Ann Inst Stat Math (2012) 64:835–856
DOI 10.1007/s10463-011-0342-9

Estimation and model selection in a class of semiparametric models for cluster data

Yan Sun · Jialiang Li · Wenyang Zhang

Received: 27 August 2010 / Revised: 6 June 2011 / Published online: 19 November 2011

Abstract Stimulated by a study in Bangladesh about the first birth interval, we propose a semivarying-coefficient model for cluster data analysis. We consider the estimation procedure for the proposed model and establish the asymptotic results of the proposed estimators. Furthermore, we employ the cross-validation (CV) to identify the constant coefficients. The associated asymptotic properties are rigorously examined. Simulation studies are conducted to investigate the performance of the proposed estimation and the CV-based model selection procedure for finite sample size. Finally, our methods are used to analyse the aforementioned data set to explore how several factors affect the first birth interval in Bangladesh.

Keywords Cluster data · Cross-validation · Local linear modelling · Semiparametric inference · Varying-coefficient models

1 Introduction

In a typical analysis for cluster data, researchers usually assume all clusters share the same regression function. The difference across clusters is accounted for by the within-cluster correlation modelled by a within-cluster covariance matrix. Many researchers

Y. Sun
School of Economics, Shanghai University of Finance and Economics, Shanghai 200433, China

J. Li (✉)
Department of Statistics and Applied Probability, National University of Singapore,
6 Science Drive 2, Singapore 117546, Singapore
e-mail: stalj@nus.edu.sg

W. Zhang
Department of Mathematical Sciences, The University of York, Heslington, York YO10 5DD, UK

have studied how to incorporate the within-cluster covariance matrix into the estimation procedure to improve the estimation of the regression function or how to accurately estimate the within-cluster covariance matrix (see, Chiou and Müller 2005; Fan and Li 2004; Fan et al. 2007; Fan and Wu 2008; Sun et al. 2007).

The above approach is essentially to account for the difference across clusters by random effects only. Whilst this kind of modelling is quite successful in cluster data analysis, sometimes, it is inappropriate to believe all clusters share the same regression function. To elaborate this further, let us consider a data set that stimulates this paper. The data set is from the Bangladesh Demographic and Health Survey 1996–1997. This survey follows a two-stage design in which clusters were selected at the first stage, and women were sampled from these clusters at the second stage. The clusters correspond to villages in rural areas and neighbourhoods in urban areas and may loosely be termed communities. What is of interest is how the factors which are commonly found to be associated with fertility in Bangladesh affect the first birth interval. The selected factors in this study are (1) woman's level of education; (2) type of region of residence; (3) woman's religion; (4) year of marriage; and (5) administrative area. Among these factors, type of region of residence and administrative area pertain to cluster levels and are called cluster-level variables, and the rest are called individual level variables.

We use y to denote the length of the first birth interval, X the vector of individual level variables, Z the vector of cluster-level variables. For $j = 1, \ldots, n_i, i = 1, \ldots, m$, let y_{ij} and X_{ij} be the jth observation of y and X in the ith cluster, Z_i the observation of Z at the ith cluster. If we use a linear model to fit the data and random effects only to account for the difference across clusters, it would imply that all clusters share the same coefficients in the regression function, which is equivalent to assuming the impacts of the factors concerned are the same for all clusters. Apparently, this is not very convincing. For example, it is evident that the impact of education in the cluster where Muslims predominate would be different to that in the cluster where Hindus predominate. There must be some deterministic effects taking part in the difference across clusters. We have to take such effects into account.

Simply allowing regression coefficients to vary over clusters, we would have

$$y_{ij} = X_{ij}^{\mathrm{T}}\mathbf{a}_i + Z_i^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_{ij}, \qquad j = 1, \ldots, n_i, \quad i = 1, \ldots, m, \tag{1}$$

where ϵ_{ij}'s are random errors with mean zero and variance σ^2. While it accounts for the dynamic of the impacts across clusters, (1) is not parsimonious. A sensible approach is to model the impacts $\mathbf{a}_i$ by the cluster-level variables.

$$\begin{cases} y_{ij} = X_{ij}^{\mathrm{T}}\mathbf{a}_i + Z_i^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_{ij}, & j = 1, \ldots, n_i, \quad i = 1, \ldots, m, \\ \mathbf{a}_i = \boldsymbol{\alpha}_0 + \mathbf{A}Z_i + \mathbf{e}_i, & i = 1, \ldots, m, \end{cases} \tag{2}$$

where $\mathbf{A} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_q)$, and $\mathbf{e}_i$'s are random effects with mean zero and covariance Σ. $\mathbf{e}_i$ and ϵ_{ij} are independent. This achieves the parsimony, and the within-cluster dependency is also accounted for by both random effects and deterministic pattern. In fact, the number of unknown coefficients in model (2) is $p(q + 1) + q$ which is usually much smaller than $pm + q$.

To incorporate the time effect on the impacts into modelling and take into account that some impacts may not change over time leads to the following semivarying coefficient models

$$\begin{cases} y_{ij} = X_{ij}^{\mathrm{T}}\mathbf{a}_i(U_{ij}) + T_i^{\mathrm{T}}\boldsymbol{\beta}_1 + V_i^{\mathrm{T}}\boldsymbol{\beta}_2(U_{ij}) + \epsilon_{ij}, \\ \mathbf{a}_i(U_{ij}) = \mathbf{A}(U_{ij})Z_i + \mathbf{e}_i, \end{cases} \tag{3}$$

$j = 1, \ldots, n_i, i = 1, \ldots, m$, where T_i is the vector of the cluster-level variables whose impacts do not change over time, we assume it is k_1 dimensional. V_i is the vector of the cluster-level variables whose impacts change over time; we assume it is k_2 dimensional. $Z_i = (T_i^{\mathrm{T}}, V_i^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\beta}_1$ is a k_1-dimensional unknown constant vector, and $\boldsymbol{\beta}_2(\cdot)$ is a k_2 dimensional unknown functional vector. $\mathbf{A}(U_{ij})$ is a $p \times q$ unknown matrix. Because the impacts of some individual level variables do not change over time, some entries of matrix $\mathbf{A}(U_{ij})$ may be constant. Without loss of generality, we assume $\mathbf{A}(U_{ij}) = \begin{pmatrix} A_1 & A_2(U_{ij}) \\ A_3(U_{ij}) & A_4(U_{ij}) \end{pmatrix}$, where A_1 is a $p_1 \times q_1$ unknown constant matrix, $A_2(\cdot)$, $A_3(\cdot)$ and $A_4(\cdot)$ are $p_1 \times q_2$, $p_2 \times q_1$ and $p_2 \times q_2$ unknown functional matrix, respectively. ϵ_{ij}'s are independent of X_{ij}, U_{ij} and Z_i while $\mathbf{e}_i$'s are independent of ϵ_{ij}, X_{ij}, U_{ij} and Z_i. We assume $(X_{ij}^{\mathrm{T}}, U_{ij})^{\mathrm{T}}$ are i.i.d. for all i and j, Z_i are i.i.d. for all i.

As a class of semiparametric models, (3) includes many important models. For example, the semivarying coefficient models (Fan and Zhang 1999, 2000a,b; Zhang et al. 2002, 2009; Wang et al. 2009) are a special case of (3) with $Z_i = 1$ and $\mathbf{e}_i = 0$; the functional mixed-effect models of Sun et al. (2007) are (3) with all constant coefficients being zero; and the well-known growth curve models (Demidenko 2004), Ch. 4) are also (3) with all coefficients being constant.

As (3) involves both functional and constant coefficients, in reality, we have to identify which coefficients are functional, which are constant when using model (3). In this paper, we will systemically investigate how to use the cross-validation (CV) to identify the constant coefficients and examine how powerful the CV is on this effect. It is worth noticing that Xia et al. (2004) and Li and Zhang (2011) applied the CV to identify the constant components in similar semivarying coefficient models.

2 Estimation procedure

We first estimate the covariance matrix Σ of the random effect $\mathbf{e}_i$ and the variance σ^2 of the random error ϵ_{ij}. This is because the estimators of Σ and σ^2 will be used when constructing the estimators of the constant coefficients.

2.1 Estimation of σ^2 and Σ

We treat all coefficients (regardless of constant or functional) in model (3) as functional when estimating σ^2 and Σ. The reasons for us to do so are (1) in reality, we do not know which coefficients are constant. If we mistakenly treat a functional coefficient as constant, the estimators of σ^2 and Σ would be very poor and they are not even

consistent. Although to treat the constant coefficients as functional may cost on the variance side of the estimators of the constant coefficients, it would have little effect on the estimators of σ^2 and Σ as the loss on the variance side of the estimators of the constant coefficients will be eliminated in the estimation procedure of σ^2 and Σ. (2) When using the CV to identify the constant coefficients, we need to compute the CV, which involves the estimators of σ^2 and Σ, for each candidate model. To treat all coefficients as functional will enable us to use the same estimators of σ^2 and Σ for all candidate models without paying any price. This will largely reduce the computation involved.

By the Taylor's expansion we have $A_k(U_{ij}) \approx A_k(u) + \dot{A}_k(u)(U_{ij} - u), k = 1, 2, 3, 4$, $\boldsymbol{\beta}_l(U_{ij}) \approx \boldsymbol{\beta}_l(u) + \dot{\boldsymbol{\beta}}_l(u)(U_{ij} - u), l = 1, 2$, when U_{ij} is in a small neighbourhood of u, which leads to the following local least squares estimation procedure:

$$L = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - X_{ij}^{\mathrm{T}}[\{B + C(U_{ij} - u)\}Z_i] - T_i^{\mathrm{T}}\{\mathbf{c}_1 + \mathbf{d}_1(U_{ij} - u)\} - V_i^{\mathrm{T}}\{\mathbf{c}_2 + \mathbf{d}_2(U_{ij} - u)\})^2 K_h(U_{ij} - u), \tag{4}$$

where $B = \begin{pmatrix} B_1 & B_2 \\ B_3 & B_4 \end{pmatrix}, C = \begin{pmatrix} C_1 & C_2 \\ C_3 & C_4 \end{pmatrix}$, $K_h(\cdot) = K(\cdot/h)/h$, h is bandwidth and $K(\cdot)$ is a kernel function.

We minimise L with respect to $B_i, C_i, i = 1, 2, 3, 4$, and $\mathbf{c}_l, \mathbf{d}_l, l = 1, 2$, to get the minimiser. The initial estimator $A_i^*(u)$ of $A_i(u)$ is the part of the minimiser corresponding to B_i, and the initial estimator $\tilde{\boldsymbol{\beta}}_l(u)$ is the part of the minimiser corresponding to $\mathbf{c}_l, l = 1, 2$.

Let $\mathbf{x}_i = (X_{i1}, \ldots, X_{in_i})^{\mathrm{T}}$, $R_i = (Z_i^{\mathrm{T}} \otimes \mathbf{x}_i, \mathbf{1}_{n_i} \otimes Z_i^{\mathrm{T}})$, $R = (R_1^{\mathrm{T}}, \ldots, R_m^{\mathrm{T}})^{\mathrm{T}}, \mathcal{U}_i = \mathrm{diag}((U_{11} - u)^i, \ldots, (U_{1n_1} - u)^i, \ldots, (U_{m1} - u)^i, \ldots, (U_{mn_m} - u)^i)$, $W = \mathrm{diag}(K_h(U_{11} - u), \ldots, K_h(U_{1n_1} - u), \ldots, K_h(U_{m1} - u), \ldots, K_h(U_{mn_m} - u))Y = (y_{11}, \ldots, y_{1n_1}, \ldots, y_{m1}, \ldots, y_{mn_m})$, $\mathbf{X} = (R, \mathcal{U}_1 R)$, where $\mathbf{1}_d$ is a d dimensional vector with each component being 1. Further, let I_p be a size p identity matrix, $\mathbf{0}_{p\times q}$ be a size $p \times q$ matrix with all entries 0, and $H_1 = (I_{p_1}, \mathbf{0}_{p_1\times p_2})$, $H_2 = (\mathbf{0}_{p_2\times p_1}, I_{p_2})$, $H_3 = \begin{pmatrix} I_{q_1} \\ \mathbf{0}_{q_2\times q_1} \end{pmatrix}$, $H_4 = \begin{pmatrix} \mathbf{0}_{q_1\times q_2} \\ I_{q_2} \end{pmatrix}$, $n = \sum_{i=1}^m n_i, l_1 = (p+1)q + q, l_2 = (p+1)q + k_2$, $D_i = (\mathbf{0}_{p\times((i-1)p)}, I_p, \mathbf{0}_{p\times((q-i)p)})$, $D = (D_1, \ldots, D_q)$. By simple calculation, we have

$$\begin{cases} A_1^*(u) = H_1\mathcal{T}H_3,\ A_2^*(u) = H_1\mathcal{T}H_4,\ A_3^*(u) = H_2\mathcal{T}H_3, \quad A_4^*(u) = H_2\mathcal{T}H_4, \\ \boldsymbol{\beta}_1^*(u) = (\mathbf{0}_{k_1\times(pq)}, I_{k_1}, \mathbf{0}_{k_1\times l_2})(\mathbf{X}^{\mathrm{T}}W\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}WY, \\ \boldsymbol{\beta}_2^*(u) = (\mathbf{0}_{k_2\times(pq+k_1)}, I_{k_2}, \mathbf{0}_{k_2\times(pq+q)})(\mathbf{X}^{\mathrm{T}}W\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}WY, \end{cases} \tag{5}$$

where $\mathcal{T} = D(I_q \otimes \{(I_{pq}, \mathbf{0}_{(pq)\times l_1})(\mathbf{X}^{\mathrm{T}}W\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}WY\})$.

Let $\mathbf{A}^*(U_{ij})$ be the $\mathbf{A}(U_{ij})$ with A_i being replaced by $A_i^*(U_{ij}) i = 1, 2, 3, 4$, $\mathbf{a}_i^*(U_{ij}) = \mathbf{A}(U_{ij})Z_i$, $\hat{\mathbf{a}}_i(U_{ij}) = \mathbf{A}^*(U_{ij})Z_i$, and $\mathbf{r}_i = (r_{i1}, \ldots, r_{in_i})^{\mathrm{T}}, r_{ij} = y_{ij} - X_{ij}^{\mathrm{T}}\mathbf{a}_i^*(U_{ij}) - T_i^{\mathrm{T}}\boldsymbol{\beta}_1 - V_i^{\mathrm{T}}\boldsymbol{\beta}_2(U_{ij})$, $\hat{\mathbf{r}}_i = (\hat{r}_{i1}, \ldots, \hat{r}_{in_i})^{\mathrm{T}}, \hat{r}_{ij} = y_{ij} - X_{ij}^{\mathrm{T}}\hat{\mathbf{a}}_i(U_{ij}) - T_i^{\mathrm{T}}\boldsymbol{\beta}_1^*(U_{ij}) - V_i^{\mathrm{T}}\boldsymbol{\beta}_2^*(U_{ij})$, $\mathbf{x}_i = (X_{i1}, \ldots, X_{in_i})^{\mathrm{T}}$, $P_i = \mathbf{x}_i(\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_i)^{-1}\mathbf{x}_i^{\mathrm{T}}$. From model (3), we have the following synthetic linear model:

$$\mathbf{r}_i = \mathbf{x}_i \mathbf{e}_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{in_i})^{\mathrm{T}}. \tag{6}$$

The residual sum of squares of this model $\mathrm{rss}_i = \mathbf{r}_i^{\mathrm{T}}(I_{n_i} - P_i)\mathbf{r}_i$ would be the raw material for estimating σ^2. The synthetic degree of freedom of rss_i is $n_i - p$ since (6) may be regarded as a linear regression model with n_i observations and p predictors. Let RSS_i be rss_i with $\mathbf{r}_i$ replaced by $\hat{\mathbf{r}}_i$. RSS_i is a natural estimator for rss_i. Pooling all $\mathrm{RSS}_i, i = 1, \ldots, m$, together naturally leads to the estimator of σ^2 as $\hat{\sigma}^2 = (n - mp)^{-1} \sum_{i=1}^m \mathrm{RSS}_i$. Finally, we estimate Σ. From (6), we have the least squares estimator of $\mathbf{e}_i$ as $\mathbf{e}_i^* = (\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_i)^{-1}\mathbf{x}_i^{\mathrm{T}}\mathbf{r}_i = \mathbf{e}_i + (\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_i)^{-1}\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\epsilon}_i$. It is easy to see

$$\begin{aligned}\sum_{i=1}^m \mathbf{e}_i^* \tilde{\mathbf{e}}_i^{*\mathrm{T}} = \sum_{i=1}^m \mathbf{e}_i \mathbf{e}_i^{\mathrm{T}} + \sum_{i=1}^m (\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_i)^{-1}\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^{\mathrm{T}}\mathbf{x}_i(\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_i)^{-1} + \sum_{i=1}^m (\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_i)^{-1}\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\epsilon}_i\mathbf{e}_i^{\mathrm{T}} \\ + \sum_{i=1}^m \mathbf{e}_i\boldsymbol{\epsilon}_i^{\mathrm{T}}\mathbf{x}_i(\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_i)^{-1}.\end{aligned} \tag{7}$$

The last two terms are of order $O_P(m^{1/2})$ and negligible. This leads to

$$m^{-1}\sum_{i=1}^m \mathbf{e}_i\mathbf{e}_i^{\mathrm{T}} \approx m^{-1}\left\{\sum_{i=1}^m \mathbf{e}_i^*\mathbf{e}_i^{*\mathrm{T}} - \sum_{i=1}^m (\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_i)^{-1}\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^{\mathrm{T}}\mathbf{x}_i(\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_i)^{-1}\right\} \tag{8}$$

$$\approx m^{-1}\left\{\sum_{i=1}^m \mathbf{e}_i^*\mathbf{e}_i^{*\mathrm{T}} - \sigma^2\sum_{i=1}^m (\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_i)^{-1}\right\}. \tag{9}$$

Let $\hat{\mathbf{e}}_i$ be $\mathbf{e}_i^*$ with $\mathbf{r}_i$ replaced by $\hat{\mathbf{r}}_i$. We obtain

$$\hat{\Sigma} = m^{-1}\sum_{i=1}^m \hat{\mathbf{e}}_i\hat{\mathbf{e}}_i^{\mathrm{T}} - m^{-1}\hat{\sigma}^2\sum_{i=1}^m (\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_i)^{-1} \tag{10}$$

2.2 Final estimation for the coefficients

We first estimate constant coefficients following the profile likelihood idea (Fan and Huang 2005). By simple calculation, (3) can be written to

$$\begin{cases} y_{ij}^* = X_{ij}^{\mathrm{T}}\mathbf{a}_i^c(U_{ij}) + V_i^{\mathrm{T}}\boldsymbol{\beta}_2(U_{ij}) + \epsilon_{ij}, \\ \mathbf{a}_i^c(U_{ij}) = \mathbf{A}^c(U_{ij})Z_i + \mathbf{e}_i, \end{cases} \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, m, \tag{11}$$

where $y_{ij}^* = y_{ij} - X_{ij1}^{\mathrm{T}}A_1Z_{i1} - T_i^{\mathrm{T}}\boldsymbol{\beta}_1$, $\mathbf{A}^c(U_{ij}) = \begin{pmatrix} \mathbf{0}_{p_1\times q_1} & A_2(U_{ij}) \\ A_3(U_{ij}) & A_4(U_{ij}) \end{pmatrix}$, X_{ij1} is the vector of the first p_1 components of X_{ij}. Pretending both A_1 and $\boldsymbol{\beta}_1$ are known, we apply the local linear modelling to model (11). As the detail is almost the same as Sect. 2.1, we present the following results without detailed derivation.

For any matrix M, let $\mathrm{vec}(M)$ be the vector by simply stacking the column vectors of matrix M below one another. Moreover, let $\mathbf{x}_{i1}$ be the matrix of the first p_1

columns of $\mathbf{x}_i$, Z_{i1} the first q_1 components of Z_i, $\mathbf{x}_i = (\mathbf{x}_{i1}, \mathbf{x}_{i2})$, $Z_i = (Z_{i1}^{\mathrm{T}}, Z_{i2}^{\mathrm{T}})^{\mathrm{T}}$, $G_i = (Z_{i1}^{\mathrm{T}} \otimes \mathbf{x}_{i2}, Z_{i2}^{\mathrm{T}} \otimes \mathbf{x}_i, \mathbf{1}_{n_i} \otimes V_i^{\mathrm{T}})$, $d = p_2q_1 + pq_2 + k_2\boldsymbol{\theta}(u) = ((\mathrm{vec}(A_3(u)))^{\mathrm{T}}, (\mathrm{vec}((A_2^{\mathrm{T}}(u), A_4^{\mathrm{T}}(u))^{\mathrm{T}}))^{\mathrm{T}}, \boldsymbol{\beta}_2(u)^{\mathrm{T}})^{\mathrm{T}}$. The estimator of $\boldsymbol{\theta}(u)$ is

$$\boldsymbol{\theta}^*(u) = (I_d, \mathbf{0}_{d\times d})(\Delta^{\mathrm{T}} W \Delta)^{-1} \Delta^{\mathrm{T}} W Y^*, \tag{12}$$

where $\Delta = (G, \mathcal{U}_1 G)$, $G = (G_1^{\mathrm{T}}, \ldots, G_m^{\mathrm{T}})^{\mathrm{T}}$, $Y^* = Y - \mathbf{F}\mathbf{b}$, $\mathbf{b} = ((\mathrm{vec}(A_1))^T, \boldsymbol{\beta}_1^{\mathrm{T}})^{\mathrm{T}}$, $\mathbf{F} = (F_1^{\mathrm{T}}, \ldots, F_m^{\mathrm{T}})^{\mathrm{T}}$, $F_i = (Z_{i1}^{\mathrm{T}} \otimes \mathbf{x}_{i1}, \mathbf{1}_{n_i} \otimes T_i^{\mathrm{T}})$.

Let Δ_{ij} and W_{ij} be the Δ and W with u being replaced by U_{ij}, respectively. Replacing the $\mathbf{A}^c(U_{ij})$ and $\boldsymbol{\beta}_2(U_{ij})$ by the corresponding components of $\boldsymbol{\theta}(U_{ij})$, we have the following synthetic regression model:

$$Y - \mathbf{F}\mathbf{b} = \mathbf{S}(Y - \mathbf{F}\mathbf{b}) + \mathbf{r}, \tag{13}$$

where $\mathbf{r} = (\mathbf{r}_1^{\mathrm{T}}, \ldots, \mathbf{r}_m^{\mathrm{T}})^{\mathrm{T}}$,

$$\mathbf{S} = \begin{pmatrix} (Z_{11}^{\mathrm{T}} \otimes X_{112}^{\mathrm{T}}, Z_{12}^{\mathrm{T}} \otimes X_{11}^{\mathrm{T}}, V_1^{\mathrm{T}})(I_d, \mathbf{0}_{d\times d})(\Delta_{11}^{\mathrm{T}} W_{11} \Delta_{11})^{-1} \Delta_{11}^{\mathrm{T}} W_{11} \\ \vdots \\ (Z_{11}^{\mathrm{T}} \otimes X_{1n_12}^{\mathrm{T}}, Z_{12}^{\mathrm{T}} \otimes X_{1n_1}^{\mathrm{T}}, V_1^{\mathrm{T}})(I_d, \mathbf{0}_{d\times d})(\Delta_{1n_1}^{\mathrm{T}} W_{1n_1} \Delta_{1n_1})^{-1} \Delta_{1n_1}^{\mathrm{T}} W_{1n_1} \\ \vdots \\ (Z_{m1}^{\mathrm{T}} \otimes X_{m12}^{\mathrm{T}}, Z_{m2}^{\mathrm{T}} \otimes X_{m1}^{\mathrm{T}}, V_m^{\mathrm{T}})(I_d, \mathbf{0}_{d\times d})(\Delta_{m1}^{\mathrm{T}} W_{m1} \Delta_{m1})^{-1} \Delta_{m1}^{\mathrm{T}} W_{m1} \\ \vdots \\ (Z_{m1}^{\mathrm{T}} \otimes X_{mn_m2}^{\mathrm{T}}, Z_{m2}^{\mathrm{T}} \otimes X_{mn_m}^{\mathrm{T}}, V_m^{\mathrm{T}})(I_d, \mathbf{0}_{d\times d})(\Delta_{mn_m}^{\mathrm{T}} W_{mn_m} \Delta_{mn_m})^{-1} \Delta_{mn_m}^{\mathrm{T}} W_{mn_m} \end{pmatrix},$$

X_{ij2} is the vector of the last p_2 components of X_{ij}. Applying the weighted least squares estimation to model (13) with weight$\Lambda = \mathrm{diag}(\mathbf{x}_1 \hat{\Sigma} \mathbf{x}_1^{\mathrm{T}}, \ldots, \mathbf{x}_m \hat{\Sigma} \mathbf{x}_m^{\mathrm{T}}) + \hat{\sigma}^2 I_n$, we obtain the final estimator of $\mathbf{b}$:

$$\hat{\mathbf{b}} = \{\mathbf{F}^{\mathrm{T}}(I_n - \mathbf{S})^{\mathrm{T}} \Lambda^{-1} (I_n - \mathbf{S})\mathbf{F}\}^{-1} \mathbf{F}^{\mathrm{T}}(I_n - \mathbf{S})^{\mathrm{T}} \Lambda^{-1} (I_n - \mathbf{S}) Y. \tag{14}$$

Substituting $\hat{\mathbf{b}}$ for $\mathbf{b}$ in (12) and changing the bandwidth h to a slightly larger one h_1, we arrive at the final estimator of $\boldsymbol{\theta}(u)$

$$\hat{\boldsymbol{\theta}}(u) = (I_d, \mathbf{0}_{d\times d})(\Delta^{\mathrm{T}} W_1 \Delta)^{-1} \Delta^{\mathrm{T}} W_1 (Y - \mathbf{F}\hat{\mathbf{b}}), \tag{15}$$

where W_1 is the W with h being replaced by h_1.

From the asymptotic properties presented in Sect. 4, we can see that the optimal bandwidth for the estimators of constant coefficients is smaller than that for functional coefficients, and that is why we need to replace h by a slightly larger bandwidth h_1 when constructing the estimators of functional coefficients through (12).

3 Model selection

Suppose (3) is the true model. We may delete the ith cluster and estimate the $\mathbf{A}(\cdot)$, $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2(\cdot)$, Σ and σ^2 based on the remaining $m-1$ clusters. Denote the resulting estimators of $\mathbf{A}(\cdot)$, $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2(\cdot)$, Σ and σ^2 by $\hat{\mathbf{A}}^{\setminus i}(\cdot)$, $\hat{\boldsymbol{\beta}}_1^{\setminus i}$, $\hat{\boldsymbol{\beta}}_2^{\setminus i}(\cdot)$, $\hat{\Sigma}^{\setminus i}$ and $\hat{\sigma}^{2\setminus i}$, respectively. The cross-validation sum of squares is defined as

$$\mathrm{CV} = m^{-1}\sum_{i=1}^{m} \mathbf{r}_i^{*\mathrm{T}}(\mathbf{x}_i \hat{\Sigma}^{\setminus i}\mathbf{x}_i^{\mathrm{T}} + \hat{\sigma}^{2\setminus i} I_{n_i})^{-1}\mathbf{r}_i^*, \tag{16}$$

where $\mathbf{r}_i^* = (r_{i1}^*, \ldots, r_{in_i}^*)^{\mathrm{T}}$, $r_{ij}^* = y_{ij} - X_{ij}^{\mathrm{T}}\hat{\mathbf{A}}^{\setminus i}(U_{ij})Z_i - T_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_1^{\setminus i} - V_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_2^{\setminus i}(U_{ij})$. We compute the CVs for all candidate models; the selected model is the one with the smallest CV.

Let L be the number of the coefficients in the model. Denote the vector consisting of all coefficients in the model by $\boldsymbol{\alpha}(\cdot) = (\alpha_1(\cdot), \ldots, \alpha_L(\cdot))$, the model with coefficients $\alpha_{i_l}(\cdot), l = 1, \ldots, k$, being functional, others being constant by $\{i_1, \ldots, i_k\}$. In the following, we will present the following backward elimination algorithm for the model selection:

(1) We start with the full model, $\{1, \ldots, L\}$, and compute its CV by (16). We denote the full model by $\mathcal{M}_L$, its CV by CV_L.
(2) Suppose the model is $\mathcal{M}_k = \{i_1, \ldots, i_k\}$ at some step. We define $\mathcal{M}_{k-1}$ as the model with the smallest CV among the models $\{i_1, \ldots, i_{j-1}, i_{j+1}, \ldots, i_k\}$, $j = 1, \ldots, k$. We use CV_k to denote the CV of model $\mathcal{M}_k$, which can be computed by (16). If $\mathrm{CV}_k < \mathrm{CV}_{k-1}$, the chosen model is $\mathcal{M}_k$, and the model selection is ended; otherwise, compute $\mathcal{M}_{k-2}$ and CV_{k-2}, and continue until reaching some l such that either $\mathrm{CV}_l < \mathrm{CV}_{l-1}$ or $l = 0$.

4 Asymptotic properties

Throughout this paper, for any matrix C, we use $C > 0$ to denote that C is positive definite, and $C \geq 0$ to denote that C is semi-positive definite. Further, denote $\mu_2 = \int t^2 K(t)dt$, $\nu_0 = \int K^2(t)dt$ and let $f(\cdot)$ be the density of U_{11}.

Theorem 1 *Under the conditions (1)–(7) in Appendix, when $nh^4 \to 0$, we have $n^{\frac{1}{2}}\{\hat{\mathbf{b}} - \mathbf{b}\} \xrightarrow{D} N(\mathbf{0}, \Pi_1^{-1} + \Pi_1^{-1}\Pi_2\Pi_1^{-1})$, where $\Pi_1 > 0$, $\Pi_2 \geq 0$ and are defined in Appendix.*

Theorem 1 implies that we have to choose $h = o(n^{-1/4})$, which is at a higher order of the optimal bandwidth for functional estimation, to make the estimators of constant coefficients achieve the convergence rate of $n^{-1/2}$.

Theorem 2 *Under the conditions* (1)–(8) *in Appendix, when* $h_1 = O(n^{-1/5})$ *and* $h/h_1 \to 0$, *we have*

$$\sqrt{nh_1 f(u)}\left\{\hat{\boldsymbol{\theta}}(u) - \boldsymbol{\theta}(u) - \frac{1}{2}\mu_2 h_1^2 \boldsymbol{\theta}''(u)\right\} \xrightarrow{D} N(\mathbf{0}, \nu_0\{\Omega_1(u)^{-1}\Omega_2(u)\Omega_1(u)^{-1} + \sigma^2\Omega_1(u)^{-1}\}), \tag{17}$$

where $\Omega_i(u) > 0, i = 1, 2$, *and are defined in Appendix.*

Comparing previous asymptotic results (Zhang and Lee 2000) for standard varying-coefficient estimation with Theorem 2, we note that the asymptotic distributions of the proposed estimators of the functional coefficients are the same as those obtained when the constant coefficients are known.

Theorem 3 *When the working model is the true model, under the conditions* (1)–(8) *in Appendix, if* $n^{\frac{1}{2}}h_1^3 \to 0, h/h_1 \to 0$, *and* $\{nh_1\}^{\frac{1}{2}}h^2 \to 0$, *we have* $CV = m^{-1}\sum_{i=1}^m \mathbf{r}_i^T\Lambda_i^{-1}\mathbf{r}_i + \frac{1}{4}\mu_2^2 h_1^4\pi_1 + \frac{\nu_0}{mh_1}\{\sigma^2\lambda_1 + \lambda_2\} + o_P\{h_1^4 + \frac{1}{mh_1}\}$, *where* $\pi_1 > 0, \Lambda_i > 0$ *are defined in Appendix, and* $\lambda_1 = n^{-1}\sum_{i=1}^m\sum_{j=1}^{n_i} E\{f(U_{ij})^{-1}G_{ij}^T\Omega_1(U_{ij})^{-1}G_{ij}[\Lambda_{0i}^{-1}]_{jj}\}, \lambda_2 = n^{-1}\sum_{i=1}^m\sum_{j=1}^{n_i} E\{f(U_{ij})^{-1}G_{ij}^T\Omega_1(U_{ij})^{-1}\Omega_2(U_{ij})\Omega_1(U_{ij})^{-1}G_{ij}[\Lambda_{0i}^{-1}]_{jj}\}$, *with* $[\Lambda_{0i}^{-1}]_{jj}$ *being the* j*th element on the diagonal of* $\Lambda_{0i}^{-1} = (\mathbf{x}_i\Sigma\mathbf{x}_i^T + \sigma^2 I_{n_i})^{-1}$.

We are now presenting the asymptotic form of the CV when the working model mistakenly treats some constant coefficients as functional. Without loss of generality, we assume the first element on the diagonal of A_1 is mistakenly treated as functional.

Theorem 4 *When the working model mistakenly treats the first element on the diagonal of* A_1 *as functional, under the conditions* (1)–(8) *in Appendix, if* $n^{\frac{1}{2}}h_1^3 \to 0, h/h_1 \to 0$, *and* $\{nh_1\}^{\frac{1}{2}}h^2 \to 0$, *we have* $CV = m^{-1}\sum_{i=1}^m \mathbf{r}_i^T\Lambda_i^{-1}\mathbf{r}_i + \frac{1}{4}\mu_2^2 h_1^4\pi_1 + \frac{\nu_0}{mh_1}\{\sigma^2\lambda_1^{(1)} + \lambda_2^{(1)}\} + o_P\{h_1^4 + \frac{1}{mh_1}\}$, *where* $\lambda_1^{(1)} = n^{-1}\sum_{i=1}^m\sum_{j=1}^{n_i} E\{f(U_{ij})^{-1}(G_{ij}^T, Z_{i11}X_{ij11})\Psi_1(U_{ij})^{-1}(G_{ij}^T, Z_{i11}X_{ij11})^T[\Lambda_{0i}^{-1}]_{jj}\}$, $\lambda_2^{(1)} = n^{-1}\sum_{i=1}^m\sum_{j=1}^{n_i} E\{f(U_{ij})^{-1}(G_{ij}^T, Z_{i11}X_{ij11})\Psi^*(U_{ij})(G_{ij}^T, Z_{i11}X_{ij11})^T[\Lambda_{0i}^{-1}]_{jj}\}$, *and* $\Psi^*(U_{ij}) = \Psi_1(U_{ij})^{-1}\Psi_2(U_{ij})\Psi_1(U_{ij})^{-1}, \Psi_1(u) = \begin{pmatrix}\Omega_1(u) & \Upsilon_1(u)\\ \Upsilon_1(u)^T & \rho_1(u)\end{pmatrix}, \Psi_2(u) = \begin{pmatrix}\Omega_2(u) & \Upsilon_2(u)\\ \Upsilon_2(u)^T & \rho_2(u)\end{pmatrix}$, $\Upsilon_1(u) = E[G_{11}Z_{111}X_{1111}|U_{11} = u], \rho_1(u) = E[\{Z_{111}X_{1111}\}^2|U_{11} = u], \Upsilon_2(u) = E[X_{11}^T\Sigma X_{11}G_{11}Z_{111}X_{1111}|U_{11} = u], \rho_2(u) = E[X_{11}^T\Sigma X_{11}\{Z_{111}X_{1111}\}^2|U_{11} = u]$, *with* Z_{111} *and* X_{1111} *representing the first component of* Z_{11} *and* X_{111}.

Remark 1 As we can see, by simple calculation, that

$$\begin{aligned}&(G_{ij}^{\mathrm{T}}, Z_{i11}X_{ij11})\Psi_1(U_{ij})^{-1}(G_{ij}^{\mathrm{T}}, Z_{i11}X_{ij11})^{\mathrm{T}} - G_{ij}^{\mathrm{T}}\Omega_1(U_{ij})^{-1}G_{ij}\\ &= \rho_3(U_{ij})^{-1}(G_{ij}^{\mathrm{T}}\Omega_1(U_{ij})^{-1}\Upsilon_1(U_{ij}) - Z_{i11}X_{ij11})^2\end{aligned} \tag{18}$$

where $\rho_3(U_{ij}) = \rho_1(U_{ij}) - \Upsilon_1(U_{ij})^{\mathrm{T}}\Omega_1(U_{ij})^{-1}\Upsilon_1(U_{ij}) > 0$ since $\Psi_1(U_{ij}) > 0$. Therefore, $\lambda_1^{(1)} - \lambda_1 > 0$, and $\lambda_1^{(1)} - \lambda_1 = O(1)$. It follows that

$$\begin{aligned}
&(G_{ij}^{\mathrm{T}}, Z_{i11}X_{ij11})\Psi_1(U_{ij})^{-1}\Psi_2(U_{ij})\Psi_1(U_{ij})^{-1}(G_{ij}^{\mathrm{T}}, Z_{i11}X_{ij11})^{\mathrm{T}} \\
&\quad -G_{ij}^{\mathrm{T}}\Omega_1(U_{ij})^{-1}\Omega_2(U_{ij})\Omega_1(U_{ij})^{-1}G_{ij} \\
&= (G_{ij}^{\mathrm{T}}, Z_{i11}X_{ij11})L(U_{ij})(G_{ij}^{\mathrm{T}}, Z_{i11}X_{ij11})^{\mathrm{T}} + l(U_{ij})^{\mathrm{T}}\Psi_2(U_{ij})l(U_{ij}), \qquad (19)
\end{aligned}$$

where $L(U_{ij}) = \rho_3(U_{ij})^{-1}[\Psi_3(U_{ij}) + \Psi_3(U_{ij})^{\mathrm{T}}]$, $\Psi_3(U_{ij}) = \begin{pmatrix} \Omega_1(U_{ij})^{-1} & \mathbf{0}_{d\times 1} \\ \mathbf{0}_{1\times d} & 0 \end{pmatrix}$ $\Psi_2(U_{ij})\Upsilon_3(U_{ij})\Upsilon_3(U_{ij})^{\mathrm{T}}$, $\Upsilon_3(U_{ij})^{\mathrm{T}} = (\Upsilon_1(U_{ij})^{\mathrm{T}}\Omega_1(U_{ij})^{-1}, -1)$, and $l(U_{ij}) = \rho_3(U_{ij})^{-1}\Upsilon_3(U_{ij})\Upsilon_3(U_{ij})^{\mathrm{T}}(G_{ij}^{\mathrm{T}}, Z_{i11}X_{ij11})^{\mathrm{T}}$.

It can be shown that $\Psi_3(U_{ij}) + \Psi_3(U_{ij})^{\mathrm{T}} \geq 0$, hence $\lambda_2^{(1)} - \lambda_2 > 0$, and $\lambda_2^{(1)} - \lambda_2 = O(1)$. Therefore, $\lambda_1^{(1)} - \lambda_1 > 0$ and $\lambda_2^{(1)} - \lambda_2 > 0$. This together with Theorems 3 and 4 indicates that the increment in the CV is detectable up to $O(\{mh_1\}^{-1})$ when the working model mistakenly treats a constant coefficient as functional.

We next give the asymptotic form of the CV when we mistakenly treat some functional coefficients as constant. Without loss of generality, we assume that the $(1, 1)$th element of $A_2(\cdot)$ is mistakenly treated as constant.

Theorem 5 *When the working model mistakenly treats the* $(1, 1)$*th element of* $A_2(\cdot)$ *as constant, under the conditions* (1)–(8) *in Appendix, we have* $CV{=}m^{-1}\sum_{i=1}^{m} \mathbf{r}_i^T\Lambda_i^{-1}\mathbf{r}_i + \pi_2 + o_P(1)$, *where* $\pi_2 > 0$, $\pi_2 = O(1)$, *and is defined in Appendix.*

Remark 2 Comparing Theorem 5 with Theorem 3, we can see that the increment in the CV is detectable up to $O(1)$ when the working model mistakenly treats some functional coefficients as constant.

5 Simulation study

The kernel function involved in the estimation is taken to be the Epanechnikov kernel $K(t) = 0.75(1 - t^2)_+$. The bandwidth h is selected by the cross validation when all coefficients are assumed to be functional. The selection is among a set of 20 equally spaced grid points over the support of U. The h that minimizes CV is taken for the estimation of constant coefficients as described in 2.2. To estimate the final functional coefficients, we set $h_1 = 1.25h$. It has been shown in Li et al. (2011) that CV can provide optimal bandwidth that satisfies the order requirements in general semiparametric regression models.

We generate data from Model (3) with $p_1 = p_2 = q_1 = q_2 = k_1 = k_2 = 1$, $A_1 = 5$, $A_2(U) = -9U(1 - U)$, $A_4(U) = 3\sin(2\pi U)$, $A_3(U) = 3.5[\exp\{-(3U - 1)^2\} + \exp\{-(4U - 3)^2\}] - 1.5$, and $\beta_1 = 3$, $\beta_2(U) = 3\sin(6\pi(U - .5)^2)$.

We set the number of clusters to be $m = 120$. For each cluster, the cluster size, n_i, is generated through 3 plus a random variable with binomial distribution $B(5, 0.5)$. U_{ij} are generated from uniform distribution $U[0, 1]$. Components of X_{ij} and Z_i are

Table 1 The MSEs and MISEs of the estimators

Parameter	TRUE	EST	MSE	Function	MISE
A_1	5	4.98	1.6×10^{-4}	$A_2(\cdot)$	0.45
β_1	3	3.01	1.0×10^{-5}	$A_3(\cdot)$	0.45
σ^2	1	1.04	1.14×10^{-5}	$A_4(\cdot)$	0.46
σ_{11}	3	2.59	6.96×10^{-4}	$\boldsymbol{\beta}_2(\cdot)$	0.35
σ_{22}	2.5	2.12	5.67×10^{-4}		
σ_{12}	1	0.73	2.59×10^{-4}		

The parameter column is the unknown parameters, the TRUE column is the true values of the unknown parameters, the EST column is the estimators of the unknown parameters obtained from the simulation with median performance, the MSE column is the MSEs of the estimators of the unknown parameters, the Function column is the unknown functions, and the MISE column is the MISEs of the estimators of the unknown functions

independently generated from uniform distribution $[-1, 1]$. The random effects $\mathbf{e}_i$ are generated from bivariate normal distribution with mean zero and covariance matrix Σ. The random errors ϵ_{ij} are generated from normal distribution $N(0, \sigma^2)$. We set $\sigma^2 = 1$ and $\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix} = \begin{pmatrix} 3 & 1 \\ 1 & 2.5 \end{pmatrix}$.

We conduct 100 simulations and use the mean squared error (MSE) and mean integrated squared error (MISE) to assess the accuracy of the estimator of a constant parameter and the estimator of a functional coefficient, respectively. The MSEs of the estimators of A_1, σ^2 and Σ are presented in Table 1, which shows our estimators are very accurate. We have also computed the MISEs of the functional coefficients $A_j(\cdot)$, $j = 2, 3, 4$, and $\boldsymbol{\beta}_2(\cdot)$, and presented them in Table 1. From Table 1, we can see our estimators of functional coefficients are also doing very well.

To give a more visible picture about how well our estimation is, we single out the one with median performance among the 100 simulations and report the estimators of constant parameters in Table 1 and estimators of functional coefficients in Fig. 1. From Fig. 1 and Table 1, we can see the estimators are indeed very good.

We have also examined how well the proposed CV coupled with the backward elimination algorithm for the model selection works. It turns out, out of 100 simulations, 92 times the CV picks the correct model, which suggests the CV coupled with the backward elimination works reasonably well.

6 Analysis of the first birth interval in Bangladesh

The data come from the BDHS of 1996–1997 (Mitra et al. 1997), which is a cross-sectional, nationally representative survey of ever-married women aged between 10 and 49. The analysis is based on a sample of 8,189 women nested within 296 primary sampling units or clusters, with sample sizes ranging from 16 to 58. We allow for the hierarchical structure of the data by fitting a two-level model with women at level 1 nested within clusters at level 2. A further hierarchical level is the administrative division; Bangladesh is divided into six administrative divisions which are Barisal,

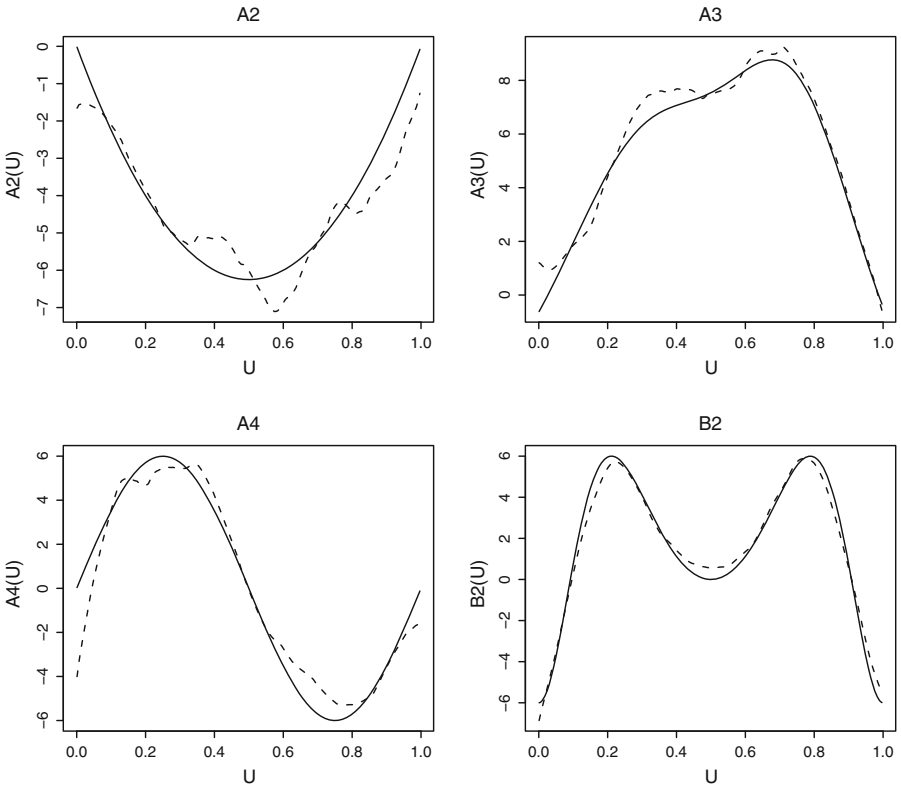

Fig. 1 Nonparametric estimates for functional coefficients in simulations

Chittagong, Dhaka, Kulna, Rajshahi and Sylhet. Effects at this level are represented in the model by fixed effects since there are only six divisions.

The dependent variable, y_{ij}, is the duration in months between marriage and the first birth for the jth woman in the ith cluster. We consider several covariates which are commonly found to be associated with fertility in Bangladesh. The selected individual-level categorical covariates include the woman's level of education (x_{ij1}) (none coded by 0, primary or secondary plus coded by 1), religion (x_{ij2}) (Muslim coded by 1, Hindu or other coded by 0). Another individual-level covariate is year of marriage (U_{ij}). We also consider two cluster-level variables, administrative division and type of region of residence. We take rural as the reference and the differences between urban and rural clusters is modeled by a dummy variables (z_{i2}). We take Barisal as the reference and the differences between Barisal and Chittagong, Barisal and Dhaka, Barisal and Kulna, Barisal and Rajshahi, and Barisal and Sylhet are modelled by dummy variables $z_{i3}, \ldots, z_{i7}$, respectively. We set $z_{i1} = 1$ to include the intercept into the model.

The proposed model (3) with $X_{ij} = (x_{ij1}, x_{ij2})^{\mathrm{T}}$ and $Z_i = (z_{i1}, \ldots, z_{i7})^{\mathrm{T}}$ is used to fit the data set. The kernel function involved in the estimation is still taken to be the Epanechnikov kernel. The bandwidth is selected by cross-validation when all coefficients are assumed to be functional.

We first use the proposed algorithm to identify which coefficients are constant and end up with the following choice:

$$\mathbf{A}(U_{ij}) = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13}(U_{ij}) & \alpha_{14}(U_{ij}) & \alpha_{15} & \alpha_{16}(U_{ij}) & \alpha_{17} \\ \alpha_{21}(U_{ij}) & \alpha_{22}(U_{ij}) & \alpha_{23}(U_{ij}) & \alpha_{24}(U_{ij}) & \alpha_{25}(U_{ij}) & \alpha_{26}(U_{ij}) & \alpha_{27}(U_{ij}) \end{pmatrix}$$

and $\boldsymbol{\beta}(U_{ij}) = (\beta_1(U_{ij}), \beta_2, \beta_3, \beta_4, \beta_5(U_{ij}), \beta_6, \beta_7(U_{ij}))^{\mathrm{T}}$.

The proposed estimation is applied to obtain the estimators of the unknown constant or functional coefficients. The estimated functional coefficients are displayed in Figs. 2 and 3, and the estimated constant coefficients are reported in Table 2.

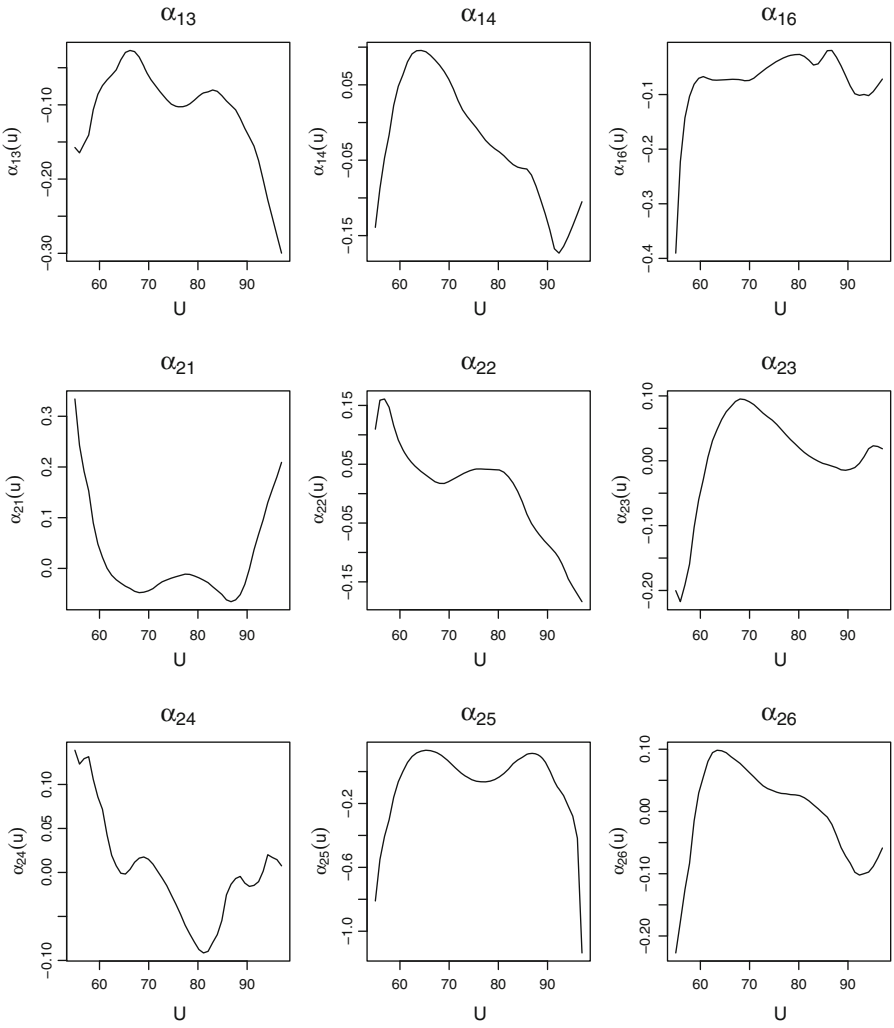

Fig. 2 The estimated functional coefficients for the first birth interval data in Bangladesh

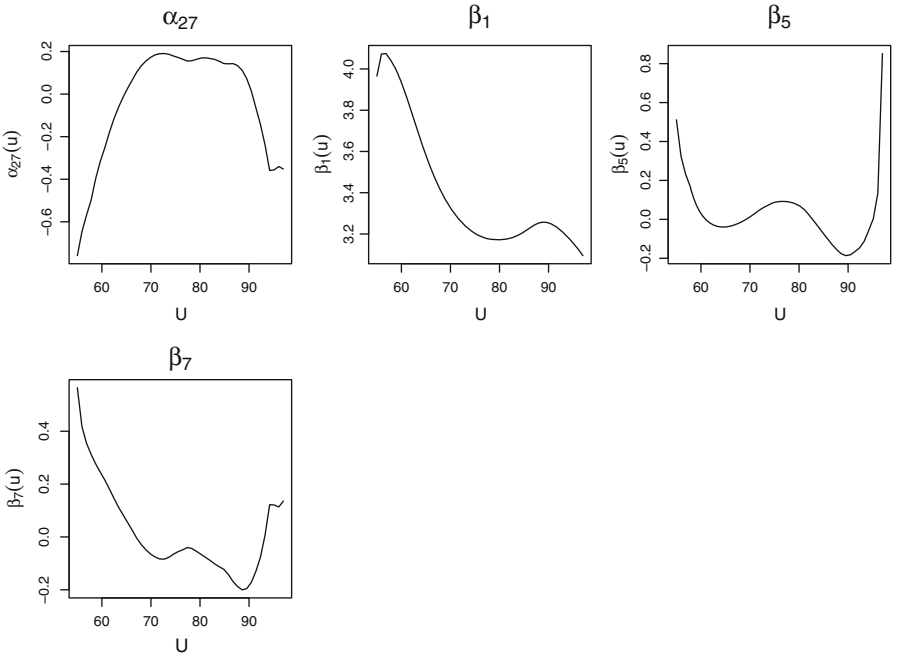

Fig. 3 The estimated functional coefficients for the first birth interval data in Bangladesh (continued)

Table 2 The estimated constant coefficients for the first birth interval data in Bangladesh

Coefficients	Estimates	Standard errors
α_{11}	0.053	0.109
α_{12}	−0.036	0.084
α_{15}	−0.067	0.107
α_{17}	0.050	0.103
β_2	−0.069	0.112
β_3	−0.015	0.044
β_4	0.026	0.056
β_6	0.034	0.102

From Fig. 3, we can see the estimator of the function $\beta_1(\cdot)$, which is the intercept and corresponding to the trend of the length of the first birth interval, is decreasing over time. This is largely due to an increase in the age at first marriage in Bangladesh; a nationally representative survey of women in 1996–1997 (Mitra et al. 1997) found that the median age at marriage was 13.3 years among respondents who were aged 45–49 years at the time of survey, compared with 15.3 years for respondents aged 20–24 years. There has been a slower increase in the age at first birth; the median age at first birth was 16.9 years among women who were aged 45–59 and 18.4 years among the younger cohort. These trends in age at marriage and age at first birth imply that the length of the first birth interval has become shorter over time.

Table 2 shows the estimator of β_2 is negative which suggests the women in rural areas have longer first birth interval than those in urban areas. We notice that the estimator of β_3 is negative, which suggests the intervals are shorter in Chittagong than in the other divisions. This regional effect is as expected and is most probably explained by lower use of contraceptives in Chittagong (the most religiously conservative part of Bangladesh) compared with other divisions. The estimators of β_4 and β_6 are positive, which indicates the intervals are longer in Dhaka and Rajshahi than in Barisal. The estimate of β_6 also suggests the first birth intervals in Rajshahi are longer than in the other divisions. The regional effect ($\beta_5(\cdot)$) of Kulna is changing over time, so does the Sylhet's ($\beta_7(\cdot)$). We can see clearly the dynamic patterns of the changes in Fig. 3. The standard errors in Table 2 may be used to construct asymptotic significance tests for the regression coefficients. We notice that none of the coefficients is significantly different from zero in this case.

Let $\hat{\mathbf{A}}(u)$ be the $\mathbf{A}(u)$ with each entry replaced by its estimator. The estimated impacts of the individual-level covariates on the length of the first birth interval in a specific cluster, say the ith cluster, can be obtained through $\hat{\mathbf{a}}_i(u) = \hat{\mathbf{A}}(u)Z_i$. For example, the estimated impact of a woman's education in a rural area in Chittagong is $\hat{\alpha}_{11} + \hat{\alpha}_{13}(u)$, u is the time. Notice that we take Barisal as reference when modelling the regional effect. It is interesting to note, from Fig. 3, that the function $\hat{\alpha}_{13}(u)$ is mainly negative, which suggests that education has less impact in Chittagong than in Barisal.

Appendix

Conditions

(1) The function $K(t)$ is a symmetric bounded density function with a compact support.
(2) $E\epsilon_{11}^4 < \infty$, $E\|\mathbf{e}_1\|^4 < \infty$, $E\|Z_1\|^{2(2+s)} < \infty$, $E\|X_{11}\|^{2(2+s)} < \infty$, $E(\|Z_1\|^{2(2+s)}\|X_{11}\|^{2(2+s)}) < \infty$, and $\max_{1\le i\le m} E\|\Lambda_{0i}^{-1}\|^{2(2+s)} < \infty$, for some $s > 0$, where $\Lambda_{0i} = \mathbf{x}_i \Sigma \mathbf{x}_i^{\mathrm{T}} + \sigma^2 I_{n_i}$, and $\|\mathbf{d}\|$ denotes Euclidean norm if $\mathbf{d}$ is a vector and Frobenius norm if $\mathbf{d}$ is a matrix.
(3) The marginal density $f(\cdot)$ of U is continuous and positive on its compact support.
(4) $A_k(\cdot)$, $k = 2, 3, 4$, and $\boldsymbol{\beta}_2(\cdot)$ have continuous second derivatives.
(5) $E\{R_{11}R_{11}^{\mathrm{T}}|U_{11} = u\}$ is continuous, where R_{11} is the first column of R_1^{T}. Further assume that $E\{R_{11}R_{11}^{\mathrm{T}}|U_{11} = u\}$ is positive definite.
(6) $n_i, i = 1, \ldots, m$, are bounded, $n \to \infty, h \to 0, h_1 \to 0$, and $nh^2 \to \infty, nh_1^2 \to \infty, nh^8 \to 0$.
(7) $E\{R_{ij}R_{il}^{\mathrm{T}}[\Lambda_{0i}^{-1}]_{jl}|U_{ij} = u\}$, $j, l = 1, \ldots, n_i, i = 1, \ldots, m$, are continuous, where R_{ij} is the jth column of R_i^{T}, and $[\Lambda_{0i}^{-1}]_{jl}$ is the (j, l)th element of Λ_{0i}^{-1}.
(8) $E\{X_{11}^{\mathrm{T}}\Sigma X_{11}R_{11}R_{11}^{\mathrm{T}}|U_{11} = u\}$ is continuous and positive definite. Moreover, $E[X_{il}^{\mathrm{T}}\Sigma X_{ir}R_{il}R_{ir}^{\mathrm{T}}|U_{il} = u, U_{ir} = v]$, $r, l = 1, \ldots, n_i, i = 1, \ldots, m$ are continuous, respectively, for $r \neq l$.

Notations in Theorem 1 To present the expressions of Π_1 and Π_2 in Theorem 1 clearly, we first introduce the following notations: $\Omega_1(u) = E[G_{11}G_{11}^{\mathrm{T}}|U_{11} = u]$, $\Omega_3(u) = E[G_{11}F_{11}^{\mathrm{T}}|U_{11} = u]$, $Q_{ij}^{\mathrm{T}} = F_{ij}^{\mathrm{T}} - G_{ij}^{\mathrm{T}}\Omega_1(U_{ij})^{-1}\Omega_3(U_{ij})$, $Q_i^{\mathrm{T}} = (Q_{i1}, \ldots, Q_{in_i})$, where F_{ij} is the jth column of F_i^{T}, $j = 1, \ldots, n_i$, $i = 1, \ldots, m$.

It is easy to see that conditions (1)–(7) guarantee the following limits exist and are finite. $\Pi_1 = \lim_{n\to\infty} n^{-1}\sum_{i=1}^m E[Q_i^{\mathrm{T}}\Lambda_{0i}^{-1}Q_i]$, $\Pi_2 = \lim_{n\to\infty} n^{-1}\sum_{i=1}^m E[\Gamma_i^{\mathrm{T}}\Lambda_{0i}\Gamma_i]$, where $\Gamma_i = (\Gamma_{i1}, \ldots, \Gamma_{in_i})^{\mathrm{T}}$ with $\Gamma_{ij}^{\mathrm{T}} = G_{ij}^{\mathrm{T}}\Omega_1(U_{ij})^{-1}\tilde{\Gamma}(U_{ij})$, and $\tilde{\Gamma}(u) = \lim_{n\to\infty} n^{-1}\sum_{i=1}^m\sum_{j=1}^{n_i}\sum_{l=1}^{n_i} E\{G_{ij}Q_{il}^{\mathrm{T}}[\Lambda_{0i}^{-1}]_{jl}|U_{ij} = u\}$.

Notations in Theorem 2 The $\Omega_1(u)$ is defined in Theorem 1, and $\Omega_2(u) = E[X_{11}^{\mathrm{T}}\Sigma X_{11}G_{11}G_{11}^{\mathrm{T}}|U_{11} = u]$.

Notations in Theorem 3 $\Lambda_i = \mathbf{x}_i\hat{\Sigma}^{\backslash i}\mathbf{x}_i^{\mathrm{T}} + \hat{\sigma}^{2\backslash i}I_{n_i}$, $i = 1, \ldots, m$, $\pi_1 = m^{-1}\sum_{i=1}^m E\{\boldsymbol{\gamma}_{1i}^{\mathrm{T}}\Lambda_{0i}^{-1}\boldsymbol{\gamma}_{1i}\}$, where $\boldsymbol{\gamma}_{1i} = (G_{i1}^{\mathrm{T}}\boldsymbol{\theta}''(U_{i1}), \ldots, G_{in_i}^{\mathrm{T}}\boldsymbol{\theta}''(U_{in_i}))^{\mathrm{T}}$.

Notations in Theorem 4 Denote the constant which the working model mistakenly treats the (1, 1)th element of $A_2(\cdot)$ as by $A_{2(1,1)}$. Let G_{ij}^* be the vector of G_{ij} with its $(p_2q_1 + 1)$th component deleted, $F_{ij}^* = (F_{ij}^{\mathrm{T}}, Z_{i21}X_{ij11})^{\mathrm{T}}$ with Z_{i21} being the first component of Z_{i2}, $\Omega_1^*(u) = E[G_{11}^*G_{11}^{*\mathrm{T}}|U_{11} = u]$, $\Omega_3^*(u) = E[G_{11}^*F_{11}^{*\mathrm{T}}|U_{11} = u]$, $\Upsilon_4(u) = E[G_{11}^*Z_{i21}X_{ij11}|U_{11} = u]$, $\boldsymbol{\gamma}_{2i} = (\gamma_{2i1}, \ldots, \gamma_{2in_i})^{\mathrm{T}}$, where $\gamma_{2ij} = [Z_{i21}X_{ij11} - G_{ij}^{*\mathrm{T}}\Omega_1^*(U_{ij})^{-1}\Upsilon_4(U_{ij})]\{A_{2(1,1)}(U_{ij}) - A_{2(1,1)}\}$. Moreover, we assume that the following limits exist and are finite, $\Upsilon_5 = \mathrm{plim}_{m\to\infty} m^{-1}\sum_{i=1}^m Q_i^{*\mathrm{T}}\Lambda_{0i}^{-1}\boldsymbol{\gamma}_{2i}$, $\Pi_3 = \mathrm{plim}_{m\to\infty} m^{-1}\sum_{i=1}^m Q_i^{*\mathrm{T}}\Lambda_{0i}^{-1}Q_i^*$, where $\grave{}$ denotes convergence in probability, and Q_i^* is the Q_i with F_{ij}, $\Omega_3(\cdot)$, $\Omega_1(\cdot)$ and G_{ij} replaced by F_{ij}^*, $\Omega_3^*(\cdot)$, $\Omega_1^*(\cdot)$ and G_{ij}^*, respectively. Let $\boldsymbol{\gamma}_{3i} = (\gamma_{3i1}, \ldots, \gamma_{3in_i})^{\mathrm{T}}$ with

$$\begin{aligned}\gamma_{3ij} = {} & [G_{ij}^{*\mathrm{T}}\tilde{\Omega}_1^*(U_{ij})^{-1}\Omega_3^*(U_{ij}) - F_{ij}^{*\mathrm{T}}]\Pi_3^{-1}\Upsilon_5 \\ & + [Z_{i21}X_{ij11} - G_{ij}^{*\mathrm{T}}\Omega_1^*(U_{ij})^{-1}\Upsilon_4(U_{ij})]\{A_{2(1,1)}(U_{ij}) - A_{2(1,1)}\}. \quad (20)\end{aligned}$$

The π_2 in Theorem 5 is $\pi_2 = m^{-1}\sum_{i=1}^m E[\boldsymbol{\gamma}_{3i}^{\mathrm{T}}\Lambda_{0i}^{-1}\boldsymbol{\gamma}_{3i}]$.

Proofs

For easy description, we write $H = \begin{pmatrix} 1 & 0 \\ 0 & h \end{pmatrix} \otimes I_d$, $l_3 = p_1q_1 + k_1$, $\boldsymbol{\gamma}_1 = (\boldsymbol{\gamma}_{11}^{\mathrm{T}}, \ldots, \boldsymbol{\gamma}_{1m}^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\Phi}_i = (G_{i1}^{\mathrm{T}}\boldsymbol{\theta}(U_{i1}), \ldots, G_{in_i}^{\mathrm{T}}\boldsymbol{\theta}(U_{in_i}))^{\mathrm{T}}$, $i = 1, \ldots, m$, $\boldsymbol{\Phi} = (\boldsymbol{\Phi}_1^{\mathrm{T}}, \ldots, \boldsymbol{\Phi}_m^{\mathrm{T}})^{\mathrm{T}}$.

Lemma 1 *Let $\{U_{ij}\}$ be i.i.d. random variables, $\{\xi_{ij}\}$ be identically distributed random variables, and ξ_{ij} be independent of ξ_{lk} for $i \neq l$. Further, assume that $E(\xi_{11}^2) < \infty$, and $K(\cdot)$ be a bounded positive function with a bounded support. When $nh^2 \to \infty$, for any nonnegative integer λ, we have*

$$\sup_u n^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}|\eta_{ij,\lambda}(u)K_h(U_{ij}-u)-E[\eta_{ij,\lambda}(u)K_h(U_{ij}-u)]|=O_P((nh^2)^{-1/2}),$$

where $\eta_{ij,\lambda}(u)=\xi_{ij}\{h^{-1}(U_{ij}-u)\}^{\lambda}$.

Proof As $K(\cdot)$ is a bounded function with a bounded support, $\sup_u|u^{\lambda}K(u)|$ is bounded, then it follows by Jensen's inequality that $\mathrm{var}(\sup_u S_{n,\lambda}) = O((nh^2)^{-1})$, where $S_{n,\lambda}=n^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\eta_{ij,\lambda}(u)K_h(U_{ij}-u)$. Therefore, the result follows. □

Lemma 2 *Under the conditions* (1)–*8, we have* $\hat{\sigma}^2=\sigma^2+O_P(n^{-1/2})$, $\hat{\Sigma}=\Sigma+O_P(n^{-1/2})$. *Moreover, neither of the asymptotic distributions of* $\hat{\sigma}^2$ *and* $\hat{\Sigma}$ *depends on whether the coefficients in the true model are functional or constant.*

Proof The lemma follows from the same arguments as that in Sun et al. (2007) for the proofs of Theorem 1 and 2 there. □

Lemma 3 *Under the conditions* (1)–(3), *(5) and (6), we have* $n^{-1}\mathbf{F}^T(I_n-\mathbf{S})^T\Lambda^{-1}(I_n-\mathbf{S})\mathbf{F}\xrightarrow{P}\Pi_1$.

Proof It follows from Lemma 1 that

$$H^{-1}\Delta^{\mathrm{T}}W\Delta H^{-1}=nf(u)\begin{pmatrix}1&0\\0&\mu_2\end{pmatrix}\otimes\Omega_1(u)\{1+o_P(1)\},\tag{21}$$

and

$$H^{-1}\Delta^{\mathrm{T}}W\mathbf{F}=nf(u)(1,0)^{\mathrm{T}}\otimes\Omega_3(u)\{1+o_P(1)\},\tag{22}$$

uniformly for u. Combining these two results yields that $(I_d,0_{d\times d})(\Delta^{\mathrm{T}}W\Delta)^{-1}\Delta^{\mathrm{T}}W\mathbf{F}=\Omega_1(u)^{-1}\Omega_3(u)\{1+o_P(1)\}$ uniformly for u.

Equivalently, we have $\mathbf{SF}=(M_{11},\ldots,M_{1n_1},\ldots,M_{m1},\ldots,M_{mn_m})^{\mathrm{T}}\{1+o_P(1)\}$, where $M_{ij}^{\mathrm{T}}=G_{ij}^{\mathrm{T}}\Omega_1(U_{ij})^{-1}\Omega_3(U_{ij})$. Therefore, it is easy to show that $n^{-1}\mathbf{F}^{\mathrm{T}}(I_n-\mathbf{S})^{\mathrm{T}}\Lambda^{-1}(I_n-\mathbf{S})\mathbf{F}=n^{-1}\sum_{i=1}^{m}Q_i^{\mathrm{T}}\Lambda_{0i}^{-1}Q_i\{1+o_P(1)\}=\Pi_1+o_P(1)$ by Lemma 2 and the Markov inequality. □

Lemma 4 *Under the conditions* (1)–(6), *we have* $n^{-1}\mathbf{F}^T(I_n-\mathbf{S})^T\Lambda^{-1}(I_n-\mathbf{S})\mathbf{\Phi}=O_P(h^2)$.

Proof By Taylor's expansion of $\boldsymbol{\theta}(v)$ with respect to v around $|v-u|<h$, Lemma 1, (21) and straightforward calculation, we have

$$(I_n-\mathbf{S})\mathbf{\Phi}=-\frac{1}{2}h^2\mu_2\boldsymbol{\gamma}_1\{1+o_P(1)\}.\tag{23}$$

Therefore, $n^{-1}\mathbf{F}^{\mathrm{T}}(I_n-\mathbf{S})^{\mathrm{T}}\Lambda^{-1}(I_n-\mathbf{S})\mathbf{\Phi}=n^{-1}\sum_{i=1}^{m}Q_i^{\mathrm{T}}\Lambda_{0i}^{-1}\{-\frac{1}{2}h^2\mu_2\boldsymbol{\gamma}_{1i}\}\{1+o_P(1)\}=O_P(h^2)$, by Lemma 2 and the Markov inequality. □

Proof of Theorem 1 It can be seen from Lemmas 3 and 4 that the bias term in $\sqrt{n}(\hat{\mathbf{b}}-\mathbf{b})$ is $\sqrt{n}\{\mathbf{F}^{\mathrm{T}}(I_n-\mathbf{S})^{\mathrm{T}}\Lambda^{-1}(I_n-\mathbf{S})\mathbf{F}\}^{-1}\mathbf{F}^{\mathrm{T}}(I_n-\mathbf{S})^{\mathrm{T}}\Lambda^{-1}(I_n-\mathbf{S})\mathbf{\Phi}=O_P(\sqrt{n}h^2)$. Obviously, it is negligible when $nh^4\to 0$.

It follows from similar straightforward calculation and Lemma 2 that

$$n^{-1}\mathrm{cov}\{\mathbf{F}^{\mathrm{T}}(I_n-\mathbf{S})^{\mathrm{T}}\Lambda^{-1}(I_n-\mathbf{S})\mathbf{r}\}=\Pi_1+\Pi_2+o(1). \tag{24}$$

Hence the theorem follows from Lemma 3, the Lindeberg Feller Theorem and Slutsky Theorem. □

Proof of Theorem 2 Since

$$\begin{aligned}
&\sqrt{nh_1f(u)}(\hat{\boldsymbol{\theta}}(u)-\boldsymbol{\theta}(u))\\
&=\sqrt{nh_1f(u)}(I_d,\mathbf{0}_{d\times d})(\Delta^{\mathrm{T}}W_1\Delta)^{-1}\Delta^{\mathrm{T}}W_1\mathbf{F}(\mathbf{b}-\hat{\mathbf{b}})\\
&\quad+\sqrt{nh_1f(u)}(I_d,\mathbf{0}_{d\times d})(\Delta^{\mathrm{T}}W_1\Delta)^{-1}\Delta^{\mathrm{T}}W_1\left(\mathbf{\Phi}-\Delta\begin{pmatrix}\boldsymbol{\theta}(u)\\ \boldsymbol{\theta}'(u)\end{pmatrix}\right)\\
&\quad+\sqrt{nh_1f(u)}(I_d,\mathbf{0}_{d\times d})(\Delta^{\mathrm{T}}W_1\Delta)^{-1}\Delta^{\mathrm{T}}W_1\mathbf{r}\equiv L_{n1}+L_{n2}+L_{n3},
\end{aligned}$$

by (21), (22) and the proof of Theorem 1, we have $L_{n1}=\sqrt{nh_1f(u)}\Omega_1(u)^{-1}\Omega_3(u)\mathbf{1}_{l_3}O_P(h^2+n^{-1/2})=o_P(1)$ when $nh_1^5=O(1)$ and $h/h_1\to 0$.

Using the same arguments as establishing (23), we can show that $L_{n2}=\frac{1}{2}\sqrt{nh_1f(u)}\mu_2h_1^2\boldsymbol{\theta}''(u)\{1+o_P(1)\}$. Further, by (21) and straightforward calculation, we get $L_{n3}=\{[nf(u)]^{-1}h_1\}^{1/2}\Omega_1(u)^{-1}\sum_{i=1}^m\sum_{j=1}^{n_i}G_{ij}r_{ij}K_{h_1}(U_{ij}-u)\{1+o_P(1)\}$. It can be shown that $E\{\sqrt{[nf(u)]^{-1}h_1}\sum_{i=1}^m\sum_{j=1}^{n_i}G_{ij}r_{ij}K_{h_1}(U_{ij}-u)\}=0$, and $\mathrm{var}\{\sqrt{[nf(u)]^{-1}h_1}\sum_{i=1}^m\sum_{j=1}^{n_i}G_{ij}r_{ij}K_{h_1}(U_{ij}-u)\}=\nu_0\{\Omega_2(u)+\sigma^2\Omega_1(u)\}+o(1)$. Therefore, by the Lindeberg Feller Theorem and Slutsky Theorem, we have $L_{n3}\xrightarrow{D}N(\mathbf{0},\nu_0\{\Omega_1(u)^{-1}\Omega_2(u)\Omega_1(u)^{-1}+\sigma^2\Omega_1(u)^{-1}\})$.

Combining the results on L_{n1}, L_{n2} and L_{n3} leads to the theorem. □

Proof of Theorem 3 Let $\Delta_{ij}^{\backslash i}$, $W_{1ij}^{\backslash i}$ be the Δ, W_1 obtained when the ith cluster is deleted and u replaced by U_{ij}, respectively, and $Y^{\backslash i}$, $\mathbf{F}^{\backslash i}$ be the Y and $\mathbf{F}$ obtained when the ith cluster is deleted, $\hat{\mathbf{b}}^{\backslash i}=((\mathrm{vec}(\hat{A}_1^{\backslash i}))^{\mathrm{T}},\hat{\boldsymbol{\beta}}_1^{\backslash iT})^{\mathrm{T}}$, and $\hat{\boldsymbol{\theta}}^{\backslash i}(u)=((\mathrm{vec}(\hat{A}_3^{\backslash i}(u)))^{\mathrm{T}},(\mathrm{vec}((\hat{A}_2^{\backslash iT}(u),\hat{A}_4^{\backslash iT}(u))^{\mathrm{T}}))^{\mathrm{T}},\hat{\boldsymbol{\beta}}_2^{\backslash i}(u)^{\mathrm{T}})^{\mathrm{T}}$. It can be seen that

$$r_{ij}^*=y_{ij}-F_{ij}^{\mathrm{T}}\mathbf{b}-G_{ij}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_0^{\backslash i}(U_{ij})+\{F_{ij}^{\mathrm{T}}-G_{ij}^{\mathrm{T}}\boldsymbol{\eta}(U_{ij})\}(\mathbf{b}-\hat{\mathbf{b}}^{\backslash i}) \tag{25}$$

where $\hat{\boldsymbol{\theta}}_0^{\backslash i}(\cdot)$ is $\hat{\boldsymbol{\theta}}^{\backslash i}(\cdot)$ with $\hat{\mathbf{b}}^{\backslash i}$ replaced by $\mathbf{b}$, and

$$\begin{aligned}
\boldsymbol{\eta}(U_{ij})&=(I_d,\mathbf{0}_{d\times d})(\Delta_{ij}^{\backslash iT}W_{1ij}^{\backslash i}\Delta_{ij}^{\backslash i})^{-1}\Delta_{ij}^{\backslash iT}W_{1ij}^{\backslash i}\mathbf{F}^{\backslash i}\\
&=\Omega_1(U_{ij})^{-1}\Omega_3(U_{ij})\{1+o_P(1)\}
\end{aligned} \tag{26}$$

uniformly; therefore,

$$\mathrm{CV} = m^{-1}\sum_{i=1}^{m}\hat{\mathbf{y}}_i^{\mathrm{T}}\Lambda_i^{-1}\hat{\mathbf{y}}_i + m^{-1}\sum_{i=1}^{m}(\mathbf{b}-\hat{\mathbf{b}}^{\setminus i})^{\mathrm{T}}R_i^{*\mathrm{T}}\Lambda_i^{-1}R_i^{*}(\mathbf{b}-\hat{\mathbf{b}}^{\setminus i}) + 2m^{-1}\sum_{i=1}^{m}(\mathbf{b}-\hat{\mathbf{b}}^{\setminus i})^{\mathrm{T}}R_i^{*\mathrm{T}}\Lambda_i^{-1}\hat{\mathbf{y}}_i, \tag{27}$$

where $\hat{\mathbf{y}}_i = (\hat{y}_{i1},\ldots,\hat{y}_{in_i})^{\mathrm{T}}$ with $\hat{y}_{ij} = y_{ij} - F_{ij}^{\mathrm{T}}\mathbf{b} - G_{ij}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_0^{\setminus i}(U_{ij})$, and $R_i^{*} = (R_{i1}^{*},\ldots,R_{in_i}^{*})^{\mathrm{T}}$ with $R_{ij}^{*\mathrm{T}} = F_{ij}^{\mathrm{T}} - G_{ij}^{\mathrm{T}}\boldsymbol{\eta}(U_{ij})$.

It can be seen, from the proof of Theorem 1, that $(\hat{\mathbf{b}}-\mathbf{b})$ is of convergence rate $\{h^2+n^{-1/2}\}$. When $n^{\frac{1}{2}}h_1^3\to 0$, $h/h_1\to 0$ and $\{nh_1\}^{\frac{1}{2}}h^2\to 0$, the asymptotic form of the CV is the same as that when $\mathbf{b}$ is known. Thus, we only need to consider the first term in (27) in detail, and we denote it by CV_1.

Let $\hat{\boldsymbol{\theta}}_0(\cdot)$ be the $\hat{\boldsymbol{\theta}}(\cdot)$ with $\hat{\mathbf{b}}$ replaced by $\mathbf{b}$. For two matrices A and B, it is easy to see $(A+hB)^{-1} = A^{-1} - hA^{-1}BA^{-1} + O(h^2)$. Also, $\hat{\boldsymbol{\theta}}_0(u) - \boldsymbol{\theta}(u) = o_P(1)$, $h_1[\hat{\boldsymbol{\theta}}_0'(u) - \boldsymbol{\theta}'(u)] = o_P(1)$, uniformly for u by Lemma 1 and straightforward calculation, where $\hat{\boldsymbol{\theta}}_0'(u) = (\mathbf{0}_{d\times d}, I_d)(\Delta^{\mathrm{T}}W_1\Delta)^{-1}\Delta^{\mathrm{T}}W_1(Y-\mathbf{F}\mathbf{b})$. So, we have

$$\hat{\boldsymbol{\theta}}_0^{\setminus i}(U_{ij}) - \hat{\boldsymbol{\theta}}_0(U_{ij}) = -\frac{\Omega_1(U_{ij})^{-1}}{nf(U_{ij})}\sum_{l=1}^{n_i}G_{il}K_{h_1}(U_{il}-U_{ij})r_{il} + o_P(\{nh_1\}^{-1}) \tag{28}$$

holds uniformly. Let $\hat{\boldsymbol{\Phi}}_i$ and $\hat{\boldsymbol{\Phi}}_i^{\setminus i}$ be the $\boldsymbol{\Phi}_i$ with $\boldsymbol{\theta}(\cdot)$ replaced by $\hat{\boldsymbol{\theta}}_0(\cdot)$ and $\hat{\boldsymbol{\theta}}_0^{\setminus i}(\cdot)$, respectively; we have

$$\begin{aligned}\mathrm{CV}_1 &= m^{-1}\sum_{i=1}^{m}\{Y_i - F_i\mathbf{b} - \hat{\boldsymbol{\Phi}}_i\}^{\mathrm{T}}\Lambda_i^{-1}\{Y_i - F_i\mathbf{b} - \hat{\boldsymbol{\Phi}}_i\}\\ &\quad + m^{-1}\sum_{i=1}^{m}\{\hat{\boldsymbol{\Phi}}_i - \hat{\boldsymbol{\Phi}}_i^{\setminus i}\}^{\mathrm{T}}\Lambda_i^{-1}\{\hat{\boldsymbol{\Phi}}_i - \hat{\boldsymbol{\Phi}}_i^{\setminus i}\}\\ &\quad + 2m^{-1}\sum_{i=1}^{m}\{Y_i - F_i\mathbf{b} - \hat{\boldsymbol{\Phi}}_i\}^{\mathrm{T}}\Lambda_i^{-1}\{\hat{\boldsymbol{\Phi}}_i - \hat{\boldsymbol{\Phi}}_i^{\setminus i}\}\\ &\equiv J_{n1} + J_{n2} + 2J_{n3}.\end{aligned} \tag{29}$$

It is easy to see

$$\begin{aligned}J_{n1} &= m^{-1}\boldsymbol{\Phi}^{\mathrm{T}}(I_n - \mathbf{S}_1)^{\mathrm{T}}(\Lambda^{\setminus i})^{-1}(I_n - \mathbf{S}_1)\boldsymbol{\Phi}\\ &\quad + m^{-1}\mathbf{r}^{\mathrm{T}}(I_n - \mathbf{S}_1)^{\mathrm{T}}(\Lambda^{\setminus i})^{-1}(I_n - \mathbf{S}_1)\mathbf{r}\\ &\quad + 2m^{-1}\boldsymbol{\Phi}^{\mathrm{T}}(I_n - \mathbf{S}_1)^{\mathrm{T}}(\Lambda^{\setminus i})^{-1}(I_n - \mathbf{S}_1)\mathbf{r}\\ &\equiv J_{n11} + J_{n12} + 2J_{n13},\end{aligned} \tag{30}$$

where $\mathbf{S}_1$ is the $\mathbf{S}$ with h replaced by h_1 and $\Lambda^{\setminus i} = \mathrm{diag}(\Lambda_1, \ldots, \Lambda_m)$. It follows from Lemma 2 and (23) that $J_{n11} = \frac{1}{4}h_1^4\mu_2^2 m^{-1}\sum_{i=1}^m \boldsymbol{\gamma}_{1i}^{\mathrm{T}}\Lambda_{0i}^{-1}\boldsymbol{\gamma}_{1i}\{1+o_P(1)\} = \frac{1}{4}h_1^4\mu_2^2\pi_1 + o_P(h_1^4)$, and $J_{n13} = -\frac{1}{2}h_1^2\mu_2 m^{-1}\boldsymbol{\gamma}_1^{\mathrm{T}}\Lambda_0^{-1}(I_n - \mathbf{S}_1)\mathbf{r}\{1+o_P(1)\}$. Since $E[m^{-1}\boldsymbol{\gamma}_1^{\mathrm{T}}\Lambda_0^{-1}(I_n - \mathbf{S}_1)\mathbf{r}] = 0$, and $\mathrm{var}[m^{-1}\boldsymbol{\gamma}_1^{\mathrm{T}}\Lambda_0^{-1}(I_n - \mathbf{S}_1)\mathbf{r}] = O(m^{-1})$, we have $J_{n13} = O_P(h_1^2 m^{-\frac{1}{2}}) = o_P(\{mh_1\}^{-1})$. With regard to J_{n12}, $J_{n12} = m^{-1}\mathbf{r}^{\mathrm{T}}(\Lambda^{\setminus i})^{-1}\mathbf{r} + m^{-1}\mathbf{r}^{\mathrm{T}}\mathbf{S}_1^{\mathrm{T}}(\Lambda^{\setminus i})^{-1}\mathbf{S}_1\mathbf{r} - 2m^{-1}\mathbf{r}^{\mathrm{T}}(\Lambda^{\setminus i})^{-1}\mathbf{S}_1\mathbf{r}$. By Lemma 2, (21), and the Markov inequality, we have

$$
\begin{aligned}
&m^{-1}\mathbf{r}^{\mathrm{T}}(\Lambda^{\setminus i})^{-1}\mathbf{S}_1\mathbf{r} \\
&\quad = \frac{1}{mn}\sum_{i=1}^m\sum_{j=1}^{n_i}\sum_{t=1}^{n_i}\sum_{l=1}^{n_i}\frac{G_{it}^{\mathrm{T}}\Omega_1(U_{it})^{-1}G_{il}}{f(U_{it})}K_{h_1}(U_{il}-U_{it})r_{ij}r_{il}[\Lambda_{0i}^{-1}]_{jt} \\
&\qquad + o_P(\{mh_1\}^{-1}).
\end{aligned}
$$

Also,

$$
\begin{aligned}
&m^{-1}\mathbf{r}^{\mathrm{T}}\mathbf{S}_1^{\mathrm{T}}(\Lambda^{\setminus i})^{-1}\mathbf{S}_1\mathbf{r} \\
&= \Bigg[\frac{1}{n^2m}\sum_{i=1}^m\sum_{j=1}^{n_i}\sum_{t=1}^{n_i}\sum_{k=1}^m\sum_{l=1}^{n_k}\sum_{s=1}^{n_k}\frac{G_{ij}^{\mathrm{T}}\Omega_1(U_{ij})^{-1}G_{kl}}{f(U_{ij})}\frac{G_{ks}^{\mathrm{T}}\Omega_1(U_{it})^{-1}G_{it}}{f(U_{it})}r_{kl}r_{ks} \\
&\quad\times K_{h_1}(U_{kl}-U_{ij})K_{h_1}(U_{ks}-U_{it})[\Lambda_{0i}^{-1}]_{jt} \\
&\quad + \frac{2}{n^2m}\sum_{k=1}^m\sum_{l=1}^{n_k}\sum_{j=1}^{n_k}\sum_{t=1}^{n_k}\sum_{v>k}^m\sum_{s=1}^{n_v}\frac{G_{kj}^{\mathrm{T}}\Omega_1(U_{kj})^{-1}G_{kl}}{f(U_{kj})}\frac{G_{vs}^{\mathrm{T}}\Omega_1(U_{kt})^{-1}G_{kt}}{f(U_{kt})}r_{kl}r_{vs} \\
&\quad\times K_{h_1}(U_{kl}-U_{kj})K_{h_1}(U_{vs}-U_{kt})[\Lambda_{0k}^{-1}]_{jt} \\
&\quad + \frac{2}{n^2m}\sum_{k=1}^m\sum_{l=1}^{n_k}\sum_{v>k}^m\sum_{s=1}^{n_v}\sum_{j=1}^{n_v}\sum_{t=1}^{n_v}\frac{G_{vj}^{\mathrm{T}}\Omega_1(U_{vj})^{-1}G_{kl}}{f(U_{vj})}\frac{G_{vs}^{\mathrm{T}}\Omega_1(U_{vt})^{-1}G_{vt}}{f(U_{vt})}r_{kl}r_{vs} \\
&\quad\times K_{h_1}(U_{kl}-U_{vj})K_{h_1}(U_{vs}-U_{vt})[\Lambda_{0v}^{-1}]_{jt} + \frac{2}{n^2m}\sum_{k=1}^m\sum_{l=1}^{n_k}\sum_{v>k}^m\sum_{s=1}^{n_v}r_{kl}r_{vs} \\
&\quad\times G_{kl}^{\mathrm{T}}\sum_{i\neq k,v}^m\sum_{j=1}^{n_i}\sum_{t=1}^{n_i}\{N_{klvs,ijt} - E(N_{klvs,ijt}|U_{kl},U_{vs})\}G_{vs} \\
&\quad + \frac{2}{n^2m}\sum_{k=1}^m\sum_{l=1}^{n_k}\sum_{v>k}^m\sum_{s=1}^{n_v}r_{kl}r_{vs}G_{kl}^{\mathrm{T}}\sum_{i\neq k,v}^m\sum_{j=1}^{n_i}\sum_{t=1}^{n_i}E(N_{klvs,ijt}|U_{kl},U_{vs})G_{vs}\Bigg] \\
&\quad\times\{1+o_P(1)\} \equiv \{L_{n1}+2L_{n2}+2L_{n3}+2L_{n4}+2L_{n5}\}\{1+o_P(1)\}
\end{aligned}
$$

where $N_{klvs,ijt}$ is $[f(U_{ij})f(U_{it})]^{-1}\Omega_1(U_{ij})^{-1}G_{ij}G_{it}^{\mathrm{T}}\Omega_1(U_{it})^{-1}K_{h_1}(U_{kl}-U_{ij})K_{h_1}(U_{vs}-U_{it})[\Lambda_{0i}^{-1}]_{jt}$. By straightforward calculations and the Markov inequality, we obtain $L_{n1} = \frac{\nu_0}{mh_1}\{\sigma^2\lambda_1 + \lambda_2\} + o_P(\{mh_1\}^{-1})$.

By simple calculation, we get $EL_{n2} = 0$, and $\mathrm{var}(L_{n2}) = O_P(\{nh_1\}^{-4})$; therefore, $L_{n2} = o_P(\{mh_1\}^{-1})$. Similarly $L_{n3} = o_P(\{mh_1\}^{-1})$. Obviously,

$$E\left[n^{-1}G_{kl}^{\mathrm{T}}\sum_{i\neq k,v}^{m}\sum_{j=1}^{n_i}\sum_{t=1}^{n_i}\{N_{klvs,ijt} - E(N_{klvs,ijt}|U_{kl}, U_{vs})\}G_{vs}\right]^2$$
$$\leq \mathrm{tr}\left\{n^{-2}\sum_{i\neq k,v}^{m} n_i^2\sum_{j=1}^{n_i}\sum_{t=1}^{n_i} E[N_{klvs,ijt}G_{vs}G_{vs}^{\mathrm{T}}N_{klvs,ijt}G_{kl}G_{kl}^{\mathrm{T}}]\right\} = O(n^{-1}h_1^{-2}).$$

Hence, $\mathrm{var}(L_{n5}) = O(n^{-3}h_1^{-2})$, this together with $EL_{n4} = 0$ leads to $L_{n4} = o_P(\{mh_1\}^{-1})$. Further, it can be easily shown that $EL_{n5} = 0$, and $\mathrm{var}(L_{n5}) = O(n^{-2}h_1^{-1})$, which implies $L_{n5} = o_P(\{mh_1\}^{-1})$. Combining all the above results relating J_{n1}, we have that

$$\begin{aligned} J_{n1} = {} & m^{-1}\sum_{i=1}^{m}\mathbf{r}_i^{\mathrm{T}}\Lambda_i^{-1}\mathbf{r}_i + \frac{1}{4}\mu_2^2h_1^4\pi_1 + \frac{\nu_0}{mh_1}\{\sigma^2\lambda_1 + \lambda_2\} \\ & - \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\sum_{t=1}^{n_i}\sum_{l=1}^{n_i}\frac{G_{it}^{\mathrm{T}}\Omega_1(U_{it})^{-1}G_{il}}{f(U_{it})}K_{h_1}(U_{il} - U_{it})r_{ij}r_{il}[\Lambda_{0i}^{-1}]_{jt} \\ & + o_P(h_1^4 + \{mh_1\}^{-1}). \end{aligned} \tag{31}$$

Next, we consider J_{n2} and J_{n3}. It follows from Lemma 2, (28) and the Markov inequality that $J_{n2} = o_P(\{mh_1\}^{-1})$. Moreover, $J_{n3} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\sum_{t=1}^{n_i}\sum_{l=1}^{n_i}\frac{G_{it}^{\mathrm{T}}\Omega_1(U_{it})^{-1}G_{il}}{f(U_{it})}K_{h_1}(U_{il} - U_{it})r_{ij}r_{il}[\Lambda_{0i}^{-1}]_{jt} + o_P(\{mh_1\}^{-1})$. This together with the result about J_{n1} leads to $\mathrm{CV}_1 = m^{-1}\sum_{i=1}^{m}\mathbf{r}_i^{\mathrm{T}}\Lambda_i^{-1}\mathbf{r}_i + \frac{1}{4}\mu_2^2h_1^4\pi_1 + \frac{\nu_0}{mh_1}\{\sigma^2\lambda_1 + \lambda_2\} + o_P\{h_1^4 + \frac{1}{mh_1}\}$. Therefore, when $n^{\frac{1}{2}}h_1^3 \to 0$, $h/h_1 \to 0$ and $\{nh_1\}^{\frac{1}{2}}h^2 \to 0$, we have $\mathrm{CV} = m^{-1}\sum_{i=1}^{m}\mathbf{r}_i^{\mathrm{T}}\Lambda_i^{-1}\mathbf{r}_i + \frac{1}{4}\mu_2^2h_1^4\pi_1 + \frac{\nu_0}{mh_1}\{\sigma^2\lambda_1 + \lambda_2\} + o_P\{h_1^4 + \frac{1}{mh_1}\}$. □

Proof of Theorem 4 The working model mistakenly treated the first element on the diagonal of A_1, denoted by $A_{1(1,1)}$, as functional $A_{1(1,1)}(\cdot)$. That is, the working model is $y_{ij} = F_{ij}^{*\mathrm{T}}\mathbf{b}^* + Z_{i11}X_{ij11}A_{1(1,1)}(U_{ij}) + G_{ij}^{\mathrm{T}}\boldsymbol{\theta}(U_{ij}) + r_{ij}$ where $F_{ij}^*, \mathbf{b}^*$ are $F_{ij}, \mathbf{b}$ with their first component deleted, respectively, while the true model is $y_{ij} = F_{ij}^{\mathrm{T}}\mathbf{b} + G_{ij}^{\mathrm{T}}\boldsymbol{\theta}(U_{ij}) + r_{ij}$.

Then by similar standard arguments as the proof of Theorems 1, 2 and 3, we get that $\mathrm{CV} = m^{-1}\sum_{i=1}^{m}\mathbf{r}_i^{\mathrm{T}}\Lambda_i^{-1}\mathbf{r}_i + \frac{1}{4}\mu_2^2h_1^4\pi_1 + \frac{\nu_0}{mh_1}\{\sigma^2\lambda_1^{(1)} + \lambda_2^{(1)}\} + o_P\{h_1^4 + \frac{1}{mh_1}\}$. □

Proof of Theorem 5 The working model is

$$y_{ij} = F_{ij}^{\mathrm{T}}\mathbf{b} + Z_{i21}X_{ij11}A_{2(1,1)} + G_{ij}^{*\mathrm{T}}\boldsymbol{\theta}^{**}(U_{ij}) + r_{ij}, \tag{32}$$

where $\boldsymbol{\theta}^{**}(u)$ is the $\boldsymbol{\theta}(u)$ with its $(p_2q_1 + 1)$th functional coefficient deleted, while the true model is $y_{ij} = F_{ij}^{*\mathrm{T}}\mathbf{b}^{**} + G_{ij}^{*\mathrm{T}}\boldsymbol{\theta}^{**}(U_{ij}) + r_{ij} + Z_{i21}X_{ij11}\{A_{2(1,1)}(U_{ij}) -$

$A_{2(1,1)}\}$, where $\mathbf{b}^{*\mathrm{T}} = (\mathbf{b}^{\mathrm{T}}, A_{2(1,1)})$. By the same arguments as that in the proof of Theorem 1, it can be shown that $\hat{\mathbf{b}}^{**} - \mathbf{b}^{**} - \boldsymbol{\delta}_1 = O_P(h^2 + n^{-1/2})$, where $\boldsymbol{\delta}_1 = \{\mathbf{F}^{**\mathrm{T}}(I_n - \mathbf{S}^{**})^{\mathrm{T}}\Lambda^{-1}(I_n - \mathbf{S}^{**})\mathbf{F}^{**}\}^{-1}\mathbf{F}^{**\mathrm{T}}(I_n - \mathbf{S}^{**})^{\mathrm{T}}\Lambda^{-1}(I_n - \mathbf{S}^{**})\mathbf{v}$, with $\mathbf{v} = (v_{11}, \ldots, v_{1n_1}, \ldots, v_{m1}, \ldots, v_{mn_m})^{\mathrm{T}}$, $v_{ij} = Z_{i21}X_{ij11}[A_{2(1,1)}(U_{ij}) - A_{2(1,1)}]$, and $\mathbf{F}^{**}$, $\mathbf{S}^{**}$, $\boldsymbol{\Phi}^{**}$ are defined in the same way as $\mathbf{F}$, $\mathbf{S}$, $\boldsymbol{\Phi}$ but based on the working model (32). It is easy to see $\hat{\boldsymbol{\theta}}^{**}(u) - \boldsymbol{\theta}^{**}(u) = \boldsymbol{\delta}_2(u) + O_P(h_1^2 + \{nh_1^2\}^{-1/2})$ uniformly for u, where Δ^{**} is defined in the same way as Δ but based on the working model (32), and $\boldsymbol{\delta}_2(u) = (I_{d-1}, \mathbf{0}_{(d-1)\times(d-1)})(\Delta^{**\mathrm{T}}W_1\Delta^{**})^{-1}\Delta^{**\mathrm{T}}W_1\{\mathbf{v} - \mathbf{F}^{**}\boldsymbol{\delta}_1\}$. The estimated residual r_{ij}^* defined in (16) based on the working model (32) is $r_{ij}^* = r_{ij}^{**} + v_{ij} - F_{ij}^{*\mathrm{T}}\boldsymbol{\delta}_1^{\setminus i} - G_{ij}^{*\mathrm{T}}\boldsymbol{\delta}_2^{\setminus i}(U_{ij})$, where $r_{ij}^{**} = F_{ij}^{*\mathrm{T}}(\mathbf{b}^{**} - \hat{\mathbf{b}}^{**\setminus i} + \boldsymbol{\delta}_1^{\setminus i}) + G_{ij}^{*\mathrm{T}}[\boldsymbol{\theta}^{**}(U_{ij}) - \hat{\boldsymbol{\theta}}^{**\setminus i}(U_{ij}) + \boldsymbol{\delta}_2^{\setminus i}(U_{ij})] + r_{ij}$ is the estimated residual when the working model (32) is the true model, and $\boldsymbol{\delta}_1^{\setminus i}$, $\boldsymbol{\delta}_2^{\setminus i}(U_{ij})$ are, respectively, the $\boldsymbol{\delta}_1$, $\boldsymbol{\delta}_2(U_{ij})$ obtained when the ith cluster is deleted. It can be shown that $\boldsymbol{\delta}_1 - \boldsymbol{\delta}_1^{\setminus i} = o_p(1)$, $\boldsymbol{\delta}_1 \xrightarrow{P} \Pi_3^{-1}\Upsilon_5$, $\boldsymbol{\delta}_2(u) - \boldsymbol{\delta}_2^{\setminus i}(u) = o_P(1)$, and $\boldsymbol{\delta}_2(u) \xrightarrow{P} \Omega_1^*(u)^{-1}\Upsilon_4(u)\{A_{2(1,1)}(u) - A_{2(1,1)}\} - \Omega_1^*(u)^{-1}\Omega_3^*(u)\Pi_3^{-1}\Upsilon_5$ uniformly with respect to u by the straightforward calculation and Lemma 1, hence $r_{ij}^* = r_{ij}^{**} + \gamma_{3ij} + o_P(1)$. By Theorem 3, Lemma 2 and the Markov inequality, we have

$$\begin{aligned}\mathrm{CV} &= m^{-1}\sum_{i=1}^{m}\mathbf{r}_i^{**\mathrm{T}}\Lambda_i^{-1}\mathbf{r}_i^{**} + m^{-1}\sum_{i=1}^{m}\boldsymbol{\gamma}_{3i}^{\mathrm{T}}\Lambda_i^{-1}\boldsymbol{\gamma}_{3i} + 2m^{-1}\sum_{i=1}^{m}\boldsymbol{\gamma}_{3i}^{\mathrm{T}}\Lambda_i^{-1}\mathbf{r}_i^{**} + o_P(1)\\ &= m^{-1}\sum_{i=1}^{m}\mathbf{r}_i^{\mathrm{T}}\Lambda_i^{-1}\mathbf{r}_i + \pi_2 + o_P(1), \qquad (33)\end{aligned}$$

where $\mathbf{r}_i^{**} = (r_{i1}^{**}, \ldots, r_{in_i}^{**})^{\mathrm{T}}$. □

Acknowledgments We thank the Editor, the Associate Editor and referees for helpful suggestions that have improved the original submission. We acknowledge the grants from National Science Foundation of China No.10801093 and Academic Research Funding in Singapore R-155-000-109-112 for supporting our research. We also acknowledge the grants support of National Medical Research Council NMRC/NIG/0054/2009.

References

Chiou, J.-M., Müller, H-G. (2005). Estimated estimating equations: semiparametric inference for clustered/longitudinal data. *Journal of the Royal Statistical Society, Series B, 67*, 531–553.

Demidenko, E. (2004). *Mixed models: Theory and applications*. Hoboken: Wiley.

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

Fan, J., Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *Journal of the American Statistical Association, 99*, 710–723.

Fan, J., Wu, Y. (2008). Semiparametric estimation of covariance matrixes for longitudinal data. *Journal of the American Statistical Association, 103*, 1520–1533.

Fan, J., Zhang, W. (1999). Statistical estimation in varying coefficient models. *The Annals of Statistics, 27*, 1491–1518.

Fan, J., Zhang, J.-T. (2000a). Two-step estimation of functional linear models with applications to longitudinal data, *Journal of the Royal Statistical Society, Series B, 62*, 303–322.

Fan, J., Zhang, W. (2000b). Simultaneous confidence bands and hypothesis testing in varying-coefficient models. *Scandinavian Journal of Statistics, 27*, 715–731.

Fan, J., Huang, T., Li, R. (2007). Analysis of longitudinal data with semiparametric estimation of covariance function. *Journal of American Statistical Association, 102*, 632–641.

Li, J., Zhang, W. (2011). A semiparametric threshold model for censored longitudinal data analysis. *Journal of the American Statistical Association, 106*, 685–696.

Li, J., Zhang, W., Wu, Z. (2011). Optimal zone for bandwidth selection in semiparametric models. *Journal of Nonparametric Statistics, 23*, 701–717.

Mitra, S. N., Al-Sabir, A., Cross, A. R., Jamil, K. (1997). Bangladesh and Demographic Health Survey 1996–1997. Dhaka and Calverton, MD: National Institute of Population Research and Training (NIPORT), Mitra and Associates, and Macro International Inc. Bangladesh.

Sun, Y., Zhang, W., Tong, H. (2007). Estimation of the covariance matrix of random effects in longitudinal studies. *The Annals of Statistics, 35*, 2795–2814.

Wang, L., Bo, K., Li, R. (2009). Local rank inference for varying coefficient models. *Journal of American Statistical Association, 104*, 1631–1645.

Xia, Y., Zhang, W., Tong, H. (2004) Efficient estimation for semivarying-coefficient models. *Biometrika, 91*, 661–681.

Zhang, W., Lee, S. Y. (2000) Variable bandwidth selection in varying-coefficient models. *Journal of Multivariate Analysis, 74*, 116–134.

Zhang, W., Lee, S. Y., Song, X. (2002). Local polynomial fitting in semivarying coefficient models, *Journal of Multivariate Analysis, 82*, 166–188.

Zhang, W., Fan, J., Sun, Y. (2009). A semiparametric model for cluster data. *Annals of Statistics, 37*, 2377–2408.