*The Canadian Journal of Statistics*
Vol. xx, No. yy, 2020, Pages 1–23
*La revue canadienne de statistique*



# Semiparametric additive frailty hazard model for clustered failure time data

Peng Liu[1], Shanshan Song[2] and Yong Zhou[3]*

[1]*School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, United Kingdom*
[2]*School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, China*
[3]*Academy of Statistics and Interdisciplinary Sciences, East China Normal University, Shanghai, China*

*Key words and phrases:* Additive frailty hazard model; clustered failure time data; local linear technique ; semiparametric model.

*MSC 2010:* Primary 62N01; secondary 62N02

*Abstract:* In this paper, we propose a flexible semiparametric additive frailty hazard model under clustered failure time data, where frailty is assumed to have an additive effect on the hazard function. When there is no frailty, this model degenerates to semiparametric additive hazard model. Our method can deal with time-varying covariate effect and constant covariate effect simultaneously and the estimate of the covariate effects will not rely on the frailty distribution. The time-varying coefficient is estimated by utilizing the local linear technique, while $\sqrt{n}$-consistency convergence rate of constant coefficient estimate can be obtained by integration. Another advantage of the estimator is that it has a closed-form and can be easily implemented in practice. The large sample properties of the estimator have been established and simulation studies under various scenarios are conducted to demonstrate the performance of the proposed methods. A real data is applied for illustration purpose. *The Canadian Journal of Statistics* xx: 1–23; 2020 © 2020 Statistical Society of Canada

*Résumé:* Insérer votre résumé ici. We will supply a French abstract for those authors who can't prepare it themselves. *La revue canadienne de statistique* xx: 1–23; 2020 © 2020 Société statistique du Canada

## 1. INTRODUCTION

The clustered failure time data is frequently encountered in biomedical and clinical studies recently. A common feature of this kind of data is that they are often correlated within the same cluster, for example, in the well known Diabetic Retinopathy Study, which begun in 1971 and aimed to study whether laser photocoagulation was effective in delay the onset of blindness in patients with diabetic retinopathy, the failure times for the two eyes of a patient were proved to be strong positive associated (Huster, Brookmeyer, & Self, 1989). In addition, when study the survival time of those who suffer from heart disease, individuals from the same family are also correlated with each other.

Several attempts of modelling have been made for this type of data on regression analysis, one straightforward approach is to build a marginal model, which ignores the dependence structure among individuals within a cluster, for example the marginal proportional hazard model (Lin, 1994; Spiekerman & Lin, 1998; Lu & Wang, 2005; Cai et al., 2007), as well as the marginal additive hazard model (Yin & Cai, 2004; Martinussen & Scheike, 2007). Marginal approach can model hazard function directly, is robust to dependence structure, and is suitable when we only care the population level covariate effects. However, it cannot address the joint quantities, and the assumed 'working independence' can not reflect the true underlying correlation structure.

* *Author to whom correspondence may be addressed.*
*E-mail: yzhou@amss.ac.cn*

Some techniques have been proposed to integrate the correlation structure, for example, Li & Yin (2009) developed a generalized method of moment approach to handle marginal accelerate failure time model, while took the dependence structure into consideration by incorporating an unknown weight matrix. A more general and intuitive way to introduce the intra-cluster dependence is the frailty, which represents an unobserved random effect that could lead to hidden heterogeneity of the mortality in a study population (Vaupel, Manton, & Stallard, 1979). In clustered failure time data, the unobserved random effect in frailty model explains the individual variations. One popular example is the gamma-frailty Cox model, where the frailty item is included in the intensity of Cox regression, and is assumed to follow gamma distribution with an unknown parameter (Murphy, 1994, 1995; Parner, 1998; Nielsen et al., 1992; Fan & Li, 2002).

It is well known that the Cox model focuses on modelling the ratio of the intensities, however, the multiplicative structure may not model data well in practical situations. In addition, the multiplicative frailty term in the Cox model will make the estimating procedure and asymptotic theory hard to obtain. The additive model provides an alternative to the Cox model, due to the reason that additive effect instead of multiplicative could be a more reasonable association in some applications (Aalen, 1989). Cai & Zeng (2011) extended the additive hazard model to clustered failure time data by including an additive frailty term on the hazard, their model can be seen as an parallel development of the well known gamma-frailty model in additive model structures. Different from gamma-frailty model, their model still induces an marginal additive model so that the estimators for covariate effects are also straightforward by constructing estimation equations. However, in Cai & Zeng (2011)'s model, their covariate effect is constant, which may be violated in practice (Hastie & Tibshirani, 1993). Martinussen, Scheike, & Zucker (2011) proposed the Aalen time varying coefficient additive gamma frailty hazard model which allows non-constant covariate effect, however, the frailty acts multiplicative on the hazards. In additive mixed effect model, no similar results have been developed to incorporate the time-varying covariate effect, we intend to fill this gap.

In this paper, we propose a partial linear additive frailty hazard model, in which some of the covariate effect are time-varying and the rest are constant. The hazard function is defined as:

$$\lambda_{ij}(t|\boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}) = \lambda(t) + \boldsymbol{X}_{ij}^{\top}(t)\beta(t) + \boldsymbol{Z}_{ij}^{\top}\gamma + \xi_i, i = 1, \cdots, n, j = 1, \cdots, n_i, \qquad (1)$$

where $\xi_i$ is the additive frailty term. This model has three major advantages: First, it serves as an additive counterpart with the usual gamma-frailty Cox model; Second, the frailty term does not need to be gamma distribution, and more importantly, estimating procedure for covariate effects will not change while other frailty distributions are assumed, while for gamma-frailty Cox model, the estimating procedure need to be changed for non-gamma distribution (Nielsen et al., 1992); Third, it includes existing additive mixed effect model (Cai & Zeng, 2011) and is therefore more flexible.

One of the motivating example for model (1) is the western Kenya parasitemia study (McElroy et al., 1997). In this study, 542 children from 309 households were enrolled, and the failure time was defined as the time from an individual's enrollment to the infection of parasitemia. Observations from the same household were extremely likely to be correlated, due to the fact that they are exposed to similar environment and share similar gene, thus we can treat this data as clustered failure time data. Frailty can be introduced to incorporate those genetic and environmental factors within each family. The covariates included baseline parasitemia density (BPD) and exposure to mosquito bites (BITE), Age and Gender. Recent studies such as Yu & Lin (2010) indicated that the effect of BPD was diminishing over time, while those of BITE, Age and Gender were constant. In model (1), we utilize the local linear method (Li, Yin, & Zhou, 2007) to deal with the time-varying coefficient. The asymptotic consistency and normality of the proposed estimator have been established. Meanwhile, by the weight technique (Yin, Li, & Zeng, 2008),

we establish the $\sqrt{n}$-consistency of the weighted estimate of constant coefficients.

This paper is organized as follows. In Section 2, we introduce some notations and give the details of the inference procedure. The large sample properties are shown in Section 3 and simulation results are provided in Section 4. Section 5 applies the proposed method to the western Kenya parasitemia data. Section 6 contains the discussion and the paper closes by the Appendix which contains all technical proof details.

## 2. MODEL AND INFERENCE PROCEDURE

### 2.1. DATA AND MODEL

The data is observed from $n$ i.i.d. clusters, $n_i$ observations are contained in $i$th cluster, which lead to the total sample size being $\sum_{i=1}^{n} n_i$, where each $n_i$ do not need to be same. For $j$th individual in $i$th cluster, $T_{ij}^*$ is the failure time, $\boldsymbol{X}_{ij}(t)$ and $\boldsymbol{Z}_{ij}$ are associated covariate vectors of dimension equal $p \times 1$ and $q \times 1$. Here we use $\boldsymbol{X}_{ij}(t)$ instead of $\boldsymbol{X}_{ij}$ to emphasize the time-varying covariate effects, instead of time-dependent covariates, because the time-dependent covariates may be internal or externel which will end in completely different estimating procedures (Kalbfleisch & Prentice, 2002). The covariate effects are $\beta(t)$ and $\gamma$, which are $p \times 1$ and $q \times 1$ column vectors, respectively. The observed data are the i.i.d. quadruples $\{(T_{ij} = T_{ij}^* \wedge C_{ij}, \boldsymbol{X}_{ij}(t), \boldsymbol{Z}_{ij}, \Delta_{ij} = I(T_{ij}^* \leq C_{ij})), j = 1, \cdots, n_i,\ i = 1, \cdots, n\}$, where $C_{ij}$ is the censoring time. The study duration is denoted as $\tau$. The frailty term $\xi_i$ in model (1) is assumed to have a zero mean and a finite moment generating function, while the density function of $\xi_i$ is $f(\cdot; \theta_0)$, here $\theta_0$ is an unknown one-dimension parameter.

### 2.2. NOTATIONS AND INFERENCE PROCEDURE

The parameters of interest in model (1) are $(\boldsymbol{\beta}(t), \gamma)$, the baseline hazard function $\lambda(t)$ as well as $\theta_0$. To estimate them, a start point is the marginal survival function:

$$
\begin{aligned}
& P(T_{ij}^* > t | \boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}) \\
&= E_{\xi_i}\left[\exp\left\{-\Lambda(t) - \int_0^t \boldsymbol{X}_{ij}(u)^\top \boldsymbol{\beta}(u) du - \int_0^t \boldsymbol{Z}_{ij}^\top \gamma du - \xi_i t\right\} \middle| \boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}\right] \qquad (2) \\
&= \exp\left\{-\Lambda(t) - \int_0^t \boldsymbol{X}_{ij}(u)^\top \boldsymbol{\beta}(u) du - \int_0^t \boldsymbol{Z}_{ij}^\top \gamma du - G(t; \theta_0)\right\},
\end{aligned}
$$

where $\Lambda(t) = \int_0^t \lambda(s) ds$ is the cumulative hazard function, $\exp\{-G(t;\theta_0)\} = \int e^{-xt} f(x; \theta_0) dx$. Let $N_{ij}(t) = \Delta_{ij} I(T_{ij} \leq t)$ be the counting process and $Y_{ij}(t) = I(T_{ij} \geq t)$ be the at-risk process. Then Equation $(2)$ is equivalent to

$$
E\left[dN_{ij}(t) \middle| \boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}, T_{ij} \geq t\right] = Y_{ij}(t)\left[dH(t) + \boldsymbol{X}_{ij}(t)^\top \beta(t) dt + \boldsymbol{Z}_{ij}^\top \gamma dt\right], \qquad (3)
$$

where $H(t) := H(t; \theta_0) = \Lambda(t) + G(t; \theta_0)$ is strictly increasing.

Let $dM_{ij}(t) = dN_{ij}(t) - Y_{ij}(t)\left[dH(t) + \boldsymbol{X}_{ij}(t)^\top \beta_0(t) dt + \boldsymbol{Z}_{ij}^\top \gamma_0 dt\right]$, by Equation (3) and assumptions C2), C3), C4), C6) in the Appendix, $M_{ij}(\cdot)$ is a local square integrable martingale. Use the idea in Lin & Ying (1994), following estimating equations can be easily constructed

from Equation (3):

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\int_0^\tau Y_{ij}(t)I(t\le s)\left[dN_{ij}(t)-dH(t)-\boldsymbol{Z}_{ij}^\top\gamma dt-\boldsymbol{X}_{ij}(t)^\top\beta(t)dt\right]=0, s>0 \quad (4)$$

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\int_0^\tau Y_{ij}(t)\begin{pmatrix}\boldsymbol{X}_{ij}(t)\\ \boldsymbol{Z}_{ij}\end{pmatrix}\left[dN_{ij}(t)-dH(t)-\boldsymbol{Z}_{ij}^\top\gamma dt-\boldsymbol{X}_{ij}(t)^\top\beta(t)dt\right]=0. \quad (5)$$

Due to the time-varying coefficient $\beta(t)$, it is difficult to estimate the parameters from Equations $(4)-(5)$. The idea of local linear technique is that we can approximate $\beta(t)$ by a linear function $\beta(t_0)+(t-t_0)\beta'(t_0)$ given $t_0$ is close to $t$. Thus localized version of Equation $(5)$ at a pre-specified time point $t_0$ (Li, Yin & Zhou, 2007) can be obtained, then the new parameter $\beta(t_0)$ can be estimated. The localized version of Equation $(5)$ at $t_0$ ($t_0\in[0,\tau]$) is:

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\int_0^\tau K_h(t-t_0)Y_{ij}(t)\begin{pmatrix}\boldsymbol{X}_{ij}(t)\\ \boldsymbol{Z}_{ij}\\ \boldsymbol{X}_{ij}(t)(t-t_0)\end{pmatrix} \quad (6)\\ &\times\left\{dN_{ij}(t)-dH(t)-\boldsymbol{Z}_{ij}^\top\gamma dt-\boldsymbol{X}_{ij}(t)^\top\left[\beta(t_0)+\beta'(t_0)(t-t_0)\right]dt\right\}=0.\end{aligned}$$

where $K_h(\cdot)=h^{-1}K(\cdot/h)$, $K(u)$ is the kernel function satisfies $\int K(u)du=1$. From Equation $(4)$ we have

$$\begin{aligned}dH(t)&=\frac{\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n_i}Y_{ij}(t)\left[dN_{ij}(t)-\boldsymbol{Z}_{ij}^\top\gamma dt-\boldsymbol{X}_{ij}(t)^\top\beta(t)dt\right]}{\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n_i}Y_{ij}(t)} \quad (7)\\ &=d\bar{N}(t)-\bar{\boldsymbol{Z}}(t)^\top\gamma dt-\bar{\boldsymbol{X}}(t)^\top\beta(t)dt,\end{aligned}$$

where

$$\bar{\boldsymbol{X}}(t)=\frac{\sum_{i=1}^{n}\sum_{j=1}^{n_i}Y_{ij}(t)\boldsymbol{X}_{ij}(t)}{\sum_{i=1}^{n}\sum_{j=1}^{n_i}Y_{ij}(t)},\ \bar{\boldsymbol{Z}}(t)=\frac{\sum_{i=1}^{n}\sum_{j=1}^{n_i}Y_{ij}(t)\boldsymbol{Z}_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{n_i}Y_{ij}(t)},\ \bar{N}(t)=\frac{\sum_{i=1}^{n}\sum_{j=1}^{n_i}Y_{ij}(t)N_{ij}(t)}{\sum_{i=1}^{n}\sum_{j=1}^{n_i}Y_{ij}(t)}.$$

Let $\zeta_0(t_0)=[\beta_0(t_0);\gamma_0(t_0);(\beta_0'(t_0))]$ be the true parameter vector and $\hat{\zeta}(\cdot)$ the corresponding estimate. Incorporate Equation $(7)$ into Equation $(6)$ and substitute $\beta(t)$ with $\beta(t_0)+\beta'(t_0)(t-t_0)$, $\zeta_0(t_0)$ can be estimated by

$$\begin{aligned}\hat{\zeta}(t_0)=\begin{bmatrix}\hat{\beta}(t_0)\\ \hat{\gamma}(t_0)\\ \hat{\beta}'(t_0)\end{bmatrix}&=\left[\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\int_0^\tau K_h(t-t_0)Y_{ij}(t)\begin{pmatrix}\boldsymbol{X}_{ij}(t)-\bar{\boldsymbol{X}}(t)\\ \boldsymbol{Z}_{ij}-\bar{\boldsymbol{Z}}(t)\\ (\boldsymbol{X}_{ij}(t)-\bar{\boldsymbol{X}}(t))(t-t_0)\end{pmatrix}^{\otimes 2}dt\right]^{-1} \quad (8)\\ &\times\left[\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\int_0^\tau K_h(t-t_0)Y_{ij}(t)\begin{pmatrix}\boldsymbol{X}_{ij}(t)-\bar{\boldsymbol{X}}(t)\\ \boldsymbol{Z}_{ij}-\bar{\boldsymbol{Z}}(t)\\ (\boldsymbol{X}_{ij}(t)-\bar{\boldsymbol{X}}(t))(t-t_0)\end{pmatrix}dN_{ij}(t)\right]\end{aligned}$$

Notice that here the estimate of $\gamma_0$ is written as $\hat{\gamma}(t_0)$ instead of $\hat{\gamma}$, due to the reason that the estimate of $\gamma$ use local data points. Given Equation $(7)$, it's natural to estimate $H(t_0)$ by (notice

that $\theta_0$ in $H(t;\theta_0)$ is dropped because when estimate $H$, $\theta_0$ is not needed):

$$\hat{H}(t_0) = \int_0^{t_0} \frac{\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n_i} Y_{ij}(t)\left[dN_{ij}(t) - \boldsymbol{Z}_{ij}^{\top}\hat{\gamma}(t)dt - \boldsymbol{X}_{ij}(t)^{\top}\hat{\beta}(t)dt\right]}{\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n_i} Y_{ij}(t)}. \tag{9}$$

Since $H(t_0) = \Lambda(t_0) + G(t_0;\theta_0)$, we only need to estimate $\theta_0$ to obtain the baseline hazard estimator, where $\hat{\Lambda}(t_0) = \hat{H}(t_0) - G(t_0;\hat{\theta}_0)$.

Denote $\boldsymbol{X}_{i\cdot}$ and $\boldsymbol{Z}_{i\cdot}$ as $(\boldsymbol{X}_{i1},\cdots,\boldsymbol{X}_{in_i})$ and $(\boldsymbol{Z}_{i1},\cdots,\boldsymbol{Z}_{in_i})$, respectively. For $i$th cluster, denote the maginal residual process as $d\varepsilon_{ij}(t) = dN_{ij}(t) - dH(t) - \boldsymbol{Z}_{ij}^{\top}\gamma dt - \boldsymbol{X}_{ij}(t)^{\top}\beta(t)dt$. Then for $j \neq l$ $(j,l \in \{1,\cdots,n_i\})$:

$$\begin{aligned}
&E\left[d\varepsilon_{ij}(t)d\varepsilon_{il}(s)|T_{ij}\geq t, T_{il}\geq s, \boldsymbol{X}_{i\cdot}, \boldsymbol{Z}_{i\cdot}\right] \\
&= \frac{E\left[(-dG(t;\theta_0)+\xi_i dt)\,(-dG(s;\theta_0)+\xi_i ds)\,I\left(T_{ij}\geq t\right)I\left(T_{il}\geq s\right)|\boldsymbol{X}_{i\cdot},\boldsymbol{Z}_{i\cdot},\right]}{E\left[I\left(T_{ij}\geq t\right)I\left(T_{il}\geq s\right)|\boldsymbol{X}_{i\cdot},\boldsymbol{Z}_{i\cdot}\right]} \\
&= \frac{P\left(C_{ij}\geq t, C_{il}\geq s|\boldsymbol{X}_{i\cdot},\boldsymbol{Z}_{i\cdot}\right)}{P\left(C_{ij}\geq t, C_{il}\geq s|\boldsymbol{X}_{i\cdot},\boldsymbol{Z}_{i\cdot}\right)P\left(T_{ij}^{*}\geq t, T_{il}^{*}\geq s|\boldsymbol{X}_{i\cdot},\boldsymbol{Z}_{i\cdot}\right)} \\
&\quad\times E\left[(-dG(t;\theta_0)+\xi_i dt)\,(-dG(s;\theta_0)+\xi_i ds)\,I\left(T_{ij}^{*}\geq t\right)I\left(T_{il}^{*}\geq s\right)|\boldsymbol{X}_{i\cdot},\boldsymbol{Z}_{i\cdot}\right] \\
&= \frac{E\left[(-dG(t;\theta_0)+\xi_i dt)\,(-dG(s;\theta_0)+\xi_i ds)\,I\left(T_{ij}^{*}\geq t\right)I\left(T_{il}^{*}\geq s\right)|\boldsymbol{X}_{i\cdot},\boldsymbol{Z}_{i\cdot}\right]}{P\left(T_{ij}^{*}\geq t, T_{il}^{*}\geq s|\boldsymbol{X}_{i\cdot},\boldsymbol{Z}_{i\cdot}\right)}.
\end{aligned} \tag{10}$$

The numerator of Equation (10) equals:

$$\begin{aligned}
&E\left[(-dG(t;\theta_0)+\xi_i dt)\,(-dG(s;\theta_0)+\xi_i ds)\,I(T_{ij}^{*}\geq t)I(T_{il}^{*}\geq s)|\boldsymbol{X}_{i\cdot},\boldsymbol{Z}_{i\cdot}\right] \\
&= E\left[(-dG(t;\theta_0)+\xi_i dt)\,(-dG(s;\theta_0)+\xi_i ds)\,EI(T_{ij}^{*}\geq t|\,\boldsymbol{X}_{i\cdot},\boldsymbol{Z}_{i\cdot})EI(T_{il}^{*}\geq s|\boldsymbol{X}_{i\cdot},\boldsymbol{Z}_{i\cdot})\right] \\
&= E_{\xi_i}\left[(-dG(t;\theta_0)+\xi_i dt)\,(-dG(s;\theta_0)+\xi_i ds)\,e^{-\xi_i(t+s)}\right] \\
&\quad\times\exp\left\{-\left[\Lambda(t)+\int_0^t \boldsymbol{Z}_{ij}^{\top}\gamma dv + \int_0^t \boldsymbol{X}_{ij}(v)^{\top}\beta(v)dv\right]\right. \\
&\quad\left.-\left[\Lambda(s)+\int_0^s \boldsymbol{Z}_{ij}^{\top}\gamma dv + \int_0^s \boldsymbol{X}_{ij}(v)^{\top}\beta(v)dv\right]\right\}.
\end{aligned} \tag{11}$$

The denominator of Equation (10) equals:

$$\begin{aligned}
&P\left(T_{ij}^{*}\geq t, T_{il}^{*}\geq s|\boldsymbol{X}_{i\cdot},\boldsymbol{Z}_{i\cdot}\right) \\
&= E_{\xi_i}\left\{\exp\left[-\Lambda(t) - \int_0^t (\boldsymbol{Z}_{ij}^{\top}\gamma + \boldsymbol{X}_{ij}(v)^{\top}\beta(v))dv - \xi_i dt\right.\right. \\
&\quad\left.\left.-\ \Lambda(s) - \int_0^s (\boldsymbol{Z}_{il}^{\top}\gamma + \boldsymbol{X}_{il}(v)^{\top}\beta(v))dv - \xi_i ds\right]\right\}.
\end{aligned} \tag{12}$$

Combine Equations $(10)-(12)$, we have

$$\begin{aligned} &E\left[d\varepsilon_{ij}(t)d\varepsilon_{il}(s)|T_{ij}\geq t,T_{il}\geq s,\boldsymbol{X}_{i\cdot},\boldsymbol{Z}_{i\cdot}\right] \\ &=\frac{E_{\xi_i}\left[\left(-dG(t;\theta_0)+\xi_i dt\right)\left(-dG(s;\theta_0)+\xi_i ds\right)e^{-\xi_i(t+s)}\right]}{E_{\xi_i}\left(e^{-\xi_i(t+s)}\right)}. \end{aligned} \tag{13}$$

Through simple calculations, we have:

$$\begin{aligned} E\left(\xi_i e^{-t\xi_i}\right)&=G'(t;\theta_0)\exp\left[-G(t;\theta_0)\right], \\ E\left({\xi_i}^2 e^{-t\xi_i}\right)&=\left(G'(t;\theta_0)^2-G''(t;\theta_0)\exp\left[-G(t;\theta_0)\right]\right), \end{aligned} \tag{14}$$

where $G'(t;\theta_0)$ and $G''(t;\theta_0)$ are the first and second derivatives of $G(t;\theta_0)$ with respect to $t$.

Substitute Equation $(14)$ into Equation $(13)$, then

$$\begin{aligned} &E\left[d\varepsilon_{ij}(t)d\varepsilon_{il}(s)|T_{ij}\geq t,T_{il}\geq s,\boldsymbol{X}_{i\cdot},\boldsymbol{Z}_{i\cdot}\right] \\ &=dG(t;\theta_0)dG(s;\theta_0)-G'(t+s;\theta_0)dG(t;\theta_0)ds-G'(t+s;\theta_0)dG(s;\theta_0)dt \\ &\quad+\left[G'(t+s;\theta_0)^2-G''(t+s;\theta_0)\right]dtds. \end{aligned} \tag{15}$$

Based on Equation (15), we can construct another estimating equation to estimate $\theta_0$:

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq l,j,l=1}^{n_i}\int_0^\tau\int_0^\tau Y_{ij}(t)Y_{il}(s)\left[d\hat{\epsilon}_{ij}(t)d\hat{\epsilon}_{il}(s)-Q(t,s;\theta)dtds\right]=0, \tag{16}$$

where

$$\begin{aligned} &d\hat{\varepsilon}_{ij}(t)=dN_{ij}(t)-d\hat{H}(t)-\boldsymbol{Z}_{ij}^\top\hat{\gamma}dt-\boldsymbol{X}_{ij}(t)^\top\hat{\beta}(t)dt, \\ &Q(t,s;\theta)=G'(t;\theta)G'(s;\theta)-G'(t+s;\theta)\left[G'(t;\theta)+G'(s;\theta)\right]+G'(t+s;\theta)^2-G''(t+s;\theta). \end{aligned}$$

Denote the solution of Equation $(16)$ by $\hat{\theta}$, then $\Lambda(t_0)$ can be estimated by $\hat{\Lambda}(t_0)=\hat{H}(t_0)-G(t_0;\hat{\theta})$.

## 3. ASYMPTOTIC PROPERTIES

In this section we will derive the asymptotic properties of proposed estimator. First we give some notations that is needed in the following theorems.

Let $\boldsymbol{W}_{ij}^{(1)}(t)=\boldsymbol{X}_{ij}(t),\bar{\boldsymbol{W}}^{(1)}(t)=\bar{\boldsymbol{X}}(t),\quad \boldsymbol{W}_{ij}^{(2)}(t)=\boldsymbol{Z}_{ij},\bar{\boldsymbol{W}}^{(2)}(t)=\bar{\boldsymbol{Z}}(t)$. Denote $a_{ij}(t|\boldsymbol{X}_{ij},\boldsymbol{Z}_{ij})dt=dA_{ij}(t|\boldsymbol{X}_{ij},\boldsymbol{Z}_{ij})=\boldsymbol{X}_{ij}(t)^\top\beta(t)dt+\boldsymbol{Z}_{ij}^\top\gamma dt+dH(t)$. For $p,q=1,2$, define:

$$\boldsymbol{c}^{(pq)}(t)=E[\sum_{j=1}^{n_i}Y_{ij}(t)(\boldsymbol{W}_{ij}^{(p)}(t)-\bar{\boldsymbol{W}}^{(p)}(t))(\boldsymbol{W}_{ij}^{(q)}(t)-\bar{\boldsymbol{W}}^{(q)}(t))^\top],$$

$$\boldsymbol{d}^{(pq)}(t)=E[\sum_{j=1}^{n_i}Y_{ij}(t)(\boldsymbol{W}_{ij}^{(p)}(t)-\bar{\boldsymbol{W}}^{(p)}(t))(\boldsymbol{W}_{ij}^{(q)}(t)-\bar{\boldsymbol{W}}^{(q)}(t))^\top a_{ij}(t|\boldsymbol{X}_{ij},\boldsymbol{Z}_{ij})].$$

For $p = q = 0$, define:

$$\boldsymbol{c}^{(00)}(t) = E[\sum_{j=1}^{n_i} Y_{ij}(t)], \boldsymbol{d}^{(00)}(t) = E[\sum_{j=1}^{n_i} Y_{ij}(t) a_{ij}(t|\boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij})],$$

at last, we define

$$\boldsymbol{\Sigma}(t) = \begin{bmatrix} \boldsymbol{d}^{(11)}(t)\kappa_{20} & \boldsymbol{d}^{(12)}(t)\kappa_{20} & \boldsymbol{d}^{(11)}(t)\kappa_{21} \\ \boldsymbol{d}^{(21)}(t)\kappa_{20} & \boldsymbol{d}^{(22)}(t)\kappa_{20} & \boldsymbol{d}^{(21)}(t)\kappa_{21} \\ \boldsymbol{d}^{(11)}(t)\kappa_{21} & \boldsymbol{d}^{(12)}(t)\kappa_{21} & \boldsymbol{d}^{(11)}(t)\kappa_{22} \end{bmatrix}.$$

Then we have following theorems:

**Theorem 1.** *Let $\boldsymbol{M}$ be a $(2p + r) \times (2p + r)$-diagonal matrix with the first $p + r$ elements equal to 1 and the remaining $p$ elements equal to $h$, let $\kappa_{ij} = \int K(u)^i u^j du$, where $i, j = 0, 1, 2, \cdots$. Under conditions A1)-A3), B1)-B2) and C1)-C6) in the Appendix, we have that, for any $t_0 \in (0, \tau)$:*

$$\sqrt{nh}\{\boldsymbol{M}(\hat{\zeta}(t_0) - \zeta_0(t_0)) - \frac{1}{2}h^2\kappa_{12}\boldsymbol{D}^{-1}(t_0)\boldsymbol{b}(t_0)\} \xrightarrow{\mathcal{D}} N(0, \boldsymbol{D}^{-1}(t_0)\boldsymbol{\Sigma}(t_0)\boldsymbol{D}^{-1}(t_0)).$$

*where*

$$\boldsymbol{D}(t_0) = \begin{bmatrix} \boldsymbol{c}^{(11)}(t_0)\kappa_{10} & \boldsymbol{c}^{(12)}(t_0)\kappa_{10} & \boldsymbol{c}^{(11)}(t_0)\kappa_{11} \\ \boldsymbol{c}^{(21)}(t_0)\kappa_{10} & \boldsymbol{c}^{(22)}(t_0)\kappa_{10} & \boldsymbol{c}^{(21)}(t_0)\kappa_{11} \\ \boldsymbol{c}^{(11)}(t_0)\kappa_{11} & \boldsymbol{c}^{(12)}(t_0)\kappa_{11} & \boldsymbol{c}^{(11)}(t_0)\kappa_{12} \end{bmatrix}, \boldsymbol{b}(t_0) = \begin{bmatrix} \boldsymbol{c}^{(11)}(t_0) \\ \boldsymbol{c}^{(22)}(t_0) \\ \boldsymbol{0}_p \end{bmatrix} \boldsymbol{\beta}_0''(t_0),$$

*with $\boldsymbol{0}_p$ denoting a zero column vector of length $p$. Furthermore, since $\kappa_{11} = 0$, we can obtain $[\boldsymbol{D}(t_0)]^{-1}\boldsymbol{b}(t_0) = [\beta_0''(t_0); \boldsymbol{0}_{p+r}]$. Thus for the constant part, the bias of the kernel estimator is 0.*

Obviously, we only utilize the local data to estimate the constant coefficients, as a result, the convergence rate of the constant coefficients is slower than $\sqrt{n}$-consistency. However, we can circumvent this drawback and still achieve $\sqrt{n}$-consistency for the following term:

$$\tilde{\gamma} = \int \Gamma(t)\hat{\gamma}(t)dt, \tag{17}$$

where the weight matrix satisfies $\int \Gamma(t)dt = I_{r\times r}$, and $I_{r\times r}$ is an identity matrix of size $r \times r$ (Yin, Li, & Zeng, 2008).

**Theorem 2.** *Under conditions A1)-A3), B1)-B2) and C1)-C6) in the Appendix, $\sqrt{n}(\tilde{\gamma} - \gamma_0)$ converges in distribution to a mean-zero normal distribution.*

In the following, we will show the uniform consistency of $\hat{H}(t)$:

**Theorem 3.** *Under conditions A1)-A3), B1)-B2) and C1)-C6) in the Appendix, we have that, for all $t \in [0, \tau]$:*

$$\hat{H}(t) \xrightarrow{P} H(t).$$

Last theorem establishes the consistency of estimator in frailty distribution as well as the baseline hazard:

**Theorem 4.** *Under conditions A1)-A3), B1)-B2) and C1)-C6) in the Appendix, uniformly for* $t \in (0, \tau)$*:*

$$\{\hat{\theta}, \hat{\Lambda}(t)\} \xrightarrow{\mathcal{P}} \{\theta_0, \Lambda(t)\}.$$

## 4. SIMULATION

In this part, we perform various simulation studies to assess the performance of the proposed method. Through out the simulation, we use the Gaussian kernel function. To choose the optimal bandwidth, we minimize the asymptotic weighted mean squared error (Yin, Li, & Zeng, 2008). To be specific, for $k$th component of $\hat{\zeta}(t_0)$, let $\phi_k(t_0)$ be the $k$th component of $\boldsymbol{D}^{-1}(t_0)$, $\sigma_{kk}$ be the $k$th diagonal element of $\boldsymbol{D}^{-1}(t_0)\boldsymbol{\Sigma}(t_0)\boldsymbol{D}^{-1}(t_0)$, and $\Psi(\cdot)$ is a nonnegative and integrable weighted function. Theoretically, we choose the optimal bandwidth by minimizing the following quantity with respect to $h$:

$$\int_0^\tau (\frac{1}{4}h^4\kappa_{12}^2\phi_k(t_0)^2 + \frac{1}{nh}\sigma_{kk}(t_0))\Psi(t_0)dt_0,$$

which lead to the optimal $h$ as

$$h_{opt,k} = \left[\frac{\int \sigma_{kk}(t_0)\Psi(t_0)dt_0}{\int \kappa_{12}^2\phi_k^2(t_0)\Psi(t_0)dt_0}\right]^{1/5} n^{-1/5}. \tag{18}$$

However, since (18) involves unknown parameters $\sigma_{kk}(t_0)$, $\phi_k(t_0)$ and $\Psi(t_0)$, practically, we use the rule of thumb to choose the bandwidth, where $h = \hat{\sigma}^* n^{-1/5}$, and $\hat{\sigma}^*$ stands for the standard deviation of the observed failure time.

For all the simulation scenarios, the estimation procedure was repeated 500 times and $h$ is chosen as the optimal bandwidth according to rule of thumb. To estimate the asymptotic variance, the clusters are selected by random sampling with replacement but leave the observations for each cluster unchanged. The bootstrap cluster size is 100.

The first model is

$$\lambda_{ij}(t|X_{ij}) = \lambda_0(t) + X_{ij}\beta(t) + \xi_i, \tag{19}$$

which only contains one covariate. Let $\lambda_0(t) = 1$, $X_{ij}(t)$ obeys uniform distribution from 0 to 1, $\beta(t) = t$, $\xi_i + \theta$ obeys exponential distribution with mean $\theta$, the true value for $\theta$ is 1. The censoring variable follows uniform distribution from 0 to $\tau$, here $\tau$ is set to be 5 to result in moderate censoring rate between 10% to 20%. The number of clusters are set as 20, 50 and 100, respectively. Within each cluster, the cluster size is randomly generated between 2-4. As suggested by one of the referees, we also performed simulation under the setting that the cluster size is large, in order to figure out whether the cluster size or the number of clusters will be more sensitive to the outcome. In the large cluster size setting, we choose cluster size as 25, while the number of clusters are set as 5,10 and 20, respectively. The results are shown in Table 1. From the table, we can see that under the small cluster size scenario the biases for parameter $\hat{\beta}_n(t)$ and $\hat{H}_n(t)$ are both small, and the SEs (the empirical standard deviation based on the observed sample) and SDs (the average of the estimated standard errors based on the asymptotic distribution) are close, the coverage probabilities are close to the nominal level 95%. However, when the cluster size is large, the estimator's behavior will be more sensitive, the coverage probabilities is lower than

TABLE 1: Simulation results for model (19) under independent censoring

| t | n | Cluster Size | $\hat{\beta}(t)$ | bias | SE | SD | cov(%) | $\hat{H}(t)$ | bias |
|---|---|---|---|---|---|---|---|---|---|
| | 20 | 2-4 | 0.157 | -0.043 | 0.915 | 0.910 | 92.8 | 0.187 | 0.005 |
| 0.20 | 50 | 2-4 | 0.135 | -0.065 | 0.502 | 0.520 | 94.2 | 0.189 | 0.007 |
| | 100 | 2-4 | 0.199 | -0.001 | 0.347 | 0.362 | 94.8 | 0.184 | 0.002 |
| | 20 | 2-4 | 0.463 | -0.037 | 0.697 | 0.735 | 94.0 | 0.423 | 0.017 |
| 0.50 | 50 | 2-4 | 0.481 | -0.019 | 0.443 | 0.428 | 94.8 | 0.424 | 0.018 |
| | 100 | 2-4 | 0.508 | 0.008 | 0.292 | 0.302 | 95.0 | 0.408 | 0.003 |
| | 20 | 2-4 | 0.784 | -0.017 | 0.847 | 0.929 | 93.6 | 0.615 | 0.027 |
| 0.80 | 50 | 2-4 | 0.815 | 0.015 | 0.520 | 0.527 | 94.4 | 0.641 | 0.026 |
| | 100 | 2-4 | 0.826 | 0.026 | 0.353 | 0.372 | 95.2 | 0.593 | 0.005 |
| | 5 | 25 | 0.210 | 0.010 | 0.523 | 0.482 | 89.6 | 0.189 | 0.007 |
| 0.20 | 10 | 25 | 0.194 | -0.006 | 0.366 | 0.344 | 92.0 | 0.185 | 0.003 |
| | 20 | 25 | 0.182 | -0.018 | 0.249 | 0.247 | 93.0 | 0.189 | 0.006 |
| | 5 | 25 | 0.508 | 0.008 | 0.448 | 0.449 | 89.2 | 0.428 | 0.023 |
| 0.50 | 10 | 25 | 0.490 | -0.010 | 0.324 | 0.304 | 92.8 | 0.414 | 0.008 |
| | 20 | 25 | 0.497 | -0.003 | 0.228 | 0.224 | 94.8 | 0.418 | 0.012 |
| | 5 | 25 | 0.837 | 0.037 | 0.564 | 0.686 | 93.2 | 0.627 | 0.039 |
| 0.80 | 10 | 25 | 0.794 | -0.006 | 0.392 | 0.381 | 91.4 | 0.602 | 0.014 |
| | 20 | 25 | 0.800 | 0.000 | 0.267 | 0.272 | 93.2 | 0.606 | 0.018 |

the nominal level 95% when number of clusters is small, though it can be improved when the number of clusters is increased.

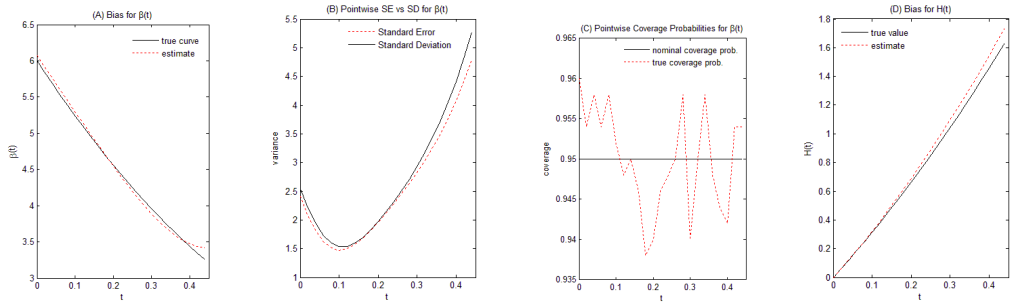


FIGURE 1: Estimation of $\beta(t)$ and $H(t)$ curves in $[0, 0.45]$ with 25 grid points. The cluster size is 100 and censor rate is 12%. The true curves are solid lines with black color, the estimate curves are dash lines with red color, $h = 0.15$

We did further simulation study under model (19), in order to study the estimator's behavior when censoring is dependent on the failure time. All the parameter settings are the same as before except the way of censoring time generation. The censoring time variable $c$ for $j$th individual in $i$th cluster is generated according to the following model:

$$\lambda_{ij}(c|X_{ij}) = \bar{\lambda}(c)\exp(X_{ij}\bar{\gamma}),$$

TABLE 2: Simulation results for model (19) under dependent censoring

| t | n | Cluster Size | $\hat{\beta}(t)$ | bias | SE | SD | cov(%) | $\hat{H}(t)$ | bias |
|---|---|---|---|---|---|---|---|---|---|
| | 20 | 2-4 | 0.186 | -0.014 | 1.008 | 0.970 | 92.4 | 0.186 | 0.004 |
| 0.20 | 50 | 2-4 | 0.135 | -0.065 | 0.538 | 0.553 | 93.2 | 0.189 | 0.007 |
| | 100 | 2-4 | 0.193 | -0.007 | 0.376 | 0.388 | 96.0 | 0.184 | 0.002 |
| | 20 | 2-4 | 0.480 | -0.020 | 1.040 | 1.196 | 92.6 | 0.418 | 0.013 |
| 0.50 | 50 | 2-4 | 0.498 | -0.002 | 0.617 | 0.634 | 93.8 | 0.423 | 0.018 |
| | 100 | 2-4 | 0.504 | 0.004 | 0.426 | 0.439 | 94.4 | 0.408 | 0.003 |
| | 20 | 2-4 | 0.860 | 0.060 | 1.644 | 2.223 | 93.0 | 0.615 | 0.027 |
| 0.80 | 50 | 2-4 | 0.885 | 0.085 | 0.951 | 0.978 | 93.6 | 0.610 | 0.022 |
| | 100 | 2-4 | 0.849 | 0.049 | 0.592 | 0.645 | 94.6 | 0.597 | 0.009 |
| | 5 | 25 | 0.208 | 0.008 | 0.557 | 0.514 | 87.8 | 0.189 | 0.006 |
| 0.20 | 10 | 25 | 0.198 | -0.002 | 0.398 | 0.371 | 90.8 | 0.186 | 0.004 |
| | 20 | 25 | 0.178 | -0.022 | 0.263 | 0.267 | 93.2 | 0.189 | 0.006 |
| | 5 | 25 | 0.515 | 0.015 | 0.688 | 0.692 | 85.0 | 0.432 | 0.026 |
| 0.50 | 10 | 25 | 0.504 | 0.004 | 0.445 | 0.444 | 91.2 | 0.412 | 0.006 |
| | 20 | 25 | 0.496 | -0.004 | 0.331 | 0.322 | 92.4 | 0.419 | 0.014 |
| | 5 | 25 | 0.887 | 0.087 | 1.179 | 1.419 | 83.0 | 0.633 | 0.045 |
| 0.80 | 10 | 25 | 0.807 | 0.007 | 0.672 | 0.658 | 89.2 | 0.598 | 0.011 |
| | 20 | 25 | 0.804 | 0.004 | 0.440 | 0.453 | 93.6 | 0.608 | 0.020 |

where $\bar{\lambda}(c) = 1$ is the baseline, $\bar{\gamma}$ is set to be 0.2 which lead to heavy censoring rate around 53%. Since both $C_{ij}$ and $T_{ij}$ rely on the covariate $X_{ij}$, thus the censoring is dependent under this scenario. Both small and large cluster size are considered here. The results are shown in Table 2. The behavior of the estimator is still good, except that the coverage probabilities are more sensitive when cluster size is large and number of clusters is small, but it can also be improved when number of clusters is increased. In Figure 1, we aim to show that the performance of the estimator is uniform consistent along the interval, here a fixed bandwidth of 0.15 is chosen, while $\lambda_0(t) = 4t + 3$, $\beta(t) = 4(t-1)^2 + 2$, $\xi_i$ and $X_{ij}$ is the same as model (19). We estimate the parameters at 25 pre-specified uniform grid points in $[0, 0.45]$, then draw a line connecting the 25 points sequentially. From this figure, it can be seen that the estimated and true curves for $\beta(t)$ are close to each other uniformly, and the behavior for SE and SD curves are also similar, while the true coverage probabilities fluctuated slightly around the nominal coverage probabilities 95%, the difference for estimated and true $H(t)$ curves is also very small along the whole interval. In Figure 1(B), it is shown that the variances are small in the middle, and large at the two tails due to lack of data.

The second model we consider is

$$\lambda_{ij}(t|X_{ij}) = \lambda_0(t) + X_{ij}\beta(t) + Z_{ij}\gamma + \xi_i, \tag{20}$$

which contains one time-varying coefficient and one constant coefficient. Here $\lambda_0(t) = 1$, $X_{ij}$ obeys uniform distribution from 0 to 1, $\beta(t) = t$, $Z_{ij}$ obeys binomial distribution with $P(Z_{ij} = 1) = P(Z_{ij} = 0) = 0.5$, $\gamma = 1$, $\xi_i + \theta$ obeys exponential distribution with mean $\theta$, the true value for $\theta$ is 1. The censoring variable follows uniform distribution from 0 to $\tau$, here we choose the light censoring rate about 11%. Both small cluster size and large cluster size settings are considered here, for the former setting, cluster size was randomly chose between 2-4 and number

TABLE 3: Simulation results for model (20)

| t | n | Cluster Size | $\hat{\beta}(t)$ | bias | SE | SD | cov(%) | $\hat{\gamma}$ | bias | SE | SD | cov(%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 20 | 2-4 | 0.278 | 0.078 | 1.107 | 1.173 | 92.6 | 1.041 | 0.041 | 0.636 | 0.632 | 92.8 |
| 0.20 | 50 | 2-4 | 0.225 | 0.025 | 0.634 | 0.662 | 94.8 | 0.918 | -0.082 | 0.426 | 0.381 | 96.2 |
| | 100 | 2-4 | 0.197 | -0.003 | 0.410 | 0.441 | 96.8 | 0.898 | -0.102 | 0.391 | 0.286 | 95.0 |
| | 20 | 2-4 | 0.642 | 0.142 | 1.037 | 1.127 | 94.6 | 1.046 | 0.046 | 0.630 | 0.619 | 94.0 |
| 0.50 | 50 | 2-4 | 0.528 | 0.028 | 0.585 | 0.598 | 96.0 | 0.921 | -0.079 | 0.409 | 0.370 | 94.6 |
| | 100 | 2-4 | 0.489 | -0.011 | 0.414 | 0.418 | 95.2 | 0.907 | -0.093 | 0.373 | 0.279 | 94.6 |
| | 20 | 2-4 | 0.999 | 0.199 | 1.569 | 1.745 | 92.4 | 1.049 | 0.049 | 0.717 | 0.744 | 92.2 |
| 0.80 | 50 | 2-4 | 0.845 | 0.045 | 0.726 | 0.770 | 96.4 | 0.942 | -0.058 | 0.456 | 0.435 | 95.4 |
| | 100 | 2-4 | 0.796 | -0.004 | 0.546 | 0.531 | 94.0 | 0.956 | -0.044 | 0.372 | 0.328 | 92.2 |
| | 5 | 25 | 0.220 | 0.022 | 0.656 | 0.594 | 87.6 | 0.986 | -0.014 | 0.413 | 0.338 | 84.8 |
| 0.20 | 10 | 25 | 0.230 | 0.030 | 0.416 | 0.425 | 93.2 | 0.926 | -0.074 | 0.372 | 0.271 | 92.2 |
| | 20 | 25 | 0.196 | -0.004 | 0.281 | 0.297 | 94.4 | 0.831 | -0.170 | 0.420 | 0.234 | 93.2 |
| | 5 | 25 | 0.486 | -0.014 | 0.594 | 0.619 | 87.0 | 0.976 | -0.024 | 0.397 | 0.336 | 85.4 |
| 0.50 | 10 | 25 | 0.506 | 0.006 | 0.414 | 0.424 | 93.4 | 0.930 | -0.070 | 0.362 | 0.271 | 91.2 |
| | 20 | 25 | 0.492 | -0.008 | 0.288 | 0.304 | 94.8 | 0.848 | -0.152 | 0.364 | 0.226 | 93.0 |
| | 5 | 25 | 0.785 | -0.015 | 0.824 | 0.942 | 83.0 | 1.001 | 0.001 | 0.482 | 0.444 | 82.8 |
| 0.80 | 10 | 25 | 0.782 | -0.019 | 0.595 | 0.559 | 90.2 | 0.971 | -0.029 | 0.382 | 0.325 | 89.8 |
| | 20 | 25 | 0.800 | 0.000 | 0.396 | 0.380 | 92.2 | 0.940 | -0.060 | 0.287 | 0.246 | 92.0 |

of clusters equals 20, 50, 100, respectively, while for the latter setting, cluster size was fixed at 25 with number of clusters equal to 5, 10, 20 respectively. The results are shown in Table 3, from this table, we can see that the biases for time-varying and constant coefficient are small, the corresponding SEs and SDs are close, the empirical coverage probabilities are close to nominal level 95% for small cluster size setting, when cluster size is large and number of clusters is small, the empirical coverage probabilities tend to be smaller than 95%, however, it can be improved when increasing the number of clusters. Due to the space limit, we omit the biases for estimator of $H(t)$, they are still very small and the results can be obtained upon request.

Figure 2 shows the estimator's behavior along a whole curve. In this curve, 40 grid points are uniformly distributed in $[0, 0.4]$, the bandwidth is fixed at 0.2. We estimate the parameters at the grid points and draw a line connecting them sequentially. The parameter setting is the same as the previous simulation for model (20) except for the values of $\lambda_0(t)$ and $\beta(t)$, here $\lambda_0(t) = 4t + 3$ and $\beta(t) = 4(t-1)^2 + 3$. From Figure 2, we can see that all the estimated curves closely match the true curves, and SE and SD curves close to each other as well, the empirical coverage probabilities curves are also close to the nominal 95% level, and the biases for $\hat{H}(t)$ are also uniformly small. It indicates that the estimator's empirical performance is uniformly good.

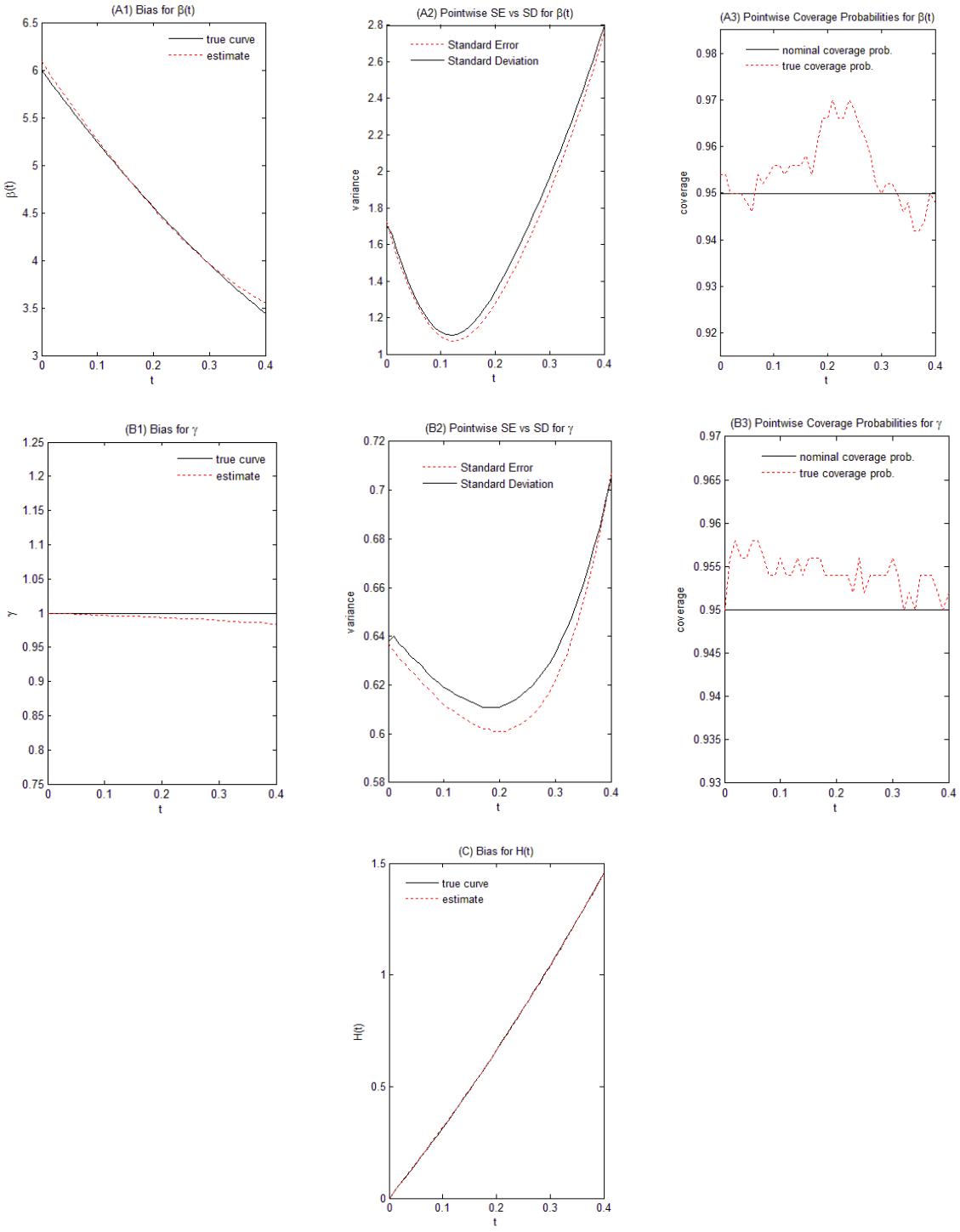


FIGURE 2: Estimation of $\beta(t)$, $\gamma(t)$ and $H(t)$ curves in $[0, 0.4]$ with 40 grid points under model (20). The cluster size 200 and censor rate is 12% . The true curves are solid lines with black color, the estimate curves are dash lines with red color, $h = 0.2$

## 5. REAL DATA

In this section, we apply the proposed method to the Kenya parasitemia data set. 607 children were recruited in the study. At the date of recruitment, each child was received therapies no matter whether being infected by parasitemia or not. Then the researchers examined the blood of these children, those with positive blood films were exclude from the study. As a result, the data contains 542 children from 309 families. The time we concern is from the examination to infection of parasitemia. The risk factors include Age, Gender, Daily Mean Bite, and Baseline Parasitemia Density (BPD). Following Yu & Lin (2010), we log-transform the baseline parasitemia density and denote the new variable as LNBPD, and also take the quartic root of Daily Mean Bite and denote new variable as BITE. The censoring rate for this data is about $11\%$. Figure 3 shows the histogram of Age, BITE, Gender, LNBPD, $T_{ij}$, and Number of Clusters in Kenya parasitemia dataset. About a half of the clusters contain more than 2 individuals, which suggests that marginal approach may not be appropriate here.

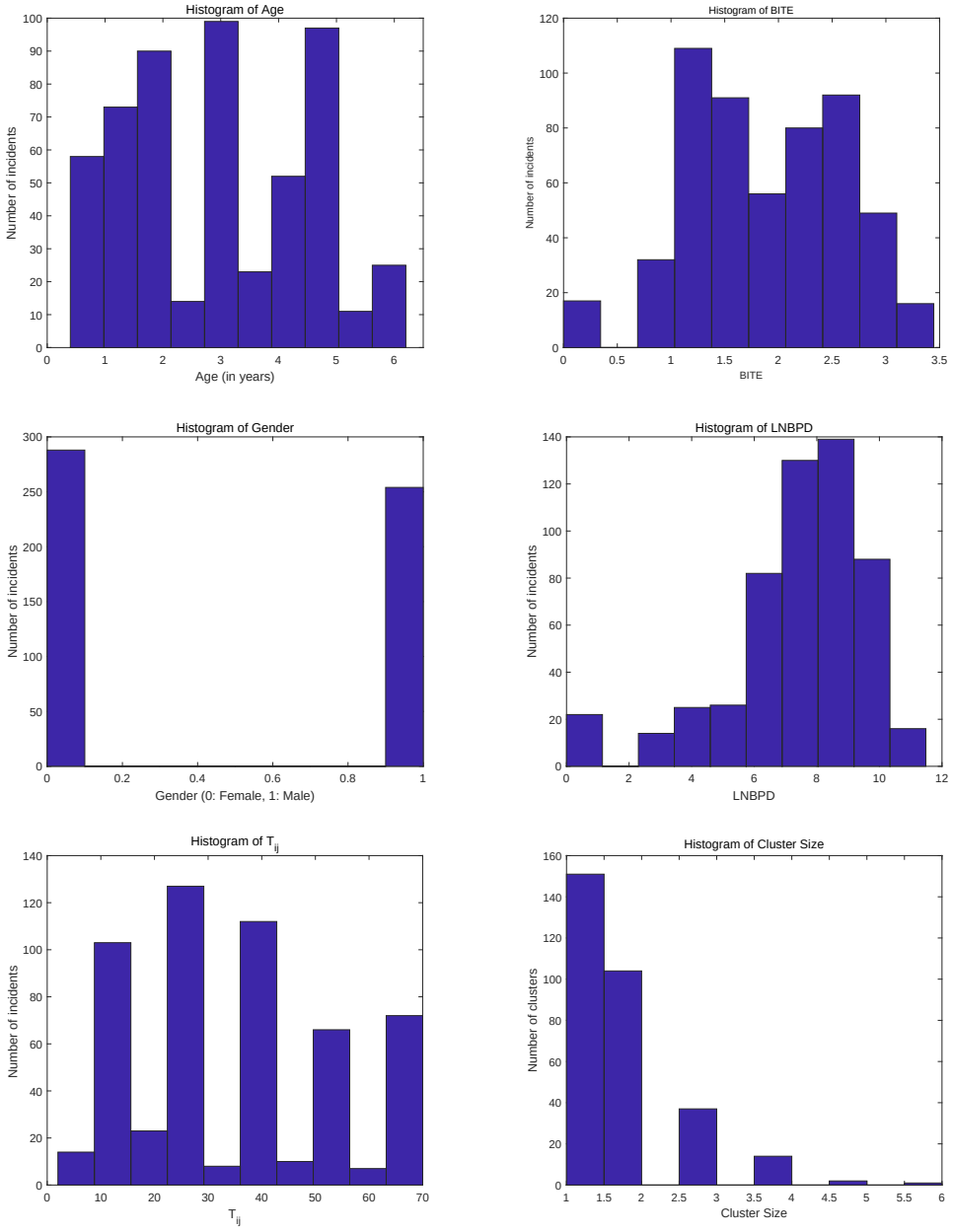


FIGURE 3: Histogram of Age, BITE, Gender, LNBPD, $T_{ij}$, and Number of Clusters in the Kenya parasitemia data

As being mentioned in Section 1, we choose to set the coefficient of LNBPD as time varying and the coefficients of Age, Gender and BITE as constant, which results in the following model:

$$\lambda_{ij}(t|\cdot) = \lambda(t) + \text{LNBPD} \times \beta_1(t) + \text{Age} \times \beta_2 + \text{Gender} \times \beta_3 + \text{BITE} \times \beta_4 + \xi_i, \quad (21)$$

TABLE 4: Constant estimate for model (21) under the Kenya parasitemia data

| t | Age | sd | Gender | sd | BITE | sd |
|---|---|---|---|---|---|---|
| 15 | .0006 | .0006 | .0008 | .0020 | -.0005 | .0003 |
| 20 | .0007 | .0006 | .0008 | .0020 | -.0005 | .0003 |
| 25 | .0007 | .0006 | .0008 | .0020 | -.0003 | .0004 |
| 30 | .0007 | .0006 | .0007 | .0020 | -.0001 | .0004 |
| 35 | .0006 | .0007 | .0005 | .0025 | -.0001 | .0005 |

here $\lambda(t)$ is the baseline hazard function, $\beta_1(t)$, $\beta_2$, $\beta_3$ and $\beta_4$ represent the covariate effect of LNBPD, Age, Gender and BITE, respectively, $\xi_i$ represents the frailty. For the time varying part, we choose bandwidth $h = 15$ because we found it fits data well. Figure 4 represents the estimated curve of coefficients of LNBPD using the initial estimator as time varies from 0 to 80. This figure indicates that the LNBPD is significantly positively associated with the risk of parasitemia before approximately day 24, however, after day 24, the association is not significant any more, finally, the effect of LNBPD decreases to zero at day 80.

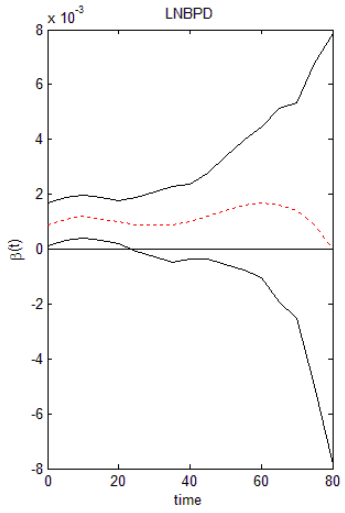


FIGURE 4: Kernel estimate and 95% confidence interval of the time varying coefficient of LNBPD for the Kenya parasitemia data, h=15

Table 4 shows the estimates of constant coefficients calculated under different $t$, where $t = 15, 20, 25, 30, 35$. The reason that we choose those time points is due to observation of the Histogram of $T_{ij}$ in Figure 3, which contains sufficient amount of data around those points. We can see that the estimate do not differ too much at different time point (also notice that sd for each covariate at different time $t$ are close to each other as well), thus the constant covariate effect assumption is reasonable, all the results are not significant, which means there is no difference between boys and girls, older children and younger children to catch parasitemia, as well as number of daily mean bite. Though Yu & Lin (2010)'s results can not be simply compared with ours due to different model settings, there are still some aspects that can point out. Both methods can capture the same trend of effect of LNBPD, which provides evidence to the common sense that the effect of baseline parasitemia density on the time of onset of parasitemia is likely to decrease

TABLE 5: Comparison of model (22) and (23) for the Kenya parasitemia data

| | LNBPD | sd | Age | sd | Gender | sd | BITE | sd |
|---|---|---|---|---|---|---|---|---|
| frailty Cox model | .0520 | .0006 | .0380 | .0010 | .0160 | .0090 | -.0420 | .0003 |
| frailty Additive model | .0010 | .0004 | .0007 | .0006 | .0001 | .0020 | -.0008 | .0004 |

over time. However, their method produces the result under a working independence assumption, while our method takes the inner-cluster correlation into consideration. Our coefficients are much smaller than Yu & Lin (2010), which is as expected due to intrinsic nature of additive hazards (Yin & Cai, 2004), while Yu & Lin (2010) fit this data under a marginal semiparametric Cox model.

We also compare our results with the gamma-frailty Cox model (Gorfine, Zucker, & Hsu, 2006) as requested by two anonymous referees and the Associate Editor, the gamma-frailty Cox model has following form:

$$\lambda_{ij}(t|\cdot) = \xi_i \lambda(t) \exp(\text{LNBPD} \times \beta_1 + \text{Age} \times \beta_2 + \text{Gender} \times \beta_3 + \text{BITE} \times \beta_4). \quad (22)$$

The parameters $\lambda(t)$, $\beta_1$, $\beta_2$, $\beta_3$, $\beta_4$ and $\xi_i$ in model (22) have the same meanings as model (21). The analysis was conducted under the R package 'frailtySurv' (Monaco, Gorfine, & Hsu, 2017). Since no parallel semiparametric version under the frailty Cox model has been developed, here we just treat all the covariates effect as constant. In addition, we also did the analysis under the additive mixed effect model (Cai & Zeng, 2011):

$$\lambda_{ij}(t|\cdot) = \lambda(t) + \text{LNBPD} \times \beta_1 + \text{Age} \times \beta_2 + \text{Gender} \times \beta_3 + \text{BITE} \times \beta_4 + \xi_i, \quad (23)$$

which is a special case of model (21). The results are shown in Table 5.

As can be seen from Table 5, both models produce estimates with the same signs, which indicates the same direction of covariate effect, while the significance for covariate effect might differ. However, we find that the effect of BITE is negative and significant for both methods, which means that higher daily mean bite will lead to smaller hazard rate, which is a contradiction to common knowledge. While in our model, the effect of BITE is not significant, which indicate that our model is more reasonable to fit this data while can also capture the dynamic effect of baseline parasitemia density to the hazard rate.

## 6. DISCUSSION

In this paper, we propose a semiparametric additive frailty hazard model for clustered failure time data which can deal with time-varying covariate effect and constant covariate effect simultaneously. The model is flexible and computational simple when compared with many existing frailty models. In the Kenya parasitemia dataset, we choose the effect of LNBPD as time dependent due to previous studies (McEroy et al., 1997; Yu & Lin, 2010), practically, how to determine which covariates effect is time dependent and which is constant? Normally speaking, it will depend on expert's opinion. However, we have an intuitive way to roughly determine: we can just set all the covariates effect as time dependent at first, then fit an estimate curve for each covariate. If the curve is close to a horizontal line, then we can set the covariate effect of this variable as constant. Actually we have already did this in the simulation study, the sub-figure B2 of Figure 2 presents the curve estimate of $\gamma$ with respect to time $t$ even if we know it is constant, and the curve is still close to the true horizontal line of $\gamma$, which shows that our method is still promising to detect the time dependency of covariate effect. Similar techniques have also been developed in partly parametric Aalen's additive model (McKeague & Sasieni, 1994). In the future, we will develop a formal model checking procedure to determine which part of coefficients is constant. Notice in

the simulation we did not report the results for $\hat{\theta}$, because it can be easily obtained following Cai & Zeng (2011).

## ACKNOWLEDGEMENTS

The authors are grateful to the Editor, Associate Editor and referees for their valuable suggestions that greatly improved the manuscript. Zhou's work was supported by the State Key Program of National Natural Science Foundation of China and the State Key Program in the Major Research Plan of National Natural Science Foundation of China.

## BIBLIOGRAPHY

Aalen, O. O. (1989). A linear regression model for the analysis of life times. *Statistics in medicine*, 8(8), 907–925.

Cai, J., Fan, J., Zhou, H., & Zhou, Y. (2007). Hazard models with varying coefficients for multivariate failure time data. *The Annals of Statistics*, 35(1), 324–354.

Cai, J. & Zeng, D. (2011). Additive mixed effect model for clustered failure time data. *Biometrics*, 67(4), 1340–1351.

Cai, Z. & Sun, Y. (2003). Local linear estimation for time-dependent coefficients in Cox's regression models. *Scandinavian Journal of Statistics*, 30(1), 93–111.

Fan, J. & Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *The Annals of Statistics*, 30, 74–99.

Fleming, T. R. & Harrington, D. P. (1991). *Counting processes and survival analysis*, John Wiley & Sons, New York.

Gorfine, M., Zucker, D. M., & Hsu, L. (2006). Prospective survival analysis with a general semiparametric shared frailty model: A pseudo full likelihood approach. *Biometrika*, 93(3), 735–741.

Hastie, T. & Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(4), 757–779.

Huster, W. J., Brookmeyer, R., & Self, S. G. (1989). Modelling paired survival data with covariates. *Biometrics*, 45, 145–156.

Kalbfleisch, J. D. & Prentice, R. L. (2011). *The statistical analysis of failure time data*, John Wiley & Sons, New York.

Li, H. & Yin, G. (2009). Generalized method of moments estimation for linear regression with clustered failure time data. *Biometrika*, 96(2), 293–306.

Li, H., Yin, G., & Zhou, Y. (2007). Local likelihood with time–varying additive hazards model. *Canadian Journal of Statistics*, 35(2), 321–337.

Lin, D. Y. (1994). Cox regression analysis of multivariate failure time data: the marginal approach. *Statistics in medicine*, 13(21), 2233–2247.

Lin, D. Y. & Ying, Z. (1994). Semiparametric analysis of the additive risk model. *Biometrika*, 81(1), 61–71.

Lu, S. E. & Wang, M. C. (2005). Marginal analysis for clustered failure time data. *Lifetime Data Analysis*, 11(1), 61-79.

Martinussen, T. & Scheike, T. H. (2007). *Dynamic regression models for survival data*. Springer Science & Business Media, New York.

Martinussen, T., Scheike, T. H., & Zucker, D. M. (2011). The Aalen additive gamma frailty hazards model. *Biometrika*, 98(4), 831–843.

McElroy, P. D., Beier, J. C., Oster, C. N., Onyango, F. K., Oloo, A. J., Lin, X., Beadle, C., & Hoffman, S. L. (1997). Dose-and time-dependent relations between infective Anopheles inoculation and outcomes of Plasmodium falciparum parasitemia among children in western Kenya. *American journal of epidemiology*, 145(10), 945–956.

McKeague, I. W. & Sasieni, P. D. (1994). A partly parametric additive risk model. *Biometrika*, 81(3), 501–514.

Monaco, J. V., Gorfine, M., & Hsu, L. (2017). frailtySurv: General Semiparametric Shared Frailty Model. *R package version*, 1(2).

Murphy, S. A. (1994). Consistency in a Proportional Hazards Model Incorporating a Random Effect. *The Annals of Statistics*, 22, 712–31.

Murphy, S. A. (1995). Asymptotic theory for the frailty model. *The annals of statistics*, 23, 182–198.

Nielsen, G. G., Gill, R. D., Andersen, P. K., & Sørensen, T. I. (1992). A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian journal of Statistics*, 19, 25–43.

Pakes, A. & Pollard, D. (1989). Simulation and the asymptotics of optimization estimators. *Econometrica: Journal of the Econometric Society*, 57, 1027–1057.

Parner, E. (1998). Asymptotic theory for the correlated gamma-frailty model. *The Annals of Statistics*, 26(1), 183–214.

Sherman, R. P. (1994). Maximal inequalities for degenerate U-processes with applications to optimization estimators. *The Annals of Statistics*, 22, 439–459.

Spiekerman, C. F. & Lin, D. Y. (1998). Marginal regression models for multivariate failure time data. *Journal of the American Statistical Association*, 93(443), 1164–1175.

Vaupel, J. W., Manton, K. G., & Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16(3), 439–454.

Yin, G. & Cai, J. (2004). Additive hazards model with multivariate failure time data. *Biometrika*, 91(4), 801–818.

Yin, G., Li, H., & Zeng, D. (2008). Partially linear additive hazards regression with varying coefficients. *Journal of the American Statistical Association*, 103(483), 1200–1213.

Yu, Z. & Lin, X. (2010). Semiparametric regression with time-dependent coefficients for failure time data analysis. *Statistica Sinica*, 20(2), 853–869.

## APPENDIX

**Assumptions for the proposed model:**

A1). The true parameter $H(t) = \Lambda(t) + G(t;\theta_0)$ is increasing, absolutely continuous and satisfies $H(\tau) < \infty$.

A2). $\mathbf{\Sigma}(t)$ is positive definite for all $t \in [0,\tau]$ and $E[\sum_{j=1}^{n_i} Y_{ij}(t)] > 0$.

A3). $\beta(t)$ has two order derivatives.

**Assumptions for the kernel function:**

B1). The kernel function $K(\cdot)$ is positive and has a bounded, symmetric density with a compact bounded support, say $[-1,1]$.

B2). $h \to 0, nh \to \infty, nh^5 = O(1)$.

**Assumptions for clustered failure time data:**

C1). There exists at least one $j \in \{1,2,\cdots,n_i\}$, such that $P(C_{ij} \geq \tau) > 0$.

C2). Cluster size $n_i$ is bounded and independent of $\boldsymbol{X}_{ij}(t)$, $\boldsymbol{Z}_{ij}$, $Y_{ij}(t)$ and $\Delta_{ij}(t)$, where $t \in [0,\tau]$. And $(C_{i1}, C_{i2}, \cdots, C_{in_i})$ are independent of $(T^*_{i1}, T^*_{i2}, \cdots, T^*_{in_i})$ and $\xi_i$ given all the covariates.

C3). The cluster specific random effect $\xi_i$ is independent of covariates $\boldsymbol{X}_{ij}(t)$ and $\boldsymbol{Z}_{ij}$, where $t \in [0,\tau]$.

C4). $T^*_{i1}, T^*_{i2}, \cdots, T^*_{in_i}$ are conditionally independent given all the covariates $\boldsymbol{X}_{ij}(t)$ and $\boldsymbol{Z}_{ij}$ ($t \in [0,\tau]$) and the random effect $\xi_i$.

C5). $\xi_i$ follows a single parameter distribution with density function $f(x;\theta)$ which has mean 0 and a finite moment generating function.

C6). $\boldsymbol{X}_{ij}(t)$ and $\boldsymbol{Z}_{ij}$ are locally bounded predictable processes.

**Lemma 1.** *For any $t \in [0,\tau]$, let*

$$\boldsymbol{e}_n(t) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n_i} Y_{ij}(t)g(t,\boldsymbol{X}_{ij}(t),\boldsymbol{Z}_{ij}), \boldsymbol{e}(t) = E[\sum_{j=1}^{n_i}(P(t|\boldsymbol{X}_{ij}(t),\boldsymbol{Z}_{ij})g(t,\boldsymbol{X}_{ij}(t),\boldsymbol{Z}_{ij}))].$$

*Define $P(t|\boldsymbol{X}_{ij}(t),\boldsymbol{Z}_{ij}) = P(T_{ij} > t|\boldsymbol{X}_{ij},\boldsymbol{Z}_{ij})$, assume $\mathcal{G} = \{Y_{ij}(t)g(t,\boldsymbol{X}_{ij}(t),\boldsymbol{Z}_{ij}) : t \in [0,\tau]\}$ is an Euclidean class with a constant envelop, and further define $\vartheta(t_0,\varepsilon)$ as the $\varepsilon$-ball of $t_0$ for $\varepsilon > 0$ small enough, $t_0 \in [0,\tau]$, then we have*

$$\sup_{t\in\vartheta(t_0,\varepsilon)} |\boldsymbol{e}_n(t) - \boldsymbol{e}(t)| = O_p(n^{-1/2}).$$

Before prove our theorems, we firstly present a lemma. The proof of the following lemma is an analogue to Corollary 7 in Sherman (1994), and a similar result can be seen in Lemma 1 of Cai & Sun (2003). In our paper, before we state the lemma, we assume that the readers are familiar with the notions of Euclidean classes, as defined in Pakes & Pollard (1989):

*Proof of Theorem 1.* In this part we will prove Theorem 1. For the sake of convenience, we first give some notations:

$$\boldsymbol{\mu}_1(t) = E[\sum_{j=1}^{n_i} Y_{ij}(t)\boldsymbol{X}_{ij}(t)]/E[\sum_{j=1}^{n_i} Y_{ij}(t)], \boldsymbol{\mu}_2(t) = E[\sum_{j=1}^{n_i} Y_{ij}(t)\boldsymbol{Z}_{ij}]/E[\sum_{j=1}^{n_i} Y_{ij}(t)].$$

$\boldsymbol{V}_{ij}(t) = [\boldsymbol{X}_{ij}(t);\boldsymbol{Z}_{ij};(\boldsymbol{X}_{ij}(t)(t-t_0))]$, $\bar{\boldsymbol{V}}(t) = [\bar{\boldsymbol{X}}(t);\bar{\boldsymbol{Z}}(t);(\bar{\boldsymbol{X}}(t)(t-t_0))]$, $\boldsymbol{\phi}(t) = [\boldsymbol{\mu}_1(t);\boldsymbol{\mu}_2(t);(\boldsymbol{\mu}_1(t)(t-t_0))]$ as the limit of $\bar{\boldsymbol{V}}(t)$, $\boldsymbol{G}_{ij}(t) = \boldsymbol{M}^{-1}\boldsymbol{V}_{ij}(t)$, $\bar{\boldsymbol{G}}(t) = \boldsymbol{M}^{-1}\bar{\boldsymbol{V}}(t)$, $\bar{\boldsymbol{G}}^*(t) = \boldsymbol{M}^{-1}\boldsymbol{\phi}(t)$. Meanwhile, some other notations have been given in Section 2.2. By Lemma 6, $\sup_{t\in\vartheta(t_0,\varepsilon)} |\bar{\boldsymbol{V}}(t) - \boldsymbol{\phi}(t)| = O_p(n^{-1/2})$.

First, define

$$\begin{aligned} U_n(\zeta(t_0)) &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\int_0^\tau K_h(t-t_0)Y_{ij}(t)\begin{bmatrix}\boldsymbol{X}_{ij}(t)-\bar{\boldsymbol{X}}(t)\\ \boldsymbol{Z}_{ij}-\bar{\boldsymbol{Z}}(t)\\ (\boldsymbol{X}_{ij}(t)-\bar{\boldsymbol{X}}(t))(t-t_0)\end{bmatrix} \\ &\quad\times\{dN_{ij}(t) - dH(t) - \boldsymbol{Z}_{ij}^\top\gamma dt - \boldsymbol{X}_{ij}(t)^\top\beta(t_0)dt - \boldsymbol{X}_{ij}(t)^\top(t-t_0)\beta'(t_0)dt\} \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\int_0^\tau K_h(t-t_0)Y_{ij}(t)[\boldsymbol{V}_{ij}(t)-\bar{\boldsymbol{V}}(t)]\{dN_{ij}(t) - dH(t) - \boldsymbol{V}_{ij}(t)^\top\zeta(t_0)dt\}, \end{aligned}$$

for an arbitrary $\zeta(t_0)$. Since $U_n(\hat{\zeta}(t_0)) = 0$, we have

$$\boldsymbol{M}^{-1}U_n(\zeta(t_0)) = \boldsymbol{D}_n(t_0)\boldsymbol{M}[\hat{\zeta}(t_0) - \zeta(t_0)] = \boldsymbol{D}(t_0)\boldsymbol{M}[\hat{\zeta}(t_0) - \zeta(t_0)] + O_p(n^{-1/2}), \quad (1)$$

where

$$\boldsymbol{D}_n(t_0) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\int_0^\tau K_h(t-t_0)Y_{ij}(t)[\boldsymbol{G}_{ij}(t) - \bar{\boldsymbol{G}}(t)]^{\otimes 2}dt.$$

We will explain the limit $\boldsymbol{D}(t_0)$ of $\boldsymbol{D}_n(t_0)$ later. Similarly,

$$\begin{aligned} U_n(\zeta_0(t_0)) &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\int_0^\tau K_h(t-t_0)Y_{ij}(t)[\boldsymbol{V}_{ij}(t)-\bar{\boldsymbol{V}}(t)] \\ &\times \{dN_{ij}(t)-dH(t)-\boldsymbol{V}_{ij}(t)^\top\zeta_0(t_0)dt\} \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\int_0^\tau K_h(t-t_0)Y_{ij}(t)[\boldsymbol{V}_{ij}(t)-\bar{\boldsymbol{V}}(t)] \\ &\times \{dM_{ij}(t)+[\boldsymbol{X}_{ij}(t)^\top\beta_0(t)dt+\boldsymbol{Z}_{ij}^\top\gamma_0 dt-\boldsymbol{V}_{ij}(t)^\top\zeta_0(t_0)dt]\}. \end{aligned}$$

Combining Assumption A3), by Taylor expansion:

$$\boldsymbol{X}_{ij}(t)^\top\beta_0(t)+\boldsymbol{Z}_{ij}^\top\gamma_0 = \boldsymbol{V}_{ij}(t)^\top\zeta_0(t_0)+\boldsymbol{X}_{ij}(t)^\top\frac{1}{2}\boldsymbol{\beta}_0''(t_0)(t-t_0)^2+o_p((t-t_0)^2).$$

It follows that

$$\begin{aligned} &\boldsymbol{M}^{-1}U_n(\zeta_0(t_0)) \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\int_0^\tau K_h(t-t_0)Y_{ij}(t)[\boldsymbol{G}_{ij}(t)-\bar{\boldsymbol{G}}(t)]dM_{ij}(t) \\ &\quad+\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\int_0^\tau K_h(t-t_0)Y_{ij}(t)[\boldsymbol{G}_{ij}(t)-\bar{\boldsymbol{G}}(t)][\frac{1}{2}\boldsymbol{X}_{ij}(t)^\top\boldsymbol{\beta}_0''(t_0)(t-t_0)^2+o_p((t-t_0)^2)]dt \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\int_0^\tau K_h(t-t_0)Y_{ij}(t)[\boldsymbol{G}_{ij}(t)-\bar{\boldsymbol{G}}(t)]dM_{ij}(t) \\ &\quad+\frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\int_0^\tau K_h(t-t_0)Y_{ij}(t)[\boldsymbol{G}_{ij}(t)-\bar{\boldsymbol{G}}(t)]\boldsymbol{X}_{ij}(t)^\top\boldsymbol{\beta}_0''(t_0)(t-t_0)^2dt+o_p(h^2) \qquad (2) \\ &\triangleq \boldsymbol{A}_n(\tau,t_0)+\boldsymbol{B}_n(\tau,t_0)+o_p(h^2). \end{aligned}$$

Then $\sqrt{nh}\boldsymbol{A}_n(\tau,t_0)$ is a sum of local square-integrable martingales with the quadratic variation process given by

$$\begin{aligned} &nh\langle\boldsymbol{A}_n,\boldsymbol{A}_n\rangle(\tau,t_0) \\ &= \frac{h}{n}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\int_0^\tau K_h^2(t-t_0)Y_{ij}(t)[\boldsymbol{G}_{ij}(t)-\bar{\boldsymbol{G}}(t)]^{\otimes 2}a_{ij}(t|\boldsymbol{X}_{ij},\boldsymbol{Z}_{ij})dt. \end{aligned}$$

Then

$$
\begin{aligned}
&|nh\langle \boldsymbol{A}_n, \boldsymbol{A}_n\rangle(\tau, t_0) - \boldsymbol{\Sigma}(t_0)| \\
\leq &\left| nh\langle \boldsymbol{A}_n, \boldsymbol{A}_n\rangle(\tau, t_0) - \frac{h}{n}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\int_0^\tau K_h^2(t-t_0)Y_{ij}(t)[\boldsymbol{G}_{ij}(t) - \bar{\boldsymbol{G}}^*(t)]^{\otimes 2}a_{ij}(t|\boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij})dt \right| \\
&+ \left| \frac{h}{n}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\int_0^\tau K_h^2(t-t_0)Y_{ij}(t)[\boldsymbol{G}_{ij}(t) - \bar{\boldsymbol{G}}^*(t)]^{\otimes 2}a_{ij}(t|\boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij})dt - \boldsymbol{\Sigma}(t_0) \right|.
\end{aligned}
$$

By Lemma 6 and Assumption B2), the last expression is of order $o_p(1)$. Then utilize Theorem 5.1.1 in Flemming & Harrington (1991) we have,

$$\sqrt{nh}\boldsymbol{A}_n(\tau, t_0) \xrightarrow{\mathcal{D}} N(0, \boldsymbol{\Sigma}(t_0)).$$

In the following we will derive the limit of $\boldsymbol{B}_n(\tau, t_0)$ and $\boldsymbol{D}_n(t_0)$. Through simple calculation, under Assumption B2), we have

$$
\begin{aligned}
&\frac{1}{h^2}\boldsymbol{B}_n(\tau, t_0) \\
\to &\begin{bmatrix} E[\sum_{j=1}^{n_i} Y_{ij}(t_0)(\boldsymbol{X}_{ij}(t_0) - \bar{\boldsymbol{X}}(t_0))\boldsymbol{X}_{ij}(t_0)^\top] \\ E[\sum_{j=1}^{n_i} Y_{ij}(t_0)(\boldsymbol{Z}_{ij} - \bar{\boldsymbol{Z}}(t_0))\boldsymbol{Z}_{ij}^\top] \\ E[\sum_{j=1}^{n_i} Y_{ij}(t_0)(\boldsymbol{X}_{ij}(t_0) - \bar{\boldsymbol{X}}(t_0))\boldsymbol{X}_{ij}(t_0)^\top]h \end{bmatrix} \frac{1}{2}\kappa_{12}\boldsymbol{\beta}_0''(t_0) \\
= &\begin{bmatrix} \boldsymbol{c}^{(11)}(t_0) \\ \boldsymbol{c}^{(22)}(t_0) \\ \boldsymbol{0}_p \end{bmatrix} \frac{1}{2}\kappa_{12}\boldsymbol{\beta}_0''(t_0) \\
\equiv &\frac{1}{2}\kappa_{12}\boldsymbol{b}(t_0),
\end{aligned}
$$

and

$$
\begin{aligned}
&\boldsymbol{D}_n(t_0) \\
\to &\begin{bmatrix} \boldsymbol{c}^{(11)}(t_0)\kappa_{10}\ \boldsymbol{c}^{(12)}(t_0)\kappa_{10}\ \boldsymbol{c}^{(11)}(t_0)\kappa_{11} \\ \boldsymbol{c}^{(21)}(t_0)\kappa_{10}\ \boldsymbol{c}^{(22)}(t_0)\kappa_{10}\ \boldsymbol{c}^{(21)}(t_0)\kappa_{11} \\ \boldsymbol{c}^{(11)}(t_0)\kappa_{11}\ \boldsymbol{c}^{(12)}(t_0)\kappa_{11}\ \boldsymbol{c}^{(11)}(t_0)\kappa_{12} \end{bmatrix} \\
\equiv &\boldsymbol{D}(t_0).
\end{aligned}
$$

Thus,

$$\sqrt{nh}\{\boldsymbol{M}(\hat{\zeta}(t_0) - \zeta_0(t_0)) - \frac{1}{2}h^2\kappa_{12}\boldsymbol{D}^{-1}(t_0)\boldsymbol{b}(t_0)\} \xrightarrow{\mathcal{D}} N(0, \boldsymbol{D}^{-1}(t_0)\boldsymbol{\Sigma}(t_0)\boldsymbol{D}^{-1}(t_0)).$$

In addition, since $\kappa_{11} = 0$, we can obtain $[\boldsymbol{D}(t_0)]^{-1}\boldsymbol{b}(t_0) = [\beta_0''(t_0); \boldsymbol{0}_{p+r}]$. □

*Proof of Theorem 2*. In the following we are ready to show the proof of Theorem 2. First, combining Equations (1)-(2), we know that

$$\begin{aligned}
&\sqrt{nh}\boldsymbol{M}[\hat{\zeta}(t_0)-\zeta_0(t_0)]\\
&=\sqrt{nh}[\boldsymbol{D}_n(t_0)]^{-1}\boldsymbol{M}^{-1}U_n(\zeta_0(t_0))\\
&=\sqrt{nh}[\boldsymbol{D}_n(t_0)]^{-1}\boldsymbol{A}_n(\tau,t_0)+\sqrt{nh}[\boldsymbol{D}_n(t_0)]^{-1}\boldsymbol{B}_n(\tau,t_0)+o_p(\sqrt{nh^5})\\
&=[\boldsymbol{D}_n(t_0)]^{-1}\sqrt{\frac{h}{n}}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\int_0^{\tau}K_h(t-t_0)Y_{ij}(t)[\boldsymbol{G}_{ij}(t)-\bar{\boldsymbol{G}}(t)]dM_{ij}(t)\\
&\quad+\sqrt{nh^5}[\boldsymbol{D}(t_0)]^{-1}\boldsymbol{b}(t_0)+O_p(h^2)+o_p(\sqrt{nh^5}).
\end{aligned}$$

Let $\boldsymbol{\mathcal{I}}=[\mathcal{I}_{kl}]$ be a $r\times(2p+r)$ matrix with $\mathcal{I}_{kl}=1$ for $k=1,\cdots,r$ and $l=p+1,\cdots,p+r$, then

$$\begin{aligned}
&\sqrt{nh}(\hat{\gamma}(t_0)-\gamma_0)\\
&=\boldsymbol{\mathcal{I}}[\boldsymbol{D}_n(t_0)]^{-1}\sqrt{\frac{h}{n}}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\int_0^{\tau}K_h(t-t_0)Y_{ij}(t)[\boldsymbol{G}_{ij}(t)-\bar{\boldsymbol{G}}(t)]dM_{ij}(t)+o_p(\sqrt{nh^5}).
\end{aligned}$$

Taking integral on each side of the last equation by the same rule with Equation (17), we have

$$\begin{aligned}
&\sqrt{n}(\tilde{\gamma}-\gamma_0)\\
&=\sqrt{n}[\int_0^{\tau}\Gamma(t_0)\hat{\gamma}(t_0)dt_0-\gamma_0]\\
&=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\int_0^{\tau}\int_0^{\tau}\Gamma(t_0)K_h(t-t_0)Y_{ij}(t)\boldsymbol{\mathcal{I}}[\boldsymbol{D}_n(t_0)]^{-1}[\boldsymbol{G}_{ij}(t)-\bar{\boldsymbol{G}}(t)]dt_0dM_{ij}(t)+o_p(\sqrt{nh^4}).
\end{aligned}$$

By Lemma 6 again, we get

$$\begin{aligned}
&\sqrt{n}(\tilde{\gamma}-\gamma_0)\\
&=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\int_0^{\tau}\int_0^{\tau}\Gamma(t_0)K_h(t-t_0)Y_{ij}(t)\boldsymbol{\mathcal{I}}[\boldsymbol{D}(t_0)]^{-1}[\boldsymbol{G}_{ij}(t)-\bar{\boldsymbol{G}}^*(t)]dt_0dM_{ij}(t)+o_p(1),
\end{aligned}$$

and

$$\begin{aligned}
&\int_0^{\tau}\Gamma(t_0)K_h(t-t_0)Y_{ij}(t)\boldsymbol{\mathcal{I}}[\boldsymbol{D}(t_0)]^{-1}[\boldsymbol{G}_{ij}(t)-\bar{\boldsymbol{G}}^*(t)]dt_0\\
&=\Gamma(t)Y_{ij}(t)\boldsymbol{\mathcal{I}}[\boldsymbol{D}(t)]^{-1}\begin{pmatrix}\boldsymbol{X}_{ij}(t)-\boldsymbol{\mu}_1(t)\\ \boldsymbol{Z}_{ij}-\boldsymbol{\mu}_2(t)\\ \mathbf{0}_p\end{pmatrix}+O_p(n^{-1/2}).
\end{aligned}$$

Thus,

$$\sqrt{n}(\tilde{\gamma}-\gamma_0)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\int_0^{\tau}\Gamma(t)Y_{ij}(t)\boldsymbol{\mathcal{I}}[\boldsymbol{D}(t)]^{-1}\begin{pmatrix}\boldsymbol{X}_{ij}(t)-\boldsymbol{\mu}_1(t)\\ \boldsymbol{Z}_{ij}-\boldsymbol{\mu}_2(t)\\ \mathbf{0}_p\end{pmatrix}dM_{ij}(t)+o_p(1).$$

By martingale central limit theorem, the proof is complete. Denote the variance function as $\boldsymbol{\Sigma}_1$, then the variance function has the form

$$\boldsymbol{\Sigma}_1 = E\left[\sum_{j=1}^{n_i}\int_0^\tau Y_{ij}(t)\Gamma(t)\boldsymbol{\mathcal{I}}[\boldsymbol{D}(t)]^{-1}\begin{pmatrix}\boldsymbol{X}_{ij}(t)-\boldsymbol{\mu}_1(t)\\ \boldsymbol{Z}_{ij}-\boldsymbol{\mu}_2(t)\\ \mathbf{0}_p\end{pmatrix}^{\otimes 2}[\boldsymbol{D}(t)]^{-1}\boldsymbol{\mathcal{I}}^\top[\Gamma(t)]^\top a_{ij}(t|\boldsymbol{X}_{ij},\boldsymbol{Z}_{ij})dt\right].$$

□

*Proof of Theorem 3.* In light of Theorem 1, we only know that $\hat{\gamma}(t) \xrightarrow{\mathcal{P}} \gamma_0(t)$ and $\hat{\beta}(t) \xrightarrow{P} \beta_0(t)$ for all $t \in (0,\tau)$, so it's not enough to establish the results of Theorem 2. However, similar to Cai & Sun (2003), we can prove that the bias around the boundary point for $\hat{\beta}(t)$ is still of order $O(h^2)$ , to be specific, define $\kappa_{ij}^* = \int_{-c}^1 K(u)^i u^j du$, then for $0 < c < 1$

$$\sqrt{nh}[\hat{\beta}(c\cdot h) - \beta_0(c\cdot h) - \frac{1}{2}h^2\kappa_{12}^*\beta_0''(0+) - h\kappa_{11}^*\beta_0'(0+)] = O_p(1).$$

Combining Theorem 3 and Equation (3), it is natural to show that $\hat{\beta}(t) \xrightarrow{P} \beta(t)$ and $\hat{\gamma}(t) \xrightarrow{P} \gamma$ uniformly for all $t \in [0,\tau]$. Then from Equations (7), (9), we can know $d\hat{H}(t) \xrightarrow{P} dH(t)$ for all $t \in [0,\tau]$. □

*Proof of Theorem 4.* Following the proof of the Theorem 2, we could conclude that

$$\{\hat{\beta}(t), \hat{\gamma}(t), \hat{H}(t)\} \xrightarrow{\mathcal{P}} \{\beta(t), \gamma(t), H(t)\} \quad \text{uniformly.}$$

Then, similar to the proof of Theorem 1 in Cai & Zeng (2011), it is easy to draw the conclusion. □

*Received 22 February 2020*
*Accepted 3 October 2020*