BIOMETRICS Accepted Manuscript, 1–24 DOI: 10.1111/biom.13399

October 2020

Analysis of Clustered Interval-Censored Data using a Class of Semiparametric Partly Linear Frailty Transformation Models

Chun Yin Lee[1], Kin Yau Wong[1], K. F. Lam[2,*], and Jinfeng Xu[2]

[1]Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong

[2]Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong

[*]*email*:hrntlkf@hku.hk

SUMMARY: A flexible class of semiparametric partly linear frailty transformation models is considered for analyzing clustered interval-censored data, which arise naturally in complex diseases and dental research. This class of models features two nonparametric components, resulting in a nonparametric baseline survival function and a potential nonlinear effect of a continuous covariate. The dependence among failure times within a cluster is induced by a shared, unobserved frailty term. A sieve maximum likelihood estimation method based on piecewise linear functions is proposed. The proposed estimators of the regression, dependence and transformation parameters are shown to be strongly consistent and asymptotically normal, whereas the estimators of the two nonparametric functions are strongly consistent with optimal rates of convergence. An extensive simulation study is conducted to study the finite-sample performance of the proposed estimators. We provide an application to a dental study for illustration.

KEY WORDS: clustered data, nonparametric estimation, partly linear model, random effects model, sieve maximum likelihood estimation.

1. Introduction

Interval-censored failure time data arise commonly in biomedical or epidemiological research, where the event of interest cannot be observed directly but is only known to have occurred within a time interval. For example, in the study of human immunodeficiency virus infection and acquired immune deficiency syndrome (HIV/AIDS), the development of clinical symptoms usually takes eight to ten years since the infection of HIV. In practice, patients are followed up on a regular basis, and the exact time to onset of clinical symptoms cannot be determined but is known to fall within two consecutive follow-up times. There are many existing studies on HIV/AIDS employing various analysis methods for interval-censored data (Betensky *et al*., 2001; Xue *et al*., 2006).

Failure time data are often clustered. For example, in studies of time to tooth emergence/loss, the event times for the teeth of an individual are clustered as the teeth share the same oral condition. Frailty models are commonly used in the literature to accommodate the association among members of the same cluster. A review of frailty models can be found in Hougaard (2000). Cook *et al*. (2008) considered a four-state Markov model for characterizing the association of interval-censored failure times in the bivariate case, while Kim (2006), Zuma *et al*. (2007) and Lam *et al*. (2010) proposed methods for estimating the parameter of interest based on the log-normal or gamma frailty Cox models; the theoretical properties of the estimators were not studied. Zeng *et al*. (2017) proposed a semiparametric transformation model with normal random effects for multivariate interval-censored data. They showed that the proposed nonparametric maximum likelihood estimators of the regression and dependence parameters are consistent and asymptotically efficient. Zhou *et al*. (2017) considered a semiparametric frailty transformation model for bivariate interval-censored failure time data and proposed a sieve estimation method based on the Bernstein polynomials. The asymptotic properties of the estimators of the regression and dependence parameters were established. In the literature on transformation models, the transformation parameter, usually estimated using grid search, is typically treated as known, and inference on the regression

and frailty parameters is made without accounting for the variability involved in the estimation of the transformation parameter. Methods that properly account for this variability and allow valid inference for the transformation parameter are needed.

Generalized partly linear model has received increasing attention recently. As advocated by Lin and Carroll (2001), one can make inference on the effects of some covariates of interest $\boldsymbol{X}$ (e.g., treatment effect) while making minimal assumptions on the effects of other covariates Z (e.g., age, which is known to be an important factor but may not be of major interest) using nonparametric functions. Indeed, nonlinear covariate effects are common in practice. For example, in clinical studies, the effect of the amount of dosage of a certain medication on the time to reaction may attain its maximum at some dosage level and then retain at the maximum level or decreases beyond the dosage level.

This work is motivated by a longitudinal dental study, where one is interested in the association between the emergence of permanent teeth and various covariates among children. In the study, tooth emergence was not directly observed but was only known to occur between two consecutive dental visits. The times to emergence of teeth of a child are naturally clustered, and certain covariates are expected to exhibit strong and nonlinear effects. Our goal is to develop flexible models with reliable estimation and inference methods that accommodate the special features of the data.

In this article, we consider a class of semiparametric partly linear transformation models for the analysis of clustered interval-censored failure time data. The model involves a nonparametric baseline function that characterizes the cumulative hazard function for a subject at baseline and a nonparametric function for the effect of a continuous covariate. The dependence among the failure times within a cluster is induced by an unobserved frailty. A sieve maximum likelihood estimation method is proposed, where the nonparametric functions are approximated by piecewise linear functions. The proposed estimation method is applicable to any frailty distributions with an explicit Laplace transform, including the gamma and positive stable distributions.

We structure the article as follows. The model specifications and estimation method are described in Section 2. Theoretical properties of the estimators are presented in Section 3. The computational details are provided in Section 4. The finite-sample performance of the estimators is investigated through a large-scale simulation study, and the results are reported in Section 5. The proposed method is applied to the aforementioned dental study in Section 6. Lastly, some concluding remarks are made in Section 7. Proofs of the theoretical results are given in the Appendix.

2. Model, Likelihood, and Sieve Estimation

Suppose that there are n independent and identically distributed clusters with N_i subjects in the ith cluster for $i = 1, \ldots, n$, where N_i is possibly random. For the jth subject in the ith cluster ($j = 1, \ldots, N_i$), let T_{ij} denote the failure time, $\boldsymbol{X}_{ij}$ denote a p-dimensional vector of covariates, and Z_{ij} denote a continuous covariate with a possibly nonlinear effect on the response variable. Assume that N_i is independent of the failure times given $(\boldsymbol{X}_{ij}, Z_{ij})$. Let ξ_i denote an unobserved frailty shared among all subjects in the ith cluster. Conditional on $(\boldsymbol{X}_{ij}, Z_{ij}, \xi_i)$, $T_{i1}, \ldots, T_{i,N_i}$ are mutually independent, and T_{ij} has cumulative hazard function

$$\Lambda_{ij}(t \mid \boldsymbol{X}_{ij}, Z_{ij}, \xi_i) = \xi_i G[\Lambda(t) \exp\{\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + g(Z_{ij})\}; \rho], \tag{1}$$

where $G(\cdot; \rho)$ is an increasing parametric function indexed by the parameter ρ, $\Lambda(\cdot)$ is an unspecified increasing function, $g(\cdot)$ is an unspecified smooth function, and $\boldsymbol{\beta}$ is a vector of regression parameters. We assume that ξ_i follows a parametric distribution with a Laplace transform $\Phi(u; \gamma) \equiv \mathrm{E}(e^{-\xi_i u})$, where γ is the parameter of the frailty distribution. The frailty terms ξ_i's characterize the heterogeneity across clusters. Two commonly used frailty distributions, namely the gamma distribution Ga(γ^{-1}, γ^{-1}) and the positive stable distribution Po(γ), are considered (Lam and Kuk, 1997; Kosorok *et al.*, 2004). Their respective Laplace transforms are given by $\Phi(u; \gamma) = (1 + \gamma u)^{-1/\gamma}$ for $\gamma > 0$ and $\Phi(u; \gamma) = \exp(-u^{\gamma})$ for $0 < \gamma < 1$. Zeng and Lin (2007) considered two transformation functions G, namely the Box-Cox transformation and the logarithmic transformation, both with a single transformation parameter. In this article, we focus

on the Box-Cox transformation, which takes the form $G(x;\rho) = \{(1+x)^\rho - 1\}/\rho$ for $\rho \geqslant 0$. Note that $\lim_{\rho\to 0} G(x;\rho) = \log(1+x)$ and $G(x;1) = x$ correspond to the proportional odds model (Rossini and Tsiatis, 1996; Lam and Leung, 2001) and the Cox proportional hazards model (Cox, 1972; Andersen and Gill, 1982), respectively, in the absence of the frailty.

Under the proposed model, $\log\Lambda(T_{ij}) = -\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} - g(Z_{ij}) + \log\{G^{-1}(\epsilon_{ij};\rho)\}$, where ϵ_{ij} is a random variable that follows the exponential distribution with rate parameter ξ_i conditional on ξ_i ($j = 1,\ldots,N_i$; $i = 1,\ldots,n$). Thus, the covariates (besides Z_{ij}) act additively on a monotone, nonparametric transformation of the failure time as in the Cox model (Cheng *et al.*, 1995). Different choices of the transformation function or frailty distribution yield different distributions or association structures of the error terms $(\log\{G^{-1}(\epsilon_{i1};\rho)\},\ldots,\log\{G^{-1}(\epsilon_{iN_i};\rho)\})$ but does not alter the interpretation of the covariate effects.

Suppose that the failure time T_{ij} is not observed directly but is only known to fall within the interval $(L_{ij}, R_{ij}]$ for some $R_{ij} > L_{ij}$. For left-censored or right-censored subjects, we set $L_{ij} = 0$ or $R_{ij} = \infty$, respectively. Let $\boldsymbol{\theta} \equiv (\boldsymbol{\beta},\gamma,\rho,\Lambda,g)$ denote the set of all parameters. The log-likelihood function of $\boldsymbol{\theta}$ is

$$\begin{aligned}\ell_n(\boldsymbol{\theta}) &= \sum_{i=1}^{n}\log\int\prod_{j=1}^{N_i}\left[\exp\{-\xi_i S_{ij}(L_{ij};\boldsymbol{\theta})\} - \exp\{-\xi_i S_{ij}(R_{ij};\boldsymbol{\theta})\}\right] f_\xi(\xi_i;\gamma)\,\mathrm{d}\xi_i \\ &= \sum_{i=1}^{n}\log\left(\sum_{\boldsymbol{r}\in\mathcal{S}_{N_i}}(-1)^{|\boldsymbol{r}|}\Phi\left[\sum_{j=1}^{N_i}\left\{(1-r_j)S_{ij}(L_{ij};\boldsymbol{\theta}) + r_j S_{ij}(R_{ij};\boldsymbol{\theta})\right\}\right]\right),\end{aligned}$$

where $f_\xi(\cdot;\gamma)$ is the density of the frailty, r_j is the jth component of $\boldsymbol{r}$, $\mathcal{S}_{N_i} = \{0,1\}^{N_i}$, and $S_{ij}(t;\boldsymbol{\theta}) = G[\Lambda(t)\exp\{\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + g(Z_{ij})\};\rho]$. For example, if $\xi_i \sim \mathrm{Ga}(\gamma^{-1},\gamma^{-1})$ and $N_i = 2$ ($i = 1,\ldots,n$), then the log-likelihood contribution of the ith cluster is

$$\begin{aligned}\log\Big(&[1+\gamma\{S_{i1}(L_{i1};\boldsymbol{\theta}) + S_{i2}(L_{i2};\boldsymbol{\theta})\}]^{-1/\gamma} - [1+\gamma\{S_{i1}(L_{i1};\boldsymbol{\theta}) + S_{i2}(R_{i2};\boldsymbol{\theta})\}]^{-1/\gamma} \\ &- [1+\gamma\{S_{i1}(R_{i1};\boldsymbol{\theta}) + S_{i2}(L_{i2};\boldsymbol{\theta})\}]^{-1/\gamma} + [1+\gamma\{S_{i1}(R_{i1};\boldsymbol{\theta}) + S_{i2}(R_{i2};\boldsymbol{\theta})\}]^{-1/\gamma}\Big).\end{aligned}$$

Because the likelihood involves two nonparametric functions, namely Λ and g, maximum like-

lihood estimation of $\boldsymbol{\theta}$ is not feasible. We propose a sieve maximum likelihood approach and estimate Λ and g using piecewise linear functions. Let $[a_1, b_1]$ be the support of L_{ij} and R_{ij}, $[a_2, b_2]$ be the support of Z_{ij}. Let $\boldsymbol{\tau} \equiv (\tau_0, \ldots, \tau_{m_{1n}})$ be a set of grid points over $[a_1, b_1]$ and $\boldsymbol{\varsigma} \equiv (\varsigma_0, \ldots, \varsigma_{m_{2n}})$ be a set of grid points over $[a_2, b_2]$, where $a_1 = \tau_0 < \cdots < \tau_{m_{1n}} = b_1$, and $a_2 = \varsigma_0 < \cdots < \varsigma_{m_{2n}} = b_2$. Suppose that $m_{1n} = O(n^{\nu_1})$ and $m_{2n} = O(n^{\nu_2})$ for some fixed $\nu_1, \nu_2 \in (0, 1)$, and the grid points are chosen such that $K^{-1}n^{-\nu_1} < \tau_j - \tau_{j-1} < Kn^{-\nu_1}$ for $j = 1, \ldots, m_{1n}$ and $K^{-1}n^{-\nu_2} < \varsigma_j - \varsigma_{j-1} < Kn^{-\nu_2}$ for $j = 1, \ldots, m_{2n}$, where K is some positive constant. For a given m_{1n}-dimensional vector $\boldsymbol{\omega}$, let

$$H\left(t; m_{1n}, \boldsymbol{\omega}, \boldsymbol{\tau}\right) \equiv \sum_{j=1}^{m_{1n}} \omega_j \left\{\left(t - \tau_{j-1}\right) I\left(\tau_{j-1} \leqslant t < \tau_j\right) + \left(\tau_j - \tau_{j-1}\right) I\left(t \geqslant \tau_j\right)\right\}$$

be a piecewise linear function over the grid points $\boldsymbol{\tau}$. We define the sieve spaces for Λ and g respectively as $\mathcal{A}_{\Lambda n} = \{\Lambda_n(t) = H(t; m_{1n}, \boldsymbol{\omega}, \boldsymbol{\tau}), t \in [a_1, b_1],\ \omega_j \geqslant 0 \text{ for } j = 1, \ldots, m_{1n}\}$ and $\mathcal{A}_{gn} = \{g_n(z) = H(z; m_{2n}, \boldsymbol{\psi}, \boldsymbol{\varsigma}), z \in [a_2, b_2]\}$. The sieve maximum likelihood estimator is

$$\widehat{\boldsymbol{\theta}}_n \equiv (\widehat{\boldsymbol{\beta}}_n, \widehat{\gamma}_n, \widehat{\rho}_n, \widehat{\Lambda}_n, \widehat{g}_n) = \underset{\boldsymbol{\theta}: \Lambda \in \mathcal{A}_{\Lambda n}, g \in \mathcal{A}_{gn}}{\arg\max}\ \ell_n(\boldsymbol{\theta}).$$

In Section 3, we present some regularity conditions and the asymptotic properties of the proposed sieve maximum likelihood estimator. In particular, we show that the sieve maximum likelihood estimator is consistent, estimators of the nonparametric functions attain the optimal rates of convergence, and the estimators of the Euclidean parameters are asymptotically normal with a covariance matrix that equals the inverse of the efficient information matrix. Because the efficient information matrix does not have an explicit form, we adopt the approach of Huang (1999) and Chen *et al*. (2012) to approximate the standard error of $\widehat{\boldsymbol{\zeta}}_n \equiv (\widehat{\boldsymbol{\beta}}_n, \widehat{\gamma}_n, \widehat{\rho}_n)$. In particular, we treat the finite-dimensional sieve parameter space as the true parameter space and compute the negative Hessian matrix of $\ell_n(\boldsymbol{\theta})$ with respect to $(\boldsymbol{\beta}, \gamma, \rho, \boldsymbol{\omega}, \boldsymbol{\psi})$. The variance of $\widehat{\boldsymbol{\zeta}}_n$ is then estimated by the corresponding elements of the inverse of the negative Hessian matrix. Previous extensive simulation studies (Xue *et al*., 2004; Lam and Xue, 2005) suggested that this standard error

estimation approach is computationally efficient and numerically stable even for large values of m_{1n} and m_{2n}.

3. Asymptotic Properties of the Sieve Estimators

Suppose that the interval $(L_{ij}, R_{ij}]$ is derived from a sequence of monitoring time points $U_{ij1} < \cdots < U_{ij,M_{ij}}$ that are independent of the failure times given the observed covariates, where M_{ij} is the number of monitoring times for $i = 1, \ldots, n; j = 1, \ldots, N_i$. The interval $(L_{ij}, R_{ij}]$ is the shortest time interval that brackets T_{ij}, so that $L_{ij} = \max\{U_{ijk} : U_{ijk} < T_{ij}, k = 0, \ldots, M_{ij}\}$ and $R_{ij} = \min\{U_{ijk} : U_{ijk} \geqslant T_{ij}, k = 1, \ldots, M_{ij} + 1\}$, where $U_{ij0} = 0$, and $U_{ij,M_{ij}+1} = \infty$. Let $\Delta_{ijk} = I(L_{ij} = U_{ijk})$ for $k = 0, \ldots, M_{ij}$. The log-likelihood can be written as

$$\ell_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \left[\sum_{\boldsymbol{r} \in \mathcal{S}_{N_i}} (-1)^{|\boldsymbol{r}|} \Phi \left\{ \sum_{j=1}^{N_i} \sum_{k=0}^{M_{ij}} \Delta_{ijk} S_{ij}(U_{ij,k+r_j}; \boldsymbol{\theta}) \right\} \right].$$

Let $\boldsymbol{\zeta} \equiv (\boldsymbol{\beta}, \gamma, \rho)$ denote the set of all Euclidean parameters, and $\boldsymbol{\zeta}_0 \equiv (\boldsymbol{\beta}_0, \gamma_0, \rho_0)$, Λ_0, and g_0 denote the true values of $\boldsymbol{\zeta}$, Λ, and g, respectively. In the sequel, we use $(N, M_j, U_{jk}, \Delta_{jk}, Z_j, \boldsymbol{X}_j)$ $(j = 1, \ldots, N, k = 1, \ldots, M_j)$ to denote the observed data for a generic cluster. Without loss of generality, assume that $a_2 = 0$ and $b_2 = 1$. To simplify the technical derivations, in this section and the Appendix, we adopt an alternative identifiability constraint on g and assume that $\mathrm{E}\{g(Z_1)\} = 0$ (instead of $g(0) = 0$); similar conditions are imposed in work on additive models (Stone, 1985; Huang, 1999). Let $\mathrm{BV}[c_1, c_2]$ be the space of functions with bounded total variation on $[c_1, c_2]$. We assume the following conditions.

(C1) The parameter value $\boldsymbol{\zeta}_0$ lies in the interior of a known compact set $\mathcal{A}_\zeta \subset \mathbb{R}^{p+2}$. Also, Λ_0 is strictly increasing and twice continuously differentiable on $[a_1, b_1]$. In addition, g_0 is twice continuously differentiable on $[0, 1]$.

(C2) For some positive constant C, $\Pr(\|\boldsymbol{X}_j\| + N + M_j < C) = \Pr(Z_j \in [0, 1]) = 1$ for $j = 1, 2, \ldots$. Also, the conditional density of Z_j given $\mathcal{U}$ is twice continuously differentiable and bounded between $[C^{-1}, C]$ on $[0, 1]$ almost surely, where $\mathcal{U}$ denotes the set of all monitoring times.

In addition, given (N, M_j), $\Pr(U_{j,k+1} - U_{j,k} > C^{-1}) = 1$ for $k = 1, \ldots, M_j$ and $j = 1, \ldots, N$, the density of U_{jk} is twice continuously differentiable, and the union of the support of $(U_{j1}, \ldots, U_{jM_j})$ is $[a_1, b_1]$ for $j = 1, 2, \ldots$.

(C3) For any $j = 1, 2, \ldots$, if $h_0 + \boldsymbol{h}_1^{\mathrm{T}} \boldsymbol{X}_j + h_2(Z_j) = 0$ almost surely for some $h_0 \in \mathbb{R}$, $\boldsymbol{h}_1 \in \mathbb{R}^p$, and $h_2 \in \mathrm{BV}[0, 1]$, then $h_0 = 0$, $\boldsymbol{h}_1 = \boldsymbol{0}$, and $h_2(z) = 0$ for $z \in [0, 1]$. Also, there exists a constant $\eta \in (0, 1)$ such that for any $\boldsymbol{d} \in \mathbb{R}^p$, $\boldsymbol{d}^{\mathrm{T}} \mathrm{var}(\boldsymbol{X}_j \mid \mathcal{U}, Z_j) \boldsymbol{d} \geqslant \eta \boldsymbol{d}^{\mathrm{T}} \mathrm{E}(\boldsymbol{X}_j \boldsymbol{X}_j^{\mathrm{T}} \mid \mathcal{U}, Z_j) \boldsymbol{d}$ almost surely.

(C4) The transformation function $G(\cdot; \rho)$ is strictly increasing, three-times continuously differentiable, and with $G(0; \rho) = 0$. Also, for any positive constant K,

$$\sum_{j=1}^{3} \sup_{\rho} |G^{(j)}(x; \rho)| + \sup_{\gamma} \int |f_{\xi}^{(j)}(\xi; \gamma)| \, \mathrm{d}\xi < C$$

for all $x < K$, where the supremums are taken over the parameter spaces of ρ and γ, $G^{(j)}$ is the jth derivative of the transformation function with respect to the transformation parameter, $f_{\xi}^{(j)}(\cdot; \gamma)$ is the jth derivative of the density of ξ with respect to γ, and C is some constant that depends on K only.

(C5) For any $\boldsymbol{\theta}$ in a small neighborhood around $\boldsymbol{\theta}_0$,

$$\begin{aligned} &\mathrm{E}\left(\Phi\left[\sum_{j=1}^{N} G\{\Lambda(U_{jM_j}) e^{\boldsymbol{X}_j^{\mathrm{T}} \boldsymbol{\beta} + g(Z_j)}; \rho\}; \gamma \right] - \Phi\left[\sum_{j=1}^{N} G\{\Lambda_0(U_{jM_j}) e^{\boldsymbol{X}_j^{\mathrm{T}} \boldsymbol{\beta}_0 + g_0(Z_j)}; \rho_0\}; \gamma_0 \right] \right)^2 \\ &\gtrsim \mathrm{E}\left[\left\{ \Lambda(U_{1M_1}) e^{\boldsymbol{X}_1^{\mathrm{T}} \boldsymbol{\beta} + g(Z_1)} - \Lambda_0(U_{1M_1}) e^{\boldsymbol{X}_1^{\mathrm{T}} \boldsymbol{\beta}_0 + g_0(Z_1)} \right\}^2 \right] + |\gamma - \gamma_0|^2 + |\rho - \rho_0|^2, \end{aligned}$$

where $\gtrsim$ denotes "greater than up to a scaling factor."

REMARK 1 Condition (C1) requires the true parameters to lie in the interior of a compact set. In some cases, the parameters in the transformation function and frailty distribution may lie at the boundary of a natural parameter space. For example, if $\xi \sim \mathrm{Ga}(\gamma^{-1}, \gamma^{-1})$, then $\gamma = 0$ corresponds to independence of subjects within a cluster, and this value of γ is on the boundary of the (natural) parameter space $\{\gamma : 0 \leqslant \gamma \leqslant C\}$. To handle this issue, we may follow Kosorok *et al*. (2004) and

expand the parameter space to include the parameter value of interest as an interior point. For the gamma frailty, we can set the parameter space of γ to be $[-c_0, C]$, where

$$c_0 = \frac{1}{C \sup_{\boldsymbol{X},Z,\boldsymbol{\beta},g,\Lambda,\rho} G\{\Lambda(b_1)e^{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}+g(Z)}; \rho\}},$$

and the supremum is taken over the parameter space and the support of the variables. A negative γ cannot be interpreted as the variance of a frailty, but the resulting likelihood can still be well-defined.

REMARK 2 Condition (C2) pertains to typical regularity conditions for modeling interval-censored data. Condition (C3) guarantees that the covariates are not degenerated; this condition is necessary for model identifiability. Condition (C4) imposes regularity conditions on the transformation function and the distribution of the frailty. Condition (C5) is a technical condition, which requires that changes in the transformation or frailty parameters would result in changes in the survival probability. This condition guarantees that the transformation and frailty parameters can be identified.

Let $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0, \rho_0, \gamma_0, \Lambda_0, g_0)$ and $d(\cdot, \cdot)$ be a distance function such that

$$d(\boldsymbol{\theta}, \boldsymbol{\theta}_0)^2 = \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2 + |\gamma - \gamma_0|^2 + |\rho - \rho_0|^2 + \|\Lambda - \Lambda_0\|^2_{[a_1,b_1]} + \|g - g_0\|^2_{[0,1]},$$

where $\|\cdot\|_{[c_1,c_2]}$ is the L_2-norm over the interval $[c_1, c_2]$. In the sequel, we suppress the subscript of the norms. The following theorems give the consistency and rate of convergence of the sieve maximum likelihood estimator and the asymptotic normality of the estimators of the Euclidean parameters.

THEOREM 1 Under Conditions (C1)–(C5), $d(\widehat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0)$ converges to 0 almost surely with

$$d(\widehat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) = O_p[\max\{n^{-2\nu_1}, n^{-(1-\nu_1)/2}, n^{-2\nu_2}, n^{-(1-\nu_2)/2}\}].$$

THEOREM 2 Assume that Conditions (C1)–(C5) hold and that the grid points $\boldsymbol{\tau}$ and $\boldsymbol{\varsigma}$ are chosen such that $1/8 < \nu_1 < 1/2$, $1/8 < \nu_2 < 1/4$, and $\nu_2 < \min(4\nu_1, 1 - \nu_1)/3$. We have

$$n^{1/2}(\widehat{\boldsymbol{\zeta}}_n - \boldsymbol{\zeta}_0) \xrightarrow{\boldsymbol{d}} \mathrm{N}(0, \widetilde{\boldsymbol{I}}^{-1}),$$

where $\widetilde{\boldsymbol{I}}$ is the efficient information matrix for $\boldsymbol{\zeta}$ defined in Lemma 2 of the Appendix.

4. Computation of the Sieve Estimators

We propose to adopt a gradient-based method for the computation of $\widehat{\boldsymbol{\theta}}_n$. In the objective function, the parameters $\boldsymbol{\omega}$, ρ and γ under the gamma frailty are subject to sign constraints, while the parameter γ under the positive stable frailty is restricted to $(0, 1)$. To facilitate the computation, we reparametrize the model by applying the logarithmic or logit transformation on the constrained parameters to reduce the original constrained nonlinear optimization problem to an unconstrained one. The optimization is then performed using the Broyden-Fletcher-Goldfarb-Shanno algorithm, which is implemented by the function $optim$ in `R` (V. 3.5.2).

One crucial task in the proposed estimation approach is to determine the numbers and locations of the grid points for Λ_n and g_n, which potentially affect the estimation of the Euclidean and infinite-dimensional parameters. Huang and Rossini (1997), in the context of sieve approximation of the baseline log-odds function under the proportional odds model for interval-censored data, reported that the choice of the number of grid points would affect the precision of the estimates. When m_{1n} or m_{2n} is too small, the shape of the nonparametric functions cannot be approximated closely by piecewise linear functions, while a large value of m_{1n} or m_{2n} may cause overfitting and pose substantial computational burden. Practical guidelines for the selection of the grid points are warranted.

Under the proposed methods, asymptotic normality of $\widehat{\boldsymbol{\zeta}}_n$ holds under $m_{1n} = O(n^{\nu_1})$ for $\nu_1 \in (1/8, 1/2)$, so we set $m_{1n} = C_1 n^{1/3}$ for some positive constant C_1; the choice of C_1 is discussed in Section 5. On the other hand, the number of grid points required for g_n depends greatly on the shape of g, which is very flexible due to the absence of monotonicity constraints on g. Under a fixed set of grid points for Λ_n, we adopt the following algorithm to select the grid points for g_n. Analogous to Lam *et al*. (2018), the basic idea is to set m_{2n} to be initially large. Then, grid points are removed

in a stepwise manner, where a grid point is removed if g is approximately linear around the point. The proposed algorithm is based on the Akaike information criterion (AIC), given by

$$\text{AIC} = -2\ell_n(\widehat{\boldsymbol{\theta}}_n) + 2(p + m_{1n} + m_{2n} + 2),$$

and is summarized below.

[1] At step 0:

(i) Fix the number of grid points m_{1n} and choose a sufficiently large initial value for m_{2n}, say $m_{2n}^{(0)} = 10$.

(ii) Set the grid points for Λ_n at $\boldsymbol{\tau} = (\tau_0, \ldots, \tau_{m_{1n}})$ and that for g_n at $\boldsymbol{\varsigma}^{(0)} = (\varsigma_0^{(0)}, \ldots, \varsigma_{m_{2n}}^{(0)})$, where τ_h is the $100(h/m_{1n})$th empirical percentile of the combined values of L_{ij}'s and R_{ij}'s $(h = 0, \ldots, m_{1n})$, and $\varsigma_h^{(0)}$ is the $100(h/m_{2n}^{(0)})$th percentile of the covariate values Z_{ij}'s $(h = 0, \ldots, m_{2n}^{(0)})$. These locations of grid points are fixed throughout the following iterative procedures to ensure a sensible coverage.

(iii) Compute the sieve maximum likelihood estimator and evaluate the AIC, denoted by $\text{AIC}^{(0)}$.

[2] At step $k = 1, 2, \ldots$:

(i) Set $m_{2n}^{(k)} = m_{2n}^{(k-1)} - 1$.

(ii) For $h = 1, \ldots, m_{2n}^{(k)}$, remove the hth inner grid point from $\boldsymbol{\varsigma}^{(k-1)}$ to form $\boldsymbol{\varsigma}^{(k,h)}$. For each h, fit the model using $\boldsymbol{\tau}$ and $\boldsymbol{\varsigma}^{(k,h)}$ as the grid points for Λ_n and g_n respectively, and record the AIC value, denoted by $\text{AIC}^{(k,h)}$. Let $\text{AIC}^{(k)} = \min\left(\text{AIC}^{(k,1)}, \ldots, \text{AIC}^{(k,m_{2n}^{(k)})}\right)$ and $h^* = \left\{h^* : \text{AIC}^{(k,h^*)} \leqslant \text{AIC}^{(k,h)} \text{ for } 1 \leqslant h \leqslant m_{2n}^{(k)}\right\}$.

(iii) Delete $\varsigma_{h^*}^{(k-1)}$ from $\boldsymbol{\varsigma}^{(k-1)}$ to form a new set of grid points $\boldsymbol{\varsigma}^{(k)} = \boldsymbol{\varsigma}_{-h^*}^{(k-1)}$.

[3] Repeat [2] until $\text{AIC}^{(k)} > \text{AIC}^{(k-1)}$. The final model possesses $(m_{2n}^* - 1) = (m_{2n}^{(k-1)} - 1)$ inner grid points with the locations given by $\boldsymbol{\varsigma}^{(k-1)}$.

5. Simulation Studies

First, we conduct a simulation study for the general choice of m_{1n}. We simulate samples in a paired data setting (i.e., $N_i = 2$ for $i = 1, \ldots, n$) with $n = 250$ or 500 independent clusters. The frailty ξ_i is assumed to follow either Ga$(\gamma^{-1}, \gamma^{-1})$ ($\gamma = 0.5$ or 1) or Po(γ) ($\gamma = 0.6$ or 0.8), and the transformation function is set to be $G(x; \rho) = \{(1 + x)^\rho - 1\}/\rho$ with $\rho = 0.5$. We set $p = 2$ and generate X_{ij1} from i.i.d. Normal$(0, 0.25)$ and X_{ij2} from i.i.d. Bernoulli(0.5). The corresponding parameter values are $\beta_1 = \beta_2 = 1$. We generate Z_{ij} from i.i.d. Uniform$(0, 2\pi)$ and set $g(z) = \sin(z)$. The covariates X_{ij1}, X_{ij2}, and Z_{ij} are mutually independent. We set the baseline function $\Lambda(t) = 0.4t^{1.6}$. The interarrival times between the observation times $(U_{ijk} - U_{ij,k-1})$ follow Uniform$(0.1, 0.5)$, and an administrative censoring occurs at $t = 5$. The right-censoring proportion varies from 10% to 30% in the scenarios. We fix $m_{1n} = C_1 n^{1/3}$ ($C_1 = 1, 2, 3, 4$) and $m_{2n} = 3$ and set the interior knots of g_n at $(0.5\pi, 1.5\pi)$, which are the essential turning points of g. The simulation results are summarized in Table S1 of the Supporting Information. For both $n = 250$ and $n = 500$, the estimates for $\boldsymbol{\beta}$, γ and ρ are nearly unbiased with close agreement between the empirical standard deviation and estimated standard error in all scenarios with $C_1 > 2$. Therefore, we suggest to adopt $C_1 = 3$ for applying the proposed methods in general.

Second, we study the finite-sample performance of the proposed sieve maximum likelihood estimator. We set the total number of observations to be 1000, with $n = 500$ and a uniform cluster size of 2, or $n = 250$ and a uniform cluster size of 4. We consider $\rho = 0$, 0.5, and 1. We set $m_{1n} = 3n^{1/3}$ and select the grid points of g_n by the algorithm detailed in Section 4 with $m_{2n}^{(0)} = 10$. The other specifications are identical to the first set of simulations. The results are reported in Table 1, and the average of the estimated curves $\widehat{\Lambda}_n$ and $\widehat{g}_n$ under $n = 500$ and $N_i = 2$ ($i = 1, \ldots, n$) are plotted in Figure 1 and Figure 2, respectively. The corresponding curves under $n = 250$ and $N_i = 4$ ($i = 1, \ldots, n$) are given in Figures S1 and S2 in the Supporting Information. In all cases, the estimates for $\boldsymbol{\beta}$, γ, and ρ are nearly unbiased. The empirical standard deviations agree with their

corresponding estimated standard errors, whereas the coverage probabilities match quite closely with the 95% nominal level in each scenario; when $\rho = 0$, the coverage for ρ is not computed because $\widehat{\rho}_n$ is not asymptotically normal. We note that, however, the point and interval estimation of some parameters can be relatively poor when the true value of ρ is set at the boundary of the parameter space. The proposed methods provide good approximations for Λ and g by capturing the turns of the curves efficiently, confirming the effectiveness of the proposed grid points selection algorithm.

[Table 1 about here.]

[Figure 1 about here.]

[Figure 2 about here.]

6. Analysis of Signal Tandmobiel Data

We apply the proposed method to analyze the data from the Signal Tandmobiel study. This is a longitudinal prospective dental study performed in Flanders, Belgium between 1996 and 2001. The original dataset contains 4430 randomly sampled children. They were examined annually, with up to 6 dental observations for each child. See Vanobbergen *et al*. (2000) for more details about the study. Here, we consider a subsample of $n = 500$, which is publicly available in the `R` package `icensBKL`. The outcome of interest is the time to emergence of the four first permanent premolars, referred to as teeth 14, 24, 34, and 44 in the European dental notation. The observations are either interval-censored or right-censored, with a right-censoring proportion of 35%.

In this application, the frailty ξ_i represents the heterogeneity of the oral health condition among the children. To further illustrate the proposed methods, in addition to the gamma distribution and the positive stable distribution, we consider the inverse Gaussian distribution for the frailty. The Laplace transform of the inverse Gaussian distribution (IG(γ)) with unit mean and variance $(2\gamma)^{-1}$ is $\Phi(u;\gamma) = \exp\left\{2\gamma - 2\gamma^{1/2}(\gamma+u)^{1/2}\right\}$. Let $(L_{ij}, R_{ij}]$ be the observed time interval covering the

emergence time T_{ij} of the jth tooth of child i, and let X_{ij1} and X_{ij2} be the indicators of the right first premolar and maxillary first premolar, respectively for $i = 1, \ldots, n$ and $j = 1, 2, 3, 4$. Gender is considered to be an important factor associated with the emergence time and thus is included into the model as a covariate (X_{ij3}), which takes values 0 and 1 for males and females, respectively. The number of decayed, missing, and filled deciduous teeth (DMFT) index, which measures the prevalence of dental caries, is also a variable of interest in the dataset. Previously, Komárek and Lesaffre (2007) used a Bayesian accelerated failure time model to analyze the data and reported that, after controlling for the effect of gender, the bad status of the primary predecessor (DMFT index > 0) accelerates the emergence of the maxillary teeth (i.e., teeth 14 and 24). In the current model, we relax the linearity assumption and set the DMFT index as Z with its effect captured by the function g.

Since there are only a few observations with DMFT index > 10, we truncate the index value at 10. We select the grid points for g_n using the proposed algorithm with $m_{2n}^{(0)} = 10$. For Λ_n, we set $m_{1n} = C_1 n^{1/3}$ with $C_1 = 3$ irrespective of the frailty distributions, as suggested in the simulation study. The estimates for the Euclidean parameters are given in Table 2, whereas the estimated survival function for a subject with zero covariate values and $\widehat{g}_n$ are plotted in Figure 3. For each frailty model, it takes approximately 4 hours to complete the estimation using an ordinary desktop computer (Windows, i7 3.4GHz CPU with 8 GB RAM).

[Table 2 about here.]

[Figure 3 about here.]

The grid points selection algorithm suggests 4 inner grid points for g_n under the positive stable and gamma distributions and 2 inner grid points under the inverse Gaussian distribution. The estimate for ρ is close to 0 under all three frailty distributions, suggesting that the proportional odds model may be more appropriate than the Cox proportional hazards model in this application. Besides, the estimated baseline survival functions for all three frailty distributions agree with the

expectation that the emergence of teeth occurs around the age of 10 (Moslemi, 2004). Also, the three estimated g-functions in Figure 3 suggest that the effect of DMFT index is nonlinear with a drastic change near DMFT $= 2$. This echoes the findings of Komárek and Lesaffre (2007) that poor conditions of the primary predecessor tend to accelerate the emergence of the studied teeth. Based on the AIC, the positive stable distribution for the frailty provides the best fit. In this model, the emergence rates of contralateral first premolars do not differ significantly ($\widehat{\beta}_1 = 0.0068$ with p-value $= 0.9404$). The mandibular first premolars have comparatively higher rates as compared to the maxillary first premolars ($\widehat{\beta}_2 = -0.2972$ with p-value $= 0.0016$). Previous studies have reported that the difference in emergence times of contralateral teeth is minimal and that mandibular teeth tend to emerge faster than their maxillary counterparts (Eskeli *et al*., 1999; Leroy *et al*., 2003). The effect of gender is significant ($\widehat{\beta}_3 = 0.6176$ with p-value $= 0.0152$), suggesting that the emergence time of the first permanent premolars in female is earlier than that in male on average.

7. Discussion

In this article, we propose a class of semiparametric partly linear frailty models for the analysis of clustered interval-censored data. The two nonparametric functions, which characterize the baseline survival function and the nonlinear effect of a covariate, are estimated using sieve maximum likelihood estimation. We proved that the sieve maximum likelihood estimator of the Euclidean parameters are strongly consistent and asymptotically normal and the estimators of the two nonparametric functions are also strongly consistent and attain optimal rates of convergence. In particular, we demonstrate theoretically and through empirical studies that the transformation parameter ρ, which is typically assumed to be known in the literature, can be consistently estimated with a valid inference procedure.

A key difference between the proposed model and some existing transformation models for multivariate failure time data (Zeng and Lin, 2007; Zeng *et al*., 2017) is that the frailty acts multiplicatively on the transformation function G instead of the argument of G. When the frailty

is inside the argument of G, the frailty acts multiplicatively to the G^{-1}-transformed cumulative hazard and can be thought of as an unobserved covariate. Under our formulation, by contrast, the frailty is a latent multiplicative effect on the cumulative hazard function. While the appropriateness of the formulations depend on the actual applications, our formulation results in a closed-form expression for the likelihood whenever the frailty distribution has an explicit Laplace transform. This simplifies the computation of the sieve maximum likelihood estimator and allows for direct maximization of the log-likelihood function using gradient methods.

In the framework presented in this article, both the transformation function G and the frailty distribution are indexed by a single parameter. One can easily extend the framework to allow for, for example, the two-parameter family of frailty distributions considered by Lam and Kuk (1997). Another interesting yet challenging extension is to consider a nonparametric transformation function G. This would pose substantial computational and theoretical challenges.

Another possible direction of extension is to allow for a multivariate covariate $\boldsymbol{Z}_{ij}$. Due to the curse of dimensionality, a fully nonparametric g is generally infeasible. One may adopt an additive model with $g(\boldsymbol{Z}_{ij}) = \sum_{k=1}^{K} g_k(Z_{ijk})$, where Z_{ijk} is the kth component of $\boldsymbol{Z}_{ij}$, K is the dimension of $\boldsymbol{Z}_{ij}$, and g_k is an unspecified function $(k = 1, \ldots, K)$. Alternatively, one may consider a single index model and set $g(\boldsymbol{Z}_{ij}) = \widetilde{g}(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Z}_{ij})$ for some regression parameter vector $\boldsymbol{\alpha}$ and univariate, unspecified function $\widetilde{g}$. For both extensions, the unspecified functions can be estimated using the sieve maximum likelihood estimation approach, and the theoretical properties of the estimators can be established along the lines of the proofs of Theorems 1 and 2. Nevertheless, the computation of the estimators may be challenging due to the extra parameters to be estimated.

Acknowledgments

The authors would like to thank the Editor, Associate Editor and a reviewer for their valuable comments and suggestions. The research of K. F. Lam was supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No.17305819).

Data Availability Statement

The data that support the findings of this paper are openly available in the `R` package `icensBKL`, available on CRAN at: https://cran.r-project.org/web/packages/icensBKL/index.html.

References

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

Betensky, R. A., Rabinowitz, D., and Tsiatis, A. A. (2001). Computationally simple accelerated failure time regression for interval censored data. *Biometrika* **88,** 703–711.

Chen, D. G. D., Sun, J., and Peace, K. E. (2012). *Interval-Censored Time-to-Event Data: Methods and Applications*. CRC Press.

Cheng, S. C., Wei, L. J., and Ying, Z. (1995). Analysis of transformation models with censored data. *Biometrika* **82,** 835–845.

Cook, R. J., Zeng, L., and Lee, K. A. (2008). A multistate model for bivariate interval-censored failure time data. *Biometrics* **64,** 1100–1109.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B* **34,** 187–220.

de Boor, C. (1978). *A Practical Guide to Splines*. New York: Springer-Verlag.

Eskeli, R., Laine-Alava, M. T., Hausen, H., and Pahkala, R. (1999). Standards for permanent tooth emergence in Finnish children. *The Angle Orthodontist* **69,** 529–533.

Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. New York: Springer.

Huang, J. (1999). Efficient estimation of the partly linear additive Cox model. *Annals of Statistics* **27,** 1536–1563.

Huang, J. and Rossini, A. J. (1997). Sieve estimation for the proportional-odds failure-time regression model with interval censoring. *Journal of the American Statistical Association* **92,** 960–967.

Kim, Y. J. (2006). Regression analysis of doubly censored failure time data with frailty. *Biometrics* **62,** 458–464.

Komárek, A. and Lesaffre, E. (2007). Bayesian accelerated failure time model for correlated interval-censored data with a normal mixture as error distribution. *Statistica Sinica* **17,** 549–569.

Kosorok, M. R., Lee, B. L., and Fine, J. P. (2004). Robust inference for univariate proportional hazards frailty regression models. *Annals of Statistics* **32,** 1448–1491.

Lam, K. F. and Kuk, A. Y. (1997). A marginal likelihood approach to estimation in frailty models. *Journal of the American Statistical Association* **92,** 985–990.

Lam, K. F. and Leung, T. L. (2001). Marginal likelihood estimation for proportional odds models with right censored data. *Lifetime Data Analysis* **7,** 39–54.

Lam, K. F., Xu, J., and Xue, H. (2018). Estimation of age effect with change-points on survival of cancer patients. *Statistics in Medicine* **37,** 1732–1743.

Lam, K. F., Xu, Y., and Cheung, T. L. (2010). A multiple imputation approach for clustered interval-censored survival data. *Statistics in Medicine* **29,** 680–693.

Lam, K. F. and Xue, H. (2005). A semiparametric regression cure model with current status data. *Biometrika* **92,** 573–586.

Leroy, R., Bogaerts, K., Lesaffre, E., and Declerck, D. (2003). The emergence of permanent teeth in Flemish children. *Community Dentistry and Oral Epidemiology* **31,** 30–39.

Lin, X. and Carroll, R. J. (2001). Semiparametric regression for clustered data using generalized estimating equations. *Journal of the American Statistical Association* **96,** 1045–1056.

Moslemi, M. (2004). An epidemiological survey of the time and sequence of eruption of permanent teeth in 4–15-year-olds in Tehran, Iran. *International Journal of Paediatric Dentistry* **14,** 432–438.

Rossini, A. J. and Tsiatis, A. A. (1996). A semiparametric proportional odds regression model for the analysis of current status data. *Journal of the American Statistical Association* **91,** 713–721.

Schumaker, L. (2007). *Spline Functions: Basic Theory*. Cambridge University Press.

Shen, X. and Wong, W. H. (1994). Convergence rate of sieve estimates. *Annals of Statistics* **22,** 580–615.

Stone, C. J. (1985). Additive regression and other nonparametric models. *Annals of Statistics* **13,**

689–705.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer.

Vanobbergen, J., Martens, L., Lesaffre, E., and Declerck, D. (2000). The Signal-Tandmobiel project a longitudinal intervention health promotion study in Flanders (Belgium): baseline and first year results. *European Journal of Paediatric Dentistry* **2,** 87–96.

Wellner, J. A. and Zhang, Y. (2007). Two likelihood-based semiparametric estimation methods for panel count data with covariates. *Annals of Statistics* **35,** 2106–2142.

Xue, H., Lam, K. F., Cowling, B. J., and de Wolf, F. (2006). Semi-parametric accelerated failure time regression analysis with application to interval-censored HIV/AIDS data. *Statistics in Medicine* **25,** 3850–3863.

Xue, H., Lam, K. F., and Li, G. (2004). Sieve maximum likelihood estimator for semiparametric regression models with current status data. *Journal of the American Statistical Association* **99,** 346–356.

Zeng, D., Gao, F., and Lin, D. Y. (2017). Maximum likelihood estimation for semiparametric regression models with multivariate interval-censored data. *Biometrika* **104,** 505–525.

Zeng, D. and Lin, D. Y. (2007). Maximum likelihood estimation in semiparametric regression models with censored data. *Journal of the Royal Statistical Society: Series B* **69,** 507–564.

Zhou, Q., Hu, T., and Sun, J. (2017). A sieve semiparametric maximum likelihood approach for regression analysis of bivariate interval-censored failure time data. *Journal of the American Statistical Association* **112,** 664–672.

Zuma, K., Lurie, M., and Jorgensen, M. (2007). Analysis of interval-censored data from circular migrant and non-migrant sexual partnerships using the EM algorithm. *Statistics in Medicine* **26,** 309–319.

Supporting Information

Tables and Figures referenced in Section 5, proofs of theoretical results referenced in the Appendix, and the computer code for the real data analysis are available with this paper at the *Biometrics* website on Wiley Online Library.

APPENDIX

TECHNICAL PROOFS

Before proving the theorems, we present the following lemmas. Proofs of the lemmas are provided in Section S1 of the Supporting Information. Let $\mathcal{F}_n = \{\ell(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta_n\}$, where $\ell(\boldsymbol{\theta})$ is the log-likelihood for a generic cluster of subjects, and $\Theta_n = \mathcal{A}_\zeta \times \mathcal{A}_{\Lambda n} \times \mathcal{A}_{gn}$. Let $\mathbb{P}_n$ and $\mathbb{P}$ denote the empirical and true probability measures, respectively.

LEMMA 1 Under Conditions (C1)–(C4),

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}_n} = \sup_{\boldsymbol{\theta}\in\Theta_n} \left|(\mathbb{P}_n - \mathbb{P})\ell(\boldsymbol{\theta})\right| \to_{\text{a.s.}} 0.$$

Let $\dot{\boldsymbol{\ell}}_\zeta(\boldsymbol{\theta})$ denote the score statistic for $\boldsymbol{\zeta}$, $\dot{\ell}_\Lambda(\boldsymbol{\theta})[h_\Lambda]$ denote the score statistic for Λ along the direction h_Λ, and $\dot{\ell}_g(\boldsymbol{\theta})[h_g]$ denote the score statistic for g along the direction h_g. For a vector of functions $\boldsymbol{h}_\Lambda = (h_{1,\Lambda}, \ldots, h_{p,\Lambda})$, $\dot{\boldsymbol{\ell}}_\Lambda(\boldsymbol{\theta})[\boldsymbol{h}_\Lambda]$ denotes the vector $(\dot{\ell}_\Lambda(\boldsymbol{\theta})[h_{1,\Lambda}], \ldots, \dot{\ell}_\Lambda(\boldsymbol{\theta})[h_{p,\Lambda}])$. The vector $\dot{\boldsymbol{\ell}}_g(\boldsymbol{\theta})[\boldsymbol{h}_g]$ is defined similarly for a vector of functions $\boldsymbol{h}_g$. Let $\mathcal{A}_\Lambda = \{\Lambda \in \ell^\infty[a_1, b_1] : \Lambda(a_1) = 0, \Lambda \text{ is monotone nondecreasing}, \Lambda(b_1) < C\}$, $\mathcal{A}_g = \{g \in \ell^\infty[0,1] : \|g\|_{\text{TV}} < C\}$, and $\Theta = \mathcal{A}_\zeta \times \mathcal{A}_\Lambda \times \mathcal{A}_g$, where $\|\cdot\|_{\text{TV}}$ denotes the total variation norm, and C is some large enough constant.

LEMMA 2 Under Conditions (C1)–(C5), there exist $\boldsymbol{h}^*_\Lambda \in L_2[a_1, b_1]^{p+2}$ and $\boldsymbol{h}^*_g \in L_2[0,1]^{p+2}$ such that

$$\mathbb{P}\big(\dot{\ell}_\Lambda(\boldsymbol{\theta}_0)[h_\Lambda]\big\{\dot{\boldsymbol{\ell}}_\zeta(\boldsymbol{\theta}_0) - \dot{\boldsymbol{\ell}}_\Lambda(\boldsymbol{\theta}_0)[\boldsymbol{h}^*_\Lambda] - \dot{\boldsymbol{\ell}}_g(\boldsymbol{\theta}_0)[\boldsymbol{h}^*_g]\big\}\big) = \mathbf{0}$$
$$\mathbb{P}\big(\dot{\ell}_g(\boldsymbol{\theta}_0)[h_g]\big\{\dot{\boldsymbol{\ell}}_\zeta(\boldsymbol{\theta}_0) - \dot{\boldsymbol{\ell}}_\Lambda(\boldsymbol{\theta}_0)[\boldsymbol{h}^*_\Lambda] - \dot{\boldsymbol{\ell}}_g(\boldsymbol{\theta}_0)[\boldsymbol{h}^*_g]\big\}\big) = \mathbf{0}$$

for all $h_\Lambda \in L_2[a_1, b_1]$ and $h_g \in L_2[0,1]$. Also, the classes $\mathcal{G}_1 = \{\dot{\boldsymbol{\ell}}_\zeta(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$, $\mathcal{G}_2 = \{\dot{\boldsymbol{\ell}}_\Lambda(\boldsymbol{\theta})[h_\Lambda] : \boldsymbol{\theta} \in \Theta, h_\Lambda \in \text{BV}[a_1, b_1]\}$, and $\mathcal{G}_3 = \{\dot{\boldsymbol{\ell}}_g(\boldsymbol{\theta})[h_g] : \boldsymbol{\theta} \in \Theta, h_g \in \text{BV}[0,1]\}$ are Donsker. In addition, the efficient information matrix for $\boldsymbol{\zeta}$, defined by

$$\widetilde{\boldsymbol{I}} \equiv \mathbb{P}\big\{\dot{\boldsymbol{\ell}}_\zeta(\boldsymbol{\theta}_0) - \dot{\boldsymbol{\ell}}_\Lambda(\boldsymbol{\theta}_0)[\boldsymbol{h}^*_\Lambda] - \dot{\boldsymbol{\ell}}_g(\boldsymbol{\theta}_0)[\boldsymbol{h}^*_g]\big\}^{\otimes 2},$$

is positive definite.

PROOF OF THEOREM 1. We show that there exists a local maximum of the log-likelihood function over the sieve space that is consistent. By Schumaker (2007), there exist functions $\widetilde{\Lambda}_n$ and $\widetilde{g}_n$, such that $\|\widetilde{\Lambda}_n - \Lambda_0\|_\infty = O(n^{-2\nu_1})$ and $\|\widetilde{g}_n - g_0\|_\infty = O(n^{-2\nu_2})$. Let $\widetilde{\boldsymbol{\theta}}_n = (\boldsymbol{\zeta}_0, \widetilde{\Lambda}_n, \widetilde{g}_n)$. By definition of the sieve maximum likelihood estimator, $\mathbb{P}_n\ell(\widehat{\boldsymbol{\theta}}_n) \geqslant \mathbb{P}_n\ell(\widetilde{\boldsymbol{\theta}}_n)$, such that

$$\mathbb{P}\ell(\widehat{\boldsymbol{\theta}}_n) - \mathbb{P}\ell(\boldsymbol{\theta}_0) \geqslant \mathbb{P}\{\ell(\widetilde{\boldsymbol{\theta}}_n) - \ell(\boldsymbol{\theta}_0)\} + (\mathbb{P}_n - \mathbb{P})\{\ell(\widetilde{\boldsymbol{\theta}}_n) - \ell(\widehat{\boldsymbol{\theta}}_n)\}.$$

The first term on the right-hand side of the above inequality is $o(1)$, and by Lemma 1, the second term goes to 0 almost surely. Therefore, the right-hand side of the above display goes to 0 almost surely. Let $h(\mathcal{U}; \boldsymbol{\theta}) = \Pr(U_{1k_1} < T_1 < U_{1,k_1+1}, \ldots, U_{Nk_N} < T_N < U_{N,k_N+1}; \boldsymbol{\theta})$ for given N and $(k_1, \ldots, k_N)$. The left-hand side of the above inequality is

$$\begin{aligned}
&\mathrm{E}\Bigg\{ \sum_{k_1=0}^{M_1} \cdots \sum_{k_N=0}^{M_N} \Delta_{1k_1} \cdots \Delta_{Nk_N} \log \frac{h(\mathcal{U}; \widehat{\boldsymbol{\theta}}_n)}{h(\mathcal{U}; \boldsymbol{\theta}_0)} \Bigg\} \\
=&\mathrm{E}\Bigg\{ \sum_{k_1=0}^{M_1} \cdots \sum_{k_N=0}^{M_N} \mathrm{E}(\Delta_{1k_1} \cdots \Delta_{Nk_N} \mid \mathcal{U}) \log \frac{h(\mathcal{U}; \widehat{\boldsymbol{\theta}}_n)}{h(\mathcal{U}; \boldsymbol{\theta}_0)} \Bigg\} \\
=&\mathrm{E}\Bigg\{ \sum_{k_1=0}^{M_1} \cdots \sum_{k_N=0}^{M_N} h(\mathcal{U}; \boldsymbol{\theta}_0) \log \frac{h(\mathcal{U}; \widehat{\boldsymbol{\theta}}_n)}{h(\mathcal{U}; \boldsymbol{\theta}_0)} \Bigg\} \\
=&\mathrm{E}\Bigg(\sum_{k_1=0}^{M_1} \cdots \sum_{k_N=0}^{M_N} \Bigg[h(\mathcal{U}; \widehat{\boldsymbol{\theta}}_n) - h(\mathcal{U}; \boldsymbol{\theta}_0) - h(\mathcal{U}; \widehat{\boldsymbol{\theta}}_n) q\Bigg\{ \frac{h(\mathcal{U}; \boldsymbol{\theta}_0)}{h(\mathcal{U}; \widehat{\boldsymbol{\theta}}_n)} \Bigg\} \Bigg] \Bigg) \\
=&-\mathrm{E}\Bigg[\sum_{k_1=0}^{M_1} \cdots \sum_{k_N=0}^{M_N} h(\mathcal{U}; \widehat{\boldsymbol{\theta}}_n) q\Bigg\{ \frac{h(\mathcal{U}; \boldsymbol{\theta}_0)}{h(\mathcal{U}; \widehat{\boldsymbol{\theta}}_n)} \Bigg\} \Bigg],
\end{aligned}$$

where $q(x) = x \log x - x + 1$, and the last equality holds because $\mathrm{E}\{\sum_{k_1=0}^{M_1} \cdots \sum_{k_N=0}^{M_N} h(\mathcal{U}; \boldsymbol{\theta})\} = 1$ for any $\boldsymbol{\theta}$. For $\widehat{\boldsymbol{\theta}}_n$ in a small enough neighborhood of $\boldsymbol{\theta}_0$, $0 \leqslant h(\mathcal{U}; \boldsymbol{\theta}_0)/h(\mathcal{U}; \widehat{\boldsymbol{\theta}}_n) \leqslant 5$, such that $q\{h(\mathcal{U}; \boldsymbol{\theta}_0)/h(\mathcal{U}; \widehat{\boldsymbol{\theta}}_n)\} \geqslant \{h(\mathcal{U}; \boldsymbol{\theta}_0)/h(\mathcal{U}; \widehat{\boldsymbol{\theta}}_n) - 1\}^2/4$. In this case, the right-hand side of the above equation is bounded above (up to a scaling factor) by

$$-\sum_{k_1=0}^{M_1} \cdots \sum_{k_N=0}^{M_N} \mathrm{E}\Bigg[\frac{\{h(\mathcal{U}; \boldsymbol{\theta}_0) - h(\mathcal{U}; \widehat{\boldsymbol{\theta}}_n)\}^2}{h(\mathcal{U}; \widehat{\boldsymbol{\theta}}_n)} \Bigg] \lesssim \sum_{k_1=0}^{M_1} \cdots \sum_{k_N=0}^{M_N} -\mathrm{E}\big[\{h(\mathcal{U}; \boldsymbol{\theta}_0) - h(\mathcal{U}; \widehat{\boldsymbol{\theta}}_n)\}^2\big]. \tag{A.1}$$

Consider the term in the summation of the right-hand side of (A.1) at $k_1 = M_1, \ldots, k_N = M_N$. By

Condition (C5), this term is (up to a scaling factor) bounded above by

$$-\mathrm{E}\left[\left\{\widehat{\Lambda}_n(U_{1M_1})e^{\boldsymbol{X}_1^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_n+\widehat{g}_n(Z_1)}-\Lambda_0(U_{1M_1})e^{\boldsymbol{X}_1^{\mathrm{T}}\boldsymbol{\beta}_0+g_0(Z_1)}\right\}^2\right]-|\widehat{\gamma}_n-\gamma_0|^2-|\widehat{\rho}_n-\rho_0|^2.$$

Therefore, we conclude that $|\widehat{\gamma}_n-\gamma_0|^2+|\widehat{\rho}_n-\rho_0|^2\to_{\mathrm{a.s.}}0$. We then follow the arguments in the proof of Theorem 3.2 in Wellner and Zhang (2007) to place a bound on the first term in the above expression by the differences between the individual parameter estimators and the corresponding true values. Following the arguments in Wellner and Zhang (2007, pp. 2126–2127), we have

$$\begin{aligned}&\mathrm{E}\left[\left\{\widehat{\Lambda}_n(U_{1M_1})e^{\boldsymbol{X}_1^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_n+\widehat{g}_n(Z_1)}-\Lambda_0(U_{1M_1})e^{\boldsymbol{X}_1^{\mathrm{T}}\boldsymbol{\beta}_0+g_0(Z_1)}\right\}^2\right]\\ \gtrsim&\,\mathrm{E}\left[\left\{q_1(U_{1M_1},\boldsymbol{X}_1,Z_1)h_1(U_{1M_1},Z_1)+q_2(U_{1M_1},Z_1)\right\}^2\right],\end{aligned}\tag{A.2}$$

where $q_1(U_{1M_1},\boldsymbol{X}_1,Z_1)=\Lambda_0(U_{1M_1})e^{g_0(Z_1)}\boldsymbol{X}_1^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0)$, $h_1(U_{1M_1},Z_1)=1+t_1\{\widehat{\Lambda}_n(U_{1M_1})e^{\widehat{g}_n(Z_1)}-\Lambda_0(U_{1M_1})e^{g_0(Z_1)}\}/\Lambda_0(U_{1M_1})e^{g_0(Z_1)}$, $q_2(U_{1M_1},Z_1)=\widehat{\Lambda}_n(U_{1M_1})e^{\widehat{g}_n(Z_1)}-\Lambda_0(U_{1M_1})e^{g_0(Z_1)}$, and t_1 is some value between 0 and 1. Under Condition (C3), we can show that

$$[\mathrm{E}\{q_1(U_{1M_1},\boldsymbol{X}_1,Z_1)q_2(U_{1M_1},Z_1)\}]^2\leqslant(1-\eta)\mathrm{E}\{q_1(U_{1M_1},\boldsymbol{X}_1,Z_1)^2\}\mathrm{E}\{q_2(U_{1M_1},Z_1)^2\},$$

so that the right-hand side of (A.2) is (up to a scaling factor) bounded below by

$$\mathrm{E}\{q_1(U_{1M_1},\boldsymbol{X}_1,Z_1)^2\}+\mathrm{E}\{q_2(U_{1M_1},Z_1)^2\}\gtrsim\|\widehat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0\|^2+\mathrm{E}\{q_2(U_{1M_1},Z_1)^2\}.$$

We conclude that $\|\widehat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0\|^2\to_{\mathrm{a.s.}}0$. By the mean-value theorem, we have

$$\begin{aligned}\mathrm{E}\{q_2(U_{1M_1},Z_1)^2\}=&\,\mathrm{E}[\{e^{\widehat{g}_n(Z_1)+\log\widehat{\Lambda}_n(U_{1M_1})}-e^{g_0(Z_1)+\log\Lambda_0(U_{1M_1})}\}^2]\\ \geqslant&\min_{z\in[0,1],u\in[C^{-1},C]t\in[0,1]}\left\{e^{t\widehat{g}_n(z)+(1-t)g_0(z)+t\log\widehat{\Lambda}_n(u)+(1-t)\log\Lambda_0(u)}\right\}^2\\ &\times\mathrm{E}\left[\left\{(\widehat{g}_n-g_0)(Z_1)+(\log\widehat{\Lambda}_n-\log\Lambda_0)(U_{1M_1})\right\}^2\right],\end{aligned}$$

where C is defined in Condition (C2), and the minimum term is bounded away from zero under Condition (C1). Under Condition (C2), we can use the arguments for the proof of Lemma 1 of Stone (1985) to show that

$$\begin{aligned}\mathrm{E}\{q_2(U_{1M_1},Z_1)^2\}\gtrsim&\,\mathrm{E}\{(\widehat{g}_n-g_0)(Z_1)^2\}+\mathrm{E}\{(\log\widehat{\Lambda}_n-\log\Lambda_0)(U_{1M_1})^2\}\\ \gtrsim&\,\|\widehat{g}_n-g_0\|^2+\mathrm{E}\{(\widehat{\Lambda}_n-\Lambda_0)(U_{1M_1})^2\},\end{aligned}$$

where the second inequality follows because the support of Z_1 covers $[0,1]$ and that $\Lambda_0(t)$ is

bounded for $t \in [a_1, b_1]$. We conclude that $\|\widehat{g}_n - g_0\|^2 \to_{\text{a.s.}} 0$. Finally, because the support of $(U_{11}, \ldots, U_{1M_1})$ covers $[a_1, b_1]$, we can use similar arguments to show that the right-hand side of (A.1) is (up to a scaling factor) bounded above by $-\|\widehat{\Lambda}_n - \Lambda_0\|^2$. The desired consistency result follows.

The proof of the rate of convergence is based on Lemma 3.4.1 of van der Vaart and Wellner (1996). For $\boldsymbol{\theta}$ in a small enough neighborhood of $\boldsymbol{\theta}_0$,

$$\begin{aligned}\mathbb{P}\ell(\boldsymbol{\theta}) - \mathbb{P}\ell(\widetilde{\boldsymbol{\theta}}_n) &= \mathbb{P}\ell(\boldsymbol{\theta}) - \mathbb{P}\ell(\boldsymbol{\theta}_0) - \{\mathbb{P}\ell(\widetilde{\boldsymbol{\theta}}_n) - \mathbb{P}\ell(\boldsymbol{\theta}_0)\} \\ &\lesssim -d(\boldsymbol{\theta}, \boldsymbol{\theta}_0)^2 + O(\|\widetilde{\Lambda}_n - \Lambda_0\|_\infty^2 + \|\widetilde{g}_n - g_0\|_\infty^2),\end{aligned}$$

where the inequality follows from the proof of consistency. Let $\delta_n = n^{-2\nu_1} + n^{-2\nu_2}$. We conclude that for $\delta > \delta_n$ and large enough n,

$$\sup_{\substack{\delta/2 < d(\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}}_n) < \delta \\ \boldsymbol{\theta} \in \Theta_n}} \mathbb{P}\ell(\boldsymbol{\theta}) - \mathbb{P}\ell(\widetilde{\boldsymbol{\theta}}_n) \lesssim -\delta^2.$$

Let $\mathcal{F}_\delta = \big\{\ell(\boldsymbol{\theta}) - \ell(\widetilde{\boldsymbol{\theta}}_n) : \delta/2 < d(\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}}_n) < \delta, \boldsymbol{\theta} \in \Theta_n\big\}$. Following the arguments of Shen and Wong (1994, pp. 597), for $0 < \epsilon < \delta$, we have

$$\log N_{[]}\{\epsilon, \mathcal{F}_\delta, L_2(\mathbb{P})\} \lesssim \max(m_{1n}, m_{2n}) \log\left(\frac{\delta}{\epsilon}\right).$$

By Lemma 3.4.2 of van der Vaart and Wellner (1996), we have

$$\begin{aligned}\|n^{1/2}(\mathbb{P}_n - \mathbb{P})\|_{\mathcal{F}_\delta} &\lesssim J_{[]}\{\delta, \mathcal{F}_\delta, L_2(\mathbb{P})\}\left[1 + \frac{J_{[]}\{\delta, \mathcal{F}_\delta, L_2(\mathbb{P})\}}{\delta^2 n^{1/2}}\right] \\ &\lesssim \max(m_{1n}, m_{2n})^{1/2}\delta\left\{1 + \frac{\max(m_{1n}, m_{2n})^{1/2}}{\delta n^{1/2}}\right\} \equiv \phi_n(\delta),\end{aligned}$$

where $J_{[]}\{\delta, \mathcal{F}_\delta, L_2(\mathbb{P})\} \equiv \int_0^\delta [1 + \log N_{[]}\{\eta, \mathcal{F}_\delta, L_2(\mathbb{P})\}]^{1/2}\, \mathrm{d}\eta$ is the bracketing entropy. Clearly, $\phi_n(\delta)/\delta$ is decreasing in δ, and $r_n^2 \phi_n(1/r_n) \lesssim n^{1/2}$ for $r_n = n^{(1-\nu_1)/2} + n^{(1-\nu_2)/2}$. Therefore, by Theorem 3.4.1 of van der Vaart and Wellner (1996), $d(\widehat{\boldsymbol{\theta}}_n, \widetilde{\boldsymbol{\theta}}_n) = O_p(r_n)$. Combined with $d(\widetilde{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) = O(n^{-2\nu_1} + n^{-2\nu_2})$, the desired result follows.

PROOF OF THEOREM 2. Clearly, we can write $\widehat{g}_n = \sum_{j=1}^{m_{2n}} \widehat{\alpha}_{nj} B_j$ and $\widetilde{g}_n = \sum_{j=1}^{m_{2n}} \widetilde{\alpha}_{nj} B_j$ for some $(\widehat{\alpha}_{n1}, \ldots, \widehat{\alpha}_{n,m_{2n}})$ and $(\widetilde{\alpha}_{n1}, \ldots, \widetilde{\alpha}_{n,m_{2n}})$, where B_j's are B-spline functions with order 1.

From the proof of Theorem 1, $\|\widehat{g}_n - \widetilde{g}_n\| = O_p[n^{-\min\{(1-\nu_1)/2,(1-\nu_2)/2,2\nu_1,2\nu_2\}}]$. We have

$$\|\widehat{g}_n - \widetilde{g}_n\|_{\mathrm{TV}} = \left\| \sum_{j=1}^{m_{2n}} (\widehat{\alpha}_{nj} - \widetilde{\alpha}_{nj}) B_j' \right\| \leqslant \sum_{j=1}^{m_{2n}} |\widehat{\alpha}_{nj} - \widetilde{\alpha}_{nj}| \|B_j'\|_\infty \leqslant O(m_{2n}^{3/2}) \Big\{ \sum_{j=1}^{m_{2n}} (\widehat{\alpha}_{nj} - \widetilde{\alpha}_{nj})^2 \Big\}^{1/2}.$$

By de Boor (1978, p. 155), the L_2-norm between $\widehat{g}_n$ and $\widetilde{g}_n$ is bounded below by the Euclidean norm of the corresponding coefficient vectors up to a scaling factor. Therefore, the right-hand side of the second inequality above is $O_p[n^{3\nu_2/2-\min\{(1-\nu_1)/2,(1-\nu_2)/2,2\nu_1,2\nu_2\}}]$, which is $o_p(1)$ by the choices of ν_1 and ν_2. Therefore, $\widehat{g}_n$ belongs to the space of bounded total variation $\mathrm{BV}[0,1]$.

By definition of the sieve maximum likelihood estimator, $\mathbb{P}_n \dot{\boldsymbol{\ell}}_\zeta(\widehat{\boldsymbol{\theta}}_n) = \mathbf{0}$ and $\mathbb{P}_n \dot{\boldsymbol{\ell}}_g(\widehat{\boldsymbol{\theta}}_n)[\boldsymbol{h}_{n,g}^*] = \mathbf{0}$, where $\boldsymbol{h}_{n,g}^*$ is the (componentwise) projection of $\boldsymbol{h}_g^*$ onto $\mathcal{A}_{gn}$. Also, following the arguments in the proof of Theorem 5.3 in Huang and Rossini (1997), we can show that $\mathbb{P}_n \dot{\boldsymbol{\ell}}_\Lambda(\widehat{\boldsymbol{\theta}}_n)[\boldsymbol{h}_{n,\Lambda}^*] = o_p(n^{-1/2})$, where $\boldsymbol{h}_{n,\Lambda}^*$ is the projection of $\boldsymbol{h}_\Lambda^*$ onto $\mathcal{A}_{\Lambda n}$. From the equations for solving $\boldsymbol{h}_\Lambda^*$ and $\boldsymbol{h}_g^*$ given in the proof of Lemma A2 and Condition 2, both $\boldsymbol{h}_\Lambda^*$ and $\boldsymbol{h}_g^*$ are twice continuously differentiable. Therefore, $\|\boldsymbol{h}_{n,\Lambda}^* - \boldsymbol{h}_\Lambda^*\|_\infty = O(n^{-2\nu_1})$ and $\|\boldsymbol{h}_{n,g}^* - \boldsymbol{h}_g^*\|_\infty = O(n^{-2\nu_2})$. By the properties of the score statistic, $\mathbb{P}\dot{\boldsymbol{\ell}}_\zeta(\boldsymbol{\theta}_0) = \mathbf{0}$, $\mathbb{P}\dot{\boldsymbol{\ell}}_\Lambda(\boldsymbol{\theta}_0)[\boldsymbol{h}_\Lambda^*] = \mathbf{0}$, and $\mathbb{P}\dot{\boldsymbol{\ell}}_g(\boldsymbol{\theta}_0)[\boldsymbol{h}_g^*] = \mathbf{0}$. We have

$$\begin{aligned}\mathbb{P}_n \dot{\boldsymbol{\ell}}_\Lambda(\widehat{\boldsymbol{\theta}}_n)[\boldsymbol{h}_\Lambda^*] =& \mathbb{P}_n \dot{\boldsymbol{\ell}}_\Lambda(\widehat{\boldsymbol{\theta}}_n)[\boldsymbol{h}_{n,\Lambda}^*] + \mathbb{P}\dot{\boldsymbol{\ell}}_\Lambda(\boldsymbol{\theta}_0)[\boldsymbol{h}_\Lambda^* - \boldsymbol{h}_{n,\Lambda}^*] + (\mathbb{P}_n - \mathbb{P})\dot{\boldsymbol{\ell}}_\Lambda(\widehat{\boldsymbol{\theta}}_n)[\boldsymbol{h}_\Lambda^* - \boldsymbol{h}_{n,\Lambda}^*] \\ &+ \mathbb{P}\{\dot{\boldsymbol{\ell}}_\Lambda(\widehat{\boldsymbol{\theta}}_n)[\boldsymbol{h}_\Lambda^* - \boldsymbol{h}_{n,\Lambda}^*] - \dot{\boldsymbol{\ell}}_\Lambda(\boldsymbol{\theta}_0)[\boldsymbol{h}_\Lambda^* - \boldsymbol{h}_{n,\Lambda}^*]\}.\end{aligned}$$

The first two terms of the right-hand side above are $o_p(n^{-1/2})$ (or zero). By Lemma 2, $\mathcal{G}_2$ is Donsker, so that the third term is $o_p(n^{-1/2})$. In addition, by Theorem 1 and applying a first-order linear expansion of $\dot{\boldsymbol{\ell}}_\Lambda(\widehat{\boldsymbol{\theta}})$ at $\boldsymbol{\theta}_0$, we can show that the fourth term is $O_p[\{n^{-(1-\nu_1)/2} + n^{-2\nu_1} + n^{-(1-\nu_2)/2} + n^{-2\nu_2}\}n^{-2\nu_1}]$, which is $o_p(n^{-1/2})$ by the choices of ν_1 and ν_2. Likewise, $\mathbb{P}_n \dot{\boldsymbol{\ell}}_g(\widehat{\boldsymbol{\theta}}_n)[\boldsymbol{h}_g^*] = o_p(n^{-1/2})$. We have

$$\begin{aligned}\mathbb{P}_n\{\dot{\boldsymbol{\ell}}_\zeta(\widehat{\boldsymbol{\theta}}_n) - \dot{\boldsymbol{\ell}}_\zeta(\boldsymbol{\theta}_0)\} &= -(\mathbb{P}_n - \mathbb{P})\dot{\boldsymbol{\ell}}_\zeta(\boldsymbol{\theta}_0) \\ \mathbb{P}_n\{\dot{\boldsymbol{\ell}}_\Lambda(\widehat{\boldsymbol{\theta}}_n)[\boldsymbol{h}_\Lambda^*] - \dot{\boldsymbol{\ell}}_\Lambda(\boldsymbol{\theta}_0)[\boldsymbol{h}_\Lambda^*]\} &= -(\mathbb{P}_n - \mathbb{P})\dot{\boldsymbol{\ell}}_\Lambda(\boldsymbol{\theta}_0)[\boldsymbol{h}_\Lambda^*] + o_p(n^{-1/2}) \\ \mathbb{P}_n\{\dot{\boldsymbol{\ell}}_g(\widehat{\boldsymbol{\theta}}_n)[\boldsymbol{h}_g^*] - \dot{\boldsymbol{\ell}}_g(\boldsymbol{\theta}_0)[\boldsymbol{h}_g^*]\} &= -(\mathbb{P}_n - \mathbb{P})\dot{\boldsymbol{\ell}}_g(\boldsymbol{\theta}_0)[\boldsymbol{h}_g^*] + o_p(n^{-1/2}).\end{aligned}$$

Let $\ddot{\boldsymbol{\ell}}_{\zeta\zeta}$, $\ddot{\boldsymbol{\ell}}_{\Lambda\zeta}[\boldsymbol{h}_{1,\Lambda}]$, and $\ddot{\boldsymbol{\ell}}_{g\zeta}[\boldsymbol{h}_{1,g}]$ be respectively the derivatives of $\dot{\boldsymbol{\ell}}_\zeta$, $\dot{\boldsymbol{\ell}}_\Lambda[\boldsymbol{h}_{1,\Lambda}]$, and $\dot{\boldsymbol{\ell}}_g[\boldsymbol{h}_{1,g}]$ with respect to $\boldsymbol{\zeta}$, $\ddot{\boldsymbol{\ell}}_{\zeta\Lambda}[h_{2,\Lambda}]$, $\ddot{\boldsymbol{\ell}}_{\Lambda\Lambda}[\boldsymbol{h}_{1,\Lambda}, h_{2,\Lambda}]$, and $\ddot{\boldsymbol{\ell}}_{g\Lambda}[\boldsymbol{h}_{1,g}, h_{2,\Lambda}]$ be respectively the derivatives of $\dot{\boldsymbol{\ell}}_\zeta$, $\dot{\boldsymbol{\ell}}_\Lambda[\boldsymbol{h}_{1,\Lambda}]$, and $\dot{\boldsymbol{\ell}}_g[\boldsymbol{h}_{1,g}]$ with respect to Λ along the direction $h_{2,\Lambda}$, and $\ddot{\boldsymbol{\ell}}_{\zeta g}[h_{2,g}]$, $\ddot{\boldsymbol{\ell}}_{\Lambda g}[\boldsymbol{h}_{1,\Lambda}, h_{2,g}]$, and

$\ddot{\boldsymbol{\ell}}_{gg}[\boldsymbol{h}_{1,g}, h_{2,g}]$ be respectively the derivatives of $\dot{\boldsymbol{\ell}}_{\zeta}$, $\dot{\boldsymbol{\ell}}_{\Lambda}[\boldsymbol{h}_{1,\Lambda}]$, and $\dot{\boldsymbol{\ell}}_{g}[\boldsymbol{h}_{1,g}]$ with respect to g along the direction $h_{2,g}$. Because the classes $\mathcal{G}_1$, $\mathcal{G}_2$, and $\mathcal{G}_3$ are Donsker, the empirical measures on the left-hand sides of the above display can be replaced by $\mathbb{P}$ with an additional $o_p(n^{-1/2})$ term. Also, by the Taylor series expansion, the boundedness of the derivatives of the score statistics, and that $d(\widehat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) = o_p(n^{-1/4})$, we conclude that

$$\begin{aligned}
\mathbb{P}\ddot{\boldsymbol{\ell}}_{\zeta\zeta}(\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0) + \mathbb{P}\ddot{\boldsymbol{\ell}}_{\zeta\Lambda}[\widehat{\Lambda}_n - \Lambda_0] + \mathbb{P}\ddot{\boldsymbol{\ell}}_{\zeta g}[\widehat{g}_n - g_0] &= -(\mathbb{P}_n - \mathbb{P})\dot{\boldsymbol{\ell}}_{\zeta}(\boldsymbol{\theta}_0) + o_p(n^{-1/2}) \\
\mathbb{P}\ddot{\boldsymbol{\ell}}_{\Lambda\zeta}[\boldsymbol{h}^*_{\Lambda}](\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0) + \mathbb{P}\ddot{\boldsymbol{\ell}}_{\Lambda\Lambda}[\boldsymbol{h}^*_{\Lambda}, \widehat{\Lambda}_n - \Lambda_0] + \mathbb{P}\ddot{\boldsymbol{\ell}}_{\Lambda g}[\boldsymbol{h}^*_{\Lambda}, \widehat{g}_n - g_0] &= -(\mathbb{P}_n - \mathbb{P})\dot{\boldsymbol{\ell}}_{\Lambda}(\boldsymbol{\theta}_0)[\boldsymbol{h}^*_{\Lambda}] + o_p(n^{-1/2}) \\
\mathbb{P}\ddot{\boldsymbol{\ell}}_{g\zeta}[\boldsymbol{h}^*_{g}](\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0) + \mathbb{P}\ddot{\boldsymbol{\ell}}_{g\Lambda}[\boldsymbol{h}^*_{g}, \widehat{\Lambda}_n - \Lambda_0] + \mathbb{P}\ddot{\boldsymbol{\ell}}_{gg}[\boldsymbol{h}^*_{g}, \widehat{g}_n - g_0] &= -(\mathbb{P}_n - \mathbb{P})\dot{\boldsymbol{\ell}}_{g}(\boldsymbol{\theta}_0)[\boldsymbol{h}^*_{g}] + o_p(n^{-1/2}).
\end{aligned}$$

Subtracting the second and third equalities in the above display from the first equality, we have

$$\mathbb{P}(\ddot{\boldsymbol{\ell}}_{\zeta\zeta} - \ddot{\boldsymbol{\ell}}_{\Lambda\zeta}[\boldsymbol{h}^*_{\Lambda}] - \ddot{\boldsymbol{\ell}}_{g\zeta}[\boldsymbol{h}^*_{g}])(\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0) = -(\mathbb{P}_n - \mathbb{P})\{\dot{\boldsymbol{\ell}}_{\zeta}(\boldsymbol{\theta}_0) - \dot{\boldsymbol{\ell}}_{\Lambda}(\boldsymbol{\theta}_0)[\boldsymbol{h}^*_{\Lambda}] - \dot{\boldsymbol{\ell}}_{g}(\boldsymbol{\theta}_0)[\boldsymbol{h}^*_{g}]\} + o_p(n^{-1/2}).$$

The desired result follows from the central limit theorem and the invertibility of the efficient information matrix $\widetilde{\boldsymbol{I}}$.

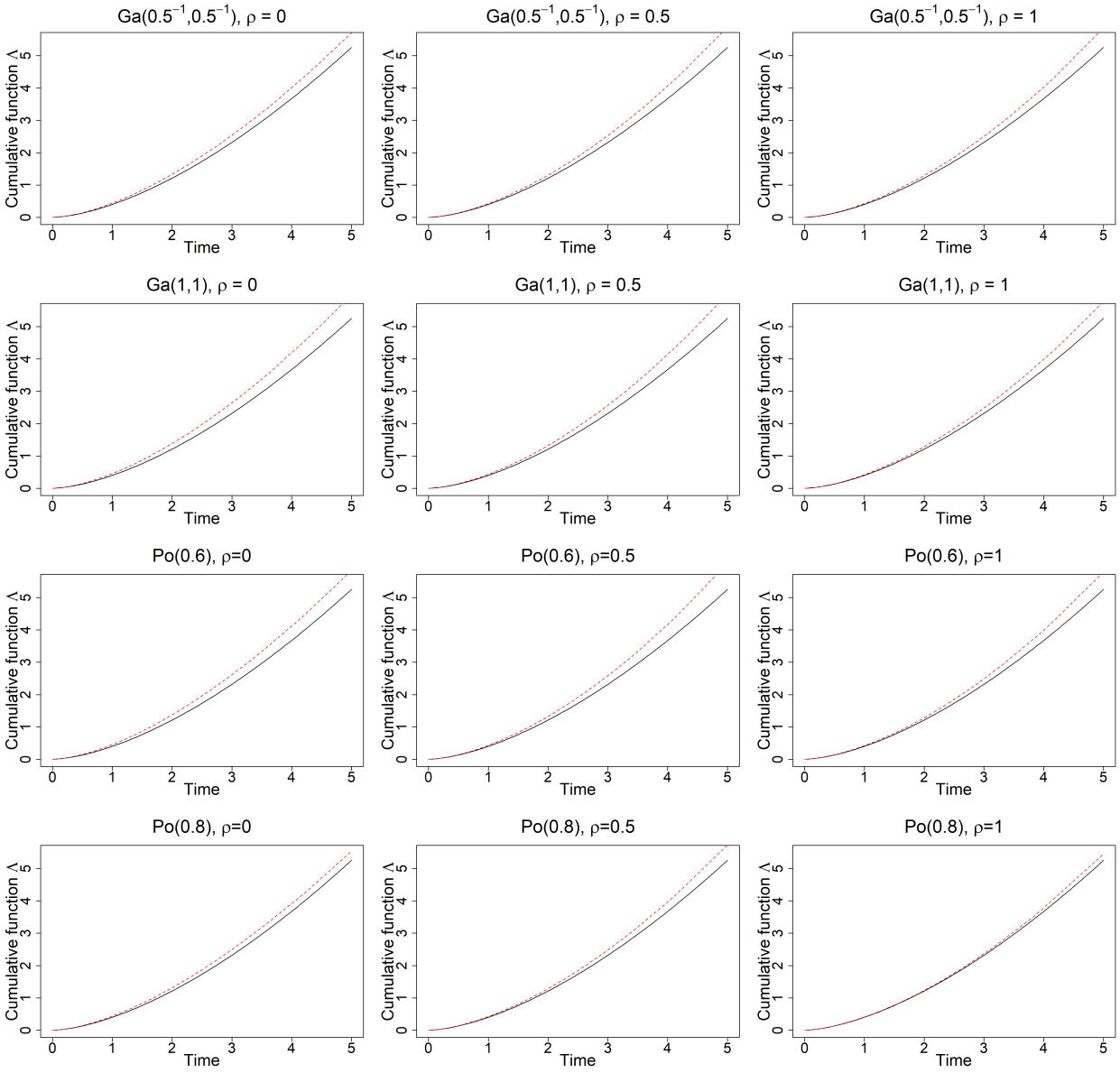

Figure 1. Estimation of Λ under the gamma and positive stable frailty distributions for $n = 500$ and $N_i = 2$ $(i = 1, \ldots, n)$. The solid black lines represent the true values and the dashed red lines are the averaged estimates. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

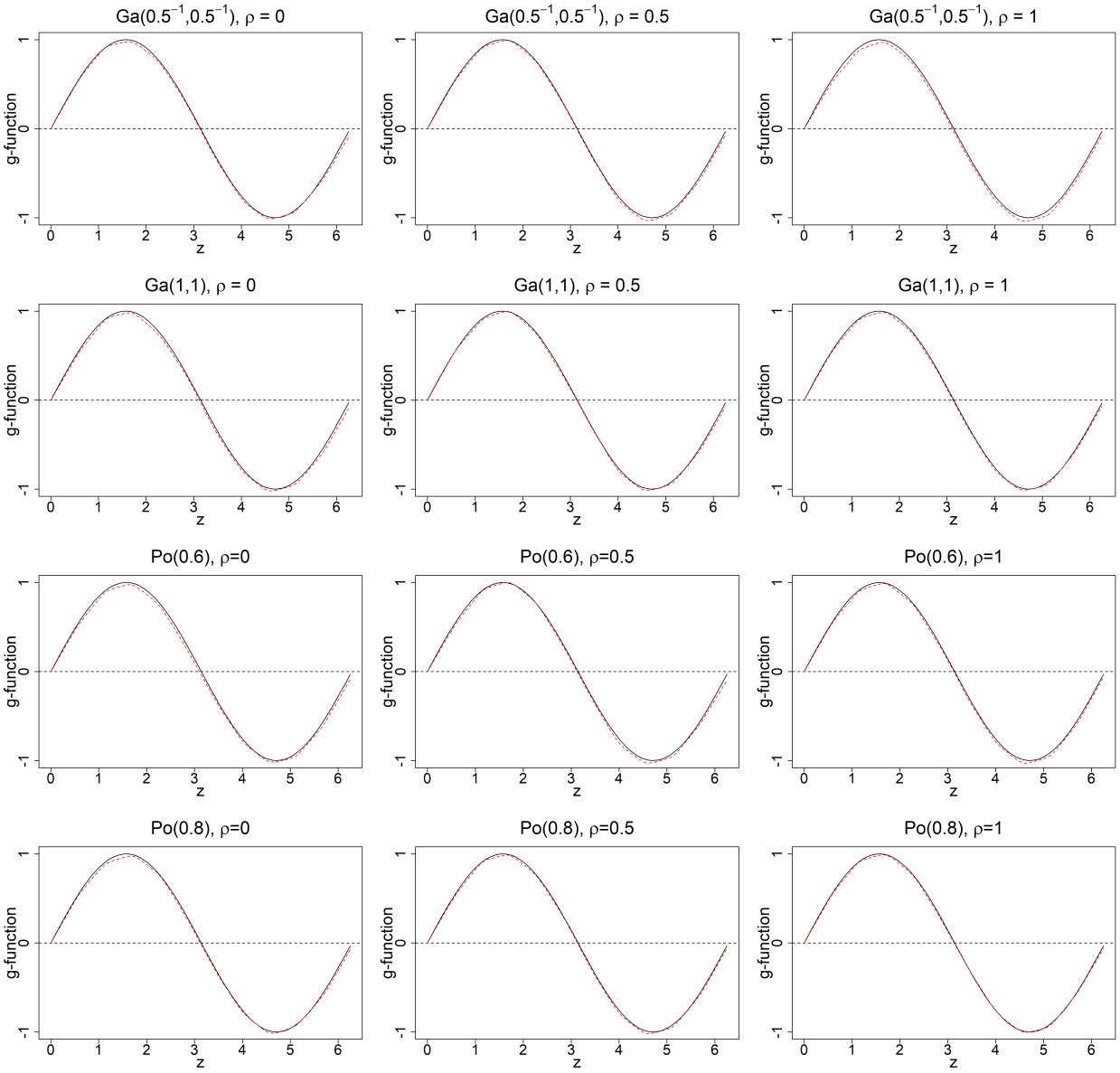

Figure 2. Estimation of g under the gamma and positive stable frailty distributions for $n = 500$ and $N_i = 2$ $(i = 1, \ldots, n)$. The solid black line represents the true values and the dashed red lines are the averaged estimates. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

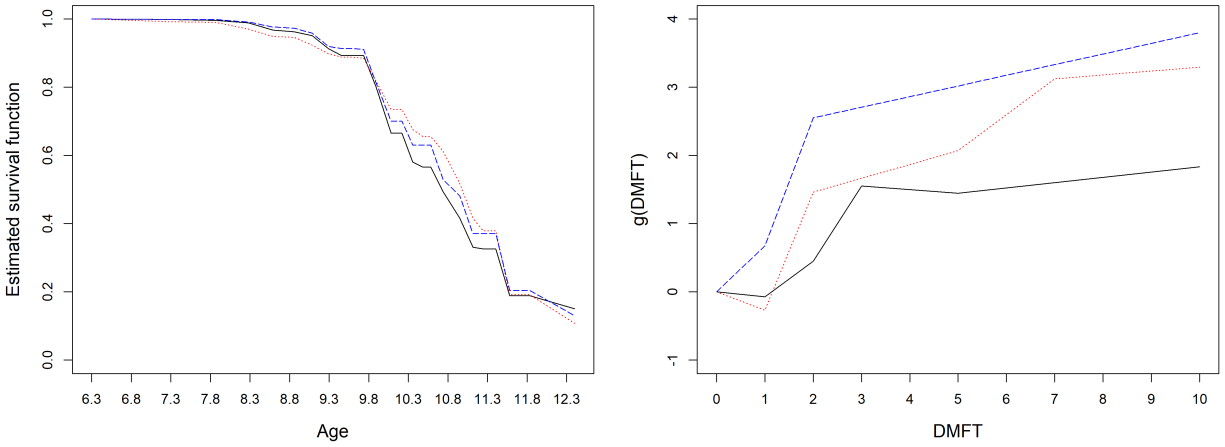

Figure 3. Estimated survival functions and g-functions for the dental dataset. The solid black, dotted red, and dashed blue lines represent the results based on the $\text{Ga}(\gamma^{-1}, \gamma^{-1})$, $\text{Po}(\gamma)$, and $\text{IG}(\gamma)$ respectively. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

Table 1
Simulation results under $C_1 = 3$ with 1000 replicates in each scenario

ξ	Par.	True	Bias (ESD, ASE, CP)	Par.	True	Bias (ESD, ASE, CP)	Par.	True	Bias (ESD, ASE, CP)
$N_i = 2, n = 500$									
$\xi \sim$ Ga	β_1	1	0.003 (0.142, 0.142, 0.95)	β_1	1	0.004 (0.135, 0.127, 0.94)	β_1	1	0.004 (0.121, 0.119, 0.94)
	β_2	1	0.000 (0.146, 0.141, 0.94)	β_2	1	0.004 (0.131, 0.127, 0.94)	β_2	1	0.004 (0.123, 0.119, 0.95)
	γ	0.5	0.026 (0.091, 0.093, 0.91)	γ	0.5	0.007 (0.093, 0.090, 0.94)	γ	0.5	0.006 (0.087, 0.087, 0.94)
	ρ	0	0.048 (0.073, 0.080, −)	ρ	0.5	0.023 (0.170, 0.158, 0.96)	ρ	1	0.054 (0.287, 0.263, 0.96)
	β_1	1	0.002 (0.156, 0.155, 0.96)	β_1	1	0.003 (0.138, 0.139, 0.96)	β_1	1	0.005 (0.125, 0.127, 0.96)
	β_2	1	0.007 (0.163, 0.155, 0.94)	β_2	1	0.006 (0.140, 0.138, 0.94)	β_2	1	0.001 (0.129, 0.126, 0.94)
	γ	1	−0.021 (0.128, 0.130, 0.93)	γ	1	−0.002 (0.131, 0.130, 0.94)	γ	1	−0.003 (0.126, 0.124, 0.94)
	ρ	0	0.052 (0.083, 0.088, −)	ρ	0.5	0.025 (0.162, 0.157, 0.96)	ρ	1	0.065 (0.261, 0.246, 0.95)
$\xi \sim$ Po	β_1	1	0.001 (0.161, 0.161, 0.96)	β_1	1	0.016 (0.148, 0.144, 0.94)	β_1	1	0.006 (0.143, 0.131, 0.92)
	β_2	1	0.003 (0.159, 0.159, 0.95)	β_2	1	0.013 (0.143, 0.143, 0.96)	β_2	1	0.006 (0.137, 0.130, 0.94)
	γ	0.6	−0.006 (0.028, 0.028, 0.93)	γ	0.6	−0.006 (0.029, 0.028, 0.94)	γ	0.6	−0.003 (0.029, 0.028, 0.94)
	ρ	0	0.048 (0.073, 0.083, −)	ρ	0.5	0.025 (0.149, 0.142, 0.96)	ρ	1	0.063 (0.262, 0.235, 0.95)
	β_1	1	−0.005 (0.146, 0.141, 0.95)	β_1	1	0.000 (0.131, 0.126, 0.95)	β_1	1	−0.003 (0.124, 0.120, 0.95)
	β_2	1	−0.006 (0.144, 0.141, 0.95)	β_2	1	0.003 (0.127, 0.126, 0.95)	β_2	1	−0.005 (0.124, 0.119, 0.94)
	γ	0.8	−0.004 (0.033, 0.031, 0.94)	γ	0.8	−0.002 (0.032, 0.031, 0.93)	γ	0.8	0.000 (0.032, 0.031, 0.94)
	ρ	0	0.042 (0.068, 0.066, −)	ρ	0.5	0.035 (0.156, 0.147, 0.95)	ρ	1	0.090 (0.284, 0.261, 0.95)
$N_i = 4, n = 250$									
$\xi \sim$ Ga	β_1	1	−0.004 (0.138, 0.137, 0.95)	β_1	1	0.005 (0.126, 0.122, 0.94)	β_1	1	−0.006 (0.113, 0.113, 0.94)
	β_2	1	−0.010 (0.137, 0.136, 0.94)	β_2	1	0.007 (0.121, 0.121, 0.94)	β_2	1	−0.008 (0.113, 0.113, 0.95)
	γ	0.5	0.013 (0.081, 0.079, 0.94)	γ	0.5	0.011 (0.078, 0.076, 0.93)	γ	0.5	0.004 (0.075, 0.074, 0.94)
	ρ	0	0.043 (0.065, 0.070, −)	ρ	0.5	0.033 (0.157, 0.144, 0.95)	ρ	1	0.079 (0.265, 0.244, 0.95)
	β_1	1	0.002 (0.151, 0.147, 0.94)	β_1	1	−0.006 (0.138, 0.129, 0.93)	β_1	1	−0.004 (0.122, 0.118, 0.94)
	β_2	1	−0.005 (0.153, 0.145, 0.95)	β_2	1	0.001 (0.131, 0.129, 0.94)	β_2	1	−0.007 (0.122, 0.117, 0.93)
	γ	1	0.001 (0.130, 0.126, 0.93)	γ	1	0.004 (0.121, 0.122, 0.95)	γ	1	0.000 (0.124, 0.117, 0.94)
	ρ	0	0.043 (0.068, 0.078, −)	ρ	0.5	0.033 (0.150, 0.142, 0.95)	ρ	1	0.063 (0.239, 0.223, 0.95)
$\xi \sim$ Po	β_1	1	0.003 (0.144, 0.141, 0.96)	β_1	1	0.004 (0.128, 0.127, 0.95)	β_1	1	−0.005 (0.118, 0.116, 0.94)
	β_2	1	0.011 (0.143, 0.140, 0.94)	β_2	1	0.007 (0.130, 0.126, 0.94)	β_2	1	−0.004 (0.119, 0.116, 0.95)
	γ	0.6	−0.005 (0.028, 0.027, 0.93)	γ	0.6	−0.002 (0.026, 0.027, 0.95)	γ	0.6	−0.002 (0.026, 0.026, 0.95)
	ρ	0	0.045 (0.070, 0.082, −)	ρ	0.5	0.022 (0.147, 0.137, 0.95)	ρ	1	0.071 (0.241, 0.228, 0.95)
	β_1	1	−0.005 (0.130, 0.132, 0.95)	β_1	1	−0.003 (0.124, 0.119, 0.93)	β_1	1	−0.006 (0.119, 0.113, 0.93)
	β_2	1	−0.003 (0.133, 0.131, 0.94)	β_2	1	−0.003 (0.121, 0.118, 0.94)	β_2	1	−0.008 (0.114, 0.113, 0.95)
	γ	0.8	−0.004 (0.026, 0.026, 0.95)	γ	0.8	−0.002 (0.027, 0.026, 0.94)	γ	0.8	0.000 (0.026, 0.026, 0.95)
	ρ	0	0.041 (0.062, 0.069, −)	ρ	0.5	0.035 (0.151, 0.143, 0.95)	ρ	1	0.099 (0.277, 0.259, 0.95)

ESD: empirical standard deviation of the parameter estimator; ASE: averaged standard error of the parameter estimator; CP: coverage probability of the nominal 95% confidence intervals.

Table 2
Estimation results for the dental dataset

m^*_{2n}	Frailty	$\widehat{\beta}_1$ (SE)	$\widehat{\beta}_2$ (SE)	$\widehat{\beta}_3$ (SE)	$\widehat{\gamma}$ (SE)	$\widehat{\rho}$ (SE)	AIC
5	Ga$(\gamma^{-1},\gamma^{-1})$	0.0135 (0.1064)	−0.2449 (0.1075)	0.8281 (0.2175)	2.1360 (0.1888)	0.2107 (0.1170)	4367
5	Po(γ)	0.0068 (0.0909)	−0.2972 (0.0944)	0.6176 (0.2545)	0.4927 (0.0203)	0.0001 (0.0004)	4278
3	IG(γ)	0.1094 (0.1421)	−0.3510 (0.1451)	0.5041 (0.3155)	0.0139 (0.0034)	0.2092 (0.0318)	4291