A SEMIPARAMETRIC MODEL FOR CLUSTER DATA

BY WENYANG ZHANG JIANQING FAN* AND YAN SUN †

University of Bath, Princeton University and Shanghai University of Finance and Economics

In the analysis of cluster data the regression coefficients are frequently assumed to be the same across all clusters. This hampers the ability to study the varying impacts of factors on each cluster. In this paper, a semiparametric model is introduced to account for varying impacts of factors over clusters by using cluster-level covariates. It achieves the parsimony of parametrization and allows the explorations of nonlinear interactions. The random effect in the semiparametric model accounts also for within cluster correlation. Local linear based estimation procedure is proposed for estimating functional coefficients, residual variance, and within cluster correlation matrix. The asymptotic properties of the proposed estimators are established and the methods for constructing simultaneous confidence bands are proposed and studied. In addition, relevant hypothesis testing problems are addressed. Simulation studies are carried out to demonstrate the methodological power of the proposed methods in the finite sample. The proposed model and methods are used to analyse the second birth interval in Bangladesh, leading to some interesting findings.

1. Introduction.

1.1. Preamble. Longitudinal data analysis has attracted considerable attention from statisticians recently. For longitudinal data, the data from the same cluster are dependent with each other. As far as modelling is concerned, this within cluster dependency is usually accounted by random cluster effects and modelled by a within cluster correlation matrix. The within cluster correlation matrix plays a very important role in longitudinal data analysis as it can be used to improve the efficiency of the estimation. Actually, most of the existing literature are devoted to address how to make use of within cluster correlation matrix to improve the estimation for unknown parameters or functions.

*Partially supported by the NIH grant R01-GM07261, and NSF grant DMS-0704337 and DMS-0714554.
†Supported by National Science Foundation of China (grant 10801093).

AMS 2000 subject classifications: Primary 62G08; secondary 62G10; 62G15

Keywords and phrases: Varying-coefficient models, local linear modelling, cluster level variable, cluster effect

1
The methodology for parametric based longitudinal data analysis is quite mature; see e.g. Diggle, Heagerty, Liang and Zeger [6] and the references therein. The situation with nonparametric based longitudinal data analysis is very different. One of the main difficulties is how to incorporate the within cluster correlation structure into the estimation procedure. Lin and Carroll [14] recommend that we ignore the within cluster correlation when kernel smoothing is employed. Welsh, Lin and Carroll [23] investigate the possibility of using weighted least squares based on the within cluster correlation structure when spline smoothing is used. They suggest that the weighted least squares estimator based on the true within cluster correlation structure works better than the estimator based on working independence when spline smoothing is used. Other literature about nonparametric longitudinal regression includes Zeger and Diggle [27], Brumback and Rice [2], Hoover et al. [11], Wu et al. [24], Martinussen and Scheike [17], Chiang et al. [3], Huang et al. [12], Wang [21], Chiu and Müller [4], Wang, Carroll, and Lin [22], Qu and Li [19], Lin and Carroll [15], Sun et al. [20] and others.

Most of the literature assumes the regression parameters or functions are the same across all clusters. However, when the regression effects of some particular clusters are of interest, it is unreasonable to assume the regression parameters or functions are the same across all clusters. The interactions of regression effects with clusters are of interest. One naive method to address this is to let the regression coefficients or functions vary freely over clusters. But, this naive method will not be parsimonious particularly when the number of clusters are large and the issue of estimability arises. In addition, the within cluster dependency may be addressed by the random effect. This leads us to model these cluster-dependent regression coefficients or functions by using cluster level variables. It addresses simultaneously the parsimony and cluster dependency of modeling.

1.2. A motivating example. The data which stimulates this project is from Bangladesh about the second birth interval which is defined as the duration between the first birth and the second birth. The data come from the Bangladesh Demographic and Health Survey (BDHS) of 1996-97 (Mitra et al. [18]), which is a cross-sectional, nationally representative survey of ever-married women aged between 10 and 49. What we are interested in is how some factors which are commonly found to be associated with contraceptive use in Bangladesh, such as education and religion, affect the second birth interval. The data were collected from different districts (clusters) in the six different divisions of Bangladesh. Bangladesh is divided into six administrative divisions which are Barisal, Chittagong, Dhaka, Kulna, Rajshahi and
Sylhet. The data from the same cluster are correlated with each other due to cluster-level factors such as cultural norms and access to family planning programmes. What is of particularly interest is how the factors of interest affect the second birth interval in some particular clusters. For example how these factors affect the second birth interval in a rural area in Chittagong division.

Some of the factors of interest are defined on the individual level such as education and are called individual level variables. Some of the factors are defined on the cluster level such as type of region of residence and are called cluster level variables. We use \( y_{ij} \) to denote the length of the second birth interval of the \( j \)th woman in the \( i \)th cluster, \( X_{ij} \) to denote the vector of the corresponding individual level variables, \( Z_i \) to denote the vector of the cluster level variables.

Frequently, the linear model

\[
(1.1) \quad y_{ij} = X_{ij}^T a + Z_i^T \beta + \epsilon_{ij}, \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, m.
\]

would be used to fit the data. The within cluster dependency would be accounted by that \( \epsilon_{ij}, j = 1, \ldots, n_i \), are correlated. The covariance matrix of \( \epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{in_i})^T \) can be incorporated into the estimation procedure.

Model (1.1) would be fine if the interest focuses only on the global impact of the factors. However, if the picture for a particular cluster is of interest, (1.1) would not be adequate. Let’s take education as an example. It is evident that the impact of education in the cluster where Muslims predominate would be different to the cluster where Hindus predominate. To take the difference of this kind into account, we may relax the assumption imposed on (1.1) and allow the regression coefficient to vary over clusters. This leads to

\[
(1.2) \quad y_{ij} = X_{ij}^T a_i + Z_i^T \beta + \epsilon_{ij}, \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, m.
\]

While it accounts for the varying impact across clusters, (1.2) is not parsimonious. In fact, (1.2) involves \( pm + q \) regression coefficients, where \( p \) and \( q \) are the dimensions of \( X_{ij} \) and \( Z_i \) respectively. When \( m \), the number of clusters, is large, there would be too many unknown parameters in model (1.2) for us to get reasonably accurate estimators. In longitudinal data analysis, we often come across large number of clusters. For example, there are 296 clusters in the second birth interval data set that stimulates this paper. If we use (1.2) to fit the data, we would face \( 296p + q \) unknown coefficients, and would certainly pay a big price on variances of the resulting estimators.

A sensible approach is to model the factor loadings \( a_i \) by using cluster
level variables. A reasonable model is

\[
\begin{aligned}
    y_{ij} &= X_{ij}^T \alpha_i + Z_{ij}^T \beta + \epsilon_{ij}, \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, m, \\
    a_i &= \alpha_0 + AZ_i + e_i, \quad i = 1, \ldots, m,
\end{aligned}
\]

(1.3)

where \( A = (\alpha_1, \ldots, \alpha_q) \), and \( e_i \) \((i = 1, \ldots, m)\) are random effects with mean zero. This achieves simultaneously the parsimony and within cluster dependency and the cluster-dependent factor loadings are allowed. In fact, The number of unknown coefficients in model (1.3) is \( p(q + 1) + q \) which is usually much smaller than \( pm + q \).

A further extension of model (1.3) is to let the factor loadings vary with time as the society and technology evolve with time. By allowing the impacts varying with time, we come up with the model

\[
\begin{aligned}
    y_{ij} &= X_{ij}^T \alpha_i(U_{ij}) + Z_{ij}^T \beta(U_{ij}) + \epsilon_{ij}, \\
    a_i(U_{ij}) &= \alpha_0(U_{ij}) + A(U_{ij})Z_i + e_i,
\end{aligned}
\]

(1.4)

where \( U_{ij} \) is time, \( A(U_{ij}) = (\alpha_1(U_{ij}), \ldots, \alpha_q(U_{ij})) \). Model (1.4) is the model which we are going to address. It is a kind of varying coefficient model (Xia and Li [26]; Fan and Zhang [8], [9]; Zhang et al. [28]; Li and Liang [13]). To make (1.4) mathematically more clear and general, from now on \( U_{ij} \) is not necessarily to be time, it can be any continuous covariate. This allows the nonlinear interaction of individual variables \( X_{ij} \) and cluster level variable \( Z_i \) with \( U_{ij} \). We assume \( \epsilon_{ij} \) is measurement error with mean 0 and variance \( \sigma^2 \), and independent of \( X_{ij}, U_{ij} \) and \( Z_i \), and \{\( e_i \)\} are i.i.d. random effects with mean \( 0_{p\times1} \) and covariance matrix \( \Sigma \), and independent of all other random variables. We assume \{(\( X_{ij}^T, U_{ij} \))\} are i.i.d., and so are \{\( Z_i \)\}.

In (1.4), \( \beta(\cdot), \alpha_k(\cdot), k = 0, \ldots, q \), are unknown functions to be estimated. \( \sigma^2 \) is unknown parameter and \( \Sigma \) is unknown matrix to be estimated.

Although (1.4) is stimulated by the second birth interval data, the modeling concept and estimation methodology, which this paper aims to explore, are equally applicable to other kinds of data, such as the data obtained from medicine and engineering.

The paper is organised as follows. We begin in Section 2 with a description of the estimation procedure for the proposed model (1.4). In Section 3 we establish the asymptotic properties of the proposed estimators. Hypothesis test associated with model (1.4) is discussed in section 4. The performance of the model is assessed by a simulation study in Section 5. In Section 6, we use the proposed model and estimation procedure to analyse the data mentioned about the second birth intervals in Bangladesh, and explore how
the impacts of the factors of interest on the length of second birth intervals in some particular clusters change over time.

2. Estimation procedure. In this section, we are going to construct the estimation procedure for the proposed model (1.4). We estimate the unknown functional coefficients first, then, $\sigma^2$ and $\Sigma$.

2.1. Estimation of functional coefficients. By Taylor’s expansion we have

\[ \alpha_k(U_{ij}) \approx \alpha_k(u) + \dot{\alpha}_k(u)(U_{ij} - u), \quad k = 0, \cdots, q, \]

\[ \beta(U_{ij}) \approx \beta(u) + \dot{\beta}(u)(U_{ij} - u), \]

when $U_{ij}$ is in a small neighbourhood of $u$, which leads to the following local least squares estimation procedure

\[
L = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left( y_{ij} - X_{ij}^T \left[ \sum_{k=0}^{q} \left\{ b_k + c_k(U_{ij} - u) \right\} z_{ik} \right] \right. \\
\left. - Z_i^T \left\{ J + d(U_{ij} - u) \right\} \right)^2 K_h(U_{ij} - u), \tag{2.1}
\]

where $z_{i0} = 1$, $K_h(\cdot) = K(\cdot/h)/h$, $h$ is bandwidth and $K(\cdot)$ is kernel function. We minimise $L$ with respect to $J$, $d$, $b_k$, $c_k$, $k = 0, \cdots, q$, to get the minimiser $\hat{J}$, $d$, $\hat{b}_k$, $\hat{c}_k$, $k = 0, \cdots, q$. We use $\hat{b}_k$ to estimate $\alpha_k(u)$, and $\hat{J}$ to estimate $\beta(u)$. From now on, we denote $\hat{b}_k$ by $\hat{\alpha}_k(u)$, and $\hat{J}$ by $\hat{\beta}(u)$.

Let

\[ x_i = (X_{i1}, \cdots, X_{in_i})^T, \quad \Gamma_i = \left( x_i, \left\{ (1, Z_i^T) \otimes I_p \right\}, 1_{n_i} \otimes Z_i^T \right), \]

\[ \Gamma = (\Gamma_1^T, \cdots, \Gamma_m^T)^T, \quad X = (\Gamma, U_1 \Gamma), \]

\[ U_i = \text{diag} \left( (U_{11} - u)^i, \cdots, (U_{1n_1} - u)^i, \cdots, (U_{m1} - u)^i, \cdots, (U_{mn_m} - u)^i \right) \]

\[ W = \text{diag} \left( K_h(U_{11} - u), \cdots, K_h(U_{1n_1} - u), \cdots, K_h(U_{m1} - u), \cdots, K_h(U_{mn_m} - u) \right) \]

\[ Y = (y_{11}, \cdots, y_{1n_1}, \cdots, y_{m1}, \cdots, y_{mn_m}) \]

where $I_p$ is size $p$ identity matrix, $1_d$ is a $d$ dimensional vector with each component being 1. By simple calculation, we have

\[
\hat{\alpha}_k(u) = A_k Y \quad (k = 0, \cdots, q), \quad \hat{\beta}(u) = B Y, \tag{2.2}
\]

where

\[ A_k = \left( e_{(k+1),(q+1)}^T \otimes I_p, 0_{p \times (q+1)} \right) \left( X^T W X \right)^{-1} X^T W, \]
\[
\mathbf{B} = (\mathbf{0}_{q \times ((q+1)p)}, I_q, \mathbf{0}_{q \times s}) \left( \mathbf{X}^T \mathbf{W} \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{W},
\]

with \( e_{k,p} \) denoting the unit vector of length \( p \) with 1 at position \( k \), \( \mathbf{0}_{p \times q} \) the size \( p \times q \) matrix with all entries 0 and \( s = (q+1)p + q \).

In practice, some times, some coefficients in (1.4) are constant. Under such situation, model (1.4) becomes a semivarying coefficient mixed effects model. An interesting question is how to estimate the constant coefficients. Fan and Zhang [8] studied a varying coefficient model with coefficients having different degrees of smoothness. They proposed a two-stages estimation procedure. Based on their idea, we propose a very simple estimation procedure for the unknown constant coefficients. Suppose the \( j \)th component \( \alpha_{kj}(u) \) of \( \alpha_{k}(u) \) is a constant which is denoted by \( C_{kj} \). We first pretend \( \alpha_{kj}(u) \) is a function and get the estimator \( \hat{\alpha}_{kj}(U_{il}) \) of \( \alpha_{kj}(u) \) at \( U_{il} \), \( l = 1, \cdots, n_i \), \( i = 1, \cdots, m \), by the above estimation procedure. Then, take average of \( \hat{\alpha}_{kj}(U_{il}) \) over \( l = 1, \cdots, n_i \), \( i = 1, \cdots, m \). This average

\[
\hat{C}_{kj} = \frac{1}{n} \sum_{i=1}^{m} \sum_{l=1}^{n_i} \hat{\alpha}_{kj}(U_{il}), \quad n = \sum_{i=1}^{m} n_i
\]

is our estimator of \( C_{kj} \). We will show it later that the convergence rate of this estimator is of order \( O_P(n^{-1/2}) \) when the bandwidth is properly selected. This provides a simple method for estimating the constant coefficients.

2.2. Estimation of \( \sigma^2 \) and \( \Sigma \). Let \( \tilde{a}_i(U_{ij}) = \alpha_0(U_{ij}) + \sum_{k=1}^{q} \alpha_k(U_{ij}) z_{ik} \) and \( \mathbf{r}_i = (r_{i1}, \cdots, r_{in_i})^T \), where \( r_{ij} = y_{ij} - X_{ij}^T \tilde{a}_i(U_{ij}) - Z_{ij}^T \beta(U_{ij}) \). Correspondingly, let \( \tilde{a}_i(\cdot) \) and \( \hat{r}_i \) be their substitution estimators. Set

\[
\mathbf{x}_i = (X_{i1}, \cdots, X_{in_i})^T \quad \text{and} \quad P_i = \mathbf{x}_i (\mathbf{x}_i^T \mathbf{x}_i)^{-1} \mathbf{x}_i^T.
\]

For each given \( i \), we have the following linear model

\[
\mathbf{r}_i = \mathbf{x}_i \mathbf{e}_i + \mathbf{e}_i, \quad \mathbf{e}_i = (\epsilon_{i1}, \cdots, \epsilon_{in_i})^T.
\]

The residual sum of squares of this linear model

\[
\text{rss}_i = \mathbf{r}_i^T (I_{n_i} - P_i) \mathbf{r}_i
\]

would be the raw material for estimating \( \sigma^2 \). The degree of freedom of rss \( i \) is \( n_i - p \). Let RSS \( \text{i} \) be rss \( i \) with \( \mathbf{r}_i \) replaced by \( \hat{\mathbf{r}}_i \). Pooling all \{RSS \( \text{i} \)\} together leads to the estimator of \( \sigma^2 \) as

\[
\hat{\sigma}^2 = (n - mp)^{-1} \sum_{i=1}^{m} \text{RSS}_i, \quad n = \sum_{i=1}^{m} n_i.
\]
Finally, we estimate $\Sigma$. From (2.4), we have the least squares estimator of $e_i$ as

$$\hat{e}_i = (x_i^T x_i)^{-1} x_i^T r_i = e_i + (x_i^T x_i)^{-1} x_i^T \epsilon_i$$

which leads to

$$\sum_{i=1}^m \hat{e}_i \hat{e}_i^T = \sum_{i=1}^m e_i e_i^T + \sum_{i=1}^m (x_i^T x_i)^{-1} x_i^T \epsilon_i e_i^T x_i (x_i^T x_i)^{-1} + \sum_{i=1}^m (x_i^T x_i)^{-1} x_i^T \epsilon_i \epsilon_i^T x_i (x_i^T x_i)^{-1} + \sum_{i=1}^m e_i e_i^T x_i (x_i^T x_i)^{-1}.$$

The last two terms are of order $O_P(m^{1/2})$, so they are negligible. Hence,

$$m^{-1} \sum_{i=1}^m e_i e_i^T \approx m^{-1} \left\{ \sum_{i=1}^m \hat{e}_i \hat{e}_i^T - \sum_{i=1}^m (x_i^T x_i)^{-1} x_i^T \epsilon_i e_i^T x_i (x_i^T x_i)^{-1} \right\} \approx m^{-1} \left\{ \sum_{i=1}^m \hat{e}_i \hat{e}_i^T - \sigma^2 \sum_{i=1}^m (x_i^T x_i)^{-1} \right\}. $$

Therefore, we have

$$(2.5) \quad \hat{\Sigma} = m^{-1} \sum_{i=1}^m \hat{e}_i \hat{e}_i^T - m^{-1} \sigma^2 \sum_{i=1}^m (x_i^T x_i)^{-1}$$

to estimate $\Sigma$. In (2.5), $\hat{e}_i$ is $\hat{e}_i$ with $r_i$ replaced by $\hat{r}_i$.

3. **Asymptotic properties.** In this section, we are going to present the asymptotic properties of the proposed estimators. For any $p \times p$ symmetric matrix $A$, we use $\text{vech}(A)$ to denote the vector consisting of all elements on and below the diagonal of the matrix $A$, $\text{vec}(A)$ to denote the vector by simply stacking the column vectors of matrix $A$ below one another. Obviously, there exists a unique $p^2 \times p(p+1)/2$ matrix $R_p$ such that $\text{vec}(A) = R_p \text{vech}(A)$.

We first introduce some notation. For any function or function vector $g(\cdot)$, we use $\dot{g}(\cdot)$ and $\ddot{g}(\cdot)$ to denote its first and second derivatives respectively. We use $\mathcal{D}$ to denote the collections of all individual and cluster level covariates. Let $\mu_i = \int t^i K(t) dt$, $\nu_i = \int t^i K^2(t) dt$, and

$$\Omega(u) = E \left\{ \left( X^T, Z^T \otimes X^T, Z^T \right)^T \left( X^T, Z^T \otimes X^T, Z^T \right) \middle| U = u \right\} = \begin{pmatrix} \Omega_1(u) & \Omega_2(u) \\ \Omega_2(u)^T & \Omega_3(u) \end{pmatrix},$$

where $\Omega_1(u)$ and $\Omega_2(u)$ are $p \times p$ matrices, and $\Omega_3(u)$ is a $p \times p$ matrix.
where \( \Omega_1(u) \) and \( \Omega_3(\cdot) \) are respectively \( p(q+1) \times p(q+1) \) and \( q \times q \) submatrix of \( \Omega(u) \). Without loss of generality, we will assume that \( \mu_0 = 1 \).

Our main asymptotic results are presented through the following 6 theorems. We leave the proofs of these theorems in Appendix. The first two theorems give the asymptotic normality of the estimated functional coefficients.

**Theorem 1.** Under the conditions (1)-(6) in Appendix, when \( nh^5 \) is bounded, for \( k = 0, \cdots, q \), we have

\[
\sqrt{nhf(u)} \left\{ \hat{\alpha}_k(u) - \alpha_k(u) - h^2 \frac{\mu_2}{2} \hat{\alpha}_k(u) \right\} \xrightarrow{D} N_p \left( 0, \nu_0 \{ e_{(k+1),(q+1)}^T \otimes I_p \} \right),
\]

where

\[
\Lambda_1(u) = \Omega_1(u) - \Omega_2(u)\Omega_3(u)^{-1}\Omega_2(u)^T, \quad \Theta_1(u) = \Upsilon_1(u)\Xi_1(u)\Upsilon_1(u)^T
\]

\[
\Upsilon_1(u) = (\Lambda_1(u)^{-1}, \Lambda_2(u)), \quad \Lambda_2(u) = -\Lambda_1(u)^{-1}\Omega_2(u)\Omega_3(u)^{-1},
\]

and

\[
\Xi_1(u) = E \left\{ X^T \Sigma X (X^T, Z^T \otimes X^T, Z^T)^T (X^T, Z^T \otimes X^T, Z^T) \big| U = u \right\}.
\]

**Theorem 2.** Under the conditions (1)-(6) in Appendix, when \( nh^5 \) is bounded, we have

\[
\sqrt{nhf(u)} \left\{ \hat{\beta}(u) - \beta(u) - h^2 \frac{\mu_2}{2} \hat{\beta}(u) \right\} \xrightarrow{D} N_q \left( 0, \nu_0 \left[ \sigma^2 \Lambda_3(u) + \Theta_2(u) \right] \right),
\]

where \( \Lambda_3(u) = \Omega_3(u)^{-1} + \Omega_3(u)^{-1}\Omega_2(u)^T\Lambda_1(u)^{-1}\Omega_2(u)\Omega_3(u)^{-1}, \Theta_2(u) = \Upsilon_2(u)\Xi_1(u)\Upsilon_2(u)^T, \) and \( \Upsilon_2(u) = \left( \Lambda_2(u)^T, \Lambda_3(u) \right) \)

We now present the asymptotic normality for the parametric component. To present the asymptotic property of \( \sigma^2 \), we assume that the following limits exist and finite: \( c_1 = \lim_{m \to \infty} n/(n - mp) \), \( c_2 = \lim_{m \to \infty} m/(n - mp) \) and

\[
\gamma = \lim_{n \to \infty} (n - mp)^{-1} \sum_{i=1}^{m} \sum_{j=1}^{n_1} E \left[ X_{ij}^T (x_i^T x_i)^{-1} x_{ij} \right]^2.
\]
Theorem 3. Under the conditions (1)-(6) in Appendix, when \( nh^8 \to 0 \), we have
\[
\sqrt{n} \left\{ \hat{\sigma}^2 - \sigma^2 \right\} \xrightarrow{D} N\left(0, 2\sigma^4 c_1(c_1 - 1 - \gamma) + \text{var}(\hat{c}_{11})c_1(2 - c_1 + \gamma)\right).
\]

Theorem 3 suggests the estimator \( \hat{\sigma}^2 \) is of convergence rate \( O_P(n^{-1/2}) \) which is the optimal convergence rate of parametric estimator.

Additional notation is needed for presenting the asymptotic normality of \( \Sigma \). Write
\[
\Delta_1 = \lim_{m \to \infty} m^{-1} \sum_{i=1}^{m} \mathbb{E} \left[ (x_i^T x_i)^{-1} \right],
\]
\[
\Delta_2 = \lim_{m \to \infty} m^{-1} \sum_{i=1}^{m} \mathbb{E} \left[ (x_i^T x_i)^{-1} \otimes (x_i^T x_i)^{-1} \right],
\]
\[
\Delta_3 = \begin{pmatrix}
\Sigma \otimes \Delta_1^{(1)} + \Delta_1 \otimes \Sigma_{(1)} \\
\vdots \\
\Sigma \otimes \Delta_1^{(p)} + \Delta_1 \otimes \Sigma_{(p)}
\end{pmatrix},
\]
where \( \Delta_1^{(r)}, \Sigma_{(r)} \) \( (r = 1, \cdots, p) \) denote the \( r \)th row of \( \Delta_1 \), \( \Sigma \) respectively.

\[
\Delta_4 = \lim_{m \to \infty} m^{-1} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \mathbb{E} \left[ \text{vec} \left( (x_i^T x_i)^{-1} X_{ij} X_{ij}^T (x_i^T x_i)^{-1} \right) \right. \\
\times \left. \text{vec} \left( (x_i^T x_i)^{-1} X_{ij} X_{ij}^T (x_i^T x_i)^{-1} \right)^T \right],
\]
\[
\Delta_5 = \lim_{m \to \infty} m^{-1} \sum_{i=1}^{m} \mathbb{E} \left[ (x_i^T x_i)^{-1} x_i^T \tilde{P}_i x_i (x_i^T x_i)^{-1} \right],
\]
where \( \tilde{P}_i \) is a diagonal matrix generated from the diagonal elements of \( P_i \).

Theorem 4. Under the conditions (1)-(6) in Appendix, when \( nh^8 \to 0 \), we have
\[
\sqrt{n} \ \text{vech} \left( \hat{\Sigma} - \Sigma \right) \xrightarrow{D} N_r \left( \mathbf{0}_{r \times 1}, (1/c_2 + p)(R_p^T R_p)^{-1} R_p^T \Delta R_p \left( R_p^T R_p \right)^{-1} \right),
\]
where \( r = p(p + 1)/2 \) and
\[
\Delta = \mathbb{E} \left\{ \left( \mathbf{e}_1 \mathbf{e}_1^T \right) \otimes \left( \mathbf{e}_1 \mathbf{e}_1^T \right) \right\} - \text{vec}(\Sigma) \text{vec}(\Sigma)^T + 2\sigma^4 \Delta_2 + \sigma^2 \left\{ \Sigma \otimes \Delta_1 \right\}.
\]
\[ + \Delta_1 \otimes \Sigma + \Delta_3 \} + \{ \text{var}(\epsilon_{11}^2) - 2\sigma^4 \} \{ \Delta_4 - c_2 \{ 2\text{vec}(\Delta_1)\text{vec}(\Delta_1)^T \\
\quad - \text{vec}(\Delta_1)\text{vec}(\Delta_2)^T - \text{vec}(\Delta_2)\text{vec}(\Delta_1)^T \} \}
\quad + \{ 2\sigma^4 c_1 (c_1 - 1 - \gamma) + \text{var}(\epsilon_{11}^2) c_1 (2 - c_1 + \gamma) \} \text{vec}(\Delta_1)\text{vec}(\Delta_1)^T. \]

Theorem 4 shows the estimator \( \hat{\Sigma} \) also achieves the convergence rate of \( O_P(n^{-1/2}) \).

**Theorem 5.** When \( \alpha_{kj}(u) \) is a constant, under the conditions (1)-(7) in Appendix and \( nh^4 \to 0 \), we have

\[
\sqrt{n} \left\{ \hat{C}_{kj} - C_{kj} \right\} \xrightarrow{D} N(0, \epsilon_{kp+j,(q+1)p}) \{ \sigma^2 E[\Lambda_1(U)^{-1}] + E[\Theta_1(U)] + \Xi_2 \} \epsilon_{kp+j,(q+1)p},
\]

where

\[
\Xi_2 = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{m} \sum_{l=1}^{n_i} \sum_{r=1, r \neq l}^{n_i} E \left\{ \Upsilon_1(U_{il}) E[\Gamma^T_i \Sigma \Gamma \Gamma^T_{ir} U_{il}, U_{ir}] \Upsilon_1(U_{ir})^T \right\},
\]

with \( \Gamma^T_{ij} = (X^T_{ij}, Z^T_i \otimes X^T_{ij}, Z^T_i) \) being the \( j \)th row of matrix \( \Gamma_i \).

When the \( j \)th \((j = 1, \cdots, q)\) component \( \beta_j(u) \) of \( \beta(u) \) is a constant, Theorem 5 applies to its estimator as well except that the variance is

\[
\epsilon_{j,q}^T \left\{ \sigma^2 E[\Lambda_3(U)^{-1}] + E[\Theta_2(U)] + \Xi_3 \} \epsilon_{j,q},
\]

where

\[
\Xi_3 = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{m} \sum_{l=1}^{n_i} \sum_{r=1, r \neq l}^{n_i} E \left\{ \Upsilon_2(U_{il}) E[\Gamma^T_i \Sigma \Gamma \Gamma^T_{ir} U_{il}, U_{ir}] \Upsilon_2(U_{ir})^T \right\}.
\]

**Theorem 6.** Under the conditions (1)-(8) in Appendix with \( K(t) \) having a compact support \([-c_0, c_0]\), and \( h = n^{-\rho}, 1/5 \leq \rho < 1/3 \), for all \( u \in [a, b] \) we have

\[
P\left( (-2 \log \{ h/(b-a) \})^{\frac{1}{2}} \left\{ \sup_{u \in [a,b]} \frac{\hat{\alpha}_{kj}(u) - \alpha_{kj}(u) - \text{bias}(\hat{\alpha}_{kj}(u)|D)}{\text{var}(\hat{\alpha}_{kj}(u)|D)^{\frac{1}{2}}} - \omega_n \right\} - x \right) \to \exp\{-2\exp\{-x\}\},
\]
where
\[
\omega_n = \left(-2 \log \left\{ \frac{h}{(b-a)} \right\}\right)^\frac{1}{2} + \frac{1}{\left(-2 \log \left\{ \frac{h}{(b-a)} \right\}\right)^\frac{1}{2}} \left[ \log \frac{K^2(c_0)}{\nu_0 \pi^2} + \frac{1}{2} \log \log \left\{ \frac{(b-a)/h}{(b-a) + \alpha} \right\} \right]
\]
if \( K(c_0) \neq 0 \), and
\[
\omega_n = \left(-2 \log \left\{ \frac{h}{(b-a)} \right\}\right)^\frac{1}{2} + \frac{1}{\left(-2 \log \left\{ \frac{h}{(b-a)} \right\}\right)^\frac{1}{2}} \log \left\{ \frac{1}{4\nu_0 \pi} \int (\dot{K}(t))^2 dt \right\}
\]
if \( K(c_0) = 0 \), \( K(t) \) is absolutely continuous and \( K^2(t), (\dot{K}(t))^2 \) are integrable on \( (-\infty, +\infty) \).

**Remark:** Theorem 6 gives the distribution of the maximum discrepancy between the estimated functional coefficient and true coefficient. It is the basis for constructing hypothesis test or confidence band. Theorem 6 also applies to the estimator of any component of \( \beta(\cdot) \).

4. **Confidence bands and hypothesis test.** In this section, we will investigate how to construct the confidence bands for the functional coefficients in model (1.4).

For model (1.4), we often wish to know if an estimated functional coefficient is significantly away from zero or if the estimated functional coefficient is really varying. More generally, we wish to test

\[
H_0 : \alpha_{kj}(u) = \alpha_0(u) \quad \longleftrightarrow \quad H_1 : \alpha_{kj}(u) \neq \alpha_0(u),
\]
where \( \alpha_0(u) \) is a specific function. In this section, we are going to construct a test for such kind of hypothesis.

As the proposed confidence bands and hypothesis test involve the estimation of the biases and variances of the proposed estimators of the functional coefficients, we first construct the estimation procedure for the biases and variances. Throughout this paper, for any functional vector \( \mathbf{g}(u) \), we use \( \mathbf{g}^{(i)}(u) \) to denote the \( i \)th derivative of \( \mathbf{g}(u) \).

4.1. **Estimation for bias and variance.** Following Fan and Gijbels [7], by (2.2), we have for \( k = 0, \ldots, q \), that

\[
\begin{align*}
\text{bias (} \mathbf{\alpha}(u) \mid \mathcal{D} \text{)} & = (e_{(k+1),(q+1)}^{T} \otimes I_p, \mathbf{0}_{p \times (q+s)}) \left( X^T W X \right)^{-1} X^T W R \\
\text{bias (} \mathbf{\beta}(u) \mid \mathcal{D} \text{)} & = (\mathbf{0}_{q \times ((q+1)p)}, I_q, \mathbf{0}_{q \times s}) \left( X^T W X \right)^{-1} X^T W R,
\end{align*}
\]
where $R = (R_{11}, \cdots, R_{1n}, \cdots, R_{m1}, \cdots, R_{mnm})^T$ with
\[
R_{ij} = X_{ij}^T A_i(U_{ij}) + Z_i^T \left\{ \beta(U_{ij}) - \beta(u) - \hat{\beta}(u)(U_{ij} - u) \right\},
\]
\[
A_i(U_{ij}) = \sum_{k=0}^{q} \{ \alpha_k(U_{ij}) - \alpha_k(u) - \hat{\alpha}_k(u)(U_{ij} - u) \} z_{ik}, \quad z_{i0} = 1.
\]

By Taylor’s expansion, we have
\[
\alpha_k(U_{ij}) - \alpha_k(u) - \hat{\alpha}_k(u)(U_{ij} - u) \approx 2^{-1} \tilde{\alpha}_k(u)(U_{ij} - u)^2 + 6^{-1} \alpha_k^{(3)}(u)(U_{ij} - u)^3,
\]
\[
\beta(U_{ij}) - \beta(u) - \hat{\beta}(u)(U_{ij} - u) \approx 2^{-1} \tilde{\beta}(u)(U_{ij} - u)^2 + 6^{-1} \beta^{(3)}(u)(U_{ij} - u)^3.
\]
This leads to
\[
(4.3) \quad R \approx (2^{-1} U_2 \Gamma, 6^{-1} U_3 \Gamma) \left( \begin{array}{c} \hat{\theta}(u)^T, \\ \theta^{(3)}(u)^T \end{array} \right)^T,
\]
where $\theta(u) = (\alpha_0(u)^T, \cdots, \alpha_p(u)^T, \beta(u)^T)^T$. The estimators of $\hat{\theta}(u)$ and $\theta^{(3)}(u)$ can be easily obtained by using local cubic fit with an appropriate pilot bandwidth $h^* (=O(n^{-1/7}))$ in Section 2.1 which is optimal for estimating $\hat{\theta}(u)$. We denote the estimators by $\hat{\theta}(u)$ and $\theta^{(3)}(u)$. Substituting them into (4.3), we obtain $\hat{R}$ and hence an estimated biases $\hat{\text{bias}}(\hat{\alpha}_k(u)|D)$ and $\hat{\text{bias}}(\hat{\beta}(u)|D)$.

We now derive an estimator of variance of $\hat{\alpha}_k(u)$ and $\hat{\beta}(u)$. Noticing that the estimators are linear in $Y$, according to (2.2). Hence, we need only to estimate $\text{var}(Y)$. A natural estimator is
\[
\hat{\text{var}}(Y|D) = \hat{\sigma}^2 I_n + \text{diag} \left( x_1 \hat{\Sigma} x_1^T, \cdots, x_m \hat{\Sigma} x_m^T \right).
\]
This together with (2.2) give us the following estimators
\[
\hat{\text{var}}(\hat{\alpha}_k(u)|D) = A_k \hat{\text{var}}(Y|D) A_k, \quad \text{and} \quad \hat{\text{var}}(\hat{\beta}(u)|D) = B \hat{\text{var}}(Y|D) B,
\]
with $A_k$ and $B$ defining in (2.2).

4.2. Confidence bands and hypothesis test. We first state the Theorem based on which the confidence bands and hypothesis test are constructed.

**Theorem 7.** Under the conditions (1)-(8) in Appendix with $K(t)$ having a compact support $[-c_0, c_0]$, and $h = n^{-\rho}$, $1/5 \leq \rho < 1/3$. Furthermore, if
\( \alpha_{kj}^{(3)}(\cdot) \) is continuous on \([a, b]\) and the pilot bandwidth \( h^* \) is of order \( n^{-1/7} \), then for all \( u \in [a, b] \) we have

\[
P\left( \left( -2 \log \{ h/(b-a) \} \right)^{1/2} \left\{ \sup_{u \in [a, b]} \left| \hat{\alpha}_{kj}(u) - \alpha_{kj}(u) - \text{bias}(\hat{\alpha}_{kj}(u) | D) \right| \right\} - \omega_n \right) < x \to \exp \left\{ -2 \exp \left\{ -x \right\} \right\},
\]

where \( \omega_n \) is exactly the same as that in Theorem 6.

**Remark:** If the component of \( \beta^{(3)}(\cdot) \) is continuous on \([a, b]\), Theorem 7 applies to its estimator as well.

Based on Theorem 7, the \( 1 - \alpha \) confidence bands of \( \alpha_{kj}(u) \) can be easily constructed as

\[
\left( \hat{\alpha}_{kj}(u) - \text{bias}(\hat{\alpha}_{kj}(u) | D) \pm \Delta_{1,\alpha}(u) \right),
\]

where

\[
\Delta_{1,\alpha}(u) = \left( \omega_n + \left[ \log 2 - \log \{-\log(1 - \alpha)\} \right] \left( -2 \log \{ h/(b-a) \} \right)^{-1/2} \right) \times \left\{ \text{var}(\hat{\alpha}_{kj}(u) | D) \right\}^{1/2}.
\]

It is worth to mention that Xia [25] investigated bias-corrected confidence bands for univariate nonparametric regression.

By Theorem 7, hypothesis (4.1) can be tested by the test statistic

\[
\left( -2 \log \{ h/(b-a) \} \right)^{1/2} \left\{ \sup_{u \in [a, b]} \left| \hat{\alpha}_{kj}(u) - \alpha_0(u) - \text{bias}(\hat{\alpha}_{kj}(u) | D) \right| \right\} - \omega_n \}
\]

and rejecting \( H_0 \) when the test statistic exceeds the asymptotic critical value \( c_\alpha = -\log \{-0.5 \log(1 - \alpha)\} \). Similarly, we may want to ask the question whether a specific functional coefficient is really varying. This amounts to testing the composite null hypothesis

\[
(4.4) \quad H_0 : \alpha_{kj}(u) = C_{kj} \quad \iff \quad H_1 : \alpha_{kj}(u) \neq C_{kj},
\]

Based on Theorems 5 and 7, we test the problem (4.4) by computing the statistic

\[
\left( -2 \log \{ h/(b-a) \} \right)^{1/2} \left\{ \sup_{u \in [a, b]} \left| \hat{\alpha}_{kj}(u) - \hat{C}_{kj} - \text{bias}(\hat{\alpha}_{kj}(u) | D) \right| \right\} - \omega_n \}
\]

and rejecting \( H_0 \) for large values of the test statistic.
5. Simulation study. In this section, we are going to use a simulated example to demonstrate how well the proposed estimation method works, and examine how much loss one could incur if ignoring the structure of \( a_i(\cdot) \) when the structure holds. We demonstrate the accuracy of the proposed estimators first.

In model (1.4), we take \( p = 3, q = 2 \) and \( m = 100 \). The cluster sizes \( n_i (i = 1, \ldots, m) \) are generated by the integer part of \( |2\xi| + 6, \xi \sim N(0, 1) \). The covariates \( \{X_{ij}\} \) are independently generated from \( N(0, I_p) \), \( \{Z_i\} \) are independently generated from \( N(0, I_q) \), and \( \{U_{ij}\} \) are independently generated from \( U(0, 1) \). We also set the random effect \( e_i \) following normal distribution \( N(0_2, \Sigma) \), measurement error \( \epsilon_{ij} \) following normal distribution \( N(0, \sigma^2) \) where \( \Sigma = (\sigma_{ij}) = 0_{2 \times 2} \) and \( \sigma = 0.5 \). We set \( \beta_0(u) = \sin(2\pi u) \) (intercept term), \( \alpha_0(u) = (\alpha_{01}(\cdot), \alpha_{02}(\cdot), \alpha_{03}(\cdot))^T \) a vector with each component being \( \sin(2\pi u) \), \( \alpha_1(u) = (\alpha_{11}(\cdot), \alpha_{12}(\cdot), \alpha_{13}(\cdot))^T \) a vector with each component being \( \cos(2\pi u) \), \( \alpha_2(u) = (\alpha_{21}(\cdot), \alpha_{22}(\cdot), \alpha_{23}(\cdot))^T \) a vector with each component being \( \sin(\pi u) \), \( \beta(u) = (\beta_1(\cdot), \beta_2(\cdot))^T \) a vector with each component being \( \sin(2\pi u) \).

For any function or functional vector \( g(\cdot) \), if \( \hat{g}(\cdot) \) is an estimator of \( g(\cdot) \), we define the mean integrated squared error (MISE) of \( \hat{g}(\cdot) \) as

\[
E \left[ \int \| \hat{g}(u) - g(u) \|^2 du \right],
\]

where \( \| b \|^2 = b^T b \). In this paper, we use the MISE to assess the accuracy of the estimators.

The proposed estimation method is employed to estimate the functional coefficients using the kernel function involved in the estimation procedure is taken to be Epanechnikov kernel \( 0.75(1 - t^2)_+ \) and bandwidth is taken to be 0.15. The MISEs of the proposed estimators of unknown functions are presented in Table 1, based on 100 simulations. The MSEs of the estimators of \( \Sigma \) and \( \sigma^2 \) are presented in Table 2. From Tables 1 and 2, we can see the proposed estimators are quite accurate.

To visualize the accuracy of the proposed estimators, among the 100 simulations, we single out the one with median performance, and plot the estimated functions together with their 95% confidence bands in Fig. 1. It shows again the proposed estimators are very accurate.

Now, we turn to examine how much loss one could incur when ignoring the structure of \( a_i(\cdot) \), namely examining the gain of our model (1.4).

It is impossible to estimate \( E(a_i(\cdot)) \) if ignoring the structure of \( a_i(\cdot) \) when the cluster effects exist. So, we now assume the cluster effects \( e_i = 0 \). Given the sizes of the clusters in the above simulated example, some clusters would
be too small for one to estimate the corresponding $a_i(\cdot)$ when ignoring the structure of $a_i(\cdot)$. So, we now assume all clusters share the same size of 50, and keep the number of clusters 100. Except the change on the sizes of the clusters and the assumption of $e_i = 0$, all remain the same as the above simulated example.

We use $\text{MISE}_{1,i}$ to denote the MISE of the estimator of $a_i(\cdot)$ obtained by the proposed estimation method, and $\text{MISE}_{2,i}$ the MISE of the estimator obtained without using the structure of $a_i(\cdot)$, i.e., regarding $a_i(\cdot)$ as a free unknown function and estimating it base on the first part of model (1.4). The ratio

$$\text{RMISE} = \frac{\sum_{i=1}^{m} \text{MISE}_{1,i}}{\sum_{i=1}^{m} \text{MISE}_{2,i}}$$

is used to assess the loss incurred on the estimation for $a(\cdot) = (a_1(\cdot), \cdots, a_m(\cdot))$ due to ignoring the structure of $a_i(\cdot)$, $i = 1, \cdots, m$. We compute the RMISEs for different bandwidths, and plot the obtained RMISEs against the bandwidths in Fig. 2. It is clear, the RMISE is almost 0 when the bandwidth is less than 0.3, and it never goes beyond 0.25, which suggests the loss is significant. Similarly, we define the RMISE for $\beta(\cdot)$, and plot the RMISEs against different bandwidths, it again shows the loss incurred on the estimation for $\beta(\cdot)$ due to ignoring the structure of $a_i(\cdot)$, $i = 1, \cdots, m$, is still significant.

6. Real data analysis. The data we study come from the Bangladesh Demographic and Health Survey (BDHS) of 1996-97 (Mitra et al. [18]), 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 8189 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, 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 the first birth and the second birth for the $j$th woman in the $i$th 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 (none coded by 0, primary or secondary plus coded by 1) denoted by $x_{ij1}$, religion (Hindu coded by 1, Muslim or other coded by 0) denoted by $x_{ij2}$, first child (boy coded by 1) denoted by $x_{ij3}$. 
$X_{ij} = (x_{ij1}, x_{ij2}, x_{ij3})^T$. Another individual-level covariate is the year of marriage ($U_{ij}$). We also consider two cluster-level variables, administrative division and type of region of residence (rural coded by 1, urban coded by 0). We take urban as the reference and the differences between urban and rural clusters are modeled by a dummy variables $z_{i1}$. We take Barisal as the reference and the differences among the six administrative divisions are modeled by a set of dummy variables $z_{il}$, $l = 2, \ldots, 6$.

The proposed model (1.4) is used to fit the data, and the proposed estimation procedure is employed to estimate $\alpha_j(\cdot) = (\alpha_{j1}(\cdot), \alpha_{j2}(\cdot), \alpha_{j3}(\cdot))^T$, $j = 0, \ldots, 6$, and $\beta(\cdot) = (\beta_1(\cdot), \ldots, \beta_6(\cdot))^T$. The kernel involved in the estimation is taken to be Epanechnikov kernel, and the bandwidth is chosen to be 35% of the range of $U_{ij}$.

First, we examine how strong the random effect on each cluster is. To this end, for each cluster, we estimate the random effect of this cluster. The density function of the norm of the random effects of all clusters is presented in Fig. 2. It is easy to see the random effects are close to zero. This indicates that the within cluster correlation has been mainly accounted by the deterministic cluster effect in (1.4).

As we mentioned before, if a coefficient is treated as a function when it is constant, we would pay a price on the variances of the resulting estimators. Thus, for each coefficient in the model, the proposed hypothesis test is employed to test whether it is constant or not. The P-value for each coefficient is depicted in Table 3. Table 3 shows that $\alpha_{01}(\cdot), \alpha_{02}(\cdot), \alpha_{32}(\cdot), \alpha_{42}(\cdot), \alpha_{52}(\cdot), \beta_1(\cdot), \ldots, \beta_6(\cdot)$ are nonconstant, the others are constant. From now on, we use $\alpha_{ij}$ to denote constant $\alpha_{ij}(\cdot)$, $\beta_i$ constant $\beta_i(\cdot)$.

The former indicates the presence of their nonlinear interactions with the year of marriage, while the latter shows no such interactions present.

The proposed estimation procedure is used to estimate the constant and functional coefficients. The estimated constant coefficients and their standard errors computed by the leave-one-cluster-out Jackknife are presented in Table 4, and the estimated functional coefficients together with their 95% confidence bands are presented in Fig. 3.

It is visible, from Table 4, some constant coefficients are not significantly apart from zero. This has some practical indications. For example, $\alpha_{41}$ (its estimated value is $-0.014$ with standard error $0.016$) not significantly apart from zero indicates the impact of education in the division of Kulna is not significantly different to that in Barisal.

As explained in the section of introduction, the proposed modelling and estimation methods mainly serve for the inference for a particular cluster. We are now going to use a cluster in the rural area in Chittagong to illustrate
Based on the model (1.4) and the estimators of the unknown coefficients involved, we have the impact of education on the second birth interval in a rural area in Chittagong

$$\hat{a}_1(U) = \hat{\alpha}_{01}(U) + \hat{\alpha}_{11} + \hat{\alpha}_{21}.$$ 

Similarly, we can get the impact of Hindu

$$\hat{a}_2(U) = \hat{\alpha}_{02}(U) + \hat{\alpha}_{12} + \hat{\alpha}_{22},$$

and the impact of first child in this area

$$\hat{a}_3(U) = \hat{\alpha}_{03} + \hat{\alpha}_{13} + \hat{\alpha}_{23}.$$

The functional coefficients $\hat{a}_1(\cdot)$ and $\hat{a}_2(\cdot)$ are presented in Fig. 3. The second one in the bottom panel in Fig. 3 is $\hat{a}_1(\cdot)$, which shows that in rural area in Chittagong, even the educated women still have shorter second birth interval than the uneducated women in urban area in Barisal before 1970. This indicates administrative division and type of region of residence play a very important role in fertility behavior in Bangladesh before 1970. It is also noticable that the second birth intervals of the educated women are getting longer.

From the third one in the bottom panel in Fig. 3, which is $\hat{a}_2(\cdot)$, we can see even in rural area in Chittagong, Hindus still have longer second birth interval than Muslims even in the urban area in Barisal after 1970. This suggests that religion plays an important role in fertility behavior in Bangladesh.

The impact $\hat{a}_3(\cdot)$ of boy in Rural in Chittagong is not varying with time. It is 0.128, which suggests even in rural area in Chittagong, the women with first child being boy still have longer second birth interval than the women with first child being girl or dead even in the urban area in Barisal. A possible interpretation is, like many developing countries, the Bangladesh culture always favors boy.

**Appendix.** In this section, we will prove the asymptotic distributions of the proposed estimators. For easy description, we write

$$\epsilon = (\epsilon_1^T, \ldots, \epsilon_m^T)^T, \quad x = \text{diag}(x_1, \ldots, x_m), \quad e = (e_1^T, \ldots, e_m^T)^T,$$

$$\hat{\theta}(u) = (\hat{\alpha}_0(u)^T, \hat{\alpha}_1(u)^T, \ldots, \hat{\alpha}_q(u)^T, \hat{\beta}(u)^T)^T,$$

$$\tau(u) = \{e_{(k+1),(q+1)}^T \otimes I_p\} \left[ \sigma^2 A_1(u)^{-1} + \Upsilon_1(u) \Xi_1(u) \Upsilon_1(u)^T \right] \{e_{(k+1),(q+1)} \otimes I_p\},$$
The marginal density \( E \) of an element of \( \Omega(u) \) is automatically positive semidefinite at the point \( u \), and \( \Omega(u) \) is continuous in the neighborhood of \( u \), and \( \Omega(u) \) is positive definite at the point \( u \).

The following technical conditions are imposed to establish the asymptotic results:

1. \( Ee_{11}^4 < \infty, E\|e_1\|^4 < \infty, Ez_j^2 < \infty, E\|z\|^2 < \infty \) for some \( d > 2 \), where \( \|e_1\|^2 = e_1^T e_1 \), \( z_j \) denotes the \( j \)th element of \( Z \) and \( x_i \) denotes the \( i \)th element of \( X \) for \( j = 1, \ldots, q, i = 1, \ldots, p \).
2. \( \hat{\alpha}_{kj}(\cdot) \) is continuous in a neighborhood of \( u \), for \( k = 0, \ldots, q, j = 1, \ldots, p \), where \( \hat{\alpha}_{kj}(\cdot) \) is the \( j \)th element of \( \hat{\alpha}_k(\cdot) \), and assume \( \hat{\alpha}_{kj}(u) \neq 0 \); similarly, \( \hat{\beta}_l(\cdot) \) is continuous in a neighborhood of \( u \), for \( l = 1, \ldots, q \), where \( \hat{\beta}_l(\cdot) \) is the \( l \)th element of \( \beta(\cdot) \), and \( \hat{\beta}_l(u) \neq 0 \).
3. The marginal density \( f(\cdot) \) of \( U \) has a continuous derivative in some neighborhood of \( u \), and \( f(u) \neq 0 \).
4. Each element of \( \Omega(u) \) and \( \Xi_1(u) \) are continuous in the neighborhood of \( u \), and \( \Omega(u) \) is positive definite at the point \( u \).
5. The function \( K(t) \) is a symmetric density function with a compact support.
6. \( n_i, i = 1, \ldots, m, \) are bounded. \( n \to \infty, h \to 0 \) and \( nh^2 \to \infty \).
7. \( E[X_{it}^T \Sigma X_{ir} \Gamma_i \Gamma_i^T U_{it} = u, U_{ir} = v] \) is continuous in the neighborhood of \( u \) and \( v \) respectively for \( r \neq l, r, l = 1, \ldots, n_i, i = 1, \ldots, m \).
8. \( \Omega(u) \) is bounded and has a continuous derivative on \( [a, b] \), \( \tau(u) \) has a bounded derivative on \( [a, b] \), \( 0 < \|\tau(u)\|_\infty < +\infty \), and the first derivative of \( K(t) \) has a finite number of sign changes over its support.

Note that \( \Omega(u) \) is automatically positive semidefinite at the point \( u \), so the second part of condition (4) is easily satisfied.

To obtain the proof of the theorems, the following lemmas are required.

**Lemma 1.** Let \( \{U_{ij}\} \) be i.i.d. random variables, \( \{\xi_{ij}\} \) be identically distributed, and \( \xi_{ij} \) be independent of \( \xi_{ik} \) for \( i \neq l \). Assume that the marginal density \( f(\cdot) \) of \( U \) has a continuous derivative in some neighborhood of \( u \), \( E(\xi_{11}\mid U_{11} = u) \) is continuous in the neighborhood of \( u \) and \( E(\xi_{11}^2) < \infty \).
Let $K(\cdot)$ be a bounded positive function with a bounded support. Then when $nh^2 \to \infty$, for $\lambda = 0, 1, 2, \ldots$

$$n^{-1} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \xi_{ij}(h^{-1}(U_{ij} - u))^{\lambda} K_h(U_{ij} - u) = \mu_\lambda f(u) E(\xi_{11}|U_{11} = u)(1 + o_P(1)).$$

**Proof:** Let $S_{n,\lambda} = n^{-1} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \xi_{ij}(h^{-1}(U_{ij} - u))^{\lambda} K_h(U_{ij} - u)$ and $g(v) = E(\xi_{11}|U_{11} = v)$, then

$$ES_{n,\lambda} = n^{-1} \sum_{i=1}^{m} \sum_{j=1}^{n_i} E \left\{ E \left[ \xi_{ij}(h^{-1}(U_{ij} - u))^{\lambda} K_h(U_{ij} - u) \right| U_{ij} \right\}$$

$$= \int g(v)(h^{-1}(v - u))^{\lambda} K_h(v - u)f(v) dv = \mu_\lambda f(u)g(u)(1 + o(1))$$

by continuity of both the density function $f(\cdot)$ and the conditional expectation function $g(\cdot)$ in the neighborhood of $u$. Moreover, as $K(\cdot)$ is a bounded function with a bounded support, $|u^{\lambda} K(u)|$ is bounded. Then by Jensen inequality it follows that

$$\text{var}(S_{n,\lambda}) \leq ES_{n,\lambda}^2 \leq n^{-2} \sum_{i=1}^{m} \sum_{j=1}^{n_i} E \left[ \xi_{ij}^2(h^{-1}(U_{ij} - u))^{2\lambda} K_h^2(U_{ij} - u) \right]$$

$$= O\left((nh^2)^{-1}\right).$$

Therefore

$$S_{n,\lambda} = ES_{n,\lambda} + O_P\left(\sqrt{\text{var}(S_{n,\lambda})}\right) = \mu_\lambda f(u)E(\xi_{11}|U_{11} = u)(1 + o_P(1)).$$

**Lemma 2.** Let $\{\xi_n, n \geq 1\}$ be an independent sequence with $E\xi_n = 0$, and

$$\lambda E|\xi_j|^3 \exp(\lambda|\xi_j|) \leq E\xi_j^2$$

for any $j \geq 1$ and some $\lambda > 0$, if $\sum_{i=1}^{n} E\xi_i^2 \to \infty$, then on a possibly enlarged probability space, there exists a sequence of independent random variables $\{\eta_n, n \geq 1\}$, $\eta_n \sim N(0, \text{var}(\xi_n))$ such that

$$\left|\sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \eta_i\right| \leq \frac{1}{\lambda C} \log \left(\sum_{i=1}^{n} \text{var}(\xi_i)\right),$$

where $C$ is a positive constant.
See Lin and Lu [16], P.129, Theorem 2.6.3.

**Proof of Theorem 1:** It can be shown that

\[
\sqrt{n} hf(u) \left( \hat{\alpha}_k(u) - \alpha_k(u) \right) = \sqrt{n} hf(u) \left( e^T_{(k+1),(q+1)} \otimes I_p, \ 0_{p \times (q+s)} \right) (X^T W X)^{-1} X^T W \epsilon \\
+ \sqrt{n} hf(u) \left( e^T_{(k+1),(q+1)} \otimes I_p, \ 0_{p \times (q+s)} \right) (X^T W X)^{-1} X^T W \epsilon e \\
+ \sqrt{n} hf(u) \left\{ (e^T_{(k+1),(q+1)} \otimes I_p, \ 0_{p \times (q+s)} ) (X^T W X)^{-1} X^T W E( \frac{y}{\mathcal{D}} ) \\
- \alpha_k(u) \right\} \equiv L_{n1} + L_{n2} + L_{n3}
\]

As

\[
L_{n1} = n(e^T_{(k+1),(q+1)} \otimes I_p, \ 0_{p \times (q+s)} ) (X^T W X)^{-1} H \sqrt{n} hf(u) \{ H^{-1} X^T W \epsilon \}
\]

with

\[
\sqrt{n} hf(u) E \{ H^{-1} X^T W \epsilon \} = 0_{2s \times 1},
\]

and

\[
n^{-1} hf(u) \text{cov} \{ H^{-1} X^T W \epsilon \} = \sigma^2 f(u) n^{-1} h E \{ H^{-1} X^T W^2 X H^{-1} \}
\]

\[
= \sigma^2 f^2(u) \left( \begin{array}{ccc}
\nu_0 & 0 \\
0 & \nu_2
\end{array} \right) \otimes \Omega(u) (1 + o(1)).
\]

Moreover, it follows from Lemma 1 that

(A.1) \[
\frac{1}{n} (X^T W X) = f(u) H \left\{ \left( \begin{array}{ccc}
1 & 0 \\
0 & \mu_2
\end{array} \right) \otimes \Omega(u) \right\} H (1 + o_P(1)).
\]

Hence,

\[
n(e^T_{(k+1),(q+1)} \otimes I_p, \ 0_{p \times (q+s)} ) (X^T W X)^{-1} H
\]

\[
= \frac{1}{f(u)} \left( (e^T_{(k+1),(q+1)} \otimes I_p, \ 0_{p \times (q+s)} ) \Omega(u)^{-1}, \ 0_{p \times (q+s)} \right) (1 + o_P(1)).
\]

By condition (1)(5)(6), Lindeberg Feller Theorem, Slutsky theorem and inverse of block matrix, it follows that

\[
L_{n1} \xrightarrow{D} N_p \left( 0_{p \times 1}, \ \nu_0 \sigma^2 \{ e^T_{(k+1),(q+1)} \otimes I_p \} A_1(u)^{-1} \{ e_{(k+1),(q+1)} \otimes I_p \} \right).
\]

Similarly, it can be shown that

\[
L_{n2} \xrightarrow{D} N_p \left( 0_{p \times 1}, \ \nu_0 \{ e^T_{(k+1),(q+1)} \otimes I_p \} \Theta_1(u) \{ e_{(k+1),(q+1)} \otimes I_p \} \right).
\]
Since \( L_{n1} \) and \( L_{n2} \) is independent, \((L_{n1} + L_{n2})\) has the following asymptotic distribution:

\[
N_p \left( 0_{p \times 1}, \nu_0 \{ e_{(k+1),(q+1)}^T \otimes I_p \} \right) \left[ \sigma^2 \Lambda_1(u)^{-1} + \Theta_1(u) \right] \{ e_{(k+1),(q+1)} \otimes I_p \}.
\]

By similar arguments as establishment of Lemma 1 and condition (2), we find that

\[
X^T W E(Y|D) - X^T W \begin{pmatrix} \theta(u) \\ \dot{\theta}(u) \end{pmatrix} = \left( \frac{1}{2} \sigma^2 \Lambda_1(u)^{-1} \hat{\alpha}_k(u) \right) \{ e_{(k+1),(q+1)} \otimes I_p \},
\]

This together with (A.1), we get that

\[
L_{n3} = \frac{\mu_2 \sqrt{nh^5 f(u)}}{2} \hat{\alpha}_k(u)(1 + o_P(1)).
\]

Therefore, when \( nh^5 \) is bounded, we obtain that

\[
\sqrt{nhf(u)} \left\{ \hat{\alpha}_k(u) - \alpha_k(u) - h^2 \frac{\mu_2}{2} \hat{\alpha}_k(u) \right\} \xrightarrow{D} N_p \left( 0_{p \times 1}, \nu_0 \{ e_{(k+1),(q+1)}^T \otimes I_p \} \right) \left[ \sigma^2 \Lambda_1(u)^{-1} + \Theta_1(u) \right] \{ e_{(k+1),(q+1)} \otimes I_p \}.
\]

**Proof of Theorem 2:** It follows immediately from the proof of Theorem 1.

**Proof of Theorem 3:** Let \( \Delta r_i = \hat{r}_i - r_i \), \( Q_i = I_{n_i} - P_i \), and \( \tilde{Q}_i \) be a diagonal matrix generated from the diagonal elements of \( Q_i \). Now

\[
\begin{align*}
\sigma^2 &= \frac{1}{n-mp} \sum_{i=1}^{m} \epsilon_i^T (Q_i - \tilde{Q}_i) \epsilon_i + \frac{1}{n-mp} \sum_{i=1}^{m} \epsilon_i^T \tilde{Q}_i \epsilon_i \\
&\quad + \frac{1}{n-mp} \sum_{i=1}^{m} E[\Delta r_i^T | D] Q_i E[\Delta r_i | D] \\
&\quad + \frac{1}{n-mp} \sum_{i=1}^{m} \left\{ \Delta r_i - E[\Delta r_i | D] \right\}^T Q_i \left\{ \Delta r_i - E[\Delta r_i | D] \right\} \\
&\quad + \frac{2}{n-mp} \sum_{i=1}^{m} \epsilon_i^T Q_i \{ \Delta r_i - E[\Delta r_i | D] \} \\
&\quad + \frac{2}{n-mp} \sum_{i=1}^{m} \epsilon_i^T Q_i E[\Delta r_i | D] \\
&\quad + \frac{2}{n-mp} \sum_{i=1}^{m} \epsilon_i^T Q_i \left\{ \Delta r_i - E[\Delta r_i | D] \right\} \\
&\equiv J_{n1} + J_{n2} + J_{n3} + J_{n4} + J_{n5} + J_{n6} + J_{n7}.
\end{align*}
\]
As $Q_i$ is an idempotent matrix and all the diagonal components of $Q_i - \tilde{Q}_i$ are equal to zero, by straightforward calculation it follows that

$$E(J_{n1} \mid D) = \frac{\sigma^2}{n - mp} \sum_{i=1}^{m} \text{tr}(Q_i - \tilde{Q}_i) = 0, \quad E(J_{n2} \mid D) = \frac{\sigma^2}{n - mp} \sum_{i=1}^{m} \text{tr}(\tilde{Q}_i) = \sigma^2,$$

$$\text{cov}(J_{n1}, J_{n2} \mid D) = E(J_{n1}J_{n2} \mid D) = \frac{2\sigma^4}{(n - mp)^2} \sum_{i=1}^{m} \text{tr}((Q_i - \tilde{Q}_i)\tilde{Q}_i) = 0,$$

and

$$\text{var}(J_{n1}) = E(E(J_{n1}^2 \mid D)) = \frac{2\sigma^4}{n - mp} \left\{ mp - \sum_{i=1}^{m} \sum_{j=1}^{n_i} E\left[ X_{ij}^T(x_i^T x_i)^{-1}X_{ij}\right]^2 \right\},$$

$$\text{var}(J_{n2}) = E(E(J_{n2}^2 \mid D)) - \sigma^4 = \frac{\text{var}(\epsilon_{11}^2)}{(n - mp)^2} \sum_{i=1}^{m} \sum_{j=1}^{n_i} E\left[ 1 - X_{ij}^T(x_i^T x_i)^{-1}X_{ij}\right]^2.$$

As $J_{n1} = (n - mp)^{-1} \sum \sum_{j=1}^{n_i} \sum_{r=1, r \neq i}^{m} X_{ij}^T(x_i^T x_i)^{-1}X_{ij} \epsilon_{il} \epsilon_{ir}$ is a sum of independent variables, by Lindeberg Feller Theorem, it follows that

$$n^{1/2}J_{n1} \xrightarrow{D} N\left(0, 2\sigma^4 c_1(c_1 - 1 - \gamma)\right).$$

Similarly,

$$n^{1/2}(J_{n2} - \sigma^2) \xrightarrow{D} N\left(0, \text{var}(\epsilon_{11}^2) c_1(2 - c_1 + \gamma)\right).$$

Since the two terms are uncorrelated, we have that

(A.4) \hspace{1cm} n^{1/2}(J_{n1} + J_{n2} - \sigma^2) \xrightarrow{D} N\left(0, 2\sigma^4 c_1(c_1 - 1 - \gamma) + \text{var}(\epsilon_{11}^2) c_1(2 - c_1 + \gamma)\right).$$

In the following, we will show that the remaining parts from $J_{n3}$ to $J_{n7}$ in (A.3) satisfy $n^{1/2}J_{nl} = o_P(1), \ l = 3, \ldots, 7$.

By the conditional bias of $\theta(u)$ and law of large numbers, it follows from $nh^8 \to 0$ that

(A.5) \hspace{1cm} n^{1/2}J_{n3} = o_P(1).
Since $0 \leq \phi^T Q_i \phi \leq \phi^T \phi$ for any $i$ and $n_i$ dimensional vector $\phi$, we have

$$
E\{|J_{n4}| D\} \leq (n - mp)^{-1} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \Gamma_{ij}^T \text{cov}(\hat{\theta}(U_{ij})|D) \Gamma_{ij} = O_P\left((nh)^{-1}\right).
$$

By the conditional bias and covariance matrix of $\hat{\theta}(u)$, it follows that

$$
E\left\{\frac{|J_{n5}| \mid D\right\} = \begin{cases} h^2 \mu_2 \sigma^2 & r = v \\ 0 & r \neq v \end{cases}
$$

And

$$
E(J_{n6}^2 \mid D) \leq \frac{h^4 \mu_2^2 \sigma^2}{(n - mp)^2} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left|\Gamma_{ij}^T \text{cov}(\hat{\theta}(U_{ij})|D) \Gamma_{ij}\right|^2 \left(1 + O_P(1)\right).
$$

By similar expression as (A.1) and straightforward calculation, we have

$$
\Gamma_{ij}^T \left[\hat{\theta}(U_{ij}) - E\left(\hat{\theta}(U_{ij})|D\right)\right] \epsilon_{ir} = \frac{1}{nf(U_{ij})} \Gamma_{ij}^T \Omega(U_{ij})^{-1} \left(\sum_{i=1}^{m} \sum_{l=1}^{n_i} \Gamma_{il} \epsilon_{il} K_h(U_{il} - U_{ij})\right) \epsilon_{ir} \left(1 + O_P(1)\right).
$$

Moreover, by boundness of kernel function and independence of the random errors, it can be shown that $E\left\{|J_{n7}| \mid D\right\} = O_P\left((nh)^{-1}\right)$.

Therefore using Markov inequality, when $h \to 0$, $nh^2 \to \infty$ we get

(A.6) 

$$
n^{1/2} J_{nl} = o_P(1), \ l = 4, \cdots, 7.
$$

Combining the results from (A.3) to (A.6), we have

$$
n^{1/2} \left\{\hat{\sigma}^2 - \sigma^2\right\} \overset{D}{\to} N\left(0, 2\sigma^4 c_1 (c_1 - 1 - \gamma) + \text{var}(\epsilon_{11}) c_1 (2 - c_1 + \gamma)\right).
$$
Proof of Theorem 4: Using standard arguments as in the proof of Theorem 3 and law of large numbers, when \( nh^2 \to \infty \), the conditional bias of \( \hat{\Sigma} \) is

\[
\text{bias}\left\{ \text{vec}(\hat{\Sigma})|D \right\} = O_P(h^4) + o_P\left(n^{-1/2}\right),
\]

and by straightforward but tedious calculation and Lindeberg-Feller Theorem, when \( nh^8 \to 0 \), it follows that

\[
n^{1/2}\text{vech}\left( \hat{\Sigma} - \Sigma \right) \xrightarrow{D} N_{p^2/2}(0, (1/c_2 + p)\Delta),
\]

where

\[
\Delta = E\left\{ (e_1^T \otimes (e_1^T)) - \text{vec}(\Sigma)\text{vec}(\Sigma)^T + 2\sigma^4\Delta_2 + \sigma^2\{\Sigma \otimes \Delta_1 + \Delta_1 \otimes \Sigma + \Delta_3\} + \text{var}(e_{11}^2) - 2\sigma^4\{\Delta_4 - c_2[2\text{vec}(\Delta_1)\text{vec}(\Delta_1)^T

- \text{vec}(\Delta_1)\text{vec}(\Delta_2)^T - \text{vec}(\Delta_2)\text{vec}(\Delta_1)^T]\} \}

+ \{2\sigma^4c_1(c_1 - 1 - \gamma) + \text{var}(e_{11}^2)c_1(2 - c_1 + \gamma)\}\text{vec}(\Delta_1)\text{vec}(\Delta_1)^T.
\]

Therefore, we have

\[
n^{1/2}\text{vech}\left( \hat{\Sigma} - \Sigma \right) \xrightarrow{D} N_{p(p+1)/2}\left(0_{p^2/2} \times 1, (1/c_2 + p)(R_p^T R_p)^{-1} R_p^T \Delta R_p (R_p^T R_p)^{-1}\right).
\]

Proof of Theorem 5: By simple calculation, it can be seen that

\[
\hat{C}_{kj} - C_{kj}
\]

\[
= \frac{1}{n} \sum_{l=1}^{m} \sum_{i=1}^{n_i} e_{kp+j,2s}^T(X_{(il)}^T W_{(il)} X_{(il)})^{-1} X_{(il)}^T W_{(il)} e
\]

\[
+ \frac{1}{n} \sum_{l=1}^{m} \sum_{i=1}^{n_i} e_{kp+j,2s}^T(X_{(il)}^T W_{(il)} X_{(il)})^{-1} X_{(il)}^T W_{(il)} x e
\]

\[
+ \frac{1}{n} \sum_{l=1}^{m} \sum_{i=1}^{n_i} e_{kp+j,2s}^T(X_{(il)}^T W_{(il)} X_{(il)})^{-1} X_{(il)}^T W_{(il)}
\]

\[
\times \left\{ E(Y|D) - X_{(il)}\left( \hat{\theta}(U_{il}) \theta(U_{il}) \right) \right\} = T_{n1} + T_{n2} + T_{n3}
\]

First, we obtain that \( E(T_{n1}) = 0 \), and for any \( j \) let

\[
F_{il}^T = e_{kp+j,2s}^T(X_{(il)}^T W_{(il)} X_{(il)})^{-1} X_{(il)}^T W_{(il)}, l = 1, \ldots, n_i, i = 1, \ldots, m,
\]
and \( \mathbf{F} = (F_{11}, \ldots, F_{1n_1}, \ldots, F_{m1}, \ldots, F_{mn_m})^T \), then we have

\[
\text{var}\{T_{n1}|D\} = \frac{\sigma^2}{n^2} 1_n \mathbf{F}^T 1_n \\
= \frac{\sigma^2}{n^2} \sum \sum \sum \sum e_{kp+j,2s}^T (X_{(i)}^TW_{(il)}^T) (X_{(i)})^{-1} \\
\times X_{(il)}^T W_{(ir)}^T X_{(ir)} (X_{(rv)}^T W_{(rv)} X_{(rv)})^{-1} e_{kp+j,2s} \\
= \frac{\sigma^2}{n} e_{kp+j,((q+1)p)}^T E^2 \left\{ \Lambda_1(U)^{-1} \right\} e_{kp+j,((q+1)p)} (1 + o_P(1)),
\]

by Lemma 1, straightforward but tedious calculation, and law of large numbers. Therefore using condition (1)(5)(6) and Lindeberg Feller Theorem, it follows that

\[
\sqrt{n} T_{n1} \xrightarrow{D} N(0, \sigma^2 e_{kp+j,((q+1)p)}^T E^2 \left\{ \Lambda_1(U)^{-1} \right\} e_{kp+j,((q+1)p)}).
\]

By the same way, it can be shown that

\[
\sqrt{n} T_{n2} \xrightarrow{D} N(0, e_{kp+j,((q+1)p)}^T \{ E[\Theta_1(U)] + \Xi_2 \} e_{kp+j,((q+1)p)}).
\]

Moreover, combining the results similar to (A.1) and (A.2), we get

\[
T_{n3} = h^2 \frac{\mu_2}{2} \left[ \frac{1}{n} \sum \sum \hat{\alpha}_{kj}^n (U_{iid}) \right] (1 + o_P(1)).
\]

Therefore, by independence of \( T_{n1} \) and \( T_{n2} \), when \( nh^4 \to 0 \) we have

\[
\sqrt{n} \left\{ \hat{C}_{kj} - C_{kj} \right\} \xrightarrow{D} N(0, e_{kp+j,((q+1)p)}^T \{ \sigma^2 E[\Lambda_1(U)^{-1}] + E[\Theta_1(U)] + \Xi_2 \} e_{kp+j,((q+1)p)}).
\]

**Proof of Theorem 6:** Obviously,

\[
\hat{\alpha}_{kj}(u) - \alpha_{kj}(u) - \text{bias}\{ \hat{\alpha}_{kj}(u) | D \} = e_{kp+j,2s}^T (X^T W X)^{-1} X^T W (\epsilon + xe) \equiv I_1(u).
\]

First of all, we approximate the random matrix \( I_1(u) \). As \( f(\cdot) \) and \( \Omega(\cdot) \) have continuous derivatives on \([a, b]\), by similar arguments as Lemma 1 and Neumann series expansion, it follows that

\[
(A.7) \quad nH(X^T W X)^{-1} H = f(u)^{-1} S(u)^{-1} + O_P \left( (nh^2)^{-2} + h \right),
\]
uniformly for \( u \in [a, b] \) where \( S(u) = \left( \begin{array}{cc} \mu_0 & 0 \\ 0 & \mu_2 \end{array} \right) \otimes \Omega(u) \).

By the asymptotic normality of \( \sqrt{n}^{-1} h f(u) H^{-1} X^T W(\epsilon + xe) \) in the proof of Theorem 1, we get that

\[
(A.8) \quad \left\| \frac{1}{n} H^{-1} X^T W(\epsilon + xe) \right\|_\infty = O_P(\frac{1}{\sqrt{nh}}).
\]

Therefore, using (A.7) and (A.8), it can be seen that

\[
(A.9) \quad \left\| I_1(u) - \frac{1}{n} e_{kp+j, 2s}^T f(u)^{-1} S(u)^{-1} H^{-1} X^T W(\epsilon + xe) \right\|_\infty = O_P\left( (nh^{\frac{3}{2}})^{-1} + (nh^{-1})^{-\frac{1}{2}} \right).
\]

Next, we consider

\[
I_2(u) = \frac{1}{n} e_{kp+j, 2s}^T f(u)^{-1} S(u)^{-1} H^{-1} X^T W(\epsilon + xe)
\]

\[
= \sum_{i=1}^m \sum_{l=1}^{n_i} \frac{1}{n f(u)} e_{kp+j, s}^T \Omega(u)^{-1} \Gamma_{il} K_h(U_{il} - u) (\epsilon_{il} + X_{il}^T e_i).
\]

Let \( w_{il} = e_{kp+j, s}^T \Omega(u)^{-1} \Gamma_{il} \), then

\[
{\{nh f(u) \tau(u)^{-1} \nu_0^{-1}\}^{1/2} I_2(u)}
\]

\[
(A.10) = \{nh f(u) \tau(u) \nu_0\}^{-1/2} \sum_{i=1}^m \sum_{l=1}^{n_i} K \left( \frac{U_{il} - u}{h} \right) w_{il} (\epsilon_{il} + X_{il}^T e_i)
\]

\[
\equiv \{nh f(u) \tau(u) \nu_0\}^{-1/2} \sum_{i=1}^m \sum_{l=1}^{n_i} K \left( \frac{U_{il} - u}{h} \right) w_{il} (\epsilon_{il} + X_{il}^T e_i)
\]

Divide interval \([a, b]\) into \( n \) subintervals \( J_r = [d_{r-1}, d_r), \quad r = 1, 2, \cdots, n - 1, J_n = [d_{n-1}, b] \) where \( d_r = a + \frac{b-a}{n} r \). Define \( \bar{U}_{il} = d_r I(U_{il} \in J_r), \quad r = 1, \cdots, n \), and it is obvious that \( |U_{il} - \bar{U}_{il}| = O(n^{-1}) \). Then by law of large numbers for random sequence \( \{w_{il}(\epsilon_{il} + X_{il}^T e_i)\} \), it follows that

\[
I_3(u) = \sum_{i=1}^m \sum_{l=1}^{n_i} K \left( \frac{U_{il} - u}{h} \right) w_{il} (\epsilon_{il} + X_{il}^T e_i)
\]

\[
(A.11) = \sum_{i=1}^m \sum_{l=1}^{n_i} K \left( \frac{\bar{U}_{il} - u}{h} \right) w_{il} (\epsilon_{il} + X_{il}^T e_i)
\]

uniformly for \( u \in [a, b] \). By the definition of \( \bar{U}_{il} \), we have that

\[
I_4(u) = \sum_{r=1}^n K \left( \frac{d_r - u}{h} \right) \sum_{i=1}^m \sum_{l=1}^{n_i} \bar{I}(U_{il} \in J_r) w_{il} (\epsilon_{il} + X_{il}^T e_i)
\]
Let \( \zeta_t = \sum_{r=1}^{t} \sum_{i=1}^{m_i} I(U_{it} \in J_r)w_{it}(\epsilon_{il} + X_{il}^T e_i) = \sum_{i=1}^{m} \sum_{l=1}^{n_i} I(a \leq U_{il} < d_{il})w_{il}(\epsilon_{il} + X_{il}^T e_i) \) (A.12), \( \zeta_0 \equiv 0 \), then by Lemma 2, for any \( t = 1, \cdots, n \) and \( u \in [a, b] \), we get

\[
|\zeta_t - n^{1/2}W(G(d_t))| = O(n^{1/4} \log n) \quad a.s.
\]

where \( W(\cdot) \) is a standard Wiener process and

\[
G(c) = \int_{a}^{c} \sigma^2 \left[ E(w_{11}^2|U_{11} = v) + E(X_{11}^T \Sigma X_{11}w_{11}^2|U_{11} = v) \right] f(v) dv 
\]

\[
+ \int_{a}^{c} \int_{a}^{c} \left\{ n^{-1} \sum_{i=1}^{m} \sum_{l=1}^{n_i} E[w_{il}X_{il}^T \Sigma X_{il}w_{il}|U_{il} = v_1, U_{ir} = v_2] \right\} \times f(v_1)f(v_2) dv_1 dv_2.
\]

It follows from Abel’s transform that

\[
I_4(u) = K \left( \frac{b - u}{h} \right) \zeta_n - \sum_{r=1}^{n-1} \left[ K \left( \frac{d_{r+1} - u}{h} \right) - K \left( \frac{d_r - u}{h} \right) \right] \zeta_r,
\]

and

\[
\left\| \sum_{r=1}^{n-1} \left[ K \left( \frac{d_{r+1} - u}{h} \right) - K \left( \frac{d_r - u}{h} \right) \right] \left[ \zeta_r - n^{1/2}W(G(d_r)) \right] \right\|_{\infty}
\]

\[
\leq \left\| \max_{1 \leq r \leq n} \left| \zeta_r - n^{1/2}W(G(d_r)) \right| \sum_{r=1}^{n-1} \left| K \left( \frac{d_{r+1} - u}{h} \right) - K \left( \frac{d_r - u}{h} \right) \right| \right\|_{\infty}
\]

\[
= O_P(n^{1/4} \log n)
\]

Hence

\[
I_4(u) = n^{1/2}K \left( \frac{b-u}{h} \right)W(G(b)) - n^{1/2} \sum_{r=1}^{n-1} \left[ K \left( \frac{d_{r+1} - u}{h} \right) - K \left( \frac{d_r - u}{h} \right) \right] W(G(d_r)) + O_P(n^{1/4} \log n)
\]

uniformly for \( u \in [a, b] \).

For a Wiener process, it is known that (Csörgő and Révész [5], P.44)

\[
\sup_{t \in [a, b]} |W(G(t + \delta)) - W(G(t))| = O(\{\delta \log(1/\delta)\}^{1/2}) \quad a.s.,
\]

when \( \delta \) is any small number. Using this property and the boundness of \( K(\cdot) \), we have

\[
\sum_{r=1}^{n-1} \left[ K \left( \frac{d_{r+1} - u}{h} \right) - K \left( \frac{d_r - u}{h} \right) \right] W(G(d_r)) = \int_{a}^{b} W(G(v))dK \left( \frac{v-u}{h} \right) + O_P(\{n^{-1} \log n\}^{1/2}).
\]
uniformly for \( u \in [a, b] \). This together with (A.11) and (A.12), it follows that

\[
\|(nh)^{-1/2} I_3(u) - h^{-1/2} \int_a^b K\left(\frac{v-u}{h}\right)dW(G(v))\|_\infty = O_P\left((nh^2)^{-\frac{1}{4}} \log n + (nh^3)^{-\frac{1}{2}}\right)
\]

Let

\[
Y_{1n}(u) = h^{-1/2} \int_a^b K\left(\frac{v-u}{h}\right)dW(G(v)),
\]

\[
Y_{2n}(u) = h^{-1/2} \int_a^b K\left(\frac{v-u}{h}\right)\left[\tau(v)f(v)\right]^{1/2}dW(v-a)
\]

and

\[
Y_{3n}(u) = h^{-1/2} \int_a^b K\left(\frac{v-u}{h}\right)dW(v-a).
\]

For a Gaussian process, following the similar proof of Lemmas in Hardle [10], we have

\[
\|Y_{1n}(u) - Y_{2n}(u)\|_\infty = O_P(h^{1/2}),
\]

\[
\|(f(u)\tau(u))^{-1/2} Y_{2n}(u) - Y_{3n}(u)\|_\infty = O_P(h^{1/2}).
\]

Therefore, by (A.13) and (A.14),

\[
\|\{nhf(u)\tau(u)\}^{-\frac{1}{2}} I_3(u) - Y_{3n}(u)\|_\infty = O_P\left((nh^2)^{-\frac{1}{4}} \log n + (nh^3)^{-\frac{1}{2}} + h^{\frac{1}{2}}\right)
\]

From Theorem 2 and Theorem 3.1 of Bickel and Rosenblatt [1], when \( h = n^{-\rho}, \ 1/5 \leq \rho < 1/3 \), we have

\[
P\left((-2 \log\{h/(b-a)\})^{\frac{1}{2}} \left\{\nu_0^{-\frac{1}{2}} \|\{nhf(u)\tau(u)\}^{-\frac{1}{2}} I_3(u)\|_\infty - \omega_n\right\} < x\right)
\]

\[
\rightarrow \exp\{-2\exp\{-x\}\}.
\]

where \( \omega_n \) is defined in Theorem 6. As

\[
\text{var}\{\hat{\alpha}_{kj}(u)|\mathcal{D}\} = \{nhf(u)\tau(u)^{-1}\nu_0^{-1}\}^{-1}(1 + o_P(1)).
\]

uniformly for \( u \in [a, b] \) by straightforward calculation and similar arguments as Lemma 1. Using the same proof of the first part of Theorem 2 of Fan and Zhang [9], the result of Theorem 6 is easily obtained.

**Proof of Theorem 7:** To prove the theorem, we first derive the rate of convergence for the bias and variance estimators. By (A.7) and its similar arguments, we have

\[
\|\text{bias}(\hat{\alpha}_{kj}(u)|\mathcal{D}) - \text{bias}(\hat{\alpha}_{kj}(u)|\mathcal{D})\|_\infty = O_P\left(h^2\{n^{-1/7} + o(h)\}\right),
\]
where the rate $n^{-1/7}$ comes from the pilot estimation of $\hat{\theta}(\cdot)$ and the term $o(h)$ comes from the coefficient in front of $\hat{\theta}^{(3)}(\cdot)$.

Furthermore, by similar proof as Lemma 1, we get

$$
\left\| \frac{h}{n} H^{-1} X^T W^2 X H^{-1} - f(u) \tilde{S}(u) \right\|_\infty = o_P(1),
$$

where $\tilde{S}(u) = \begin{pmatrix} \nu_0 & 0 \\ 0 & \nu_2 \end{pmatrix} \otimes \Omega(u)$, and

$$
\left\| \frac{h}{n} H^{-1} X^T W x x^T W X H^{-1} \right\|_\infty = O_P(1).
$$

These results together with (A.7), results of Theorem 3 and 4, we have

$$
\| \text{var}(\hat{\alpha}_{kj}(u)|D) - \text{var}(\hat{\alpha}_{kj}(u)|D) \|_\infty = O_P\left( (nh)^{-1}\left\{ n^{-\frac{1}{2}} + (nh^8)^{-\frac{1}{2}} \right\} \right).
$$

Using (A.15) and the same proof of the second part of Theorem 2 of Fan and Zhang [9], the result of Theorem 7 is obtained.

REFERENCES


### Table 1
The MISEs of the Estimators

<table>
<thead>
<tr>
<th>Estimator</th>
<th>$\hat{\beta}_0$</th>
<th>$\hat{\alpha}_{01}$</th>
<th>$\hat{\alpha}_{02}$</th>
<th>$\hat{\alpha}_{03}$</th>
<th>$\hat{\alpha}_{11}$</th>
<th>$\hat{\alpha}_{12}$</th>
<th>$\hat{\alpha}_{13}$</th>
<th>$\hat{\alpha}_{21}$</th>
<th>$\hat{\alpha}_{22}$</th>
<th>$\hat{\alpha}_{23}$</th>
<th>$\hat{\alpha}_{31}$</th>
<th>$\hat{\alpha}_{32}$</th>
<th>$\hat{\alpha}_{33}$</th>
<th>$\hat{\alpha}_{41}$</th>
<th>$\hat{\alpha}_{42}$</th>
<th>$\hat{\alpha}_{43}$</th>
<th>$\hat{\alpha}_{51}$</th>
<th>$\hat{\alpha}_{52}$</th>
<th>$\hat{\alpha}_{53}$</th>
<th>$\hat{\alpha}_{61}$</th>
<th>$\hat{\alpha}_{62}$</th>
<th>$\hat{\alpha}_{63}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MISE</td>
<td>0.013</td>
<td>0.019</td>
<td>0.019</td>
<td>0.020</td>
<td>0.019</td>
<td>0.020</td>
<td>0.019</td>
<td>0.017</td>
<td>0.016</td>
<td>0.016</td>
<td>0.014</td>
<td>0.014</td>
<td>0.016</td>
<td>0.014</td>
<td>0.014</td>
<td>0.014</td>
<td>0.014</td>
<td>0.014</td>
<td>0.014</td>
<td>0.014</td>
<td>0.014</td>
<td>0.014</td>
</tr>
</tbody>
</table>

### Table 2
The MSEs of the Estimators

<table>
<thead>
<tr>
<th>Estimator</th>
<th>$\hat{\sigma}_{11}$</th>
<th>$\hat{\sigma}_{12}$</th>
<th>$\hat{\sigma}_{13}$</th>
<th>$\hat{\sigma}_{22}$</th>
<th>$\hat{\sigma}_{23}$</th>
<th>$\hat{\sigma}_{33}$</th>
<th>$\hat{\sigma}_{43}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSE</td>
<td>0.0055</td>
<td>0.0013</td>
<td>0.0013</td>
<td>0.0052</td>
<td>0.0015</td>
<td>0.0054</td>
<td>0.0029</td>
</tr>
</tbody>
</table>

### Table 3
The P-Values of the Coefficients being Constant

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>$\hat{\beta}_0$</th>
<th>$\hat{\alpha}_{01}$</th>
<th>$\hat{\alpha}_{02}$</th>
<th>$\hat{\alpha}_{03}$</th>
<th>$\hat{\alpha}_{11}$</th>
<th>$\hat{\alpha}_{12}$</th>
<th>$\hat{\alpha}_{13}$</th>
<th>$\hat{\alpha}_{21}$</th>
<th>$\hat{\alpha}_{22}$</th>
<th>$\hat{\alpha}_{23}$</th>
<th>$\hat{\alpha}_{31}$</th>
<th>$\hat{\alpha}_{32}$</th>
<th>$\hat{\alpha}_{33}$</th>
<th>$\hat{\alpha}_{41}$</th>
<th>$\hat{\alpha}_{42}$</th>
<th>$\hat{\alpha}_{43}$</th>
<th>$\hat{\alpha}_{51}$</th>
<th>$\hat{\alpha}_{52}$</th>
<th>$\hat{\alpha}_{53}$</th>
<th>$\hat{\alpha}_{61}$</th>
<th>$\hat{\alpha}_{62}$</th>
<th>$\hat{\alpha}_{63}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>P-Value</td>
<td>0.102</td>
<td>0.005</td>
<td>0.008</td>
<td>0.248</td>
<td>0.094</td>
<td>0.217</td>
<td>0.082</td>
<td>0.074</td>
<td>0.213</td>
<td>0.060</td>
<td>0.618</td>
<td>0.038</td>
<td>0.283</td>
<td>0.235</td>
<td>0.031</td>
<td>0.135</td>
<td>0.263</td>
<td>0.020</td>
<td>0.052</td>
<td>0.052</td>
<td>0.148</td>
<td>0.269</td>
</tr>
<tr>
<td>-------------</td>
<td>-----------------</td>
<td>---------------------</td>
<td>---------------------</td>
<td>---------------------</td>
<td>---------------------</td>
<td>---------------------</td>
<td>---------------------</td>
<td>---------------------</td>
<td>---------------------</td>
<td>---------------------</td>
<td>---------------------</td>
<td>---------------------</td>
<td>---------------------</td>
<td>---------------------</td>
<td>---------------------</td>
<td>---------------------</td>
<td>---------------------</td>
<td>---------------------</td>
<td>---------------------</td>
<td>---------------------</td>
<td>---------------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>Coefficient</td>
<td>$\hat{\sigma}_{53}$</td>
<td>$\hat{\sigma}_{61}$</td>
<td>$\hat{\sigma}_{62}$</td>
<td>$\hat{\sigma}_{63}$</td>
<td>$\hat{\beta}_1$</td>
<td>$\hat{\beta}_2$</td>
<td>$\hat{\beta}_3$</td>
<td>$\hat{\beta}_4$</td>
<td>$\hat{\beta}_5$</td>
<td>$\hat{\beta}_6$</td>
<td>$\hat{\alpha}_{33}$</td>
<td>$\hat{\alpha}_{43}$</td>
<td>$\hat{\alpha}_{53}$</td>
<td>$\hat{\alpha}_{63}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P-Value</td>
<td>0.069</td>
<td>0.035</td>
<td>0.023</td>
<td>0.073</td>
<td>0.069</td>
<td>0.035</td>
<td>0.023</td>
<td>0.073</td>
<td>0.069</td>
<td>0.035</td>
<td>0.023</td>
<td>0.073</td>
<td>0.069</td>
<td>0.035</td>
<td>0.023</td>
<td>0.073</td>
<td>0.069</td>
<td>0.035</td>
<td>0.023</td>
<td>0.073</td>
<td>0.069</td>
<td>0.035</td>
</tr>
</tbody>
</table>

---

**Wenyang Zhang**  
Department of Mathematical Sciences  
University of Bath, UK  
E-mail: W.Zhang@bath.ac.uk

**Jianqing Fan**  
Department of Operation Research and Financial Engineering  
Princeton University, USA  
E-mail: jqfan@princeton.edu

**Yan Sun**  
School of Economics  
Shanghai University of Finance and Economics  
P.R.China  
E-mail: yansun2002cn@yahoo.com.cn
Table 4

Estimated Constant Coefficients

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>$\beta_0$</th>
<th>$\alpha_{03}$</th>
<th>$\alpha_{11}$</th>
<th>$\alpha_{12}$</th>
<th>$\alpha_{13}$</th>
<th>$\alpha_{21}$</th>
<th>$\alpha_{22}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate</td>
<td>3.530</td>
<td>0.016</td>
<td>-0.037</td>
<td>0.105</td>
<td>0.011</td>
<td>0.036</td>
<td>0.134</td>
</tr>
<tr>
<td>SE</td>
<td>0.004</td>
<td>0.005</td>
<td>0.010</td>
<td>0.004</td>
<td>0.003</td>
<td>0.013</td>
<td>0.006</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>$\alpha_{23}$</th>
<th>$\alpha_{31}$</th>
<th>$\alpha_{33}$</th>
<th>$\alpha_{41}$</th>
<th>$\alpha_{43}$</th>
<th>$\alpha_{51}$</th>
<th>$\alpha_{53}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate</td>
<td>0.101</td>
<td>-0.060</td>
<td>0.042</td>
<td>-0.014</td>
<td>0.111</td>
<td>-0.091</td>
<td>0.104</td>
</tr>
<tr>
<td>SE</td>
<td>0.005</td>
<td>0.006</td>
<td>0.005</td>
<td>0.016</td>
<td>0.007</td>
<td>0.009</td>
<td>0.005</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>$\alpha_{61}$</th>
<th>$\alpha_{62}$</th>
<th>$\alpha_{63}$</th>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
<th>$\beta_3$</th>
<th>$\beta_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate</td>
<td>-0.048</td>
<td>0.171</td>
<td>-0.005</td>
<td>-0.043</td>
<td>-0.103</td>
<td>-0.022</td>
<td>-0.093</td>
</tr>
<tr>
<td>SE</td>
<td>4.149</td>
<td>0.010</td>
<td>0.007</td>
<td>0.002</td>
<td>0.004</td>
<td>0.004</td>
<td>0.006</td>
</tr>
</tbody>
</table>

*SE stands for standard error of the estimator.*
Fig 1. The solid lines are the true curves, the dashed lines are the estimators, and the dotted lines are 95% confidence bands.
Fig 2. The left panel is the plot of the RMISEs of $a(\cdot)$ against bandwidths, the middle panel is the plot of the RMISEs of $\beta(\cdot)$, and the right panel is the density function of the estimated random effects in real data analysis.
Fig 3. The solid lines are the estimated functional coefficients, and dotted lines are the 95% confidence bands.