Statistica Sinica **21** (2011), 1913-1929
doi:http://dx.doi.org/10.5705/ss.2009.201

IDENTIFICATION OF THE CONSTANT COMPONENTS IN GENERALISED SEMIVARYING COEFFICIENT MODELS BY CROSS-VALIDATION

Wenyang Zhang

Southwestern University of Finance and Economics
and The University of Bath

Abstract: In practice, some coefficients in generalised varying coefficient models may be constant. We pay a price on the variance side of an estimator when constant coefficients are treated as special functions. This prompts the question of how to identify the constant coefficients. This is basically a model selection problem. In this paper, we use cross-validation (CV) as a criterion for model selection to identify the constant coefficients. We investigate the asymptotic properties of the proposed CV-based model selection approach. We report on a simulation study conducted to show how well the proposed method works when sample size is finite. Finally, the proposed method is used to analyse a data set from China about contraceptive use, which leads to some interesting findings.

Key words and phrases: Cross-validation, generalised semivarying-coefficient models, generalised varying-coefficient models, local linear modelling, local maximum likelihood estimation.

1. Introduction

Semiparametric modelling is a promising modelling strategy, it makes use of the prior information through its parametric component and the flexibility of model specification is enhanced by the nonparametric ingredient. Among many semiparametric models, the varying coefficient models are appealing due to their flexibility and interpretability. There is much literature addressing varying coefficient models, to include Fan and Huang (2005), Fan and Zhang (1999, 2000), Fan, Huang and Li (2007), Li and Liang (2008), Qu and Li (2006), Wang, Li, and Huang (2008), Wang and Xia (2009), Wang, Kai, and Li (2009), Sentürk and Müller (2006), Sun, Zhang and Tong (2007), Xia and Li (1999), Zhang, Lee, and Song (2002), and the references therein.

Varying coefficient models extend naturally to generalised varying coefficient models, for which there is strong demand from practice. In particular, these models are used in the analysis of the data set that stimulates this paper. The data set is from China about contraceptive use there during January 1980 to July 1988, when women were encouraged to use contraceptives to postpone giving birth. The womens attitude towards contraceptive use in China is varying across different age groups, levels of education, occupations, ethnic groups. Also, whether a woman previously used the contraceptive may affect her rate of contraceptive failure. Age, education, occupation, and ethnic group are some of the many factors contributing to the failure rate of contraceptive in China. To explore how the factors affect the failure rate, logistic regression models are traditionally employed. However, these may not work well when coefficients, which can be interpreted as the impacts of the factors concerned, are assumed to be constant when they are not. China has been changing, and the impacts of the factors involved likely change over time. To take time into account, for example, the simplest way is to replace the constant coefficients in the traditional logistic regression models by functions of time; this leads to the varying coefficient logistic regression model

$$\log \frac{\pi(U,\ X)}{1-\pi(U,\ X)} = X^{\mathrm{T}}\mathbf{a}(U), \tag{1.1}$$

where X is the vector of all factors concerned, U is time, and $\pi(U,\ X)$ is the failure rate of contraceptive given X and U. The model (1.1) is a specific case of the generalised varying coefficient models that are, in turn, a specific case of the generalised semivarying-coefficient models defined by (1.3), with $q = 0$.

It can of course be the case that the impacts of some factors may not change over time at all, which means some components of $\mathbf{a}(\cdot)$ in (1.1) may be constant. We formulate this problem as that of identifying the constant components in the class of models

$$\log \frac{\pi(U,\ X,\ Z)}{1-\pi(U,\ X,\ Z)} = X^{\mathrm{T}}\mathbf{a}(U) + Z^{\mathrm{T}}\boldsymbol{\beta}, \tag{1.2}$$

where X is the vector of the factors whose impacts change, Z is the vector of the factors whose impacts are constant.

To make our models more general, from now on, we take U as a scalar, not necessarily time, $X = (x_1, \ldots, x_p)^{\mathrm{T}}$ is a p-dimensional covariate, and $Z = (z_1, \ldots, z_q)^{\mathrm{T}}$ is a q-dimensional covariate. The log conditional density function of the response variable y given U, X and Z is

$$\ell\left[g^{-1}\left\{X^{\mathrm{T}}\mathbf{a}(U) + Z^{\mathrm{T}}\boldsymbol{\beta}\right\},\ y\right], \tag{1.3}$$

where $\mathbf{a}(\cdot) = (a_1(\cdot), \ldots, a_p(\cdot))^{\mathrm{T}}$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)^{\mathrm{T}}$ are unknowns to be estimated, $g(\cdot)$ is a known link function, and $\ell(\cdot,\ \cdot)$ is known as well; $(U_i,\ X_i^{\mathrm{T}},\ Z_i^{\mathrm{T}},\ y_i)$, $i = 1, \ldots, n$, is a i.i.d. sample from $(U,\ X^{\mathrm{T}},\ Z^{\mathrm{T}},\ y)$. We term (1.3) a generalised semivarying coefficient model. This is a large class of models that includes many common models like generalised linear models, generalised varying coefficient models, varying coefficient models, semivarying coefficient models, and partially-linear models (Ma, Chiou and Wang (2006)).

We are going to use cross-validation (CV) as a model selection criterion to identify the constant components in (1.3). We propose two algorithms to implement the CV-based model selection procedure without computing the CVs of all 2^{p+q} models, and show the procedure consistent. We also conduct simulation study to show how well the CV-based model selection works with finite samples.

The paper is organised as follows. Section 2 describes an estimation procedure that can be used to estimate the unknown functional coefficients and constant coefficients in model (1.3). In Section 3 we show how to construct the CV for (1.3) and describe two algorithms to implement the CV-based model selection procedure. Asymptotic properties of the CV-based model selection are presented in Section 4. The performance of the CV-based model selection when sample size is finite is assessed by a simulation study in Section 5. In Section 6, we explore how the impacts of several factors on the failure rate of contraceptive in China change over time, based on the proposed model selection and estimation procedure.

2. Estimation Procedure

Our estimation procedure is based on local maximum likelihood estimation. We estimate the constant coefficients first, then the functional coefficients. After obtaining the estimators of the constant coefficients, we estimate the functional coefficients based on the model with the constant coefficients replaced by their estimators.

2.1. Estimation of the constant coefficients

The estimation of the constant coefficient, $\boldsymbol{\beta}$, consists of two-steps: the local maximum likelihood estimator, $\tilde{\boldsymbol{\beta}}(U_i)$, of $\boldsymbol{\beta}$ is obtained at each U_i, $i = 1, \ldots, n$; $\tilde{\boldsymbol{\beta}}(U_i)$ is averaged over $i = 1, \ldots, n$ to get the final estimator of $\boldsymbol{\beta}$. Details are as follows.

For any given u, let $\dot{\mathbf{a}}(u)$ be the first derivative of $\mathbf{a}(u)$. By a Taylor's expansion, we have $\mathbf{a}(U_i) \approx \mathbf{a}(u) + \dot{\mathbf{a}}(u)(U_i - u)$ when U_i is in a small neighbourhood of u; this leads to the local log likelihood function

$$\sum_{i=1}^{n} \ell\left[g^{-1}\left\{X_i^{\mathrm{T}}\mathbf{a} + X_i^{\mathrm{T}}\mathbf{b}(U_i - u) + Z_i^{\mathrm{T}}\boldsymbol{\beta}\right\},\ y_i\right] K_{h_1}(U_i - u), \tag{2.1}$$

where $K_{h_1}(\cdot) = K(\cdot/h_1)/h_1$, $K(\cdot)$ is a kernel function and h_1 is a bandwidth.

Maximise (2.1) with respect to $(\mathbf{a}^{\mathrm{T}},\ \mathbf{b}^{\mathrm{T}},\ \boldsymbol{\beta}^{\mathrm{T}})$, and denote the maximiser by $(\tilde{\mathbf{a}}^{\mathrm{T}}(u),\ \tilde{\mathbf{b}}^{\mathrm{T}}(u),\ \tilde{\boldsymbol{\beta}}^{\mathrm{T}}(u))$. $\tilde{\boldsymbol{\beta}}(u)$ is the local maximum likelihood estimator of $\boldsymbol{\beta}$ and, because it only makes use of the information provided by the data in a small neighbourhood of u, it has large variance. To reduce the variance and get the final estimator of $\boldsymbol{\beta}$, we compute $\tilde{\boldsymbol{\beta}}(u)$ at each U_i, $i = 1, \ldots, n$, and average $\tilde{\boldsymbol{\beta}}(U_i)$ over $i = 1, \ldots, n$ to get $\hat{\boldsymbol{\beta}} = (1/n) \sum_{i=1}^n \tilde{\boldsymbol{\beta}}(U_i)$ as the final estimator of $\boldsymbol{\beta}$.

The proposed estimation procedure for $\boldsymbol{\beta}$ is easy to implement, and the estimator is asymptotic normal with convergence rate of order $O(n^{-1/2})$ when the bandwidth h_1 is properly selected, see Zhang and Peng (2010).

Theorem.(Zhang and Peng (2010)) *Under the conditions* (1)−(5) *in the Appendix, if* $h_1 \to 0$ *and* $nh_1^2/(-\log h_1) \to \infty$*, then* $\sqrt{n}\left(\hat{\beta}_q - \beta_q\right) \xrightarrow{D} N(0,\ \sigma^2)$, *where* $\sigma^2 = E\left(e_{p+q,p+q}^T \left[E\left\{\rho(U,\ \mathcal{Z})\mathcal{Z}\mathcal{Z}^T|U\right\}\right]^{-1} e_{p+q,p+q}\right)$, $e_{k,k}$ *is an unit vector of length* k *with the* k*th component being* 1, $\mathcal{Z} = (X^T,\ Z^T)^T$, $\rho(U,\ \mathcal{Z}) = -Q_2[g\{E(y|U,\ \mathcal{Z})\},\ E(y|U,\ \mathcal{Z})]$, *and* $Q_k(t,\ y) = \partial^k \ell\left\{g^{-1}(t),\ y\right\}/\partial t^k$.

From the theorem, we see that the selection of the bandwidth h_1 is not an issue, the convergence rate of $\hat{\boldsymbol{\beta}}$ is of order $O(n^{-1/2})$ as long as $h_1 \longrightarrow 0$ and $nh_1^2/\log h_1 \longrightarrow \infty$. This indicates that any bandwidth in a fairly wide range would work well; in practice, we select h_1 by cross-validation, see Zhang and Peng (2010).

2.2. Estimation of the functional coefficients

After the estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is obtained, substituting $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ in (2.1) leads to

$$\sum_{i=1}^n \ell\left[g^{-1}\left\{X_i^{\mathrm{T}}\mathbf{a} + X_i^{\mathrm{T}}\mathbf{b}(U_i - u) + Z_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}\right\},\ y_i\right] K_h(U_i - u). \tag{2.2}$$

Maximise (2.2) with respect to $(\mathbf{a}^{\mathrm{T}},\ \mathbf{b}^{\mathrm{T}})$, and denote the maximiser by $(\hat{\mathbf{a}}^{\mathrm{T}},\ \hat{\mathbf{b}}^{\mathrm{T}})$. The estimator $\hat{\mathbf{a}}(u)$ of $\mathbf{a}(u)$ is taken to be $\hat{\mathbf{a}}$. As the estimator $\hat{\boldsymbol{\beta}}$ is of convergence rate $O(n^{-1/2})$, the asymptotic behavior of the estimator $\hat{\mathbf{a}}(u)$ is the same as that of the estimator of $\mathbf{a}(u)$ obtained when $\boldsymbol{\beta}$ is known.

After $\boldsymbol{\beta}$ in the model (1.3) is replaced by $\hat{\boldsymbol{\beta}}$, (1.3) becomes a generalised varying coefficient model; the optimal bandwidth for the estimation of $\mathbf{a}(u)$ should be of order $O(n^{-1/5})$. In practice, we could use cross-validation to select an initial bandwidth h^*; because the bandwidth selected by cross-validation tends to narrower than it should be, we use $h = 1.1h^*$.

Note that another estimator of $\mathbf{a}(u)$ can be obtained during the estimation of the constant coefficient $\boldsymbol{\beta}$, the part of the maximiser of (2.1) corresponding to $\mathbf{a}$. The convergence rate of this estimator is the same as that of the proposed estimator, but they do not share the same asymptotic bias or variance. Theoretically speaking, the proposed estimator is better. Zhang, Lee, and Song

(2002) had a detailed discussion on this issue for semivarying coefficient models, a special case of the models addressed in this paper.

3. Model Selection

In this section, we present the criterion for model selection, the algorithms to implement the model selection procedure, and a brief discussion of bandwidth selection in the model selection.

3.1. Criterion of selection

We use cross-validation (CV) as the criterion for model selection. The CV for model (1.3) is constructed as follows: for each i, $i = 1, \ldots, n$, we delete the ith observation and estimate $\mathbf{a}(U_i)$ and $\boldsymbol{\beta}$ based on the other observations by the estimation procedure of Section 2. Denote the resulting estimators by $\hat{\mathbf{a}}^{\backslash i}(U_i)$ and $\hat{\boldsymbol{\beta}}^{\backslash i}$, respectively. A natural estimator of the log conditional density function of y at y_i given $U = U_i$, $X = X_i$, and $Z = Z_i$ is $L_i = \ell\left[g^{-1}\left\{X_i^{\mathrm{T}}\hat{\mathbf{a}}^{\backslash i}(U_i) + Z_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}^{\backslash i}\right\},\ y_i\right]$, and the cross-validation sum is

$$\mathrm{CV} = -n^{-1}\sum_{i=1}^{n} L_i. \tag{3.1}$$

Our formal model selection procedure is: for each possible model, compute its CV by (3.1); the selected model is the one with the smallest CV among all possible models.

3.2. Algorithms

To compute the CVs of all possible models is not practically feasible, with L covariates, there are 2^L possible models. In this section, we present two algorithms to reduce the computational burden.

Let L be the number of the coefficients in the model, and the coefficients in the model be $\boldsymbol{\alpha}(\cdot) = \big(\alpha_1(\cdot), \ldots, \alpha_L(\cdot)\big)$, the model with coefficients $\alpha_{i_l}(\cdot)$, $l = 1, \ldots, k$, being functional, others being constant by $\{i_1, \ldots, i_k\}$.

3.2.1. Backward elimination

Instead of computing the CVs of all models, we use backward elimination to find the chosen model. Details are as follows.

(1) Start with the full model, $\{1, \ldots, L\}$, and compute its CV by (3.1). Denote the full model by $\mathcal{M}_L$, its CV by CV_L.

(2) For any integer k, suppose the current model is $\mathcal{M}_k = \{i_1, \ldots, i_k\}$ with CV given by CV_k. Take $\mathcal{M}_{k-1}$ to be the model with the smallest residual sum

of squares (RSS) among the models $\{i_1, \dots, i_{j-1},\ i_{j+1}, \dots, i_k\}$, $j = 1, \dots, k$. If $\mathrm{CV}_k < \mathrm{CV}_{k-1}$, the chosen model is $\mathcal{M}_k$, and the model selection is ended; otherwise, continue to compute $\mathcal{M}_l$ and CV_l until either $\mathrm{CV}_l < \mathrm{CV}_{l-1}$ or $l = 0$.

3.2.2. Discrepancy from average

A more aggressive way to reduce the computation involved in the model selection procedure is based on the discrepancy of the estimated function from its average. Explicitly, we first treat all $\alpha_i(\cdot)$, $i = 1, \dots, L$, as functional. For each i, $i = 1, \dots, L$, we compute the discrepancy of the estimated function $\hat{\alpha}_i(\cdot)$ from its average:

$$S_i = \sum_{k=1}^{n} \left\{\hat{\alpha}_i(U_k) - \bar{\alpha}_i\right\}^2, \quad \bar{\alpha}_i = n^{-1} \sum_{k=1}^{n} \hat{\alpha}_i(U_k), \quad i = 1, \dots, L.$$

We sort S_i, $i = 1, \dots, L$, in an increasing order, say $S_{i_1} \le \cdots \le S_{i_L}$, then compute the CVs for the models $\{i_k, \dots, i_L\}$ from $k = 1$ to the turning point k_0 where the CV starts to increase. The chosen model is $\{i_{k_0}, \dots, i_L\}$.

The algorithm based on the discrepancy from average is much faster than the backward elimination based algorithm; however, from simulations, we find it less accurate.

3.3. Bandwidth issue

The bandwidth used for model selection is different than that used for estimation. From the asymptotic properties of the proposed CV in Section 4, we can see the increment in CV is of order $O(1)$ if we mistakenly treat a functional coefficient as constant, and the increment is of order $O\left((nh)^{-1}\right)$ if we mistakenly treat a constant coefficient as functional. To avoid mistakenly treating a constant coefficient as functional, we use a bandwidth as small as possible. Although it is understandable that we may risk mistakenly treating some functional coefficients as constant if the bandwidth is too small, this is not a big problem since we risk $O(1)$ on bias side with a gain on variance side of order $O\left((nh)^{-1}\right)$, and this is small compared with the loss on bias side. So, we do not risk mistakenly treat a functional coefficient as constant as long as $nh \longrightarrow \infty$.

We conclude that, as far as the model selction is concerned, the bandwidth selection is not as crucial as that in estimation procedure; any reasonably small bandwidth would work well. This conclusion held up in our simulation studies.

4. Asymptotic Properties

Before presenting the asymptotic properties of the proposed CV, we introduce some notation. Let $f(u)$ be the density function of U, $\mathbf{a}_i = \mathbf{a}(U_i)$, $\mu_i = \int u^i K(u)du$, and $\nu_i = \int u^i K^2(u)du$. We denote the second derivative of $\mathbf{a}(u)$ by $\ddot{\mathbf{a}}(u)$.

Theorem 1. *Under Conditions* (1)−(6) *in the Appendix, if the working model is the true model* (1.3)*, the asymptotic form of the CV is*

$$\begin{aligned} CV = & -n^{-1}\sum_{i=1}^{n} \ell\left[g^{-1}\left(X_i^T\mathbf{a}_i + Z_i^T\boldsymbol{\beta}\right),\ y_i\right] + 2^{-3}\mu_2^2 h^4 E\left[\zeta\left\{X^T\ddot{\mathbf{a}}(U)\right\}^2\right] \\ & -2^{-1}\nu_0(nh)^{-1}E\left\{f(U)^{-1}\zeta X^T\Lambda^{-1}X\right\} + o_P(n^{-1}h^{-1} + h^4), \end{aligned}$$

where $\zeta = Q_2\left(X^T\mathbf{a}(U) + Z^T\boldsymbol{\beta},\ y\right),\ \Lambda = E\left(\zeta X X^T|U\right).$

In the asymptotic form of the CV, the first term represents the random error, the second term reflects the asymptotic bias, and the third term reflects the asymptotic variance.

We next consider the CV when the working model mistakenly treats some constant coefficients as functional, say we assume β_1 in the model is mistakenly treated as functional.

Theorem 2. *Under Conditions* (1)−(6) *in the Appendix, if the working model mistakenly treats β_1 as functional, the asymptotic form of the CV is*

$$\begin{aligned} CV = & -n^{-1}\sum_{i=1}^{n} \ell\left[g^{-1}\left(X_i^T\mathbf{a}_i + Z_i^T\boldsymbol{\beta}\right),\ y_i\right] + 2^{-3}\mu_2^2 h^4 E\left[\zeta\left\{X^T\ddot{\mathbf{a}}(U)\right\}^2\right] \\ & -2^{-1}\nu_0(nh)^{-1}E\left\{f(U)^{-1}\zeta \mathcal{X}^T\Lambda_*^{-1}\mathcal{X}\right\} + o_P(n^{-1}h^{-1} + h^4), \end{aligned}$$

where $\mathcal{X} = (X^T,\ z_1)^T$, *and* $\Lambda_* = E\left(\zeta\mathcal{X}\mathcal{X}^T|U\right).$

Remark. From Theorems 1 and 2, we can see the increment in CV due to the working model mistakenly treating β_1 as functional is

$$2^{-1}\nu_0(nh)^{-1}E\left\{f(U)^{-1}\zeta\left(X^{\mathrm{T}}\Lambda^{-1}X - \mathcal{X}^{\mathrm{T}}\Lambda_*^{-1}\mathcal{X}\right)\right\}.$$

By simple calculation,

$$\begin{aligned} & X^{\mathrm{T}}\Lambda^{-1}X - \mathcal{X}^{\mathrm{T}}\Lambda_*^{-1}\mathcal{X} \\ & = -\left\{E(\zeta z_1^2|U) - E(\zeta z_1 X^{\mathrm{T}}|U)\Lambda^{-1}E(\zeta z_1 X|U)\right\}^{-1}\left\{X^{\mathrm{T}}\Lambda^{-1}E(\zeta z_1 X|U) - z_1\right\}^2 < 0. \end{aligned}$$

So, $E\left\{f(U)^{-1}\zeta\left(X^{\mathrm{T}}\Lambda^{-1}X - \mathcal{X}^{\mathrm{T}}\Lambda_*^{-1}\mathcal{X}\right)\right\} > 0$, which implies that the increment in CV due to the working model mistakenly treating a constant coefficient as functional is detectable up to $O\left((nh)^{-1}\right)$. By standard arguments, we have the increment in CV caused by mistakenly treating a functional coefficient as constant is $O_P(1)$. This means that a model selection procedure based on CV is sensible since $O\left((nh)^{-1}\right)$ is the dominant term aside from random error, which does not change with the model.

Theorem 3. *Under Conditions* (1)−(6) *in the Appendix,* $\lim_{n\to\infty} P(\mathcal{S}_* = \mathcal{S}_0) = 1$, *where* $\mathcal{S}_0$ *is the true model, and* $\mathcal{S}_*$ *is the model with the smallest CV.*

5. Simulation Study

In this Section, we use simulations to examine how well our proposed procedure works. We compare the two algorithms of Section 3.2, and we examine how well the hypothesis test-based model selection procedure works.

Example 1. We generated a sample $(y_i,\ X_i,\ U_i)$, $i = 1,\ldots,$ 1,500, from the logistic regression model

$$\log\left\{\frac{P(y=1|X,\ U)}{1-P(y=1|X,\ U)}\right\} = x_1a_1(U) + x_2a_2(U) + x_3a_3 + x_4a_4.$$

Here the $X_i = (x_{i1},\ x_{i2},\ x_{i3},\ x_{i4})^{\mathrm{T}}$ were independently generated from a normal distribution $N(\mathbf{0},\ I_4)$, and the U_i were independently generated from a uniform distribution $U(0,\ 1)$. We set $a_1(u) = \sin(2\pi u)$, $a_2(u) = \cos(2\pi u)$, $a_3 = 2$, and $a_4 = 1$. The Epanechnikov kernel $K(t) = 0.75(1-t^2)_+$ was used in the estimation procedure, and the bandwidth was taken to be 0.03 for the model selection.

As the computation involved is expensive, we only carried out 100 simulations. We found, in the 100 simulations, the ratio of picking the right model was chosen 95% of the time by the best subset approach, 92% of the time by the backward elimination approach, and 81% of the time by the discrepancy from average approach. This backward elimination worked reasonably well, and better than the discrepancy from average approach.

Among the 100 simulations we conducted, there were no chosen models mistakenly taking any functional coefficient as constant; all wrong models mistakenly took some constant coefficient as functional. This is in line with the theoretical analysis in Section 3.3.

For assessing the performance of the hypothesis test-based model selection, the generalised maximum likelihood ratio test developed in Zhang and Peng (2010), coupled with the backward elimination algorithm, was used. We used the significant levels 0.01, 0.05 and 0.1. Among 100 simulations, the hypothesis test-based model selection chose the right model 82% of the time when the significant level was 0.01, 74% of the time when the significant level was 0.05, and

Table 1. The MSEs or MISEs of the Estimators.

Estimator	$\hat{a}_1(\cdot)$	$\hat{a}_2(\cdot)$	Estimator	$\hat{a}_3$	$\hat{a}_4$
MISE	0.039	0.036	MSE	0.018	0.007

59% of the time when the significant level was 0.1. Thus, regardless of the significant level, the CV-based model selection procedure always worked better than the hypothesis test-based one. Performance of the hypothesis test-based model selection did depend on the selection of significant level, and presents a problem as to the appropriate level. The solution may come back to cross-validation. Then too, if a reasonable significance level is identified, the hypothesis test may still need calibration lest the power be actual power rather than nominal power.

To examine the performances of the proposed estimation for functional coefficients and the proposed estimation for constant coefficients, we use mean squared error (MSE) to assess the accuracy of an estimator of a constant coefficient, and mean integrated squared error (MISE) to assess the accuracy of an estimator of a functional coefficient. We used the same bandwidth, 0.20, to estimate both functional coefficients and constant coefficients. The MSEs or MISEs of the estimators are reported in Table 1

To examine the robustness of the proposed estimation for constant coefficients against bandwidth selection, we also computed the MSEs of the estimators of the constant coefficients at different bandwidths. Our results show the MSE of $\hat{a}_3$ was always less than 0.02 and the MSE of $\hat{a}_4$ was always less than 0.0075, as long as the bandwidth was in the interval [0.17, 0.5]. This suggests that our estimation for constant coefficients works well as long as the bandwidth is in a reasonable range, and undersmoothing is not necessary. This is in line with the theory in Section 2.1.

6. Data Analysis

We return to the data set from China about the contraceptive use in China. Interest is focused on the following factors affecting contraceptive use: age, region of residence, education, occupation, ethnicity, previous use of a contraceptive, previous failure in use of a contraceptive, and motivation to contraceptive use.

Age is grouped as "less than 24", "25 to 29" (x_2), "30 to 34" (x_3) and "over 35" (x_4). We take "less than 24" as reference, and the difference in the impact on contraceptive use among different age groups are modelled by the dummy variables x_i, $i = 2,\ 3,\ 4$. We take "urban" as reference, the difference between urban and rural is modelled by a dummy variable x_5; education is categorised as "primary -" or "junior +", we take "primary -" as reference, and the difference between "primary -" or "junior +" is modelled by a dummy variable x_6;

occupation is "agriculture", "industry" (x_7), "service" (x_8), "professional" (x_9), or "other non-agriculture" (x_{10}), we take "agriculture" as reference, and the difference among different occupations as modelled by the dummy variables x_i, $i = 7, \dots, 10$; we take "non-Han" as reference, the difference between "Han" and "non-Han" as modelled by a dummy variable x_{11}; we use dummy variables x_{12} and x_{13} to model "previous use of contraceptive" and "previous failure of contraceptive", respectively; motivation to contraceptive use is categorised as "self motivated" or "response to campaign", we take "self motivated" as reference, and use x_{14} to model the difference between "self motivated" and "response to campaign". Chronological time is denoted by U, and we set $x_1 = 1$ to incorporate an intercept into the modelling. The dependent variable, Y, is 1 if the contraceptive fails, 0 otherwise. The sample size is $14,639$.

For the identification of which coefficients are constant, which are functional, we use the Epanechnikov kernel as the kernel function, and take the bandwidth to be 1% of the range of U. The backward elimination algorithm is used to implement the model selection procedure, and it appears that only the coefficient of x_4 is constant.

The model (1.2) with $X = (x_1,\ x_2,\ x_3,\ x_5,\ \dots\ x_{14})^{\mathrm{T}}$ and $Z = x_4$ is now used to fit the data set. The estimation procedure in Section 2 is used to estimate the constant coefficient and functional coefficients. The kernel function is still the Epanechnikov kernel, but the bandwidth is taken to be 15% of the range of U. The obtained estimate of the constant coefficient is -2.001, and the estimates of the functional coefficients are presented in Figure 1.

From Figure 1, one sees that the failure rate of contraceptive in China was decreasing with time in general, and from 1986 to 1988 saw a very sharp decrease. This is generally attributed to more and more effective contraceptive methods being introduced in China during those years.

The obtained results also indicate women aged less than 24 were significantly more likely to have contraceptive failure than the women in other age groups, and that women aged over 35 had the smallest rate of contraceptive failure. The difference in failure rate between women aged less than 24 and women aged between 30 to 34 was increasing until 1981, then began to decrease, reaching a minimum in 1986, before increasing again.

While a sophisticated analysis of relevant factors is hardly possible here, it is noticeable that before 1987 women with prior contraceptive use were less likely to fail, after 1987 they became more likely to do so.

The impact of motivation is interesting. There was little difference in the rate of contraceptive failure between the women self-motivated to use contraceptives and those that responsed to the campaigns before 1985, which suggests the campaigns to encourage women using contraceptive did have some effects on

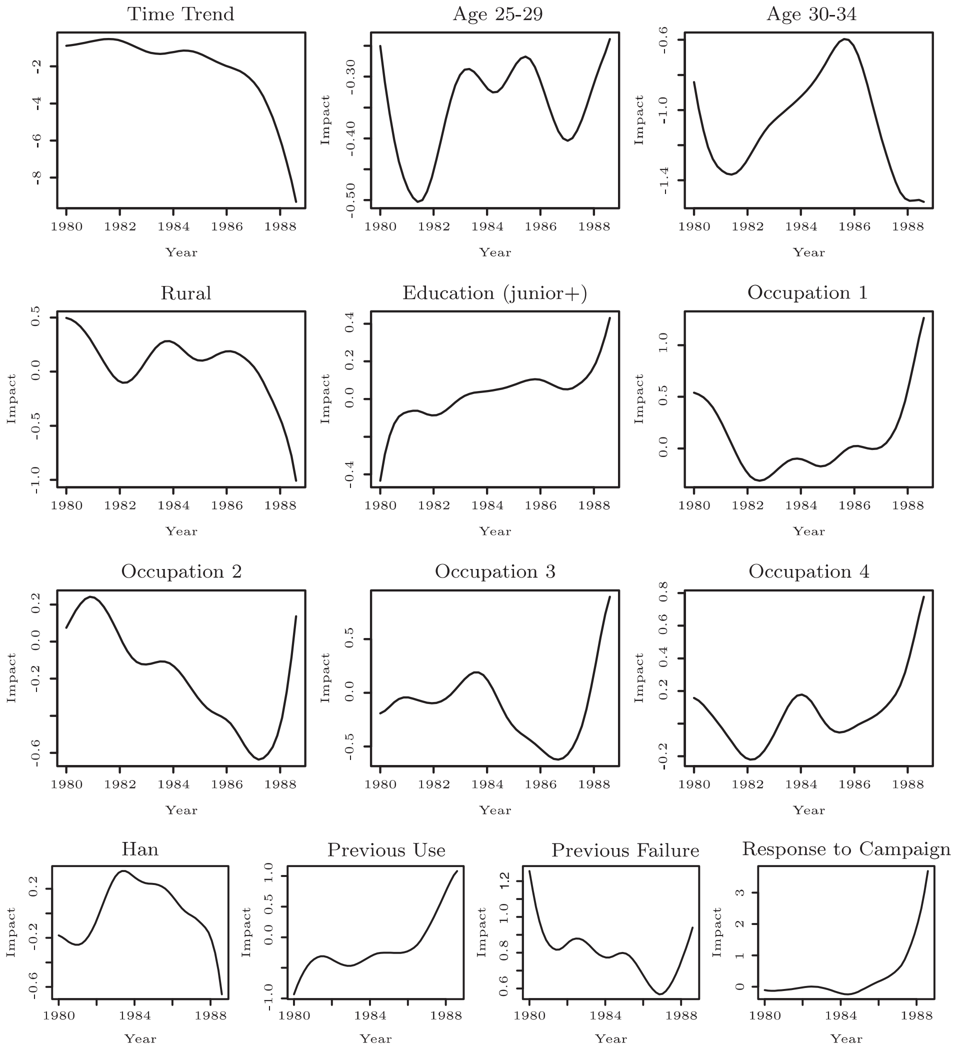

Figure 1. Occupations 1, 2, 3, and 4 represent industry, service, professional, and other non-agriculture, respectively.

women's attitude towards contraceptive use before 1985. However, the picture after 1985 is quite different. After 1985, the difference increased steadily, women using contraceptives in response to the campaigns were more and more likely to fail than self-motivated women during the later period.

Appendix

Throughout, we use $\mathbf{0}_{p\times q}$ to denote a matrix of size $p \times q$ with each entry 0, and $\mathbf{0}_p$ to denote $\mathbf{0}_{p\times p}$. Let $\mathbf{b}_i = \dot{\mathbf{a}}(U_i)$, $\ddot{\mathbf{a}}_i = \ddot{\mathbf{a}}(U_i)$.

The following technical conditions are to be imposed.

(1) The function $Q_2(s,\ y) < 0$ for $s \in R$ and y in the range of the response variable.

(2) $f(u)$ is positive on its support, $[0,\ 1]$; $f(u)$ and $\mathbf{a}(\cdot)$ have continuous second derivatives; the link function $g(\cdot)$ has a continuous third derivative.

(3) $E(XX^{\mathrm{T}}|U) > 0$ and $E(ZZ^{\mathrm{T}}|U) > 0$.

(4) The kernel function $K(\cdot)$ is a symmetric density function, and is absolutely continuous on its support set $[-A,\ A]$.

  (4a) $K(A) \neq 0$ or

  (4b) $K(A) = 0$, $K(t)$ is absolutely continuous, and $K^2(t)$ and $\dot{K}^2(t)$ are integrable on $(-\infty,\ +\infty)$.

  (4c) the function $t^3K(t)$ and $t^3\dot{K}(t)$ are bounded and $\int t^4\dot{K}(t) < \infty$.

(5) For some $s > 2$, $E(|X|^{2s}|U) < \infty$ is continuous and $E(Y^{2s}|U,\ X) < \infty$.

(6) $h = o(n^{-1/6})$, $nh \longrightarrow \infty$, $h_1 = O(n^{-1/4})$.

Lemma 1. *Let $(\xi_1,\ \eta_1), \ldots, (\xi_n,\ \eta_n)$ be i.i.d. random vectors, where the η_i's are scalars. Suppose $E|\eta_1|^s < \infty$ and $\sup_x \int |y|^s G(x,\ y)dy < \infty$, where G denotes the joint density of $(\xi_1,\ \eta_1)$, and let K be a bounded positive function with bounded support, satisfying a Lipschitz condition. Then*

$$\sup_{x\in D} |n^{-1}\sum_{i=1}^{n}\{K_h(\xi_i - x)\eta_i - E[K_h(\xi_i - x)\eta_i]\}| = O_P\left[\left\{\frac{nh}{\log(1/h)}\right\}^{-1/2}\right]$$

provided $n^{2\varepsilon-1}h \longrightarrow \infty$ for some $\varepsilon < 1 - s^{-1}$.

Proof. This follows immediately from a result in Mack and Silverman (1982).

Proof of Theorem 1. With

$$\Delta = n^{-1}\sum_{i=1}^{n}\Big(\ell\left[g^{-1}\left(X_i^{\mathrm{T}}\mathbf{a}_i + Z_i^{\mathrm{T}}\boldsymbol{\beta}\right),\ y_i\right] - \ell\left[g^{-1}\left\{X_i^{\mathrm{T}}\hat{\mathbf{a}}^{\backslash i}(U_i) + Z_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}^{\backslash i}\right\},\ y_i\right]\Big),$$

the cross-validation term can be written as

$$\mathrm{CV} = -n^{-1}\sum_{i=1}^{n}\ell\left[g^{-1}\left(X_i^{\mathrm{T}}\mathbf{a}_i + Z_i^{\mathrm{T}}\boldsymbol{\beta}\right),\ y_i\right] + \Delta.$$

With a Taylor expansion and Theorem 2 in Zhang and Peng (2010), we have

$$\Delta = n^{-1}\sum_{i=1}^{n}Q_1\left(X_i^{\mathrm{T}}\mathbf{a}_i + Z_i^{\mathrm{T}}\boldsymbol{\beta},\ y_i\right)\left[X_i^{\mathrm{T}}\left\{\mathbf{a}_i - \hat{\mathbf{a}}^{\backslash i}(U_i)\right\} + Z_i^{\mathrm{T}}\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{\backslash i}\right)\right]$$

$$
\begin{aligned}
&+(2n)^{-1}\sum_{i=1}^{n} Q_2\left(X_i^{\mathrm{T}}\mathbf{a}_i + Z_i^{\mathrm{T}}\boldsymbol{\beta},\ y_i\right)\left[X_i^{\mathrm{T}}\left\{\mathbf{a}_i - \hat{\mathbf{a}}^{\backslash i}(U_i)\right\} + Z_i^{\mathrm{T}}\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{\backslash i}\right)\right]^2\\
&+O_P\left\{(nh)^{-3/2}(-\log h)^{3/2}\right\}\\
&\overset{\triangle}{=} \Delta_1 + \Delta_2 + O_P\left\{(nh)^{-3/2}(-\log h)^{3/2}\right\}.
\end{aligned}
$$

Let $e_i = Q_1\left(X_i^{\mathrm{T}}\mathbf{a}_i + Z_i^{\mathrm{T}}\boldsymbol{\beta},\ y_i\right)$. By simple calculation, we have $E(\Delta_1) = 0$, and

$$
\begin{aligned}
\mathrm{Var}\,(\Delta_1) = n^{-2}&\sum_{i=1}^{n}\sum_{j=1}^{n} E\left\{e_i e_j Z_i^{\mathrm{T}}\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{\backslash i}\right) Z_j^{\mathrm{T}}\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{\backslash j}\right)\right\}\\
&+2n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n} E\left[e_i e_j X_i^{\mathrm{T}}\left\{\mathbf{a}_i - \hat{\mathbf{a}}^{\backslash i}(U_i)\right\} Z_j^{\mathrm{T}}\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{\backslash j}\right)\right]\\
&+n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n} E\left[e_i e_j X_i^{\mathrm{T}}\left\{\mathbf{a}_i - \hat{\mathbf{a}}^{\backslash i}(U_i)\right\} X_j^{\mathrm{T}}\left\{\mathbf{a}_j - \hat{\mathbf{a}}^{\backslash j}(U_i)\right\}\right]\\
&\overset{\triangle}{=} V_{1,1} + V_{1,2} + V_{1,3}.
\end{aligned}
$$

Applying a Taylor expansion to (2.2), and using the Theorems 1 and 2 in Zhang and Peng (2010), we have

$$
\begin{aligned}
&\mathbf{a}_i - \hat{\mathbf{a}}^{\backslash i}(U_i)\\
&\quad= (I_p,\ \mathbf{0}_p)H_i^{-1}\sum_{j\neq i} Q_1\left\{X_j^{\mathrm{T}}\mathbf{a}_i + X_j^{\mathrm{T}}\mathbf{b}_i(U_j - U_i) + Z_j^{\mathrm{T}}\boldsymbol{\beta},\ y_j\right\} K_h(U_j - U_i)\\
&\quad\quad\times\begin{pmatrix} X_j \\ (U_j - U_i)X_j \end{pmatrix}(1 + o_P(1)),
\end{aligned}
$$

where

$$
\begin{aligned}
H_i = &\sum_{j\neq i} Q_2\left\{X_j^{\mathrm{T}}\mathbf{a}_i + X_j^{\mathrm{T}}\mathbf{b}_i(U_j - U_i) + Z_j^{\mathrm{T}}\boldsymbol{\beta},\ y_j\right\} K_h(U_j - U_i)\begin{pmatrix} 1 & U_j - U_i \\ U_j - U_i & (U_j - U_i)^2 \end{pmatrix}\\
&\otimes X_j X_j^{\mathrm{T}}.
\end{aligned}
$$

By standard arguments and Lemma 1 we have, uniformly,

$$
H_i = nf(U_i)E\left\{Q_2\left(X^{\mathrm{T}}\mathbf{a}_i + Z^{\mathrm{T}}\boldsymbol{\beta},\ y\right)\mathrm{diag}(1,\ \mu_2)\otimes XX^{\mathrm{T}}|U = U_i\right\}(1 + o_P(1)).
$$

Let ζ_j be the ζ with $(U,\ X^{\mathrm{T}},\ Z^{\mathrm{T}},\ y)$ replaced by $(U_j,\ X_j^{\mathrm{T}},\ Z_j^{\mathrm{T}},\ y_j)$. With a Taylor expansion and Theorem 2 in Zhang and Peng (2010), we have

$$
\sum_{j\neq i} Q_1\left\{X_j^{\mathrm{T}}\mathbf{a}_i + X_j^{\mathrm{T}}\mathbf{b}_i(U_j - U_i) + Z_j^{\mathrm{T}}\boldsymbol{\beta},\ y_j\right\} K_h(U_j - U_i)\begin{pmatrix} X_j \\ (U_j - U_i)X_j \end{pmatrix}
$$

$$= \sum_{j\neq i} e_j K_h(U_j - U_i) \begin{pmatrix} X_j \\ (U_j - U_i)X_j \end{pmatrix}$$
$$-2^{-1} \sum_{j\neq i} \zeta_j X_j^{\mathrm{T}} \ddot{\mathbf{a}}_i K_h(U_j - U_i) \begin{pmatrix} (U_j - U_i)^2 X_j \\ (U_j - U_i)^3 X_j \end{pmatrix} (1 + o_P(1)).$$

Let Λ_i be the Λ with U being replaced by U_i. We have then, uniformaly,

$$\mathbf{a}_i - \hat{\mathbf{a}}^{\backslash i}(U_i) = \{nf(U_i)\}^{-1}\Lambda_i^{-1} \sum_{j\neq i} e_j K_h(U_j - U_i) X_j (1 + o_P(1))$$
$$-\{2nf(U_i)\}^{-1}\Lambda_i^{-1} \sum_{j\neq i} \zeta_j X_j^{\mathrm{T}} \ddot{\mathbf{a}}_i (U_j - U_i)^2 X_j K_h(U_j - U_i)(1 + o_P(1)).$$

By a similar, more tedious calculation, we have

$$\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{\backslash i}$$
$$= n^{-2} \sum_{k\neq i} f(U_k)^{-1}\Gamma_k \sum_{j\neq i} e_j K_{h_1}(U_j - U_k) \begin{pmatrix} X_j \\ Z_j \end{pmatrix} (1 + o_P(1))$$
$$-2^{-1}n^{-2} \sum_{k\neq i} f(U_k)^{-1}\Gamma_k \sum_{j\neq i} \zeta_j X_j^{\mathrm{T}} \ddot{\mathbf{a}}_k (U_j - U_k)^2 \begin{pmatrix} X_j \\ Z_j \end{pmatrix} K_{h_1}(U_j - U_k)(1 + o_P(1))$$
$$= n^{-1} \sum_{j\neq i} e_j G_j (1 + o_P(1)) - 2^{-1}n^{-1} \sum_{j\neq i} \mu_2 \zeta_j X_j^{\mathrm{T}} \ddot{\mathbf{a}}_j h_1^2 G_j (1 + o_P(1)),$$

where

$$\Gamma_j = (\Gamma_{1,j},\ \Gamma_{2,j}), \quad G_j = (\Gamma_{1,j} X_j + \Gamma_{2,j} Z_j), \quad \Gamma_{1,j} = -\Gamma_{2,j} D_{1,2,j}^{\mathrm{T}} \Lambda_j^{-1},$$
$$\Gamma_{2,j} = \left(D_{2,2,j} - D_{1,2,j}^{\mathrm{T}} \Lambda_j^{-1} D_{1,2,j} \right)^{-1},$$
$$D_{1,2,j} = E\left\{ Q_2\left(X^{\mathrm{T}}\mathbf{a}_j + Z^{\mathrm{T}}\boldsymbol{\beta},\ y \right) X Z^{\mathrm{T}} | U = U_j \right\},$$
$$D_{2,2,j} = E\left\{ Q_2\left(X^{\mathrm{T}}\mathbf{a}_j + Z^{\mathrm{T}}\boldsymbol{\beta},\ y \right) Z Z^{\mathrm{T}} | U = U_j \right\}.$$

Calculation, Theorem 1 in Zhang and Peng (2010), and condition (6) yield

$$V_{1,1} = 2n^{-2} \sum_{i=1}^{n} \sum_{j=i+1}^{n} E\left\{ e_i e_j Z_i^{\mathrm{T}} \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{\backslash i} \right) Z_j^{\mathrm{T}} \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{\backslash j} \right) \right\} + O_P(n^{-2})$$
$$= 2n^{-4} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \sum_{k\neq i} \sum_{l\neq j} E\left\{ e_i e_j e_k e_l Z_i^{\mathrm{T}} G_k Z_j^{\mathrm{T}} G_l \right\}$$
$$+2^{-1}n^{-4}\mu_2^2 h_1^4 \sum_{i=1}^{n} \sum_{j=i+1}^{n} \sum_{k\neq i} \sum_{l\neq j} E\left\{ e_i e_j \zeta_k \zeta_l X_k^{\mathrm{T}} \ddot{\mathbf{a}}_k X_l^{\mathrm{T}} \ddot{\mathbf{a}}_l Z_i^{\mathrm{T}} G_k Z_j^{\mathrm{T}} G_l \right\}$$

$$-2n^{-4}\mu_2 h_1^2 \sum_{i=1}^{n} \sum_{j=i+1}^{n} \sum_{k\neq i} \sum_{l\neq j} E\left\{e_i e_j e_k \zeta_l X_l^{\mathrm{T}} \ddot{\mathbf{a}}_l Z_i^{\mathrm{T}} G_k Z_j^{\mathrm{T}} G_l\right\} + O_P(n^{-2})$$

$$= O_P(n^{-2}).$$

Similarly, $V_{1,2} = O_P(n^{-2}h^{-1/2} + n^{-3/2}h^2)$, and $V_{1,3} = O_P(n^{-2}h^{-1} + n^{-1}h^4)$. Thus, $\Delta_1 = o_P(n^{-1}h^{-1})$.

We can write

$$\begin{aligned}\Delta_2 &= (2n)^{-1}\sum_{i=1}^{n}\zeta_i\left[X_i^{\mathrm{T}}\left\{\mathbf{a}_i - \hat{\mathbf{a}}^{\backslash i}(U_i)\right\}\right]^2\\ &\quad + n^{-1}\sum_{i=1}^{n}\zeta_i X_i^{\mathrm{T}}\left\{\mathbf{a}_i - \hat{\mathbf{a}}^{\backslash i}(U_i)\right\} Z_i^{\mathrm{T}}\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{\backslash i}\right) + (2n)^{-1}\sum_{i=1}^{n}\zeta_i\left\{Z_i^{\mathrm{T}}\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{\backslash i}\right)\right\}^2\\ &\quad \triangleq \Delta_{2,1} + \Delta_{2,2} + \Delta_{2,3}.\end{aligned}$$

By Theorems 1 and 2 in Zhang and Peng (2010), $\Delta_{2,2} = O_P(n^{-1}h^{-1/2}\log n) = o_P(n^{-1}h^{-1})$, and $\Delta_{2,3} = O_P(n^{-1}) = o_P(n^{-1}h^{-1})$. Obviously,

$$\begin{aligned}\Delta_{2,1} &= 2^{-1}n^{-3}\sum_{i=1}^{n} f(U_i)^{-2}\zeta_i\left\{X_i^{\mathrm{T}}\Lambda_i^{-1}\sum_{j\neq i} e_j K_h(U_j - U_i)X_j\right\}^2 (1 + o_P(1))\\ &\quad -2^{-1}n^{-3}\sum_{i=1}^{n}\left\{f(U_i)^{-2}\zeta_i X_i^{\mathrm{T}}\Lambda_i^{-1}\sum_{j\neq i} e_j K_h(U_j - U_i)X_j\right.\\ &\qquad \left.\times X_i^{\mathrm{T}}\Lambda_i^{-1}\sum_{l\neq i}\zeta_l X_l^{\mathrm{T}}\ddot{\mathbf{a}}_i(U_l - U_i)^2 X_l K_h(U_l - U_i)\right\}(1 + o_P(1))\\ &\quad + (2n)^{-3}\sum_{i=1}^{n} f(U_i)^{-2}\zeta_i\left\{X_i^{\mathrm{T}}\Lambda_i^{-1}\sum_{j\neq i}\zeta_j X_j^{\mathrm{T}}\ddot{\mathbf{a}}_i(U_j - U_i)^2 X_j K_h(U_j - U_i)\right\}^2\\ &\quad \times(1 + o_P(1))\\ &\quad \triangleq \Delta_{2,1,1} - \Delta_{2,1,2} + \Delta_{2,1,3}.\end{aligned}$$

By standard arguments, we have

$$\begin{aligned}\Delta_{2,1,1} &= 2^{-1}n^{-3}\sum_{i=1}^{n} f(U_i)^{-2}\zeta_i\sum_{j\neq i} e_j^2 K_h^2(U_j - U_i)\left(X_i^{\mathrm{T}}\Lambda_i^{-1}X_j\right)^2 (1 + o_P(1))\\ &= -2^{-1}\nu_0(nh)^{-1}E\left\{f(U)^{-1}\zeta X^{\mathrm{T}}\Lambda^{-1}X\right\}(1 + o(1)).\end{aligned}$$

It is easy to see

$$\Delta_{2,1,2} = 2^{-1}n^{-3}\sum_{j=1}^{n} e_j \sum_{i\neq j}\sum_{l\neq i}\big\{f(U_i)^{-2}\zeta_i X_i^{\mathrm{T}}\Lambda_i^{-1}X_j \\ \times\zeta_l X_i^{\mathrm{T}}\Lambda_i^{-1}X_l X_l^{\mathrm{T}}\ddot{\mathbf{a}}_i(U_l-U_i)^2K_h(U_l-U_i)K_h(U_j-U_i)\big\}\,(1+o_P(1)).$$

By tedious calculation and Lemma 1, $E(\Delta_{2,1,2}) = O(n^{-1}h)$, and $\mathrm{Var}\,(\Delta_{2,1,2}) = O(n^{-1}h^4)$, $\Delta_{2,1,2} = o_P(n^{-1}h^{-1})$.

By standard arguments, from

$$\Delta_{2,1,3} = (2n)^{-3}\sum_{i=1}^{n} f(U_i)^{-2}\zeta_i\left\{X_i^{\mathrm{T}}\Lambda_i^{-1}\sum_{j\neq i}\zeta_j X_j^{\mathrm{T}}\ddot{\mathbf{a}}_i(U_j-U_i)^2X_jK_h(U_j-U_i)\right\}^2 \\ \times(1+o_P(1)) \\ = 2^{-3}\mu_2^2h^4n^{-1}\sum_{i=1}^{n}\zeta_i\left[X_i^{\mathrm{T}}\Lambda_i^{-1}E\left\{\zeta X^{\mathrm{T}}\ddot{\mathbf{a}}(U)X|U=U_i\right\}\right]^2(1+o_P(1)) \\ = 2^{-3}\mu_2^2h^4E\left(\zeta\left[X^{\mathrm{T}}\Lambda^{-1}E\left\{\zeta X^{\mathrm{T}}\ddot{\mathbf{a}}(U)X|U\right\}\right]^2\right)(1+o(1)) \\ = 2^{-3}\mu_2^2h^4E\left[\zeta\left\{X^{\mathrm{T}}\ddot{\mathbf{a}}(U)\right\}^2\right](1+o(1)),$$

$$\mathrm{CV} = -n^{-1}\sum_{i=1}^{n}\ell\left[g^{-1}\left(X_i^{\mathrm{T}}\mathbf{a}_i+Z_i^{\mathrm{T}}\boldsymbol{\beta}\right),\ y_i\right]+2^{-3}\mu_2^2h^4E\left[\zeta\left\{X^{\mathrm{T}}\ddot{\mathbf{a}}(U)\right\}^2\right] \\ -2^{-1}\nu_0(nh)^{-1}E\left\{f(U)^{-1}\zeta X^{\mathrm{T}}\Lambda^{-1}X\right\}+o_P(n^{-1}h^{-1}+h^4).$$

Proof of Theorem 2. Theorem 2 follows straightforwardly from Theorem 1.

Proof of Theorem 3. If $\mathcal{S}_k$, $k=1,\ldots,2^{p+q}-1$, are the incorrect models, we have

$$\{\mathcal{S}_*\neq\mathcal{S}_0\} = \bigcup_{k=1}^{2^{p+q}-1}\{\mathrm{CV}(\mathcal{S}_k)<\mathrm{CV}(\mathcal{S}_0)\},$$

where $\mathrm{CV}(\mathcal{S}_k)$ is the CV of model $\mathcal{S}_k$. From the remark following Theorem 2, we have

$$P\{\mathrm{CV}(\mathcal{S}_k)<\mathrm{CV}(\mathcal{S}_0)\}\le P\left\{\left|o_P(n^{-1}h^{-1})\right|>C_0n^{-1}h^{-1}\right\}\longrightarrow 0,$$

where C_0 is a positive constant. This leads to

$$P(\mathcal{S}_*\neq\mathcal{S}_0)\le\sum_{k=1}^{2^{p+q}-1}P\{\mathrm{CV}(\mathcal{S}_k)<\mathrm{CV}(\mathcal{S}_0)\}\longrightarrow 0,$$

which implies $P(\mathcal{S}_* = \mathcal{S}_0) \longrightarrow 1$.

Acknowledgement

This research is supported by the GRF grant of Hong Kong Research Grant Council (Grant No. HKBU201610).

References

Fan, J. and Huang, T. (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli.* **11**, 1031-1057.

Fan, J., Huang, T., and Li, R. (2007). Analysis of longitudinal data with semiparametric estimation of covariance function. *J. Amer. Statist. Assoc.* **35**, 632-641.

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

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

Li, R. and Liang, H. (2008). Variable selection in semiparametric regression modeling. *Ann. Statist.*. **36**, 261-286.

Ma, Y., Chiou, J.-M. and Wang, N. (2006). Efficient semiparametric estimator for heteroscedastic partially-linear models, *Biometrika* **93**, 75-84.

Mack, Y. P. and Silverman, B. W. (1982). Weak and strong uniform consistency of kernel regression estimates. *Z. Wahrsch. Verw. Gebiete* **61**, 405-415.

Qu, A. and Li, R. (2006). Quadratic inference functions for varying-coefficient models with longitudinal data, *Biometrics* **62**, 379-391.

Sentürk, D. and Müller, H. G. (2006). Inference for covariate adjusted regression via varying coefficient models, *Ann. Statist.* **34**, 654-679.

Sun, Y., Zhang, W. and Tong, H. (2007). Estimation of the covariance matrix of random effects in longitudinal studies. *Ann. Statist.* **35**, 2795-2814.

Wang, H. and Xia, Y. (2009). Regularized estimation of the varying coefficient model. *J. Amer. Statist. Assoc.* **104**, 747-757.

Wang, L., Kai, B. and Li, R. (2009). Local rank inference for varying coefficient models. *J. Amer. Statist. Assoc.* **104**, 1631-1645.

Wang, L., Li, H. and Huang, J. Z. (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *J. Amer. Statist. Assoc.* **103**, 1556-1569.

Xia, Y. and Li, W. K. (1999). On the estimation and testing of functional-coefficient linear models. *Statist. Sinica* **9**, 735-58.

Zhang, W., Lee, S. Y. and Song, X. (2002). Local polynomial fitting in semivarying coefficient models. *J. Multivariate Anal.* **82**, 166-188.

Zhang, W. and Peng, H. (2010). Simultaneous confidence band and hypothesis test in generalised varying-coefficient models. *J. Multivariate Anal.* **101**, 1656-1680.

Department of Mathematical Sciences, University of Bath, Bath BA2 7AY, UK.

E-mail: W.Zhang@bath.ac.uk

(Received August 2009; accepted May 2010)