Efficient Estimation for Semivarying-Coefficient Models

Yingcun Xia
Department of Zoology, University of Cambridge
ycxia@zoo.cam.ac.uk

Wenyang Zhang
Institute of Mathematics & Statistics, University of Kent at Canterbury
W.Zhang@ukc.ac.uk

Howell Tong
The University of Hong Kong & The London School of Economics & Political Science
htong@hku.hk

Abstract

Motivated by two practical problems, we investigate a semivarying-coefficient model in this paper. We propose a new procedure to estimate the model. Asymptotic properties are established, which show that the bias of the parameter estimator is of order h^3 when a symmetric kernel is used, where h is the used bandwidth, and the variance is of order n^{-1} and efficient in the semiparametric sense. Undersmoothing is unnecessary for the root-n consistency of the parameters. Therefore commonly used bandwidth selection method can be employed. A model selection method is also developed. Simulations are included to demonstrate how the proposed method works. Some insights are obtained for the two motivating problems by using the proposed semivarying-coefficient models.

KEY WORDS: Efficient estimate, local linear, semivarying-coefficient models, strong α-mixing, varying coefficient models.

SHORT TITLE: Semivarying-coefficient models.

1 Introduction

1.1 Preamble

Statistical inference is generally based on some model assumptions, linearity being probably the strictest. While linear models are very well understood, they are often unrealistic. It is well known that treating a linear model as a mis-specified model could lead to large bias. Relaxation of the linearity assumption leads to other kinds of parametric models as well as transformation methods within the parametric approach. Nonparametric modeling makes no assumption on the specification of the model, but may lose some prior information causing an increase in the variance of the estimate. Even worse, if the dimension of the covariates is very high, it would be impossible to fit the data by a standard nonparametric model due to the 'curse of dimensionality'. To avoid the 'curse of dimensionality', many methods have been proposed in the literature, which include Härdle and Stoker (1990), Li (1991), Xia *et al.*(2002) and the references therein. An alternative approach is to relax the conditions imposed on traditional parametric models and explore the hidden structure. Examples include additive models (Breiman and Friedman, 1995; Hastie and Tibshirani, 1990), varying-coefficient models (Hastie and Tibshirani, 1993; Fan and Zhang, 1999), low-dimensional interaction models (Friedman 1991, Gu and Wahba, 1992, Stone *et al.*1997), partially linear models (Wahba 1984; Green and Silverman 1994), and their hybrids (Carroll *et al.*1997, Fan *et al.*1998, Heckman *et al.*(1998)), among others.

Varying-coefficient models are particularly appealing in longitudinal studies where it allows us to examine the extent to which covariates affect responses over time. See Hoover *et al* (1997) and Fan and Zhang (1998) for details on novel applications of varying-coefficient models to longitudinal data. For nonlinear time series applications, see Chen and Tsay (1993) and Xia and Li (1999) where functional-coefficient AR models are proposed and studied. For application in economics, see Durlauf (2001).

An interesting issue remains and relates to the situation when some of the coefficients of the varying-coefficient model are not really varying.

1.2 Two motivating practical problems and one statistical model

Our first practical problem is about the relation of human health with air pollution and weather conditions. It is well known that air pollution is harmful to human health. It is also observed that certain weather conditions can aggravate the situation. See, for example, Smith *et al.* (1999) and Xia *et al.* (2000). It is relevant to study more precisely how weather conditions affect human health, directly or indirectly, via pollutants. More specifically, because patients suffering from

circulatory and respiratory illnesses react differently under different weather conditions, it is relevant to investigate what pollutant levels and what weather conditions would combine to aggravate the human circulatory and respiratory problems to the extent that hospital admission is necessary? In our example, the pollutant data and the weather data are the daily average levels of sulphur dioxide (x_{1t} in μgm^{-3}), nitrogen dioxide (x_{2t} in μgm^{-3}), respirable suspended particulates (x_{3t} in μgm^{-3}), ozone (x_{4t} in μgm^{-3}), relative humidity (U_t in %) and temperature (V_t in ^oC). The data were collected daily in Hong Kong from January 1, 1994 to December 31, 1995. See Figure 1(a). Because additional hospital beds were released to accommodate circulatory and respiratory patients at the beginning of 1995, there is an increasing trend in the number of admissions. We use y_t to denote the daily hospital admissions with the trend removed; see Xia *et al.*(2002) for more details. Now, we consider modelling the time series y_t. Firstly, there is strong autocorrelation in y_t. We can use autoregression (with lag q) to remove the autocorrelation. Secondly, we need to consider the day-of-the-week effect, presumably due to the hospital booking system. We use dummy variables $D_{t\ell}, \ell = 1, 2, \cdots, 6$ to denote the day-of-the-week. $D_{t\ell} = 1$ if day t is the $\ell-$th day-of-the-week; 0 otherwise. Now, one possible approach is to start with a linear approximation for the effects of pollutants on y_t, noting that it is not clear how the weather condition, say humidity (U_t), affects y_t. It might affect y_t directly by $f_0(U_t)$ or indirectly via pollutants as e.g. $f_k(U_t)x_{kt}$, $k = 1, 2, 3, 4$. Taking these factors into account, we may consider the following tentative model

$$y_t = \sum_{k=1}^{q} b_k y_{t-k} + \sum_{\ell=1}^{6} c_k D_{t\ell} + f_0(U_t) + f_1(U_t)x_{1t} + f_2(U_t)x_{2t} + f_3(U_t)x_{3t} + f_4(U_t)x_{4t} + \varepsilon_t. \quad (1.1)$$

In model (1.1), the pollutants affect y_t linearly. However, their intensities (coefficients) depend on the weather condition.

Our second practical problem is about the transmission mechanism of epidemics. The SEIR mechanism describes an epidemic in four discrete states: Susceptible, Exposed, Infectious and Recovered. Infectious individuals spread the disease to the susceptible population. Those in the susceptible population to which the disease is transmitted become exposed and after a period of time, the incubation period, these individuals become infectious. Individuals remain infectious for a period of time, the infectious period, and then these individuals recover. The SEIR mechanism assumes that the recovered individuals are considered to be immune to the disease, such as in the case of measles or whooping cough for all the remaining time.

The SIR mechanism is a simplified SEIR. This mechanism can describe the dynamics of the disease quite well. In this SIR mechanism, the transmission is strictly proportional to the product of the density of susceptibles (S) and the number of infectious hosts (I). It follows that $dI/dt \propto SI$. Liu *et al.* (1987) showed that adding exponents to the S and I, $dI/dt \propto S^\gamma I^\alpha$, can lead to more

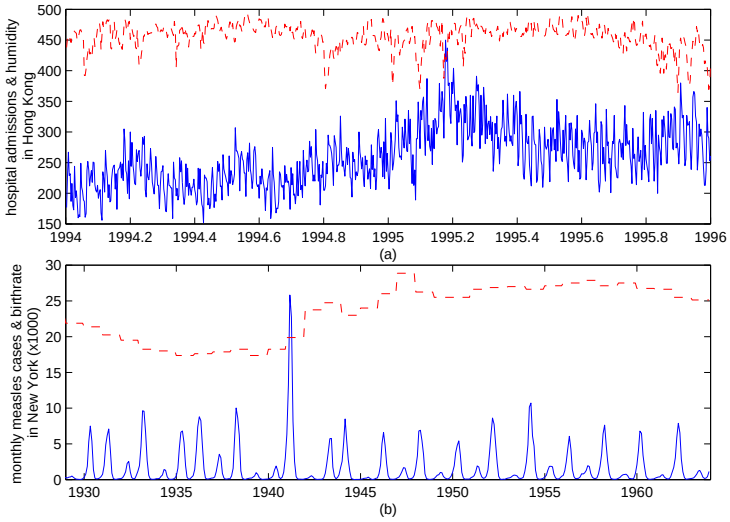

Figure 1: *(a) The lower line is the daily hospital admissions of patients suffering from circulatory and respiratory problems in Hong Kong; the upper line is the daily humidity (re-scaled for easier visualization). (b) The lower line is the measles cases in New York City. The upper line is the birth rate (re-scaled for easier visualization).*

complicated dynamics which can describe the observed data better. The mixing parameters of transmission, α and γ, give a phenomenological model for local spatial variation in the contact rate between the infectious hosts and susceptibles. Following the discrete time series model of Bailey (1975), we have

$$I_t = R_t S_{t-1}^{\gamma} I_{t-1}^{\alpha} \epsilon_t, \quad S_t = S_{t-1} + B_{t-1} - I_{t-1} + u_t, \tag{1.2}$$

where B_t is the births at time t. The basic reproductive rate R_t is essentially the average number of successful offspring that a parasite is intrinsically capable of producing; see Anderson and May (1991, p.17, p.138). In most cases, the basic reproductive rate is a function of the seasonality. The New York data was observed monthly. Let $D_{kt}, k = 1, , \cdots, 12$ be the dummy variables such that $D_{kt} = 1$ if the corresponding month is the kth month of a year and 0 otherwise. Then $R_t = \exp(\sum_{k=1}^{12} c_k D_{kt})$.

Figure 1(b) shows that monthly observed cases of measles in New York city. The dynamics displays biennial cycles after 1945 and 3 or 4-year cycles before 1945. This changes seems to coordinate with the change with the birth rate. With fixed parameters $c_k, k = 1, 2, ..., 12, \alpha$ and γ, it is very hard for model (1.2) to describe the dynamics in New York city. An immediate question is how birth rate changes the mixing parameters, i.e. which parameters will change with the birthrate? If we take logarithm on both side of first equation, we have the following model

$$y_t = \sum_{k=1}^{12} c_k D_{kt} + \gamma(b_{t-1}) \log(S_{t-1}) + \alpha(b_{t-1}) y_{t-1} + \varepsilon_t, \tag{1.3}$$

where $y_t = \log(I_t)$ and $\varepsilon_t = \log(\epsilon_t)$. The statistical task is to determine whether $\gamma(b_{t-1})$, $\alpha(b_{t-1})$ are constants or not.

The above problems suggest a common model

$$y = \sum_{j=1}^{p} a_j(U)x_j + Z^T\boldsymbol{\beta} + \varepsilon, \tag{1.4}$$

$$E(\varepsilon|U,\ X,\ Z) = 0, \quad \text{var}(\varepsilon|U,\ X,\ Z) = \sigma^2(U),$$

where $X = (x_1, \cdots, x_p)^\tau$. This is a semi-varying coefficient regression model, where $\boldsymbol{\beta}$ is a q dimension unknown vector, $a_j(\cdot)$, $j = 1, \cdots, p$, are unknown functions to be estimated.

Note that in model (1.4) if we set $\boldsymbol{\beta} = 0$, (1.4) becomes a standard varying-coefficient model. If $a_j(\cdot) = 0$, $j = 2, \cdots, p$, and $x_1 = 1$, (1.4) becomes a standard semiparametric model. So, (1.4) can be viewed as an extension of a varying-coefficient model as well as a semiparametric model. It would be naive to treat (1.4) as a special case of the varying-coefficient models because the information provided by the semiparametric structure would not be used and, as a result, we would pay an unnecessary price on the variance of the estimate. A sensible way would be to develop a new estimation procedure for (1.4).

Although there is a substantial literature on the estimation procedure for semiparametric models, (see, e.g., Speckman (1988), Wahba (1984), Green and Silverman (1994)), as far as we know, all of the existing methods require undersmoothing in order that the estimate of $\boldsymbol{\beta}$ achieves the convergence rate of $n^{-1/2}$ because the bias of the estimate is typically of the order of h^2.

Based on local linear modelling, (see, e.g., Fan and Gijbels, 1996), we are going to propose a new estimation procedure for (1.4), in which undersmooth is not necessary, the bias of the proposed estimate for $\boldsymbol{\beta}$ is of the order of h^3 and the variance is of the order of n^{-1}. Theoretically, the proposed method answered positively an important question in semiparametrics: can we estimate the parameters and the nonparametric functions of the same model at their optimal consistency rates simultaneously (i.e. using a common bandwidth)? See, e.g. Härdle *et al* (1993) and Carroll *et al.* (1997). With no need of undersmoothing, we can employ the commonly used bandwidth selection method in our estimation procedure even for the estimation of the parameters in practice. As far as we know, there is no convincing bandwidth selection method for the undersmoothing. Furthermore, for semiparametric models, the variance of the proposed estimate is the same as that of the least square estimate for a standard linear model.

The AIC (Akaike, 1973) and the cross-validation (CV) (Stone, 1973) model selection methods are discussed and compared. Based on the discussion and simulations, we suggest to use the CV method. However, the CV method is computationally intensive, especially when the number of

covariates is large. In this paper, we propose a CV-related-method to simplify the calculation procedure. Due to the possible over-fitting as Stone (1977) observed, a method is also suggested to improve the method by putting a slightly heavier penalty on model complexity. The related asymptotic property is established to justify this method.

The paper is structured as follows. The proposed estimation methodology, including the related model selection procedure, is presented in Section 2. Section 3 addresses the asymptotic properties of the proposed methodology. Simulation studies are conducted in Section 4. Section 5 applies the proposed method to the two motivating practical problems. Technical proofs are deferred to the Appendix.

2 Methodology

First, we introduce some notations and basic conditions. Let $\mathbf{1}$ be the n dimension vector with all entries 1, $\mathbf{0}_{q\times p}$ be $q \times p$ matrix with all entries 0. Let $(U_i,\ X_i^\tau,\ Z_i^\tau,\ Y_i)$, $i = 1, \cdots, n$, denote samples from $(U,\ X^\tau,\ Z^\tau,\ y)$.

$$X_i = (x_{i1}, \cdots, x_{ip})^\tau, \quad \mathbf{b}_i = (b_{i1}, \cdots, b_{ip})^\tau, \quad \mathbf{c}_i = (c_{i1}, \cdots, c_{ip})^\tau, \quad i = 1, \cdots, n.$$

$$\mathbf{a}(U) = (a_1(U), \cdots, a_p(U))^\tau, \quad Y = (Y_1, \cdots, Y_n)^\tau, \quad \mathbf{X} = (X_1, \cdots, X_n)^\tau,$$

$$\mathbf{Y} = \mathbf{1} \otimes Y, \quad \mathbf{U} = \mathrm{diag}(U_1, \cdots, U_n), \quad Z = (Z_1, \cdots, Z_n)^\tau, \quad \mathbf{Z} = \mathbf{1} \otimes Z, \quad \nu_k = \int u^k K(u) du,$$

$$W = \mathrm{diag}\Big(W_1, \cdots, W_n\Big), \quad W_i = \mathrm{diag}\Big(K_h(U_1 - U_i),\ \cdots,\ K_h(U_n - U_i)\Big), \quad \mu_i = \int u^i K^2(u) du.$$

Throughout this paper, we always assume $(U_i,\ X_i^\tau,\ Z_i^\tau,\ Y_i)$, $i = 1,\ 2, \cdots$, is a strongly mixing process, and $\sigma^2(u) = \sigma^2$. The proposed methodology and the asymptotic properties still apply for the case where $\sigma^2(u)$ is not constant but at the expense of a more complex notation.

2.1 Motivation

Traditionally, the estimation procedure for $\boldsymbol{\beta}$ goes like this. Using a Taylor's expansion, we have

$$a_j(U) \approx a_j(u) + a_j'(u)(U - u), \quad j = 1, \cdots, p,$$

which leads to the local least squares estimation procedure, namely we minimize

$$\sum_{i=1}^{n} \Big[Y_i - Z_i^\tau \boldsymbol{\beta} - \Big\{\mathbf{b} + \mathbf{c}(U_i - u)\Big\}^\tau X_i\Big]^2 K_h(U_i - u) \tag{2.1}$$

with respect to $(\mathbf{b},\ \mathbf{c})$ to get the minimizer $(\tilde{\mathbf{b}},\ \tilde{\mathbf{c}})$. Let $\tilde{\mathbf{a}}(u)$ be $\tilde{\mathbf{b}}$ and u go over U_j, $j = 1, \cdots, n$. We have

$$\tilde{\mathbf{a}}(U_j) = (I_p,\ \mathbf{0}_{p\times p})\Big\{(\mathbf{X},\ \mathbf{UX} - U_j\mathbf{X})^\tau W_j(\mathbf{X},\ \mathbf{UX} - U_j\mathbf{X})\Big\}^{-1}(\mathbf{X},\ \mathbf{UX} - U_j\mathbf{X})^\tau W_j(Y - Z\boldsymbol{\beta}).$$

Replacing $\mathbf{a}(U_j)$, $j = 1, \cdots, n$, in (1.4) by $\tilde{\mathbf{a}}(U_j)$, we get

$$Y_i = X_i^\tau \tilde{\mathbf{a}}(U_i) + Z_i^\tau \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \cdots, n.$$

Letting

$$\mathbf{H} = \begin{pmatrix} X_1^\tau(I_p,\ \mathbf{0}_{p\times p})\Big\{(\mathbf{X},\ \mathbf{UX} - U_1\mathbf{X})^\tau W_1(\mathbf{X},\ \mathbf{UX} - U_1\mathbf{X})\Big\}^{-1}(\mathbf{X},\ \mathbf{UX} - U_1\mathbf{X})^\tau W_1 \\ \vdots \\ X_n^\tau(I_p,\ \mathbf{0}_{p\times p})\Big\{(\mathbf{X},\ \mathbf{UX} - U_n\mathbf{X})^\tau W_n(\mathbf{X},\ \mathbf{UX} - U_n\mathbf{X})\Big\}^{-1}(\mathbf{X},\ \mathbf{UX} - U_n\mathbf{X})^\tau W_n \end{pmatrix}$$

we have

$$Y = \mathbf{H}(Y - Z\boldsymbol{\beta}) + Z\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.2}$$

where $\boldsymbol{\epsilon} = (\varepsilon_1, \cdots, \varepsilon_n)^\tau$. A least squares estimation procedure on (2.2) yields the estimate

$$\tilde{\boldsymbol{\beta}} = \Big\{Z^\tau(I_n - \mathbf{H})^\tau(I_n - \mathbf{H})Z\Big\}^{-1} Z^\tau(I_n - \mathbf{H})^\tau(I_n - \mathbf{H})Y.$$

As an estimate of $\boldsymbol{\beta}$, $\tilde{\boldsymbol{\beta}}$ has a bias $O_P(h^2)$ and a variance $O_P(n^{-1})$. In the literature which addresses the estimation procedure for models of the above kind, undersmoothing is typically necessary to get the convergence rate of $n^{-1/2}$ for $\tilde{\boldsymbol{\beta}}$. See e.g. Speckman,1988.

Another approach is to minimize (2.1) with respect to $(\mathbf{b},\ \mathbf{c},\ \boldsymbol{\beta})$ to get the minimizer $(\tilde{\mathbf{b}},\ \tilde{\mathbf{c}},\ \tilde{\boldsymbol{\beta}}_r(u))$, and let u go over U_j, $j = 1, \cdots, n$. In this approach, the estimate of $\boldsymbol{\beta}$ is taken to be

$$\tilde{\boldsymbol{\beta}} = n^{-1}\sum_{j=1}^{n} \tilde{\boldsymbol{\beta}}_r(U_j),$$

which still has bias $O_P(h^2)$ and variance $O_P(n^{-1})$. To achieve the convergence rate of $n^{-1/2}$, undersmoothing is again necessary.

Essentially, the above two approaches cannot estimate $\boldsymbol{\beta}$ efficiently as they have bias $O_P(h^2)$. On the other hand, $\boldsymbol{\beta}$ is an unknown vector rather than an unknown function. Therefore, there should be a global approach to estimate it with bias of higher order of $O_P(h^2)$. This thought prompts the following estimation procedure.

2.2 Estimation procedure

Note that $\boldsymbol{\beta}$ is a global constant. It could be estimated more efficiently. Now, we propose to estimate $\boldsymbol{\beta}$ by replacing u in (2.1) by U_j, and taking summation with respect to j. This leads to the following semi-local least squares estimation procedure: minimize

$$n^{-2}\sum_{j=1}^{n}\sum_{i=1}^{n}\Big[Y_i-\Big\{\mathbf{b}_j+\mathbf{c}_j(U_i-U_j)\Big\}^{\tau}X_i-Z_i^{\tau}\boldsymbol{\beta}\Big]^2K_h(U_i-U_j) \tag{2.3}$$

with respect to $\mathbf{b}_j$, $\mathbf{c}_j$, $j=1,\cdots,n$ and $\boldsymbol{\beta}$. Denote the minimizer by $(\hat{\mathbf{b}}_1,\ \hat{\mathbf{c}}_1,\ \cdots,\ \hat{\mathbf{b}}_n,\ \hat{\mathbf{c}}_n,\ \hat{\boldsymbol{\beta}})^{\tau}$. The estimate $\hat{\mathbf{a}}(U_i)$ of $\mathbf{a}(U_i)$ is $\hat{\mathbf{b}}_i$ and the estimate of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}}$, where

$$\hat{\boldsymbol{\beta}}=(\mathbf{0}_{q\times 2pn},\ I_q)(\Omega^{\tau}W\Omega)^{-1}\Omega^{\tau}W\mathbf{Y},$$

$$\Omega=(A,\ \mathbf{Z})\mathrm{diag}\Big(I_n\otimes\mathrm{diag}(I_p,\ h^{-1}I_p),\ I_q\Big),\quad A=I_n\otimes(\mathbf{X},\ \mathbf{UX})-\mathbf{U}\otimes(\mathbf{0}_{n\times p},\ \mathbf{X}).$$

We call the proposed approach a semi-local least squares estimation procedure because the $\mathbf{a}(u)$ is estimated locally, and $\boldsymbol{\beta}$ is estimated globally.

It is intuitively clear that the bias of $\hat{\boldsymbol{\beta}}$ is h^3, because the derivative of (2.3) with respect to $\boldsymbol{\beta}$ should be zero at $(\hat{\mathbf{b}}_1,\ \hat{\mathbf{c}}_1,\ \cdots,\ \hat{\mathbf{b}}_n,\ \hat{\mathbf{c}}_n,\ \hat{\boldsymbol{\beta}})^{\tau}$. That is

$$\sum_{j=1}^{n}\sum_{i=1}^{n}\Big[Y_i-\Big\{\hat{\mathbf{b}}_j+\hat{\mathbf{c}}_j(U_i-U_j)\Big\}^{\tau}X_i-Z_i^{\tau}\hat{\boldsymbol{\beta}}\Big]Z_iK_h(U_i-U_j)=0,$$

which gives us

$$\hat{\boldsymbol{\beta}}=\left\{\sum_{j=1}^{n}\sum_{i=1}^{n}Z_iZ_i^{\tau}K_h(U_i-U_j)\right\}^{-1}\sum_{j=1}^{n}\sum_{i=1}^{n}\Big[Y_i-\Big\{\hat{\mathbf{b}}_j+\hat{\mathbf{c}}_j(U_i-U_j)\Big\}^{\tau}X_i\Big\}\Big]Z_iK_h(U_i-U_j).$$

Notice that there is a bias of the order of h^2 when we approximate $\mathbf{a}(U_i)$ in Y_i by a local linear function. Meanwhile the bias of $\hat{\mathbf{b}}_j$ is also of the order of h^2. If these two terms cancel each other out in the double summation, then the bias is reduced to order h^3. Indeed, we shall show that this is the case in the section on asymptotic properties.

The proposed procedure only gives the estimate of $\mathbf{a}(\cdot)$ at U_j, $j=1,\cdots,n$. In general, $\mathbf{a}(\cdot)$ can be estimated in two steps. First, estimate $\boldsymbol{\beta}$ based on the proposed approach, and substitute $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ in (1.4). Then, (1.4) becomes a standard varying-coefficient model and the method in Fan and Zhang (1999) can be used to get the estimate of $\mathbf{a}(\cdot)$ finally. Specifically, for any u,

$$\hat{\mathbf{a}}(u)=(I_p,\ \mathbf{0}_{p\times p})\Big\{(\mathbf{X},\ \mathbf{UX}-u\mathbf{X})^{\tau}W_0(\mathbf{X},\ \mathbf{UX}-u\mathbf{X})\Big\}^{-1}(\mathbf{X},\ \mathbf{UX}-u\mathbf{X})^{\tau}W_0(Y-Z\hat{\boldsymbol{\beta}}),$$

where $W_0=\mathrm{diag}\Big(K_h(U_1-u),\cdots,K_h(U_n-u)\Big)$. As we shall see, the convergence rate of $\boldsymbol{\beta}$ is of the order of $n^{-1/2}$. Thus, the estimate of $\mathbf{a}(\cdot)$ enjoys all the properties in Fan and Zhang (1999).

2.3 Model selection

There are many model selection methods for parametric models and some special nonparametric models. Amongst them, the AIC-type (such as AIC, BIC etc.) and the cross-validation (CV) model selection methods are the most popular ones. Although the AIC-type was first derived from maximizing the likelihood function (Akaike, 1973), it is found later that it can be extended to more complicated cases that the likelihood function is very difficult to write. The difficulty for the AIC-type methods is the determination of the "degree of freedom". Most of the determinations for semi-parametric models are of somewhat arbitral because the likelihood function is not available. Following similar idea as in Fan *et al.*(2003), we can define a AIC value for our estimation method as follows. For ease of exposition, denote all the covariates here by x with different subscripts, and $P = p + q$. Suppose $\mathcal{S}_\ell$ is any subset of $\{1, 2, \cdots, P\}$, $\ell \leq \sum_{k=0}^{P} C_P^k$. We consider the semi-varying coefficient model

$$y_i = \sum_{k \in \mathcal{S}_\ell} \beta_k x_{i,k} + \sum_{k \notin \mathcal{S}_\ell} a_k(U_i) x_{i,k} + \varepsilon_i. \tag{2.4}$$

Let $RSS(\mathcal{S}_\ell)$ be the value in (2.3) corresponding to the above model. For the minimization in (2.3), the "degree of freedom" is $m_\ell = n_\ell - p_\ell$, where $n_\ell = \mathrm{tr}(W)$ and $p_\ell = \mathrm{tr}\{[\Omega^T W \Omega]^{-1} \Omega^T W^2 \Omega\}$. See Fan *et al.*(2003) for more details. We define AIC, corresponding to the above candidate model, as

$$AIC(\mathcal{S}_\ell) = \log\{RSS(\mathcal{S}_\ell)/m_\ell\} + 2p_\ell/n_\ell.$$

Suppose $AIC(\mathcal{S}_{\ell_0}) = \min_\ell AIC(\mathcal{S}_\ell)$. Then, the model with constant coefficients for $\mathcal{S}_{\ell_0}$ is the model selected by AIC methiod.

Compared with AIC-type methods, the CV method is relatively simple in that it avoids determining the "degree of freedom". The basic idea is easy to understand and free of any settings; see Stone (1977). Now, we give the procedure of CV model selection under our setting. For every i, we first estimate the β_k based on observations $\{(y_j, X_j, U_j), j \neq i\}$. Denote the estimates by $\hat{\beta}_k^{\backslash i}$ or $\hat{a}_k^{\backslash i}(u)$. Then, the CV sum of squares is defined as

$$CV(\mathcal{S}_\ell) = n^{-1} \sum_{i=1}^{n} \Big\{ y_i - \sum_{k \in \mathcal{S}_\ell} \hat{\beta}_k^{\backslash i} x_{i,k} - \sum_{k \notin \mathcal{S}_\ell} \hat{a}_k^{\backslash i}(U_i) x_{i,k} \Big\}^2.$$

Suppose $CV(\mathcal{S}_{\ell_0}) = \min_\ell CV(\mathcal{S}_\ell)$. Then, the model with constant coefficients for $\mathcal{S}_{\ell_0}$ is the 'preferred' model. Besides the advantage for CV methods mentioned above, our numerous simulations, one of them is reported in section 4, also suggest that the CV method works better here.

For linear models, several modifications on the AIC have been made, where the AIC is defined as $n \log(\hat{\sigma}) + c_n p$, where $\hat{\sigma}$ is the mean of the residual squares of the working model and p is number

of covariates, and $c_n = 2$. Most modifications of AIC increase c_n, such as $c_n = \log(n)$ (Schwartz, 1978) and $c_n = c\log\log(n)$ (Hannan and Quinn, 1979). These modifications are helpful for finite sample size simulations as well as the theoretical properties. See Burnham and Anderson (2002) for a more intensive survey.

Stone (1977) observed that the CV method tend to select too many covariates, i.e. over-fitting. A simple explanation for this phenomena is the penalty for the complexity of the model is relatively small. Extensions of the CV model selection method to semiparametric models, such as Cheng and Tong (1991) and Xia *et al.*(2002), have been effective. In their papers, the penalty against overfitting (i.e. dimension d is too high) is proportional to $(nh^d)^{-1}$ (where $h \to 0$ as $n \to \infty$), and is therefore very strong and increases with d exponentially. However, as we will see in the next section, the penalty against too many varying coefficients in the present model increase only algebraically.

Similar to the modification on AIC method, it could be helpful if we can modify the CV method appropriately. We will discuss this problem in section 4. In practice, the number of possible subsets $\mathcal{S}_\ell$ is very large when the number of covariates is large. A simplification of the CV procedure is also necessary and will be given in the same section.

3 Asymptotic properties

Let us introduce some notation first. Let

$$\Lambda(u) = E(X_1X_1^\tau|U=u), \quad \Lambda_1(u) = E(Z_1X_1^\tau|U=u), \quad \Lambda_2(u) = E(Z_1Z_1^\tau|U=u),$$

and $f(u)$ be the density function of U.

First, we study the asymptotic properties of the proposed estimation procedure.

Theorem 1. Under the conditions in Appendix, if $h \longrightarrow 0$, $\log h/(nh^3) \longrightarrow 0$, then

$$\text{bias}(\hat{\boldsymbol{\beta}}|\mathbf{X},\mathbf{U},Z) = O_P(h^3),$$

$$\begin{aligned}\text{var}(\hat{\boldsymbol{\beta}}|\mathbf{X},\mathbf{U},Z) &= \sigma^2 n^{-1}\Big\{E\Big(\Lambda_2(U_1)f(U_1) - \Lambda_1(U_1)\Lambda(U_1)^{-1}\Lambda_1(U_1)^\tau f(U_1)\Big)\Big\}^{-1} \\ &\quad \times E\Big\{\Lambda_2(U_1)f^2(U_1) - \Lambda_1(U_1)\Lambda(U_1)^{-1}\Lambda_1(U_1)^\tau f^2(U_1)\Big\} \\ &\quad \times\Big\{E\Big(\Lambda_2(U_1)f(U_1) - \Lambda_1(U_1)\Lambda(U_1)^{-1}\Lambda_1(U_1)^\tau f(U_1)\Big)\Big\}^{-1}\Big(1+O_P(h)\Big).\end{aligned}$$

Remark 1: *As $\hat{\boldsymbol{\beta}}$ is a linear function of Y, its asymptotic distribution is normal.*

Remark 2: *If U_1 is uniformly distributed, then*

$$\text{var}(\hat{\boldsymbol{\beta}}|\mathbf{X},\mathbf{U},Z) = \sigma^2 n^{-1}(E\mathbf{V})^{-1}\Big(1+O_P(h)\Big),$$

where

$$\mathbf{V} = \Lambda_2(U_1) - \Lambda_1(U_1)\Lambda(U_1)^{-1}\Lambda_1(U_1)^\tau.$$

For semiparametric models, since $\Lambda(U_1) = 1$, *we have* $\mathbf{V} = \mathrm{cov}(Z_1|U_1)$. *Therefore*

$$n\mathrm{var}(\hat{\boldsymbol{\beta}}|\mathbf{X}, \mathbf{U}, Z) = \sigma^2\Big\{E\mathrm{cov}(Z_1|U_1)\Big\}^{-1}\Big(1 + O_P(h)\Big).$$

Furthermore, if U_1 *is independent of* Z_1, *then*

$$n\mathrm{var}(\hat{\boldsymbol{\beta}}|\mathbf{X}, \mathbf{U}, Z) = \sigma^2\Big\{\mathrm{cov}(Z_1)\Big\}^{-1}\Big(1 + O_P(h)\Big).$$

This implies that the proposed estimate is as efficient as the least squares estimate for the standard linear model.

Now, we turn to the asymptotic behavior of the CV method of the proposed model selection procedure.

Because the convergence rate of $\hat{\boldsymbol{\beta}}$ is $n^{-1/2}$ when the bandwidth is taken to be the optimal bandwidth $O(n^{-1/5})$, we can see, from the proof of Theorem 2, that asymptotically the CV sum of squares of the proposed estimation procedure is the same as that of Fan and Zhang's method (1999) for a standard varying-coefficient model, which is the model (1.4) with either known $\boldsymbol{\beta}$ or $q = 0$. So, we only need to present the CV sum of squares of Fan and Zhang's method (1999) for the model (1.4) with $q = 0$.

Theorem 2. Under the conditions in Appendix and $q = 0$ (i.e. $\mathcal{S}_\ell$ is empty), if $n^{1/2}h^3 \longrightarrow 0$, $nh/\log n \longrightarrow \infty$, the CV sum of squares of Fan and Zhang's method (1999) is

$$CV = n^{-1}\boldsymbol{\epsilon}^\tau\boldsymbol{\epsilon} + 4^{-1}\nu_2^2 E\Big(X_1^\tau \mathbf{a}''(U_1)\Big)^2 h^4 + n^{-1}h^{-1}\mu_0\sigma^2 E\Big\{f(U_1)^{-1}X_1^\tau\Lambda(U_1)^{-1}X_1\Big\} + o_P(n^{-1}h^{-1} + h^4),$$

where $\mathbf{a}''(\cdot) = \Big(a_1''(\cdot), \cdots, a_p''(\cdot)\Big)^\tau$.

Remark 1: *In the formula of CV sum of squares in Theorem 2, the first term represents the effect of random error, the second term reflects the bias of the corresponding estimate method, and the third term is the variance of the method. If we treat model (1.4) as a varying-coefficient model and appeal to the method proposed by Fan and Zhang (1999), the resulting CV sum of squares would be*

$$n^{-1}\boldsymbol{\epsilon}^\tau\boldsymbol{\epsilon} + 4^{-1}\nu_2^2 E\Big(X_1^\tau \mathbf{a}''(U_1)\Big)^2 h^4 + n^{-1}h^{-1}\mu_0\sigma^2 E\Big\{f(U_1)^{-1}(X_1^\tau,\ Z_1^\tau)G^{-1}(X_1^\tau,\ Z_1^\tau)^\tau\Big\} + o_P(n^{-1}h^{-1} + h^4),$$

because the second derivative of constant $\boldsymbol{\beta}$ *is zero. Here*

$$G = \begin{pmatrix} \Lambda(U_1) & \Lambda_1(U_1)^\tau \\ \Lambda_1(U_1) & \Lambda_2(U_1) \end{pmatrix}.$$

As expected, there is no difference between the bias terms by treating model (1.4) as a varying-coefficient model or as a semivarying-coefficient model. However, simple calculation gives

$$(X_1^\tau,\ Z_1^\tau)G^{-1}(X_1^\tau,\ Z_1^\tau)^\tau - X_1^\tau\Lambda(U_1)^{-1}X_1 = \left\|\mathbf{V}^{-1/2}\Big\{\Lambda_1(U_1)\Lambda(U_1)^{-1}X_1 - Z_1\Big\}\right\|^2 > 0$$

where $\|\xi\|^2 = \xi^\tau\xi$, which indicates that the variance coming from treating model (1.4) as a varying-coefficient model is larger, and the increment is detectable up to $O(n^{-1}h^{-1})$; this leads to an increment on CV sum of squares which is detectable up to $O(n^{-1}h^{-1})$. This means that a model selection procedure based on CV sum of squares is sensible since $O(n^{-1}h^{-1})$ is the dominant term besides the the random error term $n^{-1}\boldsymbol{\epsilon}^\tau\boldsymbol{\epsilon}$, which does not change with the model. Further more,

Theorem 3. Under the conditions in Appendix, if $n^{1/2}h^3 \longrightarrow 0$, $nh/\log n \longrightarrow \infty$, then

$$\lim_{n\to\infty} P(\mathcal{S}_{\ell_0} = \mathcal{S}_0) = 1$$

where $\mathcal{S}_{\ell_0}$ is the minimizer of $CV(\mathcal{S}_\ell)$ defined in section 2.3, and $\mathcal{S}_0$ is the true model.

Theorem 3 indicates that the proposed model selection procedure is consistent.

4 Algorithm and Simulations

In this section, we illustrate the performance of the proposed methods with some simulated examples. These simulations give us some insights about how the proposed method depends on the sample size n and the signal-to-noise ratio for finite data sets. The models are designed to change with the sample size, so that we can further check the local "power" of our model selection method. In all the calculations, the cross-validation bandwidth and the Epanechnikov kernel are used. We do not have prior knowledge as to which coefficients are varying and which are not. So, all the covariates here are denoted by x with different subscripts. Recall that our model can be written as

$$y = \sum_{j=1}^{P-L} a_j(U)x_j + \sum_{j=P-L+1}^{P} \beta_j x_j + \varepsilon. \tag{4.1}$$

As we mentioned in section 2.2, implementation of the CV model selection procedure is very time-consuming when P is large. In fact, we need to estimate 2^P models with different combinations. If we estimate model (4.3) assuming that all the coefficients are varying, we have by Theorem 2,

$$\max_{1\le\ell\le P-L} \mathrm{std}(\hat{a}_\ell(U)) = O_p\{h^2 + (nh)^{-1/2}\}, \qquad \min_{P-L+1\le\ell\le P} \mathrm{std}(\hat{a}_\ell(U)) = c + o_p(1),$$

where $c > 0$. Thus,

$$\max_{1\le\ell\le P-L} \mathrm{std}(\hat{a}_\ell(U)) < \min_{P-L+1\le\ell\le P} \mathrm{std}(\hat{a}_\ell(U)). \tag{4.2}$$

Inequality (4.2) tells us, constant coefficients always have smaller standard deviation than those of of varying coefficients. Based on this, we propose the following simplified CV model selection procedure.

1. Set $L = 0$, i.e. all the coefficients are varying. For each $i = 1, 2, \cdots, n$, estimate model (4.3) based on $\{(X_\ell, y_\ell), \ell \neq i, \ell = 1, 2, \cdots, n\}$ and obtain estimates $\hat{a}_1^{\backslash i}(u), \hat{a}_2^{\backslash i}(u), ..., \hat{a}_P^{\backslash i}(u)$ using the method in Fan and Zhang (1999). Calculate $v_j = std\{\hat{a}_j^{\backslash i}(u_i),\ i = 1, 2, \cdots, n\}, j = 1, 2, \cdots, n$, and

$$CV(0) = n^{-1} \sum_{i=1}^{n} \{y_i - \sum_{j=1}^{P} \hat{a}_j^{\backslash i}(U) x_{ij}\}^2.$$

2. Look for the minimum value among $\{v_j, j = 1, ..., P-L\}$, say v_{P-L}, and then set $L = L+1$. We estimate model (4.3) using the method in section 2. Obtain estimates $\hat{a}_1(u), \hat{a}_2(u), ..., \hat{a}_{P-L}(u)$, $\hat{\beta}_{P-L+1}, ..., \hat{\beta}_P$ based on all the observations and $\hat{a}_1^{\backslash i}(u), \hat{a}_2^{\backslash i}(u)$, ..., $\hat{a}_{P-L}^{\backslash i}(u)$, $\hat{\beta}_{P-L+1}^{\backslash i}$, ..., $\hat{\beta}_P^{\backslash i}$ based on $\{(X_\ell, y_\ell), \ell \neq i, \ell = 1, 2, \cdots, n\}$ using the proposed method. Calculate $v_j = std\{a_j(u_i),\ i = 1, 2, \cdots, n\}$,

$$SSR(L) = n^{-1} \sum_{i=1}^{n} \{y_i - \sum_{j=1}^{P-L} \hat{a}_j(U) x_{ij} - \sum_{j=P-L+1}^{P} \hat{\beta}_j x_{ij}\}^2,$$

$$CV(L) = n^{-1} \sum_{i=1}^{n} \{y_i - \sum_{j=1}^{P-L} \hat{a}_j^{\backslash i}(U) x_{ij} - \sum_{j=P-L+1}^{P} \hat{\beta}_j^{\backslash i} x_{ij}\}^2.$$

3. If $CV(L) < CV(L-1)$, continue step 2. Otherwise stop. The model with current L is then the selected model.

If the model with the L constant coefficients is the true model, by Theorem 2, we have $CV(L-1) > CV(L)$. However, it is well known that the CV model selection method tends to overfit the true model because of the penalty resulted from the overfitting is not strong. See, e.g., Stone (1977). In our problem, a simple extension of CV method tends to select fewer constant coefficients. To overcome this problem, we use $\{SSR(L) + CV(L)\}/2$ to replace $CV(L)$ in step 3. By (A.2), we have $SSR(L) < CV(L)$ and

$$\{SSR(L) + CV(L)\}/2 < CV(L) < CV(L-1).$$

Thus, our criteria will increase the penalty against overfitting. By replacing $CV(L)$ in step 3 above by $\{SSR(L) + CV(L)\}/2$, we call the corresponding procedure modified CV procedure.

Another important issue for the implementation of kernel smoothing is the choice of bandwidth. As we have pointed out, under-smoothing is unnecessary in our estimation procedure, and any

consistent bandwidth selection methods can be employed. In our calculations, the Epanechnikov kernel and the cross-validation bandwidths are used.

We consider model

$$y = ca_1(U)x_I + a_2(U)x_{II} + a_3(U)x_{III} + 0.2\varepsilon, \tag{4.3}$$

$$y = b_1(U)x_1 + b_1(U)x_2 + cb_3(U)x_3 + \textstyle\sum_{k=4}^{7} b_k(U)x_k + \sigma\varepsilon, \tag{4.4}$$

where $a_0(u) = (u-0.5)^2, b_1(u) = \exp(-32(u-0.5)^2), b_2(u) = \sin(2\pi u)$ and $b_3(u) = \cos(2\pi u)$; $a_2 \equiv 1$, $a_3 \equiv 0.5$, $b_4 \equiv 1$, $b_5 \equiv -1$, $b_6 \equiv 1$, $b_7 \equiv 0$; $U \sim Uniform(0,1)$, $X = (x_1, x_2, \cdots, x_6, x_7)^T$ with $x_1 \equiv 1$ and $x_I, x_{II}, x_{III}, x_2, \cdots, x_6, \varepsilon \overset{IID}{\sim} N(0,1)$.

For the first model, we will make a full comparison among the AIC, CV and modified CV methods. The simulation results are listed in Table 1. For appropriate sample size (n=100), all the methods can pick up the true model with reasonable proportion. However, as n increase, the consistency rates for selecting a true model differs quite a lot. Amongst them, the AIC methods is the slowest; the CV method and our simplified CV have about the same performance; the modified CV method has the fast consistency rate.

TABLE 1: Calculation resuts for model (4.3): frequencies that coefficient a_k is identified as constant and that the model (Md.) is selected correctly out of 100 replications

c	n	AIC				CV				simplified CV				modified CV			
		a_1	a_2	a_3	Md.	a_1	a_2	a_3	Md.	a_1	a_2	a_3	Md.	a_1	a_2	a_3	Md.
1	50	26	69	67	47	24	83	75	61	28	79	75	55	33	80	78	59
	100	5	84	83	69	8	88	90	78	7	88	89	79	8	93	96	90
	200	0	86	88	74	0	90	93	84	0	91	93	84	1	97	99	96
2	50	42	62	60	34	50	74	66	43	50	69	66	39	51	71	66	41
	100	2	89	83	71	1	92	90	82	1	92	92	83	0	97	100	97
	200	2	88	88	74	0	93	97	89	1	94	96	89	0	99	100	99

For more extensive comparison, we consider the second model. The calculations of full CV and AIC methods are extremely heavy for this model. Based on the observations of first model, we now need to compare the performances between the simplified CV and modified. In the simulations, we set $c = 0.3$ and $c = 5/\sqrt{n}$ respectively. In the latter case, the coefficient of x_3 is varying but it tends to a constant at a root-n rate. The results are listed in Tables 2-3 and Figure 2. Table 2 shows that our modified model selection method is effective and better than the simplified CV method, which, according to our simulations as in Table 1, is equivalent to the original CV method. The performance depends on the noise level and the variability of the coefficient. Moreover, if we allow the variability to be proportional to $1/\sqrt{n}$, it seems that the frequency of misspecification tends to a constant. This suggest that our model identification approach has a strong local power for model

selection. Compared with the naive CV selection, the average of $CV(L)$ and $RSS(L)$ can really increase the penalty against overfitting and improve the efficiency of selection. Table 3 shows that our estimation method works satisfactorily. Note that in model (4.4) $E(X|U) = 0$. Therefore, our estimation method should be asymptotically as efficient as the least squares estimation assuming the nonlinear part is known. In Table 3, the efficiency of our estimation tends to that of least squares estimation as n increases. Our simulations are in line with the theoretical results. The varying coefficients are estimated quite stable and accurately as shown in Figure 2.

TABLE 2: Calculation resuts for model (4.4): frequencies that coefficient b_k is identified as constant and that the model (Md.) is selected correctly out of 100 replications

σ	n	$c = 5/\sqrt{n}$								$c = 0.3$							
		b_1	b_2	b_3	b_4	b_5	b_6	b_7	Md.	b_1	b_2	b_3	b_4	b_5	b_6	b_7	Md.
0.5	50	22 [7]*	0 [3]	0 [29]	78 [78]	73 [88]	73 [86]	70 [92]	34 [50]	15 [1]	0 [5]	39 [76]	74 [82]	72 [91]	69 [88]	65 [96]	9 [11]
	100	0 [0]	0 [0]	1 [14]	84 [99]	76 [100]	85 [100]	80 [100]	53 [85]	0 [0]	0 [0]	15 [66]	78 [99]	79 [99]	80 [100]	84 [100]	30 [34]
	200	0 [0]	0 [0]	0 [10]	94 [100]	92 [100]	90 [100]	94 [100]	78 [90]	0 [0]	0 [0]	2 [33]	92 [100]	88 [100]	92 [100]	90 [100]	72 [67]
	400	0 [0]	0 [0]	0 [11]	90 [100]	94 [100]	90 [100]	88 [100]	74 [89]	0 [0]	0 [0]	0 [2]	93 [100]	93 [100]	96 [100]	89 [100]	75 [98]
0.2	50	4 [5]	2 [5]	1 [6]	80 [97]	76 [99]	78 [98]	71 [98]	51 [92]	3 [0]	1 [0]	19 [44]	77 [94]	78 [96]	76 [99]	72 [99]	37 [49]
	100	1 [0]	0 [0]	1 [0]	87 [99]	85 [99]	83 [100]	85 [100]	66 [99]	0 [0]	0 [0]	0 [2]	87 [99]	78 [99]	81 [100]	79 [100]	60 [97]
	200	0 [0]	0 [0]	0 [3]	92 [100]	91 [100]	93 [100]	89 [100]	78 [97]	0 [0]	0 [0]	0 [0]	92 [100]	91 [100]	92 [100]	89 [100]	77 [100]
	400	0 [0]	0 [0]	0 [2]	96 [100]	97 [100]	92 [100]	95 [100]	85 [98]	0 [0]	0 [0]	0 [0]	96 [100]	96 [100]	90 [100]	97 [100]	84 [100]

*the results for the modified CV method are given in the square brackets.

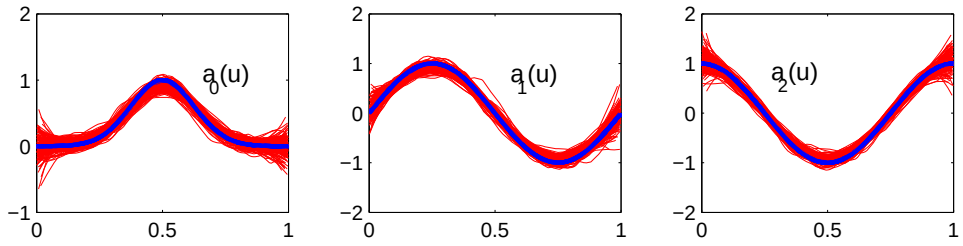

Figure 2: *Simulation results of 200 replications from model (4.4) with* $\sigma = 0.2$, $c = 1$ *and* $n = 100$. *The darkest lines are the true coefficient functions. The others are estimated coefficient functions.*

TABLE 3: Mean and standard deviations (in parentheses) of the estimated parameters for model (4.4) with 200 replications

σ	n	SEV method				LSE method			
		b_4	b_5	b_6	b_7	b_4	b_5	b_6	b_7
0.5	50	0.9924 (0.0841)	-0.9978 (0.0861)	0.9990 (0.0947)	0.0094 (0.0948)	1.0021 (0.0766)	-0.9983 (0.0671)	0.9994 (0.0743)	0.0074 (0.0720)
	100	1.0086 (0.0555)	-1.0047 (0.0634)	1.0033 (0.0593)	-0.0066 (0.0578)	1.0033 (0.0472)	-1.0039 (0.0517)	1.0039 (0.0529)	-0.0020 (0.0514)
	400	0.9986 (0.0245)	-1.0006 (0.0282)	0.9976 (0.0246)	0.0028 (0.0302)	0.9992 (0.0226)	-1.0017 (0.0263)	0.9975 (0.0229)	0.0036 (0.0290)
0.2	50	0.9946 (0.0445)	-0.9936 (0.0500)	0.9847 (0.0432)	0.0020 (0.0478)	0.9998 (0.0314)	-1.0031 (0.0287)	0.9986 (0.0281)	-0.0029 (0.0292)
	100	1.0013 (0.0232)	-0.9992 (0.0248)	0.9999 (0.0231)	-0.0020 (0.0255)	1.0003 (0.0208)	-1.0001 (0.0220)	1.0001 (0.0203)	-0.0003 (0.0219)
	400	0.9989 (0.0102)	-1.0000 (0.0096)	1.0005 (0.0105)	-0.0013 (0.0092)	0.9994 (0.0097)	-1.0003 (0.0095)	1.0008 (0.0096)	-0.0006 (0.0086)

5 Real data analysis

Now, we return to our practical problems in section 1. Using model (1.1) with $q = 6$, results of our model selection procedure are shown in table 4. Based on this table, the selected model is a semi-parametric model with either two varying coefficients for x_2 and x_3 or with one varying coefficients for x_2.

After we have determined the varying coefficients, we then use the standard method to select the explanatory variables and delete those with absolute t-values less than 2. Our final model (with varying coefficients for x_2, x_3) is

$$\begin{aligned} y_t \;=\; & \underset{(0.0372)}{-0.2333} + \underset{(0.0361)}{0.2652}y_{t-1} + \underset{(0.0370)}{0.1150}y_{t-2} + \underset{(0.0352)}{0.1318}y_{t-3} - \underset{(0.0606)}{0.3041}D_{t2} - \underset{(0.0646)}{0.1439}D_{t3} \\ & + \underset{(0.0673)}{1.0569}D_{t4} + \underset{(0.0749)}{0.4421}D_{t5} + \underset{(0.0694)}{0.6008}D_{t6} + a_2(u_t)x_{t2} + a_3(u_t)x_{t3}. \end{aligned}$$

The estimated coefficient functions are shown in Figure 3 (using bandwidth $h = 0.15$).

TABLE 4: Estimation procedure for the health problem in Hong Kong

nonlinear parts	corresponding std. of est. functions	CV-value	SSR
a_0, a_1, a_2, a_3, a_4	0.0204, 0.0504, 0.1206, 0.1184, 0.0708	0.2476	0.2317
a_1, a_2, a_3, a_4	0.0446, 0.1144, 0.1234, 0.0641	0.2447	0.2327
a_2, a_3, a_4	0.1024, 0.1253, 0.0705	0.2441	0.2349
a_2, a_3	0.1068, 0.1008	0.2439	0.2377
a_2	0.1036	0.2439	0.2413
$--$	$--$	0.2541	0.2431

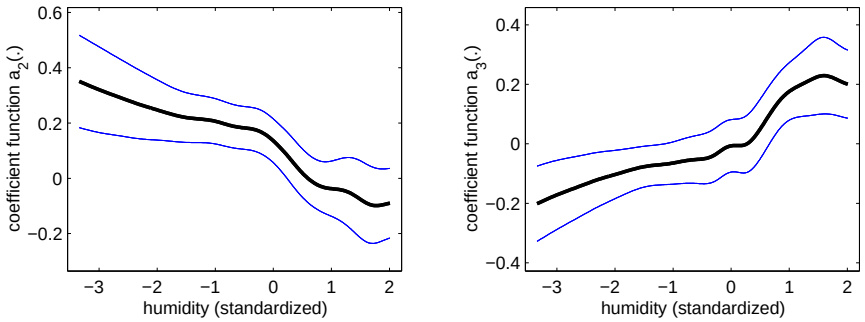

Figure 3: *The central lines in the two panels are the estimated coefficient functions $a_2(.)$ and $a_3(.)$ respectively. The upper and lower lines are the corresponding symmetric confidence intervals.*

Based on the above analysis, we may draw the following conclusions about the function of humidity. (1) The effect of humidity on the circulatory and respiratory problems is indirect via nitrogen dioxide (x_{2t}) and suspended particulates (x_{3t}); (2) As the humidity increases (i.e. the weather becomes wetter), the effect of nitrogen dioxide (x_{2t}) on the health problem reduces. A possible reason is that the increased humidity dilutes the impact of nitrogen dioxide on the circulatory and respiratory system. (3) As the humidity increases, the adverse effect of suspended particulates (x_{3t}) on human health increases. Biologically, as the humidity increases, the fungal spores and dust mite level also increase, which deteriorates the health.

Now, we consider the second practical problem. Note that the numbers of susceptible are not available. Write $S_t = \bar{S} + z_t$. We can reconstruct these variables by regression $z_t = c_0 + \sum_{k=1}^{t} b_t - \rho \sum_{k=1}^{t} I_t$, where ρ is the best fit as if z_t is the residual term. See, for example, Ellner *et al* (1998) and Finkenstädt and Grenfell (2000). Taking logarithm on both sides of the model of the first equation of (1.2) and using the Taylor expansion $\log(\bar{S} + z_t) \approx \log(\bar{S}) + z_t/\bar{S}$, we have the approximation

$$y_t = \sum_{k=1}^{12} a_k D_{tk} + a_{13}(b_t) z_t + a_{14}(b_t) y_{t-1} + \varepsilon_t, \tag{5.1}$$

where $y_t = \log(I_t)$, $a_{13} \approx \gamma(b_t)/\bar{S}_0$ and $a_{14}(b_t) = \alpha(b_t)$. The term $\log(\bar{S})$ is absorbed to the first 12 terms. Our model selection procedure is shown in table 5. The table suggests that there is only one varying coefficient for y_{t-1} as shown in the first panel of Figure 4 (with bandwidth $h = 0.01$). The fitted model is

$$\begin{aligned} y_t \;=\; & \underset{(0.0326)}{0.3029D_{1t}} + \underset{(0.0325)}{0.3838D_{2t}} + \underset{(0.0322)}{0.2511D_{3t}} + \underset{(0.0322)}{0.1935D_{4t}} - \underset{(0.0325)}{0.0163D_{5t}} - \underset{(0.0327)}{0.4951D_{6t}} - \underset{(0.0325)}{0.7445D_{7t}} \\ & \underset{(0.0323)}{-0.5924D_{8t}} - \underset{(0.0322)}{0.0278D_{9t}} + \underset{(0.0321)}{0.1749D_{10t}} + \underset{(0.0323)}{0.2620D_{11t}} + \underset{(0.0329)}{0.3083D_{12t}} + \underset{(0.0099)}{0.1438z_t} + a_{14}(b_t)y_{t-1}. \end{aligned}$$

TABLE 5: Estimation procedure for the measles dynamics in New York city

nonlinear parts	corresponding std. of est. functions	CV-value	SSR
a_{13}, a_{14}	0.0080, 0.0108	0.0383	0.0373
a_{13}	0.0105	0.0379	0.0374
$--$	–	0.0401	0.0376

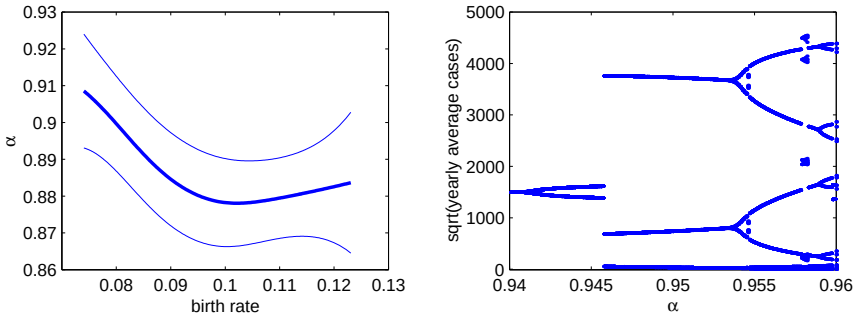

Figure 4: *In the first panel, the central line is the estimate of varying coefficient $a_{14}(b_t)$: the upper and lower lines are the 90% symmetric confidence interval. The second panel is a bifurcation analysis of the dynamics with α as the control variable.*

Based on the above analysis, we may draw the following conclusions about the function of birthrate on the dynamics of measles epidemics. (1) As the birthrate decreases, the mixing parameter α increases. When the birth rate is low (over a couple of years), the proportion of susceptibles for the school-age group is expected to be relatively high compared with the pre-school age group. Since the transmission rate in the school-age group tends to be high, therefore α is high. Similarly, when the birth rate is high over a couple of years, the proportion of susceptibles in the pre-school age group is expected to be high. Since the transmission rate in the pre-school group tends to be low, therefore α is low. Figure 4(a) lends some support to the above description. (2) To further study the impact of α, we return to equation (1.2) and consider the discrete dynamical system:

$y_t = 50\exp(\sum_{k=1}^{12}\hat{c}_k D_{tk})S_{t-1}y_{t-1}^{\alpha}$, $S_t = S_{t-1} + B - y_{t-1}$ with $B = 1500$. A bifurcation analysis is shown in the second panel of Figure 4, which suggests that as α increases, the dynamics bifurcates from a one-year cycle to a two-or-more-year cycle. This is consistent with the observations in Figure 1(b).

Acknowledgments:

We thank the referee for very helpful comments. Our thanks also go to the Research Grants Council of Hong Kong, the Hong Kong Friends of the London School of Economics and the Wellcome Trust for their support.

Appendix:

We impose the following technical conditions:

(1) $EX_{1j}^{2s} < \infty$, for some $s > 2$, $j = 1, \cdots, p$.

(2) $a_j'''(\cdot)$, $j = 1, \cdots, p$, is continuous.

(3) $f(u)$ has a continuous second derivative and a compact support set D, and $f(u) \neq 0$ on its support set.

(4) Every element in $\Lambda(u)$, $\Lambda_1(u)$ and $\Lambda_2(u)$ has third derivative; matrix $\Lambda(u)$ is invertible on $u \in D$.

(5) The conditional density of U_1 given $(X_1^\tau,\ Z_1^\tau)$ exists and is bounded. Furthermore, the conditional density of $(U_1,\ U_{1+l})$ given $(X_1^\tau,\ Z_1^\tau,\ X_{1+l}^\tau,\ Z_{1+l}^\tau)$, for all $l \geq 1$, exists and is bounded as well.

(6) For all j with $j \leq 3$, $u^j K(u)$ is a bounded positive function with a bounded support, satisfying a Lipschitz condition. $K(u)$ is symmetric.

(7) $(U_i,\ X_i^\tau,\ Z_i^\tau,\ Y_i)$, $i = 1,\ 2,\ \cdots$, is a strongly mixing process with the strong mixing coefficient $\alpha(k)$ satisfying

$$\sum_{n=1}^{\infty} \psi(n) < \infty, \tag{A.1}$$

where

$$\psi(n) = \left\{r(n)h^{1/4}\log^{1/4} n\right\}^{-1} nL_1(n)(n\eta_n^2)^{1/4}\alpha(r_1(n)), \quad r(n) = (\eta_n \log^{1/2} n)^{-1}(nh)^{1/2},$$

$$L_1(n) = (h^3 \log n)^{-1/2}(n\eta_n^2)^{1/2}, \quad \eta_n = \Big\{n \log n (\log\log n)^{1+\delta}\Big\}^s,$$

for some $0 < \delta < 1$.

(8) $n^{1-2/s} h \Big\{ \log^{1+2/s} n (\log\log n)^{2(1+\delta)/s} \Big\}^{-1} \longrightarrow \infty$ when $n \longrightarrow \infty$.

As the uniform convergence is related in the proof of the theorem, the following Lemma is needed.

Lemma 1: Let $(X_1,\ Y_1),\ (X_2,\ Y_2),\ \cdots$, be a strongly mixing process with the mixing coefficient $\alpha(k)$ satisfying (A.1), where the Y_i's are scalar random variables and $E|Y_1|^s < \infty$ for some $s > 2$. Assume further that the conditional densities of X_1 given Y_1 and the conditional density of $(X_1,\ X_{l+1})$ given $(Y_1,\ Y_{l+1})$, for all $l \geq 1$, exist and are bounded, and the kernel $K(\cdot)$ satisfies the above condition (6), and the bandwidth h satisfies the above condition (7). Then

$$\sup_{x \in D} |n^{-1} \sum_{i=1}^{n} \{K_h(X_i - x)Y_i - E[K_h(X_i - x)Y_i]\}| = O_P[\{nh/\log(n)\}^{-1/2}]$$

provided $h \longrightarrow 0$.

Proof : This follows immediately from Theorem 5 of Masry (1996).

The proof of Theorem 1:

Using Lemma 1, we have

$$\begin{aligned}
& \text{bias}(\hat{\boldsymbol{\beta}}|\mathbf{X}, \mathbf{U}, Z) \\
= \ & 2^{-1}(\mathbf{0}_{q\times 2pn},\ I_q)(\Omega^\tau W \Omega)^{-1}\Omega^\tau W \Big(\mathbf{1} \otimes (\mathbf{a}''(U_1)^\tau X_1, \cdots, \mathbf{a}''(U_n)^\tau X_n)^\tau\Big) h^2 + O_P(h^3) \\
= \ & 2^{-1} n (\mathbf{0}_{q\times 2pn},\ I_q)(\Omega^\tau W \Omega)^{-1} \times \\
& \Big(\mathbf{a}''(U_1)^\tau \Lambda(U_1) f(U_1),\ \mathbf{0}_{1\times p},\ \cdots,\ \mathbf{a}''(U_n)^\tau \Lambda(U_n) f(U_n),\ \mathbf{0}_{1\times p},\ nE(\mathbf{a}''(U_1)^\tau \Lambda_1(U_1)^\tau f(U_1))\Big)^\tau h^2 \\
& + O_P(h^3).
\end{aligned}$$

Let

$$D_{11} = \text{diag}\Big(\text{diag}(1,\ \nu_2) \otimes \Lambda(U_1) f(U_1),\ \cdots,\ \text{diag}(1,\ \nu_2) \otimes \Lambda(U_n) f(U_n)\Big),$$

$$D_{12} = \Big(\Lambda_1(U_1) f(U_1),\ \mathbf{0}_{q\times p},\ \cdots,\ \Lambda_1(U_n) f(U_n),\ \mathbf{0}_{q\times p}\Big)^\tau, \quad D_{22} = nE\Big(\Lambda_2(U_1) f(U_1)\Big),$$

By Lemma 1,

$$\begin{aligned}
& (\Omega^\tau W \Omega)^{-1} \\
= \ & n^{-1} \begin{pmatrix} D_{11}^{-1} + D_{11}^{-1} D_{12} (D_{22} - D_{12}^\tau D_{11}^{-1} D_{12})^{-1} D_{12}^\tau D_{11}^{-1} & -D_{11}^{-1} D_{12} (D_{22} - D_{12}^\tau D_{11}^{-1} D_{12})^{-1} \\ -(D_{22} - D_{12}^\tau D_{11}^{-1} D_{12})^{-1} D_{12}^\tau D_{11}^{-1} & (D_{22} - D_{12}^\tau D_{11}^{-1} D_{12})^{-1} \end{pmatrix} \\
& \times \Big(1 + O_P(h)\Big).
\end{aligned}$$

So

$$
\begin{aligned}
& \mathrm{bias}(\hat{\boldsymbol{\beta}}|\mathbf{X},\mathbf{U},Z) \\
= \;& 2^{-1}(D_{22}-D_{12}^{\tau}D_{11}^{-1}D_{12})^{-1}\Big\{nE\Big(\Lambda_1(U_1)\mathbf{a}''(U_1)f(U_1)\Big) \\
& -D_{12}^{\tau}D_{11}^{-1}\Big(\mathbf{a}''(U_1)^{\tau}\Lambda(U_1)f(U_1),\ \mathbf{0}_{1\times p},\ \cdots,\ \mathbf{a}''(U_n)^{\tau}\Lambda(U_n)f(U_n),\ \mathbf{0}_{1\times p}\Big)^{\tau}\Big\}h^2+O_P(h^3) \\
= \;& 2^{-1}\Big\{E\Big(\Lambda_2(U_1)f(U_1)\Big)-E\Big(\Lambda_1(U_1)\Lambda(U_1)^{-1}\Lambda_1(U_1)f(U_1)\Big)\Big\}^{-1} \\
& \times\left\{E\Big(\Lambda_1(U_1)\mathbf{a}''(U_1)f(U_1)\Big)-n^{-1}\sum_{i=1}^{n}\Lambda_1(U_i)\Lambda(U_i)^{-1}\Lambda(U_i)\mathbf{a}''(U_i)f(U_i)\right\}h^2+O_P(h^3) \\
= \;& O_P(h^3).
\end{aligned}
$$

It is easy to see that

$$
\mathrm{var}(\hat{\boldsymbol{\beta}}|Y,\mathbf{X},\mathbf{U})=(\mathbf{0}_{q\times 2pn},\ I_q)(\Omega^{\tau}W\Omega)^{-1}\Omega^{\tau}W\Big((\mathbf{1}\mathbf{1}^{\tau})\otimes I_n\Big)W\Omega(\Omega^{\tau}W\Omega)^{-1}(\mathbf{0}_{q\times 2pn},\ I_q)^{\tau}\sigma^2.
$$

Because

$$
\begin{aligned}
& \Omega^{\tau}W\Big((\mathbf{1}\mathbf{1}^{\tau})\otimes I_n\Big)W\Omega=\Big\{(\mathbf{1}^{\tau}\otimes I_n)W\Omega\Big\}^{\tau}(\mathbf{1}^{\tau}\otimes I_n)W\Omega \\
= \;& \left(W_1\mathbf{X},\ (W_1\mathbf{U}\mathbf{X}-U_1W_1\mathbf{X})h^{-1},\ \cdots,\ W_n\mathbf{X},\ (W_n\mathbf{U}\mathbf{X}-U_nW_n\mathbf{X})h^{-1},\ \sum_{i=1}^{n}W_i\mathbf{Z}\right)^{\tau} \\
& \times\left(W_1\mathbf{X},\ (W_1\mathbf{U}\mathbf{X}-U_1W_1\mathbf{X})h^{-1},\ \cdots,\ W_n\mathbf{X},\ (W_n\mathbf{U}\mathbf{X}-U_nW_n\mathbf{X})h^{-1},\ \sum_{i=1}^{n}W_i\mathbf{Z}\right) \\
= \;& n\begin{pmatrix}\tilde{D}_{11} & \tilde{D}_{12}\\ \tilde{D}_{12}^{\tau} & \tilde{D}_{22}\end{pmatrix}\Big(1+O_P(h)\Big),
\end{aligned}
$$

where

$$
\begin{aligned}
\tilde{D}_{11} \;=\; & n^{-1}\Big(W_1\mathbf{X},\ (W_1\mathbf{U}\mathbf{X}-U_1W_1\mathbf{X})h^{-1},\ \cdots,\ W_n\mathbf{X},\ (W_n\mathbf{U}\mathbf{X}-U_nW_n\mathbf{X})h^{-1}\Big)^{\tau} \\
& \times\Big(W_1\mathbf{X},\ (W_1\mathbf{U}\mathbf{X}-U_1W_1\mathbf{X})h^{-1},\ \cdots,\ W_n\mathbf{X},\ (W_n\mathbf{U}\mathbf{X}-U_nW_n\mathbf{X})h^{-1}\Big)
\end{aligned}
$$

$$
\tilde{D}_{12}=n\Big(\Lambda_1(U_1)f^2(U_1),\ \mathbf{0}_{q\times p},\ \cdots,\ \Lambda_1(U_n)f^2(U_n),\ \mathbf{0}_{q\times p}\Big)^{\tau},\quad \tilde{D}_{22}=n^2E\Big(\Lambda_2(U_1)f^2(U_1)\Big),
$$

we have

$$
\begin{aligned}
& \mathrm{var}(\hat{\boldsymbol{\beta}}|Y,\mathbf{X},\mathbf{U}) \\
= \;& \sigma^2n^{-1}(D_{22}-D_{12}^{\tau}D_{11}^{-1}D_{12})^{-1}\Big(\mathbf{0}_{q\times(p+q)},\ \cdots,\ \mathbf{0}_{q\times(p+q)}, \\
& \qquad n^2E(\Lambda_2(U_1)f^2(U_1))-n^2E(\Lambda_1(U_1)\Lambda(U_1)^{-1}\Lambda_1(U_1)^{\tau}f^2(U_1))\Big)(-D_{12}^{\tau}D_{11}^{-1},\ I_q)^{\tau} \\
& \times(D_{22}-D_{12}^{\tau}D_{11}^{-1}D_{12})^{-1}\Big(1+O_P(h)\Big) \\
= \;& \sigma^2n^{-1}\Big\{E\Big(\Lambda_2(U_1)f(U_1)-\Lambda_1(U_1)\Lambda(U_1)^{-1}\Lambda_1(U_1)^{\tau}f(U_1)\Big)\Big\}^{-1}
\end{aligned}
$$

$$\times E\Big\{\Lambda_2(U_1)f^2(U_1) - \Lambda_1(U_1)\Lambda(U_1)^{-1}\Lambda_1(U_1)^\tau f^2(U_1)\Big\}$$
$$\times\Big\{E\Big(\Lambda_2(U_1)f(U_1) - \Lambda_1(U_1)\Lambda(U_1)^{-1}\Lambda_1(U_1)^\tau f(U_1)\Big)\Big\}^{-1}\Big(1+O_P(h)\Big).$$

The proof of Theorem 2:

When $q=0$, the estimate of $\hat{\mathbf{a}}(U_i)$, $i=1,\cdots,n$, based on Fan and Zhang's estimation procedure is

$$\hat{\mathbf{a}}(U_i) = (I_p,\ \mathbf{0}_{p\times p})\Big\{(\mathbf{X},\ \mathbf{UX}-U_i\mathbf{X})^\tau W_i(\mathbf{X},\ \mathbf{UX}-U_i\mathbf{X})\Big\}^{-1}(\mathbf{X},\ \mathbf{UX}-U_i\mathbf{X})^\tau W_i Y.$$

Let $\mathbf{X}^{\backslash i}$ be $\mathbf{X}$ deleted the ith row, $\mathbf{U}^{\backslash i}$ be $\mathrm{diag}(U_1,\ \cdots,\ U_{i-1},\ U_{i+1},\ \cdots,\ U_n)$,

$$W^{\backslash i} = \mathrm{diag}\Big(K_h(U_1-U_i),\ \cdots,\ K_h(U_{i-1}-U_i),\ K_h(U_{i+1}-U_i),\ \cdots,\ K_h(U_n-U_i)\Big)$$

$\hat{\mathbf{a}}^{\backslash i}$ be the estimate of $\mathbf{a}(U_i)$ based on $(U_j,\ X_j,\ Y_j)$, $j=1,\cdots,i-1,\ i+1,\cdots,n$. By simple calculation, we have

$$\hat{\mathbf{a}}^{\backslash i} - \hat{\mathbf{a}}(U_i) = (nh)^{-1}K(0)(I_p,\ \mathbf{0}_{p\times p})A_i^{-1}(X_i^\tau, \mathbf{0}_{1\times p})^\tau\Big(X_i^\tau\hat{\mathbf{a}}(U_i) - Y_i\Big) + O_P(n^{-2}h^{-2}\log n)$$

holds uniformly, where

$$A_i = n^{-1}\mathrm{diag}(I_p,\ h^{-1}I_p)(\mathbf{X},\ \mathbf{UX}-U_i\mathbf{X})^\tau W_i(\mathbf{X},\ \mathbf{UX}-U_i\mathbf{X})\mathrm{diag}(I_p,\ h^{-1}I_p).$$

This leads

$$CV = n^{-1}\sum_{i=1}^n (Y_i - X_i^\tau\hat{\mathbf{a}}^{\backslash i})^2 = RSS + \Delta + O_P(n^{-2}h^{-2}\log n), \tag{A.2}$$

where

$$RSS = n^{-1}\sum_{i=1}^n\Big(Y_i - X_i^\tau\hat{\mathbf{a}}(U_i)\Big)^2,$$

$$\Delta = 2(nh)^{-1}K(0)n^{-1}\sum_{i=1}^n(X_i^\tau,\mathbf{0}_{1\times p})A_i^{-1}(X_i^\tau,\mathbf{0}_{1\times p})^\tau\Big(Y_i - X_i^\tau\hat{\mathbf{a}}(U_i)\Big)^2.$$

By Theorem 2 in Fan and Zhang (2000), we have $\hat{\mathbf{a}}(u)-\mathbf{a}(u) = o_P(1)$ uniformly holds. Furthermore, using Lemma 1, we get

$$\Delta = 2(nh)^{-1}K(0)\sigma^2E\Big(f(U_1)^{-1}X_1^\tau\Lambda(U_1)^{-1}X_1\Big) + o_P(n^{-1}h^{-1}), \tag{A.3}$$

Letting

$$\mathbf{m} = (m_1,\cdots,m_n)^\tau,\quad m_i = X_i^\tau\mathbf{a}(U_i),$$

we have

$$\begin{aligned} RSS &= n^{-1}Y^\tau(I_n-\mathbf{H})^\tau(I_n-\mathbf{H})Y \\ &= n^{-1}\mathbf{m}^\tau(I_n-\mathbf{H})^\tau(I_n-\mathbf{H})\mathbf{m} + n^{-1}\boldsymbol{\epsilon}^\tau(I_n-\mathbf{H})^\tau(I_n-\mathbf{H})\boldsymbol{\epsilon} \\ &\quad +2n^{-1}\mathbf{m}^\tau(I_n-\mathbf{H})^\tau(I_n-\mathbf{H})\boldsymbol{\epsilon}. \end{aligned} \tag{A.4}$$

By a standard argument, we have

$$n^{-1}\mathbf{m}^\tau(I_n-\mathbf{H})^\tau(I_n-\mathbf{H})\mathbf{m}=4^{-1}\nu_2^2E\Big(X_1^\tau\mathbf{a}''(U_1)\Big)^2h^4+o_P(h^4). \tag{A.5}$$

Now, turning to the second term in (A.4), we have

$$n^{-1}\boldsymbol{\epsilon}^\tau(I_n-\mathbf{H})^\tau(I_n-\mathbf{H})\boldsymbol{\epsilon}=n^{-1}\boldsymbol{\epsilon}^\tau\boldsymbol{\epsilon}-2n^{-1}\boldsymbol{\epsilon}^\tau\mathbf{H}^\tau\boldsymbol{\epsilon}+n^{-1}\boldsymbol{\epsilon}^\tau\mathbf{H}^\tau\mathbf{H}\boldsymbol{\epsilon}. \tag{A.6}$$

Using the technique in Härdle and Marron (1985) and Lemma 1, we have

$$2n^{-1}\boldsymbol{\epsilon}^\tau\mathbf{H}^\tau\boldsymbol{\epsilon}=2n^{-2}\sum_{j=1}^n\varepsilon_jX_j^\tau(I_p,\ \mathbf{0}_{p\times p})A_j^{-1}\mathrm{diag}(I_p,\ h^{-1}I_p)(\mathbf{X},\ \mathbf{UX}-U_j\mathbf{X})^\tau W_j\boldsymbol{\epsilon}=R_n+o_P(R_n),$$

where

$$R_n=2n^{-2}\sum_{j=1}^n\varepsilon_jX_j^\tau\Lambda(U_j)^{-1}f(U_j)^{-1}\sum_{i=1}^nX_iK_h(U_i-U_j)\varepsilon_i.$$

Obviously,

$$\begin{aligned}R_n&=n^{-2}\sum_{j=1}^n\sum_{i=1}^n\varepsilon_j\varepsilon_iX_j^\tau\Big(\Lambda(U_j)^{-1}f(U_j)^{-1}+\Lambda(U_i)^{-1}f(U_i)^{-1}\Big)X_iK_h(U_i-U_j)\\&=2n^{-2}K(0)h^{-1}\sum_{j=1}^n\varepsilon_j^2f(U_j)^{-1}X_j^\tau\Lambda(U_j)^{-1}X_j+Q_n\\&=2(nh)^{-1}K(0)\sigma^2E\Big(f(U_1)^{-1}X_1^\tau\Lambda(U_1)^{-1}X_1\Big)+Q_n+o_P(n^{-1}h^{-1}),\end{aligned}$$

where

$$Q_n=n^{-2}\sum_{1\le i<j\le n}\varepsilon_j\varepsilon_iX_j^\tau\Big(\Lambda(U_j)^{-1}f(U_j)^{-1}+\Lambda(U_i)^{-1}f(U_i)^{-1}\Big)X_iK_h(U_i-U_j).$$

It is not difficult to see that

$$E(Q_n)=0,\quad \mathrm{var}(Q_n)=O(n^{-2}h^{-1}).$$

This gives $Q_n=O_P(n^{-1}h^{-1/2})$, which leads to

$$2n^{-1}\boldsymbol{\epsilon}^\tau\mathbf{H}^\tau\boldsymbol{\epsilon}=2(nh)^{-1}K(0)\sigma^2E\Big(f(U_1)^{-1}X_1^\tau\Lambda(U_1)^{-1}X_1\Big)+o_P(n^{-1}h^{-1}). \tag{A.7}$$

$$n^{-1}\boldsymbol{\epsilon}^\tau\mathbf{H}^\tau\mathbf{H}\boldsymbol{\epsilon}=n^{-3}\sum_{j=1}^n\Big\{X_j^\tau(I_p,\ \mathbf{0}_{p\times p})A_j^{-1}\mathrm{diag}(I_p,\ h^{-1}I_p)(\mathbf{X},\ \mathbf{UX}-U_j\mathbf{X})^\tau W_j\boldsymbol{\epsilon}\Big\}^2=T_n+o_P(T_n),$$

where

$$T_n=n^{-3}\sum_{k=1}^n\sum_{i=1}^n\sum_{j=1}^nf(U_k)^{-2}X_j^\tau\Lambda(U_k)^{-1}X_kX_k^\tau\Lambda(U_k)^{-1}X_iK_h(U_i-U_k)K_h(U_j-U_k)\varepsilon_i\varepsilon_j.$$

Obviously,

$$\begin{aligned} T_n &= n^{-3}\sum_{k=1}^{n}\sum_{i=1}^{n} f(U_k)^{-2}X_i^\tau\Lambda(U_k)^{-1}X_kX_k^\tau\Lambda(U_k)^{-1}X_iK_h^2(U_i-U_k)\varepsilon_i^2 \\ &\quad +2n^{-3}\sum_{k=1}^{n}\sum_{1\le i<j\le n} f(U_k)^{-2}X_j^\tau\Lambda(U_k)^{-1}X_kX_k^\tau\Lambda(U_k)^{-1}X_iK_h(U_i-U_k)K_h(U_j-U_k)\varepsilon_i\varepsilon_j \\ &= T_{n,1}+T_{n,2}. \end{aligned}$$

It is easy to see that

$$\begin{aligned} T_{n,1} &= \mathrm{tr}\left\{n^{-3}\sum_{k=1}^{n}\sum_{i=1}^{n} f(U_k)^{-2}\Lambda(U_k)^{-1}X_kX_k^\tau\Lambda(U_k)^{-1}X_iX_i^\tau K_h^2(U_i-U_k)\varepsilon_i^2\right\} \\ &= n^{-1}h^{-1}\mu_0\sigma^2\mathrm{tr}\Big[E\Big\{f(U_1)^{-1}\Lambda(U_1)^{-1}X_1X_1^\tau\Big\}\Big]+o_P(n^{-1}h^{-1}) \\ &= n^{-1}h^{-1}\mu_0\sigma^2E\Big\{f(U_1)^{-1}X_1^\tau\Lambda(U_1)^{-1}X_1\Big\}+o_P(n^{-1}h^{-1}). \end{aligned}$$

Next,

$$\begin{aligned} T_{n,2} &= 2n^{-3}\sum_{1\le i<j\le n}\varepsilon_i\varepsilon_jX_j^\tau\left\{\sum_{k=1}^{n} f(U_k)^{-2}\Lambda(U_k)^{-1}X_kX_k^\tau\Lambda(U_k)^{-1}K_h(U_i-U_k)K_h(U_j-U_k)\right\}X_i \\ &= 2n^{-3}h^{-1}K(0)\sum_{1\le i<j\le n}\Big[\varepsilon_i\varepsilon_jX_j^\tau\Big\{f(U_i)^{-2}\Lambda(U_i)^{-1}X_iX_i^\tau\Lambda(U_i)^{-1}+ \\ &\qquad\qquad f(U_j)^{-2}\Lambda(U_j)^{-1}X_jX_j^\tau\Lambda(U_j)^{-1}\Big\}X_iK_h(U_i-U_j)\Big] \\ &\quad +2n^{-3}\sum_{1\le i<j\le n}\varepsilon_i\varepsilon_jX_j^\tau\left\{\sum_{k\neq i,j} f(U_k)^{-2}\Lambda(U_k)^{-1}X_kX_k^\tau\Lambda(U_k)^{-1}K_h(U_i-U_k)K_h(U_j-U_k)\right\}X_i \\ &= T_{n,2,1}+T_{n,2,2}. \end{aligned}$$

It is easy to show that

$$E(T_{n,2,1})=0,\quad \mathrm{var}(T_{n,2,1})=O(n^{-4}h^{-3}),$$

which implies that $T_{n,2,1}=o_P(n^{-1}h^{-1})$. Now, let's see $T_{n,2,2}$. Let

$$B_{i,j,k}=f(U_k)^{-2}\Lambda(U_k)^{-1}X_kX_k^\tau\Lambda(U_k)^{-1}K_h(U_i-U_k)K_h(U_j-U_k).$$

Note that

$$\begin{aligned} E\left[n^{-1}X_j^\tau\sum_{k\neq i,j}\Big\{B_{i,j,k}-E(B_{i,j,k}|U_i,U_j)\Big\}X_i\right]^2 &\le \mathrm{tr}\left\{n^{-2}\sum_{k\neq 1,2}E\Big(B_{1,2,k}X_2X_2^\tau B_{1,2,k}X_1X_1^\tau\Big)\right\} \\ &= O(n^{-1}h^{-2}). \end{aligned}$$

Therefore

$$\mathrm{var}(T_{n,2,2})=O\Big(n^{-2}E\Big\{X_1^\tau E(B_{1,2,3}|U_1,U_2)X_2\Big\}^2+n^{-3}h^{-2}\Big)=O(n^{-2}h^{-1}).$$

This together with $E(T_{n,2,2}) = 0$ leads to

$$T_{n,2,2} = O_P(n^{-1}h^{-1/2}) = o_P(n^{-1}h^{-1}).$$

Therefore $T_{n,2} = o_P(n^{-1}h^{-1})$, which implies

$$n^{-1}\boldsymbol{\epsilon}^\tau \mathbf{H}^\tau \mathbf{H}\boldsymbol{\epsilon} = n^{-1}h^{-1}\mu_0\sigma^2 E\Big\{f(U_1)^{-1}X_1^\tau \Lambda(U_1)^{-1}X_1\Big\} + o_P(n^{-1}h^{-1}). \tag{A.8}$$

Let $\ddot{\mathbf{m}} = \Big(X_1^\tau \mathbf{a}''(U_1), \cdots, X_n^\tau \mathbf{a}''(U_n)\Big)^\tau$. Using similar argument as above, we can obtain

$$n^{-1}\mathbf{m}^\tau(I_n - \mathbf{H})^\tau(I_n - \mathbf{H})\boldsymbol{\epsilon} = -2^{-1}n^{-1}\nu_2 h^2 \ddot{\mathbf{m}}^\tau(I_n - \mathbf{H})\boldsymbol{\epsilon}\Big(1 + o_P(1)\Big) = O_P(h^2 n^{-1/2}).$$

This together with (A.2), (A.3), (A.4), (A.5), (A.6), (A.7), (A.8) proves Theorem 2.

The proof of Theorem 3:

Let $\mathcal{S}_\ell$, $\ell = 1, \cdots, 2^{p+q} - 1$, be all wrong models, we have

$$\{\mathcal{S}_{\ell_0} \neq \mathcal{S}_0\} = \bigcup_{\ell=1}^{2^{p+q}-1} \{CV(\mathcal{S}_\ell) < CV(\mathcal{S}_0)\}.$$

By standard argument, we have that the increment on CV sum of squares caused by mistakenly treating functional coefficient as constant is $O_P(1)$, this together with the explanation in the remark 1 of Theorem 2 give us that

$$P\{CV(\mathcal{S}_\ell) < CV(\mathcal{S}_0)\} \leq P\{|o_p(n^{-1}h^{-1})| > C_0 n^{-1}h^{-1}\} \longrightarrow 0,$$

where C_0 is a constant and $C_0 > 0$, this leads to

$$P(\mathcal{S}_{\ell_0} \neq \mathcal{S}_0) \leq \sum_{\ell=1}^{2^{p+q}-1} P\{CV(\mathcal{S}_\ell) < CV(\mathcal{S}_0)\} \longrightarrow 0.$$

So,

$$P(\mathcal{S}_{\ell_0} = \mathcal{S}_0) \longrightarrow 1.$$

References

Anderson R. M. Anderson and R. M. May, *Infectious Disease of Humans: Dynamics and Control.* Oxford: Oxford University Press, (1991).

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, *2th International Symposium on Information Theory* (B. N. Petrov and F.Czáki, eds), pp, 276-281, Budapest: Akademiai Kiadó.

Bailey, N. T. J. (1975). *The Mathematical Theory of Infectious Diseases and Its Applications.* London: Griffin.

Breiman, L. and Friedman, J.H. (1985). Estimating optimal transformations for multiple regression and correlation (with discussion). *J. Ameri. Statist. Assoc.*, **80**, 580–619.

Carroll, R.J., Fan, J., Gijbels, I, and Wand, M.P. (1997). Generalized partially linear single-index models. *J. Ameri. Statist. Assoc.*, **92**, 477-489.

Chen, R. and Tsay, R.S. (1993). Functional-coefficient autoregressive models. *J. Ameri. Statist. Assoc.*, **88**, 298-308.

Cheng, B. and Tong, H. (1992) On consistent nonparametric order determination and chaos (with discussion), *J. R. Statist. Soc. B.* **54** 427- 449.

Durlauf, S. N. (2001). Manifesto for a growth econometrics. *J. Econometrics* , **100**, 65-69.

Ellner, S. P., B. A. Bailey, G.V. Bobashev, A.R. Gallant, B. T. Grenfell and D.W. Nychka (1998) Noise and nonlinerity in Measles Epidemics: Combining Mechanistic and Statistical Approaches to Population Modeling. *Am. Nat.* **151**, 425-440.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications.* Chapman and Hall, London.

Fan, J., Härdle, W. and Mammen, E. (1998). Direct estimation of additive and linear components for high dimensional data. *Ann. Statist.*, **26**, 943-971.

Fan, J. and Zhang, W. (1999). Statistical estimation in varying coefficient models. *Ann. Statist.*, **27**, 1491-1518.

Fan, J. and Zhang, J. T. (2000). Functional linear models for longitudinal data. *J. Roy. Statist. Soc. Ser. B*, **62**, 303-322.

Fan, J. and Zhang, W. (2000). Simultaneous confidence bands and hypothesis testing in varying-coefficient models. *Scand. J. Statist.*, **27**, 715-731.

Fan, J., Yao, Q. and Cai, Z. (2003). Adaptive varying-coefficient linear models. *J. Roy. Statist. Soc. Ser. B*, **65**, 57-80.

Finkenstädt, B. F. and Grenfell, B. T. (2000). Time series modelling of childhood diseases: a dynamical systems approach. *Appl. Statist.* **49**, 187-205.

Friedman, J.H. (1991). Multivariate adaptive regression splines (with discussion). *Ann. Statist.*, **19**, 1–141.

Green, P.J. and Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models: a Roughness Penalty Approach.* Chapman and Hall, London.

Gu, C. and Wahba, G. (1993). Smoothing spline ANOVA with component-wise Bayesian "confidence intervals". *J. Comput. Graph. Statist.*, **2**, 97–117.

Hannan, E. J. and Quinn, B. G. (1979). The determination of the order of an autoregression. *J. R. Statist. Soc. B*, **41**, 190-195.

Härdle, W., Hall, P. and Ichimura, H. (1993) Optimal smoothing in single-index models. *Ann. Statist.*, **21**, 157-178.

Härdle, W. and Marron, J. S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *Ann. Statist.*, **13**, 1465–1481.

Härdle, W. and Stoker, T.M. (1989). Investigating smooth multiple regression by the method of average derivatives. *J. Ameri. Statist. Assoc.*, **84**, 986–995.

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

Hastie, T.J. and Tibshirani, R. J. (1993). Varying-coefficient models. *J. Roy. Statist. Soc. B.*, **55**, 757-796.

Heckman, J., Ichimura, H., Smith, J. and Todd, P. (1998). Characterizing Selection Bias Using Experimental Data. *Econometrica*, **66** 1017-1098.

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

Li, K.-C. (1991). Sliced inverse regression for dimension reduction (with discussion). *J. Ameri. Statist. Assoc.*, **86**, 316–342.

Liu, W. and Levin, S.A. (1986) Influence of nonlinear incidence rate upon the behavior of SIRS epidemiological models. *J. Math. Biology*, **23** 187-204.

Masry, E. (1996). Multivariate local polynomial regression for time series: uniform strong consistency and rates. *J. Time Ser. Anal.*, **17**, 571–599.

Speckman, P. (1988). Kernel smoothing in partial linear models. *J. Roy. Statist. Soc. B.*, **50**, 413-436.

Schwartz, G. (1978). Estimating the dimensions of a model. *Ann. Statist.* **6**, 461-464.

Stone, C.J., Hansen, M., Kooperberg, C. and Truong, Y.K. (1997). Polynomial Splines and Their Tensor Products in Extended Linear Modeling. *Ann. Statist.*, **25**, 1371-1470.

Stone, M. (1977). Asymptotics for and against cross-validation. *Biometrika*, **64**, 29-35.

Wahba, G. (1984). Partial spline models for semiparametric estimation of functions of several variables. In *Statistical Analysis of Time Series*, Proceedings of the Japan U.S. Joint Seminar, Tokyo, 319–329. Institute of Statistical Mathematics, Tokyo.

Burnham, K. P., and D. R. Anderson (2002), *Model selection and multimodel inference: a practical information-theoretic approach*, Springer, New York.

Xia, Y. and Li, W. K. (1999). On the estimation and testing of functional-Coefficient linear models. *Statistica Sinica*, 9, 735-758.

Xia, Y., Tong, H., Li, W. K. and Zhu, L. (2002). An adaptive estimation of dimension reduction space (with discussions). *J. Roy. Statist. Soc. B.*, **64**, 363-410.