Bernoulli **30**(4), 2024, 3251–3275
https://doi.org/10.3150/23-BEJ1713

A robust approach for regression analysis of panel count data with time-varying covariates

DAYU SUN[1,a], YUANYUAN GUO[2,b], YANG LI[3,c], WANZHU TU[3,d] and JIANGUO SUN[4,e]

[1]*Department of Biostatistics and Bioinformatics, Emory University, GA 30322, U.S.A.,* [a]*dayu.sun@emory.edu*
[2]*Department of Biostatistics and Bioinformatics, Duke University, NC 27710, U.S.A.,* [b]*yuanyuan.guo@duke.edu*
[3]*Department of Biostatistics and Health Data Science, Indiana University, Indianapolis, IN 46202, U.S.A.,* [c]*yli18@iu.edu,* [d]*wtu1@iu.edu*
[4]*Department of Statistics, University of Missouri, MO 65211, U.S.A.,* [e]*sunj@missouri.edu*

The validity of statistical inference for panel count data with time-varying covariates depends on the correct specification of within-subject correlation structures; misspecification often leads to questionable inference. To alleviate, robust inference has been proposed for mean models, which implicitly assume monotone mean functions. When covariate values fluctuate with time, however, the assumed monotonicity becomes unrealistic. In this research, we propose a robust inference based on rate models that are free of such constraints. Since the asymptotic variance has no closed form under the rate model, we further develop computationally efficient robust variance estimators using the Expectation-Maximization (EM) algorithm, thus sidestepping the need for computationally intensive numerical methods, which could undermine the robustness. Rigorous theoretical development is provided in support of parameter estimation and inference. Extensive simulation studies demonstrate the superiority of the proposed method. We present a real clinical application to illustrate the use of the proposed method.

Keywords: EM algorithm; panel count data; robust variance estimation; semiparametric model; time-to-event data

1. Introduction

In recurrent event studies, precise event times are typically unknown when observations are made periodically at discrete points. Instead, one only obtains counts of events that occur between the consecutive observation points. Such data are referred to as panel count data. Panel count data are abundant in event history studies [2,3,14,17,34,47]. A common interest of panel count data analysis is to determine the effects of covariates on the event rate. In practice, covariates could vary with time, or there may exist an interaction between a time-invariant covariate and time, thus further complicating the analysis. An application that motivated the current research is the Young Women's Project (YWP), an epidemiological study of the sexual behaviors' effects on sexually transmitted infections (STIs) [13,36]. The YWP yielded panel count data on STIs as the exact times of infection were unknown due to its quarterly testing schedule. The subjects' sexual activities since the last visit often varied with time and were summarized as time-varying covariates.

Various models for panel count data with time-varying covariates have been proposed. Let $N(t)$ be the counting process for the total number of recurrent events up to time t; the process jumps with increment 1. For the analysis, a natural approach is to characterize the relationship between $N(t)$ and a vector of covariates denoted by $\mathbf{X}(t) \in \mathbb{R}^p$ in the following intensity process model

$$E\left[dN(t)\,|\,\mathcal{F}_{t-}\right] = \exp(\boldsymbol{\beta}^\top \mathbf{X}(t))\,d\Lambda(t), \tag{1}$$

[1]. In the above model, $dN(t) = N(t+dt-) - N(t-)$ denotes the increment of N within the small interval $[t, t+dt)$, $\mathcal{F}_t$ is the σ-field generated by $\{N(s), \mathbf{X}(s), 0 \le s \le t\}$, $\boldsymbol{\beta} \in \mathbb{R}^p$ represents covariate

effects, and $\Lambda(t)$ is an unknown baseline cumulative intensity function. Model (1) specifies the covariate effects to be proportional to the instantaneous probability of event occurrence at time t given the history of $N(\cdot)$ and $\mathbf{X}(\cdot)$ before t. Implicitly, model (1) requires a modulated Poisson process where the intensity at t depends solely on $\mathbf{X}(t)$ at time t [6]. However, the Poisson assumption is often unrealistic because it implies the event counts of the same subject from non-overlapping intervals are mutually independent, which is generally untrue in longitudinal studies such as YWP. Thus, the validity of inference in panel count data analysis is contingent on the correct specification of the within-subject correlation structures.

To relax the Poisson assumption, two semiparametric models have been considered for panel count data regression. One is the proportional rates model

$$E\left[dN(t)\,|\,\mathbf{X}(t)\right] = \exp(\boldsymbol{\beta}^\top \mathbf{X}(t))d\Lambda(t), \tag{2}$$

[24], and the other is the proportional means model

$$E\left[N(t)\,|\,\mathbf{X}(t)\right] = \mu(t)\exp(\boldsymbol{\beta}^\top \mathbf{X}(t)), \tag{3}$$

[15,33]. Here, $\boldsymbol{\beta}$ is defined as in model (1), $\Lambda(t)$ is an unspecified non-decreasing cumulative baseline rate function, and $\mu(t)$ is an unspecified non-decreasing baseline mean function. It is easy to see that model (2) is equivalent to model (1) under the Poisson assumption, and models (2) and (3) are the same with $\mu(t) = \Lambda(t)$ when the covariates are time-invariant. But with time-varying covariates, models (2) and (3) are generally not equivalent, and different inference procedures are therefore needed.

To our knowledge, the existing robust estimation procedures are all based on the means model (3). For example, Sun and Wei [33] and Hu, Sun and Wei [15] discussed the estimation for model (3), and they proposed to ascertain covariate effects by using estimating equation-based methods without relying on the Poisson assumption. Under the Poisson assumption, Wellner and Zhang [43] investigated the full and pseudo maximum likelihood estimations and proved that the resulting estimators are still consistent when the Poisson assumption does not hold. Hua, Zhang and Tu [16] later generalized Wellner and Zhang's approach by specifically modeling the overdispersion by adding a parametric Gamma-frailty and showed that the extended method is robust with respect to the parametric assumption. Similar methods were also discussed by Zhao, Tu and Yu [47] and Wang and Yu [41]. However, all of the above methods are for the situation of time-invariant covariates, thus are not readily applicable to studies such as the YWP.

For panel count data with time-varying covariates, robust methods are limited, and they are generally developed within the framework of proportional mean model (3) [7,19,20]. A specific challenge for model (3) is that with time-varying covariates, it is impossible to guarantee the monotonicity of the baseline mean function $\mu(t)$. In fact, most robust methods under model (3) produce non-monotonic estimates of the baseline mean function, and thus severely undermining their practical utility.

In contrast, the proportional rate model (2) is not subject to this limitation as it only requires the right-hand side of (2) to be positive, and the time-varying covariates can therefore be readily accommodated. Furthermore, model (2) has an appealing feature that $E\left[dN(t)\,|\,\mathbf{X}(t)\right]$ stands for the marginal instantaneous probability of an event occurring at time t given the covariates at t or the instantaneous relative risk, analogous to the hazard function in survival analysis.

The rate model (2), despite its inherent advantage, has not been carefully studied for panel count regression with time-varying covariates. Zeng and Lin [45] investigated the maximum likelihood estimation of model (1) with time-varying covariates, which could apply to model (2) under the Poisson assumption. Their method, however, relies heavily on the Poisson assumption and the correct specification of within-subject correlations, which is difficult to justify in practical data analysis. Additionally, the lack of a closed-form estimator for their asymptotic variance also presents a challenge: Without an explicit formulation for the asymptotic variances of $\hat{\boldsymbol{\beta}}$, it will be difficult to carry out inference. Numerical derivatives of the profile log-likelihood function [28,29] may provide some relief, and that approach has been used by Su and Wang [31], Zeng, Mao and Lin [46] and Zeng and Lin [45] with slight modifications. Nevertheless, the profile likelihood method is not only computationally intensive but also vulnerable to model misspecification.

To overcome these difficulties, we develop a robust likelihood-based estimation procedure for model (2) with time-varying covariates. Key contributions of the paper are:

(i) We provide rigorous proofs showing that consistency and asymptotically normality of the robust estimator can be achieved without relying on correctly specified within-subject correlation structures. For implementation, we propose an efficient Expectation-Maximization (EM) algorithm.

(ii) We propose two novel variance estimators in tractable forms, thus greatly improving computational efficiency. One is robust to the within-subject correlation assumption, and the other serves as a more efficient substitute for the profile likelihood variance estimates. Notably, the robust variance estimator is a sandwich-type estimator based on the efficient score constructed under the working Poisson assumption in a fashion analogous to Lin et al. [24] and Wellner and Zhang [43]. But unlike the published work, the forms of these two variance estimators are based on the posterior score function of $\boldsymbol{\beta}$ in the EM algorithm. Specifically, since the least-favorable direction has no explicit form, we estimate it directly by the derivative of the profiled baseline rate function.

(iii) We provide theoretical justifications for the variance estimators by establishing a general connection between the EM algorithm and the efficient score function of $\boldsymbol{\beta}$ *regardless of distribution assumptions*, based on the modern empirical process theory and semiparametric theory. To our knowledge, there is scarce literature on the consistency of variance estimators based on the EM algorithm under semiparametric settings. The existing literature, e.g. Oakes [30] and Louis [26], only derived the population forms of some estimator variance based on the EM algorithm for parametric models. The proposed variance estimation methods could be generalized to other M-estimators lacking explicit forms of asymptotic normality, which has appeared in the literature, including, for example, Zeng, Mao and Lin [46], Li et al. [22] and Li et al. [21].

(iv) Finally, we approximate the two variance estimators with the least favorable direction estimated by the mean of covariates weighted by the estimated subject-specific rate. This further improves the computational efficiency of the variance estimators, as we shall demonstrate through simulation. We theoretically justify the approximated variance estimators when the time-varying covariates are constant within each observation interval, which is the case in YWP.

The remainder of the paper is organized as follows. After introducing the notation and assumptions, we present the estimation procedure under the Poisson working assumption and develop an EM algorithm in Section 2. In Section 3, we establish the asymptotic properties of the estimators without relying on the Poisson assumption, which confirms the robustness of the estimator. In Section 4, we present the two variance estimators and show their consistency. In Section 5, we demonstrate through simulation the robustness and superiority of the proposed method and variance estimation in practical situations. In Section 6, we analyze the YWP data using the proposed methods. We conclude the paper in Section 7 with a few brief remarks. Section 8 contains proofs of the asymptotic results along with variance estimation derivation details.

2. Estimation methodology

Consider an event history study that consists of n independent subjects. For subject i, let $N_i(t)$ denote the total number of the occurrences of the recurrent event of interest up to time t and suppose that $N_i(t)$ is observed only at the sequence of time points $T_{i0} = 0 < T_{i1} < \ldots < T_{iJ_i}$. For subject

i, define the censoring indicator $\Delta_i(t) = I\left(T_{iJ_i} > t\right)$ and let $\mathbf{X}_i(t)$ denote a p-dimensional vector of the possibly time-varying covariates associated with the subject. The observed data have the form $O = \{O_i = (\{N_i(T_{ij})\}_{j=1}^{J_i}, \mathbf{X}_i(\cdot), \{T_{ij}\}_{j=1}^{J_i}), i = 1, \ldots, n\}$. That is, only panel count data are available.

Define $\Delta N_{ij} = N_i\left(T_{ij}\right) - N_i\left(T_{i(j-1)}\right)$, the number of the occurrences of the recurrent event of interest between T_{ij} and $T_{i(j-1)}$ in subject i, and assume that $N_i(t)$ satisfies model (2) and follows the non-homogeneous Poisson process, $j = 1, \ldots, J_i$, $i = 1, \ldots, n$. Then the observed likelihood function has the form

$$L(\boldsymbol{\beta}, \Lambda | O) = \prod_{i=1}^{n} \prod_{j=1}^{J_i} \frac{1}{\Delta N_{ij}!} \exp\left(-\int_{T_{i(j-1)}}^{T_{ij}} \exp\left(\boldsymbol{\beta}^\top \mathbf{X}_i(t)\right) d\Lambda(t)\right) \times \left\{\int_{T_{i(j-1)}}^{T_{ij}} \exp\left(\boldsymbol{\beta}^\top \mathbf{X}_i(t)\right) d\Lambda(t)\right\}^{\Delta N_{ij}}, \quad (4)$$

and it is natural to maximize it for the estimation of $\boldsymbol{\beta}$ and Λ. However, since $\mathbf{X}_i(t)$ is time-varying, and we do not know the exact form of $\Lambda(t)$, the integral $\int_{T_{i(j-1)}}^{T_{ij}} \exp\left(\boldsymbol{\beta}^\top \mathbf{X}_i(t)\right) d\Lambda(t)$ cannot be easily calculated. The use of numerical integration methods such as Gaussian quadrature would require a certain degree of smoothness of the integrands and also the roughness of $\mathbf{X}_i(t)$ could lead to numerical inaccuracy in integration. In other words, the maximization of (4) is complicated or not straightforward. To overcome the challenge, we develop an EM algorithm, as inspired by Zeng and Lin [45].

For the development of the EM algorithm and to make the integration in (4) tractable, we adopt the nonparametric approach by treating Λ to be a step function with a nonnegative jump size λ_k at t_k with $\lambda_0 = 0$ and $0 = t_0 < t_1 < \ldots < t_K$ denoting the ordered unique values of the T_{ij}'s for all i and j. Let $\{W_{ik}; i = 1, \ldots, n, k = 1, \ldots, K\}$ be mutually independent Poisson random variables with means $\exp\left(\boldsymbol{\beta}^\top \mathbf{X}_i(t_k)\right) \lambda_k$. Then the likelihood function given in (4) can be rewritten as

$$L(\boldsymbol{\beta}, \Lambda | O) = \prod_{i=1}^{n} \prod_{j=1}^{J_i} \frac{1}{\Delta N_{ij}!} \exp\left(-\sum_{T_{i(j-1)} < t_k \le T_{ij}} \exp\left(\boldsymbol{\beta}^\top \mathbf{X}_i(t_k)\right) \lambda_k\right) \times \left\{\sum_{T_{i(j-1)} < t_k \le T_{ij}} \exp\left(\boldsymbol{\beta}^\top \mathbf{X}_i(t_k)\right) \lambda_k\right\}^{\Delta N_{ij}} = \prod_{i=1}^{n} \prod_{j=1}^{J_i} \Pr\left(\sum_{T_{i(j-1)} < t_k \le T_{ij}} W_{ik} = \Delta N_{ij}\right).$$

This suggests that one can augment the observed data by adding W_{ik}'s as the pseudo complete data denoted by C for the development of the EM algorithm. It then follows that the pseudo complete-data log likelihood function has the form

$$\ell(\boldsymbol{\beta}, \boldsymbol{\lambda} | C) = \sum_{i=1}^{n} \sum_{k=1}^{K} \Delta_{ik} \left\{W_{ik}\left(\boldsymbol{\beta}^\top \mathbf{X}_i(t_k) + \log(\lambda_k)\right) - \exp\left(\boldsymbol{\beta}^\top \mathbf{X}_i(t_k)\right) \lambda_k\right\}, \quad (5)$$

where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_K)$ and $\Delta_{ik} = \Delta_i(t_k)$.

In the E-step of the EM algorithm, we need to find the posterior expectation of W_{ik} given the observed data and the estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ from the last iteration. Let j_{ik} be the observation time index for subject i such that $T_{i(j_{ik}-1)} < t_k \le T_{ij_{ik}}$ for any k and $A_{ik} = \{l : T_{i(j_{ik}-1)} < t_l \le T_{ij_{ik}}\}$. We write the qth iteration of $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ as $\boldsymbol{\beta}^{(q)}$ and $\boldsymbol{\lambda}^{(q)} = (\lambda_1^{(q)}, \ldots, \lambda_K^{(q)})$. In the qth iteration, the posterior mean of

W_{ik} given the observed data can be obtained as

$$\hat{E}\left[W_{ik}|\boldsymbol{\beta}^{(q)},\boldsymbol{\lambda}^{(q)}\right] = E\left[W_{ik}|\sum_{T_{i(j_{ik}-1)}<t_k\le T_{ij_{ik}}} W_{ik} = \Delta N_{ij_{ik}},\boldsymbol{\beta}^{(q)},\boldsymbol{\lambda}^{(q)}\right]$$

$$= \frac{\Delta_{ik}\exp\left(\boldsymbol{\beta}^{(q)\top}\mathbf{X}_i(t_k)\right)\lambda_k^{(q)}}{\sum_{l\in A_{ik}}\Delta_{il}\exp\left(\boldsymbol{\beta}^{(q)\top}\mathbf{X}_i(t_l)\right)\lambda_l^{(q)}}\Delta N_{ij_{ik}}, \tag{6}$$

because the conditional distribution of W_{ik} is a binomial distribution with the number of trials $\Delta N_{ij_{ik}}$ and the success probability $E[W_{ik}]/\sum_{l\in A_{ik}}E[W_{lk}]$.

In the M-step of the EM algorithm, it is straightforward to show that we can update $\boldsymbol{\lambda}$ by

$$\lambda_k^{(q+1)} = \sum_{i=1}^{n}\Delta_{ik}\hat{E}\left[W_{ik}|\boldsymbol{\beta}^{(q)},\boldsymbol{\lambda}^{(q)}\right]/\left\{\sum_{i=1}^{n}\Delta_{ik}\exp\left(\boldsymbol{\beta}^{(q)\top}\mathbf{X}_i(t_k)\right)\right\}, \tag{7}$$

$k = 1,\ldots,K$. Plugging in (7) into the posterior expectation of (5) given the observed data, one can update $\boldsymbol{\beta}$ by solving the following equation

$$\mathbf{U}\left(\boldsymbol{\beta}|\boldsymbol{\beta}^{(q)},\boldsymbol{\lambda}^{(q+1)}\right) = \sum_{i=1}^{n}\sum_{k=1}^{K}\Delta_{ik}\hat{E}\left[W_{ik}|\boldsymbol{\beta}^{(q)},\boldsymbol{\lambda}^{(q+1)}\right]\left(\mathbf{X}_i(t_k) - \bar{\mathbf{X}}(t_k;\boldsymbol{\beta})\right) = 0.$$

Here, $\bar{\mathbf{X}}(t;\boldsymbol{\beta}) = S^{(1)}(t,\boldsymbol{\beta})/S^{(0)}(t,\boldsymbol{\beta})$ with $S^{(u)}(t,\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i(t)^{\otimes u}\Delta_i(t)\exp\left(\boldsymbol{\beta}^{\top}\mathbf{X}_i(t)\right)$, where $a^{\otimes 0} = 1$, $a^{\otimes 1} = a$ and $a^{\otimes 2} = aa^{\top}$ for a vector a and $u = 0,1,2$. To reduce the computational burden, one can update $\boldsymbol{\beta}$ by using only the one-step Newton-Raphson at each M-step as $\boldsymbol{\beta}^{(q+1)} = \boldsymbol{\beta}^{(q)} - \left\{\dot{\mathbf{U}}\left(\boldsymbol{\beta}^{(q)}|\boldsymbol{\beta}^{(q)},\boldsymbol{\lambda}^{(q+1)}\right)\right\}^{-1}\mathbf{U}\left(\boldsymbol{\beta}^{(q)}|\boldsymbol{\beta}^{(q)},\boldsymbol{\lambda}^{(q+1)}\right)$, where

$$\dot{\mathbf{U}}\left(\boldsymbol{\beta}|\boldsymbol{\beta}^{(q)},\boldsymbol{\lambda}^{(q+1)}\right) = -\sum_{i=1}^{n}\sum_{k=1}^{K}\Delta_{ik}\hat{E}\left[W_{ik}|\boldsymbol{\beta}^{(q)},\boldsymbol{\lambda}^{(q+1)}\right]\left\{\frac{S^{(2)}(t_k,\boldsymbol{\beta})}{S^{(0)}(t_k,\boldsymbol{\beta})} - \bar{\mathbf{X}}(t_k;\boldsymbol{\beta})^{\otimes 2}\right\}.$$

For the implementation of the EM algorithm, it is apparent that one needs to choose a stopping or convergence criterion, and a natural one, which is used in the numerical studies below, is when the sum of the relative differences of two consecutive estimators of $\boldsymbol{\beta}$ is smaller than a small positive constant such as 10^{-4}. It is worth pointing out that the EM algorithm developed above has several important computational advantages. One is that it avoids the intractable integration in (4), which poses challenges under model (2) as discussed above. Another is that we only need to know the values of $\mathbf{X}_i(t)$ at $t_1,\ldots,t_K$ instead of the whole trajectories of covariates $\mathbf{X}_i(t)$ over time, which are rarely available in practice. More importantly, the quantities from the last iteration of EM algorithms can serve as a cornerstone for variance estimation in Section 4.

3. Asymptotic theory

In this section, we establish the consistency, rate of convergence, and asymptotic normality of the estimators. Let $\boldsymbol{\theta} = (\boldsymbol{\beta},\Lambda)$ and $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}},\hat{\Lambda})$ denote the estimators defined in the previous section with $\hat{\Lambda}(t) = \sum_{t_k<t}\hat{\lambda}_k$. Assume that the observed data $O = \{O_i\}_{i=1}^{n}$ are i.i.d. realizations of

$O = \left(\{\Delta N_{\cdot j}\}_{j=1}^{J}, \mathbf{X}, \{T_{\cdot j}\}_{j=1}^{J}\right)$, and let $\mathfrak{B}_p$ and $\mathfrak{B}$ denote the collection of Borel sets in $\mathbb{R}^p$ and $\mathbb{R}$, respectively. Let τ denote the maximum follow-up time for all subjects, $\mathfrak{B}_1[0,\tau] = \{B \cap [0,\tau] : B \in \mathfrak{B}\}$, $\mathfrak{B}_2[0,\tau] = \mathfrak{B}_1[0,\tau] \times \mathfrak{B}_1[0,\tau]$, and $\mathfrak{C}_p[0,\tau]$ denote the collection of Borel sets in $\mathbb{C}_p = \{f : [0,\tau] \to \mathbb{R}^p\}$. Furthermore, for B_1, B_2, $B_4 \in \mathfrak{B}_1[0,\tau]$ and $B_3 \in \mathfrak{C}_p[0,\tau]$, define $\nu_1(B_1 \times B_2 \times B_3) = \int_{B_3} \sum_{k=1}^{\infty} \Pr(J = k|\mathbf{X} = \mathbf{x}) \sum_{j=1}^{k} \Pr(T_{\cdot(j-1)} \in B_1, T_{\cdot j} \in B_2 | J = k, \mathbf{X} = \mathbf{x}) d\mu_{\mathbf{X}}(\mathbf{x})$, $\nu_2(B_4 \times B_3) = \int_{B_3} \sum_{k=1}^{\infty} \Pr(J = k|\mathbf{X} = \mathbf{x}) \sum_{j=1}^{k} \Pr(T_{\cdot j} \in B_4 | J = k, \mathbf{X} = \mathbf{x}) \, d\mu_{\mathbf{X}}(\mathbf{x})$, $\nu_3(B_1) = \int_{\mathcal{X}} \sum_{k=1}^{\infty} \Pr(J = k|\mathbf{X} = \mathbf{x}) \times \sum_{j=1}^{k} \Pr(T_{\cdot J} \in B_1 | J = k, \mathbf{X} = \mathbf{x}) \, d\mu_{\mathbf{X}}(\mathbf{x})$, where $\mu_{\mathbf{X}}(\mathbf{x})$ denotes the probability measure of $\mathbf{X}$, $T_{\cdot j}$ denotes the population version of T_{ij} and $\mathcal{X}$ is the sample space of $\mathbf{X}$. Also define $\mu_1(B_1 \times B_2) = \nu_1(B_1 \times B_2 \times \mathcal{X})$ and $\mu_2(B_4) = \nu_2(B_4 \times \mathcal{X})$.

For the consistency and the rate of convergence, we will consider them under the L_2 metric $d(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \sqrt{\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2^2 + \|\Delta\Lambda_1 - \Delta\Lambda_2\|_{L_2(\mu_1)}^2}$, where $\Delta\Lambda(T_{\cdot(j-1)}, T_{\cdot j}) = \Lambda(T_{\cdot j}) - \Lambda(T_{\cdot(j-1)})$. Define $\mathcal{F} = \{\Lambda : \Lambda$ is a nondecreasing function and $\Lambda(0) = 0\}$, and let $\lesssim$ ($\gtrsim$) represent the left (right) side bounded by the right (left) side multiplied by a constant. For the asymptotic properties of $\hat{\boldsymbol{\theta}}$, we need the following regularity conditions.

- (C1) The true value of $\boldsymbol{\beta}$, denoted by $\boldsymbol{\beta}_0$, lies in the interior of the compact set $\mathcal{B} \subset \mathbb{R}^p$. The true value of $\Lambda \in \mathcal{F}$, denoted by Λ_0, is continuously differentiable with positive and bounded first-order derivative Λ_0' in some interval $[\tau_0, \tau]$ where $\Pr\left(\cap_{j=1}^{J} \{T_{\cdot j} \in [\tau_0, \tau]\}\right) = 1$ and $\Lambda_0(\tau_0) > 0$.
- (C2) With probability one, the covariate $\mathbf{X}(t)$ of bounded variation over $[\tau_0, \tau]$, and its left limit exists for any t.
- (C3) The measure $\mu_l \times \mu_{\mathbf{X}}$ is absolutely continuous with respect to ν_l for $l = 1, 2$. Furthermore, $\nu_3(\{\tau\}) > 0$.
- (C4) If there exists a vector $\boldsymbol{\beta}$ such that $\boldsymbol{\beta}^\top \mathbf{X}(t) = a(t)$ for some deterministic function $a(t)$ on $t \in [\tau_0, \tau]$ almost surely (a.s.) with respect to ν_2, then $\boldsymbol{\beta} = 0$ and $a(t) = 0$ a.s. with respect to μ_2.
- (C5) The observation times are s_0-separated. That is, there exists a constant $s_0 > 0$, such that $\Pr(T_{\cdot j} - T_{\cdot(j-1)} \geq s_0 | J, \mathbf{X}) = 1$ for all $j = 1, \ldots J$.
- (C6) The number of observation times, J, is positive and finite with probability one.
- (C7) The function $M_0(O) = \sum_{j=1}^{J} \Delta N_{\cdot j} \log\{\Delta N_{\cdot j}\}$ satisfies $E[M_0(O)] < 0$.
- (C8) For some $c_0 \in (0, \infty)$, the function $\mathbf{X} \to E[\exp(c_0 N(\tau)) | \mathbf{X}(s), 0 \leq s \leq \tau]$ is uniformly bounded for $\mathbf{X} \in \mathcal{X}$.
- (C9) The measure μ_2 is absolutely continuous with respect to Lebesgue measure with a derivative $\dot{\mu}_2$ that $\dot{\mu}_2 \geq c_1 > 0$ for some positive constant c_1.
- (C10) For some $\rho \in (0, 1)$, $a^\top \mathrm{Cov}(\mathbf{X}(t), \mathbf{X}(s) | U_1, U_2) a \geq \rho a^\top E\left[\mathbf{X}(t)\mathbf{X}^\top(s) | U_1, U_2\right] a$ a.s. on ν_1 for all $a \in \mathbb{R}^p$ and $t, s \in [\tau_0, \tau]$ where $(U_1, U_2, \mathbf{X})$ follows a distribution with density $\nu_1 / \nu_1(\mathbb{R}^+ \times \mathbb{R}^+ \times \mathcal{X})$.

Conditions (C1)-(C2) are common assumptions for the true parameter $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0, \Lambda_0)$ and semi-parametric models with time-varying covariates, and in particular, Condition (C2) is needed for boundedness and square integrability of the likelihood function. The first part of Condition (C3) together with (C4) and $\Pr(\cap_{j=1}^{J} \{T_{\cdot j} \in [\tau_0, \tau]\}) = 1$ in (C1) ensures the identifiability of the parameters, while the second part is a technical condition mainly used for showing the boundedness of $\hat{\Lambda}(t)$ on $[0, \tau]$. Condition (C5) means that the two consecutive observation times should have at least s_0 difference, and Condition (C6) requires that the observation number is bounded. Condition (C7) holds when Condition (C6) is true and the second moment of $\Delta N_{\cdot j}$ is finite. Conditions (C5)-(C7) are reasonable in practice, especially when the observation times are discrete time points and the variance of $\Delta N_{\cdot j}$ is bounded.

Conditions (C8) and (C10) generalize the conditions C10 and C14 given in Wellner and Zhang [43] to accommodate time-varying covariates and could be justified similarly. Condition (C9) is the same as the second part of the condition C12 given in Wellner and Zhang [43], implying that the density of the observation times is strictly positive. As shown below, Conditions (C1)-(C7) are for the consistency of the proposed estimators, and Conditions (C8)-(C10) are additional assumptions needed for establishing the convergence rate and asymptotic normality. In the following, we first establish the strong consistency and convergence rate of $\hat{\boldsymbol{\theta}}$ and then the asymptotic normality of $\hat{\boldsymbol{\beta}}$ with their proofs relegated to Section 8.

Theorem 3.1 (Consistency). *Suppose that Conditions (C1)-(C7) and model (2) hold. Then we have that* $d(\hat{\boldsymbol{\theta}},\boldsymbol{\theta}_0) \to 0$ *a.s. as* $n \to \infty$.

Theorem 3.2 (Rate of convergence). *Suppose that Conditions (C1)-(C10) and model (2) hold. Then we have that* $n^{1/3} d(\hat{\boldsymbol{\theta}},\boldsymbol{\theta}_0) = O_p(1)$.

Theorem 3.3 (Asymptotic normality). *Under the same conditions of Theorem 3.2, we have that* $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \to_d N\left(0, \boldsymbol{\Omega}^{-1}\boldsymbol{\Sigma}(\boldsymbol{\Omega}^{-1})^\top\right)$, *where*

$$\boldsymbol{\Omega} = -E\left[\sum_{j=1}^{J} \frac{\left\{\int_{T_{\cdot(j-1)}}^{T_{\cdot j}} \exp\left(\boldsymbol{\beta}_0^\top \mathbf{X}(t)\right)\left(\mathbf{X}(t) - \bar{\mathbf{x}}(t)\right) d\Lambda_0(t)\right\}^{\otimes 2}}{\int_{T_{\cdot(j-1)}}^{T_{\cdot j}} \exp\left(\boldsymbol{\beta}_0^\top \mathbf{X}(t)\right) d\Lambda_0(t)}\right],$$

and

$$\boldsymbol{\Sigma} = E\left[\left\{\sum_{j=1}^{J}\left(\frac{\Delta N_{\cdot j}}{\int_{T_{\cdot(j-1)}}^{T_{\cdot j}} \exp\left(\boldsymbol{\beta}_0^\top \mathbf{X}(t)\right) d\Lambda_0(t)} - 1\right) \times \int_{T_{\cdot(j-1)}}^{T_{\cdot j}} \exp\left(\boldsymbol{\beta}_0^\top \mathbf{X}(t)\right)\left(\mathbf{X}(t) - \bar{\mathbf{x}}(t)\right) d\Lambda_0(t)\right\}^{\otimes 2}\right].$$

Here, $\bar{\mathbf{x}}(t) \in \mathbb{R}^p$ *is some function on* $[0,\tau]$ *implicitly defined in Section 8. Moreover, if the Poisson assumption holds,* $\hat{\boldsymbol{\beta}}$ *is asymptotically efficient.*

We remark that both the strong consistency and convergence rate given in Theorems 3.1-3.2 and the asymptotic normality given in Theorem 3.3 are true as long as model (2) holds and do not require the Poisson assumption. In other words, the proposed estimator is robust. The convergence rate $O(n^{1/3})$ is the typical optimal rate for semiparametric models [42,43,46,48]. Though it is slower than $n^{1/2}$, the common order of the convergence rate from parametric MLEs, the asymptotic normality of $\hat{\boldsymbol{\beta}}$ still follows by the convergence rate $\sqrt{n}$. In Theorem 3.3, the form of $\boldsymbol{\Sigma}$ involves $\Delta N_{\cdot j}$ because the explicit calculation of the expectation of $\boldsymbol{\Sigma}$ requires the formulation of $\operatorname{cov}(N(t), N(s) | \mathbf{X}(s))$ for $t \le s \in [0,\tau]$, which is unspecific in model (2). If the Poisson assumption holds, $\boldsymbol{\Sigma}$ would reduce to $-\boldsymbol{\Omega}$, and both $\boldsymbol{\Sigma}$ and $-\boldsymbol{\Omega}$ would be exactly the efficient information matrix of $\hat{\boldsymbol{\beta}}$, that is, $\hat{\boldsymbol{\beta}}$ would be asymptotically efficient [29,43].

It is apparent that for inference about $\boldsymbol{\beta}$, one needs to estimate the asymptotic covariance of $\hat{\boldsymbol{\beta}}$. Based on Theorem 3.3, it would be natural and typical to estimate $\boldsymbol{\Omega}$ and $\boldsymbol{\Sigma}$ by their empirical counterparts. However, since $\bar{\mathbf{x}}(t)$ does not have an explicit form, no such empirical estimators would be easily available. The profile likelihood approach would provide another choice but it would be sensitive to model specification. Furthermore, its implementation would be computationally intensive as the lower-triangular elements in the asymptotic variance matrix are estimated one at a time. To address this, in the next section, we will present two consistent estimators of $\mathbf{\Omega}$ and $\mathbf{\Sigma}$ that are free of the Poisson assumption. The proposed estimators will reuse the quantities from the last iteration of the EM algorithm and thus can significantly reduce the computational burden.

4. Two covariance estimators of $\hat{\boldsymbol{\beta}}$

We first briefly introduce the profile likelihood method and discuss what we hope to achieve in the variance estimation of $\hat{\boldsymbol{\beta}}$. Murphy, Rossini and van der Vaart [28] and Murphy and van der Vaart [29] proposed the profile likelihood method based on the numerical derivatives of the profile log-likelihood function, which has been widely utilized in, for example, Su and Wang [31], Zeng, Mao and Lin [46] and Zeng and Lin [45] with some modification. The adaptation in Zeng and Lin [45] estimates the covariance matrix of $\hat{\boldsymbol{\beta}}$ by $\hat{\mathbf{V}}^{-1}/n$, where $\hat{\mathbf{V}} = \frac{1}{n}\sum_{i=1}^{n}\left[\{\partial \mathrm{p}\ell_i(\boldsymbol{\beta}|O)/\partial\boldsymbol{\beta}|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}\}^{\otimes 2}\right]$. Here, $\mathrm{p}\ell_i(\boldsymbol{\beta}|O)$ is the profile log-likelihood function of $\boldsymbol{\beta}$ by the ith subject defined as $\ell_i(\boldsymbol{\beta}, \hat{\Lambda}_{\boldsymbol{\beta}}|O)$ where $\hat{\Lambda}_{\boldsymbol{\beta}} = \arg\max_{\Lambda\in\mathcal{F}^*}\log L_n(\boldsymbol{\beta},\Lambda)$ with $\mathcal{F}^*$ being the set of step functions with nonnegative jumps at t_k. Given $\boldsymbol{\beta}$, $\hat{\Lambda}_{\boldsymbol{\beta}}$ can be obtained by the proposed EM algorithm through iterating (7) only. We can take $\hat{\mathbf{V}}$ as the empirical covariance matrix of the gradient of $\mathrm{p}\ell_i(\boldsymbol{\beta}|O)$ with respective $\boldsymbol{\beta}$ evaluated at $\hat{\boldsymbol{\beta}}$. The lth component of the gradient is approximated by the first-order numerical derivative of $\mathrm{p}\ell_i(\boldsymbol{\beta}|O)$ as $\{\mathrm{p}\ell_i(\hat{\boldsymbol{\beta}} + h_n e_l|O) - \mathrm{p}\ell_i(\hat{\boldsymbol{\beta}}|O)\}/h_n$, where $h_n = O(n^{-1/2})$, and e_l is a p-dimensional vector with the lth element being one and all others being zero. This profile likelihood method tends to be computationally inefficient as it essentially calculates a gradient in an element-wise fashion for each subject. With large p and sample size, the computational burden could be heavy. Another drawback is the arbitrary choice of h_n. Moreover, by Corollary 3 in Murphy and van der Vaart [29], $\hat{\mathbf{V}}$ consistently estimates the asymptotic variance of $\hat{\boldsymbol{\beta}}$ only when the Poisson model holds as (4) is constructed from the Poisson process. To address these limitations, we put forward two new variance estimators based on Theorem 3.3.

The first is a robust sandwich-type variance estimator $\hat{\mathbf{\Omega}}^{-1}\hat{\mathbf{\Sigma}}(\hat{\mathbf{\Omega}}^{-1})^\top/n$. Here,

$$\hat{\mathbf{\Omega}} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J}\frac{\left\{\sum_{T_{i(j-1)}<t_k\le T_{ij}}\exp\left(\hat{\boldsymbol{\beta}}^\top\mathbf{X}_i(t_k)\right)\hat{\lambda}_k\left(\mathbf{X}_i(t_k)-\hat{\bar{\mathbf{x}}}_k\right)\right\}^{\otimes 2}}{\sum_{T_{i(j-1)}<t_k\le T_{ij}}\exp\left(\hat{\boldsymbol{\beta}}^\top\mathbf{X}_i(t_k)\right)\hat{\lambda}_k},$$

and

$$\hat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{k=1}^{K}\Delta_{ik}\left(\hat{E}\left[W_{ik}|\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\lambda}}\right] - \exp\left(\hat{\boldsymbol{\beta}}^\top\mathbf{X}_i(t_k)\right)\hat{\lambda}_k\right)\left(\mathbf{X}_i(t_k)-\hat{\bar{\mathbf{x}}}_k\right)\right\}^{\otimes 2}.$$

Here, $\hat{\bar{\mathbf{x}}}_k = -\,d\log\left(\hat{\lambda}_k(\boldsymbol{\beta})\right)/d\boldsymbol{\beta}\big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}$, where $\hat{\boldsymbol{\lambda}}(\boldsymbol{\beta}) = (\hat{\lambda}_1(\boldsymbol{\beta}),\ldots,\hat{\lambda}_K(\boldsymbol{\beta}))$ such that $\hat{\Lambda}_{\boldsymbol{\beta}}(t) = \sum_{t_k<t}\hat{\lambda}_k(\boldsymbol{\beta})$. As mentioned above, we can regard $\hat{\mathbf{\Omega}}$ and $\hat{\mathbf{\Sigma}}$ as the respective empirical counterparts and consistent estimators of $\mathbf{\Omega}$ and $\mathbf{\Sigma}$, by rewriting the integration with respect to $\hat{\boldsymbol{\lambda}}$ and by replacing $\bar{\mathbf{x}}(t_k)$ with its consistent estimator $\hat{\bar{\mathbf{x}}}_k$. This variance estimator is robust to the Poisson assumption as a direct result of Theorem 3.3. We can efficiently compute lth element of $\hat{\bar{\mathbf{x}}}_k$ by $\log(\hat{\lambda}_k(\hat{\boldsymbol{\beta}} + a_n e_l)) - \log(\hat{\lambda}_k(\hat{\boldsymbol{\beta}}))\}/a_n$ by iterating (7) as we compute $\hat{\mathbf{V}}$, where a_n can be a small positive number depending on n (e.g., $10^{-5}n^{-1/2}$). To prevent arithmetic overflow when $\hat{\lambda}_k(\hat{\boldsymbol{\beta}})$ is very close or equal to zero, we set $\hat{\bar{\mathbf{x}}}_k$ to

zero in those cases. This robust method has a much lighter computational burden than the profile likelihood method. The terms $\hat{E}\left[W_{ik}|\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\lambda}}\right]$ and $\exp\left(\hat{\boldsymbol{\beta}}^{\top}\mathbf{X}_i\left(t_k\right)\right)\hat{\lambda}_k$ are immediately available to calculate $\hat{\boldsymbol{\Omega}}$ and $\hat{\boldsymbol{\Sigma}}$ after finishing the EM algorithm. By contrast, $\hat{\mathbf{V}}$ requires computing the profile log-likelihood function of all subjects for p different $\boldsymbol{\beta}$.

For the second estimator, we first note that when the Poisson assumption holds, $\boldsymbol{\Omega}^{-1}\boldsymbol{\Sigma}(\boldsymbol{\Omega}^{\top})^{-1}$ of the first variance estimator reduces to $\boldsymbol{\Sigma}^{-1}$, which coincides with the inverse of the efficient information matrix of $\boldsymbol{\beta}$. Then we can simply estimate the asymptotic variance of $\hat{\boldsymbol{\beta}}$ by $\hat{\boldsymbol{\Sigma}}^{-1}/n$, which is further simplified computationally. This efficient information-based estimator is closely connected to the profile likelihood method: when the Poisson model holds, both $\hat{\boldsymbol{\Sigma}}$ and $\hat{\mathbf{V}}$ consistently estimate $\boldsymbol{\Sigma}$, the variance of the efficient score function of $\hat{\boldsymbol{\beta}}$ [29]. In other words, $\hat{\boldsymbol{\Sigma}}$ and $\hat{\mathbf{V}}$ are asymptotically equal, as confirmed by the simulation results in Section 5. Therefore, the efficient information-based estimator $\hat{\boldsymbol{\Sigma}}^{-1}/n$ provides a computationally efficient alternative to the profile likelihood estimator when the Poisson assumption could be justified.

We note that computational burden and complexity can be further reduced when (E1) holds.

(E1) With probability one, $\mathbf{X}(t)$ is a piecewise-constant function of t such that $\mathbf{X}(t)$ is constant for t within each internal $(T_{\cdot(j-1)}, T_{\cdot j}]$, $j = 1, \ldots, J$.

Given (E1), we can directly estimate $\bar{\mathbf{x}}(t)$ directly from $\bar{\mathbf{X}}(t;\hat{\boldsymbol{\beta}})$, which is readily available after the EM algorithm as part of $\mathbf{U}(\hat{\boldsymbol{\beta}}|\hat{\boldsymbol{\theta}})$. Thus, no additional computation and numerical differentiation are necessary. In this case, $\hat{\boldsymbol{\Sigma}}$ can be interpreted as the empirical variance of the contribution to $\mathbf{U}(\hat{\boldsymbol{\beta}}|\hat{\boldsymbol{\theta}})$ from each subject. Condition (E1) is not uncommon in real-world studies, such as YWP. The time-varying covariates in YWP are considered constant for each subject within every observation interval (a quarter) because these covariates are collected as summary measures of the sexual behaviors of the subject since the last visit. Nonetheless, these covariates could still vary across different observation intervals. In addition, we note that $\bar{\mathbf{X}}$ resembles $\bar{Z}$ in Lin et al. [24] and thus the $\bar{\mathbf{X}}$-based robust estimator is a generalized sandwich-type variance estimator, as proposed by Lin et al. [24], to panel count data. Importantly, even if Condition (E1) does not hold, the variance estimators based on $\bar{\mathbf{X}}$ could still serve as a computational-efficient approximation to the original variance estimators. In Section 5, we shall demonstrate through simulation that the variance estimators using $\bar{\mathbf{X}}$ closely resemble the original variance estimators even when (E1) is not satisfied.

Next, we prove the consistency of $\hat{\bar{\mathbf{x}}}(t) = \sum_{k=1}^{K} I(t_{k-1} < t \leq t_k)\hat{\bar{\mathbf{x}}}_k$. Lemma 1 lays out the sufficient conditions for establishing the consistency of the proposed variance estimators. We first introduce additional notation for Lemma 1. Let $\mathcal{H}_3 = \{h_3(t) \in \mathbf{R}^p, t \in [0,\tau]$: each element of $h_3(t)$ is of bounded variation over $[0,\tau]\}$. We prove the consistency of $\hat{\bar{\mathbf{x}}}(t)$ under the metric for $\mathcal{H}_3$ $\tilde{d}(h_{31},h_{32}) = \{\int \| \int_{u_2}^{u_1}(h_{31}(t) - h_{32}(t))d\Lambda_0(t)\|_2^2 d\nu_1(u_1,u_2,\mathbf{x})\}^{1/2}$ for $h_{31}, h_{32} \in \mathcal{H}_3$.

Lemma 1. *If conditions (C1)-(C10) and model (2) hold,* $\tilde{d}\left(\hat{\bar{\mathbf{x}}},\bar{\mathbf{x}}\right) \to 0$ *in probability as* $n \to \infty$. *Furthermore, if Condition (E1) additionally holds,* $\bar{\mathbf{x}}(t)$ *reduce to* $s^{(1)}(t,\boldsymbol{\beta}_0)/s^{(0)}(t,\boldsymbol{\beta}_0)$ *and* $\tilde{d}\left(\bar{\mathbf{X}}(\cdot;\hat{\boldsymbol{\beta}}),\bar{\mathbf{x}}\right) \to 0$, *where* $s^{(u)}(t,\boldsymbol{\beta}) = E\left[\Delta(t)\mathbf{X}(t)^{\otimes u}\exp(\boldsymbol{\beta}^{\top}\mathbf{X}(t))|T_{\cdot(j-1)},T_{\cdot j},J\right]$ *for* $u = 0,1,2$.

Lemma 1 establishes the connection between the EM algorithm and the asymptotic normality in Theorem 3.3 since $\hat{\bar{\mathbf{x}}}$ is based on $\hat{\boldsymbol{\lambda}}$ in EM algorithm. Given Lemma 1, we justify the aforementioned consistency of $\hat{\boldsymbol{\Omega}}$, $\hat{\boldsymbol{\Sigma}}$ and $\hat{\mathbf{V}}$ in the next theorem.

Theorem 4.1. *If conditions (C1)-(C10) and model (2) hold, then we have that* $\hat{\boldsymbol{\Omega}} \to_p \boldsymbol{\Omega}$, *and both* $\hat{\mathbf{V}}$ *and* $\hat{\boldsymbol{\Sigma}}$ *converge to* $\boldsymbol{\Sigma}$ *in probability as* $n \to \infty$.

It is important to note that the conclusions of Lemma 1 and Theorem 4.1 are not contingent on the Poisson assumption. Thus, $\hat{\mathbf{V}}$ and $\hat{\boldsymbol{\Sigma}}$ are both consistent estimators for the variance of the efficient score function and hence asymptotically equivalent regardless of whether data come from a Poisson process. The profile likelihood estimator, on the other hand, may fail to yield accurate variance estimates for $\hat{\boldsymbol{\beta}}$ when $\boldsymbol{\Sigma}$ is not the asymptotic variance of $\hat{\boldsymbol{\beta}}$ under a non-Poisson model. We elaborate the derivation of the variance estimation methods and prove Theorem 4.1 in Section 8.4.

5. A simulation study

An extensive simulation study was conducted to evaluate the finite sample performance of the estimation approach proposed in the previous sections. Herein, we considered the situation with two covariates $X_1(t)$ and X_2. The time-varying covariate $X_1(t)$ was generated through $X_1(t) = \sum_{j=1}^{10} X_{1j} I(U_{(j-1)} < t \leq U_{(j)})$. In the above, $\{X_{1j}\}_{j=1}^{10}$ were generated as independently and identically distributed (i.i.d.) random variables following the uniform distribution Unif $(0,1)$. We then generated $\{U_{(j)}\}_{j=1}^{9}$ as an ordered i.i.d. sample from Unif $(0,\tau)$, $U_{(0)} = 0$ and $U_{(10)} = \tau$, where we set $\tau = 2$ to represent the maximum follow-up time. The covariate X_2 was assumed to be a constant and generated from Unif $(0,1)$.

For the generation of panel count data, we considered two scenarios. One is that the underlying recurrent event process $N_i(t)$ follows a non-homogeneous Poisson process with the rate function $\lambda(t)\exp(\beta_1 X_1(t) + \beta_2 X_2)$. The other is to assume that $N_i(t)$ is a mixed non-homogeneous Poisson process with the rate function $r\,\lambda(t)\exp(\beta_1 X_1(t) + \beta_2 X_2)$, where r follows the Gamma distribution with both mean and and variance being 1. In the following, we set $\beta_1 = 1$, $\beta_2 = -1$ and $\lambda(t) = 16/(1+t)$ or $4(\sin(4\pi t) + 2)$. For each subject, the total number of observations J was generated from the zero-truncated Poisson distribution with mean 5, and the observation times were taken to be the order statistics of a random sample of size J from the uniform distribution over $\left(0, \bar{T}\right)$ with $\bar{T} \sim Unif(0.9\tau, \tau)$. The results given below are based on the sample size $n = 100$, 200 or 400 with 1000 replications.

Tables 1 and 2 present the results on estimation of the regression parameter $\boldsymbol{\beta}$ given by the proposed estimation approach with $\lambda(t) = 16/(1+t)$ and $4(\sin(4\pi t) + 2)$, respectively. In the tables, we calculated the average of the biases of the proposed estimates (Bias), the empirical standard error of the estimates (SE), the average of the proposed standard error estimates (SEE), and the 95% empirical coverage probability (CP). For the variance estimation, we considered two methods described in Section 4, the robust estimation (Robust) and the inverse of the estimated efficient information matrix (Info) with $a_n = 10^{-5} n^{-1/2}$. For comparison, we included the profile likelihood estimation (Profile), using $h_n = 5n^{-1/2}$ as recommended by Zeng, Mao and Lin [46]. In addition, we also computed Robust and Info with $\bar{\mathbf{x}}_k$ being estimated by $\bar{\mathbf{X}}(t_k; \hat{\boldsymbol{\beta}})$, whose SEE and CP are shown in parentheses in Tables 1 and 2. Since the setting of $X_1(t)$ does not satisfy Condition (E1), we assessed how well the $\bar{\mathbf{X}}$-based variance estimators can approximate the original Robust and Info variance estimators.

The results suggest that the proposed estimator seems to be unbiased and the robust variance estimation procedure appears to perform well, both independent of the Poisson assumption. The normal approximation to the asymptotic distribution of the proposed estimator seems to be appropriate too, and as expected, the results became better when the sample size increased.

It can be seen from the tables that under the Poisson assumption, all three variance estimation procedures yielded reasonable results and the coverage probabilities close to the 95% nominal level. However, when the Poisson assumption was violated, both the Profile and Info methods severely underestimated the standard errors, leading to drastically reduced coverage probabilities. In contrast, the Robust method was still able to provide reasonable estimates and maintain relatively satisfactory coverage probabilities, especially when the sample size was large. In other words, the proposed robust variance

Table 1. Simulation results on the estimation of regression parameters with $\lambda(t) = 16/(1+t)$.

	Poisson				Robust ($\bar{\mathbf{X}}$-based)		Info ($\bar{\mathbf{X}}$-based)		Profile	
n	Model		Bias	SE	SEE	CP	SEE	CP	SEE	CP
100	Yes	$\hat{\beta}_1$	<0.001	0.092	0.088(0.088)	92.9(92.8)	0.094(0.094)	94.9(94.9)	0.094	94.5
		$\hat{\beta}_2$	0.001	0.082	0.082(0.082)	94.9(94.9)	0.087(0.087)	96.1(96.1)	0.087	95.9
	No	$\hat{\beta}_1$	-0.037	0.315	0.300(0.299)	93.3(93.1)	0.031(0.031)	15.7(15.7)	0.031	15.7
		$\hat{\beta}_2$	-0.008	0.376	0.371(0.371)	93.7(93.6)	0.021(0.021)	7.8(7.8)	0.021	7.8
200	Yes	$\hat{\beta}_1$	-0.001	0.065	0.063(0.063)	94.1(94.1)	0.065(0.065)	95.3(95.2)	0.065	95.2
		$\hat{\beta}_2$	<0.001	0.059	0.058(0.058)	94.1(94.1)	0.060(0.060)	94.7(94.7)	0.060	94.5
	No	$\hat{\beta}_1$	-0.017	0.221	0.211(0.211)	92.5(92.5)	0.021(0.021)	14.8(14.8)	0.021	14.9
		$\hat{\beta}_2$	-0.003	0.271	0.263(0.263)	92.5(92.5)	0.014(0.014)	8.4(8.4)	0.014	8.4
400	Yes	$\hat{\beta}_1$	0.002	0.045	0.045(0.045)	94.0(94.0)	0.045(0.045)	94.4(94.4)	0.045	94.5
		$\hat{\beta}_2$	0.001	0.041	0.041(0.041)	94.7(94.7)	0.042(0.042)	94.8(94.8)	0.042	94.8
	No	$\hat{\beta}_1$	-0.009	0.150	0.150(0.150)	93.6(93.5)	0.014(0.014)	15.8(15.8)	0.014	15.7
		$\hat{\beta}_2$	-0.002	0.183	0.187(0.187)	94.1(94.1)	0.010(0.010)	7.9(7.9)	0.010	7.9

estimator seems to be robust to the Poisson assumption. Another observation from the tables is that the Profile and Info methods gave almost identical results, actually, further examination of the raw simulation results indicated that they were nearly the same for each replication. This suggests that the Info method could be used as a computationally efficient substitute when the Poisson assumption holds.

One can further observe from the table that both SEE and CP of the original Robust and Info using numerical differentiation and those based on $\bar{\mathbf{X}}$ are almost identical for all cases, even if $X_1 t$ fluctuates on $[0,\tau]$ as a random piece-wise function of 10 pieces. In the supplement materials [32], we provide more simulation results with various settings where Condition (E1) does not hold. The results from these settings all show that the $\bar{\mathbf{X}}$-based Robust and Info provide a good approximation to their original counterparts.

Then we assess the computational efficiency of Robust and Robust based on $\bar{\mathbf{X}}$ compared with Profile. We present the ratio of computational times of Robust and $\bar{\mathbf{X}}$-based Robust to those of Profile under

Table 2. Simulation results on the estimation of regression parameters with $\lambda(t) = 4(\sin(4\pi t) + 2)$.

	Poisson				Robust ($\bar{\mathbf{X}}$-based)		Info ($\bar{\mathbf{X}}$-based)		Profile	
n	Model		Bias	SE	SEE	CP	SEE	CP	SEE	CP
100	Yes	$\hat{\beta}_1$	-0.001	0.096	0.093(0.093)	93.7(93.7)	0.099(0.099)	94.7(94.7)	0.098	94.7
		$\hat{\beta}_2$	0.003	0.087	0.086(0.086)	93.7(93.7)	0.092(0.092)	95.5(95.5)	0.091	95.3
	No	$\hat{\beta}_1$	0.003	0.299	0.291(0.290)	93.3(93.3)	0.034(0.034)	18.7(18.7)	0.034	18.5
		$\hat{\beta}_2$	0.010	0.384	0.368(0.367)	92.2(92.2)	0.023(0.023)	8.6(8.6)	0.023	8.6
200	Yes	$\hat{\beta}_1$	0.001	0.068	0.066(0.066)	94.4(94.4)	0.068(0.068)	94.7(94.7)	0.068	94.8
		$\hat{\beta}_2$	0.003	0.062	0.061(0.061)	94.0(94.0)	0.063(0.063)	95.4(95.4)	0.063	95.5
	No	$\hat{\beta}_1$	-0.001	0.207	0.209(0.209)	94.9(94.9)	0.023(0.023)	15.6(15.5)	0.023	15.6
		$\hat{\beta}_2$	0.008	0.267	0.263(0.263)	92.7(92.7)	0.015(0.015)	8.1(8.1)	0.015	8.2
400	Yes	$\hat{\beta}_1$	<0.001	0.047	0.047(0.047)	95.4(95.4)	0.048(0.048)	95.6(95.7)	0.048	95.6
		$\hat{\beta}_2$	0.001	0.044	0.043(0.043)	95.1(95.1)	0.044(0.044)	95.5(95.5)	0.044	95.5
	No	$\hat{\beta}_1$	<0.001	0.145	0.147(0.147)	95.0(95.0)	0.016(0.016)	14.4(14.4)	0.016	14.4
		$\hat{\beta}_2$	<0.001	0.189	0.186(0.186)	94.5(94.5)	0.010(0.010)	7.3(7.3)	0.011	7.3

different sample sizes in the supplement materials [32]. The results illustrate that, compared with Profile, Robust is around 2 to 4 times faster, and $\bar{\mathbf{X}}$-based Robust could be up to 100 times faster. In other words, the proposed robust variance estimators not only enjoy better estimation accuracy but also are much more computationally efficient than Profile. Furthermore, the computational efficiency gain ratios of Robust and $\bar{\mathbf{X}}$-based ones decrease when the sample size increases. However, the computational efficiency gain ratios of Robust decrease slowly, and even with a very large sample size, e.g., 1600, the computational efficiency of Robust roughly doubles that of Profile. The computational efficiency ratio of $\bar{\mathbf{X}}$-based Robust increases when p, the number of parameters, becomes larger, but that of Robust does not seem to vary with p. This suggests that $\bar{\mathbf{X}}$-based Robust could be even more efficient when the number of covariates is large.

To assess the performance of the proposed estimation procedure on the baseline cumulative rate function $\Lambda(t)$, we plotted the averages of the proposed estimates $\hat{\Lambda}(t)$ corresponding to $\lambda(t) = 16/(1+t)$ and $\lambda(t) = 4(\sin(4\pi t) + 2)$ in Figure 1, respectively. In addition, the true curves of $\Lambda(t)$ were given too along with the confidence bands given by the average of the estimates plus and minus the pointwise empirical standard errors. They suggest that the proposed estimator $\hat{\Lambda}(t)$ seems to be unbiased. As expected, the estimator became more accurate when the sample size increased or under the Poisson assumption.

6. An application

We applied the proposed estimation procedure to the Young Women's Project (YWP) data mentioned in Section 1. The study examined the infections with *Chlamydia trachomatis* (CT), *Neisseria gonorrhoeae* (NG), and *Trichomonas vaginalis* (TV), which cause clinical diseases of chlamydia, gonorrhea, and trichomoniasis, respectively. The study was conducted between 1999 and 2009, and it consisted of young women aged 14 to 17 years. Basic demographic and clinical information of the study participants was collected at enrollment. Sexual behaviors and infection were assessed at all quarterly visits and thus

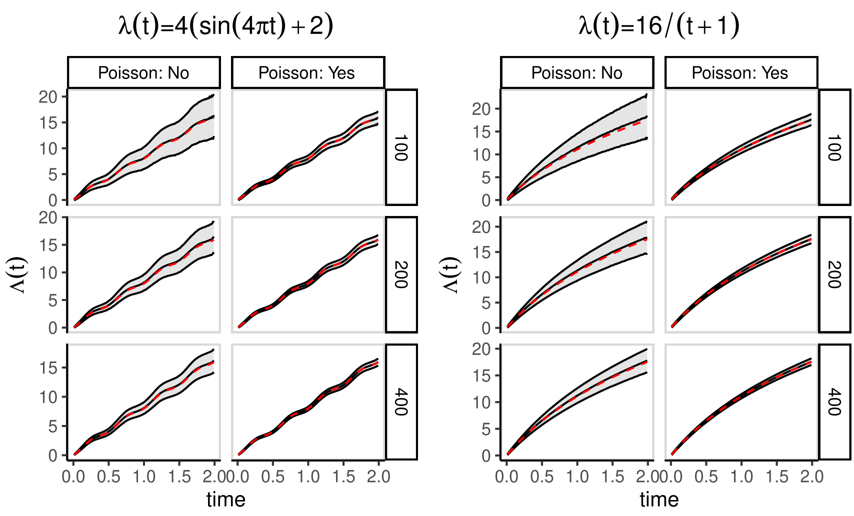

Figure 1. Estimates of the baseline cumulative rate function with $\lambda(t) = 16/(1+t)$ and $\lambda(t) = 4(\sin(4\pi t) + 2)$: black solid lines — the average of the estimates; red dashed lines — the true $\Lambda(t)$; the upper and lower boundaries of the gray ribbons — the mean plus and minus one pointwise empirical standard error of $\hat{\Lambda}(t)$, respectively.

were time-varying. At these visits, participants underwent cervical and vaginal examinations; samples were collected for testing of STI with CT, NG, and TV. The infected subjects were treated on the spot, so subsequent STI events represented new infections. The exact infection times were interval-censored, thus giving rise to panel count data on infection. Herein, we analyzed data from 271 participants that had completed at least one follow-up visit. We focused on the effects of the seven social, familial, and behavioral risk factors on each of the three types of STIs. The seven risk factors were

1. X_1: age at enrollment,
2. $X_2(t)$: 1 when having zero unprotected sexes in the previous quarter; 0 otherwise,
3. $X_3(t)$: 1 when having five or more times' unprotected sex in the previous quarter; 0 otherwise,
4. $X_4(t)$: 1 when having multiple sex partners in the previous quarter; 0 otherwise,
5. X_5: 1 when living with parents at the time of enrollment; 0 otherwise,
6. X_6: 1 when having used marijuana in the past three months prior to enrollment; 0 otherwise, and
7. X_7: 1 when having received sex education from parents prior to enrollment; 0 otherwise.

Among the seven factors, X_2, X_3 and X_4 were considered time-varying, and the rest time-invariant.

Table 3 presents the analytical results produced by the proposed method, with the application of three variance estimation procedures. As stated earlier, all the time-varying factors satisfy Condition (E1) because the summaries of sexual behaviors since the last visit were constant between two consecutive visits. In addition, the $\bar{\mathbf{X}}$-based variance estimators were identical to the original variance estimators up to 6 digits, further strengthening the conclusion of Lemma 1. Hence, we only presented the results from the $\bar{\mathbf{X}}$-based variance estimators in Table 3. For comparison, we also applied the estimating equation-based method proposed in [15] under the proportional mean model (3). The results based on the robust variance estimation suggested that having more than one sexual partner was associated with significantly increased risk for all three types of STIs. The finding is consistent with the existing medical literature [4]. The analysis also indicated that having no unprotected sex was significantly associated with the decreased risk of the infection with NG or TV but did not have a significant effect on the CT infection. On the other hand, having five or more times unprotected sex in the previous quarter had no significant association with the infection risk of all three organisms. Among the time-invariant factors, living with parents was associated with reduced CT and TV infection risks, while drug use and the received sex education were associated with increased CT risk but not NG and TV risks.

As seen in the simulation study, the profile likelihood and efficient information variance methods gave nearly identical results, which are generally similar to those given by the robust variance method, except that the two former methods also indicate a negative association between age and CT acquisition, again consistent with the current scientific understanding [23,35]. A possible explanation for this is that the immaturity of the host immune response makes younger women more susceptible to CT infection [4]. Another observation is that having five or more times' unprotected sex in the previous quarter tended to attenuate the risk for TV. Such findings should be viewed with caution since the two methods could underestimate the variance, as seen in the simulation study. In comparison, one can also see from Table 3 that the proportional mean model method failed to detect the influence of two factors, multiple partners and zero unprotected sex. We also obtained the estimated cumulative baseline rate functions corresponding to each and plotted them in the supplement materials [32].

7. Discussion and concluding remarks

This paper presents a regression model for panel count data, considering time-varying covariates, and proposes a robust likelihood-based estimation procedure within the proportional rate modeling framework. While the alternative proportional means model has been used in the literature, maintaining the

Table 3. Analysis results of Young Women's Project.

	Proposed Method							Method in Hu et al. (2003)		
		Robust		Profile		Info				
	Est	SE	p-value	SE	p-value	SE	p-value	Est	SE	p-value
					CT					
X_1	-0.053	0.038	0.164	0.026	0.040	0.026	0.039	-0.072	0.039	0.066
X_2	-0.251	0.137	0.068	0.157	0.109	0.157	0.110	-0.185	0.113	0.102
X_3	-0.152	0.137	0.267	0.159	0.340	0.159	0.338	-0.088	0.079	0.267
X_4	0.651	0.137	<0.001	0.133	<0.001	0.133	<0.001	0.234	0.129	0.070
X_5	-0.455	0.214	0.034	0.161	0.005	0.162	0.005	-0.417	0.164	0.011
X_6	0.811	0.332	0.015	0.319	0.011	0.319	0.011	1.136	0.473	0.016
X_7	-0.483	0.246	0.049	0.168	0.004	0.168	0.004	-0.409	0.201	0.042
					NG					
X_1	0.017	0.053	0.745	0.041	0.675	0.041	0.676	0.017	0.054	0.759
X_2	-0.479	0.229	0.037	0.214	0.025	0.212	0.024	-0.116	0.147	0.431
X_3	-0.219	0.217	0.313	0.188	0.244	0.188	0.244	0.026	0.120	0.832
X_4	0.966	0.189	<0.001	0.184	<0.001	0.183	<0.001	0.544	0.117	<0.001
X_5	-0.240	0.391	0.538	0.279	0.389	0.278	0.388	-0.365	0.350	0.298
X_6	0.599	0.545	0.272	0.400	0.134	0.399	0.133	1.335	0.854	0.118
X_7	-0.203	0.658	0.758	0.196	0.300	0.196	0.300	-0.506	0.485	0.297
					TV					
X_1	0.034	0.049	0.488	0.031	0.270	0.031	0.267	0.024	0.057	0.677
X_2	-0.469	0.201	0.020	0.167	0.005	0.168	0.005	-0.354	0.146	0.015
X_3	-0.335	0.187	0.073	0.157	0.033	0.158	0.034	-0.017	0.111	0.876
X_4	0.860	0.180	<0.001	0.139	<0.001	0.139	<0.001	0.396	0.154	0.010
X_5	-0.684	0.256	0.007	0.227	0.003	0.231	0.003	-0.661	0.284	0.020
X_6	-0.079	0.347	0.820	0.251	0.753	0.252	0.754	0.343	0.459	0.455
X_7	-0.181	0.364	0.619	0.257	0.482	0.258	0.484	0.021	0.280	0.940

monotonicity of the mean function is often difficult, if not impossible. Interestingly, in the analysis of YWP data, the mean model failed to detect the effects of known STI risk factors (multiple partners and practice of unprotected sex), thus raising questions about its applicability in practical situations. This research has attempted to address several key challenges that analysts face when using proportional rates models with time-varying covariates, namely the robustness of estimators when the Poisson process assumption is violated and the computational issues associated with the profile likelihood method for variance estimation. We present a robust solution with carefully developed asymptotic properties of the proposed estimators that do not rely on the Poisson process assumption. Through numerical studies, we have demonstrated the good performance of the proposed procedures in simulated situations.

One significant contribution of the proposed approach is the introduction of a variance estimation method for $\hat{\boldsymbol{\beta}}$ based on the EM algorithm. This innovation has played an essential role in overcoming the challenge arising from the absence of an explicit form for the asymptotic variance. The solution presented in this paper not only addresses the specific difficulty in panel count data analysis but also sheds light on the broader issue of handling the lack of closed-form expressions for asymptotic variance. This challenge is commonly encountered in the analysis of interval-censored failure time data under semiparametric models, where EM algorithms are frequently employed for estimation. In such situations, analysts typically resort to computationally heavy methods for variance estimation, such as profile likelihood or bootstrapping methods [11,21,22]. An important justification for the proposed variance estimation technique is that consistent estimation of $\bar{\mathbf{x}}$ is achievable through the derivatives of the discretized profiled rate function in the EM algorithm. Furthermore, a critical observation is that the efficient score function at $\boldsymbol{\theta}_0$ is econnected to its EM counterpart. Consequently, the variance estimation methods utilized in this work have the potential to be applied or extended to other models where profile likelihood methods are employed. We believe that this research has laid the initial theoretical foundation for future extensions in this direction.

As highlighted earlier, this research holds promising potential for extension in various directions. For instance, future work could encompass the analysis of mixed recurrent event and panel count data [12,44,49], modeling multivariate panel count data, the development of robust goodness-of-fit tests, and joint analysis of panel count data and interval-censored failure time data [8–10]. With these extensions in mind, we present this study as an initial yet significant stepping stone toward addressing a broader range of methodological issues concerning panel count data.

8. Proofs of the asymptotic properties and variance estimation

In this section, we will sketch the proofs for Theorems 3.1 - 3.3 and Theorem 4.1 as well as providing some details about the derivation of the proposed variance estimation. The proofs of the asymptotic properties rely on the modern empirical process theory. Let $\mathbb{P}_n$ and $\mathbf{P}$ denote the empirical measure and the true probability measure, respectively. Denote the empirical process by $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - \mathbf{P})$. For convenience, denote $D_j^0(\boldsymbol{\theta}) = \int_{T_{\cdot(j-1)}}^{T_{\cdot j}} \exp(\boldsymbol{\beta}^\top \mathbf{X}(t))d\Lambda(t)$ and $D_j^1(\boldsymbol{\theta}, h) = \int_{T_{\cdot(j-1)}}^{T_{\cdot j}} \exp(\boldsymbol{\beta}^\top \mathbf{X}(t))h(t)d\Lambda(t)$ where $h(t)$ is scalar- or vector-value function of $t \in [\tau_0, \tau]$. Let $\mathbb{M}_n(\boldsymbol{\theta}) = n^{-1}\ell(\boldsymbol{\beta}, \Lambda) = \mathbb{P}_n m_{\boldsymbol{\theta}}$ and $\mathbf{M}(\boldsymbol{\theta}) = \mathbf{P} m_{\boldsymbol{\theta}}$, where $m_{\boldsymbol{\theta}}(O) = \sum_{j=1}^J \Delta N_{\cdot j} \log\{D_j^0(\boldsymbol{\theta})\} - D_j^0(\boldsymbol{\theta})$. Here, we omit the term $-\log(\Delta N_{ij}!)$ as it does not involve the parameters. For simplicity, we suppress (O) in $m_{\boldsymbol{\theta}}(O)$ in the following.

8.1. Proof of Theorem 3.1

We begin with proving that the true parameter $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0, \Lambda_0)$ is the unique maximizer of $\mathbf{M}(\boldsymbol{\theta})$. From the definition of $\mathbf{M}(\boldsymbol{\theta})$,

$$\mathbf{M}(\boldsymbol{\theta}_0) - \mathbf{M}(\boldsymbol{\theta}) = \int \int_{u_1}^{u_2} \exp\left(\boldsymbol{\beta}^\top \mathbf{x}(t)\right) d\Lambda(t)\, h \left[\frac{\int_{u_1}^{u_2} \exp\left(\boldsymbol{\beta}_0^\top \mathbf{x}(t)\right) d\Lambda_0(t)}{\int_{u_1}^{u_2} \exp\left(\boldsymbol{\beta}^\top \mathbf{x}(t)\right) d\Lambda(t)} \right] d\nu_1(u_1, u_2, \mathbf{x}), \quad (8)$$

where $h(z) = z\log z - z + 1$. The function $h(z)$ is nonnegative for all $z > 0$ with the equality holding only at $z = 1$. Therefore, $\mathbf{M}(\boldsymbol{\theta}_0) \geq \mathbf{M}(\boldsymbol{\theta})$ and $\mathbf{M}(\boldsymbol{\theta}_0) = \mathbf{M}(\boldsymbol{\theta})$ if and only if

$$\int_{u_1}^{u_2} \exp\left(\boldsymbol{\beta}_0^\top \mathbf{x}(t)\right) d\Lambda_0(t) = \int_{u_1}^{u_2} \exp\left(\boldsymbol{\beta}^\top \mathbf{x}(t)\right) d\Lambda(t), \quad \text{a.e. with respect to } \nu_1. \quad (9)$$

Under condition (C1), (9) reduces to $\exp(\boldsymbol{\beta}_0^\top \mathbf{x}(t))\Lambda_0'(t) = \exp(\boldsymbol{\beta}^\top \mathbf{x}(t))\Lambda'(t)$, a.e. with respect to ν_2. By the similar argument to prove (5.2) in Wellner and Zhang [43, page 2123], we further have $\{1 - \exp((\boldsymbol{\beta}_0 - \boldsymbol{\beta})^\top \mathbf{x}(t))\}^2 = (\Lambda'(t)/\Lambda_0'(t) - 1)^2$ a.e. with respect to ν_2. This leads to $\boldsymbol{\beta}_0 = \boldsymbol{\beta}$ and $\Lambda_0'(t) = \Lambda'(t)$ a.e. with respect to μ_2 in view of (C3) and (C4). The latter equality also implies $\Lambda_0(t) = \Lambda(t)$ a.e. with respect to μ_2 since $\Lambda(0) = \Lambda_0(0) = 0$. It follows that $\boldsymbol{\theta}_0$ is the unique maximizer of $\mathbf{M}(\boldsymbol{\theta})$.

We next show $\hat{\Lambda}(t)$ is uniformly bounded for $t \in [0, \tau]$ a.s.. For any given $\epsilon > 0$, let $\tilde{\boldsymbol{\theta}}_\epsilon = (\hat{\boldsymbol{\beta}}, (1-\epsilon)\hat{\Lambda} + \epsilon\Lambda_0) = \hat{\boldsymbol{\theta}} + \epsilon(0, \Lambda_0 - \hat{\Lambda}_n)$. Since $\hat{\boldsymbol{\theta}}$ maximizes $\mathbb{M}_n(\cdot)$, we have $\mathbb{M}_n(\hat{\boldsymbol{\theta}}) \geq \mathbb{M}_n(\tilde{\boldsymbol{\theta}}_\epsilon) = \mathbb{M}_n(\hat{\boldsymbol{\theta}} + \epsilon(0, \Lambda_0 - \hat{\Lambda}_n))$.

Therefore, $0 \geq \lim_{\epsilon\to 0}\{\mathbb{M}_n(\hat{\boldsymbol\theta}+\epsilon(0,\Lambda_0-\hat\Lambda_n))-\mathbb{M}_n(\hat{\boldsymbol\theta})\}/\epsilon = G_n - \mathbb{P}_n(\sum_{j=1}^{J} D_j^0((\hat{\boldsymbol\beta},\Lambda_0))+\Delta N_{\cdot j})$, where $G_n = \mathbb{P}_n \sum_{j=1}^{J} \Delta N_{\cdot j} D_j^0((\hat{\boldsymbol\beta},\Lambda_0))/D_j^0(\hat{\boldsymbol\theta}) + D_j^0(\hat{\boldsymbol\theta})$. Then,

$$G_n \leq C_1 \mathbb{P}_n \left\{ \sum_{j=1}^{J} \Delta N_{\cdot j} + \Lambda_0(T_{\cdot J}) \right\} \to_{\text{a.s.}} C_1 \mathbf{P} \left\{ \sum_{j=1}^{J} \Delta N_{\cdot j} + \Lambda_0(T_{\cdot J}) \right\}, \tag{10}$$

for some finite C_1 by conditions (C1), (C2) and (C6) and the strong law of large numbers. The limit on the right hand of (10) is finite under conditions (C1), (C2) and (C6). On the other hand, $\limsup_{n\to\infty} G_n \geq \limsup_{n\to\infty} \mathbb{P}_n \sum_{j=1}^{J} D_j^0(\hat{\boldsymbol\theta})$, because $\Delta N_{\cdot j}$ and $D_j^0(\boldsymbol\theta)$ are both nonnegative by (C1) and (C5). Under conditions (C1), (C2) and (C6), for $0 < \tau_0 < \tau$, $\limsup_{n\to\infty} \mathbb{P}_n \sum_{j=1}^{J} D_j^0(\hat{\boldsymbol\theta}) \gtrsim \limsup_{n\to\infty} \mathbb{P}_n I\{\tau_0 \leq T_{\cdot J} \leq \tau\}\hat\Lambda(T_{\cdot J}) \geq \limsup_{n\to\infty} \hat\Lambda(\tau)\nu_3(\{\tau\})$. The right hand side of the last equality is obviously smaller than the left hand of (10). Hence, $\hat\Lambda(\tau)\nu_3(\{\tau\})$ is finite with probability one. Consequently, $\hat\Lambda(t)$ is uniformly bounded a.s. $t \in [0,\tau]$ because $\nu_3(\{\tau\}) > 0$ in (C3).

Finally, we prove the strong consistency of $\hat{\boldsymbol\theta}$ by the Helly selection theorem and one-sided Glivenko-Cantelli theorem stated as Theorem A.1 in Wellner and Zhang [42]. Using the arguments similar to prove Theorem 4.1 of Wellner and Zhang [42] and those on page 2125 of Wellner and Zhang [43], we obtain the strong consistency of $\hat{\boldsymbol\beta}$ and the almost everywhere convergence of $\hat\Lambda(t)$ to $\Lambda_0(t)$ on μ_2 almost surely. Due to the uniform boundedness of $\hat\Lambda$, the dominated convergence theorem yields the strong consistency of $\hat{\boldsymbol\theta}$ in the metric $d(\cdot,\cdot)$.

8.2. Proof of Theorem 3.2

We verify the conditions of Theorem 3.2.5 of van der Vaart and Wellner [39] to derive the convergence rate of $\hat{\boldsymbol\theta}$. We first show $\mathbf{M}(\boldsymbol\theta_0) - \mathbf{M}(\boldsymbol\theta) \gtrsim d^2(\boldsymbol\theta,\boldsymbol\theta_0)$. From (8), the relation between $d^2(\boldsymbol\theta,\boldsymbol\theta_0)$ and the lower bound of $\mathbf{M}(\boldsymbol\theta_0) - \mathbf{M}(\boldsymbol\theta)$ is mainly determined by $h(z) = z\log z - z + 1$. Since $h(z) \geq (1/4)(z-1)^2$ for z in a small enough neighborhood of 1, for any $\boldsymbol\theta$ in a sufficiently small neighborhood of $\boldsymbol\theta_0$, by (C1) and (C2), $\mathbf{M}(\boldsymbol\theta_0) - \mathbf{M}(\boldsymbol\theta) \gtrsim \int\{\int_{u_1}^{u_2}\exp(\boldsymbol\beta_0^\top \mathbf{x}(t))d\Lambda_0(t) - \int_{u_1}^{u_2}\exp(\boldsymbol\beta^\top\mathbf{x}(t))d\Lambda(t)\}^2 d\nu_1(u_1,u_2,\mathbf{x})$. Let $g(\xi) = \int I(U_1 < t < U_2)\exp(\boldsymbol\beta_\xi^\top \mathbf{X}(t))d\Lambda_\xi(t)$ where $\boldsymbol\beta_\xi = \xi\boldsymbol\beta + (1-\xi)\boldsymbol\beta_0$ and $\Lambda_\xi = \xi\Lambda + (1-\xi)\Lambda_0$. By the mean value theorem, there exists $\xi^* \in [0,1]$ such that $\mathbf{M}(\boldsymbol\theta_0) - \mathbf{M}(\boldsymbol\theta) \gtrsim \nu_1\{g_1h_1+g_2\}^2$ where $g_1(U_1,U_2,\mathbf{X}) = \int_{U_1}^{U_2}(\boldsymbol\beta-\boldsymbol\beta_0)^\top\mathbf{X}(t)\,d\Lambda_0(t)$, $g_2(U_1,U_2) = \Lambda(U_2)-\Lambda_0(U_2)-(\Lambda(U_1)-\Lambda_0(U_1))$ and $h_1(U_1,U_2,\mathbf{X}) = \xi\int_{U_1}^{U_2}(\boldsymbol\beta-\boldsymbol\beta_0)^\top\mathbf{X}(t)d\Lambda(t)/\int_{U_1}^{U_2}(\boldsymbol\beta-\boldsymbol\beta_0)^\top \times \mathbf{X}(t)d\Lambda_0(t) + 1 - \xi$. For convenience, we write the expectation under ν_1 as E_1. To apply Lemma 8.8 in van der Vaart [37], we need to bound $\{E_1[g_1g_2]\}^2$ by $(1-\eta)E_1[g_1^2]\,E_1[g_2^2]$ for some $\eta \in (0,1)$. By the Cauchy–Schwarz inequality, $\{E_1[g_1g_2]\}^2 = \{E_1[g_2E_1[g_1|U_1,U_2]]\}^2 \leq E_1[g_2^2]E_1[\{E_1[g_1|U_1,U_2]\}^2]$. Furthermore,

$$\begin{aligned}
&\{E_1[g_1|U_1,U_2]\}^2 \\
&= \int_{U_1}^{U_2}\int_{U_1}^{U_2}(\boldsymbol\beta-\boldsymbol\beta_0)^\top E_1[\mathbf{X}(t)|U_1,U_2]\,E_1\left[\mathbf{X}^\top(s)|U_1,U_2\right](\boldsymbol\beta-\boldsymbol\beta_0)\,d\Lambda_0(t)\,d\Lambda_0(s) \\
&\leq (1-\rho)\int_{U_1}^{U_2}\int_{U_1}^{U_2}(\boldsymbol\beta-\boldsymbol\beta_0)^\top E_1\left[\mathbf{X}(t)\mathbf{X}^\top(s)|U_1,U_2\right](\boldsymbol\beta-\boldsymbol\beta_0)\,d\Lambda_0(t)\,d\Lambda_0(s) \\
&= (1-\rho)E_1\left[g_1^2|U_1,U_2\right].
\end{aligned}$$

The last inequality is due to condition (C10) with $\rho \in (0,1)$. Consequently, Lemma 8.8 in [37] yields $\nu_1 \{g_1 h_1 + g_2\}^2 \gtrsim \nu_1 \left(g_1^2\right) + \nu_1 \left(g_2^2\right)$. Under (C1) and (C4), by Jensen's inequality, $\nu_1 \left(g_1^2\right) \gtrsim \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2^2$. Therefore, $\nu_1 \{g_1 h_1 + g_2\}^2 \gtrsim d^2 (\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ and so does $\mathbf{M}(\boldsymbol{\theta}_0) - \mathbf{M}(\boldsymbol{\theta})$.

Next, we need to derive $\phi_n(\eta)$ such that $E \sup_{d(\boldsymbol{\theta},\boldsymbol{\theta}_0)<\eta} \left|\mathbb{G}_n \left(m_{\boldsymbol{\theta}} - m_{\boldsymbol{\theta}_0}\right)\right| \lesssim \phi_n(\eta)$. Define classes $\mathcal{M}_\eta = \left\{m_{\boldsymbol{\theta}} - m_{\boldsymbol{\theta}_0} : d(\boldsymbol{\theta}, \boldsymbol{\theta}_0) < \eta, \boldsymbol{\theta} \in \mathcal{B} \times \mathcal{F}\right\}$ and $\mathcal{F}_\eta = \{\Lambda \in \mathcal{F} : \|\Delta\Lambda - \Delta\Lambda_0\|_{L_2(\mu_1)} \le \eta\}$. We start by obtaining the bracketing number of $\mathcal{M}_\eta$ in the following lemma whose proof is provided in the supplementary material Sun et al. [32].

Lemma 2. *Under the conditions (C1), (C2), (C5), (C6), (C8) and (C9),* $\log N_{[\,]}\left(\epsilon, \mathcal{M}_\eta, \|\cdot\|_{P,B}\right) \lesssim 1/\epsilon$, *where the Bernstein norm* $\|f\|_{P,B} = \{2\mathbf{P}(\exp(|f|) - 1 - |f|)\}^{1/2} \le \{\mathbf{P}(\exp(|f|)|f|^2)\}^{1/2}$.

It follows from Lemma 2, $\tilde{J}_{[\,]}\left(\eta, \mathcal{M}_\eta, \|\cdot\|_{P,B}\right) = \int_0^\eta \sqrt{1 + \log N_{[\,]}\left(\epsilon, \mathcal{M}_\eta, \|\cdot\|_{P,B}\right)}\, d\epsilon \lesssim \delta^{1/2}$. The similar arguments to derive the bracketing number of $\mathcal{M}_\eta$ can also yield that $\|f\|_{P,B} \lesssim \eta$ for every $f \in \mathcal{M}_\eta$. By Lemma 3.4.3 of [39], we have $E_{\mathbf{P}}\|\mathbb{G}_n\|_{\mathcal{M}_\eta} \lesssim \tilde{J}_{[\,]}(\eta, \mathcal{M}_\eta, \|\cdot\|_{P,B})\{1 + \tilde{J}_{[\,]}(\eta, \mathcal{M}_\eta, \|\cdot\|_{P,B})/(\eta^2\sqrt{n})\} \lesssim \eta^{1/2} + \eta^{-1}/\sqrt{n}$, leading to $\phi_n(\eta) = \eta^{1/2} + \eta^{-1}/\sqrt{n}$.

It is easy to see $\phi_n(\eta)/\eta$ is a decreasing function. If we let $r_n = n^{1/3}$, $r_n^2\phi_n(1/r_n) = n^{2/3}(n^{-1/6} + n^{-1/6}) = 2\sqrt{n}$. By Theorem 3.2.5 in van der Vaart and Wellner [39], $n^{1/3}d(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = O_p(1)$ which completes the proof.

8.3. Proof of Theorem 3.3

We consider a submodel $\left(\boldsymbol{\beta} + \epsilon h_1, \int (1 + \epsilon h_2)\, d\Lambda\right)$ where $\mathcal{H}_1 = \{h_1 : h_1 \in \mathcal{B}, \|h\|_2 \le 1\}$ and $\mathcal{H}_2 = \{h_2 : h_2$ is a function with bounded total variation in $[0,\tau]\}$. Let $\mathcal{H} = \mathcal{H}_1 \times \mathcal{H}_2$ and define a sequence of maps S_n mapping a neighborhood of $\boldsymbol{\theta}_0$, denoted by $\mathcal{U}$, in the parameter space Θ into $l^\infty(\mathcal{H})$ as $S_n(\boldsymbol{\theta})[h_1, h_2] = \frac{d}{d\epsilon}\frac{1}{n}\ell(\boldsymbol{\beta} + \epsilon h_1, \int(1 + \epsilon h_2)d\Lambda)|_{\epsilon=0} = \mathbb{P}_n m'_{\boldsymbol{\theta}}[h_1, h_2]$, where $m'_{\boldsymbol{\theta}}[h_1] = m'_{1\boldsymbol{\theta}}[h_1] + m'_{2\boldsymbol{\theta}}[h_1]$ with $m'_{1\boldsymbol{\theta}}[h_1] = \sum_{j=1}^J(\Delta N_{\cdot j}/D_j^0(\boldsymbol{\theta}) - 1)D_j^1(\boldsymbol{\theta}, h_1^\top\mathbf{X}))$ and $m'_{2\boldsymbol{\theta}}[h_2] = \sum_{j=1}^J(\Delta N_{\cdot j}/D_j^0(\boldsymbol{\theta}) - 1)D_j^1(\boldsymbol{\theta}, h_2)$. Set $S(\boldsymbol{\theta})[h_1, h_2] = \mathbf{P}m'_{\boldsymbol{\theta}}[h_1, h_2]$.

Inspired by Liu et al. [25], we need to verify the following conditions to derive the asymptotic normality of $\hat{\boldsymbol{\beta}}$:

- (a1) $\sqrt{n}(S_n - S)(\hat{\boldsymbol{\theta}})[h_1, h_2] - \sqrt{n}(S_n - S)(\boldsymbol{\theta}_0)[h_1, h_2] = o_p(1)$.
- (a2) $S(\boldsymbol{\theta}_0)[h_1, h_2] = 0$ and $S_n(\hat{\boldsymbol{\theta}})[h_1, h_2] = o_p(n^{-1/2})$.
- (a3) $\sqrt{n}(S_n - S)(\boldsymbol{\theta}_0)[h_1, h_2]$ converges in distribution to a tight Gaussian process on $l^\infty(\mathcal{H})$.
- (a4) $S(\boldsymbol{\theta})[h_1, h_2]$ is Fréchet-differentiable at $\boldsymbol{\theta}_0$ with a continuous derivative $\dot{S}(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)[h_1, h_2]$.
- (a5) $\sqrt{n}\{S(\hat{\boldsymbol{\theta}})[h_1, h_2] - S(\boldsymbol{\theta}_0)[h_1, h_2] - \dot{S}(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)[h_1, h_2]\} = o_p(1)$.

The following proposition provides a general form of the asymptotic distribution of $\hat{\boldsymbol{\theta}}$.

Proposition 1. *Under conditions (a1)-(a2) and (a4)-(a5),* $-\sqrt{n}\dot{S}(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)[h_1, h_2] = \sqrt{n}(S_n - S)(\boldsymbol{\theta}_0)[h_1, h_2] + o_p(1)$.

The proof of Proposition 1 is provided in the supplementary material Sun et al. [32].

To verify (a1), for η_n such that $d(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0)/\eta_n \to 0$, define $\Psi_n = \{\psi(\boldsymbol{\theta}; h_1, h_2) = m'_{\boldsymbol{\theta}}[h_1, h_2] - m'_{\boldsymbol{\theta}_0}[h_1, h_2], d(\boldsymbol{\theta}, \boldsymbol{\theta}_0) < \eta_n, \boldsymbol{\theta} \in \mathcal{B} \times \mathcal{F}, h_1 \times h_2 \in \mathcal{H}\}$. Since the total variation of h_2 is bounded, by Corollary 2.7.2 in van der Vaart and Wellner [39], $\log N_{[\,]}(\epsilon, \mathcal{H}_2, L_2(\mathbf{P})) \lesssim 1/\epsilon$. By the similar bracketing entropy argument in the proof of Theorem 3.2, we can deduce that $\log N_{[\,]}(\epsilon, \Psi_n, L_2(\mathbf{P})) \lesssim 1/\epsilon$.

Then, $J_{[]}(\delta_n, \Psi_n, L_2(\mathbf{P})) = \int_0^{\delta_n} \sqrt{\log N_{[]}(\epsilon, \Psi_n, L_2(\mathbf{P}))} d\epsilon \lesssim \int_0^{\delta_n} \sqrt{1/\epsilon} d\epsilon \lesssim \delta_n^{1/2}$. Clearly, $J_{[]}(\delta_n, \Psi_n, L_2(\mathbf{P})) \to 0$ for every sequence $\delta_n \to 0$. Moreover, $\sup_{(h_1,h_2)\in\mathcal{H}} \mathbf{P}(\psi(\hat{\boldsymbol{\theta}}; h_1, h_2) - \psi(\boldsymbol{\theta}_0; h_1, h_2))^2 \to 0$, in probability as $d(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) \to 0$ because $\psi(\boldsymbol{\theta}; h_1, h_2)$ is Lipschitz continuous in $\boldsymbol{\theta}$ under $L_2(\mathbf{P})$ by (C1), (C2), (C5) and (C8). It is not hard to show that $\Pr(\psi(\boldsymbol{\theta}_n; h_1, h_2) \in \Psi_n)$ goes to 1 as $n \to \infty$ by the definition of η_n. It follows from Theorem 2.2 in van der Vaart and Wellner [40] that

$$\sqrt{n}\left|(S_n - S)\left(\hat{\boldsymbol{\theta}}\right)[h_1, h_2] - (S_n - S)(\boldsymbol{\theta}_0)[h_1, h_2]\right| \leq \sup_{(h_1,h_2)\in\mathcal{H}} \sqrt{n}\left|(\mathbb{P} - \mathbf{P})\psi\left(\hat{\boldsymbol{\theta}}; h_1, h_2\right)\right| \to_p 0,$$

and hence (a1) follows.

To verify (a2), it is not hard to show $S(\boldsymbol{\theta}_0)[h_1, h_2] = 0$ after some algebra. For $S_n(\hat{\boldsymbol{\theta}})[h_1, h_2]$, let $S_n(\hat{\boldsymbol{\theta}})[h_1, h_2] = S_{1n}(\hat{\boldsymbol{\theta}})[h_1] + S_{1n}(\hat{\boldsymbol{\theta}})[h_2]$ where $S_{kn}(\boldsymbol{\theta})[h_k] = m'_{k\boldsymbol{\theta}}[h_k]$, for $k = 1, 2$. The first part $S_{1n}(\hat{\boldsymbol{\theta}})[h_1] = 0$ holds as $S_{1n}(\hat{\boldsymbol{\theta}})[h_1]$ is the score function for $\boldsymbol{\beta}$. For $S_{2n}(\boldsymbol{\theta})[h_2]$, we consider a submodel $\boldsymbol{\theta}_\epsilon = \left(\hat{\boldsymbol{\beta}}, \int (1 + \epsilon h_2) d\hat{\Lambda}\right)$ for $h_2 \in \mathcal{H}_2$. By the definition of Riemann–Stieltjes integral $\int (1 + \epsilon h_2) d\hat{\Lambda} \in \mathcal{F}^*$ for $h_2 \in \mathcal{H}_2$ and small enough ϵ. Since $\hat{\boldsymbol{\theta}}$ maximizes $\mathbb{P}_n m_{\boldsymbol{\theta}}$ over $\mathcal{B} \times \mathcal{F}^*$, we find that $0 = \lim_{\epsilon\to 0} d\mathbb{P}_n m_{\boldsymbol{\theta}_\epsilon}/d\epsilon = S_{2n}(\hat{\boldsymbol{\theta}})[h_2]$, for any $h_2 \in \mathcal{H}_2$. Therefore, $n^{1/2} S_n(\hat{\boldsymbol{\theta}})[h_1, h_2] = o_p(1)$ because $S_n(\hat{\boldsymbol{\theta}})[h_1, h_2] = 0$.

For (a3), let $\bar{\Psi} = \{m'_{\boldsymbol{\theta}_0}[h_1, h_2], h_1 \times h_2 \in \mathcal{H}\}$. It is easily seen that $\mathcal{H}_1$ and $\mathcal{H}_2$ are $\mathbf{P}$-Donsker classes van der Vaart and Wellner [39, Example 2.10.27]. It can be shown that $m'_{\boldsymbol{\theta}_0}[h_1, h_2]$ is Lipschitz continuous with respect to h_1 and h_2 under conditions (C1), (C2), (C5) and (C6). By Theorem 2.10.6 in [39], $\bar{\Psi}$ is a $\mathbf{P}$-Donsker class and we have (a3) by the definition of the Donsker class.

The Fréchet differentiability of $S(\boldsymbol{\theta})$ holds obviously by the continuity of $S(\boldsymbol{\theta})$ in $\boldsymbol{\theta}$. Let the Fréchet derivative of $S(\boldsymbol{\theta})$ at $\boldsymbol{\theta}_0$ denote by $\dot{S}(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) = d\, S(\boldsymbol{\theta}_0 + \epsilon(\boldsymbol{\theta} - \boldsymbol{\theta}_0))/d\epsilon|_{\epsilon=0}$ that is a map from the space $\{\boldsymbol{\theta} - \boldsymbol{\theta}_0 : \boldsymbol{\theta} \in \mathcal{U}\}$ to $l^\infty(\mathcal{H})$. Then, $\dot{S}(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)[h_1, h_2] = \left(\boldsymbol{\beta} - \boldsymbol{\beta}_0\right)^\top Q_1(h_1, h_2) + \int_0^\tau Q_2(h_1, h_2, t)\, d(\Lambda - \Lambda_0)$, where $Q_1(h_1, h_2) = -\mathbf{P}\sum_{j=1}^J D_j^1(\boldsymbol{\theta}_0, h_1^\top \mathbf{X} + h_2) D_j^1(\boldsymbol{\theta}_0, \mathbf{X})/D_j^0(\boldsymbol{\theta}_0)$ and $Q_2(h_1, h_2, t) = -\mathbf{P}\sum_{j=1}^J D_j^1(\boldsymbol{\theta}_0, h_1^\top \mathbf{X} + h_2)/D_j^0(\boldsymbol{\theta}_0) I(T_{\cdot(j-1)} < t < T_{\cdot j}) \exp(\boldsymbol{\beta}_0^\top \mathbf{X}(t))$.

Last for (a5), by using the mean value theorem twice, we can show under conditions (C1), (C2), (C5), (C6) and (C8), $|S(\hat{\boldsymbol{\theta}})[h_1, h_2] - S(\boldsymbol{\theta}_0)[h_1, h_2] - \dot{S}(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)[h_1, h_2]| \lesssim O(d^2(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}))$. It is seen that $\sqrt{n} d^2(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = o_p(n^{1/2-2/3}) = o_p(1)$ and thus we verify (a5).

After verifying (a1)-(a5), we next show that there exists a unique h_2^* such that $Q_2(h_1, h_2^*, t) = 0$ for any $h_1 \in \mathcal{H}_1$. Define a bilinear mapping $\varphi : \mathcal{H}_2 \times \mathcal{H}_2 \to \mathbb{R}$, such that, for $v_1, v_2 \in \mathcal{H}_2$, $\varphi(v_1, v_2) = \mathbf{P}\sum_{j=1}^J D_j^1(\boldsymbol{\theta}_0, v_1) D_j^1(\boldsymbol{\theta}_0, v_2)/D_j^0(\boldsymbol{\theta}_0)$ and a linear mapping $b : \mathcal{H}_2 \to \mathbb{R}$, $b(v; h_1) = \varphi(-h_1^\top \mathbf{X}(t), v)$. Obviously, $\varphi(v_1, v_2)$ is linear in v_1 and v_2, $\varphi(v, v) \geq 0$, and $\varphi(v_1, v_2) = \varphi(v_2, v_1)$. Furthermore, if $\varphi(v, v) = 0$ for $v \in \mathcal{H}_2$, then $\mathbf{P}\sum_{j=1}^J \{D_j^1(\boldsymbol{\theta}_0, v)\}^2/D_j^0(\boldsymbol{\theta}_0)$. This suggest that $v(t) = 0$ a.e. on v_1 because $\exp(\boldsymbol{\beta}_0^\top \mathbf{X}(t))$ and $d\Lambda_0(t)$ are positive on v_1 by condition (C1). Hence, $\varphi(v_1, v_2)$ is an inner product on $\mathcal{H}_2$ and we can define a norm on $\mathcal{H}_2$ such that $\|v\|_{\mathcal{H}_2}^2 = \varphi(v, v)$ for $v \in \mathcal{H}_2$. As a result, the mapping φ is continuous and bounded. In addition, $b(v; h_1)$ is obviously a continuous linear operator for any h_1. Thus, by the F. Ries representation theorem [5, Theorem 4.6-1], given h_1, there exists a unique $h_2^* \in \mathcal{H}_2$ such that $\varphi\left(h_2^*, h_2\right) = b(h_2; h_1)$ for any $h_2 \in \mathcal{H}_2$. Recall that $d\Lambda - d\Lambda_0$ can expressed as $\epsilon h_2 d\Lambda_0$ for $h_2 \in \mathcal{H}_2$ and some $\epsilon > 0$. Then, $\int_0^\tau Q_2\left(h_1, h_2^*, t\right) d(\Lambda - \Lambda_0) = \epsilon\left\{b(h_2; h_1) - \varphi\left(h_2^*, h_2\right)\right\} = 0$ for any $\Lambda \in \mathcal{F}$.

We now give the definition of $\bar{\mathbf{x}}$. Solving h_2^* from $\varphi\left(h_2^*, h_2\right) = b(h_2; h_1)$ for a given h_1 can be seen as a unique operator $H : \mathcal{H}_1 \to \mathcal{H}_2$ that is linear and continuous in h_1 by the F. Ries representation theorem. Hence, we can write $h_2^*(t) = Hh_1 = -h_1^\top \bar{\mathbf{x}}(t)$ where $\bar{\mathbf{x}}(t)$ is a p-dimensional function on $[0, \tau]$ whose lth element is the unique solution of $-h_1^\top \varphi(\bar{\mathbf{x}}(t), h_2) = b(h_2; -e_l)$ for any h_2. Recall that $\mathcal{H}_3 = \{h_3(t) \in \mathbf{R}^p, t \in [0, \tau]$: each element of $h_3(t)$ is of bounded variation over $[0, \tau]\}$. Since h_1 is arbitrary, we can then define $\bar{\mathbf{x}}$ as the unique solution of

$$\mathbf{P}\sum_{j=1}^{J} D_j^1(\boldsymbol{\theta}_0, \mathbf{X} - h_3) D_j^1(\boldsymbol{\theta}_0, h_2)/D_j^0(\boldsymbol{\theta}_0) = 0, \tag{11}$$

over $h_3 \in \mathcal{H}_3$ for any $h_2 \in \mathcal{H}_2$.

Next, we derive $\boldsymbol{\Omega}$ from Q_1. Plugging $h_2^*(t) = -h_1^\top \bar{\mathbf{x}}(t)$ in Q_1, we have $Q_1\left(h_1, h_2^*\right) = -h_1^\top \boldsymbol{\Omega}$ such that $\boldsymbol{\Omega} = -\mathbf{P}\sum_{j=1}^{J} D_j(\boldsymbol{\theta}_0) D_j^1(\boldsymbol{\theta}_0, \mathbf{X})^\top / D_j^0(\boldsymbol{\theta}_0)$ where we define $D_j(\boldsymbol{\theta}) = D_j^1(\boldsymbol{\theta}, \mathbf{X} - \bar{\mathbf{x}})$. Moreover, we can choose h_2 as $Hh_1' = -h_1'^\top \bar{\mathbf{x}}(t)$ for any $h_1' \in \mathcal{H}_1$ and $h_3 = \bar{\mathbf{x}}$. It follows from (11) that $\mathbf{P}\sum_{j=1}^{J} D_j(\boldsymbol{\theta}_0) D_j^1(\boldsymbol{\theta}_0, \bar{\mathbf{x}})^\top / D_j^0(\boldsymbol{\theta}_0) = 0$. Adding this matrix to $\boldsymbol{\Omega}$, after some algebra, we obtain that $\boldsymbol{\Omega} = -E[\sum_{j=1}^{J} D_j(\boldsymbol{\theta}_0)^{\otimes 2}/D_j^0(\boldsymbol{\theta}_0)]$, coinciding with the form in Theorem 3.3. By condition (C4), $-\boldsymbol{\Omega}$ is positive definite and nonsingular, i.e., $\boldsymbol{\Omega}^{-1}$ exists.

Then, by condition (a3) and Proposition 1, for any h_1,

$$\sqrt{n} h_1^\top \boldsymbol{\Omega}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) = -\sqrt{n}\,(S_n - S)(\boldsymbol{\theta}_0)\left[h_1, h_2^*\right] + o_P(1) \to_d N\left(0, \sigma^2\right), \tag{12}$$

where $\sigma^2 = E\left[m'_{\boldsymbol{\theta}_0}[h_1, h_2^*]^2\right] = h_1^\top \boldsymbol{\Sigma} h_1$ and

$$\boldsymbol{\Sigma} = E\left[\left\{\sum_{j=1}^{J}\left\{\Delta N_{\cdot j} - D_j^0(\boldsymbol{\theta}_0)\right\}\frac{D_j(\boldsymbol{\theta}_0)}{D_j^0(\boldsymbol{\theta}_0)}\right\}^{\otimes 2}\right] = E\left[\sum_{j_1=1}^{J}\sum_{j_2=1}^{J} D_{j_1 j_2}^{02}\frac{D_{j_1}(\boldsymbol{\theta}_0)}{D_{j_1}^0(\boldsymbol{\theta}_0)}\frac{D_{j_2}(\boldsymbol{\theta}_0)}{D_{j_2}^0(\boldsymbol{\theta}_0)}\right], \tag{13}$$

where $D_{j_1 j_2}^{02} = \mathrm{Cov}[\Delta N_{\cdot j_1}\Delta N_{\cdot j_2} | T_{\cdot j_1}, T_{\cdot (j_1-1)}, T_{\cdot j_2}, T_{\cdot (j_2-1)}, \mathbf{X}, J]$. Since h_1 is arbitrary, by the Cramér-Wold theorem, $\sqrt{n}\boldsymbol{\Omega}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \to_d N(0, \boldsymbol{\Sigma})$. When the Poisson assumption holds, $D_{j_1 j_2}^{02} = D_j^0(\boldsymbol{\theta}_0)$ if $j_1 = j_2 = j$ and zero otherwise. Hence, after some algebra, $\boldsymbol{\Sigma} = -\boldsymbol{\Omega}$ under the Poisson assumption.

Furthermore, we show that $\hat{\boldsymbol{\beta}}$ achieves the asymptotic efficiency if the Poisson model holds. Under the Poisson assumption, the log-likelihood function $\mathbb{P}_n m_{\boldsymbol{\theta}}$ would be the "true" log-likelihood function and we can take advantage of the existing conclusions regarding semiparametric efficiency [see, 18,38]. Denote the derivative of $m_{\boldsymbol{\theta}}$ with respect to $\boldsymbol{\beta}$ by $\dot{m}_{\boldsymbol{\theta}}[h_1] = \sum_{j=1}^{J}(\Delta N_{\cdot j}/D_j^0(\boldsymbol{\theta}) - 1) h_1^\top D_j^1(\boldsymbol{\theta}, \mathbf{X})$. Denote the score operator for Λ and its adjoint operator by $B_{\boldsymbol{\beta},\Lambda}$ and $B_{\boldsymbol{\beta},\Lambda}^*$, respectively [38, Chapter 25]. Then, with the submodel $d\Lambda_\epsilon = (1 + \epsilon h_2)\, d\Lambda_0$, Theorem 1 in Mao [27] yields that $B_{\boldsymbol{\beta},\Lambda}^* \dot{m}_{\boldsymbol{\theta}}[h_1](\cdot) = -\sum_{j=1}^{J}\int_{T_{\cdot(j-1)}}^{T_{\cdot j}} E_{\boldsymbol{\theta}}[\exp(\boldsymbol{\beta}^\top \mathbf{X}(u)) h_1^\top \mathbf{X}(u) \exp(\boldsymbol{\beta}^\top \mathbf{X}(\cdot))/D_j^0(\boldsymbol{\theta})]\, d\Lambda(u)$, and $B_{\boldsymbol{\beta},\Lambda}^* B_{\boldsymbol{\beta},\Lambda} \times h_2(\cdot) = \sum_{j=1}^{J}\int_{T_{\cdot(j-1)}}^{T_{\cdot j}} E_{\boldsymbol{\theta}}[\exp(\boldsymbol{\beta}^\top \mathbf{X}(u)) \exp(\boldsymbol{\beta}^\top \mathbf{X}(\cdot))/D_j^0(\boldsymbol{\theta})] h_2(u)\, d\Lambda(u)$. It is not hard to show that, for any $h_1 \in \mathcal{H}_1$ there exists h_2^* such that $\int_0^\tau Q_2\left(h_1, h_2^*, t\right) d\left(\Lambda - \Lambda_0\right) = 0$ for any $\Lambda \in \mathcal{F}$ implies that the normal equation $B_{\boldsymbol{\beta},\Lambda}^* \dot{m}_{\boldsymbol{\theta}}\left[h_1\right](\cdot) = B_{\boldsymbol{\beta},\Lambda}^* B_{\boldsymbol{\beta},\Lambda} h_2^*(\cdot)$ holds. It follows from $h_2^* = h_1^\top \bar{\mathbf{x}}$ that $\bar{\mathbf{x}}$ is a least favorable direction vector. Clearly, the efficient score at $\boldsymbol{\theta}_0$ is then

$$\tilde{m}_{\boldsymbol{\theta}_0} = \sum_{j=1}^{J}(\Delta N_{\cdot j} - D_j^0(\boldsymbol{\theta}_0)) D_j(\boldsymbol{\theta}_0)/D_j^0(\boldsymbol{\theta}_0). \tag{14}$$

By (12) and some algebra, $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \tilde{I}_{\boldsymbol{\beta}_0,\Lambda_0}^{-1} \mathbb{G}_n \tilde{m}_{\boldsymbol{\theta}_0} + o_P(1)$ where $\tilde{I}_{\boldsymbol{\beta}_0,\Lambda_0}$ is the efficient information matrix for $\boldsymbol{\beta}$ and $\tilde{I}_{\boldsymbol{\beta}_0,\Lambda_0} = -\mathbf{P}(\partial \tilde{m}_{\boldsymbol{\theta}}/\partial \boldsymbol{\beta}|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0, \Lambda=\Lambda_0}) = -\boldsymbol{\Omega}$. Then, $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is asymptotic linear in the efficient influence function $\tilde{I}^{-1}_{\boldsymbol{\beta}_0,\Lambda_0}\tilde{m}_{\boldsymbol{\theta}_0}$. Therefore, $\hat{\boldsymbol{\beta}}$ is asymptotic efficient under Poisson assumption [38, Lemma 25.23].

8.4. The variance estimation derivation and proof of Theorem 4.1

In this section, we derive the variance estimators proposed in Section 4. We prove their consistency in Theorem 4.1 and provide the proof of Lemma 1.

8.4.1. Proof of Theorem 4.1

In the proof of Theorem 3.3, (13), (14) and some algebra show that $\mathbf{P}\,\partial\tilde{m}_{\boldsymbol{\theta}}/\partial\boldsymbol{\beta}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \boldsymbol{\Omega}$ that $\boldsymbol{\Sigma} = \mathbf{P}\left(\tilde{m}_{\boldsymbol{\theta}_0}\right)^{\otimes 2}$ despite the validity of the Poisson assumption. In other words, both $\boldsymbol{\Omega}$ and $\boldsymbol{\Sigma}$ can be derived from the efficient score function $\tilde{m}_{\boldsymbol{\theta}_0}$ defined in (14) regardless of the Poisson model. By this observation, the main idea of the proof is to derive the alternative forms of $\boldsymbol{\Omega}$ and $\boldsymbol{\Sigma}$ expressed by t_k and λ_k by connecting $\tilde{m}_{\boldsymbol{\theta}}$ to the EM algorithm. Facilitated by the alternative forms, we then show the consistency results in Theorem 4.1.

We first show the consistency of $\hat{\boldsymbol{\Sigma}}$. Recall that $\mathcal{H}_3 = \{h_3(t) \in \mathbf{R}^p, t \in [0,\tau]$: each element of $h_3(t)$ is of bounded variation over $[0,\tau]\}$. Define $\tilde{m}^{EM}[\boldsymbol{\theta}; h_3] = \sum_{k=1}^{K}\Delta_{\cdot k}(\hat{E}[W_{\cdot k}|\boldsymbol{\beta},\boldsymbol{\lambda}] - \exp(\boldsymbol{\beta}^\top\mathbf{X}(t_k))\lambda_k)(\mathbf{X}(t_k) - h_3(t_k))$. Clearly, $\hat{\boldsymbol{\Sigma}} = \mathbb{P}_n(\tilde{m}^{EM}[\hat{\boldsymbol{\theta}};\hat{\bar{\mathbf{x}}}])^{\otimes 2}$. We next establish the connection between $\tilde{m}^{EM}[\boldsymbol{\theta}; h_3]$ and the efficient score function $\tilde{m}_{\boldsymbol{\theta}}$. Using the arguments to discretizing $\Lambda(t)$ by $\boldsymbol{\lambda}$ on $\{t_k\}_{k=1}^{K}$, we have

$$
\begin{aligned}
&\sum_{j=1}^{J}\left(\Delta N_{\cdot j}/D_j^0(\boldsymbol{\theta}) - 1\right) D_j^1(\boldsymbol{\theta},\mathbf{X} - h_3)\\
&= \sum_{k=1}^{K}\Delta_{\cdot k}\left(\frac{\Delta_{\cdot k}\exp\left(\boldsymbol{\beta}^\top\mathbf{X}(t_k)\right)\lambda_k\Delta N_{\cdot j\cdot k}}{\sum_{l\in A_{\cdot k}}\Delta_{\cdot k}\exp\left(\boldsymbol{\beta}^\top\mathbf{X}(t_l)\right)\lambda_l} - \exp\left(\boldsymbol{\beta}^\top\mathbf{X}(t_k)\right)\lambda_k\right)\left(\mathbf{X}(t_k) - h_3(t_k)\right)\\
&\overset{(6)}{=} \sum_{k=1}^{K}\Delta_{\cdot k}\left(\hat{E}\left[W_{\cdot k}|\boldsymbol{\beta},\boldsymbol{\lambda}\right] - \exp\left(\boldsymbol{\beta}^\top\mathbf{X}(t_k)\right)\lambda_k\right)\left(\mathbf{X}(t_k) - h_3(t_k)\right) = \tilde{m}^{EM}\left[\boldsymbol{\theta}; h_3\right].
\end{aligned}
$$

By setting $h_3 = \bar{\mathbf{x}}$, we have $\tilde{m}^{EM}[\boldsymbol{\theta};\bar{\mathbf{x}}] = \tilde{m}_{\boldsymbol{\theta}}$. Consequently, we obtain $\boldsymbol{\Sigma} = \mathbf{P}\left(\tilde{m}^{EM}[\boldsymbol{\theta}_0;\bar{\mathbf{x}}]\right)^{\otimes 2}$. Since each element of h_3 belongs to the bounded total variation space $\mathcal{H}_3$, we can use bracketing entropy arguments similar to those in the proof of Theorem 3.2-3.3 to show that the set $\Phi_\eta^{EM} = \{\tilde{m}^{EM}[\boldsymbol{\theta}; h_3], d(\boldsymbol{\theta},\boldsymbol{\theta}_0) < \eta, \boldsymbol{\theta} \in \mathcal{B}\times\mathcal{F}, h_3 \in \mathcal{H}_3\}$, is a $\mathbf{P}$-Donsker class and Glivenko-Cantelli class under conditions (C1), (C2), (C5), (C6) and (C8). By Lemma 1 and Theorem 3.1, $\hat{\bar{\mathbf{x}}}$ converges to $\bar{\mathbf{x}}$ in probability and $\hat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}_0$. It then follows from the continuity of $\tilde{m}^{EM}[\boldsymbol{\theta}; h_3]$ in $\boldsymbol{\theta}$ and h_3 that $\hat{\boldsymbol{\Sigma}}$ converges to $\boldsymbol{\Sigma}$ in probability.

Next, we establish the consistency of $\hat{\boldsymbol{\Omega}}$. Likewise, we need to derive the form of $\boldsymbol{\Omega}$ in terms of t_k and $\boldsymbol{\lambda}_k$ first. However, it is not straightforward due to the mathematical intractability of the explicit form of $d\hat{\lambda}_k(\boldsymbol{\beta})/d\boldsymbol{\beta}$. From the proof of Theorem 3.3 on page 3269, $-\boldsymbol{\Omega} = \boldsymbol{\Sigma}$ when the Poisson model holds and the form of $\boldsymbol{\Omega}$ stays the same no matter if the Poisson assumption holds. Thus, we can find the form of $\boldsymbol{\Omega}$ by explicitly calculating $\boldsymbol{\Sigma}$ under the Poisson assumption. Recalling the form of $\boldsymbol{\Sigma}$, under the Poisson model, some algebra yields that

$$
\boldsymbol{\Omega} = -E\left[\sum_{j=1}^{J}\frac{\left\{\sum_{T_{\cdot(j-1)}<t_k\le T_{\cdot j}}\exp\left(\boldsymbol{\beta}_0^\top\mathbf{X}(t_k)\right)\lambda_{0k}\left(\mathbf{X}(t_k) - \bar{\mathbf{x}}(t_k)\right)\right\}^{\otimes 2}}{\sum_{T_{\cdot(j-1)}<t_k\le T_{\cdot j}}\exp\left(\boldsymbol{\beta}_0^\top\mathbf{X}(t_k)\right)\lambda_{0k}}\right].
$$

Similarly to the arguments to prove the consistency of $\hat{\mathbf{\Sigma}}$, we can show $\hat{\mathbf{\Omega}}$, the empirical counterpart of $\mathbf{\Omega}$, converges to $\mathbf{\Omega}$ in probability.

Last, we prove that $\hat{\mathbf{V}}$ converges to $\mathbf{\Sigma}$ in probability and is asymptotically equal to $\hat{\mathbf{\Sigma}}$. We begin from expressing $\partial \mathrm{p}\ell_i(\boldsymbol{\beta}|O)/\partial\boldsymbol{\beta}$ in the context of the EM algorithm using equation (5) in Oakes [30]. We can write $\mathrm{p}\ell_i(\boldsymbol{\beta}|O) = \ell_i(\boldsymbol{\beta},\hat{\Lambda}_{\boldsymbol{\beta}}|O)$. Equation (5) in Oakes [30] shows that the derivatives of $\ell_i(\boldsymbol{\beta},\Lambda|O)$ with respect to (w.r.t.) $\boldsymbol{\theta}$ is the derivative of $E_{C|O,\boldsymbol{\theta}'}[\ell_i(\boldsymbol{\beta},\lambda|C)]$ w.r.t. $\boldsymbol{\theta}$ given that $\boldsymbol{\theta}'=\boldsymbol{\theta}$, where $E_{C|O,\boldsymbol{\theta}}$ denotes the conditional expectation the complete data C given the observed data O with parameter $\boldsymbol{\theta}$. Apply the chain rule and (5) in Oakes [30] to $\partial \mathrm{p}\ell_i(\boldsymbol{\beta}|O)/\partial\boldsymbol{\beta}$, it follows from $\hat{\mathbf{x}}_k = -\,d\log(\hat{\lambda}_k(\boldsymbol{\beta}))/d\boldsymbol{\beta}\big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}$ that $\partial \mathrm{p}\ell_i(\boldsymbol{\beta}|O)/\partial\boldsymbol{\beta}\big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} = \tilde{m}^{EM}[\hat{\boldsymbol{\theta}};\hat{\mathbf{x}}]$. Since $\hat{\mathbf{\Sigma}} = \mathbb{P}_n(\tilde{m}^{EM}[\hat{\boldsymbol{\theta}};\hat{\mathbf{x}}])^{\otimes 2}$, $\hat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^n\{\partial \mathrm{p}\ell_i(\boldsymbol{\beta}|O)/\partial\boldsymbol{\beta}|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}\}^{\otimes 2}$. Finally, by the definition of $\hat{\mathbf{V}}$ and Taylor's expansion, $\hat{\mathbf{V}} = \hat{\mathbf{\Sigma}} + O_p\left(h_n^2\right)$. Given $h_n = O(n^{-1/2})$, $\hat{\mathbf{V}} = \hat{\mathbf{\Sigma}} + O_p\left(n^{-1}\right)$, implying that $\hat{\mathbf{V}}$ and $\hat{\mathbf{\Sigma}}$ are asymptotically equal in probability. Thus, $\hat{\mathbf{V}}$ converges to $\mathbf{\Sigma}$ in probability due to the convergence of $\hat{\mathbf{\Sigma}}$.

8.4.2. The proof of Lemma 1

Define a class $\Phi_\eta^* = \{\varrho(\boldsymbol{\theta},h_3) = -\sum_{j=1}^J \Delta N_{\cdot j}\|D_j^1(\boldsymbol{\theta},\mathbf{X}-h_3)/D_j^0(\boldsymbol{\theta})\|_2^2, d(\boldsymbol{\theta},\boldsymbol{\theta}_0) < \eta, \boldsymbol{\theta}\in\mathcal{B}\times\mathcal{F}, h_3\in\mathcal{H}_3\}$. It follows that $\bar{\mathbf{x}}\in\mathcal{H}_3$ since $h_1\bar{\mathbf{x}}(t)\in\mathcal{H}_2$ for any $h_1\in\mathcal{H}_1$ from the proof of Theorem 3.3. Given model (2), we have that $\mathbf{P}\varrho(\boldsymbol{\theta}_0,h_3) = -\mathbf{P}\sum_{j=1}^J\|D_j^1(\boldsymbol{\theta}_0,\mathbf{X}-h_3)\|_2^2/D_j^0(\boldsymbol{\theta}_0)$ and $\mathbf{P}\varrho(\boldsymbol{\theta}_0,\bar{\mathbf{x}}) = -\mathbf{P}\sum_{j=1}^J\left\|D_j(\boldsymbol{\theta}_0)\right\|_2^2/D_j^0(\boldsymbol{\theta}_0)$. We next verify three conditions of Theorem 2.12 (i) in Kosorok [18] and then prove the convergence in probability.

Step 1. First, we establish that $\liminf_{h_3\to\bar{\mathbf{x}}}\mathbf{P}\varrho(\boldsymbol{\theta}_0,h_3)\ge\mathbf{P}\varrho(\boldsymbol{\theta}_0,\bar{\mathbf{x}})$ implies $\tilde{d}(\bar{\mathbf{x}},h_3)\to 0$ for $h_3\in\mathcal{H}_3$. It easily follows that for any $h_3\in\mathcal{H}_3$, $\mathbf{P}\varrho(\boldsymbol{\theta}_0,\bar{\mathbf{x}})-\mathbf{P}\varrho(\boldsymbol{\theta}_0,h_3) = \mathbf{P}\sum_{j=1}^J\{D_j^1(\boldsymbol{\theta}_0,\mathbf{X}-h_3)+D_j(\boldsymbol{\theta}_0)\}^\top D_j^1(\boldsymbol{\theta}_0,\bar{\mathbf{x}}-h_3)/D_j^0(\boldsymbol{\theta}_0)$. Since each element of $\bar{\mathbf{x}}(t)-h_3(t)$ belongs to $\mathcal{H}_2$ for any $h_3\in\mathcal{H}_3$, it follows from the definition of $\bar{\mathbf{x}}$ based on (11) that $\mathbf{P}\sum_{j=1}^J D_j(\boldsymbol{\theta}_0)^\top D_j^1(\boldsymbol{\theta}_0,\bar{\mathbf{x}}-h_3)/D_j^0(\boldsymbol{\theta}_0) = 0$. Subtracting the last equation from $\mathbf{P}\varrho(\boldsymbol{\theta}_0,\bar{\mathbf{x}})-\mathbf{P}\varrho(\boldsymbol{\theta}_0,h_3)$ twice, we have $\mathbf{P}\varrho(\boldsymbol{\theta}_0,\bar{\mathbf{x}})-\mathbf{P}\varrho(\boldsymbol{\theta}_0,h_3) = \mathbf{P}\sum_{j=1}^J\|D_j^1(\boldsymbol{\theta}_0,\bar{\mathbf{x}}-h_3)\|_2^2/D_j^0(\boldsymbol{\theta}_0)$. It then follows from (C1) and (C2) that $\mathbf{P}\varrho(\boldsymbol{\theta}_0,\bar{\mathbf{x}})-\mathbf{P}\varrho(\boldsymbol{\theta}_0,h_3)\gtrsim\tilde{d}(\bar{\mathbf{x}},h_3)^2$. Taking the limsup on both sides, we have $\liminf_{h_3\to\bar{\mathbf{x}}}\mathbf{P}\varrho(\boldsymbol{\theta}_0,h_3)\ge\mathbf{P}\varrho(\boldsymbol{\theta}_0,\bar{\mathbf{x}})$ implies $\limsup_{h_3\to\bar{\mathbf{x}}}\tilde{d}(\bar{\mathbf{x}},h_3)\le 0$. However, $\tilde{d}(\bar{\mathbf{x}},h_3)\ge 0$ by its definition and thus $\liminf_{h_3\to\boldsymbol{\theta}_0}\mathbf{P}\varrho(\boldsymbol{\theta}_0,h_3)\ge\mathbf{P}\varrho(\boldsymbol{\theta}_0,\bar{\mathbf{x}})$ implies $\tilde{d}(\bar{\mathbf{x}},h_3)\to 0$.

Step 2. We next prove that $\mathbb{P}_n\varrho(\hat{\boldsymbol{\theta}},\hat{\bar{\mathbf{x}}})\ge\sup_{h_3\in\mathcal{H}_3}\mathbb{P}_n\varrho(\hat{\boldsymbol{\theta}},h_3)-o_p(1)$. It is not hard to show that $\mathbb{P}_n\varrho(\hat{\boldsymbol{\theta}},h_3)$ is concave in h_3 in terms of the Gâteaux-derivative. Then, there exists h_3^* that maximizes $\mathbb{P}_n\varrho(\hat{\boldsymbol{\theta}},h_3)$ over $\mathcal{H}_3$. By the concavity of $\mathbb{P}_n\varrho(\hat{\boldsymbol{\theta}},h_3)$, for $\epsilon\in[0,1]$, $\{\mathbb{P}_n\varrho(\hat{\boldsymbol{\theta}},\hat{\bar{\mathbf{x}}}+\epsilon(h_3^*-\hat{\bar{\mathbf{x}}}))-\mathbb{P}_n\varrho(\hat{\boldsymbol{\theta}},\hat{\bar{\mathbf{x}}})\}/\epsilon\ge\mathbb{P}_n\varrho(\hat{\boldsymbol{\theta}},h_3^*)-\mathbb{P}_n\varrho(\hat{\boldsymbol{\theta}},\hat{\bar{\mathbf{x}}})$. Furthermore, discretizing $\Lambda(t)$ in $\mathbb{P}_n\varrho(\hat{\boldsymbol{\theta}},h_3)$ by λ on $\{t_k\}_{k=1}^K$ and letting $\epsilon\to 0$ we obtain that

$$\frac{2}{n}\sum_{i=1}^n\sum_{k=1}^K\left\{\frac{\sum_{l\in A_{ik}}\Delta_{il}\exp\left(\hat{\boldsymbol{\beta}}^\top\mathbf{X}_i(t_l)\right)\hat{\lambda}_l(\hat{\boldsymbol{\beta}})\left(\mathbf{X}_i(t_l)-\hat{\bar{\mathbf{x}}}(t_l)\right)}{\sum_{l\in A_{ik}}\Delta_{il}\exp\left(\hat{\boldsymbol{\beta}}^\top\mathbf{X}_i(t_l)\right)\hat{\lambda}_l(\hat{\boldsymbol{\beta}})}\right\}^\top \times\left(h_3^*(t_k)-\hat{\bar{\mathbf{x}}}(t_k)\right)\hat{E}\left[W_{ik}|\hat{\boldsymbol{\beta}},\hat{\lambda}(\hat{\boldsymbol{\beta}})\right]\ge\mathbb{P}_n\varrho\left(\hat{\boldsymbol{\theta}},h_3^*\right)-\mathbb{P}_n\varrho\left(\hat{\boldsymbol{\theta}},\hat{\bar{\mathbf{x}}}\right). \quad (15)$$

We shall prove that the left-hand side of (15) is $o_p(1)$ by using the properties of the profile log-likelihood function.

The iterative relationship in (7) implies that, for any $\boldsymbol{\beta}$ and $h_{2k} \in \mathcal{H}_2$,

$$\sum_{i=1}^{n}\sum_{k=1}^{K} h_{2k}\Delta_{ik}\left\{\hat{E}\left[W_{ik}|\boldsymbol{\beta},\hat{\lambda}(\boldsymbol{\beta})\right] - \exp\left(\boldsymbol{\beta}^{\top}\mathbf{X}_i\left(t_k\right)\right)\hat{\lambda}_k(\boldsymbol{\beta})\right\} = 0. \tag{16}$$

Thus, recalling the definition of $\hat{E}\left[W_{ik}|\boldsymbol{\beta},\hat{\lambda}(\boldsymbol{\beta})\right]$ in (6) and taking differentiation of (16) w.r.t. $\boldsymbol{\beta}$, after some algebra, we obtain that, for any $\boldsymbol{\beta}$ and $h_{2k} \in \mathcal{H}_2$,

$$\begin{aligned}
&\frac{2}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\Delta_{ik}\left\{\mathbf{X}_i\left(t_k\right) - \hat{\bar{\mathbf{x}}}(t_k)\right\}\left\{\hat{E}\left[W_{ik}|\hat{\boldsymbol{\beta}},\hat{\lambda}(\hat{\boldsymbol{\beta}})\right] - \exp\left(\hat{\boldsymbol{\beta}}^{\top}\mathbf{X}_i\left(t_k\right)\right)\hat{\lambda}_k(\hat{\boldsymbol{\beta}})\right\} h_{2k}\\
&= \frac{2}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} h_{2k}\Delta_{ik}\frac{\sum_{l\in A_{ik}}\Delta_{il}\exp\left(\boldsymbol{\beta}_0^{\top}\mathbf{X}_i\left(t_l\right)\right)\hat{\lambda}_l(\hat{\boldsymbol{\beta}})\left(\mathbf{X}_i\left(t_l\right) - \hat{\bar{\mathbf{x}}}(t_k)\right)}{\sum_{l\in A_{ik}}\Delta_{il}\exp\left(\hat{\boldsymbol{\beta}}^{\top}\mathbf{X}_i\left(t_l\right)\right)\hat{\lambda}_l(\hat{\boldsymbol{\beta}})}\hat{E}\left[W_{ik}|\hat{\boldsymbol{\beta}},\hat{\lambda}(\hat{\boldsymbol{\beta}})\right].
\end{aligned} \tag{17}$$

Since $E\left[\hat{E}\left[W_{ik}|\boldsymbol{\beta}_0,\Lambda_0\right]|\mathbf{X}_i,\{t_k\}_{k=1}^{K},\Delta_{ik}\right] = \exp\left(\boldsymbol{\beta}_0^{\top}\mathbf{X}_i\left(t_k\right)\right)\lambda_{0k}$ under model (2), we have the LHS of (17) converges to zero in probability for any $h_2 \in \mathcal{H}_2$ and $h_3 \in \mathcal{H}_3$ by the law of large number and the continuous mapping theorem. In other words, the RHS of (17) is $o_p(1)$ for any $h_2 \in \mathcal{H}_2$. It also easily follows that LHS of (15) coincides with the RHS of (17) by letting h_2 be the elements of $h_3^*(t_k) - \hat{\bar{\mathbf{x}}}(t_k)$. Consequently, $\mathbb{P}_n\varrho\left(\hat{\boldsymbol{\theta}},\hat{\bar{\mathbf{x}}}\right) \geq \mathbb{P}_n\varrho\left(\hat{\boldsymbol{\theta}},h_3^*\right) - o_p(1)$.

Step 3. Given that $h_3 \in \mathcal{H}_3$, using similar arguments in the proof of Theorem 3.3, we can show that Φ_η^* is $\mathbf{P}$-Donsker and hence Glivenko-Cantelli.

In sum, the conclusions in Steps 1-3 verify all the conditions of conclusion (i) of Theorem 2.12 in Kosorok [18]. It then follows from Theorem 2.12 in Kosorok [18], $\tilde{d}\left(\hat{\bar{\mathbf{x}}},\bar{\mathbf{x}}\right) \to 0$ in probability as $n \to \infty$.

Last, we prove the second part of Lemma 1. Given Condtion (E1), there exists $\mathbf{X}_{\cdot j}$ such that $\mathbf{X}(t) = \mathbf{X}_{\cdot j}$ for $t \in (T_{\cdot(j-1)},T_{\cdot j}]$. Then, for any $h_2 \in \mathcal{H}_2$, (11) reduces to

$$\begin{aligned}
0 =&\mathbf{P}\sum_{j=1}^{J}\int_{T_{\cdot(j-1)}}^{T_{\cdot j}}\exp\left(\boldsymbol{\beta}_0^{\top}\mathbf{X}(t)\right)\left\{\mathbf{X}(t) - h_3(t)\right\}d\Lambda_0\left(t\right)\frac{\exp\left(\boldsymbol{\beta}_0^{\top}\mathbf{X}_{\cdot j}\right)\int_{T_{\cdot(j-1)}}^{T_{\cdot j}}h_2(t)d\Lambda_0\left(t\right)}{\exp\left(\boldsymbol{\beta}_0^{\top}\mathbf{X}_{\cdot j}\right)\int_{T_{\cdot(j-1)}}^{T_{\cdot j}}d\Lambda_0\left(t\right)}\\
=&E\left[\sum_{j=1}^{J}\tilde{h}_{2j}\int_{T_{\cdot(j-1)}}^{T_{\cdot j}}E\left[\Delta(t)\exp\left(\boldsymbol{\beta}_0^{\top}\mathbf{X}(t)\right)\left\{\mathbf{X}(t) - h_3(t)\right\}|T_{\cdot(j-1)},T_{\cdot j},J\right]d\Lambda_0\left(t\right)\right],
\end{aligned} \tag{18}$$

where $\tilde{h}_{2j} = \int_{T_{\cdot(j-1)}}^{T_{\cdot j}}h_2(t)d\Lambda_0\left(t\right)/\int_{T_{\cdot(j-1)}}^{T_{\cdot j}}d\Lambda_0\left(t\right)$. It follows that $h_3(t) = s^{(1)}(t,\boldsymbol{\beta}_0)/s^{(0)}(t,\boldsymbol{\beta}_0)$. implies (18) becomes zero for any $\Lambda \in \mathcal{F}$. By the uniqueness of $\bar{\mathbf{x}}$, $\bar{\mathbf{x}}$ must be $s^{(1)}(t,\boldsymbol{\beta}_0)/s^{(0)}(t,\boldsymbol{\beta}_0)$ under Condition (E1).

Under conditions (C1) and (C2), it is not hard to show that $\{\mathbf{X}\left(t\right)^{\otimes u}\Delta\left(t\right)\exp(\boldsymbol{\beta}^{\top}\mathbf{X}\left(t\right)); t \in [\tau_0,\tau],\boldsymbol{\beta} \in \mathcal{B}\}$ is a $\mathbf{P}$-Donsker class and Glivenko-Cantelli class for $u = 0,1$. Therefore, for $u = 0,1$, $\sup_{t\in[\tau_0,\tau],\boldsymbol{\beta}\in\mathcal{B}}\sqrt{n}\left\|S^{(u)}\left(t,\boldsymbol{\beta}\right) - s^{(u)}\left(t,\boldsymbol{\beta}\right)\right\|_{u+1} \to 0$ in probability as $n \to \infty$. By the continuous mapping theorem, $\bar{\mathbf{X}}(t;\hat{\boldsymbol{\beta}})$ converges to $s^{(1)}(t,\boldsymbol{\beta}_0)/s^{(0)}(t,\boldsymbol{\beta}_0)$ uniformly for $t \in [\tau_0,\tau]$. Then, the dominated convergence theorem yields the convergence of $\bar{\mathbf{X}}(t;\hat{\boldsymbol{\beta}})$ in $\hat{d}$ and completes the second part proof of Lemma 1.

Acknowledgments

We would like to express our sincere appreciation to the Editor, Dr. Podolskij, the Associate Editor and two anonymous reviewers for their invaluable feedback and constructive suggestions that significantly improved the paper. This work was partially supported by NIH grants HL095086 and AA026969.

Supplementary Material

Supplement to "A new robust approach for regression analysis of panel count data with time-varying covariates" (DOI: 10.3150/23-BEJ1713SUPP; .pdf). The Supplement Material [32] contains some additional simulation and real data analysis results and the proofs of Proposition 1 as well as the detailed derivation of $\mathbf{\Omega}$ in terms of EM algorithm.

References

[1] Andersen, P.K. and Gill, R.D. (1982). Cox's regression model for counting processes: A large sample study. *Ann. Statist.* **10** 1100–1120. MR0673646 https://doi.org/10.1214/aos/1176345976

[2] Balakrishnan, N. and Zhao, X. (2009). New multi-sample nonparametric tests for panel count data. *Ann. Statist.* **37** 1112–1149. MR2509069 https://doi.org/10.1214/08-AOS599

[3] Balakrishnan, N. and Zhao, X. (2011). A class of multi-sample nonparametric tests for panel count data. *Ann. Inst. Statist. Math.* **63** 135–156. MR2748938 https://doi.org/10.1007/s10463-008-0209-x

[4] Batteiger, B.E., Tu, W., Ofner, S., Van Der Pol, B., Stothard, D.R., Orr, D.P., Katz, B.P. and Fortenberry, J.D. (2010). Repeated Chlamydia trachomatis genital infections in adolescent women. *J. Infect. Dis.* **201** 42–51. https://doi.org/10.1086/648734

[5] Ciarlet, P.G. (2013). *Linear and Nonlinear Functional Analysis with Applications*. Philadelphia, PA: SIAM. MR3136903

[6] Cook, R.J. and Lawless, J.F. (2007). *The Statistical Analysis of Recurrent Events*. *Statistics for Biology and Health*. New York: Springer. MR3822124

[7] Deng, S. (2013). Semiparametric regression analysis of panel count data with time-dependent covariates and informative observation and censoring times. *Comm. Statist. Theory Methods* **42** 4170–4183. MR3170990 https://doi.org/10.1080/03610926.2011.642922

[8] Du, M. and Sun, J. (2021). Statistical analysis of interval-censored failure time data. *Chinese J. Appl. Probab. Statist.* **37** 627–654. MR4410679 https://doi.org/10.3969/j.issn.1001-4268.2021.06.006

[9] Du, M. and Sun, J. (2022). Variable selection for interval-censored failure time data. *Int. Stat. Rev.* **90** 193–215. MR4481432 https://doi.org/10.1111/insr.12480

[10] Du, M. and Zhao, H. (2021). A unified approach to variable selection for Cox's proportional hazards model with interval-censored failure time data. *Stat. Methods Med. Res.* **30** 1833–1849. MR4307785 https://doi.org/10.1177/09622802211009259

[11] Gao, F. and Chan, K.C.G. (2019). Semiparametric regression analysis of length-biased interval-censored data. *Biometrics* **75** 121–132. MR3953713 https://doi.org/10.1111/biom.12970

[12] Ge, L., Zhu, L. and Sun, J. (2021). Regression analysis of mixed panel count data with informative indicator processes. *Stat. Med.* **40** 1262–1271. MR4384374 https://doi.org/10.1002/sim.8839

[13] Ghosh, P. and Tu, W. (2009). Assessing sexual attitudes and behaviors of young women: A joint model with nonlinear time effects, time varying covariates, and dropouts. *J. Amer. Statist. Assoc.* **104** 474–485. MR2751432 https://doi.org/10.1198/jasa.2009.0013

[14] Hu, X.J., Lagakos, S.W. and Lockhart, R.A. (2009). Marginal analysis of panel counts through estimating functions. *Biometrika* **96** 445–456. MR2507154 https://doi.org/10.1093/biomet/asp010

[15] Hu, X.J., Sun, J. and Wei, L.-J. (2003). Regression parameter estimation from panel counts. *Scand. J. Stat.* **30** 25–43. MR1963891 https://doi.org/10.1111/1467-9469.00316

[16] Hua, L., Zhang, Y. and Tu, W. (2014). A spline-based semiparametric sieve likelihood method for over-dispersed panel count data. *Canad. J. Statist.* **42** 217–245. MR3208337 https://doi.org/10.1002/cjs.11208
[17] Jiang, H., Su, W. and Zhao, X. (2020). Robust estimation for panel count data with informative observation times and censoring times. *Lifetime Data Anal.* **26** 65–84. MR4051838 https://doi.org/10.1007/s10985-018-09457-7
[18] Kosorok, M.R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. *Springer Series in Statistics*. New York: Springer. MR2724368 https://doi.org/10.1007/978-0-387-74978-5
[19] Li, N., Sun, L. and Sun, J. (2010). Semiparametric transformation models for panel count data with dependent observation process. *Stat. Biosci.* **2** 191–210. https://doi.org/10.1007/s12561-010-9029-7
[20] Li, N., Zhao, H. and Sun, J. (2013). Semiparametric transformation models for panel count data with correlated observation and follow-up times. *Stat. Med.* **32** 3039–3054. MR3073834 https://doi.org/10.1002/sim.5724
[21] Li, S., Hu, T., Zhao, S. and Sun, J. (2020). Regression analysis of multivariate current status data with semiparametric transformation frailty models. *Statist. Sinica* **30** 1117–1134. MR4214176
[22] Li, S., Hu, T., Zhao, X. and Sun, J. (2019). A class of semiparametric transformation cure models for interval-censored failure time data. *Comput. Statist. Data Anal.* **133** 153–165. MR3926472 https://doi.org/10.1016/j.csda.2018.09.008
[23] Li, Z., Liu, H. and Tu, W. (2015). A sexually transmitted infection screening algorithm based on semiparametric regression models. *Stat. Med.* **34** 2844–2857. MR3375984 https://doi.org/10.1002/sim.6515
[24] Lin, D.Y., Wei, L.J., Yang, I. and Ying, Z. (2000). Semiparametric regression for the mean and rate functions of recurrent events. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **62** 711–730. MR1796287 https://doi.org/10.1111/1467-9868.00259
[25] Liu, L., Su, W., Yin, G., Zhao, X. and Zhang, Y. (2022). Nonparametric inference for reversed mean models with panel count data. *Bernoulli* **28** 2968–2997. MR4474569 https://doi.org/10.3150/21-bej1444
[26] Louis, T.A. (1982). Finding the observed information matrix when using the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **44** 226–233. MR0676213
[27] Mao, L. (2020). A unified approach to the calculation of information operators in semiparametric models. *Biometrika* **107** 983–995. MR4186500 https://doi.org/10.1093/biomet/asaa037
[28] Murphy, S.A., Rossini, A.J. and van der Vaart, A.W. (1997). Maximum likelihood estimation in the proportional odds model. *J. Amer. Statist. Assoc.* **92** 968–976. MR1482127 https://doi.org/10.2307/2965560
[29] Murphy, S.A. and van der Vaart, A.W. (2000). On profile likelihood. *J. Amer. Statist. Assoc.* **95** 449–465. With comments and a rejoinder by the authors. MR1803168 https://doi.org/10.2307/2669386
[30] Oakes, D. (1999). Direct calculation of the information matrix via the EM algorithm. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **61** 479–482. MR1680298 https://doi.org/10.1111/1467-9868.00188
[31] Su, Y.-R. and Wang, J.-L. (2016). Semiparametric efficient estimation for shared-frailty models with doubly-censored clustered data. *Ann. Statist.* **44** 1298–1331. MR3485961 https://doi.org/10.1214/15-AOS1406
[32] Sun, D., Guo, Y., Li, Y., Tu, W. and Sun, J. (2024). Supplement to "A robust approach for regression analysis of panel count data with time-varying covariates." https://doi.org/10.3150/23-BEJ1713SUPP
[33] Sun, J. and Wei, L.J. (2000). Regression analysis of panel count data with covariate-dependent observation and censoring times. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **62** 293–302. MR1749540 https://doi.org/10.1111/1467-9868.00232
[34] Sun, J. and Zhao, X. (2013). *Statistical Analysis of Panel Count Data*. *Statistics for Biology and Health*. New York: Springer. MR3136574 https://doi.org/10.1007/978-1-4614-8715-9
[35] Torrone, E., Papp, J. and Weinstock, H. (2014). Prevalence of Chlamydia trachomatis genital infection among persons aged 14-39 years–United States, 2007-2012. *Morb. Mort. Wkly. Rep.* **63** 834–838.
[36] Tu, W., Batteiger, B.E., Wiehe, S., Ofner, M.S.S., Van Der Pol, B., Katz, B.P., Orr, D.P. and Fortenberry, J.D. (2009). Time from first intercourse to first sexually transmitted infection diagnosis among adolescent women. *Arch. Pediatr. Adolesc. Med.* **163** 1106–1111. https://doi.org/10.1001/archpediatrics.2009.203
[37] van der Vaart, A. (2002). Semiparametric statistics. In *Lectures on Probability Theory and Statistics (Saint-Flour, 1999)*. *Lecture Notes in Math.* **1781** 331–457. Berlin: Springer. MR1915446
[38] van der Vaart, A.W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge: Cambridge Univ. Press. MR1652247 https://doi.org/10.1017/CBO9780511802256

[39] van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes*. *Springer Series in Statistics*. New York: Springer. With applications to statistics. MR1385671 https://doi.org/10.1007/978-1-4757-2545-2
[40] van der Vaart, A.W. and Wellner, J.A. (2007). Empirical processes indexed by estimated functions. In *Asymptotics: Particles, Processes and Inverse Problems* (E.A. Cator, G. Jongbloed, C. Kraaikamp, H.P. Lopuhaä and J.A. Wellner, eds.). *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **55** 234–252. Beachwood, OH: IMS. MR2459942 https://doi.org/10.1214/074921707000000382
[41] Wang, Y. and Yu, Z. (2021). A kernel regression model for panel count data with time-varying coefficients. *Statist. Sinica* **31** 1707–1725. MR4328837 https://doi.org/10.5705/ss.202019.0220
[42] Wellner, J.A. and Zhang, Y. (2000). Two estimators of the mean of a counting process with panel count data. *Ann. Statist.* **28** 779–814. MR1792787 https://doi.org/10.1214/aos/1015951998
[43] Wellner, J.A. and Zhang, Y. (2007). Two likelihood-based semiparametric estimation methods for panel count data with covariates. *Ann. Statist.* **35** 2106–2142. MR2363965 https://doi.org/10.1214/009053607000000181
[44] Yu, G., Zhu, L., Li, Y., Sun, J. and Robison, L.L. (2017). Regression analysis of mixed panel count data with dependent terminal events. *Stat. Med.* **36** 1669–1680. MR3631986 https://doi.org/10.1002/sim.7217
[45] Zeng, D. and Lin, D.Y. (2021). Maximum likelihood estimation for semiparametric regression models with panel count data. *Biometrika* **108** 947–963. MR4341361 https://doi.org/10.1093/biomet/asaa091
[46] Zeng, D., Mao, L. and Lin, D.Y. (2016). Maximum likelihood estimation for semiparametric transformation models with interval-censored data. *Biometrika* **103** 253–271. MR3509885 https://doi.org/10.1093/biomet/asw013
[47] Zhao, H., Tu, W. and Yu, Z. (2018). A nonparametric time-varying coefficient model for panel count data. *J. Nonparametr. Stat.* **30** 640–661. MR3843044 https://doi.org/10.1080/10485252.2018.1458982
[48] Zhao, X. and Zhang, Y. (2017). Asymptotic normality of nonparametric M-estimators with applications to hypothesis testing for panel count data. *Statist. Sinica* **27** 931–950. MR3675037
[49] Zhu, L., Tong, X., Sun, J., Chen, M., Srivastava, D.K., Leisenring, W. and Robison, L.L. (2014). Regression analysis of mixed recurrent-event and panel-count data. *Biostatistics* **15** 555–568. https://doi.org/10.1093/biostatistics/kxu009

Received November 2021 and revised July 2023