A New Multilevel Modelling Approach for Clustered Survival Data *

Jinfeng Xu[a], Mu Yue[b], Wenyang Zhang[c,*]

[a] *Department of Statistics & Actuarial Science, The University of Hong Kong*
[b] *School of Mathematical Sciences, University of Electronic Science and Technology of China*
[c] *Department of Mathematics, University of York, UK*

Abstract

In multilevel modelling for clustered survival data, to account for the difference among different clusters, a commonly used approach is to introduce cluster effects, either random or fixed, in the modelling. The modelling with random effects may lead to difficulties in the implementation of the estimation procedure for the unknown parameters of interest, because numerical computation for multiple integrals may become unavoidable when the cluster effects are not scalars. On the other hand, if fixed effects are used, there would be a danger of having estimators with big variances, because there are too many nuisance parameters involved in the model used. In this paper, using the idea of homogeneity pursuit, we propose a new approach of multilevel modelling for clustered survival data. The proposed modelling does not have the potential computational problem which the modelling with random effects does, it also involves far less unknown parameters than the modelling with fixed effects. We have also established asymptotic properties to show the advantages of the proposed modelling, and conducted intensive simulation studies to demonstrate the performance of the proposed method. Finally, the proposed method is applied to analyse a data set about the second-birth interval in Bangladesh. The most interesting finding is the impact of some important factors on the length of second-birth interval varies over clusters and also has homogeneity structure.

Keywords: binary segmentation; clustered survival data; Cox models; homogeneity pursuit; multilevel modelling; partial likelihood

*This research was supported by National Natural Science Foundation of China (Grant Number 71873085)

1. Introduction

This paper is stimulated by the analysis of a real data set from Bangladesh about the second-birth interval, the time interval between the first birth and the second birth, there. The data come from the Bangladesh Demographic and Health Survey of 1996-1997 (Mitra et al. (1997)), a cross-sectional nationally representative survey. The analysis is based on a sample of 7464 women nested within 125 primary sampling units or clusters, with sample sizes ranging from 17 to 242. Some women had not had their second child when the survey took place, therefore, their second-birth intervals are censored. The data set is a typical clustered survival data. What we are interested in is how the covariates, which are commonly found to be associated with the second-birth interval, affect the length of the second-birth interval. It is well known that a failure to take into account clustering in an analysis of clustered survival data typically leads to underestimation of standard errors since clustering reduces the effective sample size. In the case of survival data, the clustering, if ignored, can also lead to substantial bias. Hence multilevel modelling, see Harvey (2003), has to be employed when analysing clustered survival data.

To facilitate statistical modelling for the data set mentioned above, let y_{ij}, $i = 1, \cdots, n_j$, $j = 1, \cdots, J$, be the length of the second-birth interval of the ith respondent in the jth primary sampling unit in the Survey. X_{ij}, a p dimensional vector, is the vector of individual-level covariates of interest corresponding to y_{ij}. In addition, the vector of the covariates, defined at the cluster level, are denoted by a q dimensional vector $W_j = (w_{j1}, \cdots, w_{jq})^{\mathrm{T}}$. The censoring times, the lengths of the time intervals between the first birth and the time when the Survey took place, are denoted by c_{ij}s. The observed data are

$$\left(t_{ij},\ (X_{ij}^T,\ W_j^T),\ \delta_{ij}\right), \quad i = 1, \cdots, n_j, \quad j = 1, \cdots, J,$$

where

$$t_{ij} = \min(y_{ij},\ c_{ij}), \quad \delta_{ij} = I(y_{ij} > c_{ij}).$$

For this data set, the cluster-level variables w_{jk}, $k = 1, \cdots, q$, are all categorical. Suppose variable w_{jk} has $c_k + 1$ categories. To model the effects of w_{jk} we create c_k (0, 1)

*The corresponding author
Email addresses: xujf@hku.hk (Jinfeng Xu), yuemu.moon@gmail.com (Mu Yue), wenyang.zhang@york.ac.uk (Wenyang Zhang)

dummy variables, $(w_{jk,1},\ \cdots,\ w_{jk,c_k})$. Denote the coefficients of these dummy variables by $(\lambda_{k,1},\ \cdots,\ \lambda_{k,c_k})$.

1.1. The commonly used multilevel modelling strategy

In multilevel modelling for clustered survival data, to account for the difference on the impacts of the covariates concerned among different clusters, a commonly used approach is to introduce cluster effects, either random or fixed, in the modelling, see Harvey (2003) and Zhang and Steele (2004), and the references therein. If random effects are used for our case, when the Cox models, see David (1972), are employed, we have the following conditional proportional hazard function

$$\begin{aligned} h(t|X_{ij}, W_j, \mathbf{e}_j) &= h_0(t) \exp\left\{X_{ij}^T(\boldsymbol{\beta} + \mathbf{e}_j) + \sum_{k=1}^{q}\sum_{l=1}^{c_k} \lambda_{k,l} w_{jk,l}\right\} \\ &= h_0(t) \exp\left(\sum_{k=1}^{q}\sum_{l=1}^{c_k} \lambda_{k,l} w_{jk,l}\right) \exp\left(X_{ij}^T\boldsymbol{\beta} + X_{ij}^T\mathbf{e}_j\right), \end{aligned}$$

for the jth cluster, where $\mathbf{e}_j$s are random effects. Denoting the resulting conditional partial likelihood function, given X_{ij}s, W_js and $\mathbf{e}_j$s, by $L(\boldsymbol{\beta}|\mathbf{e}_1,\ \cdots,\ \mathbf{e}_J)$, this typical multilevel modelling would lead to the estimator of $\boldsymbol{\beta}$ to be the maximiser of

$$E\left\{L(\boldsymbol{\beta}|\mathbf{e}_1,\ \cdots,\ \mathbf{e}_J)\right\} \tag{1.1}$$

where the expectation is taken with respect to $\mathbf{e}_1,\ \cdots,\ \mathbf{e}_J$. When the dimension of $\mathbf{e}_j$ is not 1, which is often the case, numerical computation for multiple integrals may become unavoidable in the computation of the expectation in (1.1), hence, in the computation of the estimator of $\boldsymbol{\beta}$. It is well known that numerical computation for multiple integrals can cause serious trouble even for moderate dimensions. Therefore, this approach is not really practicable when the dimension of $\mathbf{e}_j$ is not 1.

On the other hand, if we use fixed cluster effects, and still employ the Cox models, the estimator of $\boldsymbol{\beta}$ would be the maximiser of $L(\boldsymbol{\beta}|\mathbf{e}_1,\ \cdots,\ \mathbf{e}_J)$, and the maximisation is with respect to $(\boldsymbol{\beta},\ \mathbf{e}_1,\ \cdots,\ \mathbf{e}_J)$ under the condition $\sum_{j=1}^{J} \mathbf{e}_j = 0$. Apparently, this approach involves too many nuisance parameters, 744 nuisance parameters for our case, therefore, the resulting estimators may have big variances.

It is clear, when the cluster effects are not scalars, the commonly used multilevel modelling strategy has some problems. In this paper, we propose a new multilevel modelling strategy which does not involve any numerical computation for multiple integrals, and the number of unknown parameters involved is also reasonable. Furthermore, every parameter has its meaning, none of them is a nuisance parameter.

1.2. The proposed multilevel modelling strategy

The proposed multilevel modelling strategy is based on the idea of homogeneity pursuit rather than cluster effects. There is a rich literature about homogeneity pursuit, see Ke et al. (2015), Ke et al. (2016), Su et al. (2016), Su and Ju (2018), Wang et al. (2018), Wang and Su (2019), Su and Jin (2019), Ando and Bai (2017), Bonhomme and Manresa (2015), and the references therein. To make the description of the proposed methodology more generic, from now on, y_{ij} does not have to be the length of the second-birth interval, it is a survival time in the generic sense. Similarly, c_{ij}, X_{ij}, W_j, n_j and J are censoring time, individual-level covariate, cluster-level covariate, cluster size and number of clusters, respectively.

We do not employ cluster effects in the proposed multilevel modelling strategy. For each j, $j = 1, \cdots, J$, we apply the Cox models to fit the data from the jth cluster, that is the conditional hazard function $h(t|X_{ij}, W_j)$ for the jth cluster is assumed to be

$$\begin{aligned} h(t|X_{ij}, W_j) &= h_0(t) \exp\left(X_{ij}^T \boldsymbol{\beta}_j + \sum_{k=1}^{q} \sum_{l=1}^{c_k} \lambda_{k,l} w_{jk,l} \right\} \\ &= h_0(t) \exp\left(\sum_{k=1}^{q} \sum_{l=1}^{c_k} \lambda_{k,l} w_{jk,l} \right) \exp\left(X_{ij}^T \boldsymbol{\beta}_j \right), \end{aligned} \tag{1.2}$$

where $h_0(\cdot)$ is the common baseline hazard function. We embed an unknown homogeneity structure in $\boldsymbol{\beta}_j$s in the modelling to account for the information provided by different clusters about the same unknowns and reduce the number of unknown parameters, that is we assume $\boldsymbol{\beta}_j = (\beta_{1,j}, \cdots, \beta_{p,j})^{\mathrm{T}}$ have the following homogeneity structure

$$\beta_{\ell,j} = \begin{cases} \beta_{(1)} & \text{when } (\ell, j) \in \mathcal{B}_1, \\ \vdots & \vdots \\ \beta_{(H)} & \text{when } (\ell, j) \in \mathcal{B}_H, \end{cases} \tag{1.3}$$

$\{\mathcal{B}_k : \ k = 1, \ \cdots, \ H\}$ is a partition of set $\{(\ell, j) : \ \ell = 1, \ \cdots, \ p; \ j = 1, \ \cdots, \ J\}$. The model (1.2) together with the homogeneity structure (1.3) is the proposed multilevel modelling strategy for clustered survival data, in which, $h_0(\cdot)$, H, $\beta_{(i)}$, $i = 1, \ \cdots, \ H$, the partition $\{\mathcal{B}_k : \ k = 1, \ \cdots, \ H\}$, and $\lambda_{k,l}$, $l = 1, \ \cdots, \ c_k$, $k = 1, \ \cdots, \ q$, are unknown and to be estimated. We also assume the partition $\{\mathcal{B}_k : \ k = 1, \ \cdots, \ H\}$ is independent of the covariates.

The advantages of the proposed multilevel modelling over the commonly used ones are: (1) There is no numerical computation for any multiple integral needed in the estimation of the unknown parameters, which makes the implementation of the estimation much easier; (2) There is no nuisance parameter involved, and the number of unknown parameters is reasonable, which avoids the danger of having final estimators with big variances; (3) cluster level attributes of the impacts of the covariates are better accounted for and well estimated.

The reason for us to impose a homogeneity structure on the components of $\boldsymbol{\beta}_j$s rather than $\boldsymbol{\beta}_j$s is to reduce the number of unknown parameters as much as possible. This is because two different vectors may have some components in common, which represents a kind of homogeneity, and such homogeneity can not be detected by the vector based homogeneity pursuit. Therefore, the vector based homogeneity pursuit would result in more unknown parameters than the component based homogeneity pursuit such as (1.3).

Although the proposed multilevel modelling strategy is for the Cox models, the idea of the modelling applies to other kinds of survival models.

The rest of the paper is organised as follows. We begin in Section 2 with a description of an estimation procedure for the unknown parameters in the proposed model. In Section 3 we present the asymptotic properties for the proposed estimators. The performance of the proposed estimation procedure is assessed by a simulation study in Section 4. In Section 5, we explore how the covariates, which are commonly found to be associated with the second-birth interval, affect the length of the second-birth interval, based on the proposed modelling strategy and estimation procedure. All technical conditions and theoretical proofs of all theoretical results are left to the Appendix.

2. Estimation procedure

In this section, we are going to present an estimation procedure for the unknown parameters in the proposed model (1.2) and its homogeneity structure (1.3).

We first introduce some notations: for the jth cluster, we denote the distinct event times by $t_{(1),j} < \cdots < t_{(T_j),j}$ and the number of events at time $t_{(\ell),j}$ by $d_{\ell,j}$. The set of indices for the individuals at risk up to time $t_{(\ell),j}$ is denoted by $R_{\ell,j}$, and the set of indices for the events at $t_{(\ell),j}$ by $\mathcal{D}_{\ell,j}$.

2.1. Estimation of the impacts, $\boldsymbol{\beta}_j$s, of individual-level variables

The procedure for estimating $\boldsymbol{\beta}_j$ consists of three stages. In the first stage, an initial estimator for $\boldsymbol{\beta}_j$ is obtained for each cluster by the method of partial likelihood (David (1972)), where Peto's (Breslow (1972)) approximation for ties is used. We then conduct homogeneity pursuit to identify which $\beta_{i,j}$s are the same, and which are different. Finally, we reparametrize the models concerned by replacing the $\beta_{i,j}$s, which are identified to have the same value, by a single parameter, and apply the partial likelihood method to estimate the unknown parameters in the models.

Explicitly,

Stage 1 (*Initial Estimation*). For each j, $j = 1, \cdots, J$, based on the observations from the jth cluster, the partial log-likelihood function for (1.2) is

$$\mathcal{L}_j(\boldsymbol{\beta}_j) = \sum_{\ell=1}^{T_j} \left\{ \sum_{i \in \mathcal{D}_{\ell,j}} \left(X_{ij}^T \boldsymbol{\beta}_j - \log \left\{ \sum_{k \in R_{\ell,j}} \exp\left(X_{kj}^T \boldsymbol{\beta}_j \right) \right\} \right) \right\}. \tag{2.4}$$

Let $\tilde{\boldsymbol{\beta}}_j = (\tilde{\beta}_{1,j} \ \cdots, \ \tilde{\beta}_{p,j})$ maximise (2.4). $\tilde{\boldsymbol{\beta}}_j$ is an initial estimator of $\boldsymbol{\beta}_j$.

Stage 2 (*Homogeneity Pursuit*). Let $\tilde{\beta}_{i,j}$ be the ith component of $\tilde{\boldsymbol{\beta}}_j$, we sort $\tilde{\beta}_{i,j}$, $i = 1, \cdots, p$, $j = 1, \cdots, J$, in ascending order, and denote them by

$$b_{(1)} \leq \cdots \leq b_{(Jp)}$$

We use r_{ij} to denote the rank of $\tilde{\beta}_{i,j}$. Identifying the homogeneity among $\tilde{\beta}_{i,j}$, $i = 1, \cdots, p$, $j = 1, \cdots, J$, is equivalent to detecting the change points among $b_{(l)}$,

$l = 1, \cdots, Jp$. To this end, we apply the Binary Segmentation algorithm [Bai (1997); Vostrikova (1981); Venkatraman (1993)] as follows.

For any $1 \leq i < j \leq Jp$, let

$$\Delta_{ij}(\kappa) = \sqrt{\frac{(j-\kappa)(\kappa-i+1)}{j-i+1}} \left(\frac{\sum_{l=\kappa+1}^{j} b_{(l)}}{j-\kappa} - \frac{\sum_{l=i}^{\kappa} b_{(l)}}{\kappa-i+1} \right)$$

Given a threshold δ, in practice, the Binary Segmentation algorithm to detect the change points works as follows.

(1) Find $\hat{k}_1$ such that

$$\Delta_{1,Jp}(\hat{k}_1) = \max_{1 \leq \kappa < Jp} \Delta_{1,Jp}(\kappa).$$

If $\Delta_{1,Jp}(\hat{k}_1) \leq \delta$, there is no change point among $b_{(l)}$, $l = 1, \cdots, Jp$, and the process of detection ends. Otherwise, add $\hat{k}_1$ to the set of change points and divide the region $\{\kappa : 1 \leq \kappa \leq Jp\}$ into two subregions: $\{\kappa : 1 \leq \kappa \leq \hat{k}_1\}$ and $\{\kappa : \hat{k}_1 + 1 \leq \kappa \leq Jp\}$.

(2) Detect the change points in the two subregions obtained in (1), respectively. Let us deal with the region $\{\kappa : 1 \leq \kappa \leq \hat{k}_1\}$ first. Find $\hat{k}_2$ such that

$$\Delta_{1,\hat{k}_1}(\hat{k}_2) = \max_{1 \leq \kappa < \hat{k}_1} \Delta_{1,\hat{k}_1}(\kappa).$$

If $\Delta_{1,\hat{k}_1}(\hat{k}_2) \leq \delta$, there is no change point in the region $\{\kappa : 1 \leq \kappa \leq \hat{k}_1\}$. Otherwise, add $\hat{k}_2$ to the set of change points and divide the region $\{\kappa : 1 \leq \kappa \leq \hat{k}_1\}$ into two subregions: $\{\kappa : 1 \leq \kappa \leq \hat{k}_2\}$ and $\{\kappa : \hat{k}_2 + 1 \leq \kappa \leq \hat{k}_1\}$. For the region $\{\kappa : \hat{k}_1 + 1 \leq \kappa \leq Jp\}$, we find $\hat{k}_3$ such that

$$\Delta_{\hat{k}_1+1,Jp}(\hat{k}_3) = \max_{\hat{k}_1+1 \leq \kappa < Jp} \Delta_{\hat{k}_1+1,Jp}(\kappa).$$

If $\Delta_{\hat{k}_1+1,Jp}(\hat{k}_3) \leq \delta$, there is no change point in the region $\{\kappa : \hat{k}_1 + 1 \leq \kappa \leq Jp\}$. Otherwise, add $\hat{k}_3$ to the set of change points and divide the region $\{\kappa : \hat{k}_1 + 1 \leq \kappa \leq Jp\}$ into two subregions: $\{\kappa : \hat{k}_1 + 1 \leq \kappa \leq \hat{k}_3\}$ and $\{\kappa : \hat{k}_3 + 1 \leq \kappa \leq Jp\}$.

(3) For each subregion obtained in (2), we do exactly the same as that for the subregion $\{\kappa : 1 \leq \kappa \leq \hat{k}_1\}$ or $\{\kappa : \hat{k}_1 + 1 \leq \kappa \leq Jp\}$ in (2), and keep doing so until there is no subregion containing any change point.

We sort the estimated change points in ascending order and denote them by

$$\hat{k}_{(1)} < \hat{k}_{(2)} < \cdots < \hat{k}_{(\hat{H}_{-1})},$$

where $\hat{H}_{-1}$ is the number of change points detected. In addition, we denote $\hat{k}_{(0)} = 0$, $\hat{H} = \hat{H}_{-1} + 1$, and $\hat{k}_{(\hat{H})} = Jp$. We use $\hat{H}$ to estimate H. Let

$$\hat{\mathcal{B}}_\ell = \{(i,j) : \ \hat{k}_{(\ell-1)} < r_{ij} \leq \hat{k}_{(\ell)}\}, \quad 1 \leq \ell \leq \hat{H},$$

we use $\left\{\hat{\mathcal{B}}_\ell : \ 1 \leq \ell \leq \hat{H}\right\}$ to estimate the partition $\{\mathcal{B}_\ell : \ 1 \leq \ell \leq H\}$. We consider all the $\beta_{i,j}$s with the subscript (i,j) in the same member of the estimated partition having the same value.

Stage 3 (*Final Estimation*). Let $\mathcal{L}(\xi_1, \ \cdots, \ \xi_{\hat{H}})$ be

$$\sum_{j=1}^{J} \mathcal{L}_j(\boldsymbol{\beta}_j)$$

with $\beta_{i,j}$, $i = 1, \ \cdots, \ p$, $j = 1, \ \cdots, \ J$, being replaced by ξ_k if $(i,j) \in \hat{\mathcal{B}}_k$. Let $(\hat{\xi}_1, \ \cdots, \ \hat{\xi}_{\hat{H}})$ maximise $\mathcal{L}(\xi_1, \ \cdots, \ \xi_{\hat{H}})$. The final estimator $\hat{\beta}_{i,j}$ of $\beta_{i,j}$ is $\hat{\xi}_k$ if $(i,j) \in \hat{\mathcal{B}}_k$.

Remark. The threshold δ used in Stage 2 can be selected by BIC, see Volinsky and Raftery (2000). This is because the selection of δ is equivalent to the selection of the number of elements in the partition, namely the H in (1.3), which is the number of unknown parameters, therefore, the BIC criterion becomes a natural choice, that is why we use the BIC criterion to select the δ.

2.2. Estimation of the common cumulative baseline hazard function and the impacts, $\lambda_{k,l}$s, of the cluster level variables

After obtaining the estimator for $\boldsymbol{\beta}_j$ in (1.2), we estimate $\lambda_{k,l}$, $l = 1, \ \cdots, \ c_k$; $k = 1, \ \cdots, \ q$, the impact of the categorical cluster-level variables and the common cumulative baseline hazard function.

Define the baseline hazard function for the jth cluster as

$$h_{1,j}(t) = h_0(t) \exp\left(\sum_{k=1}^{q} \sum_{l=1}^{c_k} \lambda_{k,l} w_{jk,l}\right). \tag{2.5}$$

and the cumulative baseline hazard function as

$$\Lambda_{1,j}(t) = \int_0^t h_{1,j}(u)du$$

The Breslow's estimator for $\Lambda_{1,j}(t_{(\ell),j})$ is

$$\hat{\Lambda}_{1,j}(t_{(\ell),j}) = \sum_{m=1}^{\ell} \left\{ \sum_{k \in R_{m,j}} \exp\left(X_{kj}^T \hat{\boldsymbol{\beta}}_j \right) \right\}^{-1}. \tag{2.6}$$

Let

$$L_{1,j}(t) = \log\left(\Lambda_{1,j}(t)\right), \quad \Lambda_0(t) = \int_0^t h_0(u)du, \quad L_0(t) = \log\Big(\Lambda_0(t)\Big),$$

we have

$$L_{1,j}(t) = L_0(t) + \sum_{k=1}^{q} \sum_{\ell=1}^{c_k} \lambda_{k,\ell} w_{jk,\ell}.$$

This leads to the following synthetic regression model

$$\hat{L}_{1,j}(t_{(\ell),j}) = L_0(t_{(\ell),j}) + \sum_{k=1}^{q} \sum_{l=1}^{c_k} \lambda_{k,l} w_{jk,l} + \epsilon_{\ell,j}, \quad \ell = 1, \cdots, T_j, \quad j = 1, \cdots, J, \tag{2.7}$$

where $\hat{L}_{1,j}(t_{(\ell),j}) = \log\Big(\hat{\Lambda}_{1,j}(t_{(\ell),j})\Big)$.

Next we consider the estimation of (2.7). Let $t_{(1)} < t_{(2)} < \cdots < t_{(N)}$ be the distinct values of $t_{(\ell),j}$, $\ell = 1, \cdots, T_j$, $j = 1, \cdots, J$. For each $t_{(m)}$, $m = 1, \cdots, N$, applying local linear modelling, we obtain the following local least squares procedure

$$\sum_{j=1}^{J} \sum_{\ell=1}^{T_j} \left\{ \hat{L}_{1,j}(t_{(\ell),j}) - a - b(t_{(\ell),j} - t_{(m)}) - \sum_{k=1}^{q} \sum_{l=1}^{c_k} \lambda_{k,l} w_{jk,l} \right\}^2 K_h(t_{(\ell),j} - t_{(m)}), \tag{2.8}$$

where $K_h(\cdot) = K(\cdot/h)/h$, $K(\cdot)$ is a kernel function, usually taken to be the Epanechnikov kernel, $K(u) = 0.75(1-u^2)_+$. h is a bandwidth.

Let

$$\Big(\tilde{a}(t_{(m)}), \tilde{b}(t_{(m)}), \tilde{\lambda}_{1,1}(t_{(m)}), \cdots, \tilde{\lambda}_{1,c_1}(t_{(m)}), \cdots, \tilde{\lambda}_{q,1}(t_{(m)}), \cdots, \tilde{\lambda}_{q,c_q}(t_{(m)})\Big)$$

be the minimizer of (2.8). The estimators for $\lambda_{k,l}$ are taken to be

$$\hat{\lambda}_{k,l} = \frac{1}{N} \sum_{m=1}^{N} \tilde{\lambda}_{k,l}(t_{(m)}), \quad l = 1, \cdots, c_k, \quad k = 1, \cdots, q. \tag{2.9}$$

The estimator for $L_0(t_{(m)})$, $m = 1, \cdots, N$, is taken to be $\tilde{a}(t_{(m)})$. This leads to the following initial estimator for the common cumulative baseline hazard function at $t_{(m)}$

$$\tilde{\Lambda}_0(t_{(m)}) = \exp\left\{\tilde{a}(t_{(m)})\right\}, \quad m = 1,\ \cdots,\ N.$$

Viewing $\left(t_{(m)},\ \tilde{\Lambda}_0(t_{(m)})\right)$, $m = 1,\ \cdots,\ N$, as a sample from the nonparametric regression model

$$\eta = \Lambda_0(t) + \varepsilon,$$

and using a local linear modelling, we obtain the estimator $\hat{\Lambda}_0(\cdot)$ of $\Lambda_0(\cdot)$.

3. Asymptotic properties

In this section, we are going to present the asymptotic properties of the proposed estimators. First we introduce some notations. We assume H is fixed, and let $\mathcal{N} = \sum_{j=1}^{J} n_j$. For the ith subject in the jth cluster, let $N_{ij}(t) = I(t_{ij} \leq t, \delta_{ij} = 1)$ and $Y_{ij}(t) = I(t_{ij} \geq t)$ be the counting process and at risk process, respectively. We use τ to denote the study ending time as in Bradic et al. (2011). Let the σ-filtration $\mathcal{F}_t = \sigma\{N_{ij}(s), Y_{ij}(s), s \leq t, i = 1, ..., n_j, j = 1, ..., J\}$. Denote by β_j^* the true value of β_j and $\Lambda_{ij}(t) = \int_0^t Y_{ij}(u) \exp(X_{ij}^T \beta^*) h_{1,j}(u) du$. With respect to the filtration $\{\mathcal{F}_t, t \geq 0\}$, $M_{ij}(t) = N_{ij}(t) - \int_0^t Y_{ij}(u) \exp(X_{ij}^T \beta^*) h_{1,j}(u) du$, $i = 1, ..., n_j$, $j = 1, ..., J$, $t \geq 0$ are (local) martingales with predictable variation/covariation processes

$$< M_{ij}, M_{ij} > (t) = \Lambda_{ij}(t) \quad \text{and} \quad < M_{ij}, M_{i'j'} > (t) = 0, \text{ when } i \neq i' \text{ or } j \neq j'.$$

Let $\otimes$ denote the outer product. Define

$$S_j^{(\ell)}(t, \beta) = n_j^{-1} \sum_{i=1}^{n_j} X_{ij}^{\otimes \ell} Y_{ij}(t) \exp(\beta^T X_{ij}), \ \ell = 0, 1, 2,$$

$$E_j(t, \beta) = \frac{S_j^{(1)}(t, \beta)}{S_j^{(0)}(t, \beta)},$$

and

$$V_j(t, \beta) = \frac{S_j^{(2)}(t, \beta)}{S_j^{(0)}(t, \beta)} - E_j(t, \beta)^{\otimes 2}.$$

By differentiation and rearrangement of terms, it can be shown as in Andersen and Gill (1982) that the gradient of $\mathcal{L}_j(\boldsymbol{\beta})$ is

$$\dot{\mathcal{L}}_j(\boldsymbol{\beta}) \equiv \frac{\partial \mathcal{L}_j(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n_j} \int_0^t [X_{ij} - E_j(u, \boldsymbol{\beta})] dN_{ij}(u),$$

and the Hessian matrix of $\mathcal{L}_j(\boldsymbol{\beta})$ is

$$\ddot{\mathcal{L}}_j(\boldsymbol{\beta}) \equiv \frac{\partial^2 \mathcal{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = -\sum_{i=1}^{n_j} \int_0^t V_j(u, \boldsymbol{\beta}) dN_{ij}(u).$$

Let $n = \min\limits_{1 \le j \le J} n_j$ and $\Delta = \min\limits_{2 \le k \le H} |\boldsymbol{\beta}_{(k)} - \boldsymbol{\beta}_{(k-1)}|$. Assume that $n \to \infty$. Suppose K is a sufficiently large positive constant. Next we list the following regularity conditions.

CONDITION 1. (i) For any $1 \le j \le J$, the unknown parameter β_j belongs to a compact subset of $\mathcal{R}^p$, the true parameter value β_j^* lies in its interior and $\|\lambda\| \le K$.

(ii) The covariates satisfy

$$\max_{1 \le j \le J} \max_{i < i' \le n_j} \max_{1 \le k \le p} |X_{ijk} - X_{i'jk}| \le K,$$

and $\max\limits_j \|W_j\| \le K$.

(iii). There exists a positive constant c_0, and with probability tending to 1,

$$\inf_{1 \le j \le J} \inf_{\|b\|=1, b \in R^p} \int_0^{t_{(T_j),j}} b^T V_j(u, \boldsymbol{\beta}_j^*) S_j^{(0)}(u, \boldsymbol{\beta}_j^*) b h_0(u) du \ge c_0.$$

(iv). We assume

$$\frac{\log J}{n} = o(\Delta^2).$$

Remark 1. Conditions 1(i)-(iii) are standard for asymptotic analyses and 1(iv) allows the number of clusters to diverge at a slower than polynomial rate of the minimum cluster size.

CONDITION 2. (i) Assume that $Jp \log(Jp) = o(\sqrt{n})$.

(ii) There exists a positive constant c_0 such that $\min\limits_{1 \le i,j \le H} \frac{s_i}{s_j} \ge c_0$.

(iii) $\Delta = \min\limits_{2 \le \ell \le H} (\beta_{(\ell)} - \beta_{(\ell-1)}) > \frac{1}{K}$.

(iv) $\delta \ge \frac{1}{K}$ and $\delta = o(\sqrt{Jp})$.

Remark 2. Condition 2(i) allows both the number of covariates p and the number of clusters J to diverge, at a rate more stringent than Condition 1(iv) but still reasonable in most applications. Condition 2(ii) assumes that the sizes of the clusters have about the same magnitude to ensure that the limiting distribution will not be dominated by the information from a subset of the clusters with dominating sizes, for ease of exposition of asymptotic results. Condition 2(iii) is similar to the separability condition in the k-means or hierarchical clustering, requiring that the true H distinct values of regression coefficients are separable by $1/K$. Condition 2(iv) specifies the range for the rate of δ.

CONDITION 3. Assume that $\mathcal{I}_n(u) = \mathcal{N}^{-1}\sum_{j=1}^{J}\sum_{i=1}^{n_j} Y_{ij}\Psi_j^T V_j(u, \Psi_j\boldsymbol{\xi}^*)\Psi_j \exp(X_{ij}^T\Psi_j^T\boldsymbol{\xi}^* + \lambda^T W_j) \to \mathcal{I}(u)$, in probability, for almost all u in $[0,\tau]$ and that $\mathcal{I} = \int_0^\tau \mathcal{I}(u)h_0(u)du$ is positive definite.

CONDITION 4. There exist functions $s_j^{(0)}(t,\beta_j)$ and $e_j(t,\beta_j)$, $1 \le j \le J$, such that

$$\max_{1\le j\le J}\sup_{0\le t\le\tau} |S_j^{(0)}(t,\Psi_j\boldsymbol{\xi}^*) - s_j^{(0)}(t,\Psi_j\boldsymbol{\xi}^*)| \to 0$$

in probability as $n \to \infty$, and

$$\max_{1\le j\le J}\sup_{0\le t\le\tau} |E_j(t,\Psi_j\boldsymbol{\xi}^*) - e_j(t,\Psi_j\boldsymbol{\xi}^*)| \to 0$$

in probability as $n \to \infty$.

Denote the condition survival function of C_{ij} given the cluster effect by $\bar{G}_{ij}(t) = P(C_{ij} > t|W_j)$ and the conditional density function of t_{ij} by $f_{ij}(t) = dP(t_{ij} \le t|W_j)/dt$.

CONDITION 5. (i) Let $K(\cdot)$ be a symmetric and bounded kernel density function with a bounded support.

(ii) $h \to 0$ and $\mathcal{N}h^2 \to \infty$.

(iii) f_{ij} and $\bar{G}_{ij}$ have continuous derivatives in $[0,\tau]$.

(iv) As $n \to \infty$,

$$\sup_{0\le t\le\tau} |\frac{1}{\mathcal{N}}\sum_{j=1}^{J}\sum_{i=1}^{n_j} f_{ij}(t)\bar{G}_{ij}(t)\tilde{W}_j\tilde{W}_j^T - \Omega(t)| \to 0$$

in probability, where for any t, $\Omega(t)$ is a $(c+1)\times(c+1)$ symmetric and positive definite matrix.

(v) Assume that for any $0 \le t_1, t_2 \le \tau$,

$$\mathcal{N}^{-1}\sum_{j=1}^{J} n_j\tilde{W}_j\tilde{W}_j^T s_j^{(0)}(t_1,\Psi_j\boldsymbol{\xi}^*)\int_0^{t_1}[s_j^{(0)}(u,\Psi_j\boldsymbol{\xi}^*)]^{-1}h_{1,j}(u)du\, s_j^{(0)}(t_2,\Psi_j\boldsymbol{\xi}^*)\int_0^{t_2}[s_j^{(0)}(v,\Psi_j\boldsymbol{\xi}^*)]^{-1}h_{1,j}(v)dv$$

$$\to \zeta(t_1,t_2),$$

and

$$\mathcal{N}^{-1}(\sum_{j=1}^{J} n_j \tilde{W}_j s_j^{(0)}(t_1, \Psi_j \boldsymbol{\xi}^*) \int_0^{t_1} e_j^T(u, \Psi_j \boldsymbol{\xi}^*) h_{1,j}(u) du \Psi_j^T) \to \Upsilon(t_1),$$

where $\zeta(t_1, t_2)$ and $\Upsilon(t_1)$ is $(c+1) \times (c+1)$ and $(c+1) \times H$ matrices, respectively.

Remark 3. Conditions 3, 4 and 5 (iii)-(v) are standard regularity conditions in the Cox model and conditions 5 (i)-(ii) are standard for kernel smoothing.

We assume H is fixed, and let $\mathcal{N} = \sum_{j=1}^{J} n_j$.

Theorem 1. *Under the conditions 1-3, for any given j, we have*

$$\mathcal{N}^{1/2}\left(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j\right) \xrightarrow{D} N(0_p, \Psi_j \mathcal{I}^{-1} \Psi_j^T).$$

Theorem 1 shows the proposed estimator $\hat{\boldsymbol{\beta}}_j$ is asymptotic normal, and enjoys a convergence rate of order $\mathcal{N}^{-1/2}$ which is a higher order of $n_j^{-1/2}$. In fact, it is the highest order an estimator can achieve even for the case where there is no clustering. This implies the homogeneity pursuit in the estimation procedure significantly improves the accuracy of the estimators of $\boldsymbol{\beta}_j$s.

Theorem 2. *Under the conditions 1-5, when $\mathcal{N}h^4 \to 0$, we have*

$$\mathcal{N}^{1/2}\left(\hat{\lambda}_{k,l} - \lambda_{k,l}\right) \xrightarrow{D} N(0, e_{k,\ell}^T \bar{\Omega}_{11}^{-2} \bar{\nu}_{22} e_{k,\ell}).$$

Theorem 2 shows the proposed estimator $\hat{\lambda}_{k,l}$ is also asymptotic normal, and enjoys the convergence rate of order $\mathcal{N}^{-1/2}$.

Theorem 3. *Under the conditions 1-5, when $\mathcal{N}h^4 \to 0$, for any given t, we have*

$$\mathcal{N}^{1/2}\left(\hat{\Lambda}_0(t) - \Lambda_0(t)\right) \xrightarrow{D} N(0,\ \Lambda_0^2(t)\nu_{11}(t,t)).$$

Theorem 3 shows the proposed estimator $\hat{\Lambda}_0(t)$ for the common cumulative baseline hazard function is asymptotic normal, and enjoys convergence rate of order $\mathcal{N}^{-1/2}$, which is the highest order an estimator of a monotonic function can achieve.

4. Simulation studies

In this section, we are going to use a simulated example to assess the performance of the proposed estimation procedure. As the homogeneity pursuit in the estimation procedure is of importance in its own right, we are also going to examine the accuracy of the proposed homogeneity pursuit in identifying the true homogeneity structure.

We set $h_0(t) = 1$, $p = 2$, $q = 1$, $c_q = 2$, $\boldsymbol{\beta}_j = (1,\ 2)^{\mathrm{T}}$ when j is odd, $\boldsymbol{\beta}_j = (-1,\ -2)^{\mathrm{T}}$ when j is even, $\lambda_{1,1} = 1$, and $\lambda_{1,2} = -2$ in model (1.2), $H = 4$ in the homogeneity structure (1.3), then generate data from model (1.2). We first generate X_{ij}s and W_js, then generate y_{ij}, given X_{ij} and W_j, for each $(X_{ij}^{\mathrm{T}},\ W_j)$, hence, the generated $(y_{ij},\ X_{ij}^{\mathrm{T}},\ W_j)$s. Once $(y_{ij},\ X_{ij}^{\mathrm{T}},\ W_j)$s are generated, we generate the censoring times t_{ij}s, and get the generated observations $(t_{ij},\ X_{ij}^{\mathrm{T}},\ W_j,\ \delta_{ij})$s. The details about how the data are generated are as follows.

We generate the observations X_{ij}, $i = 1, \cdots, n_j$, $j = 1, \cdots, J$, of the individual-level variable from the bivariate normal distribution with mean zero, covariance matrix $\begin{pmatrix} 1.0 & 0.2 \\ 0.2 & 1.0 \end{pmatrix}$ independently and identically. The observations w_{j1}, $j = 1, \cdots, J$, of the cluster-level variable are independently and identically generated from a multinomial distribution with 3 categories and the probability for each category being 1/3, namely $Multi(1,\ 1/3,\ 1/3,\ 1/3)$. Based on the generated w_{j1}s, we can get the generated observations $w_{j1,1}$ and $w_{j1,2}$, $j = 1, \cdots, J$, of the two dummy variables created from w_{j1}s by setting the first category as reference.

Once X_{ij}s, W_js are generated, for each $(X_{ij}^{\mathrm{T}},\ W_j)$, y_{ij}, given $(X_{ij}^{\mathrm{T}},\ W_j)$, is generated as follows: from

$$h(t|X_{ij}, W_j) = h_0(t) \exp\left(X_{ij}^{\mathrm{T}}\boldsymbol{\beta}_j + w_{j1,1}\lambda_{1,1} + w_{j1,2}\lambda_{1,2}\right)$$

we have

$$\int_0^t h(u|X_{ij}, W_j)du = \Lambda_0(t) \exp\left(X_{ij}^{\mathrm{T}}\boldsymbol{\beta}_j + w_{j1,1}\lambda_{1,1} + w_{j1,2}\lambda_{1,2}\right).$$

Since

$$\int_0^t h(u|X_{ij}, W_j)du = -\log(1 - F_y(t|X_{ij}, W_j)),$$

where $F_y(t|X_{ij}, W_j)$ be the conditional distribution of y_{ij} given (X_{ij}, W_j), we have

$$\Lambda_0(y_{ij}) \exp\left(X_{ij}^{\mathrm{T}}\boldsymbol{\beta}_j + w_{j1,1}\lambda_{1,1} + w_{j1,2}\lambda_{1,2}\right) = -\log(1 - F_y(y_{ij}|X_{ij}, W_j)).$$

It is easy to see $-\log(1 - F_y(y_{ij}|X_{ij}, W_j))$, given X_{ij} and W_j, follows a standard exponential distribution, and $\Lambda_0(y_{ij}) = y_{ij}$ as $h_0(t) = 1$. So, we generate a ϵ_{ij} from a standard exponential distribution, and y_{ij}, given $(X_{ij}^{\mathrm{T}},\ W_j)$, can be generated through

$$y_{ij} = \exp\left(-X_{ij}^{\mathrm{T}}\boldsymbol{\beta}_j - w_{j1,1}\lambda_{1,1} - w_{j1,2}\lambda_{1,2}\right)\epsilon_{ij},$$

we therefore have the generated $(y_{ij},\ X_{ij}^{\mathrm{T}},\ W_j)$.

The censoring times, c_{ij}s, are independently and identically generated from the uniform distribution $U(0,\ 10)$, the observed survival times t_{ij}s and censoring indicators δ_{ij}s are generated through

$$t_{ij} = \min(y_{ij},\ c_{ij}), \quad \delta_{ij} = I(y_{ij} > c_{ij}).$$

We therefore have the generated observations $(t_{ij},\ X_{ij}^{\mathrm{T}},\ W_j,\ \delta_{ij})$s.

We use 95% confidence intervals for each component of $\boldsymbol{\beta}_j$ and $\lambda_{1,l}$ and

$$\mathrm{MSE}_\beta = \frac{1}{J}\sum_{j=1}^{J} E\left(\|\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j\|^2\right), \quad \mathrm{MSE}_\lambda = \frac{1}{2}\sum_{l=1}^{2} E(\hat{\lambda}_{1,l} - \lambda_{1,l})^2$$

to evaluate the estimators $\hat{\boldsymbol{\beta}}_j$s and $\hat{\lambda}_{1,l}$s. Furthermore, we use

$$\mathrm{MISE}(\Lambda_0) = E\left\{\int \left(\hat{\Lambda}_0(t) - \Lambda_0(t)\right)^2 dt\right\}$$

to evaluate $\hat{\Lambda}_0(\cdot)$, and the normalized mutual information (NMI), see Ke *et al.*(2015), to measure how close the estimated homogeneity structure is to the true homogeneity structure. The NMI is defined as follows:

Let $A = \{A_1,\ A_2,\ \cdots\}$ and $B = \{B_1,\ B_2,\ \cdots\}$ be two partitions of a set of cardinality k. For any set S, we use $|S|$ to denote the cardinality of S. The NMI between A and B is defined as

$$\mathrm{NMI}(A,\ B) = \frac{2I(A,\ B)}{H(A) + H(B)},$$

where

$$I(A,\ B) = \sum_{i,j} \frac{|A_i \cap B_j|}{k}\log\Big(\frac{k|A_i \cap B_j|}{|A_i||B_j|}\Big), \quad H(A) = \sum_i \frac{|A_i|}{k}\log\Big(\frac{k}{|A_i|}\Big).$$

NMI ranges between 0 and 1 with a large value indicating a higher degree of similarity between the two partitions, A and B.

In all numerical studies in this paper, we use BIC to select the threshold δ needed in the homogeneity pursuit step in the proposed estimation procedure. The Epanechnikov kernel is used and a rule-of-thumb bandwidth is adopted in estimating $\lambda_{1,1}$, $\lambda_{1,2}$ and $\Lambda_0(\cdot)$.

Our simulation studies are conducted under either balanced design or unbalanced design. For the cases of balanced design, we set the number of clusters to be either 40, 80, or 120, and the cluster size to be either 40, 80, or 120. For the cases of unbalanced design, we still set the number of clusters to be either 40, 80, or 120, but the cluster sizes, n_js, to be the absolute values of the integer parts of the random variables generated from the uniform distribution $U(\bar{n} - 10,\ \bar{n} + 10)$, where $\bar{n}$ is set to be either 40, 80, or 120. For each case, we do 100 simulations.

The average censoring rate, across the 100 simulations, for each case is presented in Table 1, which shows every case concerned shares similar censoring rates.

Table 1: **The Average Censoring Rate for Each Case Investigated**

	Balanced Design			Unbalanced Design		
	$n_j = 40$	$n_j = 80$	$n_j = 120$	$\bar{n} = 40$	$\bar{n} = 80$	$\bar{n} = 120$
$J = 40$	0.300	0.306	0.302	0.304	0.307	0.305
$J = 80$	0.301	0.300	0.303	0.307	0.303	0.304
$J = 120$	0.303	0.302	0.302	0.304	0.304	0.302

We first examine the accuracy of the proposed homogeneity pursuit in identifying the true homogeneity structure. We compare our proposed method with the K-Means estimation method using the number of clusters $k = 4$. The K-mean method is executed by *kmeans* function in R using the algorithm of Hartigan and Wong (1979). For each case, we compute the NMI between the identified homogeneity structure, by the proposed pursuit, and the true homogeneity structure for each of the 100 simulations for that case, and present the median of the obtained 100 NMIs in Table 2. Table 2 shows there is no big difference between the balanced cases and the unbalanced cases on the accuracy of the proposed homogeneity pursuit, and the proposed homogeneity pursuit works well in any case and performs better than the K-Means method. It is also interesting to see the number of clusters does not have much impact on the accuracy of the proposed homogeneity pursuit.

We now turn to examining the accuracy of the proposed estimation procedure. We

Table 2: **The Median of NMIs for Each Case Investigated**

	The Proposed Method					
	Balanced Design			Unbalanced Design		
	$n_j = 40$	$n_j = 80$	$n_j = 120$	$\bar{n} = 40$	$\bar{n} = 80$	$\bar{n} = 120$
$J = 40$	0.835	0.964	1.000	0.846	0.964	1.000
$J = 80$	0.825	0.964	1.000	0.817	0.961	1.000
$J = 120$	0.817	0.954	0.985	0.809	0.956	1.000
	The K-Means Method					
	Balanced Design			Unbalanced Design		
	$n_j = 40$	$n_j = 80$	$n_j = 120$	$\bar{n} = 40$	$\bar{n} = 80$	$\bar{n} = 120$
$J = 40$	0.744	0.802	0.813	0.758	0.804	0.810
$J = 80$	0.738	0.800	0.813	0.744	0.803	0.806
$J = 120$	0.737	0.797	0.808	0.726	0.800	0.801

are not only interested in the accuracy of the proposed estimation procedure, but also the improvement of the proposed method over the K-Means method, the overfitting method, or the underfitting method. The overfitting method is the method without homogeneity pursuit when estimating $\boldsymbol{\beta}_j$s, which is basically the initial estimation of the proposed method for estimating $\boldsymbol{\beta}_j$s. The underfitting method is the method which assumes all clusters share the same $\boldsymbol{\beta}_j$ when estimating $\boldsymbol{\beta}_j$s. The K-Means method, overfitting method and underfitting method estimate $\lambda_{1,1}$, $\lambda_{1,2}$ and $\Lambda_0(\cdot)$ in the same way as the proposed method once the estimators of $\boldsymbol{\beta}_j$s are obtained.

For each case, in each of the 100 simulations for that case, we apply either the proposed estimation procedure, the K-Means method, the overfitting method, or the underfitting method to estimate the $\boldsymbol{\beta}_j$s, $\lambda_{1,1}$, $\lambda_{1,2}$ and $\Lambda_0(\cdot)$. We report the 95 % confidence intervals for each component of $\boldsymbol{\beta}_j$s and $\lambda_{1,l}$s, and compute the MSE_β, MSE_λ and $\mathrm{MISE}(\Lambda_0)$ of the estimators obtained by each of the four methods. The results are presented in Tables 3-7 for balanced design cases, in Tables 8-12 for unbalanced design cases. From these tables, we can see the balanced design cases tell the same story as the unbalanced design cases, and the proposed estimation procedure works well for any cases.

Compared with the K-Means method, the overfitting method or underfitting method, the proposed method has significant improvement in the accuracy of the estimators of $\boldsymbol{\beta}_j$s, the impact of individual-level variables. However, when it comes to estimate the impact of the cluster-level variables, $\lambda_{1,1}$, $\lambda_{1,2}$, or the common cumulative baseline hazard function $\Lambda_0(\cdot)$, the K-Means method and the overfitting method perform as well as the proposed method. This is because the only problem with the K-Means method or the overfitting method is that the variances of the estimators of $\boldsymbol{\beta}_j$s would be big, and the estimation for $\lambda_{1,1}$, $\lambda_{1,2}$, and $\Lambda_0(\cdot)$, after the estimators of $\boldsymbol{\beta}_j$s are obtained, is essentially a smoothing operation. The effect of the variances of the estimators of $\boldsymbol{\beta}_j$s can be reduced in the smoothing, so that the variances of the estimators of $\boldsymbol{\beta}_j$s do not have much effect on the estimation for $\lambda_{1,1}$, $\lambda_{1,2}$, and $\Lambda_0(\cdot)$. That is why the K-Means method and the overfitting method perform as well as the proposed method when estimating $\lambda_{1,1}$, $\lambda_{1,2}$, and $\Lambda_0(\cdot)$. The underfitting method performs badly irrespective of the parameter being estimated. This is because underfitting comes with big bias, and bias cannot be reduced by smoothing.

5. Analysis of the second-birth interval in Bangladesh

In this section, we apply the proposed multilevel modelling together with the proposed estimation procedure to analyse the data set which motivates this paper. The data are extracted from the Bangladesh Demographic and Health Survey conducted by the government of the People's Republic of Bangladesh. The sample is nationally representative and is based on 7464 women nested within 125 primary sampling clusters, with sample sizes ranging from 17 to 242.

The second-birth interval is an important indicator for family planning. The effects of the covariates, which are commonly found to be associated with the second-birth interval, on the length of the second-birth interval is of great importance. In this section, we are going to explore these effects based on the data extracted.

Let y_{ij} be the duration in months between the first birth and the second birth for the ith woman in the jth cluster. As 19.35% of the women in the sample had not given second birth by the time of the survey, 19.35% of y_{ij}s are censored. The important covariates have been identified as potential explanatory variables based on previous research, see Zhang and Steele (2004). They are year of marriage (continuous), women's level of education (categorized as none, primary, secondary or higher, and "none" is taken to be the reference in the modelling), religion (Muslim, Others, and "Muslim" is taken to be the reference in

Table 3: **The 95 % Confidence Intervals for Each Component of $\boldsymbol{\beta}_j$s for Each Case Under Balanced Design**

		The Proposed Method			The K-Means Method		
J	$\beta_{i,j}$	$n_j = 40$	$n_j = 80$	$n_j = 120$	$n_j = 40$	$n_j = 80$	$n_j = 120$
40	1	(0.932, 1.165)	(0.958, 1.072)	(0.966, 1.047)	(1.002, 1.218)	(1.025, 1.157)	(1.003, 1.113)
	2	(1.825, 2.068)	(1.936, 2.051)	(1.967, 2.016)	(1.535, 1.916)	(1.559, 1.907)	(1.686, 1.974)
	-1	(-1.170, -0.960)	(-1.084, -0.969)	(-1.058, -0.986)	(-1.198, -0.981)	(-1.155, -1.023)	(-1.141, -1.033)
	-2	(-2.087, -1.764)	(-2.043, -1.851)	(-2.037, -1.856)	(-2.032, -1.637)	(-1.909, -1.560)	(-1.896, -1.567)
80	1	(0.995, 1.139)	(0.974, 1.049)	(0.983, 1.022)	(1.038, 1.188)	(1.036, 1.130)	(1.029, 1.102)
	2	(1.865, 2.064)	(1.958, 2.031)	(1.977, 2.015)	(1.666, 1.943)	(1.650, 1.887)	(1.685, 1.903)
	-1	(-1.139, -0.998)	(-1.043, -0.977)	(-1.021, -0.984)	(-1.193, -1.040)	(-1.130, -1.039)	(-1.125, -1.048)
	-2	(-2.062, -1.851)	(-2.029, -1.942)	(-2.016, -1.968)	(-1.913, -1.641)	(-1.866, -1.625)	(-1.831, -1.595)
120	1	(1.000, 1.122)	(0.985, 1.046)	(0.987, 1.011)	(1.038, 1.158)	(1.057, 1.132)	(1.037, 1.096)
	2	(1.881, 2.048)	(1.961, 2.024)	(1.984, 2.015)	(1.716, 1.932)	(1.631, 1.831)	(1.687, 1.869)
	-1	(-1.110, -0.995)	(-1.035, -0.981)	(-1.017, -0.988)	(-1.157, -1.037)	(-1.116, -1.042)	(-1.105, -1.043)
	-2	(-2.036, -1.863)	(-2.024, -1.957)	(-2.012, -1.979)	(-1.899, -1.678)	(-1.86, -1.664)	(-1.853, -1.668)
		Overfitting Method			Underfitting Method		
J	$\beta_{i,j}$	$n_j = 40$	$n_j = 80$	$n_j = 120$	$n_j = 40$	$n_j = 80$	$n_j = 120$
40	1	(0.915, 1.184)	(0.947, 1.116)	(0.950, 1.089)	(-1.250, -0.945)	(-1.119, -0.967)	(-1.086, -0.970)
	2	(1.914, 2.295)	(1.935, 2.188)	(1.937, 2.124)	(-2.336, -1.968)	(-2.210, -1.953)	(-2.123, -1.938)
	-1	(-1.192, -0.916)	(-1.116, -0.943)	(-1.091, -0.957)	(-1.250, -0.945)	(-1.119, -0.967)	(-1.086, -0.970)
	-2	(-2.320, -1.927)	(-2.176, -1.932)	(-2.129, -1.939)	(-2.336, -1.968)	(-2.210, -1.953)	(-2.123, -1.938)
80	1	(0.973, 1.167)	(0.969, 1.089)	(0.971, 1.063)	(-1.164, -0.981)	(-1.077, -0.962)	(-1.058, -0.959)
	2	(1.993, 2.281)	(1.976, 2.149)	(1.968, 2.098)	(-2.215, -1.985)	(-2.144, -1.971)	(-2.062, -1.945)
	-1	(-1.162, -0.972)	(-1.087, -0.966)	(-1.063, -0.970)	(-1.164, -0.981)	(-1.077, -0.962)	(-1.058, -0.959)
	-2	(-2.261, -1.988)	(-2.137, -1.964)	(-2.095, -1.965)	(-2.215, -1.985)	(-2.144, -1.971)	(-2.062, -1.945)
120	1	(0.986, 1.147)	(0.982, 1.081)	(0.977, 1.053)	(-1.154, -0.975)	(-1.060, -0.961)	(-1.050, -0.973)
	2	(2.016, 2.246)	(1.989, 2.130)	(1.983, 2.092)	(-2.173, -1.964)	(-2.143, -2.001)	(-2.070, -1.963)
	-1	(-1.136, -0.975)	(-1.074, -0.976)	(-1.055, -0.979)	(-1.154, -0.975)	(-1.060, -0.961)	(-1.050, -0.973)
	-2	(-2.242, -2.009)	(-2.127, -1.987)	(-2.088, -1.980)	(-2.173, -1.964)	(-2.143, -2.001)	(-2.070, -1.963)

the modelling) and gender or survival status of the first child (girl, boy, dead, and "girl" is taken to be the reference in the modelling). In addition, we also consider two cluster level covariates: region of residence (rural, urban, and "rural" is taken to be the reference in the modelling) and administrative division (Barisal, Chittagong, Dhaka, Kulna, Rajshahi, Sylhet, and "Barisal" is taken to be the reference in the modelling).

We apply the proposed model (1.2) together with the homogeneity structure (1.3) to fit the data. The proposed estimation procedure is used to construct the estimates of the unknown parameters, with the threshold δ used in the homogeneity pursuit set to be 4, the

Table 4: **The 95 % Confidence Intervals for Each Component of $\lambda_{1,l}$s for Each Case Under Balanced Design**

		The Proposed Method			The K-Means Method		
J	$\lambda_{1,l}$	$n_j = 40$	$n_j = 80$	$n_j = 120$	$n_j = 40$	$n_j = 80$	$n_j = 120$
40	1	(0.697, 1.083)	(0.804, 1.029)	(0.812, 1.019)	(0.637, 1.070)	(0.720, 1.013)	(0.772, 0.980)
	-2	(-2.162, -1.663)	(-2.092, -1.752)	(-2.072, -1.793)	(-2.142, -1.540)	(-2.040, -1.602)	(-2.049, -1.653)
80	1	(0.784, 1.033)	(0.852, 1.004)	(0.877, 0.995)	(0.716, 1.026)	(0.758, 0.994)	(0.785, 0.984)
	-2	(-2.108, -1.744)	(-2.040, -1.809)	(-2.015, -1.861)	(-2.049, -1.661)	(-2.012, -1.646)	(-2.019, -1.650)
120	1	(0.811, 1.003)	(0.865, 0.988)	(0.890, 0.985)	(0.759, 0.993)	(0.769, 0.974)	(0.796, 0.978)
	-2	(-2.104, -1.769)	(-2.017, -1.853)	(-2.006, -1.875)	(-2.074, -1.680)	(-1.992, -1.679)	(-2.027, -1.663)
		Overfitting Method			Underfitting Method		
J	$\lambda_{1,l}$	$n_j = 40$	$n_j = 80$	$n_j = 120$	$n_j = 40$	$n_j = 80$	$n_j = 120$
40	1	(0.580, 1.227)	(0.707, 1.138)	(0.777, 1.104)	(-0.867, 0.707)	(-0.701, 0.937)	(-0.944, 0.942)
	-2	(-2.440, -1.585)	(-2.255, -1.740)	(-2.192, -1.763)	(-0.939, 0.847)	(-0.969, 1.088)	(-0.936, 0.909)
80	1	(0.752, 1.143)	(0.775, 1.097)	(0.830, 1.050)	(-0.620, 0.532)	(-0.528, 0.544)	(-0.596, 0.644)
	-2	(-2.312, -1.760)	(-2.194, -1.763)	(-2.106, -1.809)	(-0.610, 0.563)	(-0.696, 0.610)	(-0.665, 0.657)
120	1	(0.756, 1.101)	(0.821, 1.049)	(0.860, 1.046)	(-0.503, 0.522)	(-0.482, 0.415)	(-0.469, 0.465)
	-2	(-2.352, -1.771)	(-2.125, -1.853)	(-2.103, -1.835)	(-0.418, 0.458)	(-0.549, 0.555)	(-0.541, 0.413)

Table 5: **The MSE$_\beta$s for Each Case Under Balanced Design**

	The Proposed Method			The K-Means Method		
J	$n_j = 40$	$n_j = 80$	$n_j = 120$	$n_j = 40$	$n_j = 80$	$n_j = 120$
40	0.088	0.027	0.009	0.180	0.113	0.095
80	0.083	0.016	0.004	0.158	0.117	0.090
120	0.091	0.014	0.003	0.154	0.111	0.092
	Overfitting Method			Underfitting Method		
J	$n_j = 40$	$n_j = 80$	$n_j = 120$	$n_j = 40$	$n_j = 80$	$n_j = 120$
40	0.174	0.058	0.035	5.695	5.243	5.138
80	0.157	0.060	0.036	5.410	5.214	5.196
120	0.158	0.059	0.035	5.677	5.176	5.060

bandwidth for producing the estimates of the coefficients of the cluster level variables set to be 25% of the range of the second-birth interval across all clusters, and the kernel function involved taken to be the Epanechnikov kernel. Figure 1 shows the graph of the sorted initial estimates for each individual-level variables. It is very interesting that the range of initial estimates for year of marriage is much smaller than the estimates for other variables. The obtained final results are presented in Table 13.

To explain Table 13, we use the category "Primary" of the covariate "Education" as an

Table 6: **The MSE$_\lambda$s for Each Case Under Balanced Design**

	The Proposed Method			The K-Means Method		
J	$n_j = 40$	$n_j = 80$	$n_j = 120$	$n_j = 40$	$n_j = 80$	$n_j = 120$
40	0.023	0.014	0.009	0.039	0.031	0.025
80	0.015	0.008	0.005	0.032	0.029	0.022
120	0.011	0.006	0.005	0.025	0.025	0.022
	Overfitting Method			Underfitting Method		
J	$n_j = 40$	$n_j = 80$	$n_j = 120$	$n_j = 40$	$n_j = 80$	$n_j = 120$
40	0.048	0.019	0.014	2.572	2.641	2.653
80	0.018	0.010	0.009	2.544	2.592	2.537
120	0.014	0.007	0.006	2.600	2.554	2.452

Table 7: **The MISE(Λ_0)s for Each Case Under Balanced Design**

	The Proposed Method			The K-Means Method		
J	$n_j = 40$	$n_j = 80$	$n_j = 120$	$n_j = 40$	$n_j = 80$	$n_j = 120$
40	0.635	0.186	0.119	0.873	0.459	0.430
80	0.362	0.127	0.058	0.620	0.304	0.288
120	0.309	0.129	0.065	0.587	0.433	0.313
	Overfitting Method			Underfitting Method		
J	$n_j = 40$	$n_j = 80$	$n_j = 120$	$n_j = 40$	$n_j = 80$	$n_j = 120$
40	0.719	0.178	0.110	7.579	7.695	7.606
80	0.311	0.106	0.059	7.645	7.695	7.705
120	0.266	0.108	0.066	7.889	7.820	7.856

example. The 28.08% in the brackets following "Primary" means 28.08% of the women in the data have Primary education. In the column "Estimate", there are 2 values corresponding to "Primary", that means, as far as the coefficient of the dummy variable, which takes value 1 when the woman concerned has Primary education, 0 otherwise, is concerned, the clusters in the data are grouped to 2 groups by the homogeneity pursuit in the proposed estimation procedure, and the clusters in the same group share the same coefficient. Specifically, the coefficient is -0.280 for group 1, 0.251 for group 2. The entries in the column "Standard Error" are the standard errors of the corresponding estimates. The corresponding entry in the column "% of sample", say for example 57.06, means group 1 accounts for 57.06% of the total number of clusters in the data.

From Table 13, we can see that, compared with women with no education, women who

Table 8: **The 95 % Confidence Intervals for Each Component of β_js for Each Case Under Unbalanced Design**

		The Proposed Method			The K-Means Method		
J	$\beta_{i,j}$	$n_j = 40$	$n_j = 80$	$n_j = 120$	$n_j = 40$	$n_j = 80$	$n_j = 120$
40	1	(0.949, 1.156)	(0.950, 1.078)	(0.977, 1.034)	(1.003, 1.205)	(1.014, 1.157)	(1.032, 1.139)
	2	(1.843, 2.080)	(1.924, 2.043)	(1.980, 2.028)	(1.614, 1.993)	(1.587, 1.932)	(1.568, 1.900)
	-1	(-1.179, -0.969)	(-1.081, -0.970)	(-1.049, -0.983)	(-1.208, -0.993)	(-1.154, -1.016)	(-1.122, -1.014)
	-2	(-2.093, -1.753)	(-2.047, -1.850)	(-2.050, -1.873)	(-2.015, -1.605)	(-1.933, -1.599)	(-1.951, -1.630)
80	1	(0.998, 1.156)	(0.977, 1.042)	(0.987, 1.023)	(1.034, 1.202)	(1.032, 1.116)	(1.046, 1.119)
	2	(1.840, 2.061)	(1.953, 2.027)	(1.986, 2.017)	(1.671, 1.958)	(1.668, 1.894)	(1.627, 1.862)
	-1	(-1.135, -0.996)	(-1.036, -0.972)	(-1.019, -0.986)	(-1.191, -1.047)	(-1.114, -1.023)	(-1.123, -1.045)
	-2	(-2.065, -1.843)	(-2.029, -1.946)	(-2.014, -1.979)	(-1.880, -1.604)	(-1.902, -1.672)	(-1.846, -1.616)
120	1	(1.003, 1.127)	(0.984, 1.043)	(0.990, 1.018)	(1.060, 1.188)	(1.056, 1.131)	(1.053, 1.114)
	2	(1.871, 2.046)	(1.960, 2.024)	(1.991, 2.015)	(1.635, 1.863)	(1.634, 1.830)	(1.647, 1.837)
	-1	(-1.119, -1.007)	(-1.033, -0.983)	(-1.015, -0.988)	(-1.170, -1.052)	(-1.109, -1.037)	(-1.095, -1.035)
	-2	(-2.050, -1.880)	(-2.019, -1.945)	(-2.016, -1.980)	(-1.903, -1.677)	(-1.873, -1.683)	(-1.875, -1.695)
		Overfitting Method			Underfitting Method		
J	$\beta_{i,j}$	$n_j = 40$	$n_j = 80$	$n_j = 120$	$n_j = 40$	$n_j = 80$	$n_j = 120$
40	1	(0.927, 1.187)	(0.938, 1.109)	(0.956, 1.086)	(-1.096, -0.874)	(-1.121, -0.961)	(-1.093, -0.933)
	2	(1.932, 2.314)	(1.931, 2.177)	(1.948, 2.132)	(-2.246, -1.880)	(-2.184, -1.966)	(-2.190, -1.952)
	-1	(-1.203, -0.919)	(-1.112, -0.941)	(-1.083, -0.951)	(-1.096, -0.874)	(-1.121, -0.961)	(-1.093, -0.933)
	-2	(-2.336, -1.932)	(-2.164, -1.928)	(-2.134, -1.946)	(-2.246, -1.880)	(-2.184, -1.966)	(-2.190, -1.952)
80	1	(0.972, 1.177)	(0.968, 1.084)	(0.973, 1.065)	(-1.217, -0.988)	(-1.094, -0.976)	(-1.078, -0.989)
	2	(1.987, 2.285)	(1.969, 2.135)	(1.972, 2.104)	(-2.401, -2.075)	(-2.157, -1.980)	(-2.086, -1.956)
	-1	(-1.163, -0.961)	(-1.080, -0.962)	(-1.065, -0.972)	(-1.217, -0.988)	(-1.094, -0.976)	(-1.078, -0.989)
	-2	(-2.286, -1.993)	(-2.141, -1.968)	(-2.100, -1.968)	(-2.401, -2.075)	(-2.157, -1.980)	(-2.086, -1.956)
120	1	(0.986, 1.154)	(0.980, 1.078)	(0.981, 1.057)	(-1.140, -0.985)	(-1.060, -0.962)	(-1.030, -0.955)
	2	(2.004, 2.244)	(1.989, 2.127)	(1.987, 2.095)	(-2.336, -2.048)	(-2.100, -1.955)	(-2.055, -1.936)
	-1	(-1.146, -0.987)	(-1.074, -0.977)	(-1.054, -0.978)	(-1.140, -0.985)	(-1.060, -0.962)	(-1.030, -0.955)
	-2	(-2.251, -2.017)	(-2.120, -1.979)	(-2.092, -1.981)	(-2.336, -2.048)	(-2.100, -1.955)	(-2.055, -1.936)

were educated to primary level or beyond have longer second-birth interval for more than half the clusters. This can be interpreted as educated women tend to be more devoted to their career, therefore, more likely to delay giving second birth. However, we can also see, from Table 13, there is a small number of clusters, where women who were educated to primary level have shorter second-birth interval. For most of the clusters, Muslims have shorter second-birth interval. However, there are some clusters where Muslims have longer second-birth interval. This indicates there may be some cultural difference between these clusters and others. Compared with the first child being a girl, if the first child is a boy,

Table 9: **The 95 % Confidence Intervals for Each Component of $\lambda_{1,l}$s for Each Case Under Unbalanced Design**

		The Proposed Method			The K-Means Method		
J	$\lambda_{1,l}$	$n_j = 40$	$n_j = 80$	$n_j = 120$	$n_j = 40$	$n_j = 80$	$n_j = 120$
40	1	(0.734, 1.060)	(0.806, 1.030)	(0.829, 1.026)	(0.708, 1.032)	(0.729, 1.010)	(0.751, 1.005)
	-2	(-2.153, -1.638)	(-2.102, -1.744)	(-2.061, -1.789)	(-2.157, -1.522)	(-2.084, -1.627)	(-2.035, -1.631)
80	1	(0.804, 1.014)	(0.839, 1.003)	(0.873, 0.987)	(0.743, 0.999)	(0.754, 0.999)	(0.771, 0.972)
	-2	(-2.114, -1.721)	(-2.046, -1.816)	(-2.033, -1.854)	(-2.063, -1.620)	(-2.045, -1.658)	(-2.009, -1.671)
120	1	(0.818, 1.014)	(0.866, 0.993)	(0.896, 0.978)	(0.732, 1.001)	(0.784, 0.972)	(0.789, 0.978)
	-2	(-2.084, -1.770)	(-2.013, -1.838)	(-2.018, -1.881)	(-2.037, -1.682)	(-1.998, -1.680)	(-2.018, -1.684)
		Overfitting Method			Underfitting Method		
J	$\lambda_{1,l}$	$n_j = 40$	$n_j = 80$	$n_j = 120$	$n_j = 40$	$n_j = 80$	$n_j = 120$
40	1	(0.619, 1.285)	(0.732, 1.163)	(0.777, 1.112)	(-0.797, 0.918)	(-0.844, 0.998)	(-0.947, 0.874)
	-2	(-2.440, -1.566)	(-2.266, -1.657)	(-2.189, -1.764)	(-0.89, 0.863)	(-0.862, 0.906)	(-0.940, 0.811)
80	1	(0.723, 1.167)	(0.779, 1.062)	(0.825, 1.063)	(-0.554, 0.604)	(-0.504, 0.496)	(-0.630, 0.628)
	-2	(-2.360, -1.709)	(-2.193, -1.787)	(-2.108, -1.827)	(-0.625, 0.627)	(-0.617, 0.534)	(-0.640, 0.598)
120	1	(0.790, 1.119)	(0.823, 1.056)	(0.834, 1.055)	(-0.395, 0.503)	(-0.547, 0.438)	(-0.542, 0.507)
	-2	(-2.299, -1.748)	(-2.152, -1.801)	(-2.132, -1.832)	(-0.471, 0.426)	(-0.559, 0.489)	(-0.504, 0.470)

Table 10: **The MSE$_\beta$s for Each Case Under Unbalanced Design**

	The Proposed Method			The K-Means Method		
J	$n_j = 40$	$n_j = 80$	$n_j = 120$	$n_j = 40$	$n_j = 80$	$n_j = 120$
40	0.088	0.024	0.006	0.164	0.112	0.099
80	0.097	0.015	0.004	0.163	0.095	0.105
120	0.093	0.014	0.004	0.172	0.106	0.110
	Overfitting Method			Underfitting Method		
J	$n_j = 40$	$n_j = 80$	$n_j = 120$	$n_j = 40$	$n_j = 80$	$n_j = 120$
40	0.167	0.058	0.036	5.432	5.176	5.031
80	0.180	0.060	0.036	5.258	5.203	5.107
120	0.170	0.061	0.036	5.394	5.210	5.122

the second-birth interval becomes significantly longer for some clusters, and it does not have significant difference for the rest. This reflects the culture of favouring boys in some parts of Bangladesh. When the first child is dead, the second-birth interval becomes even shorter for all of the clusters. It is also noticeable that women in urban areas tend to have longer second-birth interval than those in rural areas, which makes sense, this is because use of contraceptives in rural areas is lower than urban areas. The second-birth intervals are shorter in Chittagong than in the other divisions. This regional effect is as expected, because

Table 11: **The MSE$_\lambda$s for Each Case Under Unbalanced Design**

	The Proposed Method			The K-Means Method		
J	$n_j = 40$	$n_j = 80$	$n_j = 120$	$n_j = 40$	$n_j = 80$	$n_j = 120$
40	0.021	0.013	0.007	0.034	0.031	0.026
80	0.013	0.007	0.005	0.032	0.023	0.025
120	0.010	0.006	0.005	0.025	0.023	0.025
	Overfitting Method			Underfitting Method		
J	$n_j = 40$	$n_j = 80$	$n_j = 120$	$n_j = 40$	$n_j = 80$	$n_j = 120$
40	0.034	0.019	0.014	2.578	2.561	2.646
80	0.019	0.009	0.007	2.577	2.513	2.548
120	0.017	0.008	0.005	2.551	2.547	2.499

Table 12: **The MISE(Λ_0)s for Each Case Under Unbalanced Design**

	The Proposed Method			The K-Means Method		
J	$n_j = 40$	$n_j = 80$	$n_j = 120$	$n_j = 40$	$n_j = 80$	$n_j = 120$
40	0.620	0.176	0.097	0.666	0.415	0.344
80	0.335	0.112	0.062	0.596	0.360	0.344
120	0.360	0.115	0.058	0.658	0.397	0.362
	Overfitting Method			Underfitting Method		
J	$n_j = 40$	$n_j = 80$	$n_j = 120$	$n_j = 40$	$n_j = 80$	$n_j = 120$
40	0.520	0.177	0.110	7.348	7.727	7.695
80	0.328	0.104	0.070	7.813	7.881	7.839
120	0.296	0.114	0.060	7.755	7.933	7.864

Chittagong is the most religiously conservative part of Bangladesh, use of contraceptives there is rare.

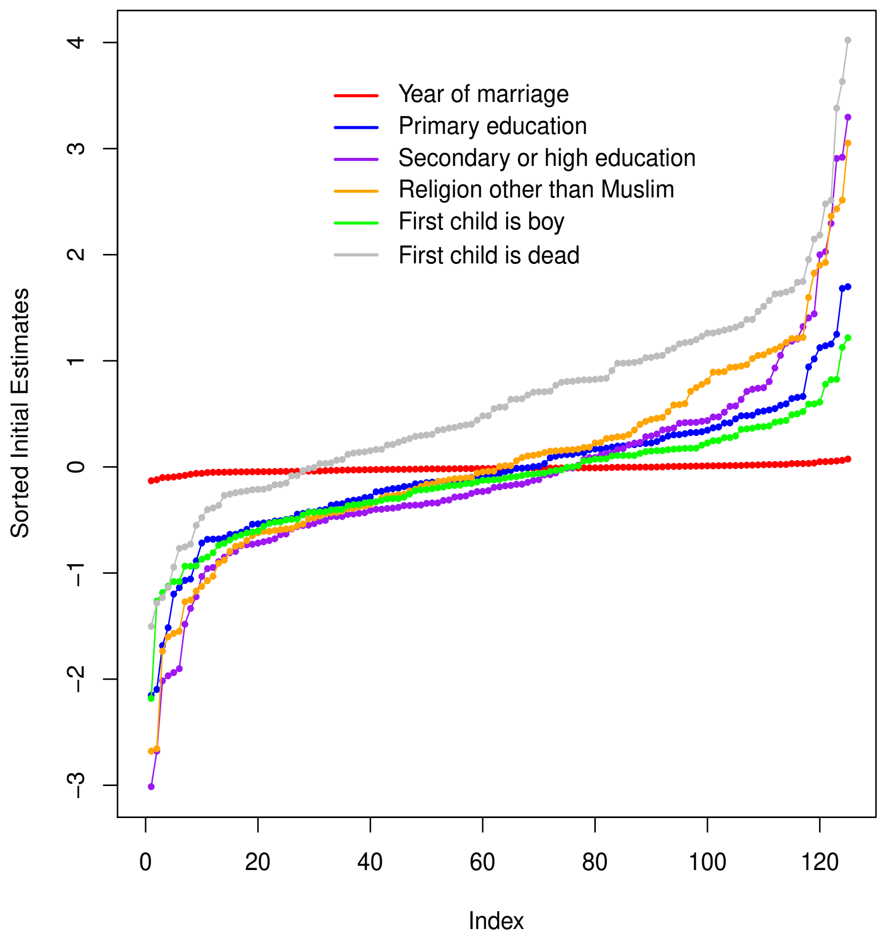

Figure 1: Plot of the sorted initial estimates for each individual-level variables.

Table 13: **Results of the Analysis of the Second Birth Interval in Bangladesh**

Covariates of Interest	Estimate	Standard Error	% of sample
Individual Level			
Year of marriage	-0.013	0.002	100.00
Education			
None (54.11 %)	0.000	-	100.00
Primary (28.08 %)	-0.280	0.042	57.06
	0.251	0.047	42.94
Secondary or higher (17.81 %)	-0.936	0.189	5.03
	-0.286	0.048	68.23
	0.475	0.089	23.46
	1.874	0.367	3.28
Religion			
Muslim (88.21 %)	0.000	-	100.00
Others (11.79 %)	-1.068	0.179	7.00
	-0.200	0.054	65.68
	0.725	0.092	27.32
Gender or survival status of 1st child			
Girl (42.43 %)	0.000	-	100.00
Boy (43.18 %)	-0.429	0.051	31.74
	0.059	0.034	68.26
Dead (14.39 %)	0.784	0.051	52.66
	0.974	0.058	44.37
	1.994	0.310	2.97
Cluster Level			
Type of region of residence			
Rural (83.99 %)	0.000	-	100.00
Urban (16.01 %)	-0.063	0.042	100.00
Administrative division			
Barisal (10.61 %)	0.000	-	100.00
Chittagong (15.11 %)	0.126	0.070	100.00
Dhaka (27.80 %)	-0.078	0.099	100.00
Kulna (11.78 %)	-0.111	0.080	100.00
Rajshahi (24.96 %)	-0.152	0.061	100.00
Sylhet (9.74 %)	0.046	0.143	100.00

6. Concluding remarks

In this paper, we propose a new multilevel modelling strategy for clustered survival data analysis. The methodological advantage of the proposed modelling is that it has successfully avoided computation of multiple integrals and abundance of nuisance parameters, which makes the implementation of the proposed modelling much easier than traditional methods. In applications, the proposed modelling strategy enables investigators to explore individual/subgroup attributes of covariates, which traditional methods cannot because they assume the impact of covariates is constant over clusters and they only use cluster effects to account for differences among clusters. Our application to second birth intervals shows that the impact of some factors on the birth interval is homogeneous within each cluster and varies across clusters, thereby revealing cultural differences between some clusters. Such findings cannot be obtained by the traditional multilevel modelling strategy. In conclusion, the proposed multilevel modelling strategy facilitates implementation and offers the advantage of distinguishing cluster coefficients in applications.

References

Andersen, P.K., Gill, R.D., 1982. Cox's regression model for counting processes: a large sample study. The Annals of Statistics 10, 1100–1120.

Ando, T., Bai, J., 2017. Clustering huge number of financial time series: A panel data approach with high-dimensional predictors and factor structures. Journal of the American Statistical Association 112, 1182–1198.

Azuma, K., 1967. Weighted sums of certain dependent random variables. Tohoku Mathematical Journal, Second Series 19, 357–367.

Bai, J., 1997. Estimating multiple breaks one at a time. Econometric Theory 13, 315–352.

Bonhomme, S., Manresa, E., 2015. Grouped patterns of heterogeneity in panel data. Econometrica 83, 1147–1184.

Bradic, J., Fan, J., Jiang, J., 2011. Regularization for cox's proportional hazards model with np-dimensionality. Annals of Statistics 39, 3092–3210.

Breslow, N., 1972. Comment on "regression and life tables" by DR Cox. Journal of the Royal Statistical Society, Series B 34, 216–217.

David, C.R., 1972. Regression models and life tables (with discussion). Journal of the Royal Statistical Society, Series B 34, 187–220.

Fleming, T.R., Harrington, D.P., 2011. Counting processes and survival analysis. John Wiley & Sons.

Hartigan, J.A., Wong, M.A., 1979. Algorithm as 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C 28, 100–108.

Harvey, G., 2003. Multilevel statistical models. 3rd edition. Arnold, London .

Hoeffding, W., 1963. Probability inequalities for sums of bounded random variables, Taylor & Francis. volume 58, pp. 13–30.

Ke, Y., Li, J., Zhang, W., 2016. Structure identification in panel data analysis. The Annals of Statistics 44, 1193–1233.

Ke, Z.T., Fan, J., Wu, Y., 2015. Homogeneity pursuit. Journal of the American Statistical Association 110, 175–194.

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

Ortega, J.M., Rheinboldt, W.C., 1970. Iterative solution of nonlinear equations in several variables. Academic Press, San Diego.

Su, L., Ju, G., 2018. Identifying latent grouped patterns in panel data models with interactive fixed effects. Journal of Econometrics 206, 554–573.

Su, L., Shi, Z., Phillips, P.C., 2016. Identifying latent structures in panel data. Econometrica 84, 2215–2264.

Su, Liangjun, W.X., Jin, S., 2019. Sieve estimation of time-varying panel data models with latent structures. Journal of Business & Economic Statistics 37, 334–349.

Venkatraman, E.S., 1993. Consistency results in multiple change-point problems. Ph.D. thesis, Stanford Univ., ProQuest LLC, Ann Arbor, MI .

Volinsky, C.T., Raftery, A.E., 2000. Bayesian information criterion for censored survival models. Biometrics 56, 256–262.

Vostrikova, L., 1981. Detection of the disorder in multidimensional random-processes. Doklady Akademii Nauk SSSR 259, 270–274.

Wang, W., Phillips, P.C., Su, L., 2018. Homogeneity pursuit in panel data models: Theory and application. Journal of Applied Econometrics 33, 797–815.

Wang, W., Su, L., 2019. Identifying latent group structures in nonlinear panels. Journal of Econometrics , to appear.

Zhang, W., Steele, F., 2004. A semiparametric multilevel survival model. Journal of the Royal Statistical Society: Series C (Applied Statistics) 53, 387–404.

Appendix

We first prove three lemmas and then prove Theorems 1-3.

Lemma 1. Assume that Condition 1 holds. Then $\max_{1\le j\le J} \|n_j^{1/2}(\tilde{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*)\| = O_p((\log J)^{1/2})$.

Proof. We prove the statement in the following three steps:
(i) We first show that $\max_{1\le j\le J} n_j^{-1/2}\dot{\mathcal{L}}_j(\boldsymbol{\beta}_j^*) = O_P((\log J)^{1/2})$. By definition, $\dot{\mathcal{L}}_j(\boldsymbol{\beta}_j^*) = \sum_{i=1}^{n_j}\int_0^\tau[X_{ij}-E_j(u,\boldsymbol{\beta}_j^*)]dN_{ij}(u) = \sum_{i=1}^{n_j}\int_0^\tau[X_{ij}-E_j(u,\boldsymbol{\beta}_j^*)]dM_{ij}(u)$. Let $a_{ijk}(u) = X_{ijk}-E_{jk}(u,\boldsymbol{\beta}_j^*), k = 1,...,p$. For the jth cluster, let $t_{(0),j}=0$ and denote the distinct event times by $t_{(1),j}<\cdots<t_{(T_j),j}$. Then $t_{(\ell),j}$ are stopping times. For $\ell = 0,...,T_j$, define

$$Z_{\ell,jk} = \sum_{i=1}^{n_j}\int_0^{t_{(\ell),j}} a_{ijk}(u)dN_{ij}(u) = \sum_{i=1}^{n_j}\int_0^{t_{(\ell),j}} a_{ijk}(u)dM_{ij}(u).$$

Note that $\dot{\mathcal{L}}_{jk}(\boldsymbol{\beta}_j^*) = Z_{T_j,jk}$. Since $M_{ij}(u)$ are martingales and $a_{ijk}(u)$ are predictable, $\{Z_{\ell,jk},\ell=0,1,...\}$ is a martingale with the difference $|Z_{\ell,jk}-Z_{\ell-1,jk}|\le \max_{i,j,k,u}|a_{ijk}(u)|\le K$. By the martingale version of the Hoeffding (1963) inequality [Azuma (1967)], for any $x>0$,

$$P(|Z_{T_j,jk}|>n_jx)\le 2\exp\left(-n_j^2x^2/(2K^2T_j)\right)\le 2e^{-n_jx^2/(2K^2)}.$$

For any $1\le k\le p$, for any $C>0$,

$$P(\max_{1\le j\le J} n_j^{-1/2}|\dot{\mathcal{L}}_{jk}(\boldsymbol{\beta}_j^*)|\ge CK(\log J)^{1/2})\le 2Je^{-C^2(\log J)/2} = 2J^{1-C^2/2}.$$

For sufficiently large C, $2J^{1-C^2/2}$ tends to 0. Hence, $\max_{1\le j\le J} n_j^{-1/2}\dot{\mathcal{L}}_j(\boldsymbol{\beta}_j^*) = O_P((\log J)^{1/2})$.
(ii) We prove that with probability tending to 1,

$$\inf_{1\le j\le J}\inf_{\|b\|=1,b\in R^p} b^T(-n_j^{-1}\ddot{\mathcal{L}}_j(\boldsymbol{\beta}_j^*))b\ge c_0e^{-K^2}/2.$$

Let $a_{jkm}(u) = (V_j(u,\beta_j^*))_{(km)} = \sum_{i=1}^{n_j} w_{ij}(u,\beta_j^*)\{X_{ijk}-E_{jk}(u,\beta_j^*)\}\{X_{ijm}-E_{jm}(u,\beta_j^*)\}$, where $w_{ij}(u,\boldsymbol{\beta}) = Y_{ij}(u)\exp(X_{ij}^T\boldsymbol{\beta})/[n_jS_j^{(0)}(u,\boldsymbol{\beta})]$. Note that

$$-n_j^{-1}\ddot{\mathcal{L}}_j(\boldsymbol{\beta}_j^*) = n_j^{-1}\sum_{i=1}^{n_j}\int_0^t V_j(u,\boldsymbol{\beta}_j^*)dN_{ij}(u),$$

and

$$\sum_{i=1}^{n_j}\int_0^t V_j(u,\boldsymbol{\beta}_j^*)(dN_{ij}(u)-Y_{ij}(u)\exp(X_{ij}^T\boldsymbol{\beta}_j^*)h_{1,j}(u)du)$$

is a martingale. By the Azuma-Hoeffding inequality, for any $x>0$,

$$P(\max_{1\le j\le J} n_j^{-1}|\sum_{i=1}^{n_j}\int_0^{t_{(T_j),j}} a_{jkm}(u,\boldsymbol{\beta}_j^*)(dN_{ij}(u)-Y_{ij}(u)\exp(X_{ij}^T\boldsymbol{\beta}_j^*)h_{1,j}(u)du)|>K^2x)\le 2Je^{-nx^2}.$$

It follows that with probability less than $2Je^{-n(c_0e^{-K^2}/(2K^2))^2}$,

$$\sup_{\|b\|=1, b\in R^p} |b^T[-n_j^{-1}\ddot{\mathcal{L}}_j - n_j^{-1}\sum_{i=1}^{n_j}\int_0^t V_j(u,\boldsymbol{\beta}_j^*)Y_{ij}(u)\exp{(X_{ij}^T\boldsymbol{\beta}_j^*)}h_{1,j}(u)du]b| \leq c_0e^{-K^2}/2.$$

By Condition 1 (i)-(iv), we have, with probability tending to 1,

$$\inf_{1\leq j\leq J}\inf_{\|b\|=1, b\in R^p} b^T(-n_j^{-1}\ddot{\mathcal{L}}_j(\boldsymbol{\beta}_j^*))b \geq c_0e^{-K^2}/2.$$

(iii) We prove that for any $\epsilon > 0$, there exists a constant $C > 0$ such that for all n sufficiently large,

$$P(\cap_{j=1}^J\{\sup_{\|\boldsymbol{\beta}_j - \boldsymbol{\beta}_j^*\| = C(\frac{\log J}{n_j})^{1/2}} (\boldsymbol{\beta}_j - \boldsymbol{\beta}_j^*)^T\dot{\mathcal{L}}_j(\boldsymbol{\beta}_j) < 0\}) > 1-\epsilon. \tag{A.1}$$

By the concavity of $\mathcal{L}_j(\cdot)$ and Theorem 6.3.4 of Ortega and Rheinboldt (1970), Condition (A.1) is sufficient to show that $\max_{1\leq j\leq J}\|n_j^{1/2}(\tilde{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*)\| = O_p((\log J)^{1/2})$. For any $\boldsymbol{\beta}_j$ satisfying $\|\boldsymbol{\beta}_j - \boldsymbol{\beta}_j^*\| = C(\frac{\log J}{n_j})^{1/2}$, let $b_j = \boldsymbol{\beta}_j - \boldsymbol{\beta}_j^*$, $a_{ij} = a_{ij}(u) = b_j^T\{X_{ij} - E_j(u,\boldsymbol{\beta}_j^*),\}$, $w_{ij} = w_{ij}(u) = Y_{ij}(u)\exp{(X_{ij}^T\boldsymbol{\beta}_j^*)}$, $c_j = c_j(u) = (\max_i a_{ij}(u) + \min_i a_{ij}(u))/2$, and $\eta_j = KC(\frac{\log J}{n_j})^{1/2}$. Note that $\max_i|a_{ij}(u) - c_j(u)| \leq \frac{\eta_j}{2}$ and

$$\begin{aligned}
&b_j^T\{E_j(u,\boldsymbol{\beta}_j^* + b_j) - E_j(u,\boldsymbol{\beta}_j^*)\} \\
&= (\sum_{i=1}^{n_j} a_{ij}w_{ij}e^{a_{ij}})/(\sum_{i=1}^{n_j} w_{ij}e^{a_{ij}}) - (\sum_{i=1}^{n_j} a_{ij}w_{ij})/(\sum_{i=1}^{n_j} w_{ij}) \\
&= (\sum_{i,k=1}^{n_j}(a_{ij} - a_{kj})(e^{a_{ij}-c_j} - e^{a_{kj}-c_j})w_{ij}w_{kj})/(\sum_{i,k=1}^{n_j} 2w_{ij}w_{kj}e^{a_{ij}-c_j}) \\
&\geq \exp{(-2\max_i|a_{ij} - c_j|)}(\sum_{i,k=1}^{n_j}(a_{ij} - a_{kj})^2w_{ij}w_{kj})/(\sum_{i,k=1}^{n_j} 2w_{ij}w_{kj}) \\
&\geq \exp{(-\eta_j)}(\sum_i^{n_j} w_{ij}a_{ij}^2)/(\sum_i^{n_j} w_{ij}),
\end{aligned}$$

where the first inequality comes from $(e^y - e^x)/(y-x) \geq e^{-(|y|\vee|x|)}$, and the second one comes from $\sum_i^{n_j} w_{ij}a_{ij} = 0$ and $\sum_{i,k=1}^{n_j} w_{ij}w_{kj}(a_{ij} - a_{kj})^2 = (2\sum_{i,k=1}^{n_j} w_{ij}a_{ij}^2)(\sum_{i,k=1}^{n_j} w_{ij})$. It follows that

$$b_j^T[\dot{\mathcal{L}}_j(\boldsymbol{\beta}_j) - \dot{\mathcal{L}}_j(\boldsymbol{\beta}_j^*)] \leq -e^{-\eta_j}\sum_i^{n_j}\int_0^\tau(\sum_i^{n_j} w_{ij}a_{ij}^2)(\sum_i^{n_j} w_{ij})^{-1}dN_{ij}(u) = e^{-\eta_j}b_j^T\ddot{\mathcal{L}}(\boldsymbol{\beta}_j^*)b_j.$$

Hence, by (i) and (ii), uniformly for all j, we have

$$\begin{aligned}(\boldsymbol{\beta}_j - \boldsymbol{\beta}_j^*)^T \dot{\boldsymbol{\mathcal{L}}}_j(\boldsymbol{\beta}_j) &= (\boldsymbol{\beta}_j - \boldsymbol{\beta}_j^*)^T \dot{\boldsymbol{\mathcal{L}}}_j(\boldsymbol{\beta}_j^*) + (\boldsymbol{\beta}_j - \boldsymbol{\beta}_j^*)^T [\dot{\boldsymbol{\mathcal{L}}}_j(\boldsymbol{\beta}_j) - \dot{\boldsymbol{\mathcal{L}}}_j(\boldsymbol{\beta}_j^*)] \\ &\leq n_j [O_P((\frac{\log J}{n_j})^{1/2}) KC(\frac{\log J}{n_j})^{1/2} - K^2 C^2 \frac{\log J}{n_j} e^{-\eta_j} c_0 e^{-K^2}/2].\end{aligned}$$

Since $\log J/n \to 0$ as $n \to \infty$, it is easy to see that (A.1) holds and this completes the proof.

Without loss of generality, assume $\boldsymbol{\beta}_{(1)} < ... < \boldsymbol{\beta}_{(H)}$. Write $s_0 = 0$. Let $s_k = |\mathcal{B}_k|$ be the size of $\mathcal{B}_k$ and $r_k = \sum_{\ell=1}^{k} s_\ell, k = 0, ..., H$. Define $\tilde{\mathcal{B}}_k := \{(i,j), \sum_{\ell}^{k-1} s_\ell + 1 \leq r_{ij} \leq \sum_{\ell=1}^{k} s_\ell\}, k = 1, ..., H$. The following lemma shows that with probability tending to 1, the homogeneity structure in the estimated coefficients is identical to that in (2.2).

Lemma 2. If Condition 1 holds, then

$$\lim_{n \to \infty} P(\cap_{k=1}^{H} \{\tilde{\mathcal{B}}_k = \mathcal{B}_k\}) = 1.$$

Proof. Let $\Delta = \min_{2 \leq k \leq H} |\boldsymbol{\beta}_{(k)} - \boldsymbol{\beta}_{(k-1)}|$. Let $\epsilon = \Delta/2$. By Lemma 1 and Condition 1 (iv), as n goes to ∞, with probability tending to 1, $\max_{1 \leq j \leq J} \|\tilde{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j\| < \Delta/3$ and hence $\max_{1 \leq j \leq J, 1 \leq i \leq p} \|\tilde{\beta}_{ij} - \beta_{ij}\| < \Delta/2$. It is sufficient to show that for any $(i_1, j_1) \neq (i_2, j_2)$ satisfying that $\beta_{i_1 j_1} \neq \beta_{i_2 j_2}$, $(\tilde{\beta}_{i_1 j_1} - \tilde{\beta}_{i_2 j_2})(\beta_{i_1 j_1} - \beta_{i_2 j_2}) > 0$. This easily follows since

$$(\beta_{i_1 j_1} - \beta_{i_2 j_2})(\tilde{\beta}_{i_1 j_1} - \tilde{\beta}_{i_2 j_2}) \geq (|\beta_{i_1 j_1} - \beta_{i_2 j_2}|)(|\beta_{i_1 j_1} - \beta_{i_2 j_2}| - 2\Delta/3) = \Delta^2/3 > 0.$$

Next we prove the consistency of the binary segmentation procedure in identifying the homogeneity structure.

Lemma 3. If Conditions 1 and 2 hold, then

$$\lim_{n \to \infty} P(\cap_{\ell=1}^{\hat{H}-1} \{\hat{k}_{(\ell)} = r_\ell\} \cap \{\hat{H} = H\}) = 1.$$

Proof. Let $b_{(\ell)}^0$ be the true coefficient associated with $b_{(\ell)}, \ell = 1, ..., Jp$. By Lemma 2, with probability tending to 1, $b_{(\ell)}^0 = \beta_{(d)}$, for $\sum_{k=1}^{d-1} s_k + 1 \leq \ell \leq \sum_{k=1}^{d} s_k, d = 1, ..., H$. Write $b_{(\ell)} = b_{(\ell)}^0 + e_{(\ell)}$. By definition,

$$\Delta_{1,Jp}(\hat{k}_1) = \max_{1 \leq \kappa < Jp} \Delta_{1,Jp}(\kappa),$$

where

$$\Delta_{ij}(\kappa) = \sqrt{\frac{(j-\kappa)(\kappa-i+1)}{j-i+1}} \left(\frac{\sum_{l=\kappa+1}^{j} b_{(l)}}{j-\kappa} - \frac{\sum_{l=i}^{\kappa} b_{(l)}}{\kappa-i+1} \right).$$

First we prove that with probability tending to 1, $\hat{k}_1 = r_k$ for some $1 \le k \le H-1$ by contradiction. Otherwise, there exist k and m such that $\hat{k}_1 = r_k + m$, where $0 \le k \le H-1$ and $1 \le m < s_{k+1}$. There are three cases which we will consider one by one.

Case 1: $k = 0$ and $\hat{k}_1 = m < s_1 = r_1$.

Let $\bar{\beta}_1 = \frac{\sum_{l=1}^{m} b^0_{(l)}}{m}$ and $\bar{\beta}_2 = \frac{\sum_{l=r_1+1}^{Jp} b^0_{(l)}}{Jp-r_1}$. We have $\bar{\beta}_1 < \bar{\beta}_2$,

$$\Delta_{1,Jp}(\hat{k}_1) = \sqrt{\frac{m}{Jp(Jp-m)}}(Jp-r_1)(\bar{\beta}_2 - \bar{\beta}_1) + \sqrt{\frac{(Jp-m)m}{Jp}}\left(\frac{\sum_{l=m+1}^{Jp} e_{(l)}}{Jp-m} - \frac{\sum_{l=1}^{m} e_{(l)}}{m}\right),$$

and

$$\Delta_{1,Jp}(r_1) = \sqrt{\frac{r_1(Jp-r_1)}{Jp}}(\bar{\beta}_2 - \bar{\beta}_1) + \sqrt{\frac{(Jp-r_1)r_1}{Jp}}\left(\frac{\sum_{l=r_1+1}^{Jp} e_{(l)}}{Jp-r_1} - \frac{\sum_{l=1}^{r_1} e_{(l)}}{r_1}\right).$$

By Lemma 2,

$$\Delta_{1,Jp}(r_1) - \Delta_{1,Jp}(\hat{k}_1) = [\sqrt{\frac{r_1(Jp-r_1)}{Jp}} - \sqrt{\frac{m}{Jp(Jp-m)}}(Jp-r_1)](\bar{\beta}_2 - \bar{\beta}_1) + O_p(\sqrt{Jp}\frac{\log Jp}{\sqrt{n}}).$$

By Condition (i),

$$\Delta_{1,Jp}(r_1) - \Delta_{1,Jp}(\hat{k}_1) = [\sqrt{\frac{r_1(Jp-r_1)}{Jp}} - \sqrt{\frac{m}{Jp(Jp-m)}}(Jp-r_1)](\bar{\beta}_2 - \bar{\beta}_1) + o_p(\frac{1}{\sqrt{Jp}}).$$

By calculation, uniformly for $1 \le m < r_1$,

$$\sqrt{r_1(Jp-r_1)} - \sqrt{\frac{m}{(Jp-m)}}(Jp-r_1) = (Jp-r_1)[\sqrt{\frac{r_1}{(Jp-r_1)}} - \sqrt{\frac{m}{(Jp-m)}}]$$

$$\ge (Jp-r_1)[\sqrt{\frac{r_1}{(Jp-r_1)}} - \sqrt{\frac{r_1-1}{(Jp-r_1+1)}}]$$

$$\ge (Jp-r_1)\frac{\frac{r_1}{(Jp-r_1)} - \frac{r_1-1}{(Jp-r_1+1)}}{2\sqrt{\frac{r_1}{(Jp-r_1)}}} = \frac{Jp}{2(Jp-r_1+1)}\sqrt{\frac{Jp-r_1}{r_1}} = \frac{Jp(Jp-r_1)}{2(Jp-r_1+1)}\frac{1}{Jp-r_1}\sqrt{\frac{Jp-r_1}{r_1}}$$

$$\ge \frac{1}{4}\frac{Jp}{\sqrt{(Jp-r_1)r_1}} \ge \frac{1}{2}.$$

Thus, with probability tending to 1,

$$\Delta_{1,Jp}(r_1) - \Delta_{1,Jp}(\hat{k}_1) > 0$$

which yields the contradiction.

Case 2: $k = H-1$ and $\hat{k}_1 = r_{H-1} + m$, where $1 \le m < s_H$.

Let $\bar{\beta}_1 = \frac{\sum_{l=1}^{r_{H-1}} b^0_{(l)}}{r_{H-1}}$ and $\bar{\beta}_2 = \frac{\sum_{l=r_{H-1}+1}^{Jp} b^0_{(l)}}{Jp-r_{H-1}}$. We have $\bar{\beta}_1 < \bar{\beta}_2$,

$$\Delta_{1,Jp}(\hat{k}_1) = \sqrt{\frac{(s_H-m)(Jp-s_H+m)}{Jp}} \left(\frac{\sum_{l=r_{H-1}+m+1}^{Jp} b^0_{(l)}}{s_H-m} - \frac{\sum_{l=1}^{r_{H-1}+m} b^0_{(l)}}{Jp-s_H+m} \right)$$

$$+\sqrt{\frac{(s_H-m)(Jp-s_H+m)}{Jp}} \left(\frac{\sum_{l=r_{H-1}+m+1}^{Jp} e^0_{(l)}}{s_H-m} - \frac{\sum_{l=1}^{r_{H-1}+m} e^0_{(l)}}{Jp-s_H+m} \right)$$

$$= \sqrt{\frac{s_H-m}{Jp(Jp-s_H+m)}}(\bar{\beta}_2-\bar{\beta}_1)r_{H-1} + o_p(\frac{1}{\sqrt{Jp}}),$$

and

$$\Delta_{1,Jp}(r_{H-1}) = \sqrt{\frac{r_{H-1}(Jp-r_{H-1})}{Jp}}(\bar{\beta}_2-\bar{\beta}_1) + o_p(\frac{1}{\sqrt{Jp}}).$$

Hence,

$$\Delta_{1,Jp}(r_{H-1}) - \Delta_{1,Jp}(\hat{k}_1) = \frac{1}{\sqrt{Jp}}\{r_{H-1}[\sqrt{\frac{s_H}{Jp-s_H}} - \sqrt{\frac{s_H-m}{Jp-s_H+m}}](\bar{\beta}_2-\bar{\beta}_1) + o_p(1)\}$$

Since

$$r_{H-1}[\sqrt{\frac{s_H}{Jp-s_H}} - \sqrt{\frac{s_H-m}{Jp-s_H+m}}] \ge (Jp-s_H)[\sqrt{\frac{s_H}{Jp-s_H}} - \sqrt{\frac{s_H-1}{Jp-s_H+1}}]$$

$$\ge (Jp-s_H)\frac{\frac{s_H}{Jp-s_H} - \frac{s_H-1}{Jp-s_H+1}}{2\sqrt{\frac{s_H}{Jp-s_H}}} = \frac{Jp}{Jp-S_H+1}\frac{1}{2\sqrt{\frac{s_H}{Jp-s_H}}} \ge 1/2,$$

uniformly for $1 \le m < s_H$, it follows that with probability tending to 1, $\Delta_{1,Jp}(r_{H-1}) > \Delta_{1,Jp}(\hat{k}_1)$ which yields the contradiction.

Case 3: $1 \le k \le H-2$. Let

$$\bar{\beta}_1 = \frac{\sum_{i=1}^{k} s_i b^0_{(i)}}{r_k}, \quad \bar{\beta}_2 = \frac{\sum_{i=k+2}^{H} s_i b^0_{(i)}}{Jp-r_{k+1}}, \quad \bar{\beta} = b^0_{(r_k+1)}.$$

We have $\bar{\beta}_2 > \bar{\beta} > \bar{\beta}_1$,

$$\Delta_{1,Jp}(\hat{k}_1) = \sqrt{\frac{(Jp-r_k-m)(r_k+m)}{Jp}} \left(\frac{\sum_{l=r_k+m+1}^{Jp} b^0_{(l)}}{Jp-r_k-m} - \frac{\sum_{l=1}^{r_k+m} b^0_{(l)}}{r_k+m} \right)$$

$$+\sqrt{\frac{(Jp-r_k-m)(r_k+m)}{Jp}}\left(\frac{\sum_{l=r_k+m+1}^{Jp}e^0_{(l)}}{Jp-r_k-m}-\frac{\sum_{l=1}^{r_k+m}e^0_{(l)}}{r_k+m}\right)$$

$$=\sqrt{\frac{(Jp-r_k-m)(r_k+m)}{Jp}}\left(\frac{(s_{k+1}-m)\bar{\beta}+(Jp-r_{k+1})\bar{\beta}_2}{Jp-r_k-m}-\frac{r_k\bar{\beta}_1+m\bar{\beta}}{r_k+m}\right)+o_p(\frac{1}{\sqrt{Jp}}),$$

$$\Delta_{1,Jp}(r_k)=\sqrt{\frac{(Jp-r_k)(r_k)}{Jp}}\left(\frac{s_{k+1}\bar{\beta}+(Jp-r_{k+1})\bar{\beta}_2}{Jp-r_k}-\bar{\beta}_1\right)+o_p(\frac{1}{\sqrt{Jp}}),$$

and

$$\Delta_{1,Jp}(r_{k+1})=\sqrt{\frac{(Jp-r_{k+1})(r_{k+1})}{Jp}}\left(\bar{\beta}_2-\frac{r_k\bar{\beta}_1+s_{k+1}\bar{\beta}}{r_{k+1}}\right)+o_p(\frac{1}{\sqrt{Jp}}).$$

Define the function

$$f(u)=(Jp-r_{k+1})(\bar{\beta}_2-\bar{\beta})\sqrt{u}+r_k(\bar{\beta}-\bar{\beta}_1)\frac{1}{\sqrt{u}},$$

for any $u>0$. Let $u_m=\frac{r_k+m}{Jp-r_k-m}$, $u_1=\frac{r_k}{Jp-r_k}$, $u_2=\frac{r_{k+1}}{Jp-r_{k+1}}$, $u_3=\frac{r_k+1}{Jp-r_k-1}$, and $u_4=\frac{r_{k+1}-1}{Jp-r_{k+1}+1}$. By Condition (ii), $\frac{c_0}{H}\le\frac{kc_0}{(H-k)}\le u_1<u_3\le u_m\le u_4<u_2\le\frac{(k+1)}{(H-k-1)c_0}\le\frac{H}{c_0}$. By Lemma 2 and Condition (i), $\Delta_{1,Jp}(\hat{k}_1)=\frac{1}{\sqrt{Jp}}(f(u_m)+o_p(1))$, $\Delta_{1,Jp}(r_k)=\frac{1}{\sqrt{Jp}}(f(u_1)+o_p(1))$, $\Delta_{1,Jp}(r_{k+1})=\frac{1}{\sqrt{Jp}}(f(u_2)+o_p(1))$, $\Delta_{1,Jp}(r_k+1)=\frac{1}{\sqrt{Jp}}(f(u_3)+o_p(1))$, and $\Delta_{1,Jp}(r_{k+1}-1)=\frac{1}{\sqrt{Jp}}(f(u_4)+o_p(1))$.

(i). If

$$\sqrt{u_1u_2}\ge\frac{r_k(\bar{\beta}-\bar{\beta}_1)}{(Jp-r_{k+1})(\bar{\beta}_2-\bar{\beta})},$$

then

$$u_2-\frac{r_k(\bar{\beta}-\bar{\beta}_1)}{(Jp-r_{k+1})(\bar{\beta}_2-\bar{\beta})}>\frac{r_k(\bar{\beta}-\bar{\beta}_1)}{(Jp-r_{k+1})(\bar{\beta}_2-\bar{\beta})}-u_1.$$

Since $u_2-u_1=\frac{r_{k+1}}{Jp-r_{k+1}}-\frac{r_k}{Jp-r_k}=\frac{Jp(r_{k+1}-r_k)}{(Jp-r_{k+1})(Jp-r_k)}\ge\frac{s_k}{Jp}\ge\frac{c_0}{H}$,

$$\sqrt{u_2u_4}-\frac{r_k(\bar{\beta}-\bar{\beta}_1)}{(Jp-r_{k+1})(\bar{\beta}_2-\bar{\beta})}>\sqrt{u_2u_4}-u_2+\frac{c_0}{2H}.$$

By calculation,

$$f(u_2)-f(u_m)=((Jp-r_{k+1})(\bar{\beta}_2-\bar{\beta})-\frac{r_k(\bar{\beta}-\bar{\beta}_1)}{\sqrt{u_2u_m}})(\sqrt{u_2}-\sqrt{u_m})$$

$$\ge(\frac{r_k(\bar{\beta}-\bar{\beta}_1)}{\sqrt{u_2u_1}}-\frac{r_k(\bar{\beta}-\bar{\beta}_1)}{\sqrt{u_2u_m}})(\sqrt{u_2}-\sqrt{u_m})$$

$$\geq \frac{r_k(\bar{\beta}-\bar{\beta}_1)}{\sqrt{u_2u_1u_m}}(\sqrt{u_m}-\sqrt{u_1})(\sqrt{u_2}-\sqrt{u_m})$$

$$\geq \frac{r_k(\bar{\beta}-\bar{\beta}_1)}{u_2u_m\sqrt{u_1}}(u_m-u_1)(u_2-u_m)$$

$$\geq \frac{r_k(\bar{\beta}-\bar{\beta}_1)}{2u_2u_m\sqrt{u_1}}(u_2-u_1)\min(u_3-u_1,u_2-u_4).$$

By Condition 2 (ii),

$$r_k(u_3-u_1)=r_k(\frac{r_k+1}{Jp-r_k-1}-\frac{r_k}{Jp-r_k})\geq \frac{r_k(Jp)}{(Jp)^2}\geq \frac{c_0}{H},$$

and

$$r_k(u_2-u_4)=r_k(\frac{r_{k+1}}{Jp-r_{k+1}}-\frac{r_{k+1}-1}{Jp-r_{k+1}+1})\geq \frac{r_k(Jp)}{(Jp)^2}\geq \frac{c_0}{H}.$$

Thus,

$$f(u_2)-f(u_m)\geq \frac{c_0^2(\bar{\beta}-\bar{\beta}_1)}{2H^2}\frac{H^{5/2}}{c_0^{5/2}}=\frac{(\bar{\beta}-\bar{\beta}_1)H^{1/2}}{2c_0^{1/2}}.$$

It follows that with probability tending to 1, $\Delta_{1,Jp}(\hat{k}_1)<\Delta_{1,Jp}(r_{k+1})$ which yields the contradiction.

(ii). If

$$\sqrt{u_1u_2}<\frac{r_k(\bar{\beta}-\bar{\beta}_1)}{(Jp-r_{k+1})(\bar{\beta}_2-\bar{\beta})},$$

$$f(u_1)-f(u_m)=(\frac{r_k(\bar{\beta}-\bar{\beta}_1)}{\sqrt{u_1u_m}}-(Jp-r_{k+1})(\bar{\beta}_2-\bar{\beta}))(\sqrt{u_m}-\sqrt{u_1})$$

$$\geq (\frac{r_k(\bar{\beta}-\bar{\beta}_1)}{\sqrt{u_mu_1}}-\frac{r_k(\bar{\beta}-\bar{\beta}_1)}{\sqrt{u_1u_2}})(\sqrt{u_m}-\sqrt{u_1})$$

$$\geq \frac{r_k(\bar{\beta}-\bar{\beta}_1)}{\sqrt{u_2u_1u_m}}(\sqrt{u_m}-\sqrt{u_1})(\sqrt{u_2}-\sqrt{u_m})$$

$$\geq \frac{r_k(\bar{\beta}-\bar{\beta}_1)}{u_2u_m\sqrt{u_1}}(u_m-u_1)(u_2-u_m)$$

$$\geq \frac{r_k(\bar{\beta}-\bar{\beta}_1)}{2u_2u_m\sqrt{u_1}}(u_2-u_1)\min(u_3-u_1,u_2-u_4)\geq \frac{(\bar{\beta}-\bar{\beta}_1)H^{1/2}}{2c_0^{1/2}}.$$

It follows that with probability tending to 1, $\Delta_{1,Jp}(\hat{k}_1)<\Delta_{1,Jp}(r_k)$ which yields the contradiction.

Hence, with probability tending to 1, $\hat{k}_1=r_k$ for some k. Inductively, for $\ell=2,...,H-1$, with probability tending to 1, there exists k such that $\hat{k}_\ell = r_k$. Next we show that with probability tending to 1, $\hat{H} = H$. For any $0 \le k_1 < k_2 \le H$, if $k_1 + 1 < k_2$, then with probability tending to 1, there exists $k_1 < \tilde{k} < k_2$ and

$$\Delta_{r_{k_1}+1, r_{k_2}}(r_{\tilde{k}}) = \max_{r_{k_1}+1 \le \kappa \le r_{k_2}} \Delta_{r_{k_1}+1, r_{k_2}}(\kappa) \ge \frac{\sqrt{Jp}}{H^2 K^2}[\frac{1}{K} + o_P(\frac{1}{\sqrt{Jp}})].$$

By Condition 2 (iv), the change point $r_{\tilde{k}}$ will be detected by the algorithm. If $k_2 = k_1 + 1$, then

$$\max_{r_{k_1}+1 \le \kappa \le r_{k_2}} \Delta_{r_{k_1}+1, r_{k_2}}(\kappa) = \frac{\sqrt{Jp}}{H^2 K^2} o_P(\frac{1}{\sqrt{Jp}}) = o_P(1).$$

Again by Condition 2 (iv), with probability tending to 1, the algorithm will stop. This completes the proof.

Proof of Theorem 1. Let $\Psi_j = (\psi_{ik,j})$ be a $p \times H$ matrix, where $\psi_{ik,j} = I(\beta_{i,j} \in \mathcal{B}_k)$, $1 \le i \le p, 1 \le k \le H, 1 \le j \le J$. By Lemma 3, with probability tending to 1, $\hat{\boldsymbol{\xi}} = (\hat{\xi}_1, ..., \hat{\xi}_H)^T$ maximizes

$$\mathcal{L}(\boldsymbol{\xi}) = \sum_{j=1}^{J} \mathcal{L}_j(\Psi_j \boldsymbol{\xi}),$$

and $\hat{\boldsymbol{\beta}}_j = \Psi_j \hat{\boldsymbol{\xi}}$. Let $\boldsymbol{\xi}^*$ denote the true parameter value of $\boldsymbol{\xi}$ and $\boldsymbol{\beta}_j = \Psi_j \boldsymbol{\xi}^*$. For any s in a compact set of $\mathcal{R}^H$, define

$$G(s) = \mathcal{L}(\boldsymbol{\xi}^* + s/\sqrt{\mathcal{N}}) - \mathcal{L}(\boldsymbol{\xi}^*)$$

$$= \sum_{j=1}^{J} \sum_{i=1}^{n_j} \int_0^\tau \{\mathcal{N}^{-1/2} X_{ij}^T s - \log S_j^{(0)}(u, \Psi_j \boldsymbol{\xi}^* + \mathcal{N}^{-1/2} \Psi_j s) + \log S_j^{(0)}(u, \Psi_j \boldsymbol{\xi}^*)\} dN_{ij}(u).$$

Write

$$\log S_j^{(0)}(u, \Psi_j \boldsymbol{\xi}^* + \mathcal{N}^{-1/2} \Psi_j s) - \log S_j^{(0)}(u, \Psi_j \boldsymbol{\xi}^*) =$$

$$\mathcal{N}^{-1/2} s^T \Psi_j^T E_j(u, \Psi_j \boldsymbol{\xi}^*) + \frac{1}{2\mathcal{N}} s^T \Psi_j^T V_j(u, \Psi_j \boldsymbol{\xi}^*) \Psi_j s + v_j(s, u).$$

Some analysis reveals that $|v_j(s, u)| \le \frac{4}{3} \mathcal{N}^{-3/2} \max_{1 \le i \le n_j} |s^T \Psi_j^T (X_{ij} - E_j(u, \Psi_j \boldsymbol{\xi}^*))|^3 = O(\mathcal{N}^{-3/2})$. It follows that

$$G_n(s) = U_n^T s - \frac{1}{2} s^T \mathcal{I}_n^* s - r_n(s),$$

where

$$U_n = \mathcal{N}^{-1/2} \sum_{j=1}^{J} \sum_{i=1}^{n_j} \int_0^\tau \Psi_j^T \{X_{ij} - E_j(u, \Psi_j \boldsymbol{\xi}^*)\} dN_{ij}(u),$$

$$\mathcal{I}_n^* = \mathcal{N}^{-1}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\int_0^\tau \Psi_j^T V_j(u,\Psi_j\boldsymbol{\xi}^*)\Psi_j dN_{ij}(u),$$

and

$$r_n(s) = \sum_{j=1}^{J}\sum_{i=1}^{n_j}\int_0^\tau v_j(s,u)dN_{ij}(u).$$

Note that

$$U_n = \mathcal{N}^{-1/2}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\int_0^\tau \Psi_j^T\{X_{ij} - E_j(u,\Psi_j\boldsymbol{\xi}^*)\}dM_{ij}(u)$$

and

$$\mathcal{I}_n^* = \mathcal{N}^{-1}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\int_0^\tau Y_{ij}\Psi_j^T V_j(u,\Psi_j\boldsymbol{\xi}^*)\Psi_j \exp{(X_{ij}^T\Psi_j^T\boldsymbol{\xi}^* + \lambda^T W_j)}h_0(u)du$$

$$+\mathcal{N}^{-1}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\int_0^\tau \Psi_j^T V_j(u,\Psi_j\boldsymbol{\xi}^*)\Psi_j dM_{ij}(u).$$

By the boundedness of the integrand, the first term of $\mathcal{I}_n^*$ goes to $\mathcal{I}$ in probability. Since

$$E\{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\int_0^\tau \Psi_j^T V_j(u,\Psi_j\boldsymbol{\xi}^*)\Psi_j dM_{ij}(u)\}^2 = E\sum_{j=1}^{J}\sum_{i=1}^{n_j}\{\Psi_j^T V_j(u,\Psi_j\boldsymbol{\xi}^*)\Psi_j\}^{\otimes 2}d < M_{i,j}, M_{i,j} > (u),$$

by the boundedness of the covariates, the second term of $\mathcal{I}_n^*$ is $O_p(\mathcal{N}^{-1/2})$. For any $\epsilon > 0$, let

$$U_n^\epsilon(t) = \mathcal{N}^{-1/2}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\int_0^t \Psi_j^T\{X_{ij} - E_j(u,\Psi_j\boldsymbol{\xi}^*)\}I\{\mathcal{N}^{-1/2}|\Psi_j^T(X_{ij} - E_j(u,\Psi_j\boldsymbol{\xi}^*)| \geq \epsilon\}dM_{ij}(u).$$

Because the predictable variation/covariation processes of $U_n(\cdot)$ and $U_n^\epsilon(\cdot)$

$$< U_n, U_n > (t) = \mathcal{N}^{-1}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\int_0^t \Psi_j^T\{X_{ij} - E_j(u,\Psi_j\boldsymbol{\xi}^*))\}^{\otimes 2}\Psi_j d < M_{ij}, M_{ij} > (u)$$

$$= \int_0^s \mathcal{I}_n(u)h_0(u)du \xrightarrow{p} \int_0^s \mathcal{I}(u)h_0(u)du,$$

and

$$< U_n^\epsilon, U_n^\epsilon > (t) = \mathcal{N}^{-1}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\int_0^t Y_{ij}\Psi_j^T\{X_{ij} - E_j(u,\Psi_j\boldsymbol{\xi}^*)\}^{\otimes 2}\Psi_j$$

$$I\{\mathcal{N}^{-1/2}|\Psi_j^T(X_{ij} - E_j(u,\Psi_j\boldsymbol{\xi}^*)| \geq \epsilon\} \exp{(X_{ij}^T\Psi_j^T\boldsymbol{\xi}^* + \lambda^T W_j)}h_0(u)du \xrightarrow{p} 0,$$

by the martingale central limit theorem (Fleming and Harrington (2011)), $U_n \to U \sim N_H(0_H, \mathcal{I})$. To summarize, for any s in a compact set S of $\mathcal{R}^H$, we have

$$G_n(s) \xrightarrow{p} G(s) = U^T s - \frac{1}{2} s^T \mathcal{I} s.$$

By the concavity of $G_n(s)$ and $G(s)$ and the concavity lemma (Page 1116, Andersen and Gill, 1982), it follows that

$$\sup_{s \in S} |G_n(s) - G(s)| \xrightarrow{p} 0.$$

We now prove that $\mathcal{N}^{1/2}(\hat{\boldsymbol{\xi}} - \boldsymbol{\xi}^*) \to N(0_H, \mathcal{I}^{-1})$. Let $\gamma_n = \mathcal{N}^{1/2}(\hat{\boldsymbol{\xi}} - \boldsymbol{\xi}^*)$ and $\alpha_n = \mathcal{I}^{-1}U$ which maximize $G_n(s)$ and $G(s)$ respectively. It is sufficient to show that for any $\delta > 0$,

$$P(|\gamma_n - \alpha_n| > \delta) \to 0.$$

For any $|s - \alpha_n| = \delta$, $G(\alpha_n) - G(s) \geq \frac{1}{2}\delta^2 \lambda_{\text{min}}(\mathcal{I})$. If $|\gamma_n - \alpha_n| = l > \delta$, then let $\eta_n = \alpha_n + (\gamma_n - \alpha_n)\frac{\delta}{l}$. By the concavity of G_n,

$$G_n(\eta_n) \geq (1 - \frac{\delta}{l})G_n(\alpha_n) + \frac{\delta}{l} G_n(\gamma_n).$$

Hence,

$$0 \leq [G_n(\gamma_n) - G_n(\alpha_n) \leq \frac{l}{\delta}[2 \sup_{|s-\alpha_n|=\delta} |G_n(s) - G(s)| - \frac{1}{2}\delta^2 \lambda_{\text{min}}(\mathcal{I})].$$

It follows that

$$P(|\gamma_n - \alpha_n| > \delta) \leq P(\sup_{|s-\alpha_n|=\delta} |G_n(s) - G(s)| \geq \frac{1}{4}\delta^2 \lambda_{\text{min}}(\mathcal{I})) \to 0.$$

Note that $\boldsymbol{\beta}_j = \Psi_j \boldsymbol{\xi}^*$ and consequently, for any $1 \leq j \leq J$, we have $\mathcal{N}^{-1/2}(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*) \to N(0_p, \Psi_j \mathcal{I}^{-1} \Psi_j^T)$. This completes the proof of Theorem 1.

Proof of Theorem 2.

Let $y_{\ell j} = \hat{L}_{1,j}(t_{(\ell),j})$, $\alpha_{\ell j} = L_0(t_{(\ell),j})$. Hence (3.4) can be written as

$$y_{\ell j} = \alpha_{\ell j} + W_j^T \lambda + \epsilon_{\ell j},$$

where $W_j = (W_{j11,\ldots,jqc_q})^T$ and $\lambda = (\lambda_{11}, \ldots, \lambda_{qc_q})^T$. Let $\Gamma_j = \mathbf{1}_{T_j} \otimes W_j^T$, where $\mathbf{1}_d$ is a d dimensional vector with each component being 1. Write $y_j = (y_{1j}, \ldots, y_{T_j j})^T$ and $\epsilon_j = (\epsilon_{1j}, \ldots, \epsilon_{T_j j})^T$. We have

$$y_j = \alpha_j + \Gamma_j \lambda + \epsilon_j.$$

Write $y = (y_1^T, ..., y_J^T)^T$, $\alpha = (\alpha_1^T, ..., \alpha_J^T)^T$, $\Gamma = (\Gamma_1^T, ..., \Gamma_J^T)^T$, and $\epsilon = (\epsilon_1^T, ..., \epsilon_J^T)^T$. Hence,

$$y = \alpha + \Gamma\lambda + \epsilon.$$

Let $c = c_1 + ... + c_q$, $\mathcal{W}(t) = \text{diag}(K_h(t_{(1),1} - t), ..., K_h(t_{(T_J),J} - t))$, $u(t) = (t_{(1),1} - t), ..., t_{(T_J),J} - t)^T$, and $X(t) = (\mathbf{1_N}, \mathbf{\Gamma}, \mathbf{u(t)})$. We estimate $L_0(t), L_0'(t) = \frac{h_0(t)}{\Lambda_0(t)}$, and λ by $\tilde{a}(t), \tilde{b}(t)$, and $\tilde{\lambda}(t)$ which minimize

$$\mathcal{M}_t(a, \lambda, b) = (y - a1_N - \Gamma\lambda - bu(t))^T \mathcal{W}(t)(y - a1_N - \Gamma\lambda - bu(t)).$$

It follows that

$$\tilde{\lambda}(t) = (0_{c\times 1}, I_c, 0_{c\times 1})(X(t)^T \mathcal{W}(t) X(t))^{-1} X(t)^T \mathcal{W}(t) y,$$

$$\tilde{a}(t) = (1, 0_{c+1}^T)(X(t)^T \mathcal{W}(t) X(t))^{-1} X(t)^T \mathcal{W}(t) y,$$

and

$$\tilde{b}(t) = (0_{c+1}^T, 1)(X(t)^T \mathcal{W}(t) X(t))^{-1} X(t)^T \mathcal{W}(t) y.$$

Let $\alpha = (L_0(t_{(1),1}), ..., L_0(t_{(T_J),J})^T$. Thus,

$$\begin{aligned}\tilde{\lambda}(t) - \lambda &= (0_{c\times 1}, I_c, 0_{c\times 1})(X(t)^T \mathcal{W}(t) X(t))^{-1} X(t)^T \mathcal{W}(t)\epsilon \\ &+ (0_{c\times 1}, I_c, 0_{c\times 1})(X(t)^T \mathcal{W}(t) X(t))^{-1} X(t)^T \mathcal{W}(t)[\alpha - L_0(t)1_{\mathcal{N}} - L_0'(t)u(t)].\end{aligned}$$

To derive the limiting distribution of $\tilde{\lambda}(t) - \lambda$, we first show that the first term is asymptotically normal with mean zero and then evaluate the magnitude of the second term which yields the bias. Let $W_0 = 1$, $\tilde{W}_j = (W_0, W_j^T)^T$. For $m = 0, 1, 2$ and $\ell = 0, 1, ..., c$, let $\mu_m = \int u^m K(u) du$. Write

$$\Omega = \begin{pmatrix} \Omega_{11} & \Omega_{21}^T \\ \Omega_{21} & \Omega_{22} \end{pmatrix},$$

where Ω_{11}, Ω_{21}, and Ω_{22} are 1×1, $c \times 1$, and $c \times c$ matrices, respectively. Let $\Omega_{1\ell}$ denote the ℓth element in the first row of Ω. Define

$$S_{\ell,m}^n(t) = \frac{1}{\mathcal{N}} \sum_{j=1}^{J} \sum_{i=1}^{n_j} \delta_{ij}(h^{-1}(t_{ij} - t))^m K_h(t_{ij} - t)\tilde{W}_{j,\ell}.$$

Then,

$$ES^n_{\ell,m}(t) = \frac{\mu_m}{\mathcal{N}} \sum_{j=1}^{J} \sum_{i=1}^{n_j} f_{ij}(t)\bar{G}_{ij}(t)\tilde{W}_{j,\ell}(1 + o_p(1)) = \mu_m \Omega_{1\ell}(t)(1 + o_p(1)).$$

As $K(\cdot)$ is a bounded function with a bounded support, for any $0 \leq m \leq 4$, $|u^m K(u)|$ is bounded. By the Jensen's inequality,

$$\mathrm{var}(S^n_{\ell,m}(t)) \leq \mathcal{N}^{-2} \sum_{j=1}^{J} n_j \sum_{i=1}^{n_j} [\delta_{ij}(h^{-1}(t_{ij} - t))^m K_h(t_{ij} - t)\tilde{W}_{j,\ell}]^2 = O((\mathcal{N}h^2)^{-1}).$$

It follows that

$$S^n_{\ell,m}(t) = ES^n_{\ell,m}(t) + O_p(\sqrt{\mathrm{var}(S^n_{\ell,m}(t))}) = \mu_m \Omega_{1\ell}(t)(1 + o_p(1)).$$

Hence,

$$\frac{1}{\mathcal{N}} X(t)^T \mathcal{W}(t) X(t) = H \left\{ \begin{array}{cc} \mu_0 \Omega(t) & 0 \\ 0 & \mu_2 \Omega_{11}(t) \end{array} \right\} H(1 + o_p(1)),$$

where $H = \left\{ \begin{array}{cc} I_{c+1} & 0 \\ 0 & h \end{array} \right\}$. Now we derive the limiting distribution of

$$\sum_{j=1}^{J} \sum_{i=1}^{n_j} \int_0^{\tau} W_j K_h(u - t)[\hat{L}_{1,j}(u) - L_{1,j}(u)] dN_{ij}(u),$$

$$\sum_{j=1}^{J} \sum_{i=1}^{n_j} \int_0^{\tau} (u - t) K_h(u - t)[\hat{L}_{1,j}(u) - L_{1,j}(u)] dN_{ij}(u),$$

and

$$\sum_{j=1}^{J} \sum_{i=1}^{n_j} \int_0^{\tau} K_h(u - t)[\hat{L}_{1,j}(u) - L_{1,j}(u)] dN_{ij}(u).$$

Let

$$\hat{\Lambda}_{1,j}(t) = \int_0^t \sum_{i=1}^{n_j} \frac{1}{n_j S_j^{(0)}(u, \Psi_j^T \hat{\xi})} dN_{ij}(u),$$

$$\tilde{\Lambda}_{1,j}(t) = \int_0^t \{\sum_{i=1}^{n_j} Y_{ij}(u) e^{X_{ij}^T \Psi_j^T \xi^*}\}^{-1} d[\sum_{i=1}^{n_j} N_{ij}(u)],$$

and

$$\Lambda^*_{1,j}(t) = \int_0^t I(\sum_{i=1}^{n_j} Y_{ij}(u) > 0) h_{1,j}(u) du.$$

We have

$$n_j^{1/2}(\hat{\Lambda}_{1,j}(t)-\Lambda_{1,j}(t)) = n_j^{1/2}(\hat{\Lambda}_{1,j}(t)-\tilde{\Lambda}_{1,j}(t))+n_j^{1/2}(\tilde{\Lambda}_{1,j}(t)-\Lambda^*_{1,j}(t))+n_j^{1/2}(\Lambda^*_{1,j}(t)-\Lambda_{1,j}(t)).$$

It is easy to see that $n_j^{1/2}(\Lambda^*_{1,j}(t)-\Lambda_{1,j}(t))$ is asymptotically negligible, $n_j^{1/2}(\tilde{\Lambda}_{1,j}(t)-\Lambda^*_{1,j}(t))$ converges to a mean zero and incremental Gaussian process with variance function

$$\int_0^t [s_j^{(0)}(u,\Psi_j\boldsymbol{\xi}^*)]^{-1}h_{1,j}(u)du,$$

and

$$n_j^{1/2}(\hat{\Lambda}_{1,j}(t)-\tilde{\Lambda}_{1,j}(t)) = [-\int_0^t n_j^{-1}\frac{S_j^{(1)}(u,\Psi_j\boldsymbol{\xi}^*)}{\{S_j^{(0)}(u,\Psi_j\boldsymbol{\xi}^*)\}^2}d\sum_{i=1}^{n_j}N_{ij}(u)+o_p(1)][n_j^{1/2}\Psi_j^T(\hat{\boldsymbol{\xi}}-\boldsymbol{\xi}^*)]$$

$$= [-\int_0^t e_j(u,\Psi_j\boldsymbol{\xi}^*)h_{1,j}(u)du+o_p(1)][n_j^{1/2}\Psi_j^T(\hat{\boldsymbol{\xi}}-\boldsymbol{\xi}^*)].$$

Note that

$$\mathcal{N}^{1/2}(\hat{\boldsymbol{\xi}}-\boldsymbol{\xi}^*) = \mathcal{I}^{-1}\mathcal{N}^{-1/2}\sum_{j=1}^J\sum_{i=1}^{n_j}\int_0^\tau \Psi_j^T\{X_{ij}-E_j(u,\Psi_j\boldsymbol{\xi}^*)\}dM_{ij}(u)+o_p(1),$$

and for any $1\le j\le J$,

$$n_j^{1/2}(\tilde{\Lambda}_{1,j}(t)-\Lambda^*_{1,j}(t)) = n_j^{-1/2}\sum_{i=1}^{n_j}\int_0^t\{S_j^{(0)}(u,\Psi_j\boldsymbol{\xi}^*)\}^{-1}dM_{ij}(u)+o_p(1).$$

Because

$$< \sum_{j=1}^J\sum_{i=1}^{n_j}\int_0^\tau \Psi_j^T\{X_{ij}-E_j(u,\Psi_j\boldsymbol{\xi}^*)\}dM_{ij}(u), \sum_{i=1}^{n_j}\int_0^t\{S_j^{(0)}(u,\Psi_j\boldsymbol{\xi}^*)\}^{-1}dM_{ij}(u) > (t)$$

$$= \sum_{i=1}^{n_j}\int_0^t \Psi_j^T\{X_{ij}-E_j(u,\Psi_j\boldsymbol{\xi}^*)\}\{S_j^{(0)}(u,\Psi_j\boldsymbol{\xi}^*)\}^{-1}Y_{ij}(u)\exp(X_{ij}^T\Psi_j\boldsymbol{\xi}^*)h_{1,j}(u)du$$

$$= 0,$$

it follows that $n_j^{1/2}(\tilde{\Lambda}_{1,j}(t)-\Lambda^*_{1,j}(t)), j=1,...,J$ and $(\hat{\boldsymbol{\xi}}-\boldsymbol{\xi}^*)$ are asymptotically independent. Furthermore,

$$\sum_{j=1}^J\sum_{i=1}^{n_j}\int_0^\tau K_h(u-t)[\hat{L}_{1,j}(u)-L_{1,j}(u)]dN_{ij}(u)$$

$$= \sum_{j=1}^{J}\sum_{i=1}^{n_j}\int_0^{\tau} Y_{ij}(u)\exp(X_{ij}^T\Psi_j\boldsymbol{\xi}^*)K_h(u-t)[\hat{L}_{1,j}(u)-L_{1,j}(u)]h_{1,j}(u)du$$

$$+\sum_{j=1}^{J}\sum_{i=1}^{n_j}\int_0^{\tau} K_h(u-t)[\hat{L}_{1,j}(u)-L_{1,j}(u)]dM_{ij}(u)$$

$$= \{\sum_{j=1}^{J}\int_0^{\tau} n_j[\hat{L}_{1,j}(u)-L_{1,j}(u)]K_h(u-t)h_{1,j}(u)s_j^{(0)}(u,\Psi_j\boldsymbol{\xi}^*)du\}(1+o_p(1))$$

$$= \{\sum_{j=1}^{J}\int_0^{\tau} n_j[\hat{\Lambda}_{1,j}(u)-\Lambda_{1,j}(u)]K_h(u-t)s_j^{(0)}(u,\Psi_j\boldsymbol{\xi}^*)du\}(1+o_p(1)).$$

By calculation,

$$\begin{aligned}
&\mathrm{Var}(\sum_{j=1}^{J}\int_0^{\tau} n_j[\hat{\Lambda}_{1,j}(u)-\Lambda_{1,j}(u)]K_h(u-t)s_j^{(0)}(u,\Psi_j\boldsymbol{\xi}^*)du)\\
= \;&\{\sum_{j=1}^{J} n_j[s_j^{(0)}(t,\Psi_j\boldsymbol{\xi}^*)]^2\int_0^t [s_j^{(0)}(u,\Psi_j\boldsymbol{\xi}^*)]^{-1}h_{1,j}(u)du+\\
&\mathcal{N}^{-1}(\sum_{j=1}^{J} n_j s_j^{(0)}(t,\Psi_j\boldsymbol{\xi}^*)\int_0^t e_j^T(u,\Psi_j\boldsymbol{\xi}^*)h_{1,j}(u)du\Psi_j^T)\mathcal{I}^{-1}(\sum_{j=1}^{J} n_j s_j^{(0)}(t,\Psi_j\boldsymbol{\xi}^*)\times\\
&\Psi_j\int_0^t e_j(u,\Psi_j\boldsymbol{\xi}^*)h_{1,j}(u)du)\}(1+o_p(1)).
\end{aligned}$$

By Lindeberg-Feller theorem, Slutsky's theorem and similar calculations as above, we have

$$\mathcal{N}^{-1/2}(I_{c+1},0_{c+1})H^{-1}X^T\mathcal{W}(t)\epsilon$$

is asymptotically normal with the covariance function

$$\zeta(t,t)+\Upsilon(t)\mathcal{I}^{-1}\Upsilon(t)^T,$$

and hence

$$\mathcal{N}^{1/2}(I_{c+1},0_{c+1})(X(t)^T\mathcal{W}(t)X(t))^{-1}X(t)^T\mathcal{W}(t)\epsilon$$

is asymptotically normal with mean zero with the covariance function

$$\Omega(t)^{-1}[\zeta(t,t)+\Upsilon(t)\mathcal{I}^{-1}\Upsilon(t)^T]\Omega(t)^{-1}.$$

Define

$$\nu(t_1,t_2)=\Omega(t_1)^{-1}[\zeta(t_1,t_2)+\Upsilon(t_1)\mathcal{I}^{-1}\Upsilon(t_2)^T]\Omega(t_2)^{-1},$$

and write

$$\nu(t_1,t_2) = \begin{pmatrix} \nu_{11}(t_1,t_2) & \nu_{12}(t_1,t_2) \\ \nu_{12}^T(t_1,t_2) & \nu_{22}(t_1,t_2) \end{pmatrix},$$

where $\nu_{11}(t_1,t_2), \nu_{12}(t_1,t_2)$, and $\nu_{22}(t_1,t_2)$ are 1×1, $1\times c$, and $c\times c$ matrices, respectively.

Next, we evaluate the bias term

$$\begin{aligned}
&\mathcal{N}^{1/2}(I_{c+1},0_{c+1})(X(t)^T\mathcal{W}(t)X(t))^{-1}X(t)^T\mathcal{W}(t)[\alpha - L_0(t)1_N - L_0^{'}(t)u(t)] \\
&= 2^{-1}L_0^{''}(t)\mathcal{N}^{1/2}(I_{c+1},0_{(c+1)\times 1})(X(t)^T\mathcal{W}(t)X(t))^{-1}X(t)^T\mathcal{W}(t)u^2(t)[1+o_p(1)] \\
&= \mathcal{N}^{1/2}2^{-1}L_0^{''}(t)\mu_2 h^2\Omega^{-1}(t)\begin{pmatrix} \Omega_{11}(t) \\ \Omega_{21}(t) \end{pmatrix}[1+o_p(1)].
\end{aligned}$$

When $\mathcal{N}h^4$ is bounded, we have

$$\sqrt{\mathcal{N}}\{(\tilde{a}(t),\tilde{\lambda}^T)^T - (L_0(t),\lambda^T)^T - \frac{h^2\mu_2 L_0^{''}(t)}{2}(1,0_c^T)^T\} \to N(0_{(c+1)\times 1},\nu(t,t)).$$

Now we proceed to obtain the limiting distribution of $\hat{\lambda}$. By definition,

$$\hat{\lambda} = \frac{1}{N}\sum_{m=1}^{N}\tilde{\lambda}(t_{(m)}) = \frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{T_j}(0_{c\times 1},I_c,0_{c\times 1})(X^T(t_{(i),j})\mathcal{W}(t_{(i),j})X(t_{(i),j}))^{-1}X^T(t_{(i),j})\mathcal{W}(t_{(i),j})y.$$

By straightforward but tedious calculation and the law of large numbers, we have

$$\begin{aligned}
&\mathrm{Var}(N\hat{\lambda}) \\
&= E\sum_{j=1}^{J}\sum_{i=1}^{T_j}\sum_{r=1}^{J}\sum_{\ell=1}^{T_r}(0_{c\times 1},I_c,0_{c\times 1})(X^T(t_{(i),j})\mathcal{W}(t_{(i),j})X(t_{(i),j}))^{-1}X^T(t_{(i),j})\mathcal{W}(t_{(i),j})\epsilon \\
&\quad \times\epsilon^T\mathcal{W}(t_{(\ell),r})X(t_{(\ell),r})(X^T(t_{(\ell),r})\mathcal{W}(t_{(\ell),r})X(t_{(\ell),r}))^{-1}(0_{c\times 1},I_c,0_{c\times 1})^T \\
&= (1+o(1))E\mathcal{N}^{-1}\sum_{j=1}^{J}\sum_{i=1}^{T_j}\sum_{r=1}^{J}\sum_{\ell=1}^{T_r}\nu_{22}(t_{(i),j},t_{(\ell),r}) \\
&= (1+o(1))E\mathcal{N}^{-1}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{r=1}^{J}\sum_{\ell=1}^{n_r}\delta_{ij}\delta_{\ell r}\nu_{22}(t_{ij},t_{\ell r}),
\end{aligned}$$

and

$$\mathrm{Bias}(\hat{\lambda}) = E\hat{\lambda} - \lambda$$

$$= (1 + o_p(1))E \frac{1}{N} \sum_{j=1}^{J} \sum_{i=1}^{n_j} \frac{h^2 \mu_2 \delta_{ij} L_0^{''}(t_{ij})}{2}.$$

Define

$$\bar{\Omega}_{11} = \int_0^\tau \Omega_{11}(u) du,$$

$$\bar{\nu}_{22} = \int_0^\tau \int_0^\tau \nu_{22}(u, v) \Omega_{11}(u) \Omega_{11}(v) du dv,$$

and

$$\Phi = \int_0^\tau L_0^{''}(u) \Omega_{11}(u) du.$$

By the law of large numbers,

$$\mathcal{N}^{-1} \sum_{j=1}^{J} \sum_{i=1}^{n_j} \delta_{ij} \to \bar{\Omega}_{11},$$

$$\mathcal{N}^{-2} \sum_{j=1}^{J} \sum_{i=1}^{n_j} \sum_{r=1}^{J} \sum_{\ell=1}^{n_r} \delta_{ij} \delta_{\ell r} \nu_{22}(t_{ij}, t_{\ell r}) \to \bar{\nu}_{22},$$

and

$$\mathcal{N}^{-1} \sum_{j=1}^{J} \sum_{i=1}^{n_j} \delta_{ij} L_0^{''}(t_{ij}) \to \Phi.$$

It follows that when $\mathcal{N}h^4$ is bounded,

$$\sqrt{\mathcal{N}}[\hat{\lambda} - \lambda - \frac{\Phi h^2 \mu_2}{2\bar{\Omega}_{11}}(1, 0_c^T)^T] \to N(0_c, \bar{\Omega}_{11}^{-2} \bar{\nu}_{22}).$$

Hence, when $\mathcal{N}h^4 \to 0$,

$$\sqrt{\mathcal{N}}(\hat{\lambda} - \lambda) \to N(0_c, \bar{\Omega}_{11}^{-2} \bar{\nu}_{22}).$$

Let $e_{k,\ell}$ denote the unit vector of length c with 1 at position corresponding to $\lambda_{k,\ell}$. Thus, for any $k = 1, ...c_q; \ell = 1, ..., q$,

$$\sqrt{\mathcal{N}}(\hat{\lambda}_{k,\ell} - \lambda_{k,\ell}) \to N(0, e_{k,\ell}^T \bar{\Omega}_{11}^{-2} \bar{\nu}_{22} e_{k,\ell}).$$

Proof of Theorem 3.

We establish the limiting distribution for $\hat{\Lambda}_0(t)$. Write

$$\Xi_1(t) = \frac{1}{N} \sum_{j=1}^{J} \sum_{i=1}^{n_j} \delta_{ij} h^{-1}(t_{ij} - t) K_h(t_{ij} - t) \exp\{(1_{1\times1}, 0_{(c+1)\times1})(X^T(t_{ij})\mathcal{W}(t_{ij})X(t_{ij}))^{-1} X^T(t_{ij})\mathcal{W}(t_{ij})y\},$$

$$\Xi_0(t)=\frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\delta_{ij}K_h(t_{ij}-t)\exp\{(1_{1\times 1},0_{(c+1)\times 1})(X^T(t_{ij})\mathcal{W}(t_{ij})X(t_{ij}))^{-1}X^T(t_{ij})\mathcal{W}(t_{ij})y\},$$

and for $\ell=0,1,2$,

$$\Theta_\ell(t)=\frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\delta_{ij}h^{-\ell}(t_{ij}-t)^\ell K_h(t_{ij}-t).$$

By the law of large numbers and Condition 5 (iv),

$$\Theta_\ell(t)=\mu_\ell(1+o_p(1)).$$

Thus,

$$\hat{\Lambda}_0(t)=(1,0)\begin{pmatrix}\Theta_0(t) & \Theta_1(t)\\ \Theta_1(t) & \Theta_2(t)\end{pmatrix}^{-1}\begin{pmatrix}\Xi_0(t)\\ \Xi_1(t)\end{pmatrix}=\Xi_0(t)(1+o_p(1)).$$

By similar arguments as in the Proof of Theorem 2, it can be shown that

$$\begin{aligned}&\mathrm{Var}(\mathcal{N}^{1/2}\Xi_0(t))\\ =\;&(1+o(1))EN^{-2}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{r=1}^{J}\sum_{\ell=1}^{n_r}\delta_{ij}\delta_{r\ell}K_h(t_{ij}-t)K_h(t_{r\ell}-t)\nu_{11}(t_{ij},t_{r\ell})\Lambda_0(t_{ij})\Lambda_0(t_{r\ell})\\ =\;&(1+o(1))\Lambda_0^2(t)\nu_{11}(t,t).\end{aligned}$$

Write

$$\mathrm{Bias}(\Xi_0(t))=E[\Xi_0(t)]-\Lambda_0(t)=\mathrm{Bias}_1+\mathrm{Bias}_2,$$

where

$$\mathrm{Bias}_1=E[\Xi_0(t)]-E\sum_{j=1}^{J}\sum_{i=1}^{n_j}N^{-1}\delta_{ij}K_h(t_{ij}-t)\Lambda_0(t_{ij}),$$

and

$$\mathrm{Bias}_2=E\sum_{j=1}^{J}\sum_{i=1}^{n_j}N^{-1}\delta_{ij}K_h(t_{ij}-t)\Lambda_0(t_{ij})-\Lambda_0(t).$$

By Theorem 2,

$$\mathrm{Bias}_1=\frac{\mu_2h^2}{2}\Lambda_0(t)L_0^{''}(t),$$

and it is easy to see that,

$$\mathrm{Bias}_2=\frac{\mu_2h^2}{2}h_0^{'}(t).$$

Hence,

$$\mathrm{Bias}(\Xi_0(t))=\frac{\mu_2h^2}{2}[\Lambda_0(t)L_0^{''}(t)+h_0^{'}(t)](1+o(1)).$$

It follows that when $\mathcal{N}h^4 \to 0$,

$$\mathcal{N}^{1/2}(\hat{\Lambda}_0(t) - \Lambda_0(t)) \to N(0, \Lambda_0^2(t)\nu_{11}(t,t)).$$