Estimation of semi-parametric additive coefficient model[*]

Lan Xue[†], Lijian Yang

September 29, 2004

Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824

Abstract

In the multivariate regression setting, we propose a flexible varying coefficient model in which the regression coefficients of some predictors are additive functions of other predictors. Marginal integration estimators of the coefficients are developed and their asymptotic properties investigated. Under $\beta-$mixing, 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. A data-driven bandwidth selection method is developed based on asymptotic considerations. Its effectiveness is confirmed in a Monte-Carlo study. The procedure is applied to the real German GNP and Wolf's Sunspot data, where the semi-parametric additive coefficient model demonstrates superior performance in terms of out-of-sample forecasts.

Keywords: German real GNP; Local polynomial; Marginal integration; Out-of-sample forecast; Rate of convergence; Varying coefficient model; Wolf's sunspot number

1 . Introduction

The nonparametric regression model has been widely used in various applications due to its ability to discover data structures that linear and parametric models fail to detect. A serious limitation of the general nonparametric model which gives statisticians reservation 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

[*] Xue's research has been supported in part by NSF grants BCS 0308420 and DMS 0405330. Yang's research has been supported in part by NSF grants SES 0127722 and DMS 0405330. The constructive comments by the coordinating editor and a referee are appreciated.

[†] Corresponding author.
E-mail address: xuelan@stt.msu.edu (L. Xue).

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

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

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

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

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

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

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 $\{\alpha_l\left(x_l\right)\}_{1\leq l\leq d}$, 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 $\{X_{il}\}_{1\leq l\leq d}$ 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 paper, we propose the following additive coefficient model which has a more flexible form, namely

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

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 April 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 January 1, 1990 to April 1, 1990). Yang & Tschernig (2002) deseasonalized this series by removing the four seasonal means from the series $\log\left(G_{t+4}/G_{t+3}\right), t = 1, ..., 120$. Denote the transformed time series as $\{Y_t\}_{t=1}^{120}$. As the nonparametric alternative to the best fitting linear autoregressive model (4.2) in subsection 4.2, we have fitted the following additive coefficient model (details in subsection 4.2)

$$Y_t = \{c_1 + \alpha_{11}\left(Y_{t-1}\right) + \alpha_{12}\left(Y_{t-8}\right)\} Y_{t-2} + \{c_2 + \alpha_{21}\left(Y_{t-1}\right) + \alpha_{22}\left(Y_{t-8}\right)\} Y_{t-4} + \sigma\varepsilon_t. \quad (1.5)$$

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 (4.2), and 0.000077 or 0.000085 for two fits of the additive coefficient model (1.5). Hence the reduction in ASPE is between 31% and 46%, see Table 1. Figure 1 clearly illustrates this improvement in prediction power. One can see that the additive coefficient model out-performs the linear autoregressive model in prediction for 8 of the 10 quarters.

(Insert Table 1 about here)

(Insert Figure 1 about here)

We organize the paper as follows. In section 2, we present the estimation procedure for the coefficient functions in model (1.4) and the asymptotic properties of the estimators. In section 3, we discuss computing issues relating to the implementation of the marginal estimation method as in section 2. In section 4, simulation results and applications to two empirical examples will be presented. Technical proofs are contained in the Appendix.

2 . The estimators

2.1 Model identification

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

$$m\left(\mathbf{X},\mathbf{T}\right)=\sum_{l=1}^{d_1}\alpha_l(\mathbf{X})T_l,\quad \alpha_l(\mathbf{X})=c_l+\sum_{s=1}^{d_2}\alpha_{ls}(X_s),\quad \forall 1\le l\le d_1 \tag{2.1}$$

with the identification conditions

$$E\left\{w(\mathbf{X})\alpha_{ls}(X_s)\right\}\equiv 0,\quad l=1,...,d_1,s=1,...,d_2, \tag{2.2}$$

for some nonnegative weight function w, with $E\left\{w(\mathbf{X})\right\}=1$. The weight function w is introduced so that estimation of the unknown functions $\left\{\alpha_l(\mathbf{X})\right\}_{1\le l\le d_1}$ will be carried out only on the support of w, $\text{supp}\,(w)$, which is compact according to assumption (A7). This is important as most of the asymptotic results for kernel type 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 (2.2) 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 (2.2), by writing

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

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

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

(A0) There exists a constant $C>0$ such that for any set of measurable functions $\left\{b_{ls}(X_s)\right\}_{1\le l\le d_1}^{1\le s\le d_2}$ that satisfy (2.2) and any set of constants $\left\{a_l\right\}_{1\le l\le d_1}$, the following holds

$$E\left[\sum_{l=1}^{d_1}\left\{a_l+\sum_{s=1}^{d_2}b_{ls}(X_s)\right\}T_l\right]^2\ge C\left[\sum_{l=1}^{d_1}a_l^2+\sum_{l=1}^{d_1}\sum_{s=1}^{d_2}E\left\{b_{ls}^2(X_s)\right\}\right]. \tag{2.3}$$

Lemma 1 *Under assumptions (A0) and (A5) in the Appendix, the representation in (2.1) subject to (2.2) is unique.*

Proof. Suppose that

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

with both the set $\{\alpha_{ls}(X_s)\}_{1\leq l\leq d_1}^{1\leq s\leq d_2}, \{c_l\}_{1\leq l\leq d_1}$ and the set $\{\widetilde{\alpha}_{ls}(X_s)\}_{1\leq l\leq d_1}^{1\leq s\leq d_2}, \{\widetilde{c}_l\}_{1\leq l\leq d_1}$ satisfying (2.2). Then upon defining for all s, l

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

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. ■

2.2 The model

Consider $\{(Y_i, \mathbf{X}_i, \mathbf{T}_i)\}_{i=1}^n$, a sample that follows (1.1) and (1.2) whose conditional mean function is described by (2.1) and the identifiability conditions (2.2), (2.3). The error terms $\{\varepsilon_i\}_{i=1}^n$ are 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 time series model, which is a useful extension of many existing nonlinear and nonparametric time series models such as exponential autoregressive model (EXPAR), threshold autoregressive model (TAR), and functional autoregressive model (FAR).

In (2.1), for every $l = 1, ..., d_2$, the coefficient of T_{il} consists of two parts, the unknown parameter c_l, and the unknown univariate functions $\{\alpha_{ls}\}_{1\leq s\leq d_2}$. The marginal integration method will be applied to estimate both. 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 (2.2), for every $l = 1, ..., d_1$ one has

$$c_l = E\left\{w(\mathbf{X})\alpha_l(\mathbf{X})\right\} = \int w(\mathbf{x})\alpha_l(\mathbf{x})\varphi(\mathbf{x})d\mathbf{x} \tag{2.4}$$

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

$$c_l + \alpha_{ls}(x_s) = E\left\{w_{-s}(\mathbf{X}_{-s})\alpha_l(x_s, \mathbf{X}_{-s})\right\} = \int w_{-s}(\mathbf{u}_{-s})\alpha_l(x_s, \mathbf{u}_{-s})\varphi_{-s}(\mathbf{u}_{-s})d\mathbf{u}_{-s} \tag{2.5}$$

where $\mathbf{u}_{-s} = (u_1, \ldots, u_{s-1}, u_{s+1}, \ldots, u_{d_2})^T, (x_s, \mathbf{u}_{-s}) = (u_1, \ldots, u_{s-1}, x_s, u_{s+1}, \ldots, u_{d_2})^T$, 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

$$c_l \approx \frac{\sum_{i=1}^n w(\mathbf{X}_i)\alpha_l(\mathbf{X}_i)}{\sum_{i=1}^n w(\mathbf{X}_i)}, \qquad \alpha_{ls}(x_s) \approx \frac{\sum_{i=1}^n w_s(\mathbf{X}_{i,-s})\alpha_l(x_s, \mathbf{X}_{i,-s})}{\sum_{i=1}^n w_s(\mathbf{X}_{i,-s})} - c_l \tag{2.6}$$

and the d_2-dimensional functions $\{\alpha_l(\mathbf{x})\}_{l=1}^{d_1}$ in the above equations (2.6) can be replaced by the usual local polynomial estimators. This is the essential idea behind the marginal integration method.

2.3 Estimators of constants $\{c_l\}$

According to the first approximation equation in (2.6), to estimate the constants $\{c_l\}_{l=1}^{d_1}$, 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 $\mathbf{x} \in \text{supp}\,(w)$, 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 $\alpha = (\alpha_1, \ldots, \alpha_{d_1})^T$

$$\sum_{i=1}^{n} \left\{ Y_i - \sum_{l=1}^{d_1} \alpha_l T_{il} \right\}^2 K_H(\mathbf{X}_i - \mathbf{x}), \tag{2.7}$$

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

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

Let $\hat{\alpha} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_{d_1})^T$ be the solution to the least squares problem in (2.7). Note that $\hat{\alpha}$ is dependent on $\mathbf{x}$, as is (2.7), 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 $\hat{\alpha} = \hat{\alpha}\,(\mathbf{x}) = (\hat{\alpha}_1\,(\mathbf{x})\,, \ldots, \hat{\alpha}_{d_1}\,(\mathbf{x}))^T$. More precisely, let

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

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

$$\hat{\alpha}_l(\mathbf{x}) = e_l^T \left\{ \mathbf{Z}^T \mathbf{W}(\mathbf{x}) \mathbf{Z} \right\}^{-1} \mathbf{Z}^T \mathbf{W}(\mathbf{x}) \mathbf{Y}. \tag{2.8}$$

By (2.6), the parameter c_l can be estimated as a weighted average of $\hat{\alpha}_l(\mathbf{X}_i)$'s, i.e.

$$\hat{c}_l = \frac{\sum_{i=1}^{n} w(\mathbf{X}_i) \hat{\alpha}_l(\mathbf{X}_i)}{\sum_{i=1}^{n} w(\mathbf{X}_i)}, \quad l = 1, \cdots, d_1. \tag{2.9}$$

Theorem 1 *Under assumptions (A1)-(A7) in the Appendix, for any $l = 1, ..., d_1$*

$$\sqrt{n}(\hat{c}_l - c_l) \xrightarrow{\mathcal{L}} N\left\{0, \sigma_l^2\right\}$$

where the asymptotic variance σ_l^2 is defined in (A.9) in the Appendix.

The rate of $1/\sqrt{n}$ at which $\hat{c}_l$ converges to c_l 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 A.5 about the term P_{1n}. 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 (2.9), see the derivation of the term P_{2n} in (A.7) and (A.8). This technique of simultaneously reducing the bias by the use of higher order kernel and "integrating out the variance" is the common feature of all marginal integration procedures.

2.4 Estimators of functions $\{\alpha_{ls}\}_{1 \leq l \leq d_1}^{1 \leq s \leq d_2}$

In the following, we illustrate the procedure for estimating the functions $\{\alpha_{ls}(x_s)\}_{1 \leq l \leq d_1}$, 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 (2.6), 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 $\alpha = (\alpha_1, \ldots, \alpha_{d_1})^T$, $\beta = (\beta_{11}, \ldots, \beta_{1p}, \ldots, \beta_{d_1 1}, \ldots, \beta_{d_1 p})^T$

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

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 Appendix, the bandwidth matrix $G_s = \text{diag}\,\{g_1, ..., g_{s-1}, g_{s+1}, ..., g_{d_2}\}$, and

$$k_{h_s}(u_s) = \frac{1}{h_s} k\left(\frac{u_s}{h_s}\right), \quad L_{G_s}(\mathbf{u}_{-s}) = \frac{1}{\prod\limits_{1 \leq s' \leq d_2, s' \neq s} g_{s'}} L\left(\frac{u_1}{g_1}, ..., \frac{u_{s-1}}{g_{s-1}}, \frac{u_{s+1}}{g_{s+1}}, ..., \frac{u_{d_2}}{g_{d_2}}\right)$$

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

$$\hat{\alpha}_l(\mathbf{x}) = e_l^T \left\{ \mathbf{Z}_s^T \mathbf{W}_s(\mathbf{x}) \mathbf{Z}_s \right\}^{-1} \mathbf{Z}_s^T \mathbf{W}_s(\mathbf{x}) \mathbf{Y}, \tag{2.10}$$

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

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

and

$$\begin{aligned} \mathbf{Z}_s &= \begin{bmatrix} \mathbf{T}_1^T, \{(X_{1s} - x_s)/h_s\} \mathbf{T}_1^T, ..., \{(X_{1s} - x_s)/h_s\}^p \mathbf{T}_1^T \\ \vdots \\ \mathbf{T}_n^T, \{(X_{ns} - x_s)/h_s\} \mathbf{T}_n^T, ..., \{(X_{ns} - x_s)/h_s\}^p \mathbf{T}_n^T \end{bmatrix} \\ &= \begin{bmatrix} [p\{(X_{1s} - x_s)/h_s\}]^T \otimes \mathbf{T}_1^T \\ \vdots \\ [p\{(X_{ns} - x_s)/h_s\}]^T \otimes \mathbf{T}_n^T \end{bmatrix} \end{aligned} \tag{2.11}$$

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

$$\hat{\alpha}_{ls}(x_s) = \frac{\sum_{i=1}^n w_{-s}(\mathbf{X}_{i,-s})\hat{\alpha}_l(x_s, \mathbf{X}_{i,-s})}{\sum_{i=1}^n w_{-s}(\mathbf{X}_{i,-s})} - \hat{c}_l, \tag{2.12}$$

where the term $\hat{c}_l$ is the $\sqrt{n}$-consistent estimator of c_l in Theorem 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 design matrix $\mathbf{Z}_s$ in (2.11). In particular, the local linear ($p = 1$) and the local cubic estimators ($p = 3$) are the most commonly used.

Theorem 2 *Under assumptions A1-A7 in the Appendix, for any* $\mathbf{x} = (x_1, ..., x_{d_2})^T \in \operatorname{supp}(w)$, *one has for* $l = 1, ..., d_1$, $s = 1, ..., d_2$

$$\sqrt{nh_s}\left\{\hat{\alpha}_{ls}(x_s) - \alpha_{ls}(x_s) - h_s^{p+1}\eta_{ls}(x_s)\right\} \xrightarrow{\mathcal{L}} N\left\{0, \sigma_{ls}^2(x_s)\right\}, \tag{2.13}$$

where $\eta_{ls}(x_s)$ *and* $\sigma_{ls}^2(x_s)$ *are defined in (A.17) and (A.19), respectively.*

Finally, based on (2.9) and (2.12), one can predict Y given any realization $(\mathbf{x}, \mathbf{t})$ of $(\mathbf{X}, \mathbf{T})$ by the predictor

$$\hat{m}(\mathbf{x}, \mathbf{t}) = \sum_{l=1}^{d_1}\left\{\hat{c}_l + \sum_{s=1}^{d_2}\hat{\alpha}_{ls}(x_s)\right\} t_l. \tag{2.14}$$

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 (2.10), which are similar to those discussed in subsection (2.3). 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 A.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 A.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 (2.13). 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 A.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 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.

3 . Implementation

Practical implementation of the estimators defined in (2.9) and (2.12) requires a rather intelligent choice of bandwidths $H = \operatorname{diag}\{h_{0,1}, ..., h_{0,d_2}\}, G_s = \operatorname{diag}\{g_1, ..., g_{s-1}, g_{s+1}, ..., g_{d_2}\}$ and $\{h_s\}_{1 \le s \le d_2}$. In the following, we discuss the choices of such bandwidths.

- Note from Theorem 1 that the asymptotic distributions of the estimators $\{\hat{c}_l\}_{l=1}^{d_1}$ 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 $\hat{h}_{01} = \cdots = \hat{h}_{0d_2} = \sqrt{\hat{\operatorname{var}}(\mathbf{X})} \log(n) n^{-1/(2q_1-1)}$, where q_1 is the order of the kernel K, required to be greater than $(d_2+1)/2$, and $\hat{\operatorname{var}}(\mathbf{X}) = \left\{\prod_{s=1}^{d_2} \hat{\operatorname{var}}(X_s)\right\}^{1/d_2}$, in which $\hat{\operatorname{var}}(X_s)$ 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 $\{\alpha_{ls}(x_s), l = 1, \ldots, d_1\}$ 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 (A.17) and (A.19). According to the definitions of $\eta_{ls}(x_s)$ and $\sigma_{ls}^2(x_s)$, $\int \eta_{ls}^2(x_s) dx_s$ and $\int \sigma_{ls}^2(x_s) dx_s$ can be approximated respectively by

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

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

where the functions K_{ls}^* are defined in (A.18).

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 $p+2$

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

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

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

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

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

3. According to the definition in (A.18), the dependence of the functions $K_{ls}^{*}(u,\mathbf{x},\mathbf{t})$ on u and $\mathbf{t}$ is explicitly known. The only unknown term $E\left(\mathbf{T}\mathbf{T}^T|\mathbf{X}=\mathbf{x}\right)$ contained in $S_\alpha^{-1}(\mathbf{x})$ is estimated by fitting matrix polynomial regression

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

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. More sophisticated bandwidth selection method requires further investigation.

- Since Theorem 2 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) $g_1=\ldots=g_{s-1}=g_{s+1}=\ldots=g_{d_2}=\hat{h}_{s,\mathrm{opt}}^{(p+1)/q_2}/\log(n)$, in which q_2, the order of the kernel function L, is required to be greater than $(d_2-1)/2$, and $\hat{h}_{s,\mathrm{opt}}$ 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 $k(u)=\frac{15}{16}\left(1-u^2\right)^2 1_{\{|u|\leq 1\}}$, 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 (2.7) is computed as $\mathbf{Z}^T\mathbf{W}(\mathbf{x})\mathbf{Z}+n^{-1}\mathbf{T}\mathbf{T}^T$, and the matrix $\mathbf{Z}_s^T\mathbf{W}_\mathbf{s}(\mathbf{x})\mathbf{Z}_s$ in (2.10) as $\mathbf{Z}_s^T\mathbf{W}_\mathbf{s}(\mathbf{x})\mathbf{Z}_s+\left(n\hat{h}_{s,\mathrm{opt}}\right)^{-1}\sqrt{\hat{\mathrm{var}}\left(\mathbf{X}\right)}\left\{\int k(u)p(u)p(u)^Tdu\right\}\bigotimes\mathbf{T}\mathbf{T}^T$, following the ridge regression idea of Seifert & Gasser (1996).

4 . Examples

4.1 A simulated example

The data are generated from the following model

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

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 $(\mathbf{X}, \mathbf{T})$. We use sample sizes $n = 100, 250$ and 500. The number of replications in the simulation is 100. First, to assess the performance of the data-driven bandwidth selector in section 3, we plot in Figure 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}. 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 estimate the functions α_{ls} on a grid of equally-spaced grid of points $x_m, m = 1, ..., n_{\text{grid}}$ with $x_1 = -0.975\pi, x_{n_{\text{grid}}} = 0.975\pi, n_{\text{grid}} = 62$. We summarize the fitting results in Table 2, which includes the means and standard errors (in the parentheses) of the $\hat{c}_l$, $l = 1, 2$ and the averaged integrated squared errors (AISE) of $\hat{\alpha}_{ls}$. By denoting the estimated function $\hat{\alpha}_{ls}$ of α_{ls} in the i-th replication by $\hat{\alpha}_{ls}^i$, we define

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

(Insert Figure 2 about here)

(Insert Table 2 about here)

Figure 3 gives the plots of $\hat{\alpha}_{11}$, $\hat{\alpha}_{12}$, $\hat{\alpha}_{21}$, $\hat{\alpha}_{22}$ obtained in the 100 replications for sample size $n = 100, 250, 500$ respectively. Also plotted are the typical estimators, whose ISE is the median of the ISEs in the 100 replications.

(Insert Figure 3 about here)

4.2 The West German real GNP

In this section, 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

$$\text{Linear AR (24):} \qquad Y_t = a_1 Y_{t-2} + a_2 Y_{t-4} + \sigma\varepsilon_t, \tag{4.2}$$

$$\text{Linear AR (248):} \qquad Y_t = b_1 Y_{t-2} + b_2 Y_{t-4} + b_3 Y_{t-8} + \sigma\varepsilon_t. \tag{4.3}$$

¿From Table 1, it is clear that besides being more parsimonious, the linear model (4.2) has smaller average squared prediction error (ASPE), compared with the model (4.3). Thus model (4.2) is the preferred linear autoregressive model. Moreover, Figures 4, 5 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}. So we have fitted the additive coefficient model (1.5) in the introduction.

(Insert Figure 4 about here)

(Insert Figure 5 about here)

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 (1.5), we use local cubic fitting (i.e. taking $p = 3$ in $\mathbf{Z}_s$ (2.11)). According to the bandwidth selection method in section 3, 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 6. 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 6, one may observe that the estimated functions have obviously nonconstant 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 (1.5) are significantly different from a constant. (Notice that by the restrictions proposed in (2.2), if a coefficient function is constant, it has to be zero.)

(Insert Figure 6 about here)

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 $\hat{b}_1 = -0.1191$, $\hat{b}_2 = 0.6458$, $\hat{b}_3 = 0.0704$. Lastly, to assess the sensitivity of marginal integration estimation method to the degree of the local polynomial, we have also fitted the model (1.5) using local linear estimation (i.e. taking $p = 1$ in $\mathbf{Z}_s$). Table 1 gives the ASEs (averaged squared estimation error) and ASPEs (averaged squared prediction error) from the above four estimations. One can see that overall the marginal integration estimation for model (1.5) is not sensitive to the order of local polynomial used. Local cubic fits outperform the local linear fits for prediction. Both local linear and local cubic fitting provide significant improvements over the linear autoregressive models.

4.3 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 (4.4), in which we use both lag 3 and lag 8 as the additive tuning variables: i.e.

$$\begin{aligned} Y_t &= \{c_1 + \alpha_{11}(Y_{t-3}) + \alpha_{12}(Y_{t-8})\} Y_{t-1} + \{c_2 + \alpha_{21}(Y_{t-3}) + \alpha_{22}(Y_{t-8})\} Y_{t-2} \\ & \quad \{c_3 + \alpha_{31}(Y_{t-3}) + \alpha_{32}(Y_{t-8})\} Y_{t-3} + \sigma\varepsilon_t. \end{aligned} \tag{4.4}$$

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. 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 are plotted in Figure 7, and the time plot of the fitted values in Figure 8. The averaged squared estimation error (ASE) is 4.18. 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

$$Y_t = \alpha_1(Y_{t-3}) Y_{t-1} + \alpha_2(Y_{t-3}) Y_{t-2} + \alpha_3(Y_{t-3}) Y_{t-3} + \alpha_6(Y_{t-3}) Y_{t-6} + \alpha_8(Y_{t-3}) Y_{t-8} + \sigma\varepsilon_t, \tag{4.5}$$

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

$$\begin{aligned} Y_t &= \phi_0\{g_4(\theta, Y_{t-3}, Y_{t-8})\} + \phi_1\{g_4(\theta, Y_{t-3}, Y_{t-8})\} Y_{t-1} + \phi_2\{g_4(\theta, Y_{t-3}, Y_{t-8})\} Y_{t-2} \\ & \quad + \phi_3\{g_4(\theta, Y_{t-3}, Y_{t-8})\} Y_{t-3} + \phi_4\{g_4(\theta, Y_{t-3}, Y_{t-8})\} Y_{t-8} + \sigma\varepsilon_t \end{aligned} \tag{4.6}$$

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

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 (4.5) has not been theoretically justified. Similarly, model (4.6) is also not theoretically justified, since according to Condition C5, p.1277 of Xia & Li (1999), the conditional density of $g_4(\theta, Y_{t-3}, Y_{t-8})$ 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 $g_4(\theta, Y_{t-3}, Y_{t-8})$ 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 (4.6) is unidentifiable. For any set of functions $\{\phi_0, \ldots, \phi_4\}$ that satisfy (4.6), one can always pick an arbitrary nonzero function $f(u)$ and define

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

It is straightforward to verify that the new set of functions $\left\{\widetilde{\phi}_0, \ldots, \widetilde{\phi}_4\right\}$ satisfy (4.6) as well. One possible fix of this problem is to drop either one of the terms $\phi_3 \left\{g_4\left(\theta, Y_{t-3}, Y_{t-8}\right)\right\} Y_{t-3}$ and $\phi_4 \left\{g_4\left(\theta, Y_{t-3}, Y_{t-8}\right)\right\} Y_{t-8}$ from (4.6), then the model is fully identifiable and satisfies Condition C5, p.1277 of Xia & Li (1999). Hence the current form of (4.6) may be considered an overfitting anomaly.

Despite the fact that models (4.5) and (4.6) 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 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 (4.6) 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.

(Insert Table 3 about here)

(Insert Figure 7 about here)

(Insert Figure 8 about here)

Appendix

A.1 Assumptions and an auxiliary lemma

We have listed below some assumptions necessary for proving Theorems 1 and 2. Throughout this appendix, 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 $l = 1, ..., d_1, s = 1, ..., d_2$.*

(A3) *The vector process $\{\zeta_i\}_{i=1}^{\infty} = \{(Y_i, \mathbf{X}_i, \mathbf{T}_i)\}_{i=1}^{\infty}$ is strictly stationary and β-mixing with the β-mixing coefficient $\beta(k) \leq c\rho^k$, for some constants $c > 0$, $0 < \rho < 1$. The β-mixing coefficient is defined as*

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

where $\mathcal{F}_{n+k}^{\infty}$ and $\mathcal{F}_0^n$ *denote the σ-algebras generated by $\{\zeta_i, i \geq n+k\}$ and $\{\zeta_0, \ldots, \zeta_n\}$* seperately.

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

$$\alpha(k) \leq c\rho^k/2. \tag{A.1}$$

(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 $\mathbf{x} \in \mathrm{supp}(w)$. 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 the bandwidth matrix $H = \mathrm{diag}\{h_{01}, \ldots, h_{0d_2}\}$ of Theorem 1, $\sqrt{n}h_{\max}^{q_1} \to 0$ and $nh_{\mathrm{prod}} \propto n^{\alpha}$ for some $\alpha > 0$, where $h_{\max} = \max\{h_{01}, \ldots, h_{0d_2}\}$, $h_{\mathrm{prod}} = \prod_{i=1}^{d_2} h_{0i}$, and $\propto$ means proportional to.

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

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

The proof of many results in this paper 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

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

$$\sup_{(i_1,....,i_m)\in S_c} \int |g_n(x_1, ..., x_m)|^{2+\delta} dF_{\xi_{i_1},...,\xi_{im}}(x_1, ..., x_m) \leq M_{n,c} < +\infty, c = 0, ..., m-1$$

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

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

allow decompositions as

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

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

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 $R^d, R_+^d = \{(y_1, ..., y_d) \in R^d | y_j \geq 0, j = 1, ..., d\}$.

Lemma A.1 *If $\beta(k) \leq C_1 k^{-(2+\delta')/\delta'}, \delta > \delta' > 0$, then*

$$EV_n^{(c)2} + EU_n^{(c)2} \leq C(m, \delta, r) n^{-c} \times \left\{ M_n^{2/(2+\delta)} \sum_{k=r+1}^{n} k\beta^{\delta/(2+\delta)}(k) + \sum_{c'=0}^{m-1} n^{-c'} M_{n,c'}'^{2/(2+\delta)} \sum_{k=1}^{r} k\beta^{\delta/(2+\delta)}(k) \right\} \quad \text{(A.2)}$$

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

$$EV_n^{(c)2} + EU_n^{(c)2} \leq C(m, \delta, r) C_2 C(\rho) n^{-c} \left\{ M_n^{2/(2+\delta)} + \sum_{c'=0}^{m-1} n^{-c'} M_{n,c'}'^{2/(2+\delta)} \right\}. \quad \text{(A.3)}$$

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 (A.2), provides an obvious venue of extension to the more general setup. Elementary arguments then establish (A.3) under geometric mixing conditions. ∎

A.2 Proofs of Theorem 1 and Theorem 2

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

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

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

$$
S_\alpha(\mathbf{x}) = \left\{\int k(u)p(u)p(u)^T du\right\} \bigotimes S(\mathbf{x}) \tag{A.4}
$$

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 A.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

$$\begin{aligned}
& b_2^{-1} \sup_{\mathbf{x}\in B} \left|\mathbf{Z}_s^T \mathbf{W}_s(\mathbf{x})\mathbf{Z}_s - E\left\{\mathbf{Z}_s^T \mathbf{W}_s(\mathbf{x})\mathbf{Z}_s\right\}\right| \\
\leq\ & b_2^{-1} \sup_j \sup_{\mathbf{x}\in B_{jn}} \left|\mathbf{Z}_s^T \mathbf{W}_s(\mathbf{x})\mathbf{Z}_s - \mathbf{Z}_s^T \mathbf{W}_s(\mathbf{x}_j)\mathbf{Z}_s\right| \\
& + b_2^{-1} \sup_j \sup_{\mathbf{x}\in B_{jn}} \left|E\mathbf{Z}_s^T \mathbf{W}_s(\mathbf{x})\mathbf{Z}_s - E\left\{\mathbf{Z}_s^T \mathbf{W}_s(\mathbf{x}_j)\mathbf{Z}_s\right\}\right| \\
& + b_2^{-1} \sup_j \left|\mathbf{Z}_s^T \mathbf{W}_s(\mathbf{x}_j)\mathbf{Z}_s - E\left\{\mathbf{Z}_s^T \mathbf{W}_s(\mathbf{x}_j)\mathbf{Z}_s\right\}\right|. \qquad \text{(A.5)}
\end{aligned}$$

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

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

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

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

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

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

almost surely, as a result of assumption (A5) (c) entails that $E\left(\mathbf{T}\mathbf{T}^T\right) < \infty$. Choosing $v_n = \left[\left(h_s g_{\text{prod}}\right)^{-3}\right]$ (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 (A.5), 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

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

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

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

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

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

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

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

where

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

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

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

and the second term

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

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

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

Then for any $\varepsilon > 0$

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

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

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

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. ■

A.2.1 Proof of Theorem 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

$$
\frac{1}{n}\sum_{i=1}^n w(\mathbf{X}_i)\{\hat{c}_l - c_l\} = I + II + III \tag{A.6}
$$

in which

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

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

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

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 A.2, one has

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

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

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

To prove Theorem 1, we need the following lemmas.

Lemma A.3 *Define*

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

Then as $n \to +\infty$

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

Lemma A.4 *For fixed* $\nu = 1, ..., v_1,$ *define*

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

Then as $n \to +\infty$

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

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

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

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

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

Then P_1 can be written as the von Mises' differential statistic $P_1 = \frac{1}{2n^2}\sum_{i,j=1}^{n} g_n\left(\xi_i, \xi_j\right)$, which can be decomposed as

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

in which

$$\theta_n\left(F\right) = \int g_n\left(u, v\right)dF_{\xi_i}\left(u\right)dF_{\xi_j}\left(v\right) = 0.$$

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 ξ_i and ξ_j. One has

$$\mathbf{V}_n^{(1)} = E_i E_{n,j} g_n\left(\xi_i, \xi_j\right) = \frac{1}{n}\sum_{j=1}^{n} g_{n,1}\left(\xi_j\right)$$

in which

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

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

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

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

$$\begin{aligned} E_1 E_2 \left| g_n \left(\xi_1, \xi_2\right)\right|^{2+\delta} &\leq M_n < +\infty \\ \sup_{i \neq j} E_{i,j} \left| g_n \left(\xi_i, \xi_j\right)\right|^{2+\delta} &\leq M_{n,0} < +\infty \\ E_i \left| g_n \left(\xi_i, \xi_i\right)\right|^{2+\delta} &\leq M_{n,1} < +\infty \end{aligned}$$

And observe that

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

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 $EP_2^2 \leq c n^{-1} b_1^2$. So we have $F_{11} = o_p\left(b_1/\sqrt{n}\right)$. ■

Lemma A.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 $K_l^*\left(\mathbf{X}, \mathbf{T}\right) = e_l^T \mathbf{S}^{-1}\left(\mathbf{X}\right) \mathbf{T}$, then

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

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

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

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

$$\begin{aligned} & \int w(\mathbf{z}) K_l^*(\mathbf{z}, \mathbf{t}) K(\mathbf{u}) \left[\sum_{l'=1}^{d_1} \sum_{s=1}^{d_2} \left\{ \alpha_{l's}(z_s + h_{0,s} u_s) - \alpha_{l's}(z_s) \right\} t_{l'} \right] \psi(\mathbf{z} + H\mathbf{u}, \mathbf{t})\, d\mathbf{u} d\mathbf{z} d\mathbf{t} \\ = \;& O\left(h_{\max}^{q_1}\right) \end{aligned}$$

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 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

$$P_{2n} = \frac{1}{n^2} \sum_{i,j=1}^{n} \frac{w(\mathbf{X}_i)}{\varphi(\mathbf{X}_i)} K_l^*(\mathbf{X}_i, \mathbf{T}_j) K_H(\mathbf{X}_j - \mathbf{X}_i) \sigma(\mathbf{X}_j, \mathbf{T}_j) \varepsilon_j \tag{A.7}$$

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

$$\begin{aligned} & \frac{1}{n} \sum_{j=1}^{n} \int w(\mathbf{X}_j - H\mathbf{u}) K_l^*(\mathbf{X}_j - H\mathbf{u}, \mathbf{T}_j) K(\mathbf{u}) \sigma(\mathbf{X}_j, \mathbf{T}_j) \varepsilon_j d\mathbf{u} + o_p\left(\frac{\log n}{n^2 h_{\text{prod}}}\right) \\ = \;& \frac{1}{n} \sum_{j=1}^{n} w(\mathbf{X}_j) K_l^*(\mathbf{X}_j, \mathbf{T}_j) \sigma(\mathbf{X}_j, \mathbf{T}_j) \varepsilon_j + o_p\left(\frac{\log n}{n^2 h_{\text{prod}}}\right) + o_p\left(h_{\max}^{q_1}\right) \\ = \;& \frac{1}{n} \sum_{j=1}^{n} w(\mathbf{X}_j) K_l^*(\mathbf{X}_j, \mathbf{T}_j) \sigma(\mathbf{X}_j, \mathbf{T}_j) \varepsilon_j + o_p\left(n^{-1/2}\right). \qquad \text{(A.8)} \end{aligned}$$

Now come back to the decomposition of $\frac{1}{n}\sum_{i=1}^{n} w(\mathbf{X}_i)(\hat{c}_l - c_l)$ as in (A.6), and by Lemmas A.2, A.3, A.4, A.5, one has

$$\frac{1}{n} \sum_{i=1}^{n} w(\mathbf{X}_i)(\hat{c}_l - c_l) = \frac{1}{n} \sum_{i=1}^{n} w(\mathbf{X}_j) \left\{ K_l^*(\mathbf{X}_j, \mathbf{T}_j) \sigma(\mathbf{X}_j, \mathbf{T}_j) \varepsilon_j + \sum_{s=1}^{d_2} \alpha_{ls}(X_{js}) \right\} + o_p\left(n^{-1/2}\right).$$

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 $\{(\mathbf{X}_i, \mathbf{T}_i)\}_{i\leq j}$, we have

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

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

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

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

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

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

$$\begin{aligned}
\sigma_l^2 &= \sum_{j=-\infty}^{+\infty} \operatorname{cov}\left(\tau_0, \tau_j\right) = 2\sum_{j=1}^{+\infty} \operatorname{cov}\left(\tau_0, \tau_j\right) + \operatorname{var}\left(\tau_0\right) \\
&= 2\sum_{j=1}^{+\infty} \operatorname{cov}\left(\tau_0, \tau_{j2}\right) + \operatorname{var}\left(\tau_0\right) \qquad \text{(A.9)}
\end{aligned}$$

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

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

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

A.2.2 Proof of Theorem 2

Following similarly as in the proof of Theorem 1, let v_2 be an integer which satisfies $b_2^{v_2} = o\left(h_s^{p+2}\right)$. Then by Lemma A.2, one has

$$\left\{\mathbf{Z}_s^T \mathbf{W}_s\left(\mathbf{x}\right) \mathbf{Z}_s\right\}^{-1} - \frac{S_\alpha^{-1}\left(\mathbf{x}\right)}{\varphi\left(\mathbf{x}\right)} = \frac{S_\alpha^{-1}\left(\mathbf{x}\right)}{\varphi\left(\mathbf{x}\right)} \sum_{v=1}^{v_2} A\left(\mathbf{x}\right)^v + Q_s\left(\mathbf{x}\right) \qquad \text{(A.10)}$$

where

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

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 1, by the equation that

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

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

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} w_{-s}(\mathbf{X}_{i,-s})\left\{\hat{\alpha}_{ls}(x_s) - \alpha_{ls}(x_s)\right\}\\
=\ &\frac{1}{n}\sum_{i=1}^{n} w_{-s}(\mathbf{X}_{i,-s})\left[e_l^T\left\{\mathbf{Z}_s^T\mathbf{W}_s(\mathbf{X}_{i,-s})\,\mathbf{Z}_s\right\}^{-1}\mathbf{Z}_s^T\mathbf{W}_s(\mathbf{X}_{i,-s})\,\mathbf{Y} - \alpha_{ls}(x_s) - \hat{c}_l\right]\\
=\ &\frac{1}{n}\sum_{i=1}^{n} w_{-s}(\mathbf{X}_{i,-s})\left[e_l^T\left\{\mathbf{Z}_s^T\mathbf{W}_s(\mathbf{X}_{i,-s})\,\mathbf{Z}_s\right\}^{-1}\mathbf{Z}_s^T\mathbf{W}_s(\mathbf{X}_{i,-s})\left\{\mathbf{Y} - \mathbf{M}\right.\right.\\
&\left.\left.+\,\mathbf{M} - \sum_{l'=1}^{d_1}\sum_{v=0}^{p}\frac{\alpha_{l's}^{(v)}(x_s)\,h_s^v}{v!}\mathbf{Z}_s e_{(d_1v+l')} - \sum_{l'=1}^{d_1}\left\{c_{l'} + \sum_{s'\neq s}^{d_2}\alpha_{l's'}(X_{is'})\right\}\mathbf{Z}_s e_{l'}\right\}\right]\\
&+\frac{1}{n}\sum_{i=1}^{n} w_{-s}(\mathbf{X}_{i,-s})\sum_{s'\neq s}\alpha_{ls'}(X_{is'}) + \frac{1}{n}\sum_{i=1}^{n} w_{-s}(\mathbf{X}_{i,-s})\,(\hat{c}_l - c_l) \qquad \text{(A.11)}
\end{aligned}$$

where $\mathbf{M}$ is the mean vector, as defined in Theorem 1. Next define

$$\begin{aligned}
\mathbf{R}_1 = \mathbf{R}_1(x_s) &= \left[\sum_{l'=1}^{d_1}\left\{\alpha_{l's}(X_{js}) - \sum_{v=0}^{p}\frac{\alpha_{l's}^{(v)}(x_s)}{v!}(X_{js} - x_s)^v\right\}T_{jl'}\right]_{j=1,\dots,n},\\
\mathbf{R}_2(\mathbf{X}_{i,-s}) &= \left[\sum_{l'=1}^{d_1}\sum_{s'\neq s}^{d_2}\left\{\alpha_{l's'}\left(\mathbf{X}_{js'}\right) - \alpha_{l's'}(\mathbf{X}_{is'})\right\}T_{jl'}\right]_{j=1,\dots,n},\\
R_3 &= \frac{1}{n}\sum_{i=1}^{n} w_{-s}(\mathbf{X}_{i,-s})\left\{\sum_{s'\neq s}\alpha_{ls'}(X_{is'})\right\},\\
R_4 &= \frac{1}{n}\left\{\sum_{i=1}^{n} w_{-s}(\mathbf{X}_{i,-s})\right\}(\hat{c}_l - c_l), \qquad \text{(A.12)}
\end{aligned}$$

$$\begin{aligned}
D_{s1}(x_s) &= \frac{1}{n}\sum_{i=1}^{n} w_{-s}(\mathbf{X}_{i,-s})\left\{e_l^T Q_s(x_s,\mathbf{X}_{i,-s})\,\mathbf{Z}_s^T\mathbf{W}_s(x_s,\mathbf{X}_{i,-s})\,\mathbf{E}\right\},\\
D_{s2}(x_s) &= \frac{1}{n}\sum_{i=1}^{n} w_{-s}(\mathbf{X}_{i,-s})\left\{e_l^T Q_s(x_s,\mathbf{X}_{i,-s})\,\mathbf{Z}_s^T\mathbf{W}_s(x_s,\mathbf{X}_{i,-s})\,\mathbf{R}_1\right\},\\
D_{s3}(x_s) &= \frac{1}{n}\sum_{i=1}^{n} w_{-s}(\mathbf{X}_{i,-s})\left\{e_l^T Q_s(x_s,\mathbf{X}_{i,-s})\,\mathbf{Z}_s^T\mathbf{W}_s(x_s,\mathbf{X}_{i,-s})\,\mathbf{R}_2(\mathbf{X}_{i,-s})\right\}, \text{(A.13)}
\end{aligned}$$

$$
\begin{aligned}
R_{r1}(x_s) &= \frac{1}{n}\sum_{i=1}^{n} w_{-s}\left(\mathbf{X}_{i,-s}\right)\left[e_l^T\left\{A\left(x_s,\mathbf{X}_{i,-s}\right)\right\}^r\mathbf{Z}_s^T\mathbf{W}_s\left(x_s,\mathbf{X}_{i,-s}\right)\mathbf{E}\right], \\
R_{r2}(x_s) &= \frac{1}{n}\sum_{i=1}^{n} w_{-s}\left(\mathbf{X}_{i,-s}\right)\left[e_l^T\left\{A\left(x_s,\mathbf{X}_{i,-s}\right)\right\}^r\mathbf{Z}_s^T\mathbf{W}_s\left(x_s,\mathbf{X}_{i,-s}\right)\mathbf{R}_1\right], \\
R_{r3}(x_s) &= \frac{1}{n}\sum_{i=1}^{n} w_{-s}\left(\mathbf{X}_{i,-s}\right)\left[e_l^T\left\{A\left(x_s,\mathbf{X}_{i,-s}\right)\right\}^r\mathbf{Z}_s^T\mathbf{W}_s\left(x_s,\mathbf{X}_{i,-s}\right)\mathbf{R}_2\left(\mathbf{X}_{i,-s}\right)\right], \quad (A.14)
\end{aligned}
$$

$$
\begin{aligned}
P_{s1}(x_s) &= \frac{1}{n}\sum_{i=1}^{n}\frac{w_{-s}\left(\mathbf{X}_{i,-s}\right)}{\varphi\left(x_s,\mathbf{X}_{i,-s}\right)}\left\{e_l^T S_\alpha^{-1}\mathbf{Z}_s^T\mathbf{W}_s\left(x_s,\mathbf{X}_{i,-s}\right)\mathbf{E}\right\}, \\
P_{s2}(x_s) &= \frac{1}{n}\sum_{i=1}^{n}\frac{w_{-s}\left(\mathbf{X}_{i,-s}\right)}{\varphi\left(x_s,\mathbf{X}_{i,-s}\right)}\left\{e_l^T S_\alpha^{-1}\mathbf{Z}_s^T\mathbf{W}_s\left(x_s,\mathbf{X}_{i,-s}\right)\mathbf{R}_1\right\}, \\
P_{s3}(x_s) &= \frac{1}{n}\sum_{i=1}^{n}\frac{w_{-s}\left(\mathbf{X}_{i,-s}\right)}{\varphi\left(x_s,\mathbf{X}_{i,-s}\right)}\left\{e_l^T S_\alpha^{-1}\mathbf{Z}_s^T\mathbf{W}_s\left(x_s,\mathbf{X}_{i,-s}\right)\mathbf{R}_2\left(\mathbf{X}_{i,-s}\right)\right\}. \quad (A.15)
\end{aligned}
$$

One can then write (A.11) as

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} w_{-s}(\mathbf{X}_{i,-s})\left\{\hat{\alpha_{sl}}(x_s)-\alpha_{sl}(x_s)\right\} \\
&= \sum_{i=1}^{3}P_{si}(x_s)+\sum_{i=1}^{3}D_{si}(x_s)+\sum_{r=1}^{v_2}\sum_{i=1}^{3}R_{ri}(x_s)+R_3+R_4. \quad (A.16)
\end{aligned}
$$

The proof of Theorem 2 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 (A.16). These asymptotic results are presented in the following lemmas:

Lemma A.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 A.7 *As $n \to +\infty$*

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

Lemma A.8 *For any fixed $r = 1, ..., v_s$, as $n \to +\infty$*

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

Lemma A.9 *As $n \to +\infty$*

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

$$P_{s2}\left(x_s\right)=h_s^{p+1}\eta_{ls}\left(x_s\right)+o_p\left(h_s^{p+1}\right),$$
$$P_{s3}\left(x_s\right)=O_p\left(g_{\max}^{q_2}\right)=o_p\left\{\left(nh_s\log n\right)^{-1/2}\right\},$$

in which

$$\eta_{ls}\left(x_s\right)=\frac{1}{(p+1)!}\sum_{l'=1}^{d_1}\alpha_{l's}^{(p+1)}\left(x_s\right)\int u^{p+1}E\left\{w_{-s}\left(\mathbf{X}_{-s}\right)T_{l'}K_{ls}^{*}\left(u,x_s,\mathbf{X}_{-s},\mathbf{T}\right)\right\}du \qquad (A.17)$$

with

$$K_{ls}^{*}\left(u,\mathbf{x},\mathbf{T}\right)=e_l^TS_\alpha^{-1}\left(\mathbf{x}\right)q^{*}\left(u,\mathbf{T}\right)k(u),\quad q^{*}(u,\mathbf{T})=\left(\mathbf{T},u\mathbf{T},...,u^p\mathbf{T}\right)^T. \qquad (A.18)$$

Furthermore

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

in which

$$\sigma_{ls}^2\left(x_s\right)=\int\frac{w_{-s}^2\left(\mathbf{z}_{-s}\right)}{\varphi^2\left(x_s,\mathbf{z}_{-s}\right)}\times$$
$$K_{ls}^{*2}\left(u,x_s,\mathbf{z}_{-s},\mathbf{t}\right)\varphi_{-s}^2\left(\mathbf{z}_{-s}\right)\sigma^2\left(x_s,\mathbf{z}_{-s},\mathbf{t}\right)\psi\left(x_s,\mathbf{z}_{-s},\mathbf{t}\right)dud\mathbf{z}_{-s}d\mathbf{t}. \qquad (A.19)$$

Proof of Lemma A.6. According to Theorem 1

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

Meanwhile, according to the identify condition (2.2) and the central limit theorem for strongly mixing process (Theorem 1.7 of Bosq 1998), we have

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

These two equations have completed the proof of lemma. ■

Proof of Lemmas A.7 and A.8. We have left these out as they are similar to Lemmas A.3, A.4. ■

Proof of Lemma A.9. From the definition in (A.15) and using the von Mises' statistic argument

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

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 $\{\xi_j, j \le i\}$, 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

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

in which the leading term $\sigma_{ls}^2\left(x_s\right)$ is as defined in (A.19). Hence we have shown that $\sqrt{nh_s}P_{s1} \xrightarrow{\mathcal{L}} N\left\{0, \sigma_{ls}^2\left(x_s\right)\right\}$.

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

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

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

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

with $\eta_{ls}\left(x_s\right)$ as defined in (A.17). Lastly, the term P_{s3} is

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

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

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

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. ■

References

Bosq, D., 1998. Nonparametric Statistics for Stochastic Processes. Springer-Verlag, New York.

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

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

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

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

Härdle, W., Liang, H., Gao, J. T., 2000. Partially Linear Models. Springer-Verlag, Heidelberg.

Hastie, T. J., Tibshirani, R. J., 1990. Generalized Additive Models. Chapman and Hall, London.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

	ASE	ASPE
Additive coefficient model (1.5), local linear fit	0.000201	0.000085
Additive coefficient model (1.5), local cubic fit	0.000205	0.000077
Linear AR (24) model (4.2)	0.000253	0.000112
Linear AR (248) model (4.3)	0.000258	0.000116

Table 1: German real GNP: the ASE's and ASPE's of four fits.

	$c_1 = 2$	$c_2 = 1$	α_{11}	α_{12}	α_{21}	α_{22}
n=100	1.9737(0.3574)	1.0406(0.2503)	0.1609	0.2541	0.1205	0.2761
n=250	2.0299(0.2410)	1.0056(0.1490)	0.0568	0.0963	0.0338	0.0649
n=500	1.9786(0.1680)	1.0026(0.1111)	0.0295	0.0483	0.0191	0.0310

Table 2: Results of the 100 simulations: the means and standard errors (in parentheses) of the estimators $\hat{c}_1$ and $\hat{c}_2$ of c_1 and c_2, and AISEs of the estimators $\hat{\alpha}_{11}$, $\hat{\alpha}_{12}$, $\hat{\alpha}_{21}$, $\hat{\alpha}_{22}$ of the coefficient functions α_{11}, α_{12}, α_{21}, α_{22}.

Year	X_t	TAR	FAR1	FAR2	SIND	Additive coefficient
1980	154.7	5.5	13.8	1.4	2.1	14.9
1981	140.5	1.3	0.0	11.4	1.7	2.4
1982	115.9	19.5	10.0	15.7	2.6	17.5
1983	66.6	4.8	3.3	10.3	2.4	1.37
1984	45.9	14.8	3.8	1.0	2.3	5.92
1985	17.9	0.2	4.6	2.6	7.6	1.96
1986	13.4	5.5	1.3	3.1	4.2	0.57
1987	29.2	0.7	21.7	12.3	13.2	0.7
AAPE		6.6	7.3	7.2	4.5	5.7
MSE		85.6	101.1	81.6	34.3	71.9

Table 3: Out-of-sample absolute prediction errors for sunspot data, under different models. Results on models TAR, FAR1 and FAR2 are from Cai, Fan & Yao (2000), Table 4, p.951, while results on model SIND are from Xia & Li (1999), Table 2, p.1280.

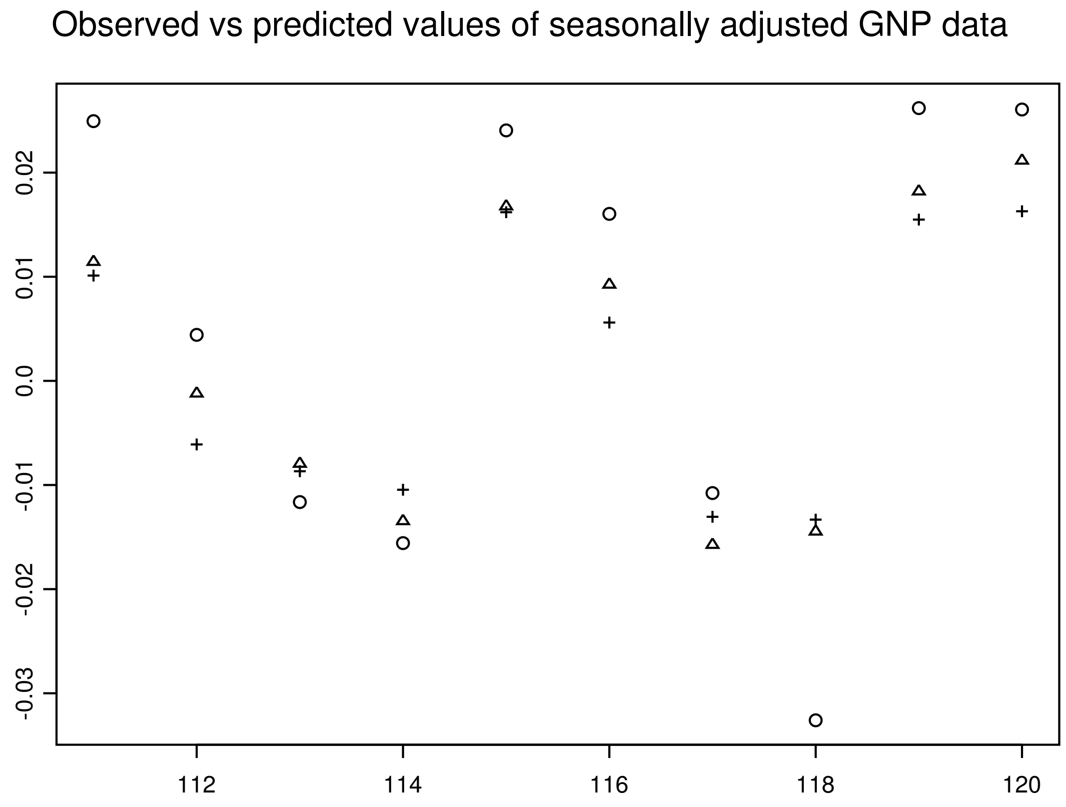


Figure 1: German real GNP: one-step prediction performance for the last ten quarters. Circle denotes the observed value, triangle denotes the prediction by additive coefficient model (1.5), and cross denotes the prediction by linear autoregressive model (4.2).

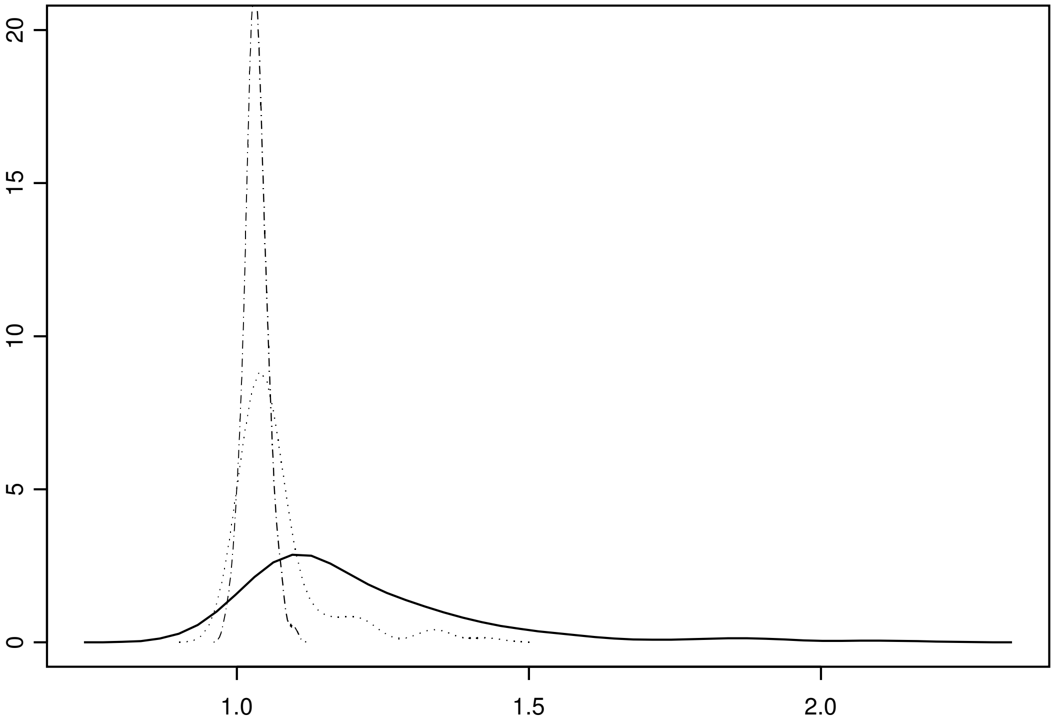


Figure 2: Bandwidth selection results of the simulated example: kernel density estimates of $\hat{h}_{1,\text{opt}}/h_{1,\text{opt}}$ ($h_{1,\text{opt}}$ is the theoretical optimal bandwidth for estimating the functions of x_1 in (4.1)). Solid curve is for $n = 100$, dotted curve is for $n = 250$, and dot-dashed curve is for $n = 500$.

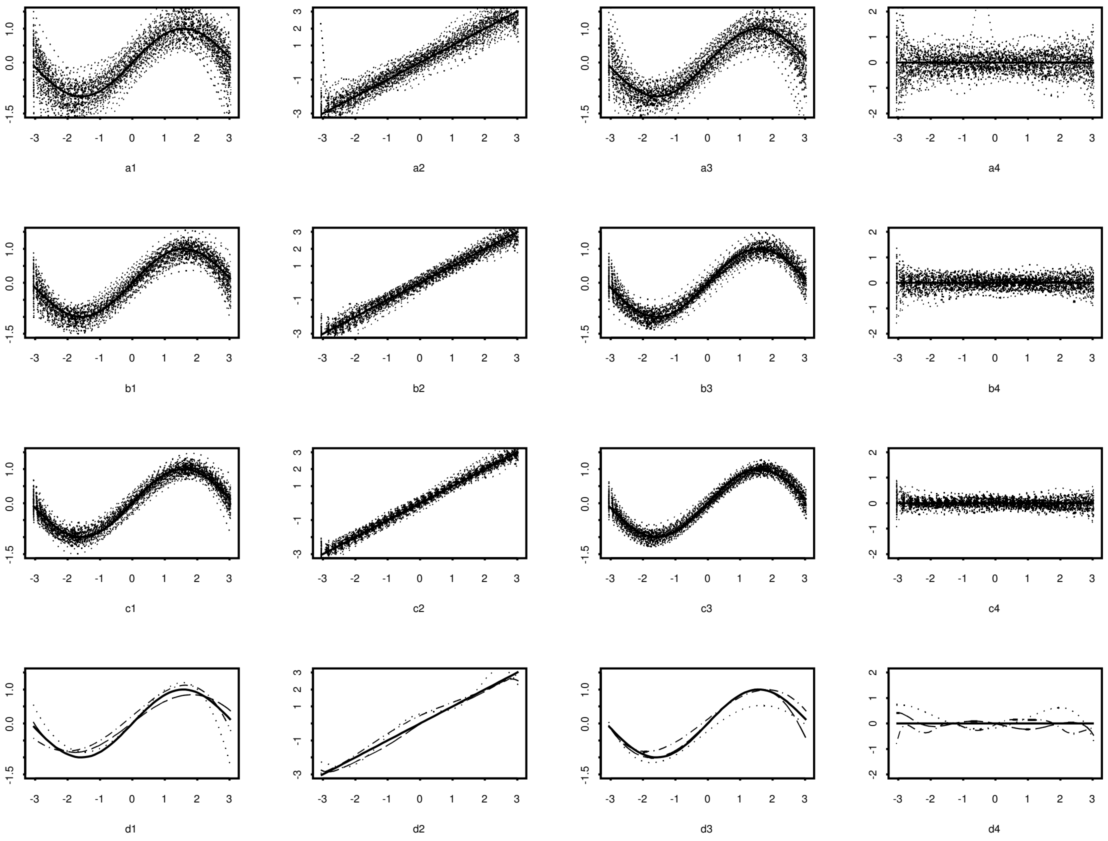
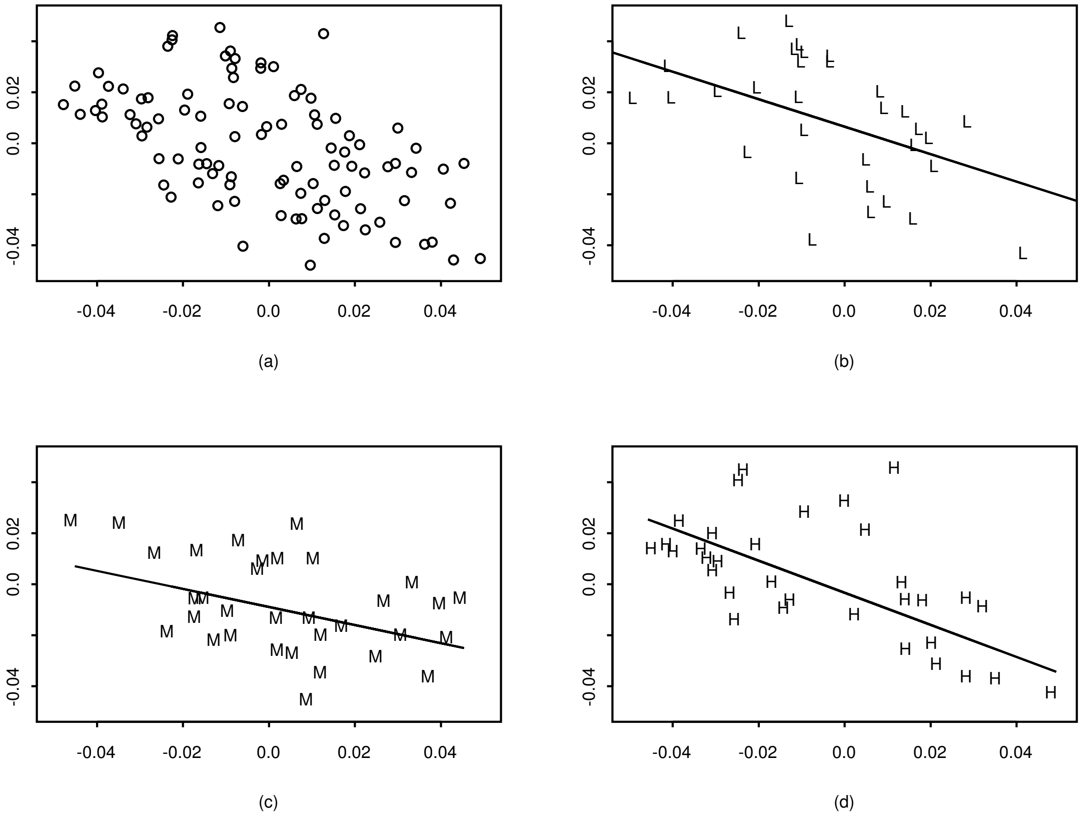


Figure 3: Plots of the estimated coefficient functions in the simulated example. (a1-a4) are plots of the 100 estimated curves for $\alpha_{11}(x_1) = \sin(x_1)$, $\alpha_{12}(x_2) = x_2$, $\alpha_{21}(x_1) = \sin(x_1)$, $\alpha_{22}(x_2) = 0$ with $n = 100$. (b1-b4) and (c1-c4) are the same as (a1-a4), but with sample size $n = 250$ and $n = 500$ respectively. (d1-d4) are plots of the typical estimators, the solid curve represents the true curve, the dotted curve is the typical estimated curve with $n = 100$, the dot-dashed curve is with $n = 250$ and the dashed curve is with $n = 500$.



(a)

(b)

(c)

(d)

Figure 4: Scatter plot of Y_t, Y_{t-2} at three levels of Y_{t-1}: H, M, L. Here the 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. (a) scatter plot of Y_t, Y_{t-2}. (b) scatter plot of Y_t, Y_{t-2} at high level of Y_{t-1}. (c) scatter plot at middle level of Y_{t-1}. (d) scatter plot at lower level of Y_{t-1}.

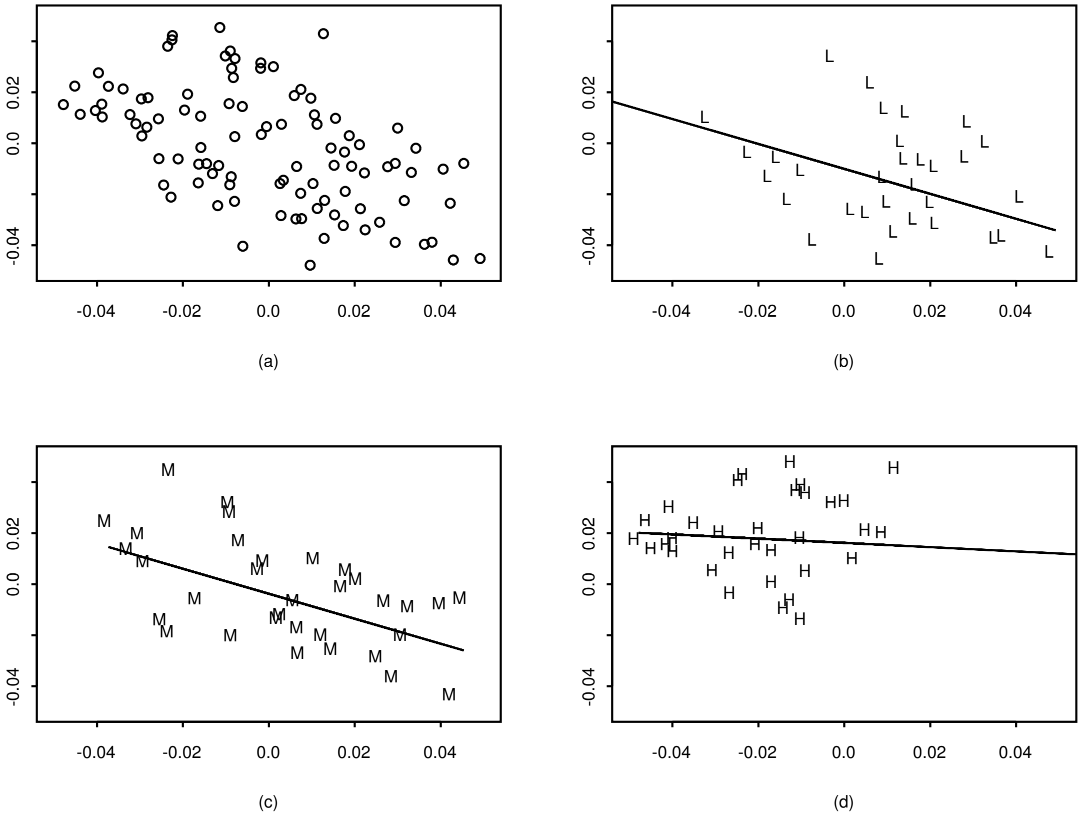


Figure 5: Scatter plot of Y_t, Y_{t-2} at three levels of Y_{t-8}: H, M, L. Here the levels are defined as: H, the high level, is the top 33% percent of the data, L, the lower level, is the lower 33% percent of the data, and M, the middle level, is the rest of the data. (a) scatter plot of Y_t, Y_{t-2}, (b) scatter plot of Y_t, Y_{t-2} at high level of Y_{t-8}, (c) scatter plot at middle level of Y_{t-8}, (d) scatter plot at lower level of Y_{t-8}.

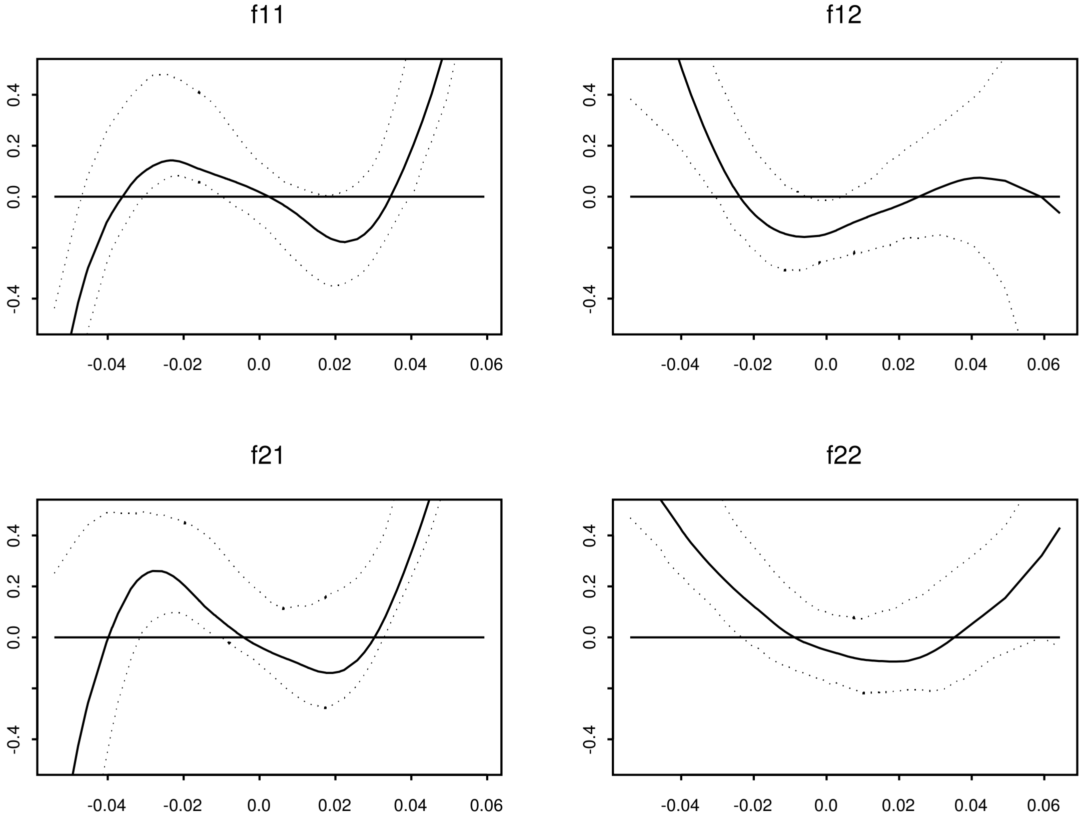


Figure 6: German real GNP: estimated functions and their point-wise wild bootstrap 95% confidence intervals, based on model (1.5): (a)$\hat{\alpha}_{11}$, (b)$\hat{\alpha}_{12}$, (c)$\hat{\alpha}_{21}$, (d)$\hat{\alpha}_{22}$.

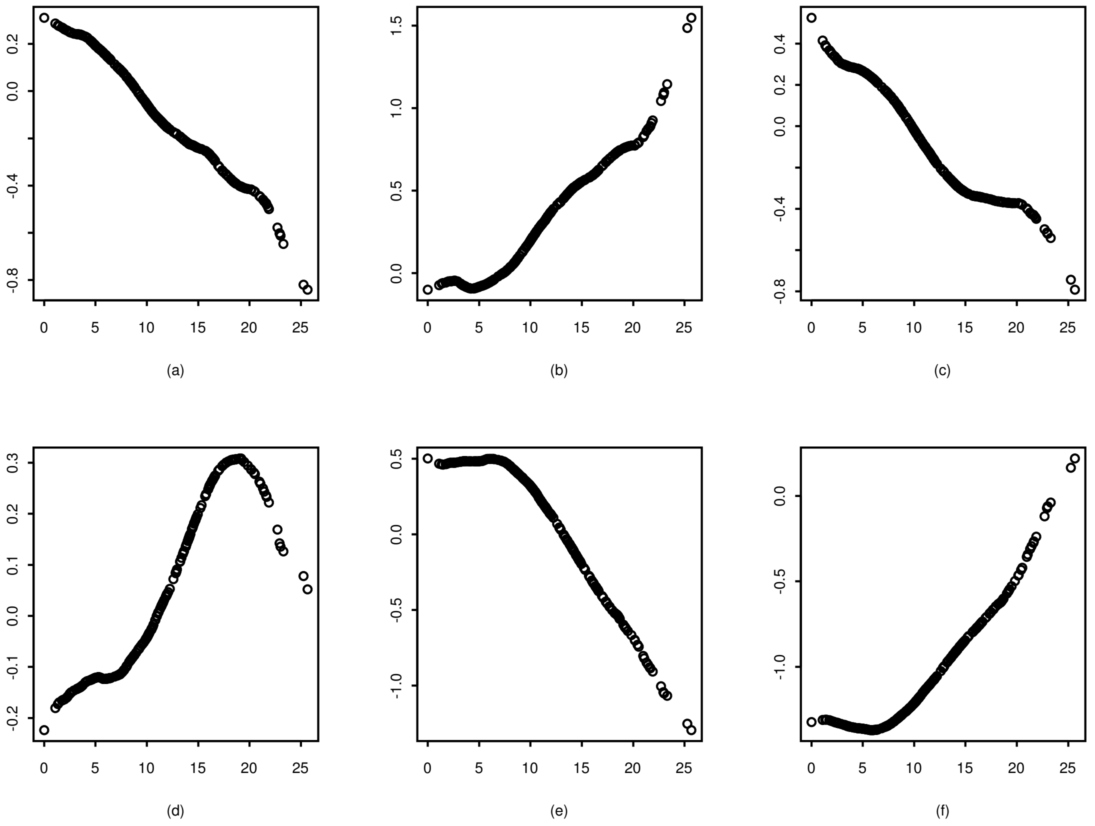


Figure 7: Wolf's Sunspot Number: estimated functions based on model (4.4): (a)$\hat{\alpha}_{11}$, (b)$\hat{\alpha}_{21}$, (c)$\hat{\alpha}_{31}$, (d)$\hat{\alpha}_{12}$, (e)$\hat{\alpha}_{22}$, (f)$\hat{\alpha}_{32}$.

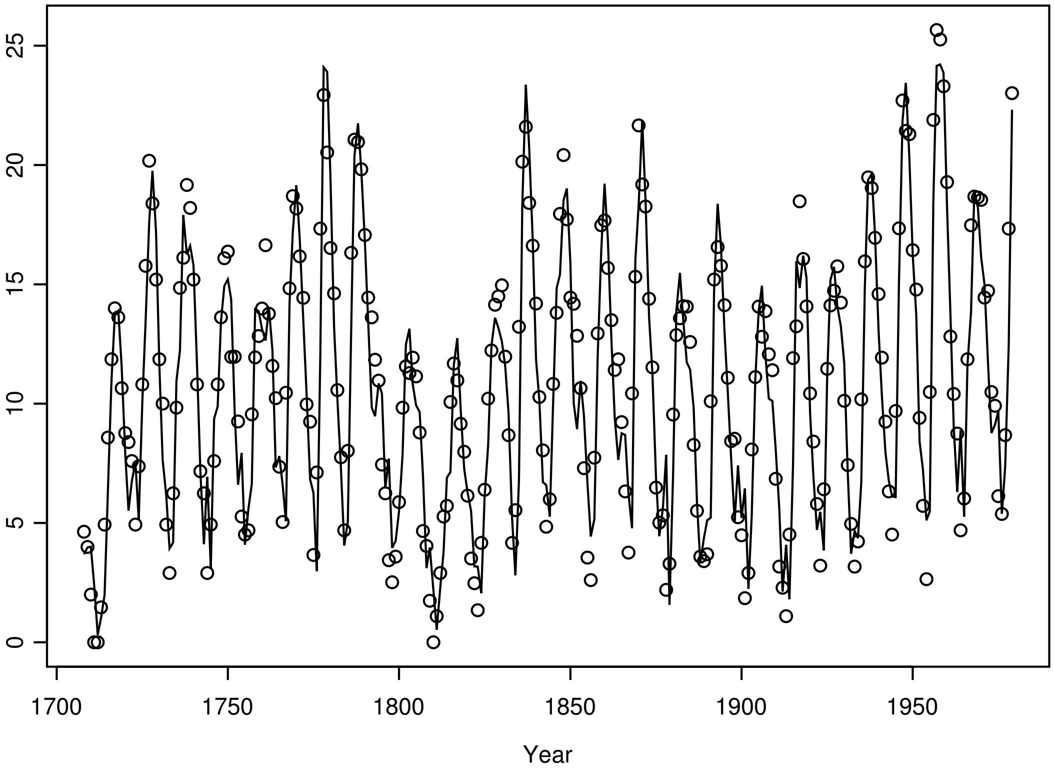


Figure 8: Wolf's Sunspot Number: time plot of the fitted values based on model (4.4) (solid line), with the observed values (circles).