum_{i=1}^n w(\mathbf{X}_i)\, e_l^T \left\{\mathbf{Z}^T\mathbf{W}(\mathbf{X}_i)\mathbf{Z}\right\}^{-1} \mathbf{Z}^T\mathbf{W}(\mathbf{X}_i) \times$$
$$\left[\mathbf{M} - \sum_{l'=1}^{d_1}\left\{c_{l'} + \sum_{s=1}^{d_2}\alpha_{l's}(X_{is})\right\}\mathbf{Z}e_{l'}\right],$$
$$III = \frac{1}{n}\sum_{i=1}^n w(\mathbf{X}_i)\sum_{s=1}^{d_2}\alpha_{ls}(X_{is})$$

where $\mathbf{M}$ is the vector of conditional means

$$\mathbf{M} = \left[\sum_{l'=1}^{d_1}\left\{c_{l'} + \sum_{s=1}^{d_2}\alpha_{l's}(X_{js})\right\}T_{jl'}\right]_{j=1,\ldots,n}$$

and $\mathbf{E} = \{\sigma(\mathbf{X}_1, \mathbf{T}_1)\varepsilon_1, \ldots, \sigma(\mathbf{X}_n, \mathbf{T}_n)\varepsilon_n\}^T$, the vector of errors. Next, observe that

$$\mathbf{M} - \sum_{l'=1}^{d_1} \left\{ c_{l'} + \sum_{s=1}^{d_2} \alpha_{l's}(X_{is}) \right\} \mathbf{Z}e_{l'}$$
$$= \left[\sum_{l'=1}^{d_1} \sum_{s=1}^{d_2} \{\alpha_{l's}(X_{js}) - \alpha_{l's}(X_{is})\} T_{jl'} \right]_{j=1,\ldots,n}.$$

Define

$$\mathbf{R}_1(\mathbf{X}_i) = \left[\sum_{l'=1}^{d_1} \sum_{s=1}^{d_2} \{\alpha_{l's}(X_{js}) - \alpha_{l's}(X_{is})\} T_{jl'} \right]_{j=1,\ldots,n}$$

one can rewrite II as

$$II = \frac{1}{n} \sum_{i=1}^{n} w(\mathbf{X}_i) \left[e_l^T \{\mathbf{Z}^T \mathbf{W}(\mathbf{X}_i) \mathbf{Z}\}^{-1} \mathbf{Z}^T \mathbf{W}(\mathbf{X}_i) \mathbf{R}_1(\mathbf{X}_i) \right].$$

Now let v_1 be the integer such that $b_1^{v_1+1} = o\left(h_{\max}^{q_1+2}\right)$. Following immediately from Lemma 2.5.2, one has

$$\{\mathbf{Z}^T\mathbf{W}(\mathbf{x})\mathbf{Z}\}^{-1} - \frac{S(\mathbf{x})^{-1}}{\varphi(\mathbf{x})} = \frac{S(\mathbf{x})^{-1}}{\varphi(\mathbf{x})} \sum_{\nu=1}^{v_1} \left\{ I_{d_1} - \frac{\mathbf{Z}^T\mathbf{W}(\mathbf{x})\mathbf{Z}S^{-1}(\mathbf{x})}{\varphi(\mathbf{x})} \right\}^{\nu} + Q_2(\mathbf{x})$$
$$= \sum_{v=1}^{v_1} Q_{1v}(\mathbf{x}) + Q_2(\mathbf{x})$$

where the matrix $Q_2(\mathbf{x})$ satisfies

$$\sup_{\mathbf{x} \in B} |Q_2(\mathbf{x})| = o\left(h_{\max}^{q_1+2}\right) \text{ w.p.1.}$$

To prove Theorem 2.2.1, we need the following lemmas.

Lemma 2.5.3. *Define*

$$D_{n1} = \frac{1}{n} \sum_{i=1}^{n} w(\mathbf{X}_i) Q_2(\mathbf{X}_i) \mathbf{Z}^T \mathbf{W}(\mathbf{X}_i) \mathbf{E},$$
$$D_{n2} = \frac{1}{n} \sum_{i=1}^{n} w(\mathbf{X}_i) Q_2(\mathbf{X}_i) \mathbf{Z}^T \mathbf{W}(\mathbf{X}_i) \mathbf{R}_1(\mathbf{X}_i).$$

Then as $n \to +\infty$

$$|D_{n1}| + |D_{n2}| = o\left(h_{\max}^{q_1+2}\right) w.p.1.$$

Lemma 2.5.4. *For fixed* $\nu = 1, ..., v_1$, *define*

$$\begin{aligned} F_{1\nu} &= \frac{1}{n}\sum_{i=1}^{n} w(\mathbf{X}_i)\, Q_{1\nu}(\mathbf{X}_i)\mathbf{Z}^T\mathbf{W}(\mathbf{X}_i)\mathbf{E}, \\ F_{2\nu} &= \frac{1}{n}\sum_{i=1}^{n} w(\mathbf{X}_i)\, Q_{1\nu}(\mathbf{X}_i)\mathbf{Z}^T\mathbf{W}(\mathbf{X}_i)\mathbf{R}_1(\mathbf{X}_i). \end{aligned}$$

Then as $n \to +\infty$

$$|F_{1\nu}| + |F_{2\nu}| = o\left(b_1^{\nu}/\sqrt{n}\right) w.p.1.$$

Proof: For simplicity of notation, we only consider the case of $F_{1\nu}$ with $\nu = 1$

$$\begin{aligned} & F_{11}(\mathbf{x}) \\ = & \frac{1}{n}\sum_{i=1}^{n} w(\mathbf{X}_i)\frac{S^{-1}(\mathbf{X}_i)}{\varphi(\mathbf{X}_i)}\left\{I_{d_1} - \frac{\mathbf{Z}^T\mathbf{W}(\mathbf{X}_i)\mathbf{Z}S^{-1}(\mathbf{X}_i)}{\varphi(\mathbf{X}_i)}\right\}\mathbf{Z}^T\mathbf{W}(\mathbf{X}_i)\mathbf{E} \\ = & \frac{1}{n}\sum_{i=1}^{n} w(\mathbf{X}_i)\, S^{-1}(\mathbf{X}_i)\left\{\frac{S(\mathbf{X}_i)}{\varphi(\mathbf{X}_i)} - \frac{\mathbf{Z}^T\mathbf{W}(\mathbf{X}_i)\mathbf{Z}}{\varphi^2(\mathbf{X}_i)}\right\}S^{-1}(\mathbf{X}_i)\mathbf{Z}^T\mathbf{W}(\mathbf{X}_i)\mathbf{E} \\ = & \frac{1}{n}\sum_{i=1}^{n} w(\mathbf{X}_i)\, S^{-1}(\mathbf{X}_i)\left[\frac{S(\mathbf{X}_i)}{\varphi(\mathbf{X}_i)} - \frac{E\left\{\mathbf{Z}^T\mathbf{W}(\mathbf{X}_i)\mathbf{Z}\right\}}{\varphi^2(\mathbf{X}_i)}\right]S^{-1}(\mathbf{X}_i)\mathbf{Z}^T\mathbf{W}(\mathbf{X}_i)\mathbf{E} \\ - & \frac{1}{n}\sum_{i=1}^{n} w(\mathbf{X}_i)\, S^{-1}(\mathbf{X}_i)\left[\frac{\mathbf{Z}^T\mathbf{W}(\mathbf{X}_i)\mathbf{Z}}{\varphi^2(\mathbf{X}_i)} - \frac{E\left\{\mathbf{Z}^T\mathbf{W}(\mathbf{X}_i)\mathbf{Z}\right\}}{\varphi^2(\mathbf{X}_i)}\right]S^{-1}(\mathbf{X}_i)\mathbf{Z}^T\mathbf{W}(\mathbf{X}_i)\mathbf{E} \\ = & P_1 - P_2. \end{aligned}$$

Let $\boldsymbol{\xi}_i = (\mathbf{X}_i, \mathbf{T}_i, \varepsilon_i)$, and define

$$\begin{aligned} g_n(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j) = & w(\mathbf{X}_i)\, S^{-1}(\mathbf{X}_i)\left[\frac{S(\mathbf{X}_i)}{\varphi(\mathbf{X}_i)} - \frac{E\left\{\mathbf{Z}^T\mathbf{W}(\mathbf{X}_i)\mathbf{Z}\right\}}{\varphi^2(\mathbf{X}_i)}\right]\times \\ & S^{-1}(\mathbf{X}_i)K_H(\mathbf{X}_j - \mathbf{X}_i)\,\mathbf{T}_j\sigma(\mathbf{X}_j, \mathbf{T}_j)\,\varepsilon_j \\ & + w(\mathbf{X}_j)\, S^{-1}(\mathbf{X}_j)\left[\frac{S(\mathbf{X}_j)}{\varphi(\mathbf{X}_j)} - \frac{E\left\{\mathbf{Z}^T\mathbf{W}(\mathbf{X}_j)Z\right\}}{\varphi^2(\mathbf{X}_j)}\right]\times \\ & S^{-1}(\mathbf{X}_j)K_H(\mathbf{X}_i - \mathbf{X}_j)\,\mathbf{T}_i\sigma(\mathbf{X}_i, \mathbf{T}_i)\,\varepsilon_i. \end{aligned}$$

Then P_1 can be written as the von Mises' differential statistic

$$P_1 = \frac{1}{2n^2} \sum_{i,j=1}^{n} g_n\left(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j\right),$$

which can be decomposed as

$$P_1 = \frac{1}{2}\left\{\boldsymbol{\theta}_n(F) + 2\mathbf{V}_n^{(1)} + \mathbf{V}_n^{(2)}\right\}$$

in which

$$\boldsymbol{\theta}_n\left(F\right) = \int g_n\left(u, v\right) dF_{\boldsymbol{\xi}_i}\left(u\right) dF_{\boldsymbol{\xi}_j}\left(v\right) = 0.$$

In order to write down the explicit expressions of $\mathbf{V}_n^{(1)}, \mathbf{V}_n^{(2)}$, let E_i denote taking expectation with respect to the random vector indexed by i and $E_{n,j}$ denote taking expectation with respect to the random vector indexed by j using the empirical measure, both under the presumption of independence between $\boldsymbol{\xi}_i$ and $\boldsymbol{\xi}_j$. One has

$$\mathbf{V}_n^{(1)} = E_i E_{n,j} g_n\left(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j\right) = \frac{1}{n} \sum_{j=1}^{n} g_{n,1}\left(\boldsymbol{\xi}_j\right)$$

in which

$$\begin{aligned} g_{n,1}\left(\boldsymbol{\xi}_j\right) \quad = \quad & \int w(\mathbf{z}) S^{-1}(\mathbf{z}) \left[\frac{S(\mathbf{z})}{\varphi(\mathbf{z})} - \frac{E\left\{\mathbf{Z}^T \mathbf{W}(\mathbf{z}) \mathbf{Z}\right\}}{\varphi^2(\mathbf{z})}\right] \times \\ & S^{-1}(\mathbf{z}) K_H\left(\mathbf{X}_j - \mathbf{z}\right) \varphi(\mathbf{z}) d\mathbf{z} \mathbf{T}_j \sigma\left(\mathbf{X}_j, \mathbf{T}_j\right) \varepsilon_j. \end{aligned}$$

Clearly $g_{n,1}$ has mean 0 and variance of order b_1^2. So $V_n^{(1)} = \frac{1}{n}\sum_{j=1}^{n} g_{n,1}\left(\boldsymbol{\xi}_j\right) = o_p\left(b_1/\sqrt{n}\right)$. Finally for $\mathbf{V}_n^{(2)}$, by Lemma 2.5.1, under assumption (A3), one has for some small $\delta > 0$

$$E\left(\mathbf{V}_n^{(2)}\right)^2 \leq cn^{-2} \left\{ M_n^{\frac{2}{2+\delta}} + M_{n,0}^{\frac{2}{2+\delta}} + M_{n,1}^{\frac{2}{2+\delta}} n^{-1} \right\}$$

where $M_n, M_{n,0}$ and $M_{n,1}$ are the quantities which satisfy the following inequalities

$$\begin{aligned} E_1 E_2 \left| g_n \left(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2 \right) \right|^{2+\delta} &\leq M_n < +\infty \\ \sup_{i \neq j} E_{i,j} \left| g_n \left(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j \right) \right|^{2+\delta} &\leq M_{n,0} < +\infty \\ E_i \left| g_n \left(\boldsymbol{\xi}_i, \boldsymbol{\xi}_i \right) \right|^{2+\delta} &\leq M_{n,1} < +\infty \end{aligned}$$

And observe that

$$\begin{aligned} E_{i,j} \left| g_n \left(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j \right) \right|^{2+\delta} &\leq c b_1^{2+\delta} E \left| w \left(\mathbf{X}_i \right) K_H \left(\mathbf{X}_j - \mathbf{X}_i \right) T_j \sigma \left(\mathbf{X}_j, \mathbf{T}_j \right) \varepsilon_j \right|^{2+\delta} \\ &\leq c b_1^{2+\delta} c(\rho) \left\{ E \left| K_H \left(\mathbf{X}_j - \mathbf{x} \right) \mathbf{T}_j \sigma \left(\mathbf{X}_j, \mathbf{T}_j \right) \varepsilon_j \right|^{2+\delta} \right\}^{(2+\delta)/(2+2\delta)} \\ &\leq \left(\frac{1}{h_{\text{prod}}^{1+2\delta}} \right)^{(2+\delta)/(2+2\delta)} c b_1^{2+\delta} c(\rho). \end{aligned}$$

So we can take $M_{n,0} = h_{\text{prod}}^{-(1+2\delta)(2+\delta)/(2+2\delta)} c b_1^{2+\delta}$, and by setting the mixing coefficient ρ to 0, one also gets $M_n = h_{\text{prod}}^{-(1+2\delta)(2+\delta)/(2+2\delta)} c b_1^{2+\delta}$. Similarly, we can show that $M_{n,1} = c b_1^{2+\delta} h_{\text{prod}}^{-(2+\delta)}$. So by taking δ small, one has

$$\begin{aligned} & E \left(P_1^2 \right) \\ \leq\ & c n^{-2} \left(h_{\text{prod}}^{-(1+2\delta)(2+\delta)/(2+2\delta)} b_1^{2+\delta} \right)^{2/(2+\delta)} + c n^{-3} \left(b_1^{2+\delta} h_{\text{prod}}^{-(2+\delta)} \right)^{2/(2+\delta)} + c b_1^2 / n \\ \leq\ & c n^{-2} b_1^2 h_{\text{prod}}^{-2(1+2\delta)/(2+2\delta)} + c n^{-3} b_1^2 h_{\text{prod}}^{-2(2+\delta)/(2+\delta)} + c b_1^2 / n \\ \leq\ & c n^{-1} b_1^2. \end{aligned}$$

Similarly, we can show that $E P_2^2 \leq c n^{-1} b_1^2$. So we have $F_{11} = o_p \left(b_1 / \sqrt{n} \right)$. ■

Lemma 2.5.5. *Define*

$$P_{1n} = \frac{1}{n} \sum_{i=1}^{n} \frac{w \left(\mathbf{X}_i \right)}{\varphi(\mathbf{X}_i)} \left\{ e_l^T \mathbf{S}^{-1}(\mathbf{X}_i) \mathbf{Z}^T \mathbf{W}(\mathbf{X}_i) \mathbf{R}_1(\mathbf{X}_i) \right\}$$

then $P_{1n} = O_p \left(h_{\max}^{q_1} \right) = o_p \left(n^{-1/2} \right)$ *as* $n \to \infty$.

Proof: Let $K_l^*(\mathbf{X},\mathbf{T}) = e_l^T \mathbf{S}^{-1}(\mathbf{X})\,\mathbf{T}$, then

$$P_{1n} = \frac{1}{n^2}\sum_{i,j=1}^{n} \frac{w(\mathbf{X}_i)}{\varphi(\mathbf{X}_i)} K_l^*(\mathbf{X}_i,\mathbf{T}_j)\, K_H(\mathbf{X}_j - \mathbf{X}_i) \times \left[\sum_{l'=1}^{d_1}\sum_{s=1}^{d_2} \{\alpha_{l's}(X_{js}) - \alpha_{l's}(X_{is})\}\, T_{jl'}\right]$$

which is again a von Mises' statistic. Its θ_n is of the form

$$\int \frac{w(\mathbf{z})}{\varphi(\mathbf{z})} K_l^*(\mathbf{z},\mathbf{t})\, K_H(\mathbf{x}-\mathbf{z}) \left[\sum_{l'=1}^{d_1}\sum_{s=1}^{d_2} \{\alpha_{l's}(x_s) - \alpha_{l's}(z_s)\}\, t_{l'}\right] \times \varphi(\mathbf{z})\,\psi(\mathbf{x},\mathbf{t})\, d\mathbf{z}d\mathbf{x}d\mathbf{t}.$$

After changing of variable $\mathbf{u} = H^{-1}(\mathbf{x}-\mathbf{z})$, the above becomes

$$\int w(\mathbf{z}) K_l^*(\mathbf{z},\mathbf{t}) K(\mathbf{u}) \left[\sum_{l'=1}^{d_1}\sum_{s=1}^{d_2} \{\alpha_{l's}(z_s + h_{0,s}u_s) - \alpha_{l's}(z_s)\}\, t_{l'}\right] \times \psi(\mathbf{z} + H\mathbf{u},\mathbf{t})\, d\mathbf{u}d\mathbf{z}d\mathbf{t}$$

$$= O\left(h_{\max}^{q_1}\right)$$

where the last step is obtained by Taylor expansion of $\alpha_{ls}(z_s + h_{0,s}u_s)$ to q_1-th degree and of $\psi(\mathbf{z}+H\mathbf{u},\mathbf{t})$ to (q_1-1)-th degree, which exist according to assumptions (A2) and (A5) (a). By assumption (A1), all the terms with order smaller than $h_{\max}^{q_1}$ disappear. So the leading term left is of $h_{\max}^{q_1}$ order. It is routine to verify that $\mathbf{V}_n^{(1)}$ and $\mathbf{V}_n^{(2)}$ are $O_p\left(h_{\max}^{q_1}\right)$ as well. Hence $P_{1n} = O_p\left(h_{\max}^{q_1}\right)$ and assumption (A6) (a) entails that $O_p\left(h_{\max}^{q_1}\right) = o_p\left(n^{-1/2}\right)$. ■

Finally we can finish the proof of Theorem 2.2.1 as follows. Define

$$P_{2n} = \frac{1}{n}\sum_{i=1}^{n} \frac{w(\mathbf{X}_i)}{\varphi(\mathbf{X}_i)} \left\{e_l^T \mathbf{S}^{-1}(\mathbf{X}_i)\,\mathbf{Z}^T\mathbf{W}(\mathbf{X}_i)\,\mathbf{E}\right\}.$$

Then

$$P_{2n} = \frac{1}{n^2}\sum_{i,j=1}^{n} \frac{w(\mathbf{X}_i)}{\varphi(\mathbf{X}_i)} K_l^*(\mathbf{X}_i,\mathbf{T}_j)\, K_H(\mathbf{X}_j - \mathbf{X}_i)\,\sigma(\mathbf{X}_j,\mathbf{T}_j)\,\varepsilon_j \qquad (2.18)$$

which again, by a von Mises' statistic argument, becomes

$$\frac{1}{n}\sum_{j=1}^{n}\int \frac{w(\mathbf{x})}{\varphi(\mathbf{x})}K_l^*(\mathbf{x},\mathbf{T}_j)\,K_H(\mathbf{X}_j-\mathbf{x})\,\varphi(\mathbf{x})\,d\mathbf{x}\sigma(\mathbf{X}_j,\mathbf{T}_j)\,\varepsilon_j + o_p\left(\frac{\log n}{n^2 h_{\text{prod}}}\right)$$

which, after changing of variable $\mathbf{X}_j = \mathbf{x} + H\mathbf{u}$ becomes

$$\begin{aligned}
&\frac{1}{n}\sum_{j=1}^{n}\int w(\mathbf{X}_j - H\mathbf{u})\,K_l^*(\mathbf{X}_j - H\mathbf{u},\mathbf{T}_j)\,K(\mathbf{u})\,\sigma(\mathbf{X}_j,\mathbf{T}_j)\,\varepsilon_j d\mathbf{u} + o_p\left(\frac{\log n}{n^2 h_{\text{prod}}}\right)\\
&= \frac{1}{n}\sum_{j=1}^{n} w(\mathbf{X}_j)\,K_l^*(\mathbf{X}_j,\mathbf{T}_j)\,\sigma(\mathbf{X}_j,\mathbf{T}_j)\,\varepsilon_j + o_p\left(\frac{\log n}{n^2 h_{\text{prod}}}\right) + o_p\left(h_{\max}^{q_1}\right)\\
&= \frac{1}{n}\sum_{j=1}^{n} w(\mathbf{X}_j)\,K_l^*(\mathbf{X}_j,\mathbf{T}_j)\,\sigma(\mathbf{X}_j,\mathbf{T}_j)\,\varepsilon_j + o_p\left(n^{-1/2}\right). \qquad (2.19)
\end{aligned}$$

Now come back to the decomposition of $\frac{1}{n}\sum_{i=1}^{n} w(\mathbf{X}_i)(\hat{\alpha_{l0}} - \alpha_{l0})$ as in (2.17), and by Lemmas 2.5.2, 2.5.3, 2.5.4, 2.5.5, one has

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} w(\mathbf{X}_i)(\hat{\alpha}_{l0} - \alpha_{l0})\\
&= \frac{1}{n}\sum_{i=1}^{n} w(\mathbf{X}_j)\left\{K_l^*(\mathbf{X}_j,\mathbf{T}_j)\,\sigma(\mathbf{X}_j,\mathbf{T}_j)\,\varepsilon_j + \sum_{s=1}^{d_2}\alpha_{ls}(X_{js})\right\} + o_p\left(n^{-1/2}\right).
\end{aligned}$$

Now define

$$\begin{aligned}
\tau_j &= w(\mathbf{X}_j)\left\{K_l^*(\mathbf{X}_j,\mathbf{T}_j)\,\sigma(\mathbf{X}_j,\mathbf{T}_j)\,\varepsilon_j + \sum_{s=1}^{d_2}\alpha_{ls}(X_{js})\right\}\\
&= \tau_{j1} + \tau_{j2}.
\end{aligned}$$

Then by the condition that ε_j is independent of $\{(\mathbf{X}_i,\mathbf{T}_i)\}_{i\le j}$, we have

$$E\{w(\mathbf{X}_j)\,K_l^*(\mathbf{X}_j,\mathbf{T}_j)\,\sigma(\mathbf{X}_j,\mathbf{T}_j)\,\varepsilon_j\} = E\{w(\mathbf{X}_j)\,K_l^*(\mathbf{X}_j,\mathbf{T}_j)\,\sigma(\mathbf{X}_j,\mathbf{T}_j)\}\,E(\varepsilon_j) = 0$$

and by the identification condition that $E\left\{w(\mathbf{X})\sum_{s=1}^{d_2}\alpha_{ls}(X_s)\right\} = 0$. So $E(\tau_j) = 0$. Furthermore, by assumption (A3), $\{\tau_j\}$ is a stationary β-mixing process, with

geometric β-mixing coefficient. By Minkowski's inequality, for some $\delta > 0$

$$E|\tau_j|^{2+\delta} \leq \left\{ \left(E|\tau_{j1}|^{2+\delta} \right)^{1/(2+\delta)} + \left(E|\tau_{j2}|^{2+\delta} \right)^{1/(2+\delta)} \right\}^{2+\delta}.$$

By assumptions (A1), (A4), (A5) and (A7), we have

$$\begin{aligned} E|\tau_{j1}|^{2+\delta} &= E|w(\mathbf{X}_j) K_l^*(\mathbf{X}_j, \mathbf{T}_j) \sigma(\mathbf{X}_j, \mathbf{T}_j) \varepsilon_j|^{2+\delta} \\ &= E|w(\mathbf{X}_j) e_l S(\mathbf{X}_j) T_j \sigma(\mathbf{X}_j, \mathbf{T}_j)|^{2+\delta} E|\varepsilon_j|^{2+\delta} \\ &\leq cE \left(\sum_{l=1}^{d_1} |T_{jl}| \right)^{2+\delta} E|\varepsilon_j|^{2+\delta} \\ &\leq c \left\{ \sum_{l=1}^{d_1} \left(E|T_{jl}|^{2+\delta} \right)^{1/(2+\delta)} \right\}^{2+\delta} E|\varepsilon_j|^{2+\delta} < +\infty. \end{aligned}$$

By assumption (A7) that weight function w has compact support and the continuity of the functions w, α_{ls}, one has $E|\tau_{j2}|^{2+\delta} < +\infty$. So $E|\tau_j|^{2+\delta} < +\infty$. Next, define

$$\begin{aligned} \sigma_l^2 &= \sum_{j=-\infty}^{+\infty} \operatorname{cov}(\tau_0, \tau_j) = 2\sum_{j=1}^{+\infty} \operatorname{cov}(\tau_0, \tau_j) + \operatorname{var}(\tau_0) \\ &= 2\sum_{j=1}^{+\infty} \operatorname{cov}(\tau_0, \tau_{j2}) + \operatorname{var}(\tau_0) \end{aligned} \tag{2.20}$$

which is finite by Theorem 1.5 of Bosq (1998). Applying the central limit theorem for strongly mixing process (Theorem 1.7 of Bosq 1998), we have

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tau_j \Longrightarrow N(0, \sigma_l^2).$$

Theorem 2.2.1 now follows immediately by the assumption (A6) (a) on the bandwidths and the fact that $\frac{1}{n}\sum_{i=1}^{n} w(\mathbf{X}_i) \to 1$ a.s. ■

2.5.4 Proof of Theorem 2.3.1

Following similarly as in the proof of Theorem 2.2.1, let v_2 be an integer which satisfies $b_2^{v_2} = o\left(h_s^{p+2}\right)$. Then by Lemma 2.5.2, one has

$$\left\{\mathbf{Z}_s^T \mathbf{W}_s(\mathbf{x}) \mathbf{Z}_s\right\}^{-1} - \frac{S_\alpha^{-1}(\mathbf{x})}{\varphi(\mathbf{x})} = \frac{S_\alpha^{-1}(\mathbf{x})}{\varphi(\mathbf{x})} \sum_{v=1}^{v_2} A(\mathbf{x})^v + Q_s(\mathbf{x}) \tag{2.21}$$

where

$$A(\mathbf{x}) = I_{(p+1)d_1} - \frac{\mathbf{Z}_s^T \mathbf{W}_s(\mathbf{x}) \mathbf{Z}_s S_\alpha^{-1}(\mathbf{x})}{\varphi(\mathbf{x})}$$

and the matrix $Q_s(\mathbf{x})$ satisfies

$$\sup_{\mathbf{x} \in B} |Q_s(\mathbf{x})| = o\left(h_s^{p+2}\right) \text{ w.p. 1.}$$

Also as in the proof of Theorem 2.2.1, by the equation that

$$e_l \left\{\mathbf{Z}_s^T \mathbf{W}_s(\mathbf{X}_{is}) \mathbf{Z}_s\right\}^{-1} \mathbf{Z}_s^T \mathbf{W}_s(\mathbf{X}_{is}) \mathbf{Z}_s e_{l'} = \delta_{ll'}, \quad l' = 1, \ldots, d_1$$

for fixed $l = 1, \ldots, d_1$ and $s = 1, \ldots, d_2$, we have the following decomposition

$$\begin{aligned}
& \frac{1}{n} \sum_{i=1}^{n} w_{-s}(\mathbf{X}_{i,-s}) \left\{\hat{\alpha}_{ls}(x_s) - \alpha_{ls}(x_s)\right\} \\
= & \frac{1}{n} \sum_{i=1}^{n} w_{-s}(\mathbf{X}_{i,-s}) \left[e_l^T \left\{\mathbf{Z}_s^T \mathbf{W}_s(\mathbf{X}_{i,-s}) \mathbf{Z}_s\right\}^{-1} \mathbf{Z}_s^T \mathbf{W}_s(\mathbf{X}_{i,-s}) \mathbf{Y} - \alpha_{ls}(x_s) - \hat{\alpha}_{l0}\right] \\
= & \frac{1}{n} \sum_{i=1}^{n} w_{-s}(\mathbf{X}_{i,-s}) \left[e_l^T \left\{\mathbf{Z}_s^T \mathbf{W}_s(\mathbf{X}_{i,-s}) \mathbf{Z}_s\right\}^{-1} \mathbf{Z}_s^T \mathbf{W}_s(\mathbf{X}_{i,-s}) \{\mathbf{Y} - \mathbf{M}\right. \\
+ & \left. \mathbf{M} - \sum_{l'=1}^{d_1} \sum_{v=0}^{p} \frac{\alpha_{l's}^{(v)}(x_s) h_s^v}{v!} \mathbf{Z}_s e_{(d_1 v + l')} - \sum_{l'=1}^{d_1} \left\{c_{l'} + \sum_{s' \neq s}^{d_2} \alpha_{l's'}(X_{is'})\right\} \mathbf{Z}_s e_{l'} \right\} \Bigg] \\
+ & \frac{1}{n} \sum_{i=1}^{n} w_{-s}(\mathbf{X}_{i,-s}) \sum_{s' \neq s} \alpha_{ls'}(X_{is'}) + \frac{1}{n} \sum_{i=1}^{n} w_{-s}(\mathbf{X}_{i,-s}) (\hat{\alpha}_{l0} - \alpha_{l0})
\end{aligned} \tag{2.22}$$

where $\mathbf{M}$ is the mean vector, as defined in Theorem 2.2.1. Next define

$$\mathbf{R}_1 = \mathbf{R}_1(x_s) = \left[\sum_{l'=1}^{d_1} \left\{ \alpha_{l's}(X_{js}) - \sum_{v=0}^{p} \frac{\alpha_{l's}^{(v)}(x_s)}{v!} (X_{js} - x_s)^v \right\} T_{jl'} \right]_{j=1,\ldots,n},$$

$$\mathbf{R}_2(\mathbf{X}_{i,-s}) = \left[\sum_{l'=1}^{d_1} \sum_{s' \neq s}^{d_2} \left\{ \alpha_{l's'}(\mathbf{X}_{js'}) - \alpha_{l's'}(\mathbf{X}_{is'}) \right\} T_{jl'} \right]_{j=1,\ldots,n},$$

$$R_3 = \frac{1}{n} \sum_{i=1}^{n} w_{-s}(\mathbf{X}_{i,-s}) \left\{ \sum_{s' \neq s} \alpha_{ls'}(X_{is'}) \right\},$$

$$R_4 = \frac{1}{n} \left\{ \sum_{i=1}^{n} w_{-s}(\mathbf{X}_{i,-s}) \right\} (\hat{\alpha}_{l0} - \alpha_{l0}), \tag{2.23}$$

$$D_{s1}(x_s) = \frac{1}{n} \sum_{i=1}^{n} w_{-s}(\mathbf{X}_{i,-s}) \left\{ e_l^T Q_s(x_s, \mathbf{X}_{i,-s}) \mathbf{Z}_s^T \mathbf{W}_s(x_s, \mathbf{X}_{i,-s}) \mathbf{E} \right\}, \tag{2.24}$$

$$D_{s2}(x_s) = \frac{1}{n} \sum_{i=1}^{n} w_{-s}(\mathbf{X}_{i,-s}) \left\{ e_l^T Q_s(x_s, \mathbf{X}_{i,-s}) \mathbf{Z}_s^T \mathbf{W}_s(x_s, \mathbf{X}_{i,-s}) \mathbf{R}_1 \right\},$$

$$D_{s3}(x_s) = \frac{1}{n} \sum_{i=1}^{n} w_{-s}(\mathbf{X}_{i,-s}) \left\{ e_l^T Q_s(x_s, \mathbf{X}_{i,-s}) \mathbf{Z}_s^T \mathbf{W}_s(x_s, \mathbf{X}_{i,-s}) \mathbf{R}_2(\mathbf{X}_{i,-s}) \right\},$$

$$R_{r1}(x_s) = \frac{1}{n} \sum_{i=1}^{n} w_{-s}(\mathbf{X}_{i,-s}) \left[e_l^T \left\{ A(x_s, \mathbf{X}_{i,-s}) \right\}^r \mathbf{Z}_s^T \mathbf{W}_s(x_s, \mathbf{X}_{i,-s}) \mathbf{E} \right], \tag{2.25}$$

$$R_{r2}(x_s) = \frac{1}{n} \sum_{i=1}^{n} w_{-s}(\mathbf{X}_{i,-s}) \left[e_l^T \left\{ A(x_s, \mathbf{X}_{i,-s}) \right\}^r \mathbf{Z}_s^T \mathbf{W}_s(x_s, \mathbf{X}_{i,-s}) \mathbf{R}_1 \right],$$

$$R_{r3}(x_s) = \frac{1}{n} \sum_{i=1}^{n} w_{-s}(\mathbf{X}_{i,-s}) \left[e_l^T \left\{ A(x_s, \mathbf{X}_{i,-s}) \right\}^r \mathbf{Z}_s^T \mathbf{W}_s(x_s, \mathbf{X}_{i,-s}) \mathbf{R}_2(\mathbf{X}_{i,-s}) \right],$$

$$P_{s1}(x_s) = \frac{1}{n} \sum_{i=1}^{n} \frac{w_{-s}(\mathbf{X}_{i,-s})}{\varphi(x_s, \mathbf{X}_{i,-s})} \left\{ e_l^T S_\alpha^{-1} \mathbf{Z}_s^T \mathbf{W}_s(x_s, \mathbf{X}_{i,-s}) \mathbf{E} \right\}, \tag{2.26}$$

$$P_{s2}(x_s) = \frac{1}{n} \sum_{i=1}^{n} \frac{w_{-s}(\mathbf{X}_{i,-s})}{\varphi(x_s, \mathbf{X}_{i,-s})} \left\{ e_l^T S_\alpha^{-1} \mathbf{Z}_s^T \mathbf{W}_s(x_s, \mathbf{X}_{i,-s}) \mathbf{R}_1 \right\},$$

$$P_{s3}(x_s) = \frac{1}{n} \sum_{i=1}^{n} \frac{w_{-s}(\mathbf{X}_{i,-s})}{\varphi(x_s, \mathbf{X}_{i,-s})} \left\{ e_l^T S_\alpha^{-1} \mathbf{Z}_s^T \mathbf{W}_s(x_s, \mathbf{X}_{i,-s}) \mathbf{R}_2(\mathbf{X}_{i,-s}) \right\}.$$

One can then write (2.22) as

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} w_{-s}(\mathbf{X}_{i,-s})\left\{\hat{\alpha}_{sl}(x_s)-\alpha_{sl}(x_s)\right\} \\
&= \sum_{i=1}^{3} P_{si}(x_s)+\sum_{i=1}^{3} D_{si}(x_s)+\sum_{r=1}^{v_2}\sum_{i=1}^{3} R_{ri}(x_s)+R_3+R_4. \qquad (2.27)
\end{aligned}$$

The proof of Theorem 2.3.1 is completed by applying assumption (A6) (b) on the bandwidths h_s and G_s, and the asymptotic results on each term of the decomposition in (2.27). These asymptotic results are presented in the following lemmas:

Lemma 2.5.6. *As* $n \to +\infty$

$$\sqrt{nh_s}R_3 = O_p\left(\sqrt{h_s}\right), \sqrt{nh_s}R_4 = O_p\left(\sqrt{h_s}\right).$$

Lemma 2.5.7. *As* $n \to +\infty$

$$\sup_{x_s \in \text{supp}(w_s)} |D_{s1}(x_s)+D_{s2}(x_s)+D_{s3}(x_s)| = o\left(h_s^{p+2}\right) \ w.p.\ 1.$$

Lemma 2.5.8. *For any fixed* $r = 1, ..., v_s$, *as* $n \to +\infty$

$$\sup_{x_s \in \text{supp}(w_s)} |R_{r1}(x_s)|+|R_{r2}(x_s)|+|R_{r3}(x_s)| = o\left(b_2^r/\sqrt{nh_s}\right) \ w.p.\ 1.$$

Lemma 2.5.9. *As* $n \to +\infty$

$$\begin{aligned}
&P_{s1} \\
&= \frac{1}{n}\sum_{j=1}^{n} \frac{w_{-s}(\mathbf{X}_{j,-s})}{\varphi(x_s,\mathbf{X}_{j,-s})}\frac{1}{h_s}K_{ls}^{*}\left(\frac{X_{js}-x_s}{h_s}, x_s, \mathbf{X}_{j,-s}, \mathbf{T}_j\right)\varphi_{-s}(\mathbf{X}_{j,-s})\,\sigma(\mathbf{X}_j,\mathbf{T}_j)\,\varepsilon_j \\
&+o_p\left\{(nh_s\log n)^{-1/2}\right\},
\end{aligned}$$

$$P_{s2}(x_s) = h_s^{p+1}\eta_{ls}(x_s)+o_p\left(h_s^{p+1}\right),$$

$$P_{s3}(x_s) = O_p\left(g_{\max}^{q_2}\right) = o_p\left\{(nh_s\log n)^{-1/2}\right\},$$

in which

$$\eta_{ls}(x_s) = \frac{1}{(p+1)!} \sum_{l'=1}^{d_1} \alpha_{l's}^{(p+1)}(x_s) \int u^{p+1} E\left\{w_{-s}(\mathbf{X}_{-s}) T_{l'} K_{ls}^{*}(u, x_s, \mathbf{X}_{-s}, \mathbf{T})\right\} du \tag{2.28}$$

with

$$K_{ls}^{*}(u, \mathbf{x}, \mathbf{T}) = e_l^T S_\alpha^{-1}(\mathbf{x}) q^*(u, \mathbf{T}) k(u), \quad q^*(u, \mathbf{T}) = (\mathbf{T}, u\mathbf{T}, ..., u^p\mathbf{T})^T. \tag{2.29}$$

Furthermore

$$\sqrt{nh_s} P_{s1} \xrightarrow{\mathcal{L}} N\left\{0, \sigma_{ls}^2(x_s)\right\}$$

in which

$$\sigma_{ls}^2(x_s) = \int \frac{w_{-s}^2(\mathbf{z}_{-s})}{\varphi^2(x_s, \mathbf{z}_{-s})} \times$$
$$K_{ls}^{*2}(u, x_s, \mathbf{z}_{-s}, \mathbf{t}) \varphi_{-s}^2(\mathbf{z}_{-s}) \sigma^2(x_s, \mathbf{z}_{-s}, \mathbf{t}) \psi(x_s, \mathbf{z}_{-s}, \mathbf{t}) du d\mathbf{z}_{-s} d\mathbf{t}. \tag{2.30}$$

Proof of Lemma 2.5.6. According to Theorem 2.2.1

$$\sqrt{nh_s} R_4 = \sqrt{nh_s} \frac{1}{n} \left\{\sum_{i=1}^{n} w_{-s}(\mathbf{X}_{i,-s})\right\} (\hat{\alpha}_{l0} - \alpha_{l0}) = \sqrt{nh_s} O_p\left(\sqrt{1/n}\right) = O_p\left(\sqrt{h_s}\right).$$

Meanwhile, according to the identify condition (1.8) and the central limit theorem for strongly mixing process (Theorem 1.7 of Bosq 1998), we have

$$\sqrt{nh_s} R_3 = \sqrt{nh_s} \frac{1}{n} \sum_{i=1}^{n} w_{-s}(\mathbf{X}_{i,-s}) \left\{\sum_{s' \neq s} \alpha_{ls'}(X_{is'})\right\} = \sqrt{nh_s} O_p\left(\sqrt{1/n}\right) = O_p\left(\sqrt{h_s}\right).$$

These two equations have completed the proof of lemma. ■

Proof of Lemmas 2.5.7 and 2.5.8. We have left these out as they are similar to Lemmas 2.5.3, 2.5.4. ■

Proof of Lemma 2.5.9. From the definition in (2.26) and using the von Mises' statistic argument

$$\begin{aligned} P_{s1} &= \frac{1}{n}\sum_{i,j=1}^{n}\frac{w_{-s}\left(\mathbf{X}_{i,-s}\right)}{\varphi\left(x_s,\mathbf{X}_{i,-s}\right)}\frac{1}{h_s}K_{ls}^{*}\left(\frac{X_{js}-x_s}{h_s},x_s,\mathbf{X}_{i,-s},\mathbf{T}_j\right)\times \\ &\quad L_{G_s}\left(\mathbf{X}_{j,-s}-\mathbf{X}_{i,-s}\right)\sigma\left(\mathbf{X}_j,\mathbf{T}_j\right)\varepsilon_j \\ &= \frac{1}{nh_s}\sum_{j=1}^{n}\int\frac{w_{-s}\left(\mathbf{z}_{-s}\right)}{\varphi\left(x_s,\mathbf{z}_{-s}\right)}K_{ls}^{*}\left(\frac{X_{js}-x_s}{h_s},x_s,\mathbf{z}_{-s},\mathbf{T}_j\right)\times \\ &\quad L_{G_s}\left(\mathbf{X}_{j,-s}-\mathbf{z}_{-s}\right)\sigma\left(\mathbf{X}_j,\mathbf{T}_j\right)\varepsilon_j\times \\ &\quad \varphi_{-s}\left(\mathbf{z}_{-s}\right)d\mathbf{z}_{-s}+o_p\left\{\left(nh_s\log n\right)^{-1/2}\right\} \end{aligned}$$

which after changing of variable $\mathbf{z}_{-s}=\mathbf{X}_{j,-s}-G_s\mathbf{v}$, one has

$$\begin{aligned} P_{s1} &= \frac{1}{nh_s}\sum_{j=1}^{n}\int\frac{w_{-s}\left(\mathbf{X}_{j,-s}-G_s\mathbf{v}\right)}{\varphi\left(x_s,\mathbf{X}_{j,-s}-G_s\mathbf{v}\right)}K_{ls}^{*}\left(\frac{X_{js}-x_s}{h_s},x_s,\mathbf{X}_{j,-s}-G_s\mathbf{v},\mathbf{T}_j\right)\mathbf{L}(\mathbf{v}) \\ &\quad \times\varphi_{-s}\left(\mathbf{X}_{j,-s}-G_s\mathbf{v}\right)d\mathbf{v}\sigma\left(\mathbf{X}_j,\mathbf{T}_j\right)\varepsilon_j+o_p\left\{\left(nh_s\log n\right)^{-1/2}\right\} \\ &= \frac{1}{nh_s}\sum_{j=1}^{n}\frac{w_{-s}\left(\mathbf{X}_{j,-s}\right)}{\varphi\left(x_s,\mathbf{X}_{j,-s}\right)}K_{ls}^{*}\left(\frac{X_{js}-x_s}{h_s},x_s,\mathbf{X}_{j,-s},\mathbf{T}_j\right)\varphi_{-s}(\mathbf{X}_{j,-s})\sigma\left(\mathbf{X}_j,\mathbf{T}_j\right)\varepsilon_j \\ &\quad +o_p\left\{\left(nh_s\log n\right)^{-1/2}\right\}. \end{aligned}$$

By assumption (A4) (a) that ε_i is independent of $\left\{\xi_j, j\leq i\right\}$, the first term is the average of a sequence of martingale differences. Then by the martingale central limit theorem of Liptser and Shirjaev (1980), the term $\sqrt{nh_s}P_{s1}$, or

$$\frac{\sqrt{nh_s}}{nh_s}\sum_{j=1}^{n}\frac{w_{-s}\left(\mathbf{X}_{j,-s}\right)}{\varphi\left(x_s,\mathbf{X}_{j,-s}\right)}K_{ls}^{*}\left(\frac{X_{js}-x_s}{h_s},x_s,\mathbf{X}_{j,-s},\mathbf{T}_j\right)\varphi_{-s}(\mathbf{X}_{j,-s})\sigma\left(\mathbf{X}_j,\mathbf{T}_j\right)\varepsilon_j$$

is asymptotically normal with mean 0 and variance

$$
\begin{aligned}
& h_s^{-1} \int \frac{w_{-s}^2(\mathbf{z}_{-s})}{\varphi^2(x_s, \mathbf{z}_{-s})} K_{ls}^{*2}\left(\frac{z_s - x_s}{h_s}, x_s, \mathbf{z}_{-s}, \mathbf{t}\right) \varphi_{-s}^2(\mathbf{z}_{-s}) \sigma^2(\mathbf{z}, \mathbf{t}) \psi(\mathbf{z}, \mathbf{t}) d\mathbf{z} d\mathbf{t} \\
= & \int \frac{w_{-s}^2(\mathbf{z}_{-s})}{\varphi^2(x_s, \mathbf{z}_{-s})} K_{ls}^{*2}(u, x_s, \mathbf{z}_{-s}, \mathbf{t}) \varphi_{-s}^2(\mathbf{z}_{-s}) \sigma^2(x_s + h_s u, \mathbf{z}_{-s}, \mathbf{t}) \times \\
& \psi(x_s + h_s u, \mathbf{z}_{-s}, \mathbf{t}) du d\mathbf{z}_{-s} d\mathbf{t} \\
= & \int \frac{w_{-s}^2(\mathbf{z}_{-s})}{\varphi^2(x_s, \mathbf{z}_{-s})} K_{ls}^{*2}(u, x_s, \mathbf{z}_{-s}, \mathbf{t}) \varphi_{-s}^2(\mathbf{z}_{-s}) \sigma^2(x_s, \mathbf{z}_{-s}, \mathbf{t}) \times \\
& \psi(x_s, \mathbf{z}_{-s}, \mathbf{t}) du d\mathbf{z}_{-s} d\mathbf{t} + o(h_s) \\
= & \sigma_{ls}^2(x_s) + o(h_s)
\end{aligned}
$$

in which the leading term $\sigma_{ls}^2(x_s)$ is as defined in (2.30). Hence we have shown that

$$\sqrt{nh_s} P_{s1} \xrightarrow{\mathcal{L}} N\left\{0, \sigma_{ls}^2(x_s)\right\}.$$

For the term $P_{s2}(x_s)$

$$
\begin{aligned}
P_{s2}(x_s) &= \frac{1}{n} \sum_{i=1}^{n} \frac{w_{-s}(\mathbf{X}_{i,-s})}{\varphi(x_s, \mathbf{X}_{i,-s})} \left\{e_l^T S_\alpha^{-1} \mathbf{Z}_s^T \mathbf{W}_s(\mathbf{X}_{i,-s}) \mathbf{R}_1\right\} \\
= & \frac{1}{n} \sum_{i,j=1}^{n} \frac{w_{-s}(\mathbf{X}_{i,-s})}{\varphi(x_s, \mathbf{X}_{i,-s})} \frac{1}{h_s} K_{ls}^{*}\left(\frac{X_{js} - x_s}{h_s}, x_s, \mathbf{X}_{i,-s}, \mathbf{T}_j\right) L_{G_s}(\mathbf{X}_{j,-s} - \mathbf{X}_{i,-s}) \\
& \times \left[\sum_{l'=1}^{d_1} \left\{\alpha_{l's}(X_{js}) - \sum_{v=0}^{p} \frac{\alpha_{l's}^{(v)}(x_s)}{v!} (X_{js} - x_s)^v\right\} T_{jl'}\right] \\
= & \int\int \frac{w_{-s}(\mathbf{x}_{-s})}{\varphi(x_s, \mathbf{x}_{-s})} \frac{1}{h_s} K_{ls}^{*}\left(\frac{z_s - x_s}{h_s}, x_s, \mathbf{x}_{-s}, \mathbf{t}\right) L_{G_s}(\mathbf{z}_{-s} - \mathbf{x}_{-s}) \\
& \left[\sum_{l'=1}^{d_1} \left\{\alpha_{l's}(z_s) - \sum_{v=0}^{p} \frac{\alpha_{l's}^{(v)}(x_s)}{v!} (z_s - x_s)^v\right\} t_{l'}\right] \times \\
& \psi(\mathbf{z}, \mathbf{t}) \varphi_{-s}(\mathbf{x}_{-s}) d\mathbf{z} d\mathbf{x}_{-s} d\mathbf{t} \left\{1 + o_p(1)\right\}.
\end{aligned}
$$

After changing of variable $z_s = x_s + h_s u$ and $\mathbf{z}_{-s} = \mathbf{x}_{-s} + G_s \mathbf{v}$, the above equals to

$$\begin{aligned}
&\int\int \frac{w_{-s}(\mathbf{x}_{-s})}{\varphi(x_s, \mathbf{x}_{-s})} K^*_{ls}(u, x_s, \mathbf{x}_{-s}, \mathbf{t})\, L(\mathbf{v}) \left\{ \sum_{l'=1}^{d_1} \frac{\alpha^{(p+1)}_{l's}(x_s)}{(p+1)!} h_s^{p+1} u^{p+1} t_{l'} \right\} \times \\
&\psi(x_s + h_s u, \mathbf{x}_{-s} + G_s \mathbf{v}, \mathbf{t})\, \varphi_{-s}(\mathbf{x}_{-s})\, du d\mathbf{v} d\mathbf{t} d\mathbf{x}_{-s} \{1 + o_p(1)\} \\
= &\frac{h_s^{p+1}}{(p+1)!} \sum_{l'=1}^{d_1} \alpha^{(p+1)}_{l's}(x_s) \int\int \frac{w_{-s}(\mathbf{x}_{-s})}{\varphi(x_s, \mathbf{x}_{-s})} K^*_{ls}(u, x_s, \mathbf{x}_{-s}, \mathbf{t})\, u^{p+1} t_{l'} \\
&\times \psi(x_s, \mathbf{x}_{-s}, \mathbf{t})\, \varphi_{-s}(\mathbf{x}_{-s})\, du d\mathbf{t} d\mathbf{x}_{-s} \{1 + o_p(1)\} \\
= &\frac{h_s^{p+1}}{(p+1)!} \sum_{l'=1}^{d_1} \alpha^{(p+1)}_{l's}(x_s) \int w_{-s}(\mathbf{x}_{-s}) \times \\
&\left\{ \int K^*_{ls}(u, x_s, \mathbf{x}_{-s}, \mathbf{t})\, u^{p+1} t_{l'} \psi(\mathbf{t} | x_s, \mathbf{x}_{-s})\, du d\mathbf{t} \right\} \varphi_{-s}(\mathbf{x}_{-s})\, d\mathbf{x}_{-s} \{1 + o_p(1)\} \\
= &\frac{h_s^{p+1}}{(p+1)!} \sum_{l'=1}^{d_1} \alpha^{(p+1)}_{l's}(x_s) \int u^{p+1} E\{w_{-s}(\mathbf{X}_{-s})\, T_{l'} K^*_{ls}(u, x_s, \mathbf{X}_{-s}, \mathbf{T})\}\, du \\
&+ o_p\left(h_s^{p+1}\right) \\
= &h_s^{p+1}(x_s)\, \eta_{ls}(x_s) + o_p\left(h_s^{p+1}\right)
\end{aligned}$$

with $\eta_{ls}(x_s)$ as defined in (2.28). Lastly, the term P_{s3} is

$$\begin{aligned}
&\frac{1}{n} \sum_{i=1}^{n} \frac{w_{-s}(\mathbf{X}_{i,-s})}{\varphi(x_s, \mathbf{X}_{i,-s})} \{e_l^T \mathbf{S}_s^{-1} \mathbf{Z}_s^T \mathbf{W}_s(\mathbf{X}_{i,-s})\, \mathbf{R}_2(\mathbf{X}_{i,-s})\} \\
= &\frac{1}{n} \sum_{i,j=1}^{n} \frac{w_{-s}(\mathbf{X}_{i,-s})}{\varphi(x_s, \mathbf{X}_{i,-s})} \frac{1}{h_s} K^*_{ls}\left(\frac{X_{js} - x_s}{h_s}, x_s, \mathbf{x}_{i,-s}, \mathbf{T}_j\right) L_{G_s}(\mathbf{X}_{j,-s} - \mathbf{X}_{i,-s}) \\
&\times \left[\sum_{l'=1}^{d_1} \sum_{s' \neq s}^{d_2} \{\alpha_{l's'}(\mathbf{X}_{js'}) - \alpha_{l's'}(\mathbf{X}_{is'})\}\, T_{jl'} \right]
\end{aligned}$$

$$\begin{aligned}
= &\int\int \frac{w_{-s}(\mathbf{x}_{-s})}{\varphi(x_s, \mathbf{x}_{-s})} \frac{1}{h_s} K^*_{ls}\left(\frac{z_s - x_s}{h_s}, x_s, \mathbf{x}_{-s}, \mathbf{t}\right) L_{G_s}(\mathbf{z}_{-s} - \mathbf{x}_{-s}) \times \\
&\left[\sum_{l'=1}^{d_1} \sum_{s' \neq s}^{d_2} \{\alpha_{l's'}(z_{s'}) - \alpha_{l's'}(x_{s'})\}\, t_{l'} \right] \psi(\mathbf{z}, \mathbf{t})\, \varphi_{-s}(\mathbf{x}_{-s})\, d\mathbf{z} d\mathbf{x}_{-s} d\mathbf{t} \{1 + o_p(1)\}
\end{aligned}$$

which after changing of variable, $z_s = x_s + h_s u$ and $\mathbf{z}_{-s} = \mathbf{x}_{-s} + G_s \mathbf{v}$, equals to

$$\int\int \frac{w_{-s}(\mathbf{x}_{-s})}{\varphi(x_s, \mathbf{x}_{-s})} K_{ls}^{*}(u, x_s, \mathbf{x}_{-s}, \mathbf{t}) L(\mathbf{v}) \times$$
$$\left[\sum_{l'=1}^{d_1} \sum_{s' \neq s}^{d_2} \left\{ \alpha_{l's'}(x_{s'} + g_{s'} v_{s'}) - \alpha_{l's'}(x_{s'}) \right\} t_{l'} \right] \times$$
$$\psi(x_s + h_s u, \mathbf{x}_{-s} + G_s \mathbf{v}, \mathbf{t}) \varphi_{-s}(\mathbf{x}_{-s}) du d\mathbf{v} d\mathbf{t} d\mathbf{x}_{-s} \{1 + o_p(1)\}$$
$$= O_p\left(g_{\max}^{q_2}\right) = o_p\left\{(nh \log n)^{-1/2}\right\}$$

by Taylor expansion to q_2-th degree of $\alpha_{l's'}$ and $(q_2 - 1)$-th degree of ψ, using assumptions (A2) and (A5) (a). Then the result follows from assumption (A1) that L is a kernel function of q_2-th order. ■

Chapter 3

Polynomial Spline Estimation

3.1 Introduction

In the last chapter, we have proposed a local polynomial based marginal integration method to estimate the unknown coefficient functions. Asymptotic distributions have also been obtained. The parameters $\{\alpha_{l0}\}_{l=1}^{d_1}$ are estimated at the parametric rate $1/\sqrt{n}$, and the nonparametric functions $\{\alpha_{ls}(x_s)\}_{l=1,s=1}^{d_1,d_2}$ are estimated at the same rate as the univariate smoothing. However, due to the integration step and its 'local' nature, the kernel type method proposed in the last chapter can be quite computationally expensive. Based on a sample of size n, to estimate the coefficient functions $\{\alpha_l(\mathbf{x})\}_{l=1}^{d_1}$ in (1.4) at any fixed point $\mathbf{x}$, a total of $(d_2+1)n$ weighted least squares estimations have to be done. So the computational burden increases dramatically as sample size n and the dimension of the tuning variables d_2 increase.

In this chapter, we propose a much faster estimation method of polynomial spline for model (1.4). In contrast to the local polynomial, polynomial spline is a global

smoothing method. It characterizes the nonparametric function components by only a finite number of parameters. One needs to solve only one least squares estimation to obtain the estimators of all the components in the coefficient functions, regardless of the sample size n and the dimension of the tuning variable d_2. Thus it reduces the computation substantially.

As an attractive alternative to the local polynomial smoothing method, polynomial spline has been used to estimate various models, for example, additive model (Stone 1985), the functional ANOVA model (Huang 1998a, 1998b; Huang, Kooperberg, Stone & Truong 2000), the varying coefficient model (Huang, Wu & Zhou 2002), and additive model for weakly dependent data (Huang & Yang 2004). The asymptotic results of above polynomial spline estimators are developed for either i.i.d. data or longitudinal data, except for Huang and Yang (2004), which gives partial derivation of the asymptotic results for time series data. In contrast, we gave complete proof of the polynomial spline estimators' rate of convergence for time series data under geometrically strongly mixing condition. Another major innovation in this dissertation is the use of approximation space with unbounded basis, while all the works before have bounded basis. For example, Huang, Wu & Zhou (2002) has imposed the assumption that $\mathbf{T} = \left(T_1, \ldots, T_{d_1}\right)^T$ in (1.4) has compactly supported distribution to make their basis bounded. The method proposed in this current work only imposes some mild moment conditions on $\mathbf{T}$.

The rest of this chapter is organized as follows. Section 3.2 discusses the identification of model (1.4). Section 3.3 presents the polynomial spline estimators, their L_2 consistency and a model selection procedure based on Bayes Information Criterion

(BIC). It is worth mentioning here that the estimation and model selection procedure developed in Section 3.3 applies not only to model (1.4), but adapts automatically to all of the four submodels mentioned before: varying coefficient model (Hastie & Tibshirani 1993), functional coefficient model (Chen & Tsay 1993b), additive model (Hastie & Tibshirani 1990, Chen & Tsay 1993a), and simple linear regression model. This feature is not shared by any local polynomial and kernel estimators. Technical assumptions and proofs are given in section 3.4.

3.2 The Set-up and Notations

As introduced in the last chapter, $\{(Y_i, \mathbf{X}_i, \mathbf{T}_i)\}_{i=1}^n$ is a sequence of strictly stationary observations generated from the additive coefficient model (1.1). But differently from before, we assume the predictor vector $\mathbf{X}$ has a compact support, since most of the polynomial spline approximation is conducted on a compact set. Without lose of generality, let the compact set be $\chi = [0,1]^{d_2}$. Accordingly, the identification condition (1.8) is simplified to

$$E\{\alpha_{ls}(X_{is})\} = 0, l = 1, \ldots, d_1, s = 1, \ldots, d_2. \tag{3.1}$$

The errors $\{\varepsilon_i\}_{i=1}^n$ in (1.1) are i.i.d with $E(\varepsilon_i|\mathbf{X}_i, \mathbf{T}_i) = 0$, $E(\varepsilon_i^2|\mathbf{X}_i, \mathbf{T}_i) = 1$, and ε_i is independent of the σ-field $\mathcal{F}_i = \sigma\{(\mathbf{X}_j, \mathbf{T}_j), j \leq i\}$ for $i = 1, \ldots, n$. The conditional variance function $\sigma^2(\mathbf{x}, \mathbf{t})$ is assumed to be continuous and bounded. The variables $(\mathbf{X}_i, \mathbf{T}_i)$ can consist of either exogenous variables or lagged values of Y_i.

Following Stone (1985), p.693, the space of s-centered square integrable functions

on $[0,1]$ is

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

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

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

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

In what follows, denote by E_n the empirical expectation, $E_n\varphi = \sum_{i=1}^n \varphi\left(\mathbf{X}_i, \mathbf{T}_i\right)/n$. We introduce two inner products on $\mathcal{M}$. For functions $m_1, m_2 \in \mathcal{M}$, the theoretical and empirical inner products are defined respectively as

$$\begin{aligned} \langle m_1, m_2 \rangle &= E\left\{m_1\left(\mathbf{X}, \mathbf{T}\right) m_2\left(\mathbf{X}, \mathbf{T}\right)\right\}, \\ \langle m_1, m_2 \rangle_n &= E_n\left\{m_1\left(\mathbf{X}, \mathbf{T}\right) m_2\left(\mathbf{X}, \mathbf{T}\right)\right\}. \end{aligned}$$

The corresponding induced norms are

$$\|m_1\|_2^2 = E m_1^2\left(\mathbf{X}, \mathbf{T}\right), \quad \|m_1\|_{2,n}^2 = E_n m_1^2\left(\mathbf{X}, \mathbf{T}\right).$$

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

Lemma 3.2.1. *Under assumptions (C1) and (C2) in the subsection 3.4.1, there exists a constant $C > 0$ such that*

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

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

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

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

which, by (3.1), equals to

$$c_3\left[\sum_{l=1}^{d_1}\alpha_{l0}^2 + \sum_{l=1}^{d_1}E\left\{\sum_{s=1}^{d_2}\alpha_{ls}(X_s)\right\}^2\right].$$

Applying Lemma 1 of Stone (1985), one gets

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

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

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

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

3.3 Polynomial Spline Estimation

3.3.1 The estimators

For each of the tuning variable direction, i.e. $s = 1, \ldots, d_2$, we introduce a knot sequence $k_{s,n}$ on $[0,1]$, which has N_n interior knots and is denoted as,

$$k_{s,n} = \left\{0 = x_{s,0} < x_{s,1} < \cdots < x_{s,N_n} < x_{s,N_n+1} = 1\right\}.$$

For any nonnegative integer p, we denote $\varphi_s = \varphi^p([0,1], k_{s,n})$, the space of functions that satisfy

(i) It is a polynomial of degree p (or less) on each of the intervals $[x_{s,i}, x_{s,i+1}), i = 0, \ldots, N_n - 1$, and $\left[x_{s,N_n}, x_{s,N_n+1}\right]$,

(ii) and if $p \geq 1$, it is $p-1$ continuously differentiable on $[0,1]$.

A function that satisfies (i), (ii) is called a polynomial spline. It is a piecewise polynomial connected smoothly on the interior knots. For example, a polynomial spline with degree $p = 0$ is a piecewise constant function, and a polynomial spline with degree $p = 1$ is a piecewise linear function and continuous on $[0,1]$. The polynomial spline space φ_s is determined by the degree of the polynomial p and the knot sequence $k_{s,n}$. Let $h_s = h_{s,n} = \max_{i=0,\ldots,N_n} |x_{s,i+1} - x_{s,i}|$, which is called mesh size of $k_{s,n}$ and can be understood as the smoothness parameter like bandwidth in the local polynomial context. Define $h = \max_{s=0,\ldots,d_2} h_s$, where h measures the overall smoothness. In what follows, denote by $C^p([0,1])$ the space of p-times continuously differentiable functions.

Lemma 3.3.1. *For $1 \leq s \leq d_2$, define $\varphi_s^0 = \{g_s : g_s \in \varphi_s, E(g_s(X_s)) = 0\}$, the space of centered polynomial splines. There exists a constant $c > 0$, so that for any $\alpha_s \in \mathcal{A}_s^0 \cap C^{p+1}([0,1])$, there exists a $g_s \in \varphi_s^0$, such that $\|\alpha_s - g_s\|_\infty \leq c \left\|\alpha_s^{(p+1)}\right\|_\infty h_s^{p+1}$.*

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

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

Lemma 3.3.1 entails that if the functions $\{\alpha_{ls}(x_s)\}_{l=1,s=1}^{d_1,d_2}$ in (1.4) are smooth, they are approximated well by centered splines $\{g_{ls}(x_s) \in \varphi_s^0\}_{l=1,s=1}^{d_1,d_2}$. As the definition of φ_s^0 depends on the unknown distribution of X_s, the empirically defined space $\varphi_s^{0,n} = \{g_s : g_s \in \varphi_s, E_n(g_{ls}) = 0\}$ is used. Intuitively, function $m \in \mathcal{M}$ is approximated by some function from the approximate space

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

Given a sequence of observations $\{(Y_i, \mathbf{X}_i, \mathbf{T}_i)\}_{i=1}^n$ generated from the regression model (1.1), the estimator of the unknown regression function m is defined as its 'best' approximation from the space $\mathcal{M}_n$, i.e.

$$\hat{m} = \text{argmin}_{m_n \in \mathcal{M}_n} \sum_{i=1}^{n} \{Y_i - m_n(\mathbf{X}_i, \mathbf{T}_i)\}^2. \tag{3.2}$$

To be precise, we introduce the following basis notations. Let $J_n = N_n + p$ and $\left\{w_{s,0}, w_{s,1}, \ldots, w_{s,J_n}\right\}$ be a set of basis for the polynomial spline space φ_s, for $s = 1, \ldots, d_2$. For example, we have used the well-known truncated power basis in

the implementation

$$\left\{1, x_s, \ldots, x_s^p, (x_s - x_{s,1})_+^p, \ldots, \left(x_s - x_{s,N_n}\right)_+^p\right\} \tag{3.3}$$

in which $(x)_+^p = (x_+)^p$. Let

$$\mathbf{w} = \left\{1, w_{1,1}, \ldots, w_{1,J_n}, \ldots, w_{d_2,1}, \ldots, w_{d_2,J_n}\right\},$$

then $\left\{\mathbf{w}t_1, \ldots, \mathbf{w}t_{d_1}\right\}$ is a set of basis of $\mathcal{M}_n$, which has dimension $R_n = d_1\{d_2 J_n + 1\}$, and (3.2) amounts to

$$\hat{m}(\mathbf{x}, \mathbf{t}) = \sum_{l=1}^{d_1} \left\{\hat{c}_{l0} + \sum_{s=1}^{d_2}\sum_{j=1}^{J_n} \hat{c}_{ls,j} w_{s,j}(x_s)\right\} t_l \tag{3.4}$$

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

$$\sum_{i=1}^{n} \left(Y_i - \sum_{l=1}^{d_1} \left\{c_{l0} + \sum_{s=1}^{d_2}\sum_{j=1}^{J_n} c_{ls,j} w_{s,j}(X_{is})\right\} T_{il}\right)^2 \tag{3.5}$$

with respect to $\{c_{l0}, c_{ls,j}, l = 1, \ldots, d_1, s = 1, \ldots, d_2, j = 1, \ldots, J_n\}$. Note Lemma 3.4.5 entails that, with probability approaching one, the sum of squares in (3.5) has a unique minimizer.

For $l = 1, \ldots, d_1, s = 1, \ldots, d_2$, denote

$$\alpha_{ls}^*(x_s) = \sum_{j=1}^{J_n} \hat{c}_{ls,j} w_{s,j}(x_s). \tag{3.6}$$

Then the estimators of $\{\alpha_{l0}\}_{l=1}^{d_1}$ and $\{\alpha_{ls}(x_s)\}_{l=1,s=1}^{d_1,d_2}$ in (1.4) are given as

$$\begin{aligned} \hat{\alpha}_{l0} &= \hat{c}_{l0} + \sum_{s=1}^{d_2} E_n \alpha_{ls}^*, \quad l = 1, \ldots, d_1; \\ \hat{\alpha}_{ls}(x_s) &= \alpha_{ls}^*(x_s) - E_n \alpha_{ls}^* \quad l = 1, \ldots, d_1, s = 1, \ldots, d_2. \end{aligned} \tag{3.7}$$

where $\{\hat{\alpha}_{ls}(x_s)\}_{l=1,s=1}^{d_1,d_2}$ are empirically centered to consistently estimate the theoretically centered function components in (1.4). These estimators are determined by the knot sequences $\{k_{s,n}\}_{s=1}^{d_2}$ and the polynomial degree p, which relates to the smoothness of the regression function. We will refer to an estimator by its degree p. For example, a linear spline fit corresponds to $p = 1$.

Theorem 3.3.1. *If* $\alpha_{ls} \in C^{p+1}([0,1])$, *for* $l = 1, \ldots, d_1, s = 1, \ldots, d_2$, *and under the assumptions (C1)-(C5) in the subsection 3.4.1, one has*

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

and for $l = 1, \ldots, d_1, s = 1, \ldots, d_2$,

$$|\hat{\alpha}_{l0} - \alpha_{l0}| = O_p\left(h^{p+1} + \sqrt{1/nh}\right), \|\hat{\alpha}_{ls} - \alpha_{ls}\|_2 = O_p\left(h^{p+1} + \sqrt{1/nh}\right).$$

Following from Theorem 3.3.1, the optimal order of h is $n^{-1/(2p+3)}$, and in that case $\|\hat{\alpha}_{ls} - \alpha_{ls}\|_2 = O_p\left(n^{-1/(2p+3)}\right)$, which is the same rate of the mean square errors of the marginal integration estimators in Chapter 2 . The constants $\{\alpha_{l0}\}_{l=1}^{d_1}$, however, are estimated at a faster parametric rate of $1/\sqrt{n}$ by the marginal integration method.

3.3.2 Knot number selection

An appropriate selection of the knot sequence is important to efficiently implement the proposed polynomial spline estimation method. Stone (1986) found that the number of knots is more crucial than its location. Thus we discuss an approach to select the number of knots N_n using the Akaike Information Criterion (AIC). For knots locations, we use either equally spaced knots (the same distance between

any adjacent knots), or quantile knots (sample quantiles with the same number of observations between any two adjacent knots).

According to Theorem 3.3.1, the optimal order of N_n is $n^{1/(2p+3)}$. Thus we propose to select the 'optimal' N_n denoted as N_n^{opt} from the set of integers in $[0.5N_r, \min(5N_r, Tb)]$ with $N_r = n^{1/(2p+3)}$ and $Tb = \{n/(4d_1) - 1\}/d_2$ which ensures that the total number of parameters in the least square estimation is less than $n/4$.

To be specific, we denote the estimator for the i-th response Y_i by $\hat{Y}_i(N_n) = \hat{m}(\mathbf{X}_i, \mathbf{T}_i)$, for $i = 1, \cdots, n$. Here $\hat{m}$ depends on the knot sequence as given in (3.4). Let $q_n = (1 + d_2 N_n) d_1$ be the total number of parameters in the least square problem (3.5). Then N_n^{opt} is the one minimizing the AIC value

$$N_n^{\text{opt}} = \underset{N_n \in [0.5N_r, \min(5N_r, Tb)]}{\operatorname{argmin}} \text{AIC}(N_n) \tag{3.8}$$

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

3.3.3 Model selection

For the full model (1.4), a natural question to ask is whether all the functions $\{\alpha_{ls}(x_s)\}_{l=1,s=1}^{d_1,d_2}$ are significant. A simpler model by setting some of $\{\alpha_{ls}(x_s)\}_{l,s=1}^{d_1,d_2}$ zero may perform as well as the full model. For $1 \le l \le d_1$, let S_l denote the set of indices of the tuning variables which are significant in the coefficient function of T_l, and S the collection of indices from all the sets S_l. The set S is called the model indices. In particular, the model indices of the full model is $S_f = \left\{S_{f1}, \ldots, S_{fd_1}\right\}$, where $S_{fl} \equiv \{1, \ldots, d_2\}, 1 \le l \le d_1$. For two indices $S = \left\{S_1, \ldots, S_{d_1}\right\}, S' = \left\{S'_1, \ldots, S'_{d_1}\right\}$, we say that $S \subset S'$ if and only if $S_l \subset S'_l$, for all $1 \le l \le d_1$ and

$S_l \neq S_l'$, for some l. The goal is to select the smallest sub-model with indices $S \subset S_f$, which gives the same information as the full additive coefficient model. Following Huang & Yang (2004), both AIC and BIC are considered.

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

$$\sum_{i=1}^{n} \left(Y_i - \sum_{l=1}^{d_1} \left\{ c_{l0} + \sum_{s \in S_l} \sum_{j=1}^{J_{n,S}} c_{ls,j} w_{s,j}\left(X_{is}\right) \right\} T_{il} \right)^2 . \tag{3.9}$$

Define

$$\hat{m}_S\left(\mathbf{x}, \mathbf{t}\right) = \sum_{l=1}^{d_1} \left\{ \hat{c}_{l0} + \sum_{s \in S_l} \sum_{j=1}^{J_{n,S}} \hat{c}_{ls,j} w_{s,j}\left(x_s\right) \right\} t_l. \tag{3.10}$$

Denote $\hat{Y}_{i,s} = \hat{m}_S\left(\mathbf{X}_i, \mathbf{T}_i\right), i = 1, \cdots, n$, $\mathrm{MSE}_S = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_{i,s}\right)^2 / n$, and the total number of parameters in (3.9) as $q_S = \sum_{l=1}^{d_1} \left\{1 + \#\left(S_l\right) J_{n,s}\right\}$. Then the submodel is selected with the smallest AIC (or BIC) values, which are defined as

$$\mathrm{AIC}_S = \log\left(\mathrm{MSE}_S\right) + 2q_S/n, \quad \mathrm{BIC}_S = \log\left(\mathrm{MSE}_S\right) + \log\left(n\right) q_S/n.$$

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

Theorem 3.3.2. *Under the same conditions as in Theorem 3.3.1, and* $N_{n,S} \asymp N_{n,S_0} \asymp n^{1/(2p+3)}$, *the BIC is consistent: for any* $S \neq S_0, \lim_{n\to\infty} P\left(BIC_S > BIC_{S_0}\right) = 1$, *hence* $\lim_{n\to\infty} P\left(\hat{S} = S_0\right) = 1$.

The condition that $N_{n,S} \asymp N_{n,S_0}$ is essential for the BIC to be consistent. The number of parameters q_S depends on the number of knots and the number of additive terms used in the model function. To ensure BIC consistency, roughly the same sufficient number of knots should be used to estimate the various models so that q_S depends only on the number of functions terms. In the implementation, we have used the same number of interior knots N_n^{opt} (see (3.8), the optimal knot number for the full additive coefficient model) in the estimation of all the submodels.

3.4 Assumption and Proofs

3.4.1 Assumptions and notations

The following assumptions are needed for our theoretical results.

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

Instead of assuming that $\mathbf{T} = \left(T_1, \ldots, T_{d_1}\right)^T$ is bounded as in Huang, Wu and Zhou (2002), we impose the following (conditional) moment conditions on $\mathbf{T}$.

(C2) (i) *There exist positive constants* $0 < c_3 \leq c_4$,*such that* $c_3 I_{d_1} \leq E(\mathbf{T}\mathbf{T}^T|\mathbf{X} = \mathbf{x}) \leq c_4 I_{d_1}$ *uniformly for all* $\mathbf{x} \in \chi$. *Here* I_{d_1} *denotes the* $d_1 \times d_1$ *identity matrix.*

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

(iii) *Furthermore we assume that there exist positive constants* c_5, c_6 *such that,*

$c_5 \leq E\left\{(T_l T_{l'})^{2+\delta_0} | \mathbf{X} = \mathbf{x}\right\} \leq c_6$ *a.s. for some* $\delta_0 > 0$ and $l, l' = 1, \ldots, d_1$.

(C3) *The* d_2 *sets of knots denoted as*

$$k_{s,n} = \left\{0 = x_{s,0} \leq x_{s,1} \leq \cdots \leq x_{s,N_n} \leq x_{s,N_n+1} = 1\right\}, s = 1, \ldots, d_2,$$

are quasi-uniform, that is, there exists $c_7 > 0$

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

Furthermore the number of interior knots $N_n \asymp n^{\frac{1}{2p+3}}$, *where* p *denotes the degree of the spline space and '* $\asymp$*' denotes both sides have the same order.*

Let $h = \max_{s=1,\ldots,d_2; j=0,\ldots,Nn} |x_{s,j+1} - x_{s,j}|$. Then (C3) implies that $h \asymp n^{-\frac{1}{2p+3}}$.

(C4) *The vector process* $\{\varsigma_t\}_{t=-\infty}^{\infty} = \{(Y_t, \mathbf{X}_t, \mathbf{T}_t)\}_{t=-\infty}^{\infty}$ *is strictly stationary and geometric strongly mixing.*

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

Assumptions (C1)-(C5) are common in the nonparametric regression literature. Assumption (C1) is the same as Condition 1, p.693 of Stone (1985), assumption (c), p.468 of Huang & Yang (2004). Assumption (C2) (i) is a direct extension of condition (ii), p.531 of Huang & Shen (2004). Assumption (C2) (ii) is a direct extension of condition (v), p.531 of Huang & Shen (2004), and of the moment condition A.2 (c) p.952 of Cai, Fan & Yao (2000). Assumption (C3) is the same as in equation (6),

p.249 of Huang (1998a), and also p.59, Huang (1998b). Assumption (C4) is similar to condition (iv), p.531 of Huang & Shen (2004). Assumption (C5) is the same as p.242 of Huang (1998a), and p.465 of Huang & Yang (2004).

3.4.2 Technical lemmas

For notational convenience, we introduce, for $1 \leq l \leq d_1, 1 \leq s \leq d_2$,

$$\tilde{\alpha}_{l0} = \hat{c}_{l0} + \sum_{s=1}^{d_2} E\alpha_{ls}^*, \quad \tilde{\alpha}_{ls}(x_s) = \alpha_{ls}^*(x_s) - E\alpha_{ls}^*. \tag{3.11}$$

Then one can rewrite $\hat{m}$ defined in (3.2) as $\hat{m} = \sum_{l=1}^{d_1} \left\{ \tilde{\alpha}_{l0} + \sum_{s=1}^{d_2} \tilde{\alpha}_{ls}(x_s) \right\} t_l$. We center the $\alpha_{ls}^*(x_s)$ in (3.11) with respect to the theoretical mean, instead of the empirical mean as $\hat{\alpha}_{ls}(x_s)$ does in (3.7). The terms $\{\tilde{\alpha}_{l0}\}_{l=1}^{d_1}$, $\{\tilde{\alpha}_{ls}(x_s)\}_{l=1,s=1}^{d_1,d_2}$ are not directly observable and serve only as the intermediate step in the proof of Theorem 3.3.1. By observing that, for $1 \leq l \leq d_1, 1 \leq s \leq d_2$

$$\hat{\alpha}_{ls}(x_s) = \tilde{\alpha}_{ls}(x_s) - E_n\tilde{\alpha}_{ls}, \quad \hat{\alpha}_{l0} = \tilde{\alpha}_{l0} - \sum_{s=1}^{d_2} E_n\tilde{\alpha}_{ls}, \tag{3.12}$$

the terms $\{\tilde{\alpha}_{l0}\}_{l=1}^{d_1}$, $\{\tilde{\alpha}_{ls}(x_s)\}_{l=1,s=1}^{d_1,d_2}$ and $\{\hat{\alpha}_{l0}\}_{l=1}^{d_1}, \{\hat{\alpha}_{ls}(x_s)\}_{l=1,s=1}^{d_1,d_2}$ differ only by a constant. In section 3.4.3, we first prove the consistency of $\{\tilde{\alpha}_{l0}\}_{l=1}^{d_1}$, $\{\tilde{\alpha}_{ls}(x_s)\}_{l=1,s=1}^{d_1,d_2}$ in Theorem 3.4.1. Then Theorem 3.3.1 follows by showing $\{E_n\tilde{\alpha}_{ls}\}_{s=1,l=1}^{d_1,d_2}$ negligible.

We use the B-spline basis for the proofs, which is equivalent to the truncated power basis used in implementation, but has nice local properties that each base is supported on a finite number of the knot intervals, see de Boor (2001) for more details. With $J_n = N_n + p$, we denote the B-spline basis of φ_s by $\mathbf{b}_s = \left\{ b_{s,0}, \ldots, b_{s,J_n} \right\}$. For the polynomial spline spaces $\{\varphi_s\}_{s=1}^{d_2}$ defined in subsection 3.3.1, define the

corresponding subspaces: $\varphi_s^0 = \{g \in \varphi_s, E\{g(X_s)\} = 0\}$. Note that the functions $\{\tilde{\alpha}_{ls}(x_s), 1 \le l \le d_1\} \in \varphi_s^0$. For $1 \le s \le d_2$,denote $\mathbf{B}_s = \left\{B_{s,1}, \ldots, B_{s,J_n}\right\}$, in which

$$B_{s,j} = \sqrt{N_n}\left(b_{s,j} - \frac{E(b_{s,j})}{E(b_{s,0})} b_{s,0}\right), j = 1, \ldots, J_n. \tag{3.13}$$

Note that under assumption (C1), $E(b_{s,0}) > 0$. Thus $B_{s,j}$ is well defined.

Now, let $\mathbf{B} = (\mathbf{1}, B_{1,1}, \ldots, B_{1,J_n}, \ldots, B_{d_2,1}, \ldots, B_{d_2,J_n})^T$, in which $\mathbf{1}$ denotes the identity function defined on χ. Define

$$\mathbf{G} = \left(\mathbf{B}t_1, \ldots, \mathbf{B}t_{d_1}\right)^T = \mathbf{B} \otimes \mathbf{t},$$

where $\mathbf{t} = \left(t_1, \ldots, t_{d_1}\right)^T$. Then $\mathbf{G}$ is a set of basis for $\mathcal{M}_n$. For easy reference of the elements in $\mathbf{G}$, we write $\mathbf{G} = \left(G_1, \ldots, G_{R_n}\right)^T$, with $R_n = d_1(d_2 J_n + 1)$. Using (3.2), one gets an alternative representation of

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

in which $\left\{\hat{c}_{l0}^*, \hat{c}_{ls,j}^*, 1 \le l \le d_1, 1 \le s \le d_2, 1 \le j \le J_n\right\}$ minimizes the sum of squares as in (3.5), with $w_{s,j}$ replaced by $B_{s,j}$. Applying Lemma 3.2.1, a function $m_n \in \mathcal{M}_n$ has a unique representation as $m_n(\mathbf{x}, \mathbf{t}) = \sum_{l=1}^{d_1} \left\{ \alpha_{l0} + \sum_{s=1}^{d_2} g_{ls}^0(x_s) \right\} t_l; g_{ls}^0 \in \varphi_s^0$. Thus for $\{\tilde{\alpha}_{l0}\}_{l=1}^{d_1}$, $\{\tilde{\alpha}_{ls}(x_s)\}_{l=1,s=1}^{d_1,d_2}$ defined in (3.11), one has $\tilde{\alpha}_{l0} = \hat{c}_{l0}^*, \tilde{\alpha}_{ls} = \sum_{j=1}^{J_n} \hat{c}_{ls,j}^* B_{s,j}(x_s), 1 \le l \le d_1, 1 \le s \le d_2$.

Theorem 3.4.1. *Under assumptions (C1)-(C5), if* $\alpha_{ls} \in C^{p+1}([0,1])$, *for* $1 \le l \le d_1, 1 \le s \le d_2$, *one has*

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

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

To prove Theorem 3.4.1, we first present the properties of the basis **G** in Lemmas 3.4.1-3.4.3.

Lemma 3.4.1. *For any $1 \leq s \leq d_2$, and spline basis $B_{s,j}$ as in (3.13), one has*

(i) $E(B_{s,j}) = 0$, *for* $j = 1, \ldots, J_n$.

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

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

Proof. (i) is trivial. (ii) follows from Theorem 5.4.2 of DeVore & Lorentz (1993), and assumptions (C1), (C3). To prove (iii), we introduce the auxiliary knots for $\{k_{s,n}\}_{s=1}^{d_2}$. Recall that $k_{s,n}$ is a knot sequence on $[0,1]$ with N_n interior knots,

$$k_{s,n} = \left\{ 0 = x_{s,0} < x_{s,1} < \cdots < x_{s,N_n} < x_{s,N_n+1} = 1 \right\}.$$

For $s = 1, \ldots, d_2$, we denote the auxiliary knots $x_{s,-p} = \ldots = x_{s,-1} = x_{s,0} = 0$, and $x_{s,N_n+p+1} = \ldots = x_{s,N_n+2} = x_{s,N_n+1} = 1$. Then

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

where $\|\cdot\|_2^*$ is defined as $\|f\|_2^* = \sqrt{\int f^2(x)\,dx}$, for any square integrable function f. Let $d_{s,j} = (x_{s,j+1} - x_{s,j-p})/(p+1)$. Then by Theorem 5.4.2 of Devore & Lorentz

(1993), there exists a positive constant C, such that the above is

$$\geq c_1 C \left[\sum_{j=1}^{J_n} a_j^2 N_n d_{s,j} + \left\{ \sum_{j=1}^{J_n} \frac{a_j \sqrt{N_n} E(b_{s,j})}{E(b_{s,0})} \right\}^2 d_{s,0} \right]$$

$$\geq c_1 C \sum_{j=1}^{J_n} a_j^2 N_n d_{s,j} \geq c_1 C (p+1)/c_7 \sum_{j=1}^{J_n} a_j^2. \blacksquare$$

Lemma 3.4.2. *There exists a constant $C > 0$, such that as $n \to \infty$, for any sets of coefficients, $\{c_{l0}, c_{ls,j}, l = 1, \ldots, d_1; s = 1, \ldots, d_2; j = 1, \ldots, J_n\}$*

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

Proof. By Lemma 3.2.1, there exists a constant $C_1 > 0$ such that

$$\left\| \sum_{l=1}^{d_1} \left(c_{l0} + \sum_{s=1}^{d_2} \sum_{j=1}^{J_n} c_{ls,j} B_{s,j} \right) t_l \right\|_2^2 \geq C_1 \left\{ \sum_{l=1}^{d_1} \left(c_{l0}^2 + \sum_{s=1}^{d_2} \left\| \sum_{j=1}^{J_n} c_{ls,j} B_{s,j} \right\|_2^2 \right) \right\}.$$

Lemma 3.4.1 provides a constant $C_2 > 0$, so that the above is

$$\geq C_1 \left\{ \sum_{l=1}^{d_1} \left(c_{l0}^2 + C_2 \sum_{s=1}^{d_2} \sum_{j=1}^{J_n} c_{ls,j}^2 \right) \right\}.$$

The lemma now follows by taking $C = \min(C_2, 1) C_1$. $\blacksquare$

Lemma 3.4.3. *Let $\langle \mathbf{G}, \mathbf{G} \rangle$ be the $R_n \times R_n$ matrix defined as $\langle \mathbf{G}, \mathbf{G} \rangle = (\langle G_i, G_j \rangle)_{i,j=1}^{R_n}$. Define $\langle \mathbf{G}, \mathbf{G} \rangle_n$ similarly as $\langle \mathbf{G}, \mathbf{G} \rangle$, but replace the theoretical inner product with the empirical inner product, and let $\mathbf{D} = \operatorname{diag}(\langle \mathbf{G}, \mathbf{G} \rangle)$. Define*

$$Q_n = \sup \left| \mathbf{D}^{-1/2} \left(\langle \mathbf{G}, \mathbf{G} \rangle_n - \langle \mathbf{G}, \mathbf{G} \rangle \right) \mathbf{D}^{-1/2} \right|,$$

where sup *is taken oven all the elements in the random matrix. Then as $n \to \infty$,*

$$Q_n = O_p \left(\sqrt{n^{-1} h^{-1} \log^2(n)} \right).$$

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

$$P\left(|\xi| \geq \epsilon\sqrt{\frac{\log^2(n)}{nh}}\right) \leq P\left(\left|\tilde{\xi}\right| \geq \epsilon\sqrt{\frac{\log^2(n)}{nh}}\right) + P(\xi \neq \tilde{\xi}) \tag{3.14}$$

in which

$$P(\xi \neq \tilde{\xi}) \leq P\left(T_{il} \neq \tilde{T}_{il}, \text{ for some } i = 1, \ldots, n\right) \leq \sum_{i=1}^{n} P\left(|T_{il}| \geq n^{\delta}\right) \leq \frac{E|T_l|^m}{n^{m\delta-1}}.$$

Also note that

$$\sup_{0 \leq x_s \leq 1} |B_{s,j}(x_s)| = \sup_{0 \leq x_s \leq 1} \left|\sqrt{N_n}\left\{b_{s,j} - \frac{E(b_{s,j})}{E(b_{s,0})}b_{s,0}\right\}\right| \leq c\sqrt{N_n},$$

for some $c > 0$. Then by Minkowski's inequality, for any positive integer $k \geq 3$

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

On the other hand

$$\begin{aligned} E\left|\tilde{\xi}_i\right|^2 &\geq \frac{E\left|B_{s,j}^2(X_s)\tilde{T}_l^2\right|^2}{2} - E^2\left\{B_{s,j}^2(X_s)\tilde{T}_l^2\right\} \\ &\geq \frac{E\left|B_{s,j}^2(X_s)T_l^2\right|^2}{2} - E^2\left\{B_{s,j}^2(X_s)T_l^2\right\} - \frac{E\left|B_{s,j}^2(X_s)T_l^2 I_{\left\{|T_l| > n^{\delta}\right\}}\right|^2}{2} \end{aligned}$$

in which, under assumption (C2)

$$\begin{aligned} E\left|B_{s,j}^2(X_s)T_l^2 I_{\left\{|T_l| > n^{\delta}\right\}}\right|^2 &\leq E\left|B_{s,j}^4(X_s)E\left(\frac{1}{n^{\delta\delta_0}}T_l^{4+\delta_0}|\mathbf{X}\right)\right| \\ &\leq \frac{c_6}{n^{\delta\delta_0}}E\left|B_{s,j}^4(X_s)\right| \leq \frac{cN_n}{n^{\delta\delta_0}}, \end{aligned}$$

where δ_0 is as in assumption (C2). Furthermore

$$E^2\left\{B_{s,j}^2\left(X_s\right)T_l^2\right\} \leq c_4^2 E^2\left\{B_{s,j}^2\left(X_s\right)\right\} \leq c,$$

$$E\left|B_{s,j}^2\left(X_s\right)T_l^2\right|^2 \geq c_5 E\left|B_{s,j}\left(X_s\right)\right|^4 \geq c_1 c_5 \int \left|B_{s,j}\left(x_s\right)\right|^4 dx_s \geq c c_1 c_5 N_n.$$

Thus $E\left|\tilde{\xi}_i\right|^2 \geq cN_n - c - \frac{cN_n}{n^{\delta\delta_0}} \geq cN_n$. So there exists a constant $c > 0$, such that for all $k > 2$

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

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

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

where

$$a_1 = 2\frac{n}{q} + 2\left(1 + \frac{\epsilon^2\log^2(n)/nh}{25m_2^2 + 5\epsilon c_r\sqrt{\log^2(n)/nh}}\right), m_2^2 = E\tilde{\xi}_i^2$$

$$a_2(k) = 11n\left(1 + \frac{5m_p^{k/(2k+1)}}{\epsilon\sqrt{\log^2(n)/nh}}\right), m_p = \left\|\tilde{\xi}_i\right\|_p.$$

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

$$P\left(\frac{1}{n}\left|\sum_{i=1}^n \tilde{\xi}_i\right| \geq \epsilon\sqrt{\frac{\log^2(n)}{nh}}\right)$$

$$\leq c\log(n)\exp\left\{-\frac{\epsilon^2\log(n)}{50c_0m_2^2}\right\} + cn^{\frac{3}{2}}\exp\left\{-\log(\rho)c_0\log(n)\right\}.$$

Thus by (3.14), taking c_0, ϵ, m large enough and use assumption (C4), one has that

$$\begin{aligned}
&\sum_{n=1}^{\infty} P\left(\sup|\langle \mathbf{G},\mathbf{G}\rangle_n - \langle \mathbf{G},\mathbf{G}\rangle| \geq \epsilon\sqrt{\frac{\log^2(n)}{nh}}\right) \\
\leq\ & \sum_{n=1}^{\infty} \{d_1 d_2(N_n+2)\}^2 \left\{c\log(n)\exp\left\{-\frac{\epsilon^2\log(n)}{25c_0}\right\}\right. \\
& \left. + cn^{3/2}\exp\{-\log(\rho)c_0\log(n)\} + \frac{E|T_l|^m}{n^{m\delta-1}}\right\} \\
<\ & \sum_{n=1}^{\infty} \{d_1 d_2(N_n+2)\}^2 n^{-3} < +\infty
\end{aligned}$$

in which $N_n \asymp n^{\frac{1}{2p+3}}$. Then the lemma follows from Borel-Cantelli Lemma and Lemma 3.4.1. ■

Lemma 3.4.4. *As $n \to \infty$, one has*

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

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

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

$$\begin{aligned}
|\langle \phi_1,\phi_2\rangle_n - \langle \phi_1,\phi_2\rangle| &= \left|\left\langle \sum_{j=1}^{R_n} a_{1j}G_j, \sum_{j=1}^{R_n} a_{2j}G_j\right\rangle_n - \left\langle \sum_{j=1}^{R_n} a_{1j}G_j, \sum_{j=1}^{R_n} a_{2j}G_j\right\rangle\right| \\
= \sum_{i,j=1}^{R_n} |a_{1i}a_{2j}|\,|\langle G_i,G_j\rangle_n - \langle G_i,G_j\rangle| &\leq Q_n \sum_{i,j=1}^{R_n} |a_{1i}a_{2j}|\,\|G_i\|_2\|G_j\|_2 \\
\leq Q_n C \sum_{i,j=1}^{R_n} |a_{1i}a_{2j}| &\leq Q_n C\sqrt{\mathbf{a}_1^T\mathbf{a}_1\mathbf{a}_2^T\mathbf{a}_2}.
\end{aligned}$$

On the other hand by Lemma 3.4.2,

$$\|\phi_1\|_2^2\|\phi_2\|_2^2 = \left(\mathbf{a}_1^T\langle \mathbf{G},\mathbf{G}\rangle\mathbf{a}_1\right)\left(\mathbf{a}_2^T\langle \mathbf{G},\mathbf{G}\rangle\mathbf{a}_2\right) \geq C^2\mathbf{a}_1^T\mathbf{a}_1\mathbf{a}_2^T\mathbf{a}_2.$$

Then

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

Lemma 3.4.4 shows that the empirical and theoretical inner products are uniformly close over the approximation space $\mathcal{M}_n$. This lemma plays the crucial role analogous to that of Lemma 10 in Huang (1998a). Our result is new in that (i) the spline basis of Huang (1998a) must be bounded, whereas the term t in basis $\mathbf{G}$ makes it possibly bounded; (ii) Huang (1998a)'s setting is i.i.d. with uniform approximation rate of $o_p(1)$, while our setting is α-mixing, broadly applicable to time series data, with approximation rate the sharper $O_p\left(\sqrt{\log^2(n)/nh}\right)$. The next lemma follows immediately from Lemmas 3.4.2 and 3.4.4.

Lemma 3.4.5. *There exists constant $C > 0$ such that except on an event whose probability tends to zero as $n \to \infty$*

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

3.4.3 Proof of mean square consistency

Proof of Theorem 3.4.1. We denote

$$\mathbf{Y} = (Y_1,\ldots,Y_n)^T, \mathbf{m} = \{m(\mathbf{X}_1,\mathbf{T}_1),\ldots,m(\mathbf{X}_n,\mathbf{T}_n)\}^T,$$

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

Note that $\mathbf{Y} = \mathbf{m} + \mathbf{E}$, and projecting this relationship onto the approximation space $\mathcal{M}_n$, one has $\hat{m} = \overline{m} + \overline{e}$, where $\hat{m}$ is defined in (3.2), and $\overline{m}$, $\overline{e}$ are the solution to (3.2)

with Y_i replaced by $m(\mathbf{X}_i, \mathbf{T}_i)$ and $\sigma(\mathbf{X}_i, \mathbf{T}_i)\varepsilon_i$ respectively. Also one can uniquely represent $\overline{m}$ as $\overline{m} = \sum_{l=1}^{d_1} \left(\overline{\alpha}_{l0} + \sum_{s=1}^{d_2} \overline{\alpha}_{ls} \right) t_l, \overline{\alpha}_{ls} \in \varphi_s^0$. With these notations, one has the error decomposition $\hat{m} - m = \overline{m} - m + \overline{e}$, where $\overline{m} - m$ is the bias term, and $\overline{e}$ is the variance term. Since for $1 \leq l \leq d_1, 1 \leq s \leq d_2, \alpha_{ls} \in C^{p+1}([0,1])$, by Lemma 3.3.1, there exist $C > 0$ and spline functions $g_{ls} \in \varphi_s^0$, such that

$$\|\alpha_{ls} - g_{ls}\|_\infty \leq Ch^{p+1}. \tag{3.15}$$

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

$$\begin{aligned} \|m - m_n\|_2 &\leq \sum_{l=1}^{d_1} \sum_{s=1}^{d_2} \|\{\alpha_{ls}(x_s) - g_{ls}(x_s)\} t_l\|_2 \leq c_4 \sum_{l=1}^{d_1} \sum_{s=1}^{d_2} \|\alpha_{ls}(x_s) - g_{ls}(x_s)\|_\infty \\ &\leq c_4 C h^{p+1}. \end{aligned} \tag{3.16}$$

Also $\|m - m_n\|_{2,n} \leq Ch^{p+1}$ a.s. Then by the definition of projection, one has

$$\|m - \overline{m}\|_{2,n} \leq \|m - m_n\|_{2,n} \leq Ch^{p+1}$$

which also implies $\|\overline{m} - m_n\|_{2,n} \leq \|m - \overline{m}\|_{2,n} + \|m - m_n\|_{2,n} \leq Ch^{p+1}$. By Lemma 3.4.4

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

Together with (3.16), one has

$$\|m - \overline{m}\|_2 = O_p\left(h^{p+1}\right). \tag{3.17}$$

Next we consider the variance term $\overline{e}$. For some set of coefficients

$$\hat{\mathbf{a}} = \left(\hat{a}_1, \hat{a}_2, \cdots, \hat{a}_{R_n}\right)^T,$$

one can write $\overline{e}(\mathbf{x},\mathbf{t}) = \sum_{j=1}^{R_n} \hat{a}_j G_j(\mathbf{x},\mathbf{t})$. By the definition of projection, one has

$$\left(\langle G_j, G_j \rangle_n\right)_{j,j'=1}^{R_n} \begin{pmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \vdots \\ \hat{a}_{R_n} \end{pmatrix} = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} G_1(\mathbf{X}_i,\mathbf{T}_i)\,\sigma(\mathbf{X}_i,\mathbf{T}_i)\,\varepsilon_i \\ \frac{1}{n}\sum_{i=1}^{n} G_2(\mathbf{X}_i,\mathbf{T}_i)\,\sigma(\mathbf{X}_i,\mathbf{T}_i)\,\varepsilon_i \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} G_{R_n}(\mathbf{X}_i,\mathbf{T}_i)\,\sigma(\mathbf{X}_i,\mathbf{T}_i)\,\varepsilon_i \end{pmatrix}.$$

Multiplying both sides with the same vector, one gets

$$\begin{pmatrix} \hat{a}_1 & \hat{a}_2 & \cdots & \hat{a}_{R_n} \end{pmatrix} \left(\langle G_j, G_j \rangle_n\right)_{j,j'=1}^{R_n} \begin{pmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \vdots \\ \hat{a}_{R_n} \end{pmatrix}$$

$$= \begin{pmatrix} \hat{a}_1 & \hat{a}_2 & \cdots & \hat{a}_{R_n} \end{pmatrix} \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} G_1(\mathbf{X}_i,\mathbf{T}_i)\,\sigma(\mathbf{X}_i,\mathbf{T}_i)\,\varepsilon_i \\ \frac{1}{n}\sum_{i=1}^{n} G_2(\mathbf{X}_i,\mathbf{T}_i)\,\sigma(\mathbf{X}_i,\mathbf{T}_i)\,\varepsilon_i \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} G_{R_n}(\mathbf{X}_i,\mathbf{T}_i)\,\sigma(\mathbf{X}_i,\mathbf{T}_i)\,\varepsilon_i \end{pmatrix}.$$

Now, by Lemmas 3.4.2, 3.4.4, the LHS is $\left\|\sum_{j=1}^{R_n} \hat{a}_j G_j\right\|_{2,n}^2 \geq C(1-Q_n)\sum_{j=1}^{R_n} \hat{a}_j^2$, while the RHS is

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

Hence

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

entailing

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

and as a result

$$\|\bar{e}\|_2^2 \leq C\left(1-Q_n\right)^{-2}\left\{\sum_{j=1}^{R_n}\left(\frac{1}{n}\sum_{i=1}^{n}G_j\left(\mathbf{X}_i,\mathbf{T}_i\right)\sigma\left(\mathbf{X}_i,\mathbf{T}_i\right)\varepsilon_i\right)^2\right\}.$$

Since ε_i is independent of $\{(\mathbf{X}_j,\mathbf{T}_j),j\leq i\}$, for $i=1,\ldots,n$, one has

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

where one makes use of the boundedness of $\sigma(\mathbf{x},\mathbf{t})$ and $E\{G_j(\mathbf{X_i},\mathbf{T}_i)\}^2$ (Lemma 3.4.1 (ii) and assumptions C2, C5). Therefore $\|\bar{e}\|_2^2=O_p(n^{-1}h^{-1})$. This, together with (3.17) prove that $\|\hat{m}-m\|_2=O_p\left(h^{p+1}+\sqrt{1/nh}\right)$. Now Lemma 3.2.1 entails that for some constant $C>0$, one has

$$\|\hat{m}-m\|_2^2\geq C\left[\sum_{l=1}^{d_1}\left\{\left(\tilde{\alpha}_{l0}-\alpha_{l0}\right)^2+\sum_{s=1}^{d_2}\|\tilde{\alpha}_{ls}-\alpha_{ls}\|_2^2\right\}\right].$$

Thus for $1\leq l\leq d_1, 1\leq s\leq d_2$,

$$|\tilde{\alpha}_{l0}-\alpha_{l0}|=O_p\left(h^{p+1}+\sqrt{1/nh}\right),\|\tilde{\alpha}_{ls}-\alpha_{ls}\|_2=O_p\left(h^{p+1}+\sqrt{1/nh}\right).\blacksquare$$

Proof of Theorem 3.3.1.

By (3.12), one only needs to show $|E_n\tilde{\alpha}_{ls}|=O_p\left(h^{p+1}+\sqrt{1/nh}\right)$, for $1\leq l\leq d_1, 1\leq s\leq d_2$. Note that $|E_n\tilde{\alpha}_{ls}|\leq|E_n\{\tilde{\alpha}_{ls}-\alpha_{ls}\}|+|E_n\alpha_{ls}|$, whose first term

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

with $\|\alpha_{ls} - g_{ls}\|_{2,n} \leq \|\alpha_{ls} - g_{ls}\|_{\infty} \leq Ch^{p+1}$, and applying Lemmas 3.2.1 and 3.4.3, one has

$$\begin{aligned}
\|\overline{\alpha}_{ls} - g_{ls}\|_{2,n} &\leq (1+Q_n)\|\overline{\alpha}_{ls} - g_{ls}\|_2 \leq (1+Q_n)\|\overline{m} - m_n\|_2 = O_p\left(h^{p+1}\right), \\
\|\tilde{\alpha}_{ls} - \overline{\alpha}_{ls}\|_{2,n} &\leq (1+Q_n)\|\tilde{\alpha}_{ls} - \overline{\alpha}_{ls}\|_{2,n} \leq (1+Q_n)\|\tilde{e}\|_2 = O_p\left(\sqrt{1/nh}\right).
\end{aligned}$$

Thus $|E_n\{\tilde{\alpha}_{ls} - \alpha_{ls}\}| = O_p\left(h^{p+1} + \sqrt{1/nh}\right)$. Since $|E_n\alpha_{ls}| = O_p(1/\sqrt{n})$, one now has $|E_n\tilde{\alpha}_{ls}| = O_p\left(h^{p+1} + \sqrt{1/nh}\right)$. Theorem 3.3.1 now follows from the triangular inequality. ■

3.4.4 Proof of BIC consistency

We denote the model space $\mathcal{M}_S$ corresponding to the submodel m_S as

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

and its spline approximation space $\mathcal{M}_{n,S}$ as

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

where $\mathcal{H}_s^0 = \{\alpha_s : E\{\alpha_s^2(X_s)\} < +\infty, E\{\alpha_s(X_s)\} = 0\}$. For $S \subset S_f$, $\mathcal{M}_S \subset \mathcal{M}_{S_d}$ and $\mathcal{M}_{n,S} \subset \mathcal{M}_{n,S_d}$. Let Proj_S (and $\mathrm{Proj}_{n,S}$) be the orthogonal least square projection operator onto $\mathcal{M}_S$ (and $\mathcal{M}_{n,S}$) with respect to the empirical inner product. Then $\hat{m}_S$ defined in (3.10) can be viewed as: $\hat{m}_S = \mathrm{Proj}_{n,S}(\mathbf{Y})$. As a special case of Theorem 3.3.1, one has the following result.

Lemma 3.4.6. *Under the same conditions as in Theorem 3.3.1, one has*

$$\|\hat{m}_S - m_S\|_2 = O_p\left(1/N_S^{p+1} + \sqrt{N_S/n}\right).$$

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

Proof of Theorem 3.3.2. Notice that

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

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

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

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

$$\begin{aligned} &\mathrm{MSE}_S - \mathrm{MSE}_{S_0} = \|\hat{m}_S - \hat{m}_{S,0}\|_{2,n}^2 = \|\hat{m}_S - \hat{m}_{S,0}\|_2^2 \{1 + o_p(1)\} \\ \leq\ & \left(\|\hat{m}_S - m\|_2^2 + \|\hat{m}_{S,0} - m\|_2^2 \right) \{1 + o_p(1)\} = O_p\left(n^{-(2p+2)/(2p+3)} \right). \end{aligned}$$

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

$$\mathrm{MSE}_S - \mathrm{MSE}_{S_0} = -\frac{n^{r-1}}{E\{\sigma^2(\mathbf{X}, \mathbf{T})\}\{1 + o_p(1)\}} - n^{r-1} \log(n) \{1 + o_p(1)\},$$

which leads to $\lim_{n \to +\infty} \left\{ P\left(\mathrm{BIC}_s - \mathrm{BIC}_{S_0} < 0 \right) \right\} = 1$, instead.

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

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

which implies that $\lim_{n \to +\infty} \left\{ P\left(\mathrm{BIC}_s - \mathrm{BIC}_{S_0} > 0 \right) \right\} = 1$. ■

Chapter 4

Examples

4.1 Monto Carlo Studies

In this section, we study the finite-sample performances of the proposed methods which include: two estimation methods (integration estimation and polynomial spline estimation), the bandwidth selection procedure for the integration method, and the model selection procedures based on nonparametric AIC and BIC proposed for the polynomial spline estimation. For those purposes, two Monte Carlo studies are designed: one with an i.i.d set up and the other one with a nonlinear time series set up. In both examples, sample sizes are taken to be $n = 100, 250$ and 500, and the number of replications is 100.

To assess the performance of the estimators of function components, we introduce the averaged integrated squared error (AISE). By denoting the estimated function of

α_{ls} in the i-th replication by $\hat{\alpha}_{i,ls}$, we define

$$\text{ISE}(\hat{\alpha}_{i,ls}) = \frac{1}{n_{\text{grid}}} \sum_{m=1}^{n_{\text{grid}}} \{\hat{\alpha}_{i,ls}(x_m) - \alpha_{ls}(x_m)\}^2 \quad \text{and} \quad \text{AISE}(\hat{\alpha}_{ls}) = \frac{1}{100} \sum_{i=1}^{100} \text{ISE}(\hat{\alpha}_{i,ls}),$$

where $\{x_m\}_{m=1}^{n_{\text{grid}}}$ are the grid points where the functions are evaluated.

4.1.1 An i.i.d example

The data are generated from the following model

$$Y = \{c_1 + \alpha_{11}(X_1) + \alpha_{12}(X_2)\} T_1 + \{c_2 + \alpha_{21}(X_1) + \alpha_{22}(X_2)\} T_2 + \varepsilon \qquad (4.1)$$

with

$$c_1 = 2, \ \ c_2 = 1, \alpha_{11}(x_1) = \alpha_{21}(x_1) = \sin(x_1), \alpha_{12}(x_2) = x_2, \ \ \alpha_{22}(x_2) = 0,$$

where $\mathbf{X} = (X_1, X_2)^T$ is uniformly distributed on $[-\pi, \pi] \times [-\pi, \pi]$, and $\mathbf{T} = (T_1, T_2)^T$ follows the bivariate standard normal distribution. The vectors $\mathbf{X}, \mathbf{T}$ are generated independently. The error term ε is a standard normal random variable and independent of $(\mathbf{X}, \mathbf{T})$.

First, to assess the performance of the data-driven bandwidth selector in section 2.4, we plot in Figure 4.2 the kernel estimates of the sampling distribution density of the ratio $\hat{h}_{1,\text{opt}}/h_{1,\text{opt}}$, where $h_{1,\text{opt}}$ is the optimal bandwidth for estimating α_{11} and α_{21}. Solid curve is for $n = 100$, dotted curve is for $n = 250$, and dot-dashed curve is for $n = 500$. One can see that the sampling distribution of the ratio $\hat{h}_{1,\text{opt}}/h_{1,\text{opt}}$ converges to 1 rapidly as the sample size increases. Similar results are also obtained for $h_{2,\text{opt}}$, the optimal bandwidth for estimating α_{12} and α_{22}. The plot is omitted.

The simulation results indicate that the proposed bandwidth selection method is reliable in this instance. The fact that the distribution of the selected bandwidth seems skewed toward larger values is due to the use of simple polynomial function as a plug-in substitute of the true regression function.

Second, we use three different methods: linear spline ($p = 1$), cubic spline ($p = 3$) and the marginal integration, to estimate this additive coefficient model. In the polynomial spline estimation, we use equally spaced knots with the number of interior knots chosen by the proposed AIC procedure. For $s = 1, 2$, let $x^i_{s,\min}$, $x^i_{s,\max}$ denote the smallest and largest observation of the variable x_s in the i-th replication. Knots are placed evenly on the intervals $\left[x^i_{s,\min}, x^i_{s,\max}\right]$, with the number of interior knots N_n selected by AIC as in subsection 3.3.2.

To make fair comparison, the functions $\{\alpha_{ls}\}^{2,2}_{l=1,s=1}$ are estimated on a grid of equally-spaced points $x_m, m = 1, ..., n_{\text{grid}}$ with $x_1 = -0.975\pi, x_{n_{\text{grid}}} = 0.975\pi, n_{\text{grid}} = 62$.

Respectively, Tables 4.2.3 and 4.2.3 report the means and standard errors (in the parentheses) of $\{\hat{c}_l\}_{l=1,2}$ and the averaged integrated squared errors (AISE) of $\{\hat{\alpha}_{ls}\}^{s=1,2}_{l=1,2}$ for the three fits. One observes for all three fits, the standard errors of the constant estimators and the AISEs of the estimators of the function components decrease as samples sizes increase. This result numerically confirms our asympototic convergence results.

Also the polynomial spline method performs overall better than the marginal integration method. The two spline fits ($p = 1$, 3) are generally comparable, but clearly the cubic fit ($p = 3$) is slightly better than the linear fit ($p = 1$) for the

large sample size ($n = 250, 500$). The fitting results are also visually presented in Figures 4.3 and 4.4, which give the plots of the 100 estimated curves using marginal integration and cubic spline fitting respectively. In both figures, (a1-a4) are plots of the 100 estimated curves for $\alpha_{11}(x_1) = \sin(x_1)$, $\alpha_{12}(x_2) = x_2$, $\alpha_{21}(x_1) = \sin(x_1)$, $\alpha_{22}(x_2) = 0$ for $n = 100$. (b1-b4) and (c1-c4) are the same as (a1-a4), but for sample size $n = 250$ and $n = 500$ respectively. They clearly illustrate the estimation improvements as sample sizes increase for both fittings. (d1-d4) give the plots of their typical estimated curves, whose ISE is the median of the 100 ISEs from the replications. The solid curve represents the true curve, the dotted curve is the typical estimated curve for $n = 100$, the dot-dashed and dashed curves are for $n = 250$ and $n = 500$ respectively, which shows even for sample size as small as 100, the fits are satisfactory.

As mentioned earlier, the polynomial spline method enjoys great computational efficiency. It takes merely 20 seconds or less to run 100 simulations using polynomial spline method on a Pentium 4 PC. The computation time is almost the same for different sample sizes. However for marginal integration method, the computation burden increases dramatically as the sample size increases. For example, it takes marginal integration about 2 hours to run 100 simulations for samples size $n = 100$; and takes about 20 hours for sample size $n = 500$.

Next we test the model selection criteria proposed in the subsection 3.3.3. For each replication used for estimation, a model selection is also conducted. Polynomial splines with $p = 1, 2, 3$ are used for estimation. The model selection results are presented in Table 4.4. For each setup, the first, second and third columns give the

number of underfitting, correct fitting and overfitting over 100 simulations. It shows that the BIC gives rather accurate selection results (more than 86% correct selection rate) even when the sample size is as small as 100, and gives absolute correct selections when sample sizes increase to 250 and 500. This confirms our assertion that BIC is consistent. Compared with BIC, AIC tends to over-fit. But AIC has the advantage that it never under-fit.

4.1.2 A nonlinear autoregressive example

In this example, the data are generated from a nonlinear autoregressive time series model

$$\begin{aligned} Y_t &= \{c_1 + \alpha_{11}(Y_{t-1}) + \alpha_{12}(Y_{t-2})\} Y_{t-3} + \{c_2 + \alpha_{21}(Y_{t-1}) \\ &\quad + \alpha_{22}(Y_{t-2})\} Y_{t-4} + 0.1\varepsilon_t, \end{aligned} \tag{4.2}$$

with $c_1 = 0.2, c_2 = -0.3$ and

$$\begin{aligned} \alpha_{11}(u) &= (0.3+u)\exp(-4u^2), \alpha_{12}(u) = 0.3/\left\{1+(u-1)^4\right\}, \\ a_{21}(u) &= 0, \alpha_{22}(u) = -(0.6+1.2u)\exp(-4u^2). \end{aligned}$$

The ε_t is the i.i.d. standard normal noise.

In each replication, a total of $1000+n$ observations are generated and only the last n observations are used to ensure stationarity. An example of the simulated series with $n = 100$ is given in Figure 4.5.

For estimation, we have used linear polynomial spline ($p = 1$). We have used the quantile knot sequences, which is shown to be better than the equally spaced knots.

The coefficient functions $\{\alpha_{ls}\}_{l=1,s=1}^{2,2}$ are estimated on a grid of equally-spaced points on the interval $[-1, 1]$, with the number of grid points $n_{\text{grid}} = 41$.

Tables 4.5 and 4.6 summarizes the estimation results, which includes the means and standard errors (in the parentheses) of $\{\hat{c}_l\}_{l=1,2}$ and the averaged integrated squared errors (AISE) of $\{\hat{\alpha}_{ls}\}_{l=1,2}^{s=1,2}$. Similar to the i.i.d example, the estimation is shown to improve as sample sizes increase, which again supports the asymptotic result. For visual representation, the fitting results are also presented in Figures 4.6, which give the plots of the 100 estimated curves using marginal integration and cubic spline fitting respectively. In both figures, (a1-a4) are plots of the 100 estimated curves for $\{\hat{\alpha}_{ls}\}_{l=1,2}^{s=1,2}$ when $n = 100$. (b1-b4) and (c1-c4) are the same as (a1-a4), but when $n = 250$ and $n = 500$ respectively. (d1-d4) are give the plots of their typical estimated curves, whose ISE is the median of the 100 ISEs from the replications. The solid curve represents the true curve, the dotted curve is the typical estimated curve for $n = 100$, the dot-dashed and dashed curves are for $n = 250$ and $n = 500$ respectively, which shows even for sample size as small as 100, the fits are satisfactory.

The model selection results are presented in Table 4.7. AIC is found to tend to overfit, compared with BIC. Also as the degree of the polynomial spline is increased, the selection results improve, except the case that the sample size is small ($n = 100$). For the sample sizes $n = 250, 500$, we have obtained quite desirable model selection result.

4.2 Empirical Examples

4.2.1 West German GNP

In this subsection, we discuss in detail the West German real GNP data first mentioned in the introduction. Yang & Tschernig (2002) found that it had an autoregressive structure on lags 4, 2, 8 according to FPE and AIC, lags 4, 2 according to BIC, where the FPE, AIC and BIC are lag selection criteria for linear time series models as in Brockwell & Davis (1991). On the other hand, lags 4, 2, 8 are selected by the semi-parametric seasonal shift criterion, and lags 4, 1, 7 are selected by the semi-parametric seasonal dummy criterion for the slightly different series $\{\log(G_{t+4}/G_{t+3})\}_{t=1}^{120}$. Both semi-parametric criteria are developed in Yang & Tschernig (2002). According to Brockwell & Davis (1991), p.304, the lag selection criteria AIC and FPE of the linear time series models are asymptotically efficient but inconsistent, while BIC selects the correct set of variables consistently. Therefore one may fit a linear autoregressive model with either Y_{t-2}, Y_{t-4} or $Y_{t-2}, Y_{t-4}, Y_{t-8}$ as the regressors, with the understanding that the variable Y_{t-8} may be redundant for linear modeling

$$\text{Linear AR (24):} \qquad Y_t = a_1 Y_{t-2} + a_2 Y_{t-4} + \sigma\varepsilon_t, \tag{4.3}$$

$$\text{Linear AR (248):} \qquad Y_t = b_1 Y_{t-2} + b_2 Y_{t-4} + b_3 Y_{t-8} + \sigma\varepsilon_t. \tag{4.4}$$

From Table 4.2.3, it is clear that besides being more parsimonious, the linear model (4.3) has smaller average squared prediction error (ASPE), compared with the model (4.4). Thus model (4.3) is the preferred linear autoregressive model. Moreover,

Figures 4.9, 4.10 show that the scatter plots of Y_t against the two significant linear predictors, Y_{t-2} and Y_{t-4}, along with the least squares regression lines, actually vary significantly at different levels of Y_{t-1} and Y_{t-8}. Here the three levels are defined as: H, the high level, is the top 33% percent of the data, L, the lower level, is the lower 33% percent of the data, and M, the middle level, is the rest of the data. So we have fitted the additive coefficient model (1.6) in the introduction.

We use the first 110 observations for estimation and perform one-step prediction using the last 10 observations. When estimating the coefficient functions in model (1.6), we first use marginal integration with local cubic fittings. According to the bandwidth selection method in section 2.4, we use bandwidths 0.0031 and 0.0020 for estimating the functions of Y_{t-1} and Y_{t-8} respectively. The estimated coefficient functions are plotted in Figure 4.11. We have also generated 500 wild bootstrap (Mammen 1992) samples and obtain 95% point-wise bootstrap confidence intervals of the estimated coefficient functions. From Figure 4.11, one may observe that the estimated functions have obviously non-constant forms. In addition, their 95% confidence intervals can't completely cover a horizontal line passing zero in any of the four plots. This supports the hypothesis that the coefficient functions in (1.6) are significantly different from a constant. (Notice that by the restrictions proposed in (1.8), if a coefficient function is constant, it has to be zero.)

To assess the sensitivity of marginal integration estimation method to the degree of the local polynomial, we have also fitted the model (1.6) using local linear estimation (i.e. taking $p = 1$ in $\mathbf{Z}_s$). Table 4.2.3 shows that overall the marginal integration estimation for model (1.6) is not sensitive to the order of local polynomial used.

We have also applied the polynomial splines ($p = 1, 3$) to fit the model. The curve estimates are plotted in Figure 4.12, in which solid lines denote the estimation results using linear spline ($p = 1$), and dotted lines denote estimates using cubic spline ($p = 3$), which are generally agreeable with those obtained from the marginal integration. For the two linear autoregressive models, we estimate their constant coefficients by maximum likelihood method. The estimated coefficients are $\hat{a}_1 = -.2436$, $\hat{a}_2 = .5622$ and $\hat{b}_1 = -0.1191$, $\hat{b}_2 = 0.6458$, $\hat{b}_3 = 0.0704$.

Table 4.2.3 gives the ASEs (averaged squared estimation error) and ASPEs (averaged squared prediction error) of the above six fits. Spline fits overall are better than those from local polynomial. All four fits of the additive coefficients provide significant improvements over two linear autoregressive models in both estimation and prediction.

4.2.2 Wolf's annual sunspot number

In this example, we consider Wolf's annual sunspot number data for the period 1700-1987. Many authors have analyzed this data set. Tong (1990) used a TAR model with lag 8 as the tuning variable. Chen & Tsay (1993b) and Cai, Fan & Yao (2000) both used a FAR model with lag 3 as the tuning variable. Xia & Li (1999) proposed a single index model using a linear combination of lag 3 and lag 8 as the tuning variable. Motivated by those models, we propose our additive coefficient model (4.5),

in which we use both lag 3 and lag 8 as the additive tuning variables,

$$\begin{aligned} Y_t &= \{c_1 + \alpha_{11}(Y_{t-3}) + \alpha_{12}(Y_{t-8})\} Y_{t-1} + \{c_2 + \alpha_{21}(Y_{t-3}) + \alpha_{22}(Y_{t-8})\} Y_{t-2} \\ &\quad \{c_3 + \alpha_{31}(Y_{t-3}) + \alpha_{32}(Y_{t-8})\} Y_{t-3} + \sigma\varepsilon_t. \end{aligned} \tag{4.5}$$

Following the convention in the literature, we use the transformed data, where $Y_t = 2\left(\sqrt{1+X_t} - 1\right)$, X_t denotes the observed sunspot number at year t. We use the first 280 data points (Year 1700-1979) to estimate the coefficient functions, and leave out years 1980-1987 for prediction. We have used marginal integration with local cubic fitting (MI), linear spline (PS1) and cubic spline (PS3) to estimate the unknown coefficient functions. In the marginal integration, the bandwidths 6.87 and 6.52 are selected for estimating functions of Y_{t-3} and functions of Y_{t-8} respectively. The estimated coefficient functions with integration fits are plotted in Figure 4.13. The time plot of the fitted values is given in Figure 4.14, in which solid line represents the fitted values and circles represents the observed values. The fitting using splines are similarly, thus is omitted.

The averaged squared estimation errors (ASE) using integration, linear spline and cubic spline are 4.18, 3.64 and 3.72 respectively. Finally we use our estimated model to predict the sunspot numbers in 1980-1987, and compare these predictions with those based on the TAR model of Tong (1990), the FAR model of Chen & Tsay (1993), denoted as FAR1, and the following two models; the FAR model of Cai, Fan & Yao (2000) denoted as FAR2

$$\begin{aligned} Y_t &= \alpha_1(Y_{t-3}) Y_{t-1} + \alpha_2(Y_{t-3}) Y_{t-2} + \alpha_3(Y_{t-3}) Y_{t-3} \\ &\quad + \alpha_6(Y_{t-3}) Y_{t-6} + \alpha_8(Y_{t-3}) Y_{t-8} + \sigma\varepsilon_t, \end{aligned} \tag{4.6}$$

and the single index coefficient model of Xia & Li (1999) denoted as SIND

$$\begin{aligned} Y_t &= \phi_0\left\{g_4\left(\theta, Y_{t-3}, Y_{t-8}\right)\right\} + \phi_1\left\{g_4\left(\theta, Y_{t-3}, Y_{t-8}\right)\right\} Y_{t-1} \\ &+\phi_2\left\{g_4\left(\theta, Y_{t-3}, Y_{t-8}\right)\right\} Y_{t-2} + \phi_3\left\{g_4\left(\theta, Y_{t-3}, Y_{t-8}\right)\right\} Y_{t-3} \\ &+\phi_4\left\{g_4\left(\theta, Y_{t-3}, Y_{t-8}\right)\right\} Y_{t-8} + \sigma\varepsilon_t \end{aligned} \tag{4.7}$$

in which $g_4\left(\theta, Y_{t-3}, Y_{t-8}\right) = \cos\left(\theta\right) Y_{t-3} + \sin\left(\theta\right) Y_{t-8}$.

According to Condition (A.1) b, p.952 of Cai, Fan & Yao (2000), the conditional density of Y_{t-3} given the variables $(Y_{t-1}, Y_{t-2}, Y_{t-3}, Y_{t-6}, Y_{t-8})$ should be bounded. It is clear, however, that Y_{t-3} is completely predictable from $(Y_{t-1}, Y_{t-2}, Y_{t-3}, Y_{t-6}, Y_{t-8})$, and hence the distribution of Y_{t-3} given the variables $(Y_{t-1}, Y_{t-2}, Y_{t-3}, Y_{t-6}, Y_{t-8})$ is a probability mass at one point, not a continuous distribution with any kind of density. Thus, the use of model (4.6) has not been theoretically justified. Similarly, model (4.7) is also not theoretically justified, since according to Condition C5, p.1277 of Xia & Li (1999), the conditional density of $g_4\left(\theta, Y_{t-3}, Y_{t-8}\right)$ given the variables $(Y_{t-1}, Y_{t-2}, Y_{t-3}, Y_{t-8}, Y_t)$ should be bounded, whereas again, the distribution of $g_4\left(\theta, Y_{t-3}, Y_{t-8}\right)$ given the variables $(Y_{t-1}, Y_{t-2}, Y_{t-3}, Y_{t-8}, Y_t)$ is also a point mass. In addition, we illustrate that model (4.7) is unidentifiable. For any set of functions $\{\phi_0, \ldots, \phi_4\}$ that satisfy (4.7), one can always pick an arbitrary nonzero function $f\left(u\right)$ and define

$$\begin{aligned} \widetilde{\phi}_0\left(u\right) &= \phi_0\left(u\right) + uf\left(u\right), \widetilde{\phi}_1\left(u\right) = \phi_1\left(u\right), \widetilde{\phi}_2\left(u\right) = \phi_2\left(u\right), \\ \widetilde{\phi}_3\left(u\right) &= \phi_3\left(u\right) - \cos\left(\theta\right) f\left(u\right), \widetilde{\phi}_4\left(u\right) = \phi_4\left(u\right) - \sin\left(\theta\right) f\left(u\right). \end{aligned}$$

It is straightforward to verify that the new set of functions $\left\{\widetilde{\phi}_0, \ldots, \widetilde{\phi}_4\right\}$ satisfy (4.7) as well. One possible fix of this problem is to drop either one of the terms

$\phi_3\{g_4(\theta, Y_{t-3}, Y_{t-8})\}Y_{t-3}$ and $\phi_4\{g_4(\theta, Y_{t-3}, Y_{t-8})\}Y_{t-8}$ from (4.7), then the model is fully identifiable and satisfies Condition C5, p.1277 of Xia & Li (1999). Hence the current form of (4.7) may be considered an overfitting anomaly.

Despite the fact that models (4.6) and (4.7) suffer these theoretical deficiencies, we have listed the average absolute prediction errors (AAPE) and averaged squared prediction errors (ASPE) of model TAR, FAR1, FAR2, SIND and our proposed model in Table 4.2.3. By comparing the AAPEs and ASPEs, our model outperforms TAR, FAR1 and FAR2, while the unidentifiable SIND model has smallest AAPE and ASPE. We believe that this superior forecasting power of (4.7) is due to the prediction advantage of overfitting models. For example, in forecasting of linear time series, the overfitting AIC/FPE selects models more powerful than the consistent BIC, see, for instance, the discussion of Brockwell & Davis (1991), p.304.

4.2.3 Housing price

In this example, we consider the Tucson housing price data, which consists of 2971 sales observations during year 1998. The data contains unit-specific information on sale price (PRICE), lot size (LOT), age of dwelling in years (AGE), square footage (SQFT), and the absolute locations of housing units which are represented by Cartesian $\{x, y\}$ coordinates, derived from latitude and longitude, referenced against the southwestern-most observation. We are interested in predicting the unit housing price from its determinants: LOT, AGE, SQFT, and absolute location. In determining the housing price, the interactions between the absolute location and the other determi-

nants: AGE, SQFT, and LOT are found to be significant. Interestingly, Fik *et al.* (2003) modeled the interaction by including a polynomial expansion (to the third degree) of the property's $\{x, y\}$ coordinates in the coefficients of the other determinants. The following terms are found to be significant: AGE, AGE^2, SQFT^2, LOT^2, $\text{AGE} * \text{SQFT}$, $\text{LOT} * y^2$, $\text{LOT} * y^3$, $\text{SQFT} * y^3$, $\text{SQFT} * x^2$, $\text{SQFT} * x^2 * y$. We will refer to this model as the "parametric model" later. (For detailed discussion, see Model 3 in Fik *et al.* 2003). Instead of restricting the absolute location interaction with the other determinants through a polynomial expansion, the interaction is modeled by including additive smooth functions of the location coordinates into the coefficient functions. Considering that the term $\text{AGE} * \text{SQFT}$ is significant, we have also included AGE in the coefficient functions, naturally resulting in the following additive coefficient model

$$\begin{aligned} \log(\text{PRICE}) &= \text{LOT}^2 + \alpha_0(\text{AGE}) + \{c_1 + \alpha_{11}(x) + \alpha_{12}(y) + \alpha_{13}(\text{AGE})\}\,\text{SQFT} \\ &\quad + \{c_2 + \alpha_{21}(x) + \alpha_{22}(y) + \alpha_{23}(\text{AGE})\}\,\text{LOT} + \varepsilon. \end{aligned} \tag{4.8}$$

To see if Model (4.8) is redundant, the following two sub-models are also considered

$$\begin{aligned} \log(\text{PRICE}) &= \text{LOT}^2 + \alpha_0(\text{AGE}) + \{c_1 + \alpha_{11}(x) + \alpha_{12}(y) + \alpha_{13}(\text{AGE})\}\,\text{SQFT} \\ &\quad + \{c_2 + \alpha_{21}(x) + \alpha_{22}(y)\}\,\text{LOT} + \varepsilon, \end{aligned} \tag{4.9}$$

$$\begin{aligned} \log(\text{PRICE}) &= \text{LOT}^2 + \alpha_0(\text{AGE}) + \{c_1 + \alpha_{11}(x) + \alpha_{12}(y)\}\,\text{SQFT} \\ &\quad + \{c_2 + \alpha_{21}(x) + \alpha_{22}(y)\}\,\text{LOT} + \varepsilon. \end{aligned} \tag{4.10}$$

To estimate the above three additive coefficient models, quadratic spline ($p = 2$) is used. Table 4.2.3 gives the adjusted R-squares and BICs of the parametric model

in Fik *et al.* (2003) and three additive coefficient models. Among those four models, the full additive coefficient model (4.8) gives the highest adjusted-R^2 (0.855), and model (4.10) gives the smallest BIC (-3.62). According to either adjusted-R^2 or BIC, the semiparametric additive coefficient model is preferred over the parametric one.

As did in Fik *et al.* (2003), we have also compared the four models in terms of their prediction accuracies. Similar to Fik *et al.* (2003), a sample of 2471 observations is randomly selected from the database and used to estimate the model. Using the estimated model, we predict the housing prices of the remaining 500 samples. The prediction performance is evaluated via averaged absolute prediction error (AAPE) and the percentage of predicted prices within 10% of actual prices. As given in Table 4.2.3, the additive coefficient models show great improvements in prediction too, compared with the parametric model. Among them, the model (4.10) is most appealing, it has the simple interpretable structure, but gives good predictions as the full model (4.8) does.

	ASE	ASPE
Integration fit, $p = 1$	0.000201	0.000085
Integration fit, $p = 3$	0.000205	0.000077
Spline fit $p = 1$	0.000194	0.000076
Spline fit $p = 3$	0.000179	0.000081
Linear AR fit 24	0.000253	0.000112
Linear AR fit 248	0.000258	0.000116

Table 4.1: GNP data: the ASEs and ASPEs of six fits.

Integration fit	$c_1 = 2$	$c_2 = 1$
$n = 100$	1.9737(0.3574)	1.0406(0.2503)
$n = 250$	2.0299(0.2410)	1.0056(0.1490)
$n = 500$	1.9786(0.1680)	1.0026(0.1111)
Spline fit $p = 1$		
$n = 100$	1.9776(0.0497)	0.9862(0.0208)
$n = 250$	2.0389(0.0188)	1.0023(0.0065)
$n = 500$	2.0091(0.0136)	1.0035(0.0030)
Spline fit $p = 3$		
$n = 100$	1.9894(0.0568)	1.0061(0.0194)
$n = 250$	1.9936(0.0225)	1.0011(0.0059)
$n = 500$	1.9871(0.0074)	0.9974(0.0024)

Table 4.2: Simulated i.i.d example: estimation of constants.

Integration fit	α_{11}	α_{12}	α_{21}	α_{22}
$n = 100$	0.1609	0.2541	0.1205	0.2761
$n = 250$	0.0568	0.0963	0.0338	0.0649
$n = 500$	0.0295	0.0483	0.0191	0.0310
Spline fit $p = 1$				
$n = 100$	0.0742	0.0883	0.0824	0.0626
$n = 250$	0.0314	0.0369	0.0271	0.0214
$n = 500$	0.0138	0.0191	0.0143	0.0104
Spline fit $p = 3$				
$n = 100$	0.0944	0.1279	0.1023	0.0814
$n = 250$	0.0258	0.0396	0.0227	0.0232
$n = 500$	0.0120	0.0155	0.0128	0.0096

Table 4.3: Simulated i.i.d example: estimation of function components.

n	BIC								
	$p = 1$			$p = 2$			$p = 3$		
100	4	96	0	3	97	0	8	91	1
250	0	100	0	0	100	0	0	100	0
500	0	100	0	0	100	0	0	100	0
n	AIC								
100	0	83	17	0	87	13	0	82	18
250	0	86	14	0	89	11	0	87	13
500	0	93	7	0	93	7	0	94	6

Table 4.4: Simulated i.i.d example: model selection with BIC and AIC.

Spline fit $p = 1$	$c_1 = 0.2$	$c_2 = -0.3$
$n = 100$	0.2504(0.0481)	−0.2701(0.0374)
$n = 250$	0.1983(0.0271)	−0.2936(0.0279)
$n = 500$	0.1989(0.0202)	−0.2975(0.0209)

Table 4.5: Simulated nonlinear AR model: estimation of constants.

Spline fit $p = 1$	α_{11}	α_{12}	α_{21}	α_{22}
$n = 100$	0.0113	0.0050	0.0042	0.0195
$n = 250$	0.0030	0.0021	0.0025	0.0039
$n = 500$	0.0015	0.0011	0.0016	0.0019

Table 4.6: Simulated nonlinear AR model: estimation of function components.

n	BIC								
	$p=1$			$p=2$			$p=3$		
100	11	88	1	0	69	31	1	94	5
250	0	100	0	0	100	0	0	100	0
500	0	100	0	0	100	0	0	100	0
n	AIC								
100	0	89	11	4	95	1	0	88	12
250	0	90	10	0	89	11	0	90	10
500	0	91	9	0	93	7	0	92	8

Table 4.7: Simulated nonlinear AR model: model selection with AIC and BIC.

Year	X_t	TAR	FAR1	FAR2	SIND	MI	PS1	PS3
1980	154.7	5.5	13.8	1.4	2.1	14.9	0.2	4.5
1981	140.5	1.3	0.0	11.4	1.7	2.4	8.2	5.3
1982	115.9	19.5	10.0	15.7	2.6	17.5	9.8	9.9
1983	66.6	4.8	3.3	10.3	2.4	1.37	14.2	12.8
1984	45.9	14.8	3.8	1.0	2.3	5.92	0.3	1.6
1985	17.9	0.2	4.6	2.6	7.6	1.96	5.9	6.5
1986	13.4	5.5	1.3	3.1	4.2	0.57	2.3	4.6
1987	29.2	0.7	21.7	12.3	13.2	0.7	2.9	1.1
AAPE		6.6	7.3	7.2	4.5	5.7	5.5	5.8
ASPE		85.6	101.1	81.6	34.3	71.9	52.08	47.17

Table 4.8: Wolf's Sunspot Number: out-of-sample absolute prediction errors.

	Adjusted-R^2	BIC	AAPE	Percentage
Model (4.8)	0.855	-3.61	14932.6	60.8
Model (4.9)	0.853	-3.61	15127.49	61
Model (4.10)	0.852	-3.62	14940.36	61.2
Parametric model	0.834	-3.57	16464.12	54.6

Table 4.9: Tucson housing price: estimation and prediction results.

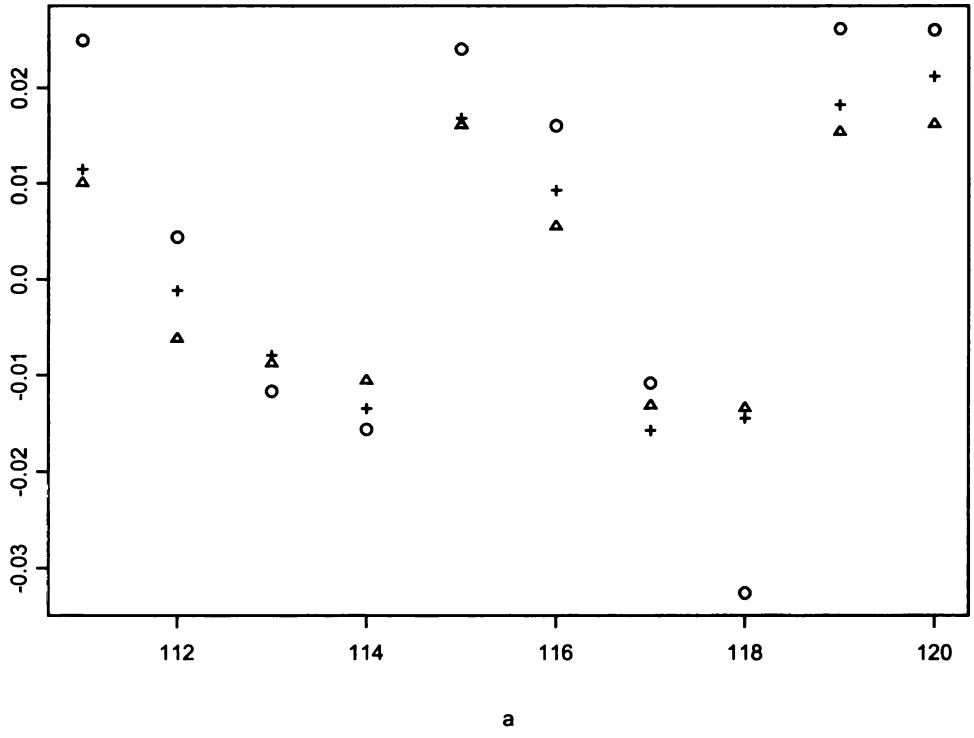

Figure 4.1: GNP data: one-step prediction performance.

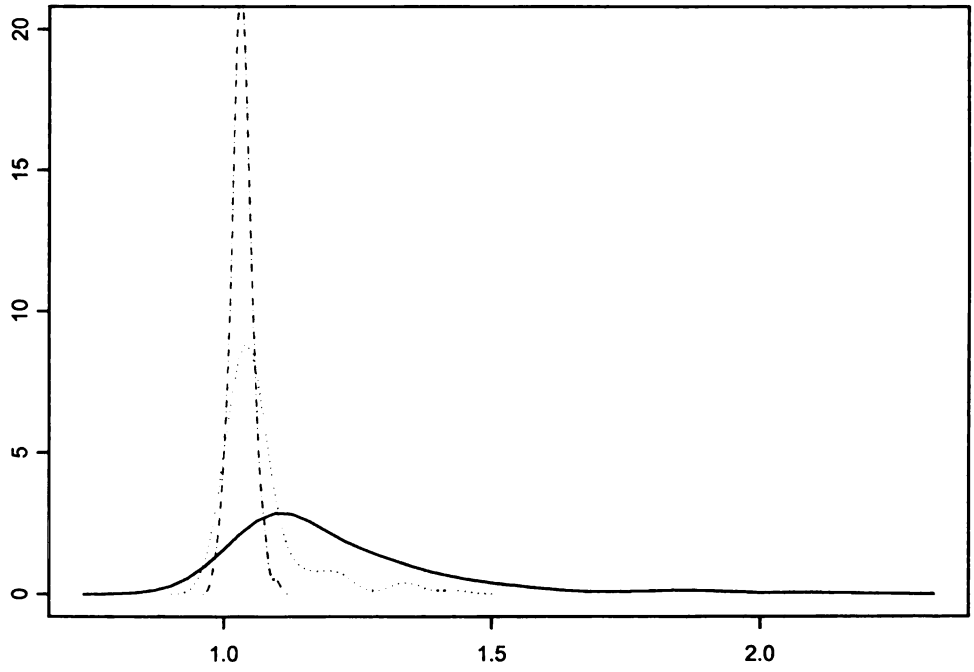

Figure 4.2: Kernel density estimates of $\hat{h}_{1,\text{opt}}/h_{1,\text{opt}}$.

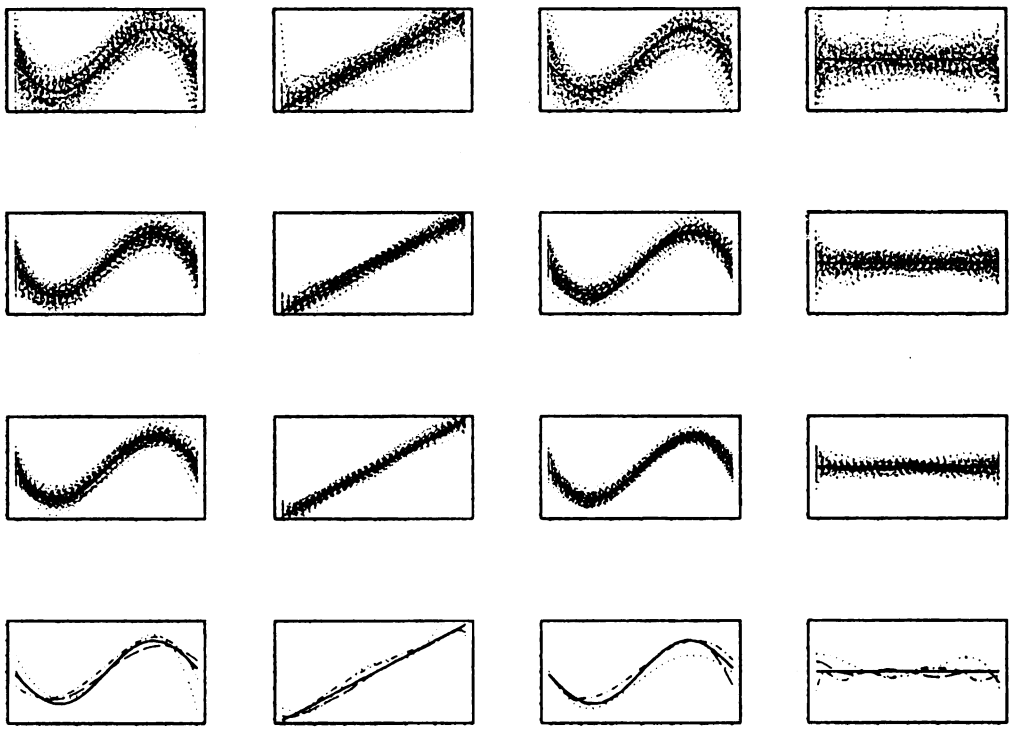

Figure 4.3: Plots of the estimated coefficient functions using marginal integration.

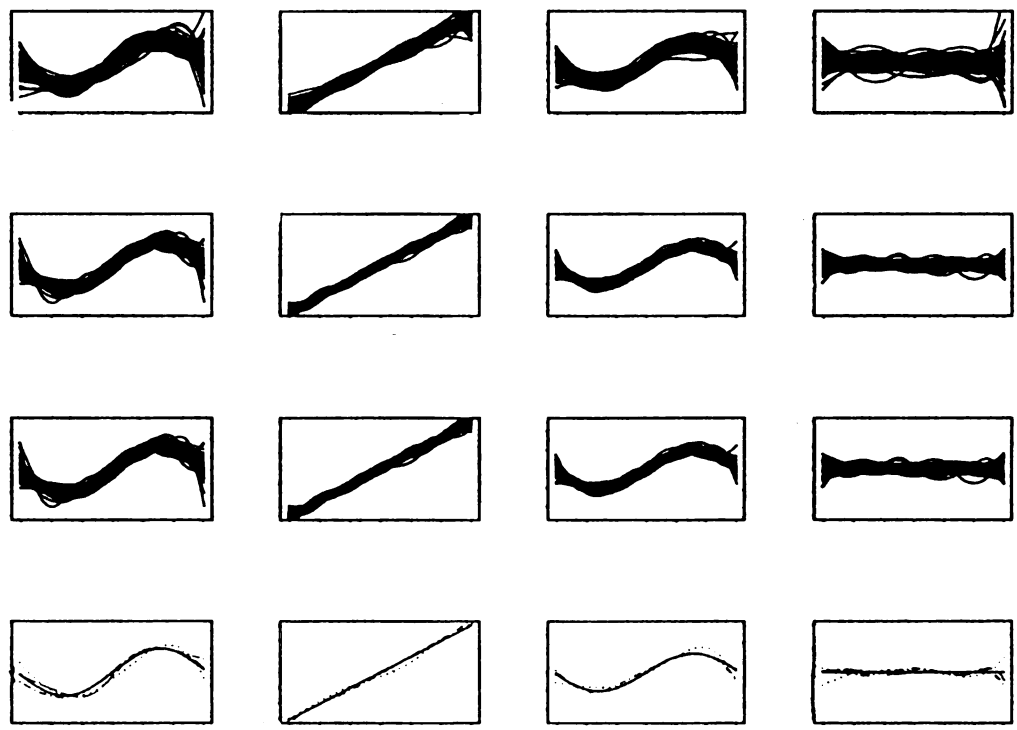

Figure 4.4: Plots of the estimated coefficient functions using cubic spline.

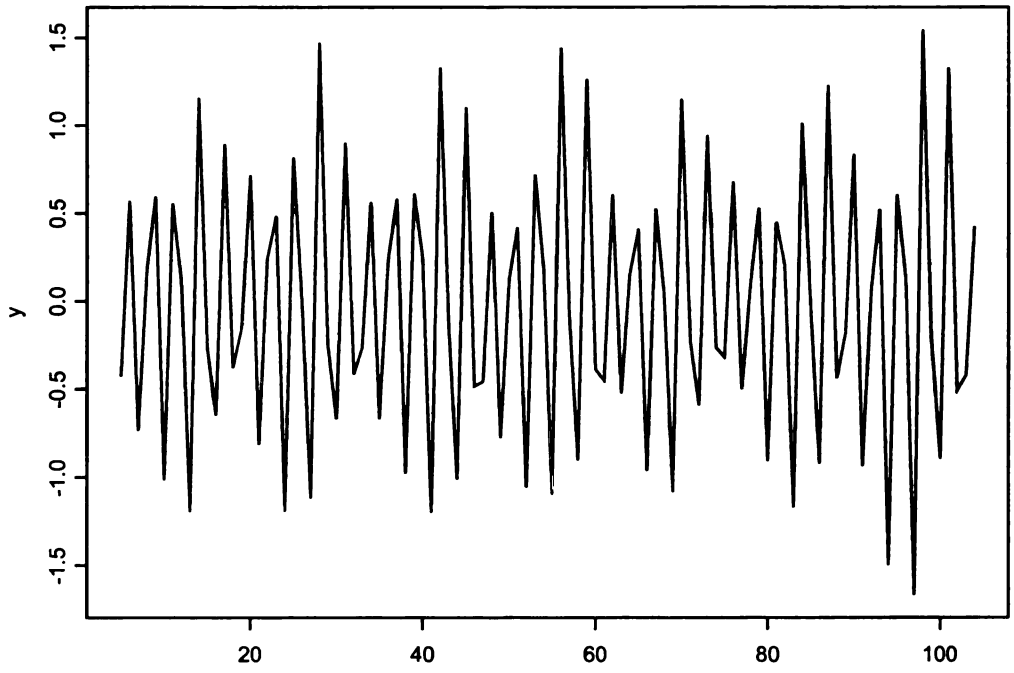

Figure 4.5: Time plot of a simulated series from model (4.2), with $n = 100$.

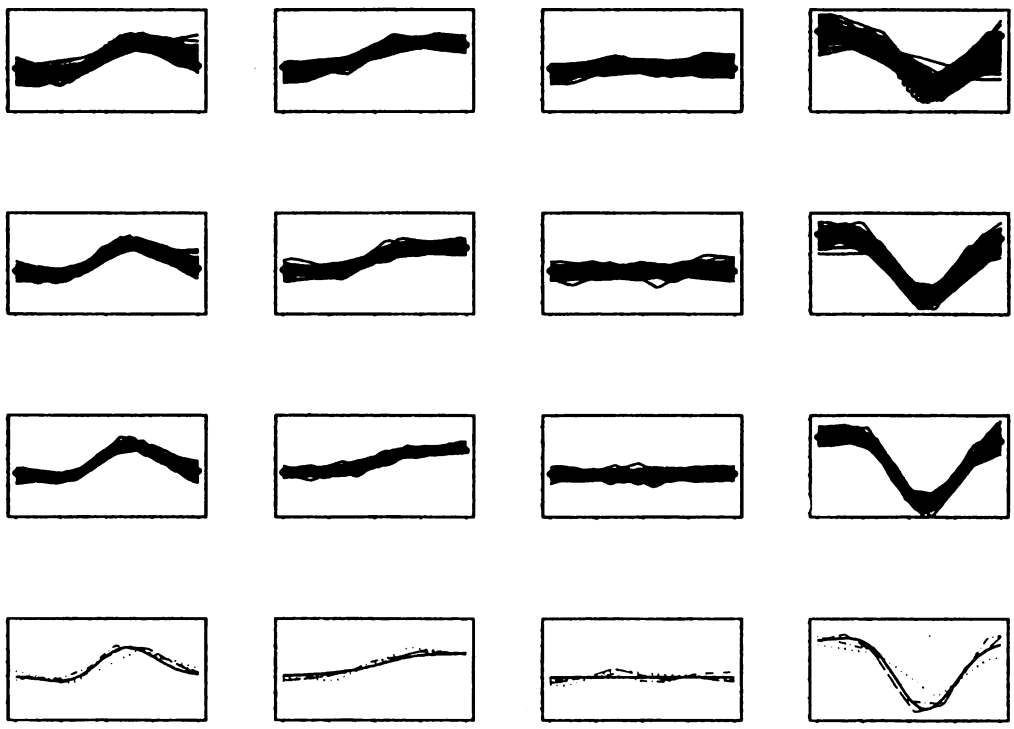

Figure 4.6: Plots of the estimated coefficient functions using linear spline.

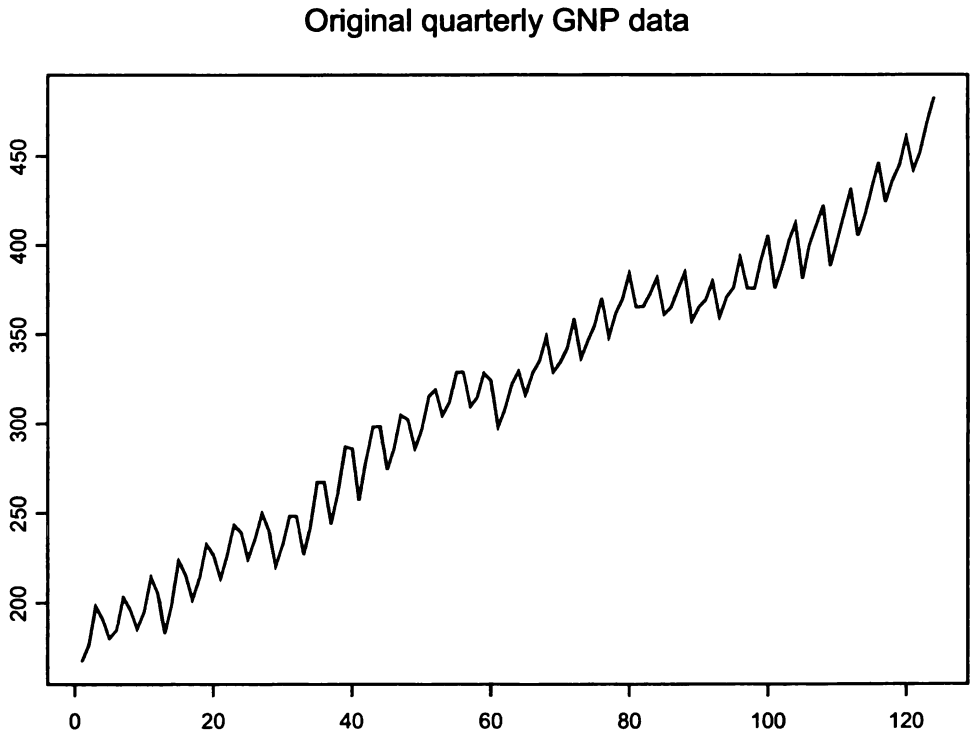

Figure 4.7: GNP data: time plot of the series $\{G_t\}_{t=1}^{124}$.

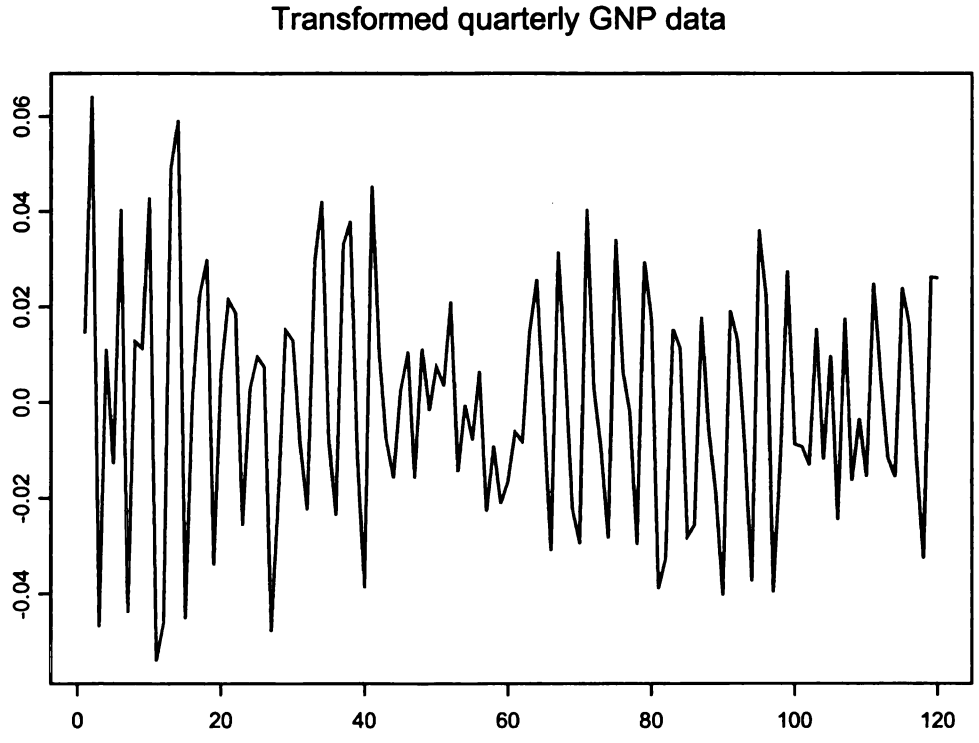

Figure 4.8: GNP data after transformation: time plot of the series $\{Y_t\}_{t=1}^{120}$.

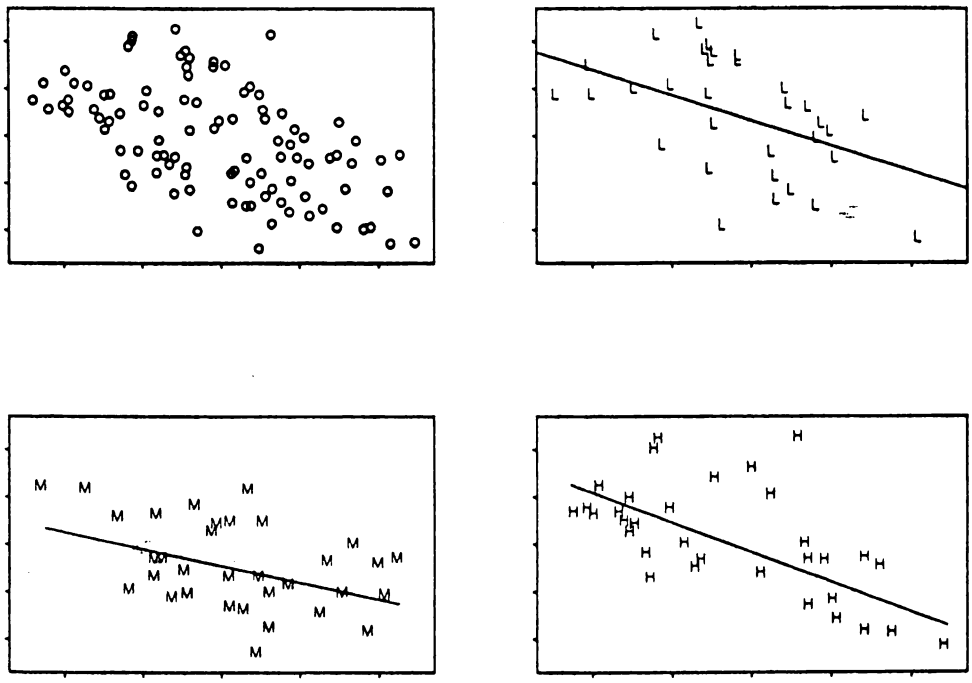

Figure 4.9: Scatter plot of Y_t, Y_{t-2} at three levels of Y_{t-1}.

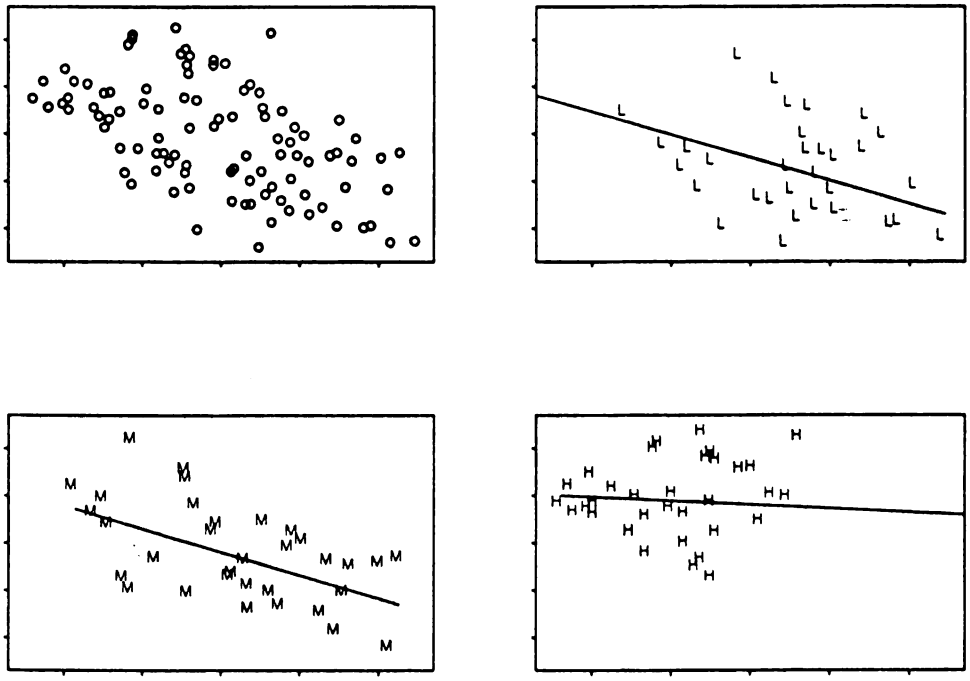

Figure 4.10: Scatter plot of Y_t, Y_{t-2} at three levels of Y_{t-8}.

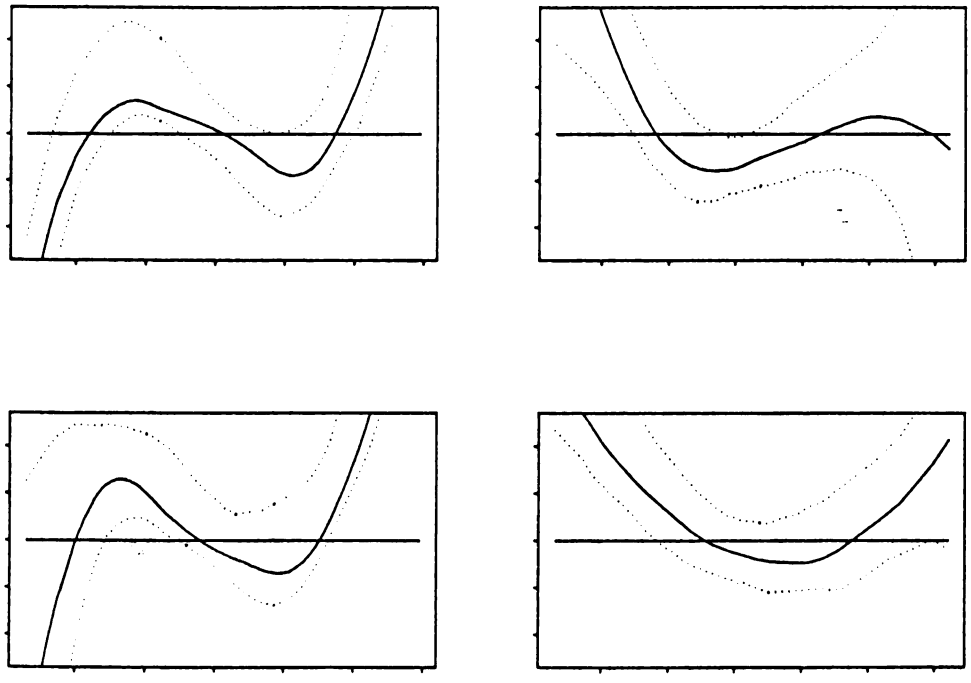

Figure 4.11: Estimated functions and their bootstrap 95% confidence intervals.

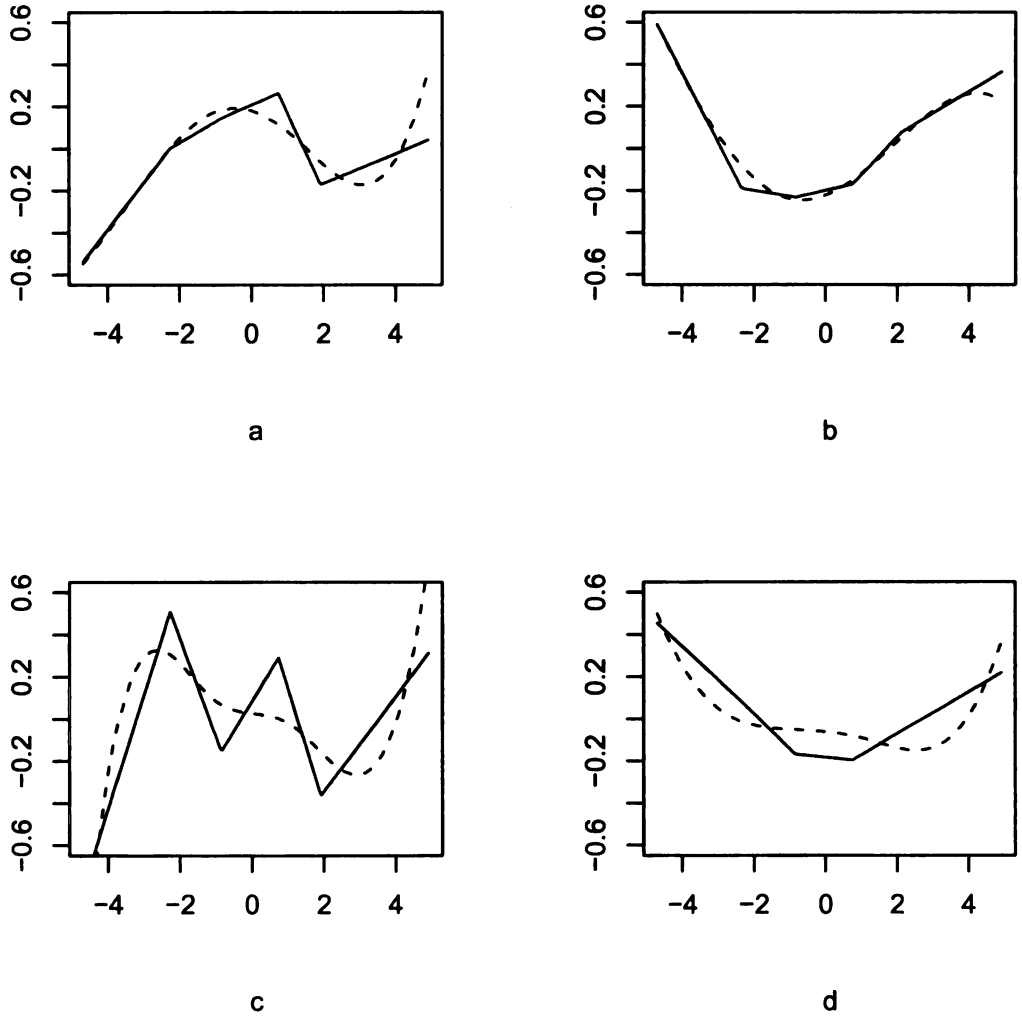

Figure 4.12: Spline approximations of the functions.

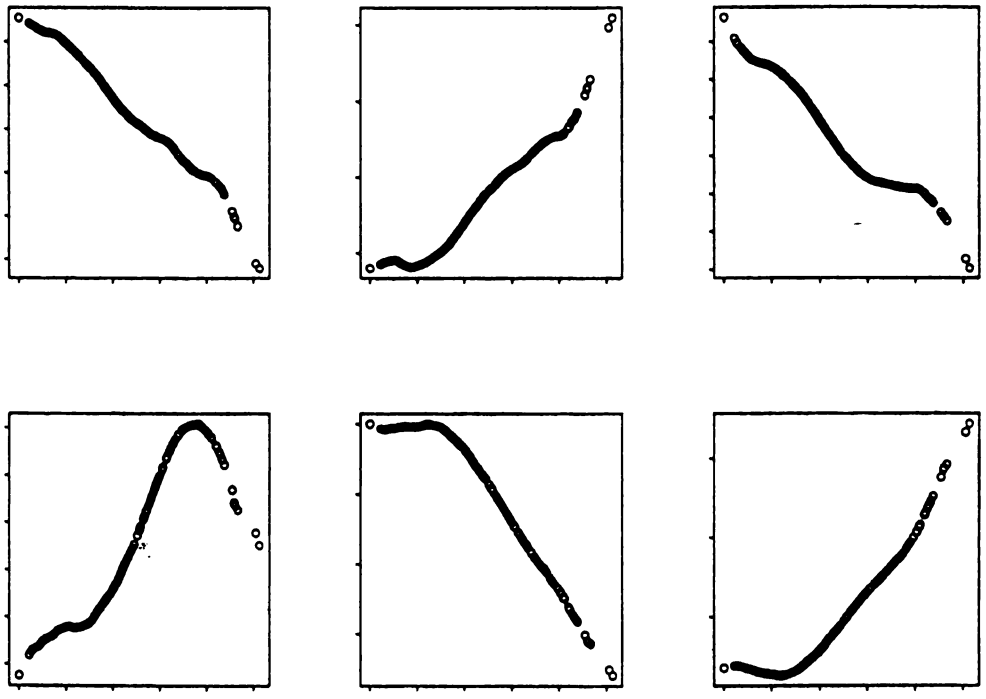

Figure 4.13: Estimated function components.

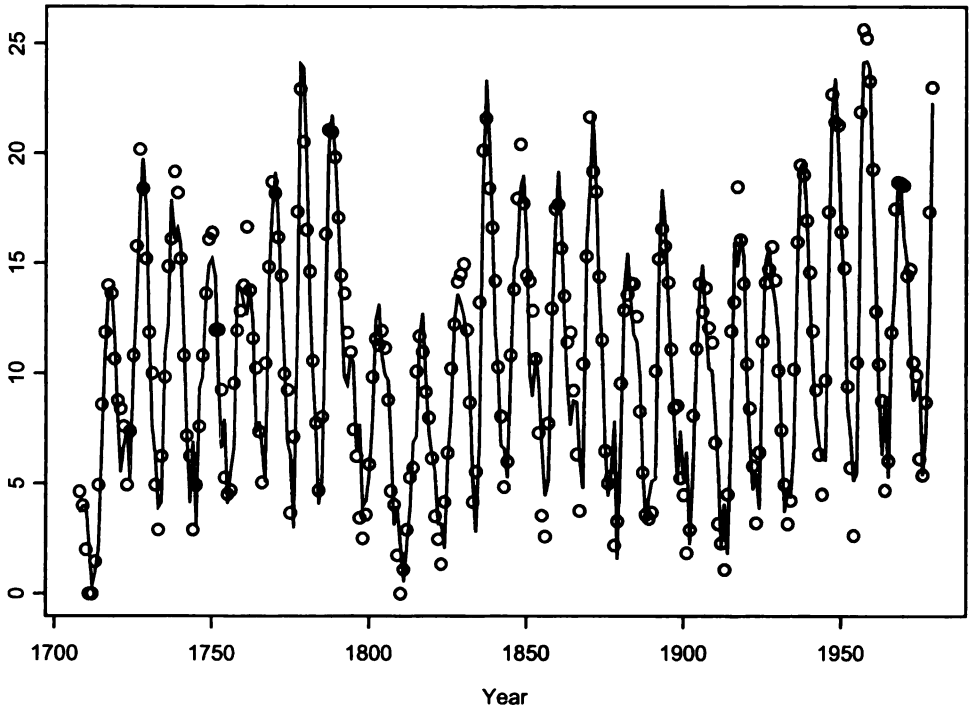

Figure 4.14: Time plot of the fitted values based on marginal integration.

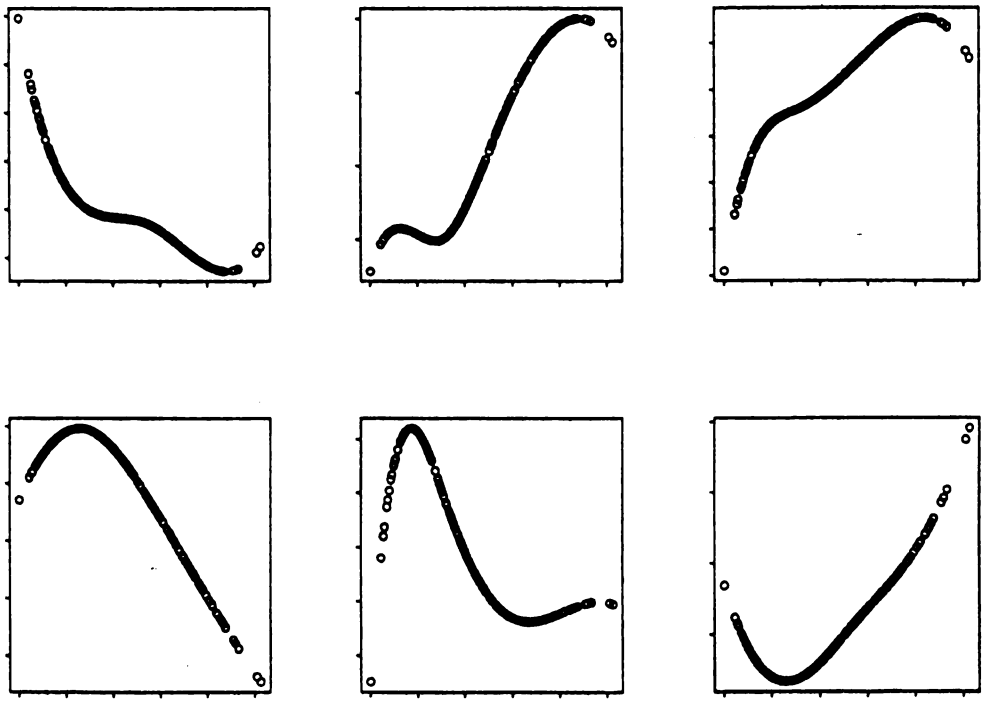

Figure 4.15: Estimated functions with cubic spline approximation.

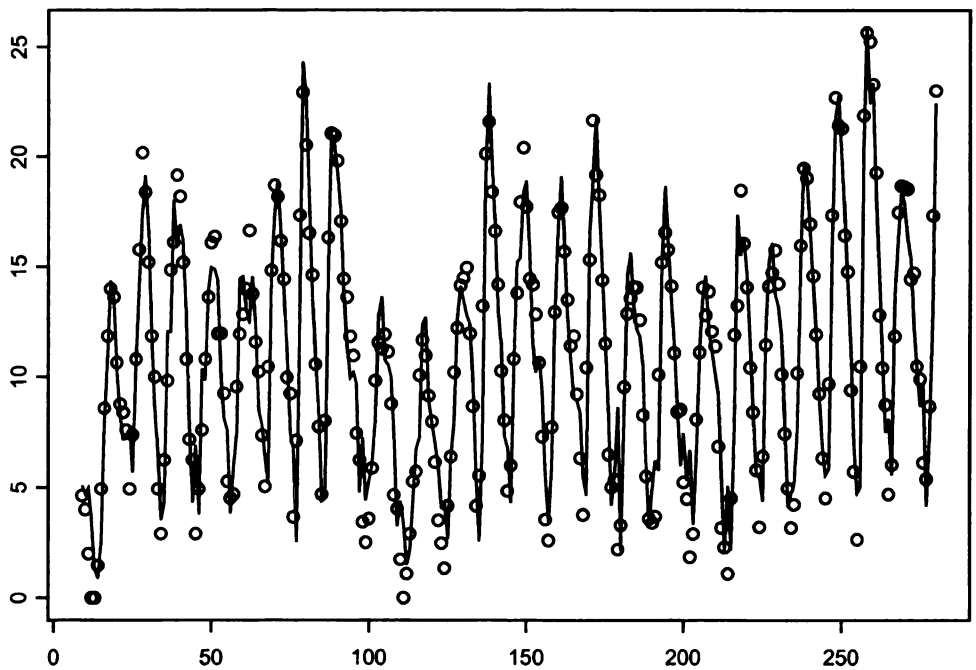

Figure 4.16: Time plot of the fitted values with cubic approximation.

Bibliography

[1] Bosq, D., 1998. *Nonparametric Statistics for Stochastic Processes.* Springer-Verlag, New York.

[2] Brockwell, P. J. and Davis, R. A., 1991. *Time Series: Theory and Methods.* Springer-Verlag, New York.

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

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

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

[6] Chui, K., 1992. *Wavelets: A tutorial in theory and applications.* Boston, Academic Press.

[7] de Boor, 2001. *A Practical Guide to Splines.* Springer, New York.

[8] Devore R. A. and Lorentz, G. G., 1993. *Constructive Approximation.* Springer-Varlag, Berlin Heidelberg.

[9] Donoho, D. L. and Johnstone, I. M., 1995. Adapting to unknown smoothness via wavelet shrinking. *J. Am. Statist. Assoc.* **90**, 1200-1224.

[10] Eilers, P. H. and Marx, B. D., 1996. Flexible smoothing with B-splines and penalties. *Statist. Sci.* **11** 89–121.

[11] Eubank, R. L., 1988. *Spline smoothing and nonparametric regression.* Marcel Dekker, Inc., New York.

[12] Fik, T. J., Ling, D. C. and Mulligan G. F., 2003. Modeling spatial variation in housing prices: A variable interaction approach. *Real Estate Economics* **V31**, 623-646.

[13] Fan, J., and Gijbels, I., 1996. *Local polynomial modelling and its applications.* Chapman & Hall, London.

[14] Härdle, W., Liang, H., and Gao, J. T., 2000. *Partially Linear Models.* Springer-Verlag, Heidelberg.

[15] Härdle, W., Kerkyacharian, G., Picard, D., and Tsybakov, A., 1998. *Wavelets, approximation, and statistical applications.* Springer-Verlag, New York.

[16] Hastie, T. J., and Tibshirani, R. J., 1990. *Generalized Additive Models.* Chapman and Hall, London.

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

[18] Hoover. D. R., Rice, J. A., Wu, C. O., and Yang. L. P., 1998. Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika.* **85**, 809-822.

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

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

[21] Huang, J. Z., Kooperberg, C., Stone, C. J., and Truong, Y. K., 2000. Functional ANOVA modeling for proportional hazards regression. *Ann. Statist.* **28**, 961-999.

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

[23] Huang, J. Z., and Yang, L., 2004. Identification of nonlinear additive autoregressive models. *Journal of the Royal Statistical Society Series B.* **66**, 463-477.

[24] Linton, O. B., and Härdle, W., 1996. Estimation of additive regression models with known links. *Biometrika.* **83**, 529-540.

[25] Linton, O. B., and Nielsen, J. P., 1995. A Kernel method of estimating structured nonparametric regression based on marginali integration. *Biometrika.* **82**, 93-100.

[26] Liptser, R. Sh., and Shirjaev, A. N., 1980. A functional central limit theorem for martingales. *Theory of Probability and Applications.* **25**, 667-688.

[27] Mammen, E., 1992. *When Does Bootstrap Work: Asymptotic Results and Simulations.* Lecture Notes in Statistics 77, Springer-Verlag, Berlin.

[28] Nadaraya, E. A., 1964. On estimating regression. *Theory Prob. Appl.*, **9**, 141-142.

[29] Ruppert, D., Wand, M. P., and Carroll, R. J., 2003. *Semiparametric regression.* Cambridge University Press, Cambridge.

[30] Seifert, B., and Gasser, T., 1996. Finite-sample variance of local polynomial: analysis and solutions. *J. Amer. Statist. Assoc.* **91**, 267-275.

[31] Sperlich, S., Tjøstheim, D., and Yang, L., 2002. Nonparametric estimation and testing of interaction in additive models. *Econom. Theory.* **18**, 197-251.

[32] Stone, C. J., 1985. Additive regression and other nonparametric models. *Ann. Statist.* **13**, 689 - 705.

[33] Tong, H., 1990. *Nonlinear Time Series: A Dynamical System Approach.* Oxford University Press, Oxford, U.K.

[34] Wahba, G., 1990. *S*pline models for observational data. SIAM, Philadelphia, PA.

[35] Wand, M. P., and Jones, M. C., 1995. *Kernel smoothing.* Chapman and Hall, Ltd., London.

[36] Watson, G. S., 1964. Smoothing regression analysis. *Sankhya Ser. A*, **26**, 359-372.

[37] Xia, Y. C., and Li, W. K., 1999. On single-index coefficient regression models. *J. Amer. Statist. Assoc.* **94**, 1275-1285.

[38] Yang, L. , Härdle, W., Park, B. U., and Xue, L., 2004. Estimation and testing of varying coefficients with marginal integration. *J. Amer. Statist. Assoc.* under revision.

[39] Yang, L., and Tschernig, R., 2002. Non- and semi-parametric identification of seasonal nonlinear autoregression models. *Econom. Theory*, **18**, 1408-1448.

[40] Yoshihara, k., 1976. Limiting behavior of U-statistics for stationary, absolutely regular processes. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, **35**, 237-252.