Estimation of semi-parametric additive coefficient model

Lan Xue*, Lijian Yang

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

Received 25 February 2004; received in revised form 29 September 2004; accepted 5 November 2004

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
A general function of a large number of variables without additional structures. Many efforts have been made to impose structures on the regression function to partly alleviate the “curse of dimensionality”, which is broadly described as dimension reduction. Some well-known dimension reduction approaches are: (generalized) additive models (Chen and Tsay, 1993a; Hastie and Tibshirani, 1990; Sperlich et al., 2002; Stone, 1985), partially linear models (Härdle et al., 2000), and varying coefficient models (Hastie and 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 \{\((Y_i, X_i, T_i)\)\}_i=1^n is drawn that satisfies

\[ Y_i = m(X_i, T_i) + \sigma(X_i, T_i)\varepsilon_i, \quad (1.1) \]

where for the response variables \(Y_i\) and predictor vectors \(X_i\) and \(T_i\), \(m\) and \(\sigma^2\) are the conditional mean and variance functions

\[ m(X_i, T_i) = E(Y_i|X_i, T_i), \quad \sigma^2(X_i, T_i) = \text{var}(Y_i|X_i, T_i) \quad (1.2) \]

and \(E(\varepsilon_i|X_i, T_i) = 0, \text{var}(\varepsilon_i|X_i, T_i) = 1\). For the varying coefficient model, the conditional mean takes the following form:

\[ m(X, T) = \sum_{l=1}^d \alpha_l(X_l)T_l \quad (1.3) \]

in which all tuning variables \(X_l, l = 1, \ldots, d\) make up the vector \(X\), and all linear predictor variables \(T_l, l = 1, \ldots, d\) are univariate and distinct. Hastie and Tibshirani (1993) proposed a backfitting algorithm to estimate the varying coefficient functions \(\{\alpha_l(x_l)\}^d_{l=1}\), 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 and Tsay (1993b) and later in the context of longitudinal data by Hoover et al. (1998), in which all the tuning variables \(X_l, l = 1, \ldots, d\) are the same and univariate. For more recent developments of the functional coefficient model, see Cai et al. (2000). In a different direction, Yang et al. (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 \(X_l, l = 1, \ldots, d\) have to be equal. On the other hand, it requires that none of the tuning variables \(X_l, l = 1, \ldots, d\) are equal.

In this paper, we propose the following additive coefficient model which has a more flexible form, namely

\[ m(X, T) = \sum_{l=1}^{d_1} \alpha_l(X)T_l, \quad \alpha_l(X) = \sum_{s=1}^{d_2} \alpha_{ls}(X_s), \quad \forall 1 \leq l \leq d_1 \quad (1.4) \]

in which the coefficient functions \(\{\alpha_l(X)\}^{d_1}_{l=1}\) are additive functions of the tuning variables \(X = (X_1, \ldots, X_{d_2})^T\). Note that without the additivity restriction on the coefficient functions
Table 1
German real GNP: the ASEs and ASPEs of four fits

<table>
<thead>
<tr>
<th>Model</th>
<th>ASE</th>
<th>ASPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Additive coefficient model (1.5), local linear fit</td>
<td>0.000201</td>
<td>0.000085</td>
</tr>
<tr>
<td>Additive coefficient model (1.5), local cubic fit</td>
<td>0.000205</td>
<td>0.000077</td>
</tr>
<tr>
<td>Linear AR (24) model (4.2)</td>
<td>0.000253</td>
<td>0.000112</td>
</tr>
<tr>
<td>Linear AR (248) model (4.3)</td>
<td>0.000258</td>
<td>0.000116</td>
</tr>
</tbody>
</table>

\[
\{ z_l(\mathbf{X}) \}_{l=1}^{d_1}, \text{ model (1.4) would be a kind of functional coefficient model with a multivariate tuning variable } \mathbf{X} \text{ instead of a univariate one as in the existing literature. The additive structure is imposed on the coefficient functions } \{ z_l(\mathbf{X}) \}_{l=1}^{d_1}, \text{ so that inference can be made on them without the “curse of dimensionality”}.
\]

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 and 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 and Tsay (1993a), Hastie and Tibshirani (1990).
3. When for any fixed \( l = 1, \ldots, d_1, \) \( x_{ls}(x_s) \equiv 0 \) for all but one \( s = 1, \ldots, d_2 \), (1.4) reduces to the varying coefficient model (1.3) of Hastie and Tibshirani (1993).
4. When \( d_1 = d_2 = d \), and \( x_{ls}(x_s) \equiv 0 \) for \( l \neq s \), (1.4) reduces to the varying-coefficient model of Yang et al. (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 124th quarter being from January 1, 1990 to April 1, 1990). Yang and Tschernig (2002) deseasonalized this series by removing the four seasonal means from the series \( \log(G_t+4/G_t+3), t = 1, \ldots, 120 \). 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 Section 4.2, we have fitted the following additive coefficient model (details in Section 4.2)

\[
Y_t = \{ c_1 + x_{11}(Y_{t-1}) + x_{12}(Y_{t-8}) \} Y_{t-2} + \{ c_2 + x_{21}(Y_{t-1}) + x_{22}(Y_{t-8}) \} Y_{t-4} + \sigma e_t.
\]

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 10 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. Fig. 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.

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 \( m(X, T) \) in (1.4) needs to be identified. One practical solution is to rewrite it as

\[
m(X, T) = \sum_{l=1}^{d_1} \alpha_l(X) T_l, \quad \alpha_l(X) = c_l + \sum_{s=1}^{d_2} \alpha_{ls}(X_s), \quad \forall 1 \leq l \leq d_1
\]  

with the identification conditions

\[
E\{w(X)\alpha_{ls}(X_s)\} \equiv 0, \quad l = 1, \ldots, d_1, \quad s = 1, \ldots, d_2
\]  

for some nonnegative weight function \( w \), with \( E\{w(X)\} = 1 \). The weight function \( w \) is introduced so that estimation of the unknown functions \( \{\alpha_l(X)\}_{1 \leq l \leq d_1} \) 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 \( X \) is not required to be compact. This relaxation is very desirable since most time series distributions are not compactly supported. See Yang and 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 \( m(X, T) = \sum_{l=1}^{d_1} \sum_{s=1}^{d_2} \alpha_{ls}(X_s) T_l \) can be reorganized to satisfy (2.2), by writing

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

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

In addition, for the functions \( \{\alpha_{ls}(X_s)\}_{1 \leq s \leq d_2} \) and parameters \( \{c_l\}_{1 \leq l \leq d_1} \) to be uniquely determined, one imposes an additional assumption (A0) There exists a constant \( C > 0 \) such that for any set of measurable functions \( \{b_{ls}(X_s)\}_{1 \leq s \leq d_2} \) that satisfy (2.2) and any set of constants \( \{a_l\}_{1 \leq l \leq d_1} \), the following holds:

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

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

Proof. Suppose that

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

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

\[
b_{ls}(X_s) \equiv \bar{\alpha}_{ls}(X_s) - \alpha_{ls}(X_s), \quad a_l \equiv \bar{c}_l - c_l
\]

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

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

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 \( s, l \). \( \square \)
2.2. The model

Consider \( \{(Y_i, X_i, T_i)\}_{i=1}^n \), 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 \( \{e_i\}_{i=1}^n \) are i.i.d with \( Ee_i = 0 \), \( Ee_i^2 = 1 \), and with the additional property that \( e_i \) is independent of \( \{(X_j, T_j)\}_{j \neq i} \), \( i = 1, \ldots, n \). With this error structure, the explanatory variable vector \( (X_i, T_i) \) can contain exogenous variables and/or lag variables of \( Y_i \). If \( (X_i, T_i) \) 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 \( l = 1, \ldots, d_2 \), the coefficient of \( T_{il} \) consists of two parts, the unknown parameter \( c_l \), and the unknown univariate functions \( \{z_{ls}\}_{1 \leq s \leq d_2} \). The marginal integration method will be applied to estimate both. The marginal integration method was first discussed in Linton and Nielsen (1995) in the context of additive models, see also the marginal integration method for generalized additive models in Linton and Härdle (1996). To see how the marginal integration method works in our context, observe that according to the identification condition (2.2), for every \( l = 1, \ldots, d_1 \) one has

\[
c_l = E\{w(X)z_l(X)\} = \int w(x)z_l(x)\varphi(x) \, dx
\]

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

\[
c_l + z_{ls}(x_s) = E\{w_{-s}(X_{-s})z_l(x_s, X_{-s})\}
= \int w_{-s}(u_{-s})z_l(x_s, u_{-s})\varphi_{-s}(u_{-s}) \, du_{-s},
\]

where \( u_{-s} = (u_1, \ldots, u_{s-1}, u_{s+1}, \ldots, u_{d_2})^T \), \( (x_s, u_{-s}) = (u_1, \ldots, u_{s-1}, x_s, u_{s+1}, \ldots, u_{d_2})^T \), the density of \( X \) is \( \varphi \), and the marginal density of \( X_{-s} = (X_1, \ldots, X_{s-1}, X_{s+1}, \ldots, X_{d_2})^T \) is \( \varphi_{-s} \), and \( w_{-s}(x_{-s}) = E\{w(X_{s}, x_{-s})\} = \int w(u, x_{-s}) \, d\varphi_{-s}(u) \). In addition, the marginal density of \( X_s \) is denoted by \( \varphi_s \). Intuitively, one has

\[
c_l \approx \frac{\sum_{i=1}^n w(X_i)z_l(X_i)}{\sum_{i=1}^n w(X_i)}, \quad z_{ls}(x_s) \approx \frac{\sum_{i=1}^n w_s(X_{i,-s})z_l(x_s, X_{i,-s})}{\sum_{i=1}^n w_s(X_{i,-s})} - c_l
\]

and the \( d_2 \)-dimensional functions \( \{z_l(x)\}_{l=1}^{d_1} \) 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 \( \{z_l(x)\}_{l=1}^{d_1} \) at those data points \( X_i \) that are in the support of the weight function \( w \). More generally, for any fixed \( x \in \text{supp}(w) \), we
approximate $z_l(x)$ locally by a constant $z_l$, and estimate \( \{z_l(x)\}_{l=1}^{d_1} \) by minimizing the following weighted sum of squares with respect to $z = (z_1, \ldots, z_{d_1})^T$: \[
\sum_{i=1}^{n} \left\{ Y_i - \sum_{l=1}^{d_1} z_l T_{il} \right\}^2 K_H(X_i - x), \quad (2.7)
\]
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 

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

Let $\hat{z} = (\hat{z}_1, \ldots, \hat{z}_{d_1})^T$ be the solution to the least-squares problem in (2.7). Note that $\hat{z}$ is dependent on $x$, as is (2.7), and the components in $\hat{z}$ give the estimators for $\{z_l(x)\}_{l=1}^{d_1}$. To emphasize the dependence on $x$, we write $\hat{z} = \hat{z}(x) = (\hat{z}_1(x), \ldots, \hat{z}_{d_1}(x))^T$. More precisely, let 

\[
W(x) = \text{diag}\{K_H(X_i - x)/n\}_{1 \leq i \leq n}, \quad Z = \begin{pmatrix} T_{11} & \cdots & T_{1d_1} \\ \vdots & \ddots & \vdots \\ T_{n1} & \cdots & T_{nd_1} \end{pmatrix},
\]

\[
Y = (Y_1, \ldots, Y_n)^T
\]

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

\[
\hat{z}_l(x) = e_l^T (Z^T W(x) Z)^{-1} Z^T W(x) Y. \quad (2.8)
\]

By (2.6), the parameter $c_l$ can be estimated as a weighted average of $\hat{z}_l(X_i)$’s, i.e., 

\[
\hat{c}_l = \frac{\sum_{i=1}^{n} w(X_i) \hat{z}_l(X_i)}{\sum_{i=1}^{n} w(X_i)}, \quad l = 1, \ldots, d_1. \quad (2.9)
\]

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

\[
\sqrt{n}(\hat{c}_l - c_l) \overset{d}{\to} N\{0, \sigma_l^2\},
\]

where the asymptotic variance $\sigma_l^2$ is defined in (A.11) 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 $\hat{z}_l(x)$. First, the bias of $\hat{z}_l(x)$ in estimating $z_l(x)$ 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 $P_{\varepsilon_1}$. Second, the usual variance of $\hat{z}_l(x)$ in estimating $z_l(x)$ 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 $P_{\varepsilon_2}$ in (A.9) and (A.10). This technique of simultaneously reducing the bias by the use of a higher-order kernel and “integrating out the variance” is the common feature of all marginal integration procedures.
2.4. Estimators of functions \( \{ \alpha_{ls} \}_{1 \leq l \leq d_1} \)

In the following, we illustrate the procedure for estimating the functions \( \{ \alpha_{ls}(x_s) \}_{1 \leq l \leq d_1} \), for any fixed \( s = 1, \ldots, d_2 \). 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 \( \{ \alpha_l(x) \}_{l=1}^{d_1} \) at those points \( (x_s, X_{i,-s}) \) that lie in the support of \( w \). For any \( x \in \text{supp}(w) \), differently from estimating the constants, we approximate the function \( \alpha_l(x) \) locally at \( x \) by \( \alpha_l(x) \approx \alpha_l + \sum_{j=1}^{p} \beta_{lj} (u_s - x_s)^j \), and estimate \( \{ \alpha_l(x) \}_{l=1}^{d_1} \) by minimizing the following weighted sum of squares with respect to \( \alpha = (\alpha_1, \ldots, \alpha_{d_1})^T \), \( \beta = (\beta_{11}, \ldots, \beta_{1p}, \ldots, \beta_{d_11}, \ldots, \beta_{d_1p})^T \)

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

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 \( G_s = \text{diag}\{g_1, \ldots, g_{s-1}, g_{s+1}, \ldots, g_{d_2}\} \), and

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

\[
L_{G_s}(u_{-s}) = \frac{1}{\prod_{1 \leq s' \leq d_2, s' \neq s} g_{s'}} \left[ \begin{array}{cccc}
g_1 & u_{s-1} & u_{s+1} & \ldots & u_{d_2} \\
g_1 & g_{s-1} & g_{s+1} & \ldots & g_{d_2} \\
g_1 & g_{s-1} & g_{s+1} & \ldots & g_{d_2} \\
\vdots & \ddots & \ddots & \ddots & \vdots \\
g_1 & u_{s-1} & u_{s+1} & \ldots & u_{d_2} \\
\end{array} \right]
\]

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

\[
\hat{\alpha_l}(x) = e_l^T [Z_s^TW_s(x)Z_s]^{-1} Z_s^TW_s(x)Y,
\]

(2.10)

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

\[
W_s(x) \equiv \text{diag}\{n^{-1} k_{h_s}(X_{is} - x_s)L_{G_s}(X_{i,-s} - x_{-s})\}_{1 \leq i \leq n}
\]

and

\[
Z_s = \begin{bmatrix}
T_1^T, \{ (X_{1s} - x_s)/h_s \} T_1^T, \ldots, \{ (X_{1s} - x_s)/h_s \}^p T_1^T \\
\vdots \\
T_n^T, \{ (X_{ns} - x_s)/h_s \} T_n^T, \ldots, \{ (X_{ns} - x_s)/h_s \}^p T_n^T \\
\end{bmatrix}
\]

\[
= \begin{bmatrix}
[p((X_{1s} - x_s)/h_s)]^T \otimes T_1^T \\
\vdots \\
[p((X_{ns} - x_s)/h_s)]^T \otimes T_n^T \\
\end{bmatrix}
\]

(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 \( \alpha_{ls} \) for \( l = 1, \ldots, d_1 \).
simultaneously, which are given by

\[
\hat{g}_{ls}(x_s) = \frac{\sum_{i=1}^{n} w_{-s}(X_{i,-s}) \hat{g}_l(x_s, X_{i,-s})}{\sum_{i=1}^{n} w_{-s}(X_{i,-s})} - \hat{c}_l,
\]

(2.12)

where the term \(\hat{c}_l\) is the \(\sqrt{n}\)-consistent estimator of \(c_l\) in Theorem 1. The estimator \(\hat{g}_{ls}(x_s)\) is referred to as the \(p\)th order local polynomial estimator, where \(p\) is the highest polynomial degree of variables \(X_{i,s} - x_s, i = 1, \ldots, n\), in the definition of the design matrix \(Z_s\) 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 \(x = (x_1, \ldots, x_{d_2})^T \in \text{supp}(w)\), one has for \(l = 1, \ldots, d_1, s = 1, \ldots, d_2\)

\[
\sqrt{n h_s} \left( \hat{g}_{ls}(x_s) - g_{ls}(x_s) - h_s^{p+1} \eta_{ls}(x_s) \right) \overset{d}{\to} N\{0, \sigma_{ls}^2(x_s)\},
\]

(2.13)

where \(\eta_{ls}(x_s)\) and \(\sigma_{ls}^2(x_s)\) are defined in (A.16) and (A.18), respectively.

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

\[
\hat{\eta}(x, t) = \sum_{l=1}^{d_1} \sum_{s=1}^{d_2} \left\{ \hat{c}_l + \hat{g}_{ls}(x_s) \right\} l_t.
\]

(2.14)

To appreciate why \(\hat{g}_{ls}(x_s)\) can be estimated by \(\hat{g}_{ls}\) at the rate of \(1/\sqrt{n h_s}\), which is the same as the rate of estimating a nonparametric function in the univariate case, we discuss two special features of \(\hat{g}_l(x)\) given in (2.10), which are similar to those discussed in Section 2.3. First the bias of \(\hat{g}_l(x)\) in estimating \(g_l(x)\) is of order \(h_s^{p+1} + g_{\text{max}}^q\), where the first term can be understood as the approximation bias caused by locally approximating \(g_l\) 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 \(\{g_{ls'}\}_{s' \neq s}\) using a constant, which is bounded by \(g_{\text{max}}^q\) since the kernel \(L\) is of order \(q_2\), see \(P_{s3}\) in Lemma A.9. The order \(g_{\text{max}}^q\) of the second bias term is negligible compared to the rescaling factor of order \(1/\sqrt{n h_s}\), according to (A6) (b). Hence, only the first bias term appears in the asymptotic distribution formula (2.13). As for the variance of \(\hat{g}_l(x)\) in estimating \(g_l(x)\), 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 \(X_{i,-s}\), the bandwidths \(g_1, \ldots, g_{s-1}, g_{s+1}, \ldots, g_{d_2}\) related to \(X_{i,-s}\) are integrated out, see \(P_{s1}\) in Lemma A.9. Then the variance of \(\hat{g}_{ls}\) is reduced to the order \(n^{-1} h_s^{-1}\). If the same bandwidth \(h_s\) is used for all variable directions in \(X\), then Assumption (A6) (b) would imply that \(n n^{-d_2/(2p+3)} \to \infty\) 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 \(g_1, \ldots, g_{s-1}, g_{s+1}, \ldots, g_{d_2}\) 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 \( H = \text{diag}\{h_{0,1}, \ldots, h_{0,d_2}\}, G_s = \text{diag}\{g_1, \ldots, g_{s-1}, g_s+1, \ldots, g_{d_2}\} \) and \( \{h_s\}_1 \leq s \leq d_2 \). In the following, we discuss the choices of such bandwidths.

- Note from Theorem 1 that the asymptotic distributions of the estimators \( \{\hat{c}_l\}_l \) depend only on the quantity \( \sigma_l^2 \), not on the bandwidths in \( H \). Hence we have only specified that \( H \) satisfy the order assumptions in (A6) (a) by taking \( \hat{h}_{01} = \cdots = \hat{h}_{0d_2} = \sqrt{\text{var}(X)} \log(n)n^{-1/(2d_l-1)} \), where \( q_1 \) is the order of the kernel \( K \), required to be greater than \( (d_2 + 1)/2 \), and \( \text{var}(X) = \{ \prod_{s=1}^{d_2} \text{var}(X_s) \}^{1/d_2} \), in which \( \text{var}(X) \) denotes the sample variance of \( X_s \), \( s = 1, \ldots, d_2 \).

- The asymptotic distributions of the estimators \( \{\hat{z}_l\}_1 \leq l \leq d_1 \) depend not only on the functions \( \eta_l \) and \( \sigma_l^2 \) but also crucially on the choice of bandwidths \( h_s \). Moreover, for each \( s = 1, \ldots, d_2 \), the coefficient functions \( \{z_{l,s}(x_s), l = 1, \ldots, d_1\} \) 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 \( \{\hat{z}_{l,s}(x_s), l = 1, \ldots, d_1\} \), which is defined as

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

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

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

in which \( \eta_{l,s}(x_s) \) and \( \sigma_{l,s}^2(x_s) \) are the asymptotic bias and variance of \( \hat{z}_{l,s} \) as in (A.16) and (A.18). According to the definitions of \( \eta_{l,s}(x_s) \) and \( \sigma_{l,s}^2(x_s) \), \( \int \eta_{l,s}^2(x_s) \, dx_s \) and \( \int \sigma_{l,s}^2(x_s) \, dx_s \) can be approximated, respectively, by

\[
\int \left[ \frac{1}{(p + 1)!} \sum_{l'=-1}^{d_1} \sum_{l=1}^{d_1} \partial_{x_{l,s}}^{(p+1)}(x_{s}) \int u^{p+1} \frac{1}{n} \sum_{i=1}^{n} \{w_{-s}(X_{i,-s})T_{il}/K_{l,s}^*\}
\right]^2 \, dx_s,
\]

\[
\frac{1}{n} \sum_{i=1}^{n} \frac{w_{-s}^2(X_{i,-s})q_{-s}^2(X_{i,-s})}{q_{-s}(X_{i})} \int K_{l,s}^*(u, X_{i}, T_{i}) \, du,
\]

where the functions \( K_{l,s}^* \) are defined in (A.17).
To implement this, one needs to evaluate terms such as $\hat{d}_{ls}^{(p+1)}(x_s)$, $\sigma^2(x, \mathbf{t})$, $\varphi(x)$, $\varphi(x_{-s})$ 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 $\hat{z}_{ls}^{(p+1)}(x_s)$ are estimated by fitting a polynomial regression model of degree $p + 2$

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

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

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

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

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

with the standard normal density $\phi$ and the rule-of-the-thumb bandwidth

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

3. According to the definition in (A.17), the dependence of the functions $K_{ls}^*(u, \mathbf{x}, \mathbf{t})$ on $u$ and $\mathbf{t}$ is explicitly known. The only unknown term $E(\mathbf{T}\mathbf{T}^T|\mathbf{X} = \mathbf{x})$ contained in $S_{-1}(\mathbf{x})$ is estimated by fitting matrix polynomial regression

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

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

In this procedure, one simply uses polynomial regression to estimate some of the unknown quantities, which is easy to implement, but may lead the estimated optimal bandwidths to be biased relative to the true optimal bandwidths. The development of a more sophisticated bandwidth selection method requires further investigation.

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

Following the above discussion, the order of the kernels $K$ and $L$ are required to be greater than $(d_2 + 1)/2$ and $(d_2 - 1)/2$, respectively. If the dimension of $X$ equals to 2, kernels $K$
and $L$ can have order 2. We have used the quadratic kernel $k(u) = 15/16(1 - u^2)^21_{|u| \leq 1}$, where $1_{|u| \leq 1}$ is the indicator function of $[-1, 1]$ and the kernels $K, L$ are product kernels.

Lastly, the matrix $Z^T W(x) Z$ in (2.7) is computed as $Z^T W(x) Z + n^{-1}TT^T$, and the matrix $Z_s^T W_s(x) Z_s$ in (2.10) as $Z_s^T W_s(x) Z_s + (n \hat{h}_{s, \text{opt}})^{-1} \sqrt{\text{var}(X)} \{ \int k(u)p(u)p(u)^T du \} \otimes TT^T$, following the ridge regression idea of Seifert and Gasser (1996).

4. Examples

4.1. A simulated example

The data are generated from the following model

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

with

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

where $X = (X_1, X_2)^T$ is uniformly distributed on $[-\pi, \pi] \times [-\pi, \pi]$, and $T = (T_1, T_2)^T$ follows the bivariate standard normal distribution. The vectors $X, T$ are generated independently. The error term $\varepsilon$ is a standard normal random variable and independent of $(X, T)$. 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 Fig. 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 $\alpha_{11}$ and $\alpha_{21}$. 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

![Fig. 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 $n = 500.$](image-url)
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 $\hat{z}_{11}$, $\hat{z}_{12}$, $\hat{z}_{21}$, $\hat{z}_{22}$ of the coefficient functions $z_{11}$, $z_{12}$, $z_{21}$, $z_{22}$.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>$c_1 = 2$ (Mean)</th>
<th>$c_2 = 1$ (Mean)</th>
<th>$z_{11}$ (Mean)</th>
<th>$z_{12}$ (Mean)</th>
<th>$z_{21}$ (Mean)</th>
<th>$z_{22}$ (Mean)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 100$</td>
<td>1.9737 (0.3574)</td>
<td>1.0406 (0.2503)</td>
<td>0.1609</td>
<td>0.2541</td>
<td>0.1205</td>
<td>0.2761</td>
</tr>
<tr>
<td>$n = 250$</td>
<td>2.0299 (0.2410)</td>
<td>1.0056 (0.1490)</td>
<td>0.0568</td>
<td>0.0963</td>
<td>0.0338</td>
<td>0.0649</td>
</tr>
<tr>
<td>$n = 500$</td>
<td>1.9786 (0.1680)</td>
<td>1.0026 (0.1111)</td>
<td>0.0295</td>
<td>0.0483</td>
<td>0.0191</td>
<td>0.0310</td>
</tr>
</tbody>
</table>

As the sample size increases, similar results are also obtained for $h_{2,\text{opt}}$, the optimal bandwidth for estimating $z_{12}$ and $z_{22}$. 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 a simple polynomial function as a plug-in substitute of the true regression function.

Second, we estimate the functions $\hat{z}_{ls}$ on a grid of equally spaced grid of points $x_m$, $m = 1, \ldots, n_{\text{grid}}$ 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 $\hat{c}_l$, $l = 1, 2$ and the averaged integrated squared errors (AISE) of $\hat{z}_{ls}$. By denoting the estimated function $\hat{z}_{ls}$ of $z_{ls}$ in the $i$th replication by $\hat{z}_{ls}^i$, we define

\[
\text{ISE}(\hat{z}_{ls}^i) = \frac{1}{n_{\text{grid}}} \sum_{m=1}^{n_{\text{grid}}} (\hat{z}_{ls}^i(x_m) - z_{ls}(x_m))^2
\]

and

\[
\text{AISE}(\hat{z}_{ls}) = \frac{1}{100} \sum_{i=1}^{100} \text{ISE}(\hat{z}_{ls}^i).
\]

Fig. 3 gives the plots of $\hat{z}_{11}$, $\hat{z}_{12}$, $\hat{z}_{21}$, $\hat{z}_{22}$ 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.

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 and 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 and 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 and Tschernig (2002). According to Brockwell and 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

Linear AR (24):

\[
Y_t = a_1 Y_{t-2} + a_2 Y_{t-4} + \sigma \varepsilon_t.
\]  

(4.2)

Linear AR (248):

\[
Y_t = b_1 Y_{t-2} + b_2 Y_{t-4} + b_3 Y_{t-8} + \sigma \varepsilon_t.
\]  

(4.3)
From Table 1, it is clear that besides being more parsimonious, the linear model (4.2) has smaller average squared prediction error, compared with model (4.3). Thus model (4.2) is the preferred linear autoregressive model. Moreover, Figs. 4 and 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.

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 $Z_i$ (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 Fig. 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 Fig. 6, one may observe that the estimated functions have obviously nonconstant forms. In addition, their 95% confidence intervals cannot 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.)
For the two linear autoregressive models, we estimate their constant coefficients by maximum likelihood method. The estimated coefficients are $\hat{a}_1 = -0.2436$, $\hat{a}_2 = 0.5622$ 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 model (1.5) using local linear estimation (i.e. taking $p = 1$ in $Z_s$). 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 and Tsay (1993b) and Cai et al. (2000) both used a FAR model with lag 3 as the tuning variable. Xia and 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.

\[
Y_t = \{c_1 + \alpha_{11}(Y_{t-3}) + \alpha_{12}(Y_{t-8})\}Y_{t-1} + \{c_2 + \alpha_{21}(Y_{t-3}) + \alpha_{22}(Y_{t-8})\}Y_{t-2} \\
\times \{c_3 + \alpha_{31}(Y_{t-3}) + \alpha_{32}(Y_{t-8})\}Y_{t-3} + \sigma \epsilon_t. \tag{4.4}
\]

Following the convention in the literature, we use the transformed data, where \( Y_t = 2(\sqrt{T + X_t} - 1) \), \( X_t \) 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 Fig. 7, and the time plot of the fitted values in Fig. 8. The 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 and Tsay (1993), denoted as FAR1, and the following two models; the FAR model of Cai et al. (2000) denoted as FAR2

\[
Y_t = \alpha_1(Y_{t-3})Y_{t-1} + \alpha_2(Y_{t-3})Y_{t-2} + \alpha_3(Y_{t-3})Y_{t-3} \\
+ \alpha_6(Y_{t-3})Y_{t-6} + \alpha_8(Y_{t-3})Y_{t-8} + \sigma \epsilon_t \tag{4.5}
\]
and the single index coefficient model of Xia and Li (1999) denoted as SIND

\[ Y_t = \phi_0\{g_4(\theta, Y_{t-3}, Y_{t-8})\} + \phi_1\{g_4(\theta, Y_{t-3}, Y_{t-8})\}Y_{t-1} + \phi_2\{g_4(\theta, Y_{t-3}, Y_{t-8})\} \]
\[ \times Y_{t-2} + \phi_3\{g_4(\theta, Y_{t-3}, Y_{t-8})\}Y_{t-3} + \phi_4\{g_4(\theta, Y_{t-3}, Y_{t-8})\}Y_{t-8} + \sigma e_t, \]

(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, of Cai et al. (2000, p. 952), 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
Table 3
Out-of-sample absolute prediction errors for sunspot data, under different models

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

Results on models TAR, FAR1 and FAR2 are from Cai et al. (2000, Table 4, p. 951), while results on model SIND are from Xia and Li (1999, Table 2, p. 1280).

use of model (4.5) has not been theoretically justified. Similarly, model (4.6) is also not theoretically justified, since according to Condition C5, of Xia and Li (1999, p. 1277), 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

$$\tilde{\phi}_0(u) = \phi_0(u) + uf(u), \quad \tilde{\phi}_1(u) = \phi_1(u), \quad \tilde{\phi}_2(u) = \phi_2(u),$$

$$\tilde{\phi}_3(u) = \phi_3(u) - \cos(\theta)f(u), \quad \tilde{\phi}_4(u) = \phi_4(u) - \sin(\theta)f(u).$$

It is straightforward to verify that the new set of functions $\{\tilde{\phi}_0, \ldots, \tilde{\phi}_4\}$ satisfy (4.6) as well. One possible fix of this problem is to drop either one of the terms $\phi_3\{g_4(\theta, Y_{t-3}, Y_{t-8})\}Y_{t-3}$ and $\phi_4\{g_4(\theta, Y_{t-3}, Y_{t-8})\}Y_{t-8}$ from (4.6), then the model is fully identifiable and satisfies Condition C5, of Xia and Li (1999, p. 1277). 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 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 and Davis (1991, p. 304).

Acknowledgements

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
Appendix A.

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. $\int K(u)\,du = 1$ while $\int K(u)u_1^{r_1}\cdots u_{d_2}^{r_{d_2}}\,du = 0$, 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 $z_i(x_s)$ have bounded continuous $p^*$th derivatives for $l = 1, \ldots, d_1, s = 1, \ldots, d_2$.

(A3) The vector process $\{\zeta_i\}_{i=1}^\infty = \{(Y_i, X_i, T_i)\}_{i=1}^\infty$ is strictly stationary and $\beta$-mixing with the $\beta$-mixing coefficient $\beta(k) \leq cp^k$, for some constants $c > 0$, $0 < p < 1$. The $\beta$-mixing coefficient is defined as

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

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

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

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

(A4) The error term satisfies:

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

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

(A5) The vector $(X, T)$ has a joint probability density $\psi(x, t)$. The marginal densities of $X, X_s$ and $X_{-s}$ are denoted by $\varphi, \varphi_s$ and $\varphi_{-s}$, respectively.

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

(b) Let $S(x) = E(TT^T | X = x)$. We assume there exists a $c > 0$, such that $S(x) \geq cI_{d_2}$ uniformly for $x \in \text{supp}(w)$. Here $I_{d_2}$ is the $d_2 \times d_2$ identity matrix.
(c) The random matrix $TT^T$ satisfies the Cramer’s moment condition, i.e., there exists a positive constant $c$, such that $E|T_jT_l|^2 \leq c^2k|E|T_jT_l|^2$, and $E|T_jT_l|^2 \leq c$ holds uniformly for $k = 3, 4, \ldots$, and $1 \leq l, l' \leq d_1$.

(A6) The bandwidths satisfy:

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

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

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

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 $\beta$-mixing coefficients $\beta(k), k = 1, 2, \ldots$, and $r$ a fixed positive integer. Let $\{\theta_n(F)\}$ denote the functionals of the distribution function $F$ of $\xi_i$

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

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

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

$$\sup_{i_1, \ldots, i_m} \int |g_n(x_1, \ldots, x_m)|^{2+\delta} \, dF_{\xi_1, \ldots, \xi_m}(x_1, \ldots, x_m) \leq M_{n,c} < +\infty,$$

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

$$\theta_n(F_n) = \int g_n(x_1, \ldots, x_m) \, dF_n(x_1) \cdots dF_n(x_m)$$

$$= \frac{1}{n^m} \sum_{i_1=1}^{n} \cdots \sum_{i_m=1}^{n} g_n(\xi_{i_1}, \ldots, \xi_{i_m}),$$

$$U_n = \frac{1}{\binom{n}{m}} \sum_{1 \leq i_1 < \cdots < i_m \leq n} g_n(\xi_{i_1}, \ldots, \xi_{i_m})$$
allow decompositions as

$$
\theta_n(F_n) = \theta_n(F) + \sum_{c=1}^{m} \binom{m}{c} V_n^{(c)},
$$

$$
V_n^{(c)} = \int g_{n,c}(x_1, \ldots, x_c) \prod_{j=1}^{c} \{dF_n(x_j) - dF(x_j)\},
$$

$$
U_n = \theta_n(F) + \sum_{c=1}^{m} \binom{m}{c} U_n^{(c)},
$$

$$
U_n^{(c)} = \frac{(n - c)!}{n!} \sum_{1 \leq i_1 < \cdots < i_c \leq n} \int g_{n,c}(x_{i_1}, \ldots, x_{i_c}) \prod_{j=1}^{c} [dI_{R^d_+}(x_j - \zeta_{i_j}) - dF(x_j)],
$$

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

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

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

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

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

(A.3)

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

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

(A.4)

**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.3), provides an obvious venue of extension to the more general setup. Elementary arguments then establish (A.4) under geometric mixing conditions. □
A.2. Proofs of Theorems 1 and 2

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

\[
Z^T W(x)Z = \frac{1}{n} \sum_{i=1}^{n} K_H(X_i - x) T_i^T,
\]

\[
Z_s^T W_s(x)Z_s = \frac{1}{n} \sum_{i=1}^{n} k_h_s(X_{is} - x_s) L_{G_s}(X_{i, -s} - x_{-s})
\times \left[ \left\{ p \left( \frac{X_{is} - x_s}{h_s} \right) p^T \left( \frac{X_{is} - x_s}{h_s} \right) \right\} \otimes (T_i^T) \right]
\]

in which, \( \otimes \) denotes the Kronecker product of matrices. Define also the following matrix

\[
S_2(x) = \left\{ \int k(u) p(u) p(u)^T \, du \right\} \otimes S(x), \tag{A.5}
\]

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

**Lemma A.2.** Let \( b_1 = \ln n (h_{\text{max}}^{q_1} + 1/\sqrt{nh_{\text{prod}}}) \), \( b_2 = \ln n (h_s + g_{\text{max}}^{q_2} + 1/\sqrt{nh_s g_{\text{prod}}}) \), and define the compact set \( B = \text{supp}(w) \subset R^{d_2} \). Under assumptions (A1)–(A6), as \( n \to \infty \), with probability one

\[
\sup_{x \in B} |Z^T W(x)Z - \varphi(x) S(x)| = o(b_1), \quad \text{and} \quad \sup_{x \in B} |Z_s^T W_s(x)Z_s - \varphi(x) S_2(x)| = o(b_2).
\]

**Proof.** The proof of this lemma makes use of the covering technique and the Bernstein’s inequality for mixing data (Bosq, 1998, Theorem 1.4). The complete proof of this lemma can be found in Xue and Yang (2004). \( \Box \)

A.2.1. Proof of Theorem 1

Now let \( v_1 \) be the integer satisfying \( b_1^{v_1+1} = o(h_{\text{max}}^{q_1+2}) \). Following immediately from Lemma A.2, one has

\[
\{Z^T W(x)Z\}^{-1} - \frac{S(x)^{-1}}{\varphi(x)} = \frac{S(x)^{-1}}{\varphi(x)} \sum_{v=1}^{v_1} \left\{ I_{d_1} - \frac{Z^T W(x)ZS^{-1}(x)}{\varphi(x)} \right\} + Q_c(x)
\]

\[
= \sum_{v=1}^{v_1} Q_{cv}(x) + Q_c(x),
\]

where the matrix \( Q_c(x) \) satisfies \( \sup_{x \in B} |Q_c(x)| = o(h_{\text{max}}^{q_1+2}) \) w.p.1.
Also by observing that, \( e_l^T \{ Z^T W(X_i)Z \}^{-1} Z^T W(X_i)Ze_{l'} = \delta_{ll'} \), where \( \delta_{ll'} \) equals to 1 if \( l = l' \) and equals to 0 otherwise, one has the following decomposition:

\[
\frac{1}{n} \sum_{i=1}^{n} w(X_i) |\hat{c}_i - c_i| = \sum_{i=1}^{2} \{ D_{ci} + P_{ci} \} + \sum_{v=1}^{v_1} F_{iv} + R_c, \tag{A.6}
\]

with \( R_c = \frac{1}{n} \sum_{i=1}^{n} w(X_i) \sum_{s=1}^{d_z} \varphi_s(X_is) \), and the terms \( D_{ci}, P_{ci}, \) and \( F_{iv} \) are given as follows:

By denoting the operator \( \Phi_c \) as

\[
\Phi_c(A, B) = \frac{1}{n} \sum_{i=1}^{n} w(X_i) A(X_i) Z^T W(X_i) B,
\]

we set, for \( i = 1, 2, \) and \( v = 1, \ldots, v_1, \)

\[
D_{ci} = \Phi_c(Q_c(x), B_i); \quad P_{ci} = \Phi_c(S^{-1}(x)/\varphi(x), B_i); \quad F_{iv} = \Phi_c(Q_{cv}(x), B_i), \tag{A.7}
\]

with

\[
B_1 = [\sigma(X_j, T_j)e_j]_{j=1,\ldots,n}, \quad B_2 = \begin{bmatrix} \sum_{l'=1}^{d_1} \sum_{s=1}^{d_z} \{ \varphi_{l's}(X_{js}) - \varphi_{l's}(X_is) \} T_{jl'} \end{bmatrix}_{j=1,\ldots,n}. \tag{A.8}
\]

To prove Theorem 1, we need the following lemmas.

**Lemma A.3.** As \( n \to +\infty \), \( |D_{c1}| + |D_{c2}| = \Theta(h_1^{q_1+2}) \) w.p.1.

**Lemma A.4.** For fixed \( v = 1, \ldots, v_1, \) as \( n \to +\infty \), \( |F_{11}| + |F_{2v}| = \Theta(b_1^v/\sqrt{n}) \) w.p.1.

**Proof.** For simplicity, consider only the case of \( F_{1v} \) with \( v = 1 \): i.e.,

\[
F_{11}(x) = \frac{1}{n} \sum_{i=1}^{n} w(X_i) S^{-1}(X_i) \left\{ I_{d_1} - \frac{Z^T W(X_i)ZS^{-1}(X_i)}{\varphi(X_i)} \right\} Z^T W(X_i)/E
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} w(X_i) S^{-1}(X_i) \left[ \frac{S(X_i)}{\varphi(X_i)} - \frac{E[Z^T W(X_i)Z]}{\varphi^2(X_i)} \right] S^{-1}(X_i) Z^T W(X_i)/E
\]

\[
- \frac{1}{n} \sum_{i=1}^{n} w(X_i) S^{-1}(X_i) \left[ \frac{Z^T W(X_i)Z}{\varphi^2(X_i)} - \frac{E[Z^T W(X_i)Z]}{\varphi^2(X_i)} \right] \times S^{-1}(X_i) Z^T W(X_i)/E = P_1 - P_2.
\]
Let $\xi_i = (X_i, T_i, e_i)$, and define

$$g_n(\xi_i, \xi_j) = w(X_i)S^{-1}(X_i) \left[ \frac{S(X_i)}{\phi(X_i)} - \frac{E[Z^TW(X_i)Z]}{\phi^2(X_i)} \right]$$

$$\times S^{-1}(X_i)K_H(X_j - X_i)T_j\sigma(X_j, T_j)e_j$$

$$+ w(X_j)S^{-1}(X_j) \left[ \frac{S(X_j)}{\phi(X_j)} - \frac{E[Z^TW(X_j)Z]}{\phi^2(X_j)} \right]$$

$$\times S^{-1}(X_j)K_H(X_i - X_j)T_i\sigma(X_i, T_i)e_i.$$  

Then $P_1$ can be written as the von Mises’ differentiable statistic $P_1 = \frac{1}{2n^2} \sum_{i,j=1}^{n} g_n(\xi_i, \xi_j)$, which can be decomposed as $P_1 = \frac{1}{2}(\theta_n(F) + 2V_n^{(1)} + V_n^{(2)})$, in which, $\theta_n(F) = 0$.

In order to write down the explicit expressions of $V_n^{(1)}$, $V_n^{(2)}$, let $E_i (E_{n,i})$ denote taking expectation with respect to the random vector indexed by $i$ using the theoretical (empirical) measure, both under the presumption of independence between $\xi_i$ and $\xi_j$. One has $V_n^{(1)} = E_i E_{n,j} g_n(\xi_i, \xi_j) = \frac{1}{n} \sum_{j=1}^{n} g_{n,1}(\xi_j)$, in which

$$g_{n,1}(\xi_j) = \int w(z)S^{-1}(z) \left[ \frac{S(z)}{\phi(z)} - \frac{E[Z^TW(z)Z]}{\phi^2(z)} \right]$$

$$\times S^{-1}(z)K_H(X_j - z)\phi(z) dzT_j\sigma(X_j, T_j)e_j.$$  

Clearly $g_{n,1}$ has mean 0 and variance of order $b_1^2$. So $V_n^{(1)} = \frac{1}{n} \sum_{j=1}^{n} g_{n,1}(\xi_j) = O_p(b_1^2/\sqrt{n})$.

For $V_n^{(2)}$, by Lemma A.1, under assumption (A3), one has for some small $\delta > 0$

$$E(V_n^{(2)})^2 \leq cn^{-2} \left\{ M_{n,0}^{2+\delta} + M_{n,1}^{2+\delta} + MN_{n,1}^{2+\delta} \right\},$$

where $M_{n,0}$, $M_{n,1}$ and $MN_{n,1}$ satisfy the inequalities as in (A2). Observe that

$$E_{i,j} g_n(\xi_i, \xi_j)^{2+\delta} \leq c b_1^{2+\delta} E|w(X_i)K_H(X_j - X_i)T_j\sigma(X_j, T_j)e_j|^{2+\delta}$$

$$\leq h_{\text{prod}}^{-(1+2\delta)(2+\delta)/(2+2\delta)} c b_1^{2+\delta}.$$  

So we can take $M_{n,0} = h_{\text{prod}}^{-(1+2\delta)(2+\delta)/(2+2\delta)} b_1^{2+\delta}$, and by setting the mixing coefficient $\rho$ 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 $\delta$ small, one has $E(P_1^2)$ is less or equal to

$$cn^{-2} \left( h_{\text{prod}}^{-(1+2\delta)(2+\delta)/(2+2\delta)} b_1^{2+\delta} \right)^{2/(2+\delta)} + cn^{-3} \left( b_1^{2+\delta} h_{\text{prod}}^{-(2+\delta)} \right)^{2/(2+\delta)} + c b_1^2/n \leq cn^{-1} b_1^2.$$  

Similarly, we can show that $E P_2^2 \leq cn^{-1} b_1^2$. So we have $F_{11} = O_p(b_1/\sqrt{n})$. □

Lemma A.5. $P_{11} = O_p(h_{\text{max}}^{q_1}) = o_p(n^{-1/2})$ as $n \to \infty$. 

ARTICLE IN PRESS

24

L. Xue, L. Yang / Journal of Statistical Planning and Inference III (III) III–III
Proof. Let $K_n^*(X, T) = e_T S^{-1}(X)T$, then $P_{c1}$ is a von Mises’ statistic, with $\theta_n$ of the form

$$
\int w(z) K_j^*(z, t) K_H(x - z) \left[ \sum_{l'=1}^{d_1} \sum_{s=1}^{d_2} \{ z_{l's}(x_s) - z_{l's}(z_s) \} t_{l'} \right] \phi(z) \psi(x, t) \, dz \, dx \, dt.
$$

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

$$
\int w(z) K_j^*(z, t) K(u) \left[ \sum_{l'=1}^{d_1} \sum_{s=1}^{d_2} \{ z_{l's}(z_s + h_{0,s} u_s) - z_{l's}(z_s) \} t_{l'} \right] \times \psi(z + Hu, t) \, du \, dz \, dt,
$$

which is $O(\Delta_{1})$ by a Taylor expansion of $z_{l's}(z_s + h_{0,s} u_s)$ to $q_1$th degree and of $\psi(z + Hu, t)$ 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 $\Delta_{1}$ disappear. So the leading term is of $\Delta_{1}$ order. It is routine to verify that $\nu_{n}^{(1)}$ and $\nu_{n}^{(2)}$ are $O(\Delta_{1})$ as well. Hence $P_{c1} = O_p(\Delta_{1})$ and assumption (A6) (a) entails that $O_p(\Delta_{1}) = o_p(n^{-1/2})$.

We can finish the proof of Theorem 1 as follows. By a von Mises’ statistic argument,

$$
P_{c2} = \frac{1}{n} \sum_{j=1}^{n} \int \frac{w(x)}{\phi(x)} K_j^*(x, T_j) K_H(X_j - x) \phi(x) \, dx \sigma(X_j, T_j)e_j + o_p \left( \frac{\log n}{n^2 \Delta_{1}} \right)
$$

which, after the change of variable $X_j = x + Hu$, becomes

$$
\frac{1}{n} \sum_{j=1}^{n} w(x_j) K_j^*(x_j, T_j) \sigma(x_j, T_j)e_j + o_p(n^{-1/2}).
$$

(A.9)

Now let us return to the decomposition of $\frac{1}{n} \sum_{i=1}^{n} w(x_i)(\hat{c}_i - c_i)$ as in (A.6). By Lemmas A.2–A.5, one has

$$
\frac{1}{n} \sum_{i=1}^{n} w(x_i)(\hat{c}_i - c_i) = \frac{1}{n} \sum_{i=1}^{n} w(x_j) \left\{ K_j^*(x_j, T_j) \sigma(x_j, T_j)e_j + \sum_{s=1}^{d_2} \xi_{l's}(X_{js}) \right\} + o_p(n^{-1/2}).
$$

(A.10)

If we define $\tau_j = w(x_j)K_j^*(x_j, T_j)\sigma(x_j, T_j)e_j + \sum_{s=1}^{d_2} \xi_{l's}(X_{js}) = \tau_{j1} + \tau_{j2}$, then by assumption (A4) (a) and the identification condition (2.2), we have $E(\tau_j) = 0$. Together with assumption (A3), $\{\tau_j\}$ is a stationary $\beta$-mixing process, with geometric $\beta$-mixing coefficient. By Minkowski’s inequality and assumptions (A1), (A4), (A5) and (A7), we have, for some $\delta > 0$, $E|\tau_j|^2 + \delta < +\infty$. Next, define

$$
\sigma_j^2 = 2 \sum_{j=1}^{+\infty} \text{cov}(\tau_0, \tau_{j2}) + \text{var}(\tau_0).
$$

(A.11)
A.2.2. Proof of Theorem 2

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

\[
\{Z_s^T W_s(x)Z_s\}^{-1} - \frac{S_s^{-1}(x)}{\phi(x)} = \sum_{v_s=1}^{v_s} Q_{sv}(x) + Q_s(x),
\]

(A.12)

where \( Q_{sv}(x) = \frac{S_s^{-1}(x)}{\phi(x)} \left( I_n + \frac{Z_s^T W_s(x)Z_s S_s^{-1}(x)}{\phi(x)} \right) \), and the matrix \( Q_s(x) \) satisfies

\[
\sup_{x \in B} |Q_s(x)| = o(h_s^{p+2}) \text{ w.p.1.}
\]

Also as in the proof of Theorem 1, by the equation that

\[
e_l(Z_s^T W_s(x_s)Z_s)^{-1}Z_s^T W_s(x_s)Z_s e_l = \delta_{ll'}, \ l = 1, \ldots, d_1 \text{ and } s = 1, \ldots, d_2,
\]

we have the following decomposition:

\[
\frac{1}{n} \sum_{i=1}^{n} w_{-s}(x_{i,-s}) \{ \tilde{\alpha}_{ls}(x_s) - \alpha_{ls}(x_s) \} = \sum_{i=1}^{d_1} \left\{ D_{si}(x_s) + P_{si}(x_s) + \sum_{v=1}^{v_s} R_{vi}(x_s) \right\} + R_1 + R_2,
\]

(A.13)

where

\[
R_1 = \frac{1}{n} \sum_{i=1}^{n} \sum_{s' \neq s} w_{-s}(x_{i,-s}) \sum_{s' \neq s} \alpha_{ls'}(x_{is'}) \}, \quad R_2 = \frac{1}{n} \left\{ \sum_{i=1}^{n} w_{-s}(x_{i,-s}) \left( \hat{c}_l - c_l \right) \right\}
\]

(A.14)

and the other terms in (A.13) are given as follows. By defining an operator

\[
\Phi_s(A, C) = \frac{1}{n} \sum_{i=1}^{n} w_{-s}(x_{i,-s}) \{ e_l^T A Z_s^T W_s(x, x_{i,-s}) C \},
\]

we set for \( i = 1, 2, 3 \), and \( v = 1, \ldots, v_s \),

\[
D_{si}(x_s) = \Phi_s(Q_s(x_s, x_{i,-s}), C_i), \quad P_{si}(x_s) = \Phi_s(S_s^{-1}/(x_s, x_{i,-s}), C_i), \quad R_{vi}(x_s) = \Phi_s(Q_{sv}(x_s, x_{i,-s})^T, C_i),
\]

(A.15)

where \( C_i \) is the vector of errors same as \( B_1 \) in (A.8), and

\[
C_2 = C_2(x_s) = \sum_{l'=1}^{d_1} \sum_{v=0}^{p} \frac{\alpha_{l's}(x_{js}) - \alpha_{l's}(x_{js})}{v!} (x_{js} - x_s)^v T_{jl'},
\]

\[
C_3 = C_3(x_s) = \sum_{l'=1}^{d_1} \sum_{s' \neq s} \alpha_{l's'}(x_{js'}) - \alpha_{l's'}(x_{is'}) T_{jl'},
\]

(A.15)
The proof of Theorem 2 is completed by applying assumption (A6) (b) and the asymptotic results on each term in (A.13), which are presented in the following lemmas:

**Lemma A.6.** As $n \to +\infty$, $\sqrt{n h_s} R_1 = O_p(\sqrt{h_s})$, $\sqrt{n h_s} R_2 = O_p(\sqrt{h_s})$.

**Lemma A.7.** As $n \to +\infty$, $\sup_{x_s \in \text{supp}(w_s)} |D_{s1}(x_s) + D_{s2}(x_s) + D_{s3}(x_s)| = o(h_s^{p+2})$ w.p. 1.

**Lemma A.8.** For any fixed $v=1, \ldots, v_s$, as $n \to +\infty$, $\sup_{x_s \in \text{supp}(w_s)} |R_{v1}(x_s)| + |R_{v2}(x_s)| + |R_{v3}(x_s)| = o(h_s^{p+2})$ w.p. 1.

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

$$P_{s1} = \frac{1}{n} \sum_{j=1}^{n} \frac{w_{-s}(X_j, -s)}{\varphi(x_s, X_j, -s)} \frac{1}{h_s} K_{1s}^{p} \left( \frac{X_{js} - x_s}{h_s}, x_s, X_{j, -s}, T_j \right)$$

$$\times \varphi_{-s}(X_{j, -s}) \sigma(X_j, T_j) e_j + o_p((h_s \log n)^{-1/2}),$$

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

in which

$$\eta_{ls}(x_s) = \frac{1}{(p + 1)!} \sum_{l'=1}^{d_l} 2 l_{s}^{(p+1)}(x_s) \int u^{p+1} E\{w_{-s}(X_{-s}) T_{l'} K_{ls}^{p}(u, x_s, X_{-s}, T)\} du$$

(A.16)

with

$$K_{ls}^{p}(u, x, T) = e_l^T S_{x}^{-1}(x) q^{*}(u, T) k(u), \quad q^{*}(u, T) = (T, uT, \ldots, u^P T)^T. \quad (A.17)$$

Furthermore $\sqrt{n h_s} P_{s1} \xrightarrow{d} N(0, \sigma_{ls}^2(x_s))$, in which

$$\sigma_{ls}^2(x_s) = \int \frac{w_{-s}^2(z_{-s})}{\varphi_{-s}^2(x_s, z_{-s})} K_{ls}^{2p}(u, x_s, z_{-s}, t) \varphi_{-s}^2(z_{-s}) \sigma^2(x_s, z_{-s}, t)$$

$$\times \frac{1}{\sqrt{D}(x_s, z_{-s}, t)} du dz_{-s} dt. \quad (A.18)$$

**Proof of Lemma A.6.** This lemma follows immediately by applying the central limit theorem for strongly mixing process (Bosq, 1998, Theorem 1.7) on $R_1$ and Theorem 1 on $R_2$, respectively. □

**Proof of Lemmas A.7 and A.8.** We have omitted these as they are similar to the proofs of Lemmas A.3, A.4. □
Proof of Lemma A.9. From the definition in (A.15) and using the von Mises’ statistic and change of variable arguments, $P_{s1}$ equals to
\[
\frac{1}{nh_s} \sum_{j=1}^{n} \frac{w_{s}(X_{j,-s})}{\varphi(x_s, X_{j,-s})} K_{h_s}^p \left( \frac{X_{js} - x_s}{h_s}, x_s, X_{j,-s}, T_j \right) \varphi_{-s}(X_{j,-s}) \sigma(X_{j}, T_j) e_j \\
+ o_p\{(nh_s \log n)^{-1/2}\}.
\]
By assumption (A4) (a), 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 $\sqrt{n}h_s P_{s1}$ is asymptotically normal with mean 0 and variance
\[
\int \frac{w_{s}^2(z_{-s})}{\varphi^2(x_s, z_{-s})} K_{h_s}^{*2} \left( u, x_s, z_{-s}, t \right) \varphi_{-s}^2(z_{-s}) \sigma_{-s}^2(x_s, z_{-s}, t) \psi(x_s, z_{-s}, t) \, du \, dz_{-s} \, dt \\
+ o(h_s) = \sigma_{h_s}^2(x_s) + o(h_s)
\]
in which the leading term $\sigma_{h_s}^2(x_s)$ is as defined in (A.18). For the term $P_{s2}(x_s)$, we have
\[
P_{s2}(x_s) = \int \int \frac{w_{s}(X_{-s})}{\varphi(x_s, X_{-s})} \frac{1}{h_s} K_{h_s}^p \left( \frac{z_s - x_s}{h_s}, x_s, X_{-s}, t \right) L_{G_s}(z_{-s} - x_{-s}) \\
\times \left[ \sum_{l'=1}^{d_1} \left\{ \alpha_{l'}(z_s) - \sum_{v=0}^{p} \frac{\alpha_{l'}^{(v)}(x_s)}{v!} (z_s - x_s)^v \right\} I_{l'} \right] \\
\times \psi(z, t) \varphi_{-s}(x_{-s}) \, dz \, dx_{-s} \, dt\{1 + o_p(1)\}.
\]
After the change of variables $z_s = x_s + h_s u$ and $z_{-s} = x_{-s} + G_s v$, it equals to $h_s^{p+1}(x_s) \eta_{l_s}(x_s) + o_p(h_s^{p+1})$, with $\eta_{l_s} (x_s)$ as defined in (A.16). Lastly, the term $P_{s3}$ is seen to be
\[
P_{s3} = \int \int \frac{w_{s}(X_{-s})}{\varphi(x_s, X_{-s})} \frac{1}{h_s} K_{h_s}^p \left( \frac{z_s - x_s}{h_s}, x_s, X_{-s}, t \right) L_{G_s}(z_{-s} - x_{-s}) \\
\times \left[ \sum_{l'=1}^{d_1} \sum_{l''=1}^{d_2} \left\{ \alpha_{l''}(z_s') - \alpha_{l''}(x_s') \right\} I_{l''} \right] \psi(z, t) \varphi_{-s}(x_{-s}) \, dz \, dx_{-s} \, dt\{1 + o_p(1)\}
\]
which after the change of variables, $z_s = x_s + h_s u$ and $z_{-s} = x_{-s} + G_s v$, is seen to be $O_p(q_2^{q_2}) = o_p\{(nh \log n)^{-1/2}\}$, by a Taylor expansion to $q_2$th degree for $x_{l'}$, $x_{l''}$ and $(q_2 - 1)$th degree for $\psi$, 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