Spline-Based Semiparametric Sieve Maximum Likelihood Method for Over-dispersed Panel Count Data

Lei Hua

Center for Biostatistics in AIDS Research, Harvard School of Public Health
FXB 514, 651 Huntington Avenue,
Boston, MA 02115, U.S.A.

Ying Zhang

Department of Biostatistics, The University of Iowa
C22 GH, 200 Hawkins Drive,
Iowa City, IA 52242, U.S.A.

Summary. In this article we propose to analyze over-dispersed panel count data using a Gamma-Frailty nonhomogeneous Poisson process model. Conditional on a Gamma distributed frailty variable, the cumulative count, $\mathbb{N}(t)$, is assumed to follow a nonhomogeneous Poisson process. Cubic B-spline functions are used to approximate the logarithm of the baseline mean function $\Lambda_0(t)$ in the semiparametric proportional mean model $E(\mathbb{N}(t)|Z) = \Lambda_0(t)e^{\beta_0^T Z}$. The regression parameters and spline coefficients are estimated by maximizing a likelihood with the nuisance over-dispersion parameter, σ^2, replaced by a method of moment estimate. The asymptotic properties of the proposed maximum likelihood estimator, including its consistency, convergence rate and the asymptotic normality of the estimated regression parameters, are studied using modern empirical process theory. A spline-based least-squares standard error estimator is developed to facilitate a robust inference of the regression parameters. Simulation studies are conducted to investigate finite sample performance of the proposed method. Finally, the proposed method is applied to the data from a bladder tumor clinical trial.

Keywords: Counting process; Gamma-Frailty; Monotone B-splines; Over-dispersion; Panel count data; Semiparametric model;

1. Introduction

Panel count data are a special type of recurrent event data in which only number of events at some discrete observation times are collected. Such data are often seen in clinical trials where it is impossible or impractical to monitor the disease progression continuously. A motivating example is the bladder tumor randomized clinical trial conducted by the Veterans Administration Cooperative Urological Research Group (Byar et al., 1980). Patients with superficial bladder tumor were randomized into one of the three arms: placebo, pyridoxine pills and thiotepa instillation. Many patients have multiple recurrence of tumor and at each follow-up time the number of recurrent tumors was counted, the tumors were removed and the original treatment was continued. The number of observations and observation times vary from subject to subject. The primary objective of this trial is to determine the treatment effect on suppressing the tumor recurrence.

Different approaches have been proposed in literature to analyze panel count data illustrated by this bladder tumor clinical trial by, for example, Byar et al. (1980), Wei et al.

(1989), Sun and Wei (2000), Zhang (2002), Wellner and Zhang (2000, 2007) and Lu et al. (2009). Specifically, semiparametric analysis with the proportional mean model

$$E(\mathbb{N}(t)|Z) = \Lambda_0\,(t)\,e^{\beta_0^T Z} \tag{1}$$

where Λ_0 is the nondecreasing baseline mean function and β_0 is a d-dimensional vector of regression parameters for time-independent covariate Z, has been widely accepted as a robust model for regression analysis of recurrent event data by, for example, Lawless and Nadeau (1995), Sun and Wei (2000), Lin et al. (2000) and Sun et al. (2005).

Using nonhomogeneous Poisson process, Wellner and Zhang (2007) proposed two likelihood-based estimators, maximum pseudo-likelihood estimator (MPLE) and maximum likelihood estimator (MLE), for the analysis of panel count data under the proportional mean model (1). They proved the consistency and derived the convergence rate of both estimators even though the true underlying counting process may be misspecified. They also showed that in spite of the fact that the nonparametric estimator of the baseline mean function converges at a slower rate $n^{1/3}$, the estimated regression parameters still converge at the standard rate $n^{1/2}$ and are asymptotically normally distributed. Incorporating some correlation between the cumulative counts, the MLE is more efficient than the MPLE at the cost of more computing burden. Lu et al. (2009) studied the spline-based sieve version of the two estimators of Wellner and Zhang (2007) by approximating the logarithm of baseline mean function using monotone B-spline functions, i.e.,

$$\log \Lambda_0\,(t) \approx \sum_{l=1}^{q_n} \alpha_l B_l\,(t)$$

subject to a monotone constraint, i.e., $\alpha_1 \le \alpha_2 \le \cdots \le \alpha_{q_n}$. The monotonicity of the spline coefficients $\alpha_l, l = 1, \cdots, q_n$ guarantees the monotone nondecreasing property of the estimated baseline mean function (Schumaker, 1981). With such approximation, (1) can be reparameterized as

$$E(\mathbb{N}(t)|Z) = \Lambda_0\,(t)\,e^{\beta_0^T Z} \approx \exp\left(\sum_{l=1}^{q_n} \alpha_l B_l\,(t) + \beta_0^T Z\right) \tag{2}$$

The number of the B-spline basis functions is often chosen much smaller than the sample size. Compared to their counterparts in Wellner and Zhang (2007), the sieve estimators based on this approximation not only show numerical advantage but also converge in a faster rate. Both the MLE and sieve-MLE were shown semiparametric efficient (Wellner and Zhang, 2007; Lu et al., 2009), if the Poisson process is the true underlying counting process for panel count data. When the Poisson process is misspecified for the data, the estimators may not be very efficient.

In this article, we adopt the spline-based approximation of the proportional mean given in (2) and consider a new semiparametric sieve-MLE using a Gamma-Frailty nonhomogeneous Poisson process for the underlying counting process. We assume that conditional on a Gamma-distributed frailty variable, the cumulative counts follow a nonhomogeneous Poisson process. The variance of the gamma distribution is referred to as over-dispersion parameter and helps accommodate the over-dispersion as well as the correlations between increments in non-overlapping intervals. This model appears more reasonable than the nonhomogeneous Poisson process model in biomedical applications of longitudinal count data.

The estimates can be computed by a two-stage algorithm: for the first stage, the over-dispersion parameter is estimated using the method of moment proposed by Zeger (1988); for the second stage, the estimates of the regression parameters and the spline coefficients are obtained by maximizing the likelihood with the over-dispersion parameter replaced by its method of moment estimate. We show that the new sieve-MLE is consistent and could converge at their optimal convergence rate in the nonparametric/semiparametric regression setting. The asymptotic normality of the estimated regression parameters is also established by modifying the theorem developed by Wellner and Zhang (2007) to handle the presence of an additional over-dispersion parameter. The asymptotic variance of the estimated regression parameters depends on the limiting value of the estimated over-dispersion parameter and can be estimated using a spline-based least-squares estimation method. The new sieve-MLE is generally more efficient than the sieve-MLE studied by Lu et al. (2009) when over-dispersion is present.

The rest of the paper is organized as follows: Section 2 introduces the model based on the Gamma-Frailty nonhomogeneous Poisson assumption; Section 3 discusses the two-stage algorithm that is applied to compute the new sieve-MLE; Section 4 describes the asymptotic properties of the new sieve-MLE; Section 5 presents a spline-based least-squares method to estimate the standard error of the estimated regression parameters; Section 6 provides numerical results including simulation studies and an application to the bladder tumor example; We conclude our paper with some remarks in Section 7. Some technical details are included in the Appendices.

2. Spline-Based Sieve Maximum Likelihood Method

Assume $\mathbb{N} = \{\mathbb{N}(t) : t \geq 0\}$ is a univariate counting process with the conditional proportional mean given by (1). There are K random observations of this counting process at $0 \equiv T_{K,0} < T_{K,1} < \cdots < T_{K,K}$. We denote $\underline{T}_K \equiv (T_{K,1}, T_{K,2}, \cdots, T_{K,K})$ and $\mathbb{N} \equiv (\mathbb{N}(T_{K,1}), \mathbb{N}(T_{K,2}), \cdots, \mathbb{N}(T_{K,K}))$. We further assume conditional on the time-independent covariates Z, $(K, \underline{T}_K)$ is independent of the underlying counting process. Panel count data compose of a random sample of $X_i, i = 1, \cdots, n$ where $X_i = \left(K_i, \underline{T}_{K_i}^{(i)}, \mathbb{N}^{(i)}, Z_i\right)$ with $\mathbb{N}^{(i)} = \left(\mathbb{N}^{(i)}\left(T_{K_i,1}^{(i)}\right), \mathbb{N}^{(i)}\left(T_{K_i,2}^{(i)}\right), \cdots, \mathbb{N}^{(i)}\left(T_{K_i,K_i}^{(i)}\right)\right)$.

Throughout the rest of this paper, we study the Gamma-Frailty nonhomogeneous Poisson process model. That is, conditional on the frailty variable γ the counting process $\mathbb{N}(t)$ follows the nonhomogeneous Poisson process with mean function $\gamma\Lambda_0(t)\, e^{\beta_0^T Z}$ and the frailty variable follows a gamma distribution with mean 1 and variance σ^2, i.e., $\gamma \sim \Gamma\left(1/\sigma^2, 1/\sigma^2\right)$. This parametrization of Gamma distribution guarantees the identifiability of the model with the unconditional proportional mean structure specified in (1). The conditional likelihood of the counts given the frailty variable can be written as

$$f(\mathbb{N}_1, \mathbb{N}_2, \cdots, \mathbb{N}_K | \gamma) = \Pi_{j=1}^{K} \frac{e^{-\gamma \Delta\Lambda_j e^{\beta^T Z}} \left(\gamma \Delta\Lambda_j e^{\beta^T Z}\right)^{\Delta\mathbb{N}_j}}{\Delta\mathbb{N}_j!}$$

where $\mathbb{N}_j = \mathbb{N}(T_{K,j}), \Delta\mathbb{N}_j = \mathbb{N}_j - \mathbb{N}_{j-1}$ and $\Lambda_j = \Lambda(T_{K,j}), \Delta\Lambda_j = \Lambda_j - \Lambda_{j-1}$ for $j =$

$1, 2, \cdots, K$. We assume $\mathbb{N}(0) = \Lambda(0) = 0$. Integrating out γ, we have

$$f(\mathbb{N}_1, \mathbb{N}_2, \cdots, \mathbb{N}_K) = \frac{\Pi_{j=1}^K \left(\Delta\Lambda_j e^{\beta^T Z}\right)^{\Delta\mathbb{N}_j} \left(1/\sigma^2\right)^{1/\sigma^2}}{\Pi_{j=1}^K \Delta\mathbb{N}_j! \Gamma\left(1/\sigma^2\right)} \frac{\Gamma\left(\mathbb{N}_K + 1/\sigma^2\right)}{\left(\Lambda_K e^{\beta^T Z} + 1/\sigma^2\right)^{\mathbb{N}_K + 1/\sigma^2}}$$

The log likelihood based on a single observation X subject to an additive constant is,

$$\begin{aligned} m\left(\beta, \Lambda, \sigma^2; X\right) = & \sum_{j=1}^K \Delta\mathbb{N}_j \log\left(\Delta\Lambda_j e^{\beta^T Z_i}\right) - \left(\mathbb{N}_K + 1/\sigma^2\right) \times \log\left(\Lambda_K e^{\beta^T Z} + 1/\sigma^2\right) \\ & + 1/\sigma^2 \times \log\left(1/\sigma^2\right) + \log\Gamma\left(\mathbb{N}_K + 1/\sigma^2\right) - log\Gamma\left(1/\sigma^2\right) \end{aligned} \qquad (3)$$

We propose to approximate the baseline mean function $\Lambda_0(t)$ using cubic B-splines. This idea of sieve-MLE was originally proposed by Geman and Hwang (1982) and the spline-based semiparametric sieve-MLE for the analysis of panel count data was adopted by Lu et al. (2009). Assume the observation times are restricted to a closed interval $[L, U]$. $[L, U]$ can be divided into $m_n + 1$ subintervals which form a sequence of spline knots

$$\Xi = \left\{L \equiv \xi_0 = \xi_1 \cdots = \xi_l < \xi_{l+1}, \cdots < \xi_{m_n+l} < \xi_{m_n+l+1} \cdots = \xi_{m_n+2l} \equiv U\right\},$$

where l is the order of B-splines and $l = 4$ corresponds to cubic B-splines. In this article, we approximate the natural logarithm of the smooth baseline mean function $\log\Lambda_0(t)$ by $\sum_{i=1}^{q_n} \alpha_i B_i(t)$ and jointly estimate the regression parameter β and spline coefficients $\alpha = (\alpha_1, \cdots, \alpha_{q_n})$ subjecting to the monotone constraints, $\alpha_1 \leq \alpha_2 \leq \cdots, \alpha_{q_n}$, by maximizing the approximated log likelihood,

$$\begin{aligned} l\left(\beta, \alpha, \sigma^2\right) = & \sum_{i=1}^n \left\{ \sum_{j=1}^{K_i} \Delta\mathbb{N}_{K_i,j}^{(i)} \log\left[\exp\left(\sum_{l=1}^{q_n} \alpha_l B_l\left(t_{K_i,j}^{(i)}\right)\right) - \exp\left(\sum_{l=1}^{q_n} \alpha_l B_l\left(t_{K_i,j-1}^{(i)}\right)\right)\right] \right. \\ & + N_{K_i,K_i}^{(i)} \beta^T Z_i - \left(\mathbb{N}_{K_i,K_i}^{(i)} + 1/\sigma^2\right) \log\left[\exp\left(\sum_{l=1}^{q_n} \alpha_l B_l\left(t_{K_i,K_i}^{(i)}\right) + \beta^T Z_i\right)\right. \\ & \left. +1/\sigma^2\right] + 1/\sigma^2 \log\left(1/\sigma^2\right) + \log\Gamma\left(\mathbb{N}_{K_i,K_i}^{(i)} + 1/\sigma^2\right) - \log\Gamma\left(1/\sigma^2\right) \Big\}. \end{aligned} \qquad (4)$$

Here $B_i(t), i = 1, \cdots q_n$ are B-spline basis functions corresponding to the spline knots defined in Ξ and q_n is the sum of m_n and the order of the B-spline functions l. Based on the variation diminishing property of B-splines (Schumaker, 1981), monotone constraints of the spline coefficients, i.e., $\alpha_1 \leq \alpha \leq \cdots, \leq \alpha_{q_n}$, can result in such B-splines approximation to be monotone as well.

3. Computing Algorithm

In the model discussed in Section 2, the parameters in the proportional mean structure are of primary interest, the over-dispersion parameter σ^2 is treated as an extra nuisance parameter. Both parameters could be computed simultaneously by maximizing the likelihood shown in (4). However this involves maximizing a gamma function and is computationally cumbersome. In this article, we propose a two-stage estimating algorithm. In the first stage, the nuisance parameter σ^2 is estimated using the method of moment proposed by

Zeger (1988), i.e.,

$$\hat{\sigma}_n^2 = \frac{\sum_{i=1}^{n} \sum_{j=1}^{K_i} \left\{ \left(\mathbb{N}_{ij}^{(i)} - \hat{\mu}_{ij} \right)^2 - \hat{\mu}_{ij} \right\}}{\sum_{i=1}^{n} \sum_{j=1}^{K_i} \hat{\mu}_{ij}^2} \tag{5}$$

where $\mathbb{N}_{ij}^{(i)} = \mathbb{N}^{(i)} \left(T_{K_i,j}^{(i)} \right)$ and $\hat{\mu}_{ij}$ is a consistent estimate of $E\left(\mathbb{N}_{ij}^{(i)}\right)$ for $j = 1, \cdots, K_i; i = 1, \cdots, n$, based on the semiparametric MPLE studied by Zhang (2002) and Wellner and Zhang (2007) due to its computational convenience. $\hat{\sigma}_n^2$ is a $\sqrt{n}$-consistent estimate of σ_0^2 satisfying the equation:

$$Var\left(\mathbb{N}(T)\right) = E\left(\mathbb{N}(T)\right) + \sigma_0^2 \left\{ E\left(\mathbb{N}(T)\right) \right\}^2,$$

that is $\sqrt{n}\left(\hat{\sigma}_n^2 - \sigma_0^2\right) = O_p(1)$. The proof is given in Appendix A.1. In the second stage, the parameters in the proportional mean structure are estimated by maximizing the spline-based sieve likelihood (4), where σ^2 is replaced by the method of moment estimate given by (5).

A hybrid algorithm of Newton-Raphson iteration and Isotonic regression (NR/IR) is used to find the maximizer of spline-based sieve likelihood subject to the monotone constraints on the spline coefficients, $\alpha_1 \leq \alpha_2 \leq \cdots, \leq \alpha_{q_n}$. Newton-Raphson (NR) algorithm is widely used in optimization of convex nonlinear functions. It converges to the true value in a quadratic rate numerically. However it cannot guarantee the monotonicity of the solution. So after each NR iteration, we project the estimates to the cone depicted by the monotone constraints using isotonic regression. That is, with each NR update $\{\tilde{\alpha}_i, i = 1, 2, \cdots, q_n\}$, we find $\{\hat{\alpha}_i, i = 1, 2, \cdots, q_n\}$ such that

$$\{\hat{\alpha}_i, i = 1, 2, \cdots, q_n\} = \underset{\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_{q_n}}{\operatorname{argmin}} \sum_{i=1}^{q_n} w_i \left(\alpha_i - \tilde{\alpha}_i\right)$$

We choose $w_i, i = 1, 2, \cdots, q_n$ to be the diagonal elements of the negative Hessian matrix that correspond to $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_{q_n})$. The solution of this optimization has a nice interpretation (Groeneboom and Wellner, 1992): it is the left derivative of the greatest convex minorant of the cumulative sum diagram $\{P_i, i = 0, 1, \cdots, n\}$ where

$$P_0 = (0, 0) \text{ and } P_i = \left(\sum_{l=1}^{i} w_l, \sum_{l=1}^{i} w_l \tilde{\alpha}_l \right);$$

and can be expressed as

$$\hat{\alpha}_i = \max_{j<i} \min_{l>i} \frac{\sum_{m=j}^{l} w_m \tilde{\alpha}_m}{\sum_{m=j}^{l} w_m}$$

The NR/IR algorithm tailored to the spline-based sieve semiparametric maximum likelihood estimation is summarized in the following steps.

Step 0 (Estimation of σ_0^2): Estimate the over-dispersion parameter σ^2 using the method of moments estimate $\hat{\sigma}_n^2$ given by (5) with $\hat{\mu}_{ij}$ computed using the semiparametric MPLE (Wellner and Zhang, 2007).

Iterate the algorithm through the following steps until convergence.

Step 1 (Newton-Raphson Update): Start with an initial point $\hat{\theta}^{(0)} = \left(\hat{\alpha}^{(0)}, \hat{\beta}^{(0)}\right)$ that satisfies the monotone constraints of the spline coefficients, i.e., $\hat{\alpha}^{(0)} = \left(\hat{\alpha}_1^{(0)}, \hat{\alpha}_2^{(0)}, \cdots, \hat{\alpha}_{q_n}^{(0)}\right), \hat{\alpha}_1^{(0)} \le \hat{\alpha}_2^{(0)} \le \cdots \le \hat{\alpha}_{q_n}^{(0)}$. Update the current estimates $\hat{\theta}^{(k)} = \left(\hat{\alpha}^{(k)}, \hat{\beta}^{(k)}\right)$ by Newton-Raphson algorithm with step-halving line search,

$$\tilde{\theta}^{(k+1)} = \left(\tilde{\alpha}^{(k+1)}, \tilde{\beta}^{(k+1)}\right) = \hat{\theta}^{(k)} - (1/2)^r H^{-1}\left(\hat{\theta}^{(k)}; \hat{\sigma}_n^2\right) \dot{l}\left(\hat{\theta}^{(k)}; \hat{\sigma}_n^2\right).$$

where $\dot{l}\left(\hat{\theta}^{(k)}; \hat{\sigma}_n^2\right)$ is the gradient and $H^{-1}\left(\hat{\theta}^{(k)}; \hat{\sigma}_n^2\right)$ is the inverse of Hessian matrix of the log likelihood in (4) evaluated at $\hat{\theta}^{(k)}$. r is the smallest integer starting from 0 such that

$$l\left(\hat{\theta}^{(k)} - (1/2)^r H^{-1}\left(\hat{\theta}^{(k)}; \hat{\sigma}_n^2\right) \dot{l}\left(\hat{\theta}^{(k)}; \hat{\sigma}_n^2\right); \hat{\sigma}_n^2\right) > l\left(\hat{\theta}^{(k)}; \hat{\sigma}_n^2\right)$$

Step 2 (Isotonic Regression): Construct the cumulative sum diagram $\{P_i, i = 0, 1, \cdots, n\}$ with

$$P_0 = (0,0) \text{ and } P_i = \left(\sum_{l=1}^{i} w_l, \sum_{l=1}^{i} w_l \tilde{\alpha}_l^{(k+1)}\right);$$

where $w_l, l = 1, 2, \cdots, q_n$ are the diagonal elements of the negative Hessian matrix, $-H\left(\hat{\theta}^{(k)}; \hat{\sigma}_n^2\right)$ that correspond to α. The monotonic update of $\hat{\alpha}_i^{(k)}$ is then calculated by

$$\hat{\alpha}_i^{(k+1)} = \max_{j<i} \min_{l>i} \frac{\sum_{m=j}^{l} w_m \tilde{\alpha}_m^{(k+1)}}{\sum_{m=j}^{l} w_m}$$

Since there is no constraints on β, let $\hat{\beta}^{(k+1)} = \tilde{\beta}^{(k+1)}$.
Step 3 (Check the convergence): Let $d = \|\hat{\theta}^{(k+1)} - \hat{\theta}^{(k)}\|$, if $d < \varepsilon$ for a small $\varepsilon > 0$ stop the algorithm, otherwise go back to step 1.

4. Asymptotic Properties

In this section, we study the asymptotic properties of the spline-based semiparametric sieve-MLE. First we introduce some notations used in Wellner and Zhang (2007) and Lu et al. (2009). Let $\mathcal{B}_d$ and $\mathcal{B}$ denote the collection of Borel sets in $\mathbb{R}^d$ and $\mathbb{R}$, respectively, and let $\mathcal{B}_{[0,\tau]} = \{B \cap [0,\tau] : \ B \in \mathcal{B}\}$. We define a measure μ and a L_2-metric d as following: for $B \in \mathcal{B}_{[0,\tau]}$ and $C \in \mathcal{B}_d$,

$$\mu(B \times C) = \int_C \sum_{k=1}^{\infty} P(K = k|Z = z) \sum_{j=1}^{k} P(T_{k,j} \in B|K = k, Z = z) dP(z)$$

and

$$d(\theta_1, \theta_2) = \left\{|\beta_2 - \beta_1|^2 + \int |\Lambda_2(t) - \Lambda_1(t)|^2 d\mu(t)\right\}^{1/2}. \tag{6}$$

We assume the following regularity conditions for the model:

Condition 1. The true parameter $\left(\beta_0, \Lambda_0, \sigma_0^2\right) \in \overset{\circ}{\mathcal{R}}^d \times \mathcal{F} \times \overset{\circ}{\mathcal{R}}^+$, where $\overset{\circ}{\mathcal{R}}^d$ and $\overset{\circ}{\mathcal{R}}^+$ are the interior of some compact set of $\mathcal{R}^d$ and $\mathcal{R}^+$ in $\mathbb{R}^d$ and $\mathbb{R}^+$, respectively. $\mathcal{F}$ is a class of monotone nondecreasing functions.

Condition 2. The observation times $T_{K,j} : j = 1, 2, \cdots, K, K = 1, 2, \cdots$ are bounded in $S[T] = \{t : \delta < t < \tau\}$ for some $\delta > 0$ and $\tau > 0$, $\Lambda_0(\delta) > 0$ and $P\left(T_{K,j} - T_{K,j-1} \geq s_0\right) = 1$ for some constant s_0. $P\left(K \leq k_0\right) = 1$ for some constant k_0.

Condition 3. The true baseline mean function Λ_0 is p^{th} differentiable and bounded. The derivatives have positive and finite lower and upper bounds in $S[T]$.

Condition 4. For some $\eta \in (0, 1), a^T Var(Z|U, V)a \geq \eta a^T E(ZZ^T|U, V)a$ a.s. for all $a \in \mathbb{R}^d$ where (U, V, Z) follows distribution $\mu/\mu(\mathbb{R}^{+2} \times \mathcal{Z})$.

Condition 5. The covariate Z is bounded, i.e., $P\left(|Z| \leq z_0\right) = 1$ for some constant z_0. And $P\left(aZ \neq c\right) > 0$ for any $a \in \mathbb{R}^d$ and $c \in \mathbb{R}$.

Condition 6. $E\left\{e^{C\mathbb{N}(t)}\right\}$ is uniformly bounded for $t \in S[T]$.

Condition 7. The measure μ is absolutely continuous with respect to Lebesgue measure with a finite lower bound in the observation interval $S[T]$.

For the spline-based sieve-MLE, we also need to properly allocate the spline knots to warrant a good approximation of B-splines to a smooth function.

Condition 8. The number of internal knots $m_n = O\left(n^{\nu}\right)$ for $0 < \nu < 1/2$. The maximum spacing of the knots satisfies $\Delta_{\max} = \max_{l+1 \leq i \leq m_n + l + 1} |\xi_i - \xi_{i-1}| = O(n^{-\nu})$. Further denote $\Delta_{\min} = \min_{l+1 \leq i \leq m_n + l + 1} |\xi_i - \xi_{i-1}|$, there exists a constant $M > 0$ such that $\Delta_{\max}/\Delta_{\min} \leq M$ uniformly in n.

Theorem 4.1 (Consistency). *Suppose that Conditions 1-3,5, 7 and 8 hold and the counting process* $\mathbb{N}$ *satisfies the proportional mean regression model. Then given an estimate of the over-dispersion parameter* $\hat{\sigma}_n^2$ *such that* $\hat{\sigma}_n^2 \rightarrow_p \sigma_0^2$*,*

$$d\left(\left(\hat{\beta}_n, \hat{\Lambda}_n\right), \left(\beta_0, \Lambda_0\right)\right) \rightarrow_p 0$$

Theorem 4.2 (Convergence Rate). *Suppose that Conditions 1-8 hold and the counting process* $\mathbb{N}$ *satisfies the proportional mean regression model. Then given a* $\sqrt{n}-$*consistent estimate of the over-dispersion parameter* σ_0^2*,* $\hat{\sigma}_n^2$*,*

$$d\left(\left(\hat{\beta}_n, \hat{\Lambda}_n\right), \left(\beta_0, \Lambda_0\right)\right) = O_p\left(n^{-min(p\nu,(1-\nu)/2)}\right).$$

Remark 4.1. *This theorem implies that when* $\nu = 1/\left(2p + 1\right)$*,*

$$d\left(\left(\hat{\beta}_n, \hat{\Lambda}_n\right), \left(\beta_0, \Lambda_0\right)\right) = O_p\left(n^{-p/(2p+1)}\right)$$

which is the optimal convergence rate in the nonparametric regression setting. When the baseline mean function is smooth, i.e. $p > 1$*, the spline-based semiparametric sieve-MLE can converge faster than the conventional semiparametric estimators using step functions to estimate the baseline mean function considered by Wellner and Zhang (2007).*

Theorem 4.3 describes the asymptotic normality of the estimated regression parameters despite a slower convergence rate of the estimated baseline mean function. We use similar notations as those in Huang (1996) and Wellner and Zhang (2007) with the objective function $m\left(\beta, \Lambda, \sigma^2; X\right)$ taken as the log likelihood specified in (3). Suppose that Λ_η is a parametric path in the monotone nondecreasing function class $\mathcal{F}$ through Λ, i.e. $\Lambda_\eta \in \mathcal{F}$, and $\Lambda_\eta|_{\eta=0} = \Lambda$. Let $\mathcal{H} = \left\{h : h = \frac{\partial \Lambda_\eta}{\partial \eta}|_{\eta=0}\right\}$ and for any $h \in \mathcal{H}$, we define

$$\begin{aligned}
m_1\left(\beta, \Lambda, \sigma^2; x\right) &= \bigtriangledown_\beta m\left(\beta, \Lambda, \sigma^2; x\right) \\
&\equiv \left(\frac{\partial m\left(\beta, \Lambda, \sigma^2; x\right)}{\partial \beta_1}, \cdots, \frac{\partial m\left(\beta, \Lambda, \sigma^2; x\right)}{\partial \beta_d}\right)^T, \\
m_2\left(\beta, \Lambda, \sigma^2; x\right)[h] &= \frac{\partial m\left(\beta, \Lambda_\eta, \sigma^2; x\right)}{\partial \eta}|_{\eta=0}, \\
m_{11}\left(\beta, \Lambda, \sigma^2; x\right) &= \bigtriangledown^2_\beta m\left(\beta, \Lambda, \sigma^2; x\right), \\
m_{12}\left(\beta, \Lambda, \sigma^2; x\right)[h] &= \frac{\partial m_1\left(\beta, \Lambda_\eta, \sigma^2; x\right)}{\partial \eta}|_{\eta=0}, \\
m_{21}\left(\beta, \Lambda, \sigma^2; x\right)[h] &= \bigtriangledown_\beta m_2\left(\beta, \Lambda, \sigma^2; x\right)[h], \\
m_{22}\left(\beta, \Lambda, \sigma^2; x\right)[h_1, h_2] &= \frac{\partial^2 m\left(\beta, \Lambda_{\eta_j}, \sigma^2; x\right)}{\partial \eta^2}|_{\eta_j=0, j=1,2} \\
&\equiv \frac{\partial m_2\left(\beta, \Lambda_{\eta_2}, \sigma^2; x\right)[h_1]}{\partial \eta_2}.
\end{aligned}$$

Let $\mathbb{P}_n$ denote the ordinary empirical measure defined by $\mathbb{P}_n f(X) = \frac{1}{n}\sum_{i=1}^n f(X_i)$ and $\mathbb{G}_n$ the centered empirical measure defined by $\mathbb{G}_n f(X) = \sqrt{n}(\mathbb{P}_n - P)f(x) = \frac{1}{\sqrt{n}}\sum_{i=1}^n (f(X_i) - Ef(X))$.

Theorem 4.3 (Asymptotic Normality). *Suppose that Conditions 1-8 hold and the counting process* $\mathbb{N}$ *satisfies the proportional mean regression model. Then given a* $\sqrt{n}-$*consistent estimate of the over-dispersion parameter* σ_0^2, $\hat{\sigma}_n^2$,

$$\begin{aligned}
\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) &= A_0^{-1}\mathbb{G}_n\left(m_1\left(\beta_0, \Lambda_0, \sigma_0^2\right) - m_2\left(\beta_0, \Lambda_0, \sigma_0^2\right)[h^*]\right) + o_p(1) \\
&\to_d N\left(0, A_0^{-1}B_0A_0^{-1}\right)
\end{aligned}$$

where

$$\begin{aligned}
A_0 = A\left(\beta_0, \Lambda_0, \sigma_0^2\right) &= -P\left(m_{11}\left(\beta_0, \Lambda_0, \sigma_0^2\right) - m_{21}\left(\beta_0, \Lambda_0, \sigma_0^2\right)[h^*]\right) \\
B_0 = B\left(\beta_0, \Lambda_0, \sigma_0^2\right) &= P\left(m_1\left(\beta_0, \Lambda_0, \sigma_0^2\right) - m_2\left(\beta_0, \Lambda_0, \sigma_0^2\right)[h^*]\right)^{\otimes 2}
\end{aligned}$$

and $h^* = (h_1^*, h_2^*, \cdots, h_d^*)^T$ *with*

$$h_s^* = \Lambda_0 \times \frac{E\left(\frac{Z_s \times 1/\sigma_0^2}{\Lambda_{0,K}e^{\beta_0^T Z} + 1/\sigma_0^2} \times e^{\beta_0^T Z}|K, \underline{T}_K\right)}{E\left(e^{\beta_0^T Z}|K, \underline{T}_K\right) - E\left(\frac{\Lambda_{0,K}e^{2\beta_0^T Z}}{\Lambda_{0,K}e^{\beta_0^T Z} + 1/\sigma_0^2}|K, \underline{T}_K\right)}, \quad s = 1, \cdots d \tag{7}$$

We prove Theorems 1, 2 and 3 by verifying the conditions of Theorem 5.7 in van der Vaart (1998), Theorem 3.4.1 of van der Vaart and Wellner (1996) and modifying Theorem 6.1 in Wellner and Zhang (2007), respectively. The sketch of the proofs are given in Appendix A.2.

5. A consistent estimate of standard error for the estimated regression parameters

In order to make inference about the regression parameters based on their asymptotic normality, we need to estimate $h^* = (h_1^*, \cdots, h_d^*)^T$. As shown in the proof of Theorem 4.3 in Appendix A.2, finding h_s^* is the projection problem of solving each component of h^* by

$$h_s^* = \underset{h \in \mathcal{H}}{\operatorname{argmin}}\, P\left(m_{1,s}\left(\beta_0, \Lambda_0, \sigma_0^2; X\right) - m_2\left(\beta_0, \Lambda_0, \sigma_0^2; X\right)[h]\right)^2 \tag{8}$$

for $s = 1, \cdots, d$ where $m_{1,s}$ is the s^{th} component of m_1. We apply the spline-based sieve method again and estimate h_s^* by a linear span of cubic B-spline basis functions, e.g., $\hat{h}_{n,s} = \sum_{j=1}^{q_n} \gamma_{j,s} B_j$ for $s = 1, 2, \cdots, d$ where $\gamma_{j,s}, j = 1, \cdots, q_n$ are estimated by minimizing the empirical version of (8), namely,

$$\mathbb{P}_n \left(m_{1,s}\left(\hat{\beta}_n, \hat{\Lambda}_n, \hat{\sigma}_n^2; X\right) - m_2\left(\hat{\beta}_n, \hat{\Lambda}_n, \hat{\sigma}_n^2; X\right)[\hat{h}_{n,s}]\right)^2,$$

where $\hat{\beta}_n, \hat{\Lambda}_n$ and $\hat{\sigma}_n^2$ are consistent estimates of β_0, Λ_0 and σ_0^2, respectively. Since m_2 is a bilinear operator, this is equivalent to solving a least-squares problem and the solution of $\underline{\gamma}_s = (\gamma_{1,s}, \gamma_{2,s}, \cdots, \gamma_{q_n,s})^T$ is given by

$$\left(m_2^T\left(\hat{\beta}_n, \hat{\Lambda}_n, \hat{\sigma}_n^2; X\right)[B] \times m_2\left(\hat{\beta}_n, \hat{\Lambda}_n, \hat{\sigma}_n^2; X\right)[B]\right)^{-1} \times \\ \left(m_2^T\left(\hat{\beta}_n, \hat{\Lambda}_n, \hat{\sigma}_n^2; X\right)[B] \times m_{1,s}\left(\hat{\beta}_n, \hat{\Lambda}_n, \hat{\sigma}_n^2; X\right)\right)$$

where $m_2\left(\hat{\beta}_n, \hat{\Lambda}_n, \hat{\sigma}_n^2; X\right)[B]$ is the $n \times q_n$ design matrix with $(i, m)^{th}$ entry being

$$\sum_{j=1}^{K_i} \frac{\mathbb{N}^{(i)}\left(t_{K_i,j}^{(i)}\right) - \mathbb{N}^{(i)}\left(t_{K_i,j-1}^{(i)}\right)}{\hat{\Lambda}_n\left(t_{K_i,j}^{(i)}\right) - \hat{\Lambda}_n\left(t_{K_i,j-1}^{(i)}\right)} \left(B_m\left(t_{K_i,j}^{(i)}\right) - B_m\left(t_{K_i,j-1}^{(i)}\right)\right) - \\ \frac{\mathbb{N}^{(i)}\left(t_{K_i,K_i}^{(i)}\right) + 1/\hat{\sigma}_n^2}{\hat{\Lambda}_n\left(t_{K_i,K_i}^{(i)}\right) e^{\hat{\beta}_n^T Z_i} + 1/\hat{\sigma}_n^2} e^{\hat{\beta}_n^T Z_i} B_m\left(t_{K_i,K_i}^{(i)}\right)$$

THEOREM 5.1. *Let* $\left(\hat{\beta}_n, \hat{\Lambda}_n, \hat{\sigma}_n^2\right)$ *be a consistent estimate of* $\left(\beta_0, \Lambda_0, \sigma_0^2\right)$ *and* $\hat{h}_{n,s}, s = 1, \cdots, d$, *be the least-squares estimate of* h_s^*. *Denote* $\hat{h}_n = \left(\hat{h}_{n,1}, \hat{h}_{n,2}, \cdots, \hat{h}_{n,d}\right)^T$, *under Conditions 1-3, 5, 6 and 8,* $\hat{h}_n$ *is a consistent estimate of* h^*. *Let*

$$\hat{A}_n = -\mathbb{P}_n\left(m_{11}\left(\hat{\beta}_n, \hat{\Lambda}_n, \hat{\sigma}_n^2; X\right) - m_{21}\left(\hat{\beta}_n, \hat{\Lambda}_n, \hat{\sigma}_n^2; X\right)[\hat{h}_n]\right)$$
$$\hat{B}_n = \mathbb{P}_n\left(m_1\left(\hat{\beta}_n, \hat{\Lambda}_n, \hat{\sigma}_n^2; X\right) - m_2\left(\hat{\beta}_n, \hat{\Lambda}_n, \hat{\sigma}_n^2; X\right)[\hat{h}_n]\right)^{\otimes 2}.$$

Then $\hat{A}_n \rightarrow_p A_0$ *and* $\hat{B}_n \rightarrow_p B_0$.

The spline-based least-squares estimate of standard error in semiparametric regression setting was originally proposed by Huang et al. (2008). The proof of Theorem 5.1 is given in Appendix A.3.

6. Numerical Results

6.1. Simulation Studies

Simulation studies are conducted to examine finite sample performance of the spline-based semiparametric sieve-MLE under the Gamma-Frailty nonhomogeneous Poisson model. For the i^{th} subject, we generate $X_i = \left(K_i, \underline{T}^{(i)}_{K_i}, \mathbb{N}^{(i)}, Z_i\right)$ in a way that may reflect actual observations in a clinical follow-up study. (i) Six follow-up times are pre-scheduled at $T^\circ = \{T^\circ_j : T^\circ_j = 2j, j = 1, \cdots, 6\}$. The actual observation times T°_{ij} are generated from a normal distribution, $N(T^\circ_j, 1/3)$. Let $\xi_{ij} = 1_{[T^\circ_{ij-1} < T^\circ_{ij}]}$, for $j = 1, \cdots, 6$ and $T^\circ_{i0} = 0$. (ii) Let $\delta_{ij} = 1$ if the j^{th} visit actually happens and zero otherwise, with $P(\delta_{ij} = 1) = \frac{1}{1+e^{T^\circ_{ij}-10}}$ indicating the probability of missing visit increasing as follow-up proceeds. The i^{th} subject has $K_i = \sum_{j=1}^{6} \xi_{ij}\delta_{ij}$ observations at $\underline{T}^{(i)}_{K_i} = \left(T^{(i)}_{K_i,1}, T^{(i)}_{K_i,2}, \cdots, T^{(i)}_{K_i,K_i}\right)$, where $T^{(i)}_{K_i,j}$ is the j^{th} ordered observation time of $\{T^\circ_{ij} : \xi_{ij}\delta_{ij} = 1, j = 1, \cdots, 6\}$. (iii) The covariate vector $Z_i = (Z_{i1}, Z_{i2}, Z_{i3})$ is simulated by $Z_{i1} \sim \text{Uniform}\,(0,1)\,, Z_{i2} \sim N\,(0,1)\,,\text{and } Z_{i3} \sim \text{Bernoulli}\,(0.5)$. The regression parameter $\beta_0 = (\beta_{0,1}, \beta_{0,2}, \beta_{0,3})^T = (-1.0, 0.5, 1.5)^T$. (iv) Given $\left(Z_i, K_i, \underline{T}^{(i)}_{K_i}\right)$, different scenarios are used to generate the panel counts

$$\mathbb{N}^{(i)} = \left(\mathbb{N}(i)\left(T^{(i)}_{K_i,1}\right), \mathbb{N}^{(i)}\left(T^{(i)}_{K_i,2}\right), \cdots, \mathbb{N}^{(i)}\left(T^{(i)}_{K_i,K_i}\right)\right).$$

Scenario 1. Data are generated from a Gamma-Frailty Poisson model. The frailty parameters $\gamma_1, \gamma_2, \cdots, \gamma_n$ are randomly drawn from the Gamma distribution, $\Gamma\,(0.5, 0.5)$, giving an over-dispersion parameter of 2. Conditional on the frailty parameter γ_i as well as the covariates Z_i, the panel counts for each subject are drawn from a Poisson process, i.e.

$$\mathbb{N}^{(i)}\left(T^{(i)}_{K_i,j}\right) - \mathbb{N}^{(i)}\left(T^{(i)}_{K_i,j-1}\right)|\gamma_i \sim \text{Poisson}\left\{2\gamma_i\left[\left(T^{(i)}_{K_i,j}\right)^{1/2} - \left(T^{(i)}_{K_i,j-1}\right)^{1/2}\right]e^{\beta_0^T Z_i}\right\} \quad (9)$$

for $j = 1, 2, \cdots, K_i$. In this scenario, the counting process given only the covariate is not a Poisson process. However, the conditional mean given the covariate vector still satisfies the proportional mean model specified in (1). The marginal distribution of the counts is a negative binomial distribution.

Scenario 2. Data are generated from a Lognormal-Frailty Poisson model. A random sample $(\gamma_1, \gamma_2, \cdots, \gamma_n)$ is generated from a lognormal distribution with mean 1 and variance 2. Conditional on each frailty term γ_i, the cumulative counts are drawn from a Poisson process by (9). In this scenario, the proportional mean model (1) still holds. The marginal distribution of the cumulative counts is not a Poisson process, nor does it follow a negative binomial distribution.

Scenario 3. Data are generated from mixture Poisson data similar to those discussed in Wellner and Zhang (2007) and Lu et al. (2009). We first generate a random sample $\eta_1, \cdots, \eta_n$ from $(-0.8, 0, 0.8)$ with probability $(0.25, 0.5, 0.25)$. Given η_i, the cumulative counts are generated from a nonhomogeneous Poisson process with mean $(2 + \eta_i)\,t^{1/2}e^{\beta^T Z_i}$. That is,

$$\mathbb{N}^{(i)}\left(T^{(i)}_{K_i,j}\right) - \mathbb{N}^{(i)}\left(T^{(i)}_{K_i,j-1}\right)|\eta_i \sim \text{Poisson}\left\{(2 + \eta_i)\left[\left(T^{(i)}_{K_i,j}\right)^{1/2} - \left(T^{(i)}_{K_i,j-1}\right)^{1/2}\right]e^{\beta_0^T Z_i}\right\}$$

for $j = 1, 2, \cdots, K_i$. In the frailty formulation, this is equivalent to generating a discrete frailty term γ_i from $\{0.6, 1, 1.4\}$ with probabilities 0.25, 0.5 and 0.25, respectively. Given γ_i, the cumulative counts follow a Poisson process similar to those in *Scenario 1.*

For all three scenarios, monotone cubic B-splines are used to approximate the logarithm of the baseline mean function. The number of interior knots is chosen to be $m_n = \lceil N^{1/3} \rceil$, the smallest integer above $N^{1/3}$, where N is the number of collected distinct observation times $\left\{ \underline{T}_{K_i}^{(i)} : \ 1 \leq i \leq n \right\}$. These knots are placed at the corresponding quantiles of the distinct observation times so Condition 8 is satisfied. In our simulation studies, we generate 1000 Monte Carlo samples with sample size of 50 and 100 for each scenario, respectively. The proposed estimator based on the Gamma-Frailty Poisson process model and the two estimators from Lu et al. (2009), the spline-based sieve-MPLE and sieve-MLE based on Poisson process model are computed and compared.

Table 6.1 summarizes simulation results for the estimated regression parameters in all three scenarios, including their bias (bias), Monte Carlo standard deviation (M-C sd), the average of the estimated standard error (ASE) based on the spline-based least-squares method, and 95% coverage of the regression parameters. Figures 1 plot the squared bias and the variance of the estimated baseline mean function at $t = 2, 2.25, \cdots, 9$ for the corresponding scenarios. When data follow a Gamma-Frailty Poisson process as simulated from *Scenario 1*, all three estimates are consistent. The biases are negligible compared to the standard errors. The estimates based on the Gamma-Frailty Poisson process model take the over-dispersion into account and apparently outperform their alternative estimates based on the Poisson process model in view of the estimation standard error. The spline-based least-squares estimates of the standard error of different maximum likelihood estimates appear to underestimate the true values a little bit when sample size is small, which attributes to the empirical coverage probability lower than the nominal level. The underestimation lessens as sample size increases. Among the three standard error estimates, the sieve-MLE based on the Gamma-Frailty Poisson process model have the least bias. For the estimates of the baseline mean function at different time points, the bias is negligible compared to its standard error for all three estimators. The estimator accounting for over-dispersion using the Gamma-Frailty Poisson process model has the smallest standard error and is apparently more efficient than the two estimators studied in Lu et al. (2009).

Simulation results from *Scenario 2* and *Scenario 3* are similar to those from *Scenario 1*, even though the underlying frailty variable is not Gamma distributed. Based on these simulation results, we conclude that (i) the spline-based semiparametric sieve-MLE based on the Gamma-Frailty Poisson process model that accounts for the over-dispersion and autocorrelation will improve the estimation efficiency when over-dispersion or autocorrelation is present in the data; (ii) the Gamma-Frailty Poisson process model is robust against the underlying distribution of the frailty variable.

Table 1. Simulations results of the over-dispersed panel count data under different assumptions

	Bias			M-C sd			ASE			95% coverage		
	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
Gamma-Frailty Poisson Data:												
Sample size n=50												
Independent Poisson	0.0392	-0.0076	0.0100	0.3589	1.0867	0.6182	0.2181	0.7828	0.4445	0.719	0.838	0.845
Poisson Process	0.0381	-0.0063	0.0099	0.3596	1.0780	0.6162	0.2333	0.8836	0.5622	0.735	0.855	0.862
Gamma-Frailty Poisson	-0.0110	-0.0123	0.0017	0.2778	0.8347	0.4851	0.2086	0.8468	0.4399	0.897	0.902	0.904
Sample size n=100												
Independent Poisson	0.0338	-0.0142	0.0168	0.2617	0.8185	0.4352	0.1709	0.6170	0.3495	0.743	0.856	0.882
Poisson Process	0.0334	-0.0163	0.0145	0.2590	0.7999	0.4314	0.1743	0.6349	0.3653	0.754	0.866	0.891
Gamma-Frailty Poisson	0.0028	0.0016	0.0029	0.1527	0.5113	0.3072	0.1425	0.5755	0.3017	0.921	0.941	0.916
Lognormal-Frailty Poisson Data:												
Sample size n=50												
Independent Poisson	0.0294	0.0035	-0.0041	0.3064	0.9480	0.5198	0.1825	0.6574	0.3755	0.740	0.848	0.855
Poisson Process	0.0295	0.0037	-0.0065	0.3078	0.9438	0.5151	0.1883	0.7012	0.4083	0.753	0.869	0.861
Gamma-Frailty Poisson	0.0095	0.0476	-0.0214	0.2046	0.6953	0.3827	0.1612	0.6680	0.3492	0.879	0.903	0.912
Sample size n=100												
Independent Poisson	0.0322	-0.0152	-0.0017	0.2355	0.7306	0.3695	0.1471	0.5431	0.3057	0.764	0.872	0.893
Poisson Process	0.0322	-0.0206	-0.0031	0.2352	0.7275	0.3673	0.1484	0.5446	0.3102	0.762	0.876	0.901
Gamma-Frailty Poisson	0.0074	-0.0030	-0.0113	0.1412	0.4833	0.2783	0.1211	0.4547	0.2560	0.913	0.904	0.915
Mixture Poisson Data:												
Sample size n=50												
Independent Poisson	0.0036	-0.0026	0.0020	0.0877	0.2550	0.1425	0.0567	0.2003	0.1182	0.771	0.850	0.890
Poisson Process	0.0036	-0.0034	0.0015	0.0865	0.2514	0.1394	0.0565	0.2016	0.1179	0.775	0.854	0.891
Gamma-Frailty Poisson	0.0013	-0.0007	0.0025	0.0659	0.2021	0.1208	0.0573	0.2013	0.1164	0.894	0.914	0.913
Sample size n=100												
Independent Poisson	0.0032	-0.0024	0.0015	0.0630	0.1823	0.1026	0.0448	0.1551	0.0892	0.815	0.892	0.918
Poisson Process	0.0030	-0.0005	0.0019	0.0627	0.1795	0.0993	0.0448	0.1548	0.0880	0.819	0.885	0.921
Gamma-Frailty Poisson	0.0019	-0.0008	0.0013	0.0442	0.1375	0.0814	0.0402	0.1311	0.0788	0.922	0.925	0.945

Table 2. The spline-based sieve semiparametric inference for bladder tumor data

	Independent Poisson			Poisson Process			Gamma-Frailty Poisson ($\hat{\sigma}_n^2 = 1.32$)		
	Est.	Std.	p-value	Est.	Std.	p-value	Est.	Std.	p-value
Z1	0.1444	0.0553	0.0090	0.2075	0.0433	<.0001	0.3289	0.0976	0.0007
Z2	-0.0447	0.0462	0.3342	-0.0353	0.0945	0.7089	0.0054	0.1310	0.9681
Z3	0.1776	0.2706	0.5117	0.0637	0.2295	0.7812	0.0213	0.4267	0.9792
Z4	-0.6966	0.3021	0.0211	-0.7960	0.3179	0.0012	-1.0692	0.3765	0.0029

6.2. Application

The proposed method is applied to the bladder tumor data introduced in Section 1. A total of 116 patients were randomized into three treatment groups, with 31 using pyridoxin pills, 38 instilled with thiotepa and 47 in placebo group. Their follow-up times vary from one week to sixty-four weeks. Four variables, including the number (Z_1) and size (Z_2) of tumor at baseline, and two indicator variables, one for pyridoxin (Z_3), one for thiotepa (Z_4), are included in the proportional mean model, i.e.,

$$E(\mathbb{N}(t)|Z_1, Z_2, Z_3, Z_4) = \Lambda_0\,(t)\,exp\,(\beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + \beta_4 Z_4)$$

Analysis results based on the proposed method along with those based on the two methods studied in Lu et al. (2009) are shown in Table 6.2. The number of tumors observed at study entrance is positively related to the recurrence of bladder tumor. Per tumor increase at the baseline, the number of tumors at follow-ups increases by 15.5%, 23.1% and 39.1% on average using the Poisson process model-based sieve-MPLE and sieve-MLE, and the Gamma-Frailty Poisson process model-based sieve-MLE, respectively. Thiotepa instillation effectively decreases the number of recurrent tumors. The number of recurrent tumors for patients with thiotepa instillation is 49.5%, 45.1% and 32.5% of those in control group according to the three methods. The tumor size and the treatment of pyridoxin pills do not significantly affect the number of recurrent tumors at follow-up visits. The estimation method based on the Gamma-Frailty Poisson process model provides an estimate of the over-dispersion parameter 1.32 which evidently supports the over-dispersion of the panel count or the potential positive correlation between non-overlapping increments in the counting process. The effect of tumor number at study entrance and the treatment of thiotepa are quantitatively more significant when accounting for the correlation between cumulative counts using the frailty variable.

7. Concluding Remark

In this article we propose to analyze panel count data using the Gamma-Frailty Poisson process model. For over-dispersed panel count data that occur frequently in longitudinal follow-up studies of biomedical research, the proposed method yields a more efficient estimation procedure compared to the established likelihood methods based on Poisson process model studied in Wellner and Zhang (2007) and Lu et al. (2009). When over-dispersion is not an issue for panel count data, Zeger's method of moments estimate of the over-dispersion parameter is often zero or negative. Once that happens, the spline-based sieve-MLE of Lu et al. (2009) will be used in the second stage estimation. Such two-stage procedure yields very similar results to the Lu-Zhang-Huang's sieve-MLE for panel counts simulated from nonhomogeneous Poisson process (results not showing here). This implies

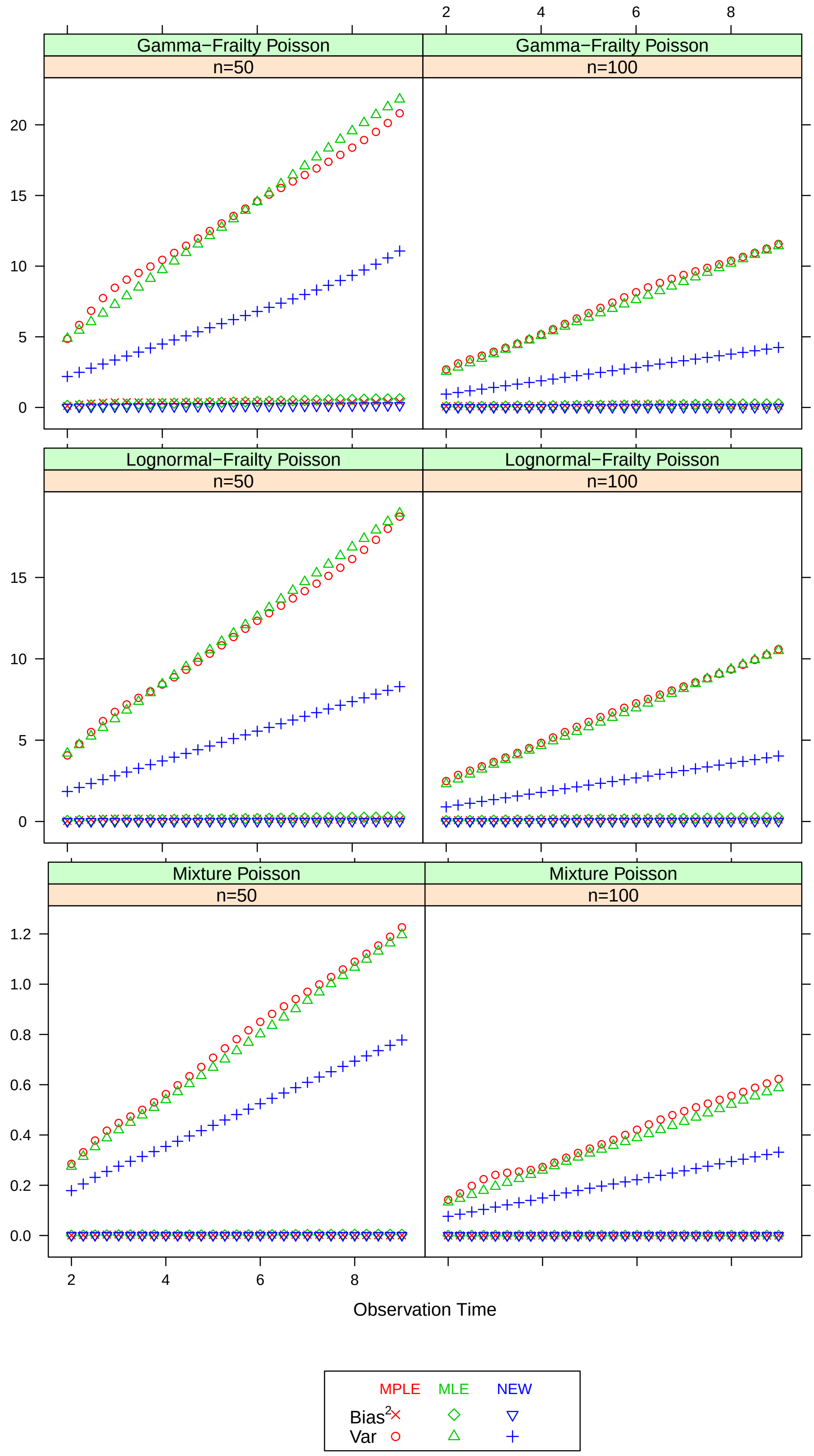

Fig. 1. Simulation results for estimations of the baseline mean function, $\Lambda_0(t) = 2t^{1/2}$

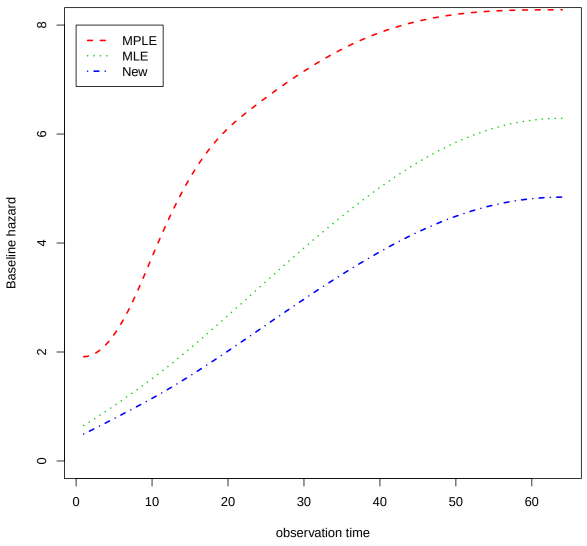

Fig. 2. Point estimates of the baseline mean function

that the proposed estimation method that accounts for over-dispersion performs as good as the semiparametric MLE for nonhomogeneous Poisson panel count data. Another strength of the proposed method rests on the fact that the Gamma-Frailty Poisson process model is only a working assumption for deriving the estimates. Both the theoretical and numerical results demonstrate that the advantage of the proposed method over the Poisson likelihood based estimates does not depend on true distribution of the frailty variable. In spite of the robustness properties of Poisson process model for panel count data demonstrated by Wellner and Zhang (2007) and the numerical efficiency of spline-based sieve-MLE under the Poisson model shown by Lu et al. (2009), we strongly recommend the use of the proposed method in the analysis of panel count data, as the over-dispersion is highly prevalent in applications of counting process data.

A. Technical Proofs

We use modern empirical process theory to study the asymptotic properties of the proposed estimate and the standard error estimate of the estimated regression parameters. Thereafter, C stands for a universal constant that may vary from place to place. Section A.1 provides the proof of the $\sqrt{n}-$consistency of the method of moment estimate of the over-dispersion parameter; Section A.2 outlines the proofs of Theorems 4.1, 4.2 and 4.3; Section A.3 sketches the proof of the consistency of spline-based sieve least-squares variance estimation method. Section A.4 gives two technical lemmas and proofs.

A.1. Proof of $\sqrt{n}$–consistency of $\hat{\sigma}_n^2$

Proof. Let μ_j denote the proportional mean at observation time t_j specified in (1). Wherever without confusion, we suppress the dependence of μ_j on (β, Λ) and let

$$\mu_{0j} = \Lambda_0\left(t_j\right) e^{\beta_0^T Z}; \quad \hat{\mu}_{nj}^{(0)} = \hat{\Lambda}_n^{(0)}\left(t_j\right) e^{\hat{\beta}_n^{(0)^T} Z}$$

where $\left(\hat{\beta}_n^{(0)}, \hat{\Lambda}_n^{(0)}\right)$ is the MPLE studied by Wellner and Zhang (2007). The method of moment estimate of the over-dispersion parameter in (5) can be rewritten as

$$\hat{\sigma}_n^2 = \frac{\mathbb{P}_n\left(\sum_{j=1}^K \left(\mathbb{N}_j - \hat{\mu}_{nj}^{(0)}\right)^2 - \hat{\mu}_{nj}^{(0)}\right)}{\mathbb{P}_n\left(\sum_{j=1}^K \hat{\mu}_{nj}^{(0)^2}\right)} \tag{10}$$

The numerator of (10) can be decomposed to

$$\begin{aligned}
&\mathbb{P}_n\left(\sum_{j=1}^K \left(\mathbb{N}_j - \hat{\mu}_{nj}^{(0)}\right)^2 - \hat{\mu}_{nj}^{(0)}\right)\\
=&\mathbb{P}_n\left(\left[\sum_{j=1}^K \left(\mathbb{N}_j - \mu_{0j}\right)^2 - \mu_{0j}\right]\right) + \mathbb{P}_n\left(\sum_{j=1}^K \left(\mu_{0j} - \hat{\mu}_{nj}^{(0)}\right)^2\right)\\
&+ 2\mathbb{P}_n\left(\sum_{j=1}^K \left(\mathbb{N}_j - \mu_{0j}\right)\left(\mu_{0j} - \hat{\mu}_{nj}^{(0)}\right)\right) + \mathbb{P}_n\left(\sum_{j=1}^K \left(\mu_{0j} - \hat{\mu}_{nj}^{(0)}\right)\right)\\
=&J_1 + J_2 + J_3 + J_4
\end{aligned}$$

where

$$J_1 = \left(\mathbb{P}_n - P\right)\left(\left[\sum_{j=1}^K \left(\mathbb{N}_j - \mu_{0j}\right)^2 - \mu_{0j}\right]\right) + \sigma_0^2 P\left(\sum_{j=1}^K \mu_{0j}^2\right) \tag{11}$$

$$J_2 = \left(\mathbb{P}_n - P\right)\left(\sum_{j=1}^K \left(\mu_{0j} - \hat{\mu}_{nj}^{(0)}\right)^2\right) + P\left(\sum_{j=1}^K \left(\mu_{0j} - \hat{\mu}_{nj}^{(0)}\right)^2\right) \tag{12}$$

$$J_3 = 2\left(\mathbb{P}_n - P\right)\left(\sum_{j=1}^K \left(\mathbb{N}_j - \mu_{0j}\right)\left(\mu_{0j} - \hat{\mu}_{nj}^{(0)}\right)\right) \tag{13}$$

$$J_4 = \left(\mathbb{P}_n - P\right)\left(\sum_{j=1}^K \left(\mu_{0j} - \hat{\mu}_{nj}^{(0)}\right)\right) + P\left(\sum_{j=1}^K \left(\mu_{0j} - \hat{\mu}_{nj}^{(0)}\right)\right) \tag{14}$$

By ordinary central limit theorem and Conditions 1, 2, 5 and 6, the first term in (11) is $O_p\left(n^{-1/2}\right)$.

Denote

$$\mathcal{J}_1(\delta) = \left\{ \sum_{j=1}^{K} (\mu_{0j} - \mu_j)^2 : \quad (\beta, \Lambda) \in \mathcal{R}^d \times \mathcal{F}, d((\beta, \Lambda), (\beta_0, \Lambda_0)) \leq \delta \right\},$$

$$\mathcal{J}_2(\delta) = \left\{ \sum_{j=1}^{K} (\mathbb{N}_j - \mu_{0j})(\mu_{0j} - \mu_j) : \quad (\beta, \Lambda) \in \mathcal{R}^d \times \mathcal{F}, d((\beta, \Lambda), (\beta_0, \Lambda_0)) \leq \delta \right\}, \text{ and}$$

$$\mathcal{J}_3(\delta) = \left\{ \sum_{j=1}^{K} (\mu_{0j} - \mu_j) : \quad (\beta, \Lambda) \in \mathcal{R}^d \times \mathcal{F}, d((\beta, \Lambda), (\beta_0, \Lambda_0)) \leq \delta \right\}.$$

Using the same technique for constructing the brackets as given in Wellner and Zhang (2007), it is easily shown that the bracketing numbers, $N_{[]}(\epsilon, \mathcal{J}_1(\delta), L_2(P))$, $N_{[]}(\epsilon, \mathcal{J}_2(\delta), L_2(P))$ and $N_{[]}(\epsilon, \mathcal{J}_3(\delta), L_2(P))$ are all bounded above by $C \exp(1/\epsilon)(1/\epsilon)^d$. It follows that $\mathcal{J}_1$, $\mathcal{J}_2$ and $\mathcal{J}_3$ are all P-Donsker. Due to the consistency of MPLE $(\hat{\beta}_n^{(0)}, \hat{\Lambda}_n^{(0)})$ given by Wellner and Zhang (2007), using Conditions 1-3, 5 and 6, the result of Lemma 7.1 in Wellner and Zhang (2007) and Dominated Convergence Theorem (DCT), it can be also easily shown that

$$P\left(\sum_{j=1}^{K} \left(\mu_{0j} - \hat{\mu}_{nj}^{(0)}\right)^2\right)^2 \rightarrow_p 0,$$

$$P\left(\sum_{j=1}^{K} (\mathbb{N}_j - \mu_{0j})\left(\mu_{0j} - \hat{\mu}_{nj}^{(0)}\right)\right)^2 \rightarrow_p 0,$$

$$\text{and } P\left(\sum_{j=1}^{K} \left(\mu_{0j} - \hat{\mu}_{nj}^{(0)}\right)\right)^2 \rightarrow_p 0$$

as $\delta \rightarrow 0$. Therefore, by the relationship between P-Donsker and asymptotic equicontinuity (Corollary 2.3.12 van der Vaart and Wellner, 1996), it follows that the first terms of (12), (13), and (14) are all $o_p\left(n^{-1/2}\right)$.

Using Conditions 1-3, 5 and 6, the result of Lemma 7.1 in Wellner and Zhang (2007) and the $n^{1/3}$ convergence rate of $(\hat{\beta}_n^{(0)}, \hat{\Lambda}_n^{(0)})$ as shown in Wellner and Zhang (2007), it follows that,

$$P\left(\sum_{j=1}^{K} \left(\mu_{0j} - \hat{\mu}_{nj}^{(0)}\right)^2\right) \leq Cd^2\left((\hat{\beta}_n^{(0)}, \hat{\Lambda}_n^{(0)}), (\beta_0, \Lambda_0)\right) = O_p\left(n^{-2/3}\right).$$

By the similar arguments as used in the proof of Theorem 2 in Zhang (2006) along with the convergence rate of $\left(\hat{\beta}_n^{(0)}, \hat{\Lambda}_n^{(0)}\right)$, it is parallel to show that $\sqrt{n}P\left(\sum_{j=1}^{K}\left(\mu_{0j} - \hat{\mu}_{nj}^{(0)}\right)\right)$ is asymptotically normal with zero mean and hence $P\left(\sum_{j=1}^{K}\left(\mu_{0j} - \hat{\mu}_{nj}^{(0)}\right)\right) = O_p\left(n^{-1/2}\right)$. Thus,

$$\mathbb{P}_n\left(\sum_{j=1}^{K}\left(\mathbb{N}_j - \hat{\mu}_{nj}^{(0)}\right)^2 - \hat{\mu}_{nj}^{(0)}\right) = \sigma_0^2 P\left(\sum_{j=1}^{K} \mu_{0j}^2\right) + O_p\left(n^{-1/2}\right)$$

The denominator of (10) can be decomposed to

$$\mathbb{P}_n\left(\sum_{j=1}^{K}\hat{\mu}_{nj}^{(0)^2}\right)=(\mathbb{P}_n-P)\left(\sum_{j=1}^{K}\hat{\mu}_{nj}^{(0)^2}\right)+P\left(\sum_{j=1}^{K}\hat{\mu}_{nj}^{(0)^2}\right) \tag{15}$$

Let $\mathcal{J}_4=\left\{\sum_{j=1}^{K}\mu_j^2:(\beta,\Lambda)\in\mathcal{R}^d\times\mathcal{F}\right\}$. it can be similarly argued that $N_{[]}(\epsilon,\mathcal{J}_4,L_1(P))$ is bounded above by $C\exp(1/\epsilon)(1/\epsilon)^d$. By Theorem 2.4.1 of van der Vaart and Wellner (1996) (Glivenko-Cantelli Theorem), $\mathcal{J}_4$ is a Glivenko-Cantelli class and hence $(\mathbb{P}_n-P)\left(\sum_{j=1}^{K}\hat{\mu}_{nj}^{(0)^2}\right)=o_p(1)$. The consistency of $\left(\hat{\beta}_n^{(0)},\hat{\Lambda}_n^{(0)}\right)$ along with Conditions 1-3 and 5, and DCT result in $P\left(\sum_{j=1}^{K}\left(\hat{\mu}_{nj}^{(0)^2}\right)\right)=P\left(\sum_{j=1}^{K}\mu_{0j}^2\right)+o_p(1)$.

Therefore,

$$\frac{\mathbb{P}_n\left(\sum_{j=1}^{K}\left(\mathbb{N}_j-\hat{\mu}_{nj}^{(0)}\right)^2-\hat{\mu}_{nj}^{(0)}\right)}{\mathbb{P}_n\sum_{j=1}^{K}\hat{\mu}_{nj}^{(0)^2}}=\sigma_0^2+O_p\left(n^{-1/2}\right)$$

This proof is complete.

A.2. Proof of the Asymptotic properties

A.2.1. Proof of Theorem 4.1 (Consistency):

To study the asymptotic properties of the proposed estimate, some notations about B-splines are needed. Let

$$\phi_{l,\Xi}=\left\{\sum_{i=1}^{q_n}a_iB_i:\quad B_i,i=1,2,\cdots,q_n \text{ are the B-spline basis functions at }\ \Xi\right\}$$

and

$$\psi_{l,\Xi}=\left\{\sum_{i=1}^{q_n}a_iB_i:\ \sum_{i=1}^{q_n}a_iB_i\in\phi_{l,\Xi}\ \text{ and }\ a_1\le a_2\le\cdots\le a_{q_n}\right\}$$

Proof. To prove the consistency of the proposed two-stage estimation method, we apply Theorem 5.7 in van der Vaart (1998) and check the three sufficient conditions for the global consistency. Let

$$\mathbb{M}_n(\beta,\Lambda)=\mathbb{P}_nm\left(\beta,\Lambda,\hat{\sigma}_n^2\right);\quad \mathbb{M}(\beta,\Lambda)=Pm\left(\beta,\Lambda,\sigma_0^2\right)$$

First,

$$\begin{aligned}\mathbb{M}_n(\beta,\Lambda)-\mathbb{M}(\beta,\Lambda)&=\mathbb{P}_nm\left(\beta,\Lambda,\hat{\sigma}_n^2\right)-Pm\left(\beta,\Lambda,\sigma_0^2\right)\\&=\mathbb{P}_n\left(m\left(\beta,\Lambda,\hat{\sigma}_n^2\right)-m\left(\beta,\Lambda,\sigma_0^2\right)\right)+(\mathbb{P}_n-P)\,m\left(\beta,\Lambda,\sigma_0^2\right)\end{aligned}$$

By Taylor expansion

$$\mathbb{P}_n\left(m\left(\beta,\Lambda,\hat{\sigma}_n^2\right)-m\left(\beta,\Lambda,\sigma_0^2\right)\right)=\mathbb{P}_n\dot{m}_{\sigma^2}\left(\beta,\Lambda,\tilde{\sigma}^2\right)\left(\hat{\sigma}_n^2-\sigma_0^2\right)$$

where $\dot{m}_{\sigma^2}\left(\beta,\Lambda,\sigma^2\right)=\frac{\partial}{\partial\sigma^2}m\left(\beta,\Lambda,\sigma^2\right)$ and $\left|\tilde{\sigma}^2-\sigma_0^2\right|\le\left|\hat{\sigma}_n^2-\sigma_0^2\right|$. By Conditions 1, 2, 3, 5 and 6, it can be easily argued that $\mathbb{P}_n\dot{m}_{\sigma^2}\left(\beta,\Lambda,\tilde{\sigma}^2\right)=O_p(1)$. Then the consistency of $\hat{\sigma}_n^2$ implies that $\mathbb{P}_n\left(m\left(\beta,\Lambda,\hat{\sigma}_n^2\right)-m\left(\beta,\Lambda,\sigma_0^2\right)\right)=o_p(1)$.

Define $\mathcal{L}_1 = \{m\left(\beta, \Lambda, \sigma_0^2\right), \beta \in \mathcal{R}^d, \log\Lambda \in \psi_{l,\Xi}\}$ and $\mathcal{L}_1^* = \{m\left(\beta, \Lambda, \sigma_0^2\right), \beta \in \mathcal{R}^d, \Lambda \in \mathcal{F}\}$. Using Theorem 2.7.5 of van der Vaart and Wellner (1996) and the same technique for constructing the brackets given by Wellner and Zhang (2007), we can show that the bracketing number, $N_{[]}(\epsilon, \mathcal{L}_1^*, L_1(P))$ is bounded by $C\exp(1/\epsilon)\left(1/\epsilon\right)^d$. Since $\exp(\psi_{l,\Xi}) \subset \mathcal{F}$ and $\mathcal{L}_1 \subset \mathcal{L}_1^*$, $N_{[]}(\epsilon, \mathcal{L}_1, L_1(P))$ is bounded by $C\exp(1/\epsilon)\left(1/\epsilon\right)^d$ as well. By Glivenko-Cantelli Theorem, $\mathcal{L}_1$ is Glivenko-Cantelli and hence $\left(\mathbb{P}_n - P\right) m\left(\beta, \Lambda, \sigma_0^2\right) = o_p\left(1\right)$. This justifies the uniform convergence condition.

Second, the separation condition has been established by Lemma A.1, Part (ii).

Last, based on the spline approximation result given by de Boor (2001, p.148), there exist a $\Lambda_{0,n}$, such that $\log\Lambda_{0,n} \in \psi_{l,\Xi}$ of order $l \geq p+2$ and $\|\Lambda_{0,n} - \Lambda_0\|_\infty \leq Cq_n^{-p} = O(n^{-p\nu})$.

$$\begin{aligned}
&\mathbb{M}_n\left(\hat{\beta}_n, \hat{\Lambda}_n\right) - \mathbb{M}_n\left(\beta_0, \Lambda_0\right)\\
&\quad= \mathbb{P}_n m\left(\hat{\beta}_n, \hat{\Lambda}_n, \hat{\sigma}_n^2\right) - \mathbb{P}_n m\left(\beta_0, \Lambda_0, \hat{\sigma}_n^2\right)\\
&\quad= \mathbb{P}_n\left(m\left(\hat{\beta}_n, \hat{\Lambda}_n, \hat{\sigma}_n^2\right) - m\left(\beta_0, \Lambda_{0,n}, \hat{\sigma}_n^2\right)\right) + \mathbb{P}_n\left(m\left(\beta_0, \Lambda_{0,n}, \hat{\sigma}_n^2\right) - m\left(\beta_0, \Lambda_0, \hat{\sigma}_n^2\right)\right)\\
&\quad\geq \mathbb{P}_n\left(m\left(\beta_0, \Lambda_{0,n}, \hat{\sigma}_n^2\right) - m\left(\beta_0, \Lambda_0, \hat{\sigma}_n^2\right)\right)\\
&\quad= \mathbb{P}_n\left(m\left(\beta_0, \Lambda_{0,n}, \sigma_0^2\right) - m\left(\beta_0, \Lambda_0, \sigma_0^2\right)\right) + \mathbb{P}_n\left(m\left(\beta_0, \Lambda_{0,n}, \hat{\sigma}_n^2\right) - m\left(\beta_0, \Lambda_{0,n}, \sigma_0^2\right)\right)\\
&\qquad- \mathbb{P}_n\left(m\left(\beta_0, \Lambda_0, \hat{\sigma}_n^2\right) - m\left(\beta_0, \Lambda_0, \sigma_0^2\right)\right)
\end{aligned}$$

Using the same arguments as above, the last two terms are both $o_p\left(1\right)$. Note that

$$\begin{aligned}
&\mathbb{P}_n\left(m\left(\beta_0, \Lambda_{0,n}, \sigma_0^2\right) - m\left(\beta_0, \Lambda_0, \sigma_0^2\right)\right)\\
&\quad= \left(\mathbb{P}_n - P\right)\left(m\left(\beta_0, \Lambda_{0,n}, \sigma_0^2\right) - m\left(\beta_0, \Lambda_0, \sigma_0^2\right)\right) + P\left(m\left(\beta_0, \Lambda_{0,n}, \sigma_0^2\right) - m\left(\beta_0, \Lambda_0, \sigma_0^2\right)\right)\\
&\quad= P\left(m\left(\beta_0, \Lambda_{0,n}, \sigma_0^2\right) - m\left(\beta_0, \Lambda_0, \sigma_0^2\right)\right) + o_p(1)
\end{aligned}$$

by Glivenko-Cantelli theorem. By Conditions 1, 2, 3, 5 and 6, Taylor expansion of $m(\beta_0, \Lambda, \sigma^2)$ at Λ_0 yields that

$$P\left(m\left(\beta_0, \Lambda_{0,n}, \sigma_0^2\right) - m\left(\beta_0, \Lambda_0, \sigma_0^2\right)\right) \geq -C\|\Lambda_{0,n} - \Lambda_0\|_{L_2(\mu)}.$$

Therefore, $\mathbb{M}_n\left(\hat{\beta}_n, \hat{\Lambda}_n\right) - \mathbb{M}_n\left(\beta_0, \Lambda_0\right) \geq -o_p\left(1\right)$. The proof is complete.

A.2.2. Proof of Theorem 4.2 (Convergence Rate):

Proof. The convergence rate is derived by verifying the conditions in Theorem 3.4.1 of van der Vaart and Wellner (1996). To apply the theorem to this problem, we denote $\theta = (\beta, \Lambda) \in \Theta_n$ with $\Theta_n = \{(\beta, \Lambda) : \quad \beta \in \mathcal{R}^d, \quad \log\Lambda \in \psi_{l,\Xi}\}$. We also denote $\theta_n = (\beta_0, \Lambda_{0,n})$ with the $\Lambda_{0,n}$ chosen as in the proof of Theorem 4.1 and $\hat{\theta}_n = (\hat{\beta}_n, \hat{\Lambda}_n) \in \Theta_n$, the proposed estimate of θ_0 in Θ_n. Hence $\mathbb{M}_n(\hat{\theta}_n) \geq \mathbb{M}_n(\theta_n)$. For the sieve estimation problem studied in this article, let

$$M_n(\theta) = \mathbb{M}(\beta, \Lambda) = Pm(\beta, \Lambda, \sigma_0^2) = Pm(\theta, \sigma_0^2) \text{ and } d_n(\theta, \theta_n) = d(\theta, \theta_n)$$

First, by the separation property given by Lemma A.1, Part (ii), it follows that $M_n(\theta) - M_n(\theta_0) \leq -Cd_n^2(\theta, \theta_0)$. Since $\|\theta_n - \theta_0\|_\infty = \|\Lambda_{0,n} - \Lambda_0\|_\infty = O(n^{-p\nu})$, for any θ such that

$\delta/2 < d_n(\theta,\theta_n) < \delta$, $d_n(\theta,\theta_0) \geq d_n(\theta,\theta_n) - d_n(\theta_n,\theta_0) \geq C\delta$ for large enough n. Therefore,

$$\begin{aligned} M_n(\theta) - M_n(\theta_n) &= M_n(\theta) - M_n(\theta_0) + M_n(\theta_0) - M_n(\theta_n) \\ &\leq -C\delta^2 - O(n^{-p\nu}) = -C\delta^2 \text{ for large enough } n \end{aligned}$$

Second, note that for any θ such that $\delta/2 < d_n(\theta,\theta_n) < \delta$,

$$\begin{aligned} &(\mathbb{M}_n - M_n)(\theta) - (\mathbb{M}_n - M_n)(\theta_n) \\ &\quad = \left[\mathbb{P}_n m(\theta,\hat{\sigma}_n^2) - Pm(\theta,\sigma_0^2)\right] - \left[\mathbb{P}_n m(\theta_n,\hat{\sigma}_n^2) - Pm(\theta_n,\sigma_0^2)\right] \\ &\quad = \mathbb{P}_n \left\{\left[m(\theta,\hat{\sigma}_n^2) - m(\theta,\sigma_0^2)\right] - \left[m(\theta_n,\hat{\sigma}_n^2) - m(\theta_n,\sigma_0^2)\right]\right\} \\ &\qquad + (\mathbb{P}_n - P)\left(m(\theta,\sigma_0^2) - m(\theta_n,\sigma_0^2)\right) \\ &\quad = (\mathbb{P}_n - P)\left(m(\theta,\sigma_0^2) - m(\theta_n,\sigma_0^2)\right) + \delta O_p(n^{-1/2}) \end{aligned}$$

by Conditions 1-3, 5 and 6, and the $\sqrt{n}$-consistency of $\hat{\sigma}_n^2$.

Let $\mathcal{L}_3(\delta) = \{m(\theta,\sigma_0^2) - m(\theta_n,\sigma_0^2) : \theta \in \Theta_n, \delta/2 < d_n(\theta,\theta_n) < \delta)\}$. Using the result of Lemma A.2 and the technique given by Wellner and Zhang (2007, p.2129), it can be easily shown that $\log N_{[]}(\epsilon,\mathcal{L}_3(\delta),\| \, \|_{P,B})$ is also bounded above by $Cq_n \log(\delta/\epsilon)$ with the 'Bernstein norm' defined in van der Vaart and Wellner (1996, p.324). Then it follows that

$$\tilde{J}_{[]}\left(\delta,\mathcal{L}_3(\delta),\|\cdot\|_{P,B}\right) = \int_0^\delta \sqrt{1 + logN_{[]}\left(\epsilon,\mathcal{L}_3(\delta),\|\cdot\|_{P,B}\right)}d\epsilon \leq Cq_n^{1/2}\delta$$

which results in

$$E_P\|\sqrt{n}(\mathbb{P}_n - P)\|_{\mathcal{L}_3(\delta)} \leq C\left(q_n^{1/2}\delta + q_n/n^{1/2}\right)$$

by Lemma 3.4.3 of van der Vaart and Wellner (1996). Hence

$$\begin{aligned} E_P \sup_{\substack{\delta/2 < d_n(\theta,\theta_n) < \delta \\ \theta\in\Theta_n}} &\sqrt{n}\left|\mathbb{P}_n m\left(\theta,\hat{\sigma}_n^2\right) - Pm\left(\theta,\sigma_0^2\right) - \right. \\ &\left. \left[\mathbb{P}_n m\left(\theta_0,\hat{\sigma}_n^2\right) - Pm\left(\theta_0,\sigma_0^2\right)\right]\right| \leq C\phi_n(\delta) \end{aligned}$$

with $\phi_n(\delta) = C\left(q_n^{1/2}\delta + q_n/n^{1/2}\right)$. An easy algebra shows that $r_n^2\phi_n(r_n^{-1}) \leq Cn^{1/2}$ with $r_n = n^{\min(p\nu,(1-\nu)/2)}$. Then it follows that

$$r_n d_n(\hat{\theta}_n,\theta_n) = O_p(1)$$

by the conclusion of Theorem 3.4.1 of van der Vaart and Wellner (1996). Moreover, since $\|\theta_n - \theta_0\|_\infty = O(n^{-p\nu})$, it follows that

$$\begin{aligned} r_n d(\hat{\theta}_n,\theta_0) &= r_n d_n(\hat{\theta}_n,\theta_0) \leq r_n d_n(\hat{\theta}_n,\theta_n) + r_n d_n(\theta_n,\theta_0) \\ &= O_p(1) + r_n O(n^{-p\nu}) = O_p(1) \end{aligned}$$

The proof is complete.

A.2.3. Proof of Theorem 4.3 (Asymptotic Normality):
Incorporating the extra over-dispersion parameter, we adopt the following notations

$$\begin{aligned} S_1(\beta,\Lambda) &= Pm_1\left(\beta,\Lambda,\sigma_0^2;X\right);\ \ S_2(\beta,\Lambda) = Pm_2\left(\beta,\Lambda,\sigma_0^2;X\right)\\ S_{1n}(\beta,\Lambda) &= \mathbb{P}_n m_1\left(\beta,\Lambda,\hat{\sigma}_n^2;X\right);\ \ S_{2n}(\beta,\Lambda) = \mathbb{P}_n m_2\left(\beta,\Lambda,\hat{\sigma}_n^2;X\right)\\ \dot{S}_{11}(\beta,\Lambda) &= Pm_{11}\left(\beta,\Lambda,\sigma_0^2;X\right);\ \ \dot{S}_{22}(\beta,\Lambda) = Pm_{22}\left(\beta,\Lambda,\sigma_0^2;X\right)\\ \dot{S}_{12}(\beta,\Lambda)[h] &= \dot{S}_{12}^T(\beta,\Lambda)[h] = Pm_{12}\left(\beta,\Lambda,\sigma_0^2;X\right). \end{aligned}$$

Proof. There are only slightly different expressions in S_{1n} and S_{2n} from those used in Theorem 6.1 of Wellner and Zhang (2007) in which a fixed quantity σ_0^2 is replaced by its estimate $\hat{\sigma}_n^2$. So Theorem 6.1 of Wellner and Zhang (2007) cannot be immediately applied. We will derive the asymptotic normality of $\hat{\beta}_n$ by modifying the proof of Wellner-Zhang's theorem under the same conditions. First, we show that Conditions A1-A6 of Wellner-Zhang's theorem hold in this problem under Conditions 1-8.

A1. The condition is satisfied with the consistency and convergence rate of $\left(\hat{\beta}_n,\hat{\Lambda}_n\right)$.

A2. $Pm_1\left(\beta_0,\Lambda_0,\sigma_0^2\right)=0$ and $Pm_2\left(\beta_0,\Lambda_0,\sigma_0^2\right)[h]=0$ as long as the proportional mean model in (1) hold.

A3. By the same technique used in information calculation as given by Wellner and Zhang (2007, p.2130), it can be easily calculated that the least favorable direction h^* is expressed by (7).

A4. Since $\left(\hat{\beta}_n,\hat{\Lambda}_n\right)$ are estimated by maximizing the likelihood $m\left(\beta,\Lambda,\hat{\sigma}_n^2\right)$, they satisfy the score equation, i.e.,

$$\mathbb{P}_n m_1\left(\hat{\beta}_n,\hat{\Lambda}_n,\hat{\sigma}_n^2;X\right)=0 \text{ and } \mathbb{P}_n m_2\left(\hat{\beta}_n,\hat{\Lambda}_n,\hat{\sigma}_n^2;X\right)[h]=0.$$

where h is any function in $\mathcal{H}$. The first part is automatically true. To prove the second part, by specifically choosing $h=\hat{\Lambda}_n S\in\mathcal{H}$, it suffices to show that

$$I=\mathbb{P}_n\left\{m_2\left(\hat{\beta}_n,\hat{\Lambda}_n,\hat{\sigma}_n^2;X\right)[\Lambda_0 S]-m_2\left(\hat{\beta}_n,\hat{\Lambda}_n,\hat{\sigma}_n^2;X\right)[\hat{\Lambda}_n S]\right\}=o_p\left(n^{-1/2}\right)$$

where

$$S=\frac{E\left(\frac{Z\times 1/\sigma_0^2}{\Lambda_{0,K}e^{\beta_0^T Z}+1/\sigma_0^2}\times e^{\beta_0^T Z}|K,\underline{T}_K\right)}{E\left(e^{\beta_0^T Z}|K,\underline{T}_K\right)-E\left(\frac{\Lambda_{0,K}e^{2\beta_0^T Z}}{\Lambda_{0,K}e^{\beta_0^T Z}+1/\sigma_0^2}|K,\underline{T}_K\right)}.$$

Now rewrite $\mathbb{P}_n\left\{m_2\left(\hat{\beta}_n,\hat{\Lambda}_n,\hat{\sigma}_n^2;X\right)[\Lambda_0 S]-m_2\left(\hat{\beta}_n,\hat{\Lambda}_n,\hat{\sigma}_n^2;X\right)[\hat{\Lambda}_n S]\right\}$ by T_1+T_2 where

$$\begin{aligned} T_1 =&\mathbb{P}_n\left\{m_2\left(\hat{\beta}_n,\hat{\Lambda}_n,\hat{\sigma}_n^2;X\right)[(\Lambda_0-\hat{\Lambda}_n)S]-m_2\left(\hat{\beta}_n,\hat{\Lambda}_n,\sigma_0^2;X\right)[(\Lambda_0-\hat{\Lambda}_n)S]\right\}\\ T_2 =&\mathbb{P}_n\left\{m_2\left(\hat{\beta}_n,\hat{\Lambda}_n,\sigma_0^2;X\right)[(\Lambda_0-\hat{\Lambda}_n)S]\right\}. \end{aligned}$$

By Taylor expansion

$$\begin{aligned} T_1 =&\mathbb{P}_n \left\{\dot{m}_{2\sigma^2}\left(\hat{\beta}_n, \hat{\Lambda}_n, \breve{\sigma}^2\right)\left[\left(\Lambda_0 - \hat{\Lambda}_n\right) S\right]\right\}\left(\hat{\sigma}_n^2 - \sigma_0^2\right) \\ =&\Big[(\mathbb{P}_n - P)\left\{\dot{m}_{2\sigma^2}\left(\hat{\beta}_n, \hat{\Lambda}_n, \breve{\sigma}^2\right)\left[\left(\Lambda_0 - \hat{\Lambda}_n\right) S\right]\right\} \\ &+ P\left\{\dot{m}_{2\sigma^2}\left(\hat{\beta}_n, \hat{\Lambda}_n, \breve{\sigma}^2\right)\left[\left(\Lambda_0 - \hat{\Lambda}_n\right) S\right]\right\}\Big]\left(\hat{\sigma}_n^2 - \sigma_0^2\right) \end{aligned}$$

where $\dot{m}_{2\sigma^2}\left(\beta, \Lambda, \sigma^2\right)[h] = \frac{\partial}{\partial \sigma^2} m_2\left(\beta, \Lambda, \sigma^2\right)[h]$ and $\left|\breve{\sigma}^2 - \sigma_0^2\right| \leq \left|\hat{\sigma}_n^2 - \sigma_0^2\right|$. By the same technique for constructing $L_1(P)$ brackets used in Wellner and Zhang (2007) for

$$\mathcal{K}_2 = \left\{\dot{m}_{2\sigma^2}\left(\beta, \Lambda, \sigma^2\right)\left[\left(\Lambda_0 - \Lambda\right) S\right] : \beta \in \mathcal{R}^d, \Lambda \in \mathcal{F}, \sigma^2 \in \mathcal{R}^+\right\},$$

it is easy argued that $N_{[]}\left(\epsilon, \mathcal{K}_2, L_1(P)\right) \leq C \exp(1/\epsilon)(1/\epsilon)^{d+1}$ using Conditions 1-3, 5 and 6. Hence $\mathcal{K}_2$ is Glivenko-Cantelli and therefore,

$$(\mathbb{P}_n - P)\left\{\dot{m}_{2\sigma^2}\left(\hat{\beta}_n, \hat{\Lambda}_n, \breve{\sigma}^2\right)\left[(\Lambda_0 - \hat{\Lambda}_n) S\right]\right\} = o_p(1).$$

By consistency of $\left(\hat{\beta}_n, \hat{\Lambda}_n\right)$, dominated convergence theorem and Conditions 1-3, 5 and 6, it can be also easily shown that

$$P\left\{\dot{m}_{2\sigma^2}\left(\hat{\beta}_n, \hat{\Lambda}_n, \breve{\sigma}^2\right)\left[(\Lambda_0 - \hat{\Lambda}_n) S\right]\right\} = o_p(1).$$

Finally the $\sqrt{n}$−consitency of $\hat{\sigma}_n^2$ leads to $T_1 = o_p\left(n^{-1/2}\right)$.

Since σ^2 is fixed at σ_0^2 in T_2, the proof of $T_2 = o_p\left(n^{-1/2}\right)$ follows the same lines as those given in (Wellner and Zhang, 2007, p.2131-2133), given Conditions 1-7. Hence A4 is justified.

A5. With the notations defined at the beginning of this section, we have

$$\begin{aligned} \left(S_{1n} - S_1\right)(\beta, \Lambda) - \left(S_{1n} - S_1\right)(\beta_0, \Lambda_0) =& R_1 + R_2 \\ \left(S_{2n} - S_2\right)(\beta, \Lambda)[h^*] - \left(S_{2n} - S_2\right)(\beta_0, \Lambda_0)[h^*] =& Q_1 + Q_2 \end{aligned}$$

where

$$\begin{aligned} R_1 =& \mathbb{P}_n\left[\left(m_1\left(\beta, \Lambda, \hat{\sigma}_n^2\right) - m_1\left(\beta, \Lambda, \sigma_0^2\right)\right) - \left(m_1\left(\beta_0, \Lambda_0, \hat{\sigma}_n^2\right) - m_1\left(\beta_0, \Lambda_0, \sigma_0^2\right)\right)\right] \\ R_2 =& (\mathbb{P}_n - P)\left(m_1\left(\beta, \Lambda, \sigma_0^2\right) - m_1\left(\beta_0, \Lambda_0, \sigma_0^2\right)\right) \\ Q_1 =& \mathbb{P}_n\left[\left(m_2\left(\beta, \Lambda, \hat{\sigma}_n^2\right)[h^*] - m_2\left(\beta, \Lambda, \sigma_0^2\right)\right)[h^*] \right. \\ & \left. - \left(m_2\left(\beta_0, \Lambda_0, \hat{\sigma}_n^2\right)[h^*] - m_2\left(\beta_0, \Lambda_0, \sigma_0^2\right)[h^*]\right)\right] \\ Q_2 =& (\mathbb{P}_n - P)\, m_2\left(\beta, \Lambda, \sigma_0^2\right)[h^*] - (\mathbb{P}_n - P)\, m_2\left(\beta_0, \Lambda_0, \sigma_0^2\right)[h^*] \end{aligned}$$

Using the same arguments as given in the proof of $T_1 = o_p\left(n^{-1/2}\right)$, it can be similarly shown that both R_1 and Q_1 are $o_p\left(n^{-1/2}\right)$ for any (β, Λ) such that $d\left((\beta, \Lambda), (\beta_0, \Lambda_0)\right) = O_p\left(r_n^{-1}\right)$. As σ^2 is fixed at σ_0^2, showing both R_2 and Q_2 being $o_p\left(n^{-1/2}\right)$ follows the same lines as those given by Wellner and Zhang (2007, p.2133-2134) using Conditions 1-7. Hence A5 is justified.

A6. Since both $S_1(\beta,\Lambda)$ and $S_2(\beta,\Lambda)[h^*]$ do not involve $\hat{\sigma}_n^2$, the justification of A6 follows exactly the same lines as those given in Wellner and Zhang (2007, p.2134-2135).

Following the same lines as those given in the proof of Theorem 6.1 of Wellner and Zhang (2007, p.2139-2140), it yields that

$$\begin{aligned}\sqrt{n}\left(\hat{\beta}-\beta_0\right) &= (A_0+o(1))^{-1}\sqrt{n}\mathbb{P}_n\{m_1(\beta_0,\Lambda_0,\hat{\sigma}_n^2)-m_2(\beta_0,\Lambda_0,\hat{\sigma}_n^2)[h^*]\}+o_p(1)\\ &= (A_0+o(1))^{-1}\sqrt{n}\mathbb{P}_n\{m_1(\beta_0,\Lambda_0,\sigma_0^2)-m_2(\beta_0,\Lambda_0,\sigma_0^2)[h^*]\}\\ &\quad+(A_0+o(1))^{-1}\mathbb{P}_n\{m_{1\sigma^2}(\beta_0,\Lambda_0,\tilde{\sigma}^2)-m_{2\sigma^2}(\beta_0,\Lambda_0,\tilde{\sigma}^2)[h^*]\}\sqrt{n}(\hat{\sigma}_n^2-\sigma_0^2)\\ &\quad+o_p(1)\;\text{ for some }\;\tilde{\sigma}^2\;\text{ such that }\;|\tilde{\sigma}^2-\sigma_0^2|\le|\hat{\sigma}_n^2-\sigma_0^2|\end{aligned}$$

For the log likelihood of the Gamma-Frailty Poisson process, it is easily seen that for any $\sigma^2>0$,

$$Pm_{1\sigma^2}(\beta_0,\Lambda_0,\sigma^2)=Pm_{2\sigma^2}(\beta_0,\Lambda_0,\sigma^2)[h^*]\equiv 0.$$

By the Glivenko-Cantelli Theorem,

$$\begin{aligned}&\mathbb{P}_n\{m_{1\sigma^2}(\beta_0,\Lambda_0,\tilde{\sigma}^2)-m_{2\sigma^2}(\beta_0,\Lambda_0,\tilde{\sigma}^2)[h^*]\}\\ &\quad=(\mathbb{P}_n-P)\{m_{1\sigma^2}(\beta_0,\Lambda_0,\tilde{\sigma}^2)-m_{2\sigma^2}(\beta_0,\Lambda_0,\tilde{\sigma}^2)[h^*]\}=o_p(1)\end{aligned}$$

Hence due to the $\sqrt{n}$-consistency of $\hat{\sigma}_n^2$, it follows that

$$\begin{aligned}\sqrt{n}\left(\hat{\beta}-\beta_0\right) &= (A_0+o(1))^{-1}\sqrt{n}\mathbb{P}_n\{m_1(\beta_0,\Lambda_0,\sigma_0^2)-m_2(\beta_0,\Lambda_0,\sigma_0^2)[h^*]\}+o_p(1)\\ &\to_d N\left(0,A_0^{-1}B_0A_0^{-1}\right)\end{aligned}$$

The proof is complete.

A.3. Proofs for the consistency of $\hat{h}_{n,s}, s=1,2,\cdots d$ *and* $\hat{A}_n,\hat{B}_n$

Denote $\theta=(\beta,\Lambda)$ and $\rho_s\left(\theta,h;\sigma^2\right)=\left(m_{1,s}\left(\theta;\sigma^2\right)-m_2\left(\theta;\sigma^2\right)[h]\right)^2, s=1,2,\cdots,d$. Note that $\hat{h}_{n,s}=\text{argmin}_{h\in\phi_{l,\Xi}}\mathbb{P}_n\rho_s\left(\hat{\theta}_n,h;\hat{\sigma}_n^2\right)$, we first show that

$$\|\hat{h}_n-h^*\|_{\mathcal{H}}=\max_{1\le s\le d}\|\hat{h}_{n,s}-h_s^*\|_{L_2(\mu)}\to_p 0.$$

By evaluating the upper bound of the bracketing entropy number of $\mathfrak{S}=\{\rho_s\left(\theta,h;\sigma^2\right):\theta\in\mathcal{R}^d\times\exp(\psi_{l,\Xi}),h\in\phi_{l,\Xi},\sigma^2\in\mathcal{R}^+\}$ with $\exp(\psi_{l,\Xi})=\{f:\log f\in\psi_{l,\Xi}\}$ and $\mathcal{R}^d\subset\mathbb{R}^d$ and $\mathcal{R}^+\subset\mathbb{R}^+$ being compact, it can be easily argued that $\mathfrak{S}$ is Glivenko-Cantelli class. Moreover, Condition 8 implies that there exists a $h_{n,s}^*\in\phi_{l,\Xi}$ of order $l\ge p+2$ such that $\|h_{n,s}^*-h_s^*\|_\infty=O(n^{-p\nu})$. (de Boor, 2001, p.145). Then using Conditions 1-3, 5 and 6, DCT, and Glivenko-Cantelli Theorem, the same arguments used in the proof of 4.1 lead to

$$\begin{aligned}\mathbb{P}_n\rho_s\left(\hat{\theta}_n,\hat{h}_{n,s};\hat{\sigma}_n^2\right)-\mathbb{P}_n\rho_s\left(\hat{\theta}_n,h_s^*;\hat{\sigma}_n^2\right) &= \mathbb{P}_n\rho_s\left(\hat{\theta}_n,\hat{h}_{n,s};\hat{\sigma}_n^2\right)-\mathbb{P}_n\rho_s\left(\hat{\theta}_n,h_{n,s}^*;\hat{\sigma}_n^2\right)\\ &\quad+\mathbb{P}_n\rho_s\left(\hat{\theta}_n,h_{n,s}^*;\hat{\sigma}_n^2\right)-\mathbb{P}_n\rho_s\left(\hat{\theta}_n,h_s^*;\hat{\sigma}_n^2\right)\\ &\le(\mathbb{P}_n-P)\left(\rho_s\left(\hat{\theta}_n,h_{n,s}^*;\hat{\sigma}_n^2\right)-\rho_s\left(\hat{\theta}_n,h_s^*;\hat{\sigma}_n^2\right)\right)\\ &\quad+P\left(\rho_s\left(\hat{\theta}_n,h_{n,s}^*;\hat{\sigma}_n^2\right)-\rho_s\left(\hat{\theta}_n,h_s^*;\hat{\sigma}_n^2\right)\right)\\ &=o_p(1).\end{aligned}$$

Since $\mathfrak{S}$ is Glivenko-Cantelli, the above inequality can be also rewritten as

$$\begin{aligned}\mathbb{P}_n\rho_s\left(\hat{\theta}_n,\hat{h}_{n,s};\hat{\sigma}_n^2\right) &\leq \mathbb{P}_n\rho_s\left(\hat{\theta}_n,h_s^*;\hat{\sigma}_n^2\right)+o_p(1)\\ &= \left(\mathbb{P}_n-P\right)\rho_s\left(\hat{\theta}_n,h_s^*;\hat{\sigma}_n^2\right)+P\rho_s\left(\hat{\theta}_n,h_s^*;\hat{\sigma}_n^2\right)+o_p(1)\\ &= P\rho_s\left(\hat{\theta}_n,h_s^*;\hat{\sigma}_n^2\right)+o_p(1). \qquad (16)\end{aligned}$$

Hence with Conditions 1-3, 5 and 6, it follows that

$$\begin{aligned}&P\left(\rho_s\left(\theta_0,\hat{h}_{n,s};\sigma_0^2\right)-\rho_s\left(\theta_0,h_s^*;\sigma_0^2\right)\right)\\ &\quad=P\left(\rho_s\left(\theta_0,\hat{h}_{n,s};\sigma_0^2\right)-\rho_s\left(\hat{\theta}_n,\hat{h}_{n,s};\hat{\sigma}_n^2\right)\right)-P\left(\rho_s\left(\theta_0,h_s^*;\sigma_0^2\right)-\rho_s\left(\hat{\theta}_n,h_s^*;\hat{\sigma}_n^2\right)\right)\\ &\qquad+P\left(\rho_s\left(\hat{\theta}_n,\hat{h}_{n,s};\hat{\sigma}_n^2\right)-\rho_s\left(\hat{\theta}_n,h_s^*;\hat{\sigma}_n^2\right)\right)\\ &\quad=o_p(1)+P\left(\rho_s\left(\hat{\theta}_n,\hat{h}_{n,s};\hat{\sigma}_n^2\right)-\rho_s\left(\hat{\theta}_n,h_s^*;\hat{\sigma}_n^2\right)\right)\\ &\qquad\text{by the consistency of } (\hat{\theta}_n,\hat{\sigma}_n^2) \text{ and DCT}\\ &\quad\leq o_p(1)-\left(\mathbb{P}_n-P\right)\rho_s\left(\hat{\theta}_n,\hat{h}_{n,s};\hat{\sigma}_n^2\right) \quad\text{by}\quad (16)\\ &\quad=o_p\left(1\right)\quad\text{by Glivenko-Cantelli Theorem.}\end{aligned}$$

With the uniqueness of h_s^*, the event $\|\hat{h}_{n,s}-h_s^*\|_{L_2(\mu)}>\epsilon$ is a subset of the event $P\rho_s\left(\theta_0,\hat{h}_{n,s};\sigma_0^2\right)>P\rho_s\left(\theta_0,h_s^*;\sigma_0^2\right)$ and the latter goes to zero in probability as $n\to\infty$ by the preseeding inequality. Let $\epsilon\to 0$ we conclude $\|\hat{h}_n-h^*\|_H\to_p 0$.

Nest, we show the consistency of both $\hat{A}_n$ and $\hat{B}_n$. Denote

$$\rho\left(\theta,h;\sigma^2\right)=\left(m_1\left(\theta;\sigma^2,X\right)-m_2\left(\theta;\sigma^2\right)[h]\right)^{\otimes 2}$$

and $\mathfrak{S}_1=\{\rho\left(\theta,h;\sigma^2\right):\theta\in\mathcal{R}^d\times\exp(\psi_{l,\Xi}),h\in\phi_{l,\Xi}^d;\sigma^2\in\mathcal{R}^+\}$, where $\phi_{l,\Xi}^d=\{h=(h_1,h_2,\cdots,h_d):h_s\in\phi_{l,\Xi},\ 1\leq s\leq d\}$. Then $\mathfrak{S}_1$ consists of d individual $\mathfrak{S}$ and hence a Glivenko-Cantelli as well. By Conditions 1-3, 5 and 6, the consistency of $(\hat{\theta}_n,\hat{\sigma}_n^2,\hat{h}_n)$, the Glivenko-Cantelli and DCT theorems, it follows that

$$\begin{aligned}\hat{B}_n &= \mathbb{P}_n\rho\left(\hat{\theta}_n,\hat{h}_n;\hat{\sigma}_n^2\right)^{\otimes 2}=\left(\mathbb{P}_n-P\right)\rho\left(\hat{\theta}_n,\hat{h}_n;\hat{\sigma}_n^2\right)^{\otimes 2}+P\rho\left(\hat{\theta}_n,\hat{h}_n;\hat{\sigma}_n^2\right)^{\otimes 2}\\ &\to_p P\rho\left(\theta_0,h^*;\sigma_0^2,X\right)^{\otimes 2}=B_0.\end{aligned}$$

Let $\rho^*\left(\theta,h;\sigma^2\right)=m_{11}\left(\theta;\sigma^2\right)-m_{21}\left(\theta;\sigma^2\right)[h]$, we can similarly show that the class $\mathfrak{S}_2=\{\rho^*\left(\theta,h;\sigma^2\right):\theta\in\mathcal{R}^d\times\exp(\psi_{l,\Xi}),h\in\phi_{l,\Xi}^d;\sigma^2\in\mathcal{R}^+\}$ is Glivenko-Cantelli. And

$$\begin{aligned}\hat{A}_n &= -\mathbb{P}_n\rho^*\left(\hat{\theta}_n,\hat{h}_n;\hat{\sigma}_n^2\right)=-\left(\mathbb{P}_n-P\right)\rho^*\left(\hat{\theta}_n,\hat{h}_n;\hat{\sigma}_n^2\right)-P\rho^*\left(\hat{\theta}_n,\hat{h}_n;\hat{\sigma}_n^2\right)\\ &\to_p -P\rho^*\left(\theta_0,h^*;\sigma_0^2\right)=A_0.\end{aligned}$$

The proof is complete.

A.4. Two technical lemmas

To study the asymptotic properties of the proposed estimate of (β_0, Λ_0), the following technical lemmas are critical.

LEMMA A.1. *Denote* $\mathbb{M}\left(\beta, \Lambda, \sigma^2\right) = Pm\left(\beta, \Lambda, \sigma^2\right) \forall \left(\beta, \Lambda, \sigma^2\right) \in \mathcal{R}^d \times \mathcal{F} \times \mathcal{R}^+$. *Suppose Conditions 1, 3-5 hold, then*

(i) $\mathbb{M}\left(\beta_0, \Lambda_0, \sigma^2\right) \geq \mathbb{M}\left(\beta, \Lambda, \sigma^2\right)$ *for any* $(\beta, \Lambda) \in \mathcal{R}^d \times \mathcal{F}, \sigma^2 \in \mathcal{R}^+$ *and the equality hold iff* $\beta = \beta_0$ *and* $\Lambda = \Lambda_0$ *a.e with respect to* μ.

(ii) There exists a constant C, such that

$$\mathbb{M}\left(\beta_0, \Lambda_0, \sigma^2\right) - \mathbb{M}\left(\beta, \Lambda, \sigma^2\right) \geq Cd^2\left(\left(\beta_0, \Lambda_0\right), (\beta, \Lambda)\right)$$

for any (β, Λ) *in a neighborhood of* (β_0, Λ_0) *and* $\sigma^2 \in \mathcal{R}^+$.

PROOF. First, we prove the uniqueness of the maximizer. A straightforward algebra shows that

$$\begin{aligned}
&\mathbb{M}\left(\beta_0, \Lambda_0, \sigma^2\right) - \mathbb{M}\left(\beta, \Lambda, \sigma^2\right) \\
&= P\left(\sum_{j=1}^{K}\left(\triangle\Lambda_{0,j} e^{\beta_0^T Z} log \frac{\triangle\Lambda_{0,j} e^{\beta_0^T Z}}{\triangle\Lambda_j e^{\beta^T Z}}\right) - \left(\Lambda_{0,K} e^{\beta_0^T Z} + 1/\sigma^2\right) log \frac{\Lambda_{0,K} e^{\beta_0^T Z} + 1/\sigma^2}{\Lambda_K e^{\beta^T Z} + 1/\sigma^2}\right) \\
&= PI_1 + PI_2
\end{aligned}$$

where

$$\begin{aligned}
I_1 &= \sum_{j=1}^{K}\left(\triangle\Lambda_{0,j} e^{\beta_0^T Z} log \frac{\triangle\Lambda_{0,j} e^{\beta_0^T Z}}{\triangle\Lambda_j e^{\beta^T Z}}\right) - \left(\Lambda_{0,K} e^{\beta_0^T Z}\right) log \frac{\Lambda_{0,K} e^{\beta_0^T Z}}{\Lambda_K e^{\beta^T Z}} \\
&= \Lambda_{0,K} e^{\beta_0^T Z} \sum_{j=1}^{K}\left(\frac{\triangle\Lambda_{0,j}}{\Lambda_{0,K}} log \frac{\triangle\Lambda_{0,j}/\Lambda_{0,K}}{\triangle\Lambda_j/\Lambda_K}\right)
\end{aligned}$$

and

$$I_2 = \left(\Lambda_{0,K} e^{\beta_0^T Z}\right) log \frac{\Lambda_{0,K} e^{\beta_0^T Z}}{\Lambda_K e^{\beta^T Z}} - \left(\Lambda_{0,K} e^{\beta_0^T Z} + 1/\sigma^2\right) log \frac{\Lambda_{0,K} e^{\beta_0^T Z} + 1/\sigma^2}{\Lambda_K e^{\beta^T Z} + 1/\sigma^2}.$$

Note that $\sum_{j=1}^{K}\left(\frac{\triangle\Lambda_{0,j}}{\Lambda_{0,K}} log \frac{\triangle\Lambda_{0,j}/\Lambda_{0,K}}{\triangle\Lambda_j/\Lambda_K}\right)$ is the Kullback-Leibler's information $K_{p_0}\left(p_0, p\right)$ with $p_{0,j} = \frac{\triangle\Lambda_{0,j}}{\Lambda_{0,K}}$ and $p_j = \frac{\triangle\Lambda_j}{\Lambda_K}$ for $j = 1, 2, \cdots, K$. So, it is nonnegative and the equality hold when $\frac{\triangle\Lambda_{0,j}}{\Lambda_{0,K}} = \frac{\triangle\Lambda_j}{\Lambda_K}, j = 1, 2, \cdots, K$. Therefore, $PI_1 \geq 0$ and $PI_1 = 0$ iff

$$\Lambda = C\Lambda_0 \text{ a.e. w.r.t } \mu. \text{ for some constant } C \tag{17}$$

To show $PI_2 \geq 0$, we denote $x = \Lambda_{0,K} e^{\beta_0^T Z} > 0, b = \Lambda_K e^{\beta^T Z} - \Lambda_{0,K} e^{\beta_0^T Z}$ for the notational simplicity and let

$$f(b) = xlog \frac{x}{x + b} - \left(x + 1/\sigma^2\right) log \frac{x + 1/\sigma^2}{x + 1/\sigma^2 + b}, \quad x > 0, x + b > 0$$

A straightforward algebra yields that $\partial f(b)/\partial b = 0$ only if $b = 0$ and

$$\frac{\partial^2 f(b)}{\partial b^2} = \begin{cases} > 0 & \text{for } \ -x < b \leq \sqrt{x\left(x + 1/\sigma^2\right)} \\ < 0 & \text{for } \ b > \sqrt{x\left(x + 1/\sigma^2\right)} \end{cases}$$

This implies that $f\left(b\right)$ reaches its minimum at $b = 0$ and $f\left(0\right) = 0$. So $PI_2 \geq 0$ and the equality hold when

$$\Lambda e^{\beta^T Z} = \Lambda_0 e^{\beta_0^T Z} \text{ a.e. w.r.t. } \mu.$$

Then the result of the first part follows using the same argument as given by Wellner and Zhang (2007).

Now we prove the second part of the lemma. I_1 can be rewritten as following,

$$\begin{aligned} I_1 &= \Lambda_{0,K} e^{\beta_0^T Z} \sum_{j=1}^{K} \left(\frac{\triangle \Lambda_{0,j}}{\Lambda_{0,K}} log \frac{\triangle \Lambda_{0,j}/\Lambda_{0,K}}{\triangle \Lambda_j/\Lambda_K} \right) \\ &= \Lambda_{0,K} e^{\beta_0^T Z} \sum_{j=1}^{K} \left[\frac{\triangle \Lambda_j}{\Lambda_K} \left(\frac{\triangle \Lambda_{0,j}/\Lambda_{0,K}}{\triangle \Lambda_j/\Lambda_K} log \frac{\triangle \Lambda_{0,j}/\Lambda_{0,K}}{\triangle \Lambda_j/\Lambda_K} - \frac{\triangle \Lambda_{0,j}/\Lambda_{0,K}}{\triangle \Lambda_j/\Lambda_K} + 1 \right) \right] \\ &\geq \frac{1}{4} \Lambda_{0,K} e^{\beta_0^T Z} \sum_{j=1}^{K} \frac{\triangle \Lambda_j}{\Lambda_K} \left(\frac{\triangle \Lambda_{0,j}/\Lambda_{0,K}}{\triangle \Lambda_j/\Lambda_K} - 1 \right)^2 = \frac{1}{4} \Lambda_{0,K} e^{\beta_0^T Z} \sum_{j=1}^{K} \frac{1}{\triangle \Lambda_j/\Lambda_K} \left(\frac{\triangle \Lambda_{0,j}}{\Lambda_{0,K}} - \frac{\triangle \Lambda_j}{\Lambda_K} \right)^2 \\ &\geq \frac{1}{4} \Lambda_{0,K} e^{\beta_0^T Z} \sum_{j=1}^{K} \left(\frac{\triangle \Lambda_{0,j}}{\Lambda_{0,K}} - \frac{\triangle \Lambda_j}{\Lambda_K} \right)^2 \end{aligned}$$

The first inequality is due to the fact that $xlogx - x + 1 \geq \frac{1}{4}(x-1)^2$ for x in a neighborhood of $x = 1$, the equality hold only when $x = 1$.

Performing Taylor expansion for $f(b)$ at 0 yields

$$f(b) = \frac{1/\sigma^2 \left[x\left(x + 1/\sigma^2\right) - \xi^2\right]}{2\left(x+\xi\right)^2 \left(x + 1/\sigma^2 + \xi\right)^2} b^2 \ \text{ for a } \ |\xi| < |b|$$

When b is in a neighborhood of zero, $|b| < |x|$ at almost everywhere in μ. It follows that the numerator of I_2 can be bounded below by

$$1/\sigma^2 \left[x\left(x + 1/\sigma^2\right) - \xi^2\right] b^2 \geq 1/\sigma^2 \left[x\left(x + 1/\sigma^2\right) - x^2\right] b^2 = \left(1/\sigma^2\right)^2 x b^2;$$

and the denominator can be bounded above by

$$2\left(x+\xi\right)^2 \left(x + 1/\sigma^2 + \xi\right)^2 \leq 2(2x)^2 (x + 1/\sigma^2 + x)^2 = 8x^2 (2x + 1/\sigma^2)^2.$$

Hence $f(b) \geq \frac{\left(1/\sigma^2\right)^2}{8x(2x+1/\sigma^2)^2} b^2$ and therefore

$$I_2 \geq \frac{\left(1/\sigma^2\right)^2}{8\Lambda_{0,K} e^{\beta_0^T Z} \left(2\Lambda_{0,K} e^{\beta_0^T Z} + 1/\sigma^2\right)^2} \left(\Lambda_{0K} e^{\beta_0^T Z} - \Lambda_K e^{\beta^T Z}\right)^2 \ \text{ a.e. w.r.t. } \mu$$

Combine the inequalities for I_1 and I_2, we have,

$$I_1 + I_2 \geq \frac{1}{4} \Lambda_{0,K} e^{\beta_0^T Z} \times \frac{1}{k^2} \sum_{j=1}^{K} \left[k^2 \left(\theta_{j1} - \theta_{j2}\right)^2 + \left(l_1 - l_2\right)^2\right] \ \text{ a.e. w.r.t. } \mu,$$

where

$$k = \frac{\sqrt{2K}\Lambda_{0,K}e^{\beta_0^T Z}\left(2\Lambda_{0,K}e^{\beta_0^T Z} + 1/\sigma^2\right)}{1/\sigma^2},$$

$$\theta_{j1} = \frac{\triangle\Lambda_{0,j}}{\Lambda_{0,K}}, \theta_{j2} = \frac{\triangle\Lambda_j}{\Lambda_K}, l_1 = \Lambda_{0,K}e^{\beta_0^T Z} \text{ and } l_2 = \Lambda_K e^{\beta^T Z}.$$

When $l_1 = l_2$, $I_1 + I_2 \geq \frac{1}{4}\Lambda_{0,K}e^{\beta_0^T Z} \times \sum_{j=1}^K (\theta_{j1} - \theta_{j2})^2$. Therefore

$$P(I_1 + I_2) \geq CP\sum_{j=1}^K \left(\triangle\Lambda_{0,j}e^{\beta_0^T Z} - \Delta\Lambda_j e^{\beta^T Z}\right)^2.$$

We now show that this inequality is also true when $l_1 \neq l_2$. We claim that for $C = \frac{1}{2} \wedge \frac{k^2}{(l_1 \wedge l_2)^2}$, we have

$$k^2(\theta_1 - \theta_2)^2 + (l_1 - l_2)^2 \geq C(l_2\theta_2 - l_1\theta_1)^2 \quad \forall 0 \leq \theta_1 \leq 1, 0 \leq \theta_2 \leq 1, l_1 \geq \gamma_1, l_2 \geq \gamma_2$$

for some $\gamma_1 > 0$ and $\gamma_2 > 0$.

First we discuss the case when l_1, l_2 and θ_1, θ_2 are concordant, e.g. $(l_1 - l_2)(\theta_1 - \theta_2) \geq 0$. Without lost of generality, we assume $l_1 > l_2$ and $\theta_1 \geq \theta_2$.

$$\begin{aligned} k^2(\theta_1 - \theta_2)^2 + (l_1 - l_2)^2 &\geq \frac{1}{2}(k(\theta_1 - \theta_2) + (l_1 - l_2))^2 \\ &\geq \frac{1}{2}(k(\theta_1 - \theta_2) + (l_1 - l_2)\theta_1)^2 \\ &= \frac{1}{2}(l_1\theta_1 - l_2\theta_2 + (k - l_2)(\theta_1 - \theta_2))^2 \end{aligned} \tag{18}$$

Since

$$\begin{aligned} k &= \frac{\sqrt{2K}\Lambda_{0,K}e^{\beta_0^T Z}\left(2\Lambda_{0,K}e^{\beta_0^T Z} + 1/\sigma^2\right)}{1/\sigma^2} \geq \sqrt{2K}\Lambda_{0,K}e^{\beta_0^T Z} \geq \Lambda_{0,K}e^{\beta_0^T Z} \\ &\geq min\left(\Lambda_{0,K}e^{\beta_0^T Z}, \Lambda_K e^{\beta^T Z}\right) = l_2. \text{ a.e. w.r.t. } \mu \end{aligned}$$

By (18), $k^2(\theta_1 - \theta_2)^2 + (l_1 - l_2)^2 \geq \frac{1}{2}(l_1\theta_1 - l_2\theta_2)^2$ a.e. w.r.t. μ.

For a discordant pair, say, $l_1 < l_2, \theta_1 \geq \theta_2$, we further discuss the claim in two cases:

(a) When $l_1\theta_1 \geq l_2\theta_2$ we have

$$\theta_1 - \theta_2 = \frac{1}{l_1}(l_1\theta_1 - l_1\theta_2) > \frac{1}{l_1}(l_1\theta_1 - l_2\theta_2) \geq 0$$

So $(\theta_1 - \theta_2)^2 > \frac{1}{l_1^2}(l_1\theta_1 - l_2\theta_2)^2$.

(b) When $l_1\theta_1 < l_2\theta_2$ we have

$$l_2 - l_1 \geq l_2\theta_2 - l_1\theta_2 \geq l_2\theta_2 - l_1\theta_1 > 0$$

So $(l_2 - l_1)^2 > (l_1\theta_1 - l_2\theta_2)^2$.

Therefore, $k^2 (\theta_1 - \theta_2)^2 + (l_1 - l_2)^2 \geq C (l_1\theta_1 - l_2\theta_2)^2$ with $C = \frac{1}{2} \wedge \frac{k^2}{(l_1 \wedge l_2)^2}$.

So,

$$\begin{aligned} P(I_1 + I_2) \geq & P\left\{ \frac{1}{4}\Lambda_{0,K} e^{\beta_0^T Z} \times \left(\frac{1}{2k^2} \wedge \frac{1}{(\Lambda_{0,K} e^{\beta_0^T Z} \wedge \Lambda_K e^{\beta^T Z})^2} \right) \sum_{j=1}^{K} (l_2\theta_{j2} - l_1\theta_{j1})^2 \right\} \\ \geq & CP \sum_{j=1}^{K} \left(\triangle\Lambda_{0,j} e^{\beta_0^T Z} - \Delta\Lambda_j e^{\beta^T Z} \right)^2 \end{aligned}$$

due to the compactness of the parameter space of β, Λ and the boundness of the $(Z, K, \underline{T}_K)$ specified in Conditions 1,2 and 5.

Finally, following the same proof as in Wellner and Zhang (2007), with Condition 4 the above inequality further implies

$$\mathbb{M}\left(\beta_0, \Lambda_0, \sigma^2\right) - \mathbb{M}\left(\beta, \Lambda, \sigma^2\right) \geq C\left\{ |\beta - \beta_0|^2 + \|\Lambda - \Lambda_0\|_{L_2(\mu)}^2 \right\}$$

Hence the proof for Lemma A.1 is complete.

Let

$$\phi_{l,\Xi}(\delta) = \left\{ \sum_{i=1}^{q_n} a_i B_i : \sum_{i=1}^{q_n} a_i B_i \in \phi_{l,\Xi} \text{ and } \sum_{i=1}^{q_n} |a_i| \leq \delta \text{ for some constant } \delta \right\}$$

LEMMA A.2. *The entropy of $\phi_{l,\Xi}(\delta)$, $\log N(\epsilon, \phi_{l,\Xi}(\delta), \|\cdot\|)$ with L_1-, L_2- and L_∞- norms can be shown bounded above by $Cq_n log(q_n^{1/2} \times \frac{\delta}{\epsilon}), Cq_n log(\frac{\delta}{\epsilon})$ and $Cq_n log\left(\frac{\delta}{q_n^{1/2}\epsilon}\right)$, respectively.*

The proof of this lemma follows exactly the same lines as those for the proof of Lemma 5 in Shen and Wong (1994) and is omitted.

Let

$$\psi_{l,\Xi}(\delta) = \left\{ \sum_{i=1}^{q_n} a_i B_i : \sum_{i=1}^{q_n} a_i B_i \in \psi_{l,\Xi} \text{ and } \sum_{i=1}^{q_n} |a_i| \leq \delta \text{ for some constant } \delta \right\}.$$

Because $\psi_{l,\Xi}(\delta) \subset \phi_{l,\Xi}(\delta)$, the entropy of $\psi_{l,\Xi}(\delta)$ with the three different norms aforementioned are also bounded by $Cq_n log(q_n^{1/2} \times \frac{\delta}{\epsilon}), Cq_n log(\frac{\delta}{\epsilon})$ and $Cq_n log\left(\frac{\delta}{q_n^{1/2}\epsilon}\right)$, respectively.

References

Byar, D., C. Blackard, and Vacurg (1980). Comparisons of placebo, pyridoxine, and topical thiotepa in preventing stage i bladder cancer. *Urology 10*(6), 556–561.

de Boor, C. (2001). *A Practical Guide to Splines.* Springer-Verlag.

Geman, A. and C. Hwang (1982). Nonparametric maximum likelihood estimation by the method of sieves. *The Annuals of Statistics 10*(2), 401–414.

Groeneboom, P. and J. A. Wellner (1992). *Information bounds and nonparametric Maximum Likelihood Estimation.* Birkhauser.

Huang, J. (1996). Efficient estimation for the proportional hazards model with interval censoringq. *The Annals of Statistics 24*(2), 540–568.

Huang, J., Y. Zhang, and L. Hua (2008). A least-squares approach to consistent information estimation in semiparametric models. *Technical Report, Dept. Biostat, Univ. of Iowa.*, Available at http://www.public–health.uiowa.edu/biostat/research/documents/2008–3–Huang–Zhang–Hua.pdf.

Lawless, J. F. and J. Nadeau (1995). Some simple robust methods for the analysis of recurrent events. *Technometrics 37*(2), 158–168.

Lin, D., L. Wei, I. Yang, and Z. Ying (2000). Semiparametric regression for the mean and rate functions of recurrent events. *J. R. Statist. Soc. B 62*(4), 711–730.

Lu, M., Y. Zhang, and J. Huang (2009). Semiparametric estimation methods for panel count data using monotone polynomial splines. *Journal of the American Statistical Association 104*(487), 1060–1070.

Schumaker, L. (1981). *Spline Functions: Basic Theory.* New York: Wiley.

Shen, X. and W. Wong (1994). Convergence rate of sieve estimates. *The Annals of Statistics 22*(2), 580–615.

Sun, J., D.-H. Park, L. Sun, and X. Zhao (2005). Semiparametric regression analysis of longitudinal data with informative observation times. *Journal of the American Statistical Association 100*(471), 882–889.

Sun, J. and L. Wei (2000). Regression analysis of panel count data with covariate-dependent observation and censoring times. *J. R. Statist. Soc. B 62*(2), 293–302.

van der Vaart, A. (1998). *Asymptotic Statistics.* Cambridge University Press.

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

Wei, L., D. Lin, and L. Weissfeld (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distribution. *Journal of the American Statistical Association 84*(408), 1065–1073.

Wellner, J. A. and Y. Zhang (2000). Two estimators of the mean of a counting process with anel count data. *The Annals of Statistics 28*(3), 779–814.

Wellner, J. A. and Y. Zhang (2007). Two likelihood-based semiparametric estimation methods for panel count data with covariates. *The Annals of Statistics 35*(5), 2106–2142.

Zeger, S. L. (1988). A regression model for time series of counts. *Biometrika 75*(4), 621–629.

Zhang, Y. (2002). A semiparametric pseudo likelihood estimation method for panel count data. *Biometrika 89*(1), 39–48.

Zhang, Y. (2006). Nonparametric k-sample tests with panel count data. *Biometrika 93*(4), 777–790.