Joint Analysis of Longitudinal Data and Informative Observation Times with Time-Dependent Random Effects

Yang Li, Xin He, Haiying Wang, and Jianguo Sun

Abstract Longitudinal data occur in many fields such as the medical follow-up studies that involve repeated measurements. For their analysis, most existing approaches assume that the observation or follow-up times are independent of the response process either completely or given some covariates. In practice, it is apparent that this may not be true. We present a joint analysis approach that allows the possible mutual correlations that can be characterized by time-dependent random effects. Estimating equations are developed for the parameter estimation and the resulting estimators are shown to be consistent and asymptotically normal.

1 Introduction

Longitudinal data occur in many fields such as the medical follow-up studies that involve repeated measurements. In these situations, study subjects are generally observed only at discrete times. Therefore, for the analysis of longitudinal data, two processes need to be considered: one is the response process, which is usually of the primary interest but not continuously observable; the other one is the observation process, which is nuisance but gives rise to the discrete times when the responses are observed.

Y. Li (✉)
Department of Mathematics and Statistics, University of North Carolina
at Charlotte, Charlotte, NC, USA
e-mail: Y.Li@uncc.edu

X. He
Department of Epidemiology and Biostatistics, University of Maryland, College Park, MD, USA
e-mail: xinhe@umd.edu

H. Wang
Department of Mathematics and Statistics, University of New Hampshire, Durham, NH, USA
e-mail: HaiYing.Wang@unh.edu

J. Sun
Department of Statistics, University of Missouri, Columbia, MO, USA
e-mail: sunj@missouri.edu

Z. Jin et al. (eds.), *New Developments in Statistical Modeling, Inference and Application*, ICSA Book Series in Statistics, DOI 10.1007/978-3-319-42571-9_2

An extensive literature exists for the analysis of longitudinal data. Sun and Kalbfleisch (1995) and Wellner and Zhang (2000) investigated nonparametric estimation of the mean function when the response process is a counting process. Cheng and Wei (2000), Sun and Wei (2000), Zhang (2002) and Wellner and Zhang (2007) developed some semiparametric approaches for regression analysis under the proportional means models. However, with respect to the observation process, most existing approaches assume that the observation times are independent of the underlying response process either completely or given some covariates. For the analysis with a correlated observation process, there is limited work and most of them assume independent censoring or require some restrictive conditions such as the Poisson assumption or specified correlation structure for dependence (He et al. 2009; Huang et al. 2006; Kim et al. 2012; Li et al. 2013; Sun et al. 2007; Zhao and Tong 2011; Zhao et al. 2013; Zhou et al. 2013).

In many situations, however, the response process, the observation and censoring times may be mutually correlated. In addition, such correlations may be time-dependent. For instance, both the observation times and longitudinal responses may depend on the stage of disease progression. Their correlation may change over time and so are their correlations with the follow-up times. He et al. (2009) considered such correlations in shared frailty models. However, their method requires the assumptions that the underlying random effect is normally distributed and the observation process is a nonhomogeneous Poisson process. Also all correlations between the three processes are assumed to be fixed over time. Zhao et al. (2013) proposed a robust estimation procedure and relaxed the Poisson assumption required in He et al. (2009). However, the follow-up times are assumed to be independent from covariates, responses and observation times; and the possible correlations between responses and observation times are time-independent. More recently, Sun et al. (2012) presented a joint model with time-dependent correlations between the response process, the observation times and a terminal event, where the random effect associated with the terminal event is fixed over time and follow a specified distribution. In practice, however, such conditions may not hold or be difficult to check when informative censoring involves.

We consider regression analysis of longitudinal data when the underlying response process, the observation and censoring times are mutually correlated and none of the correlations is restricted by specified forms or distributions. A general estimation approach is proposed. The remainder of this chapter is organized as follows: In Sect. 2, we introduce the notation and present the model. Section 3 presents the estimation procedure and establishes the asymptotic properties of the resulting estimators. In Sect. 4, a simulation study is performed to evaluate the finite sample properties of the proposed estimators. Some concluding remarks are given in Sect. 5.

2 Notation and Models

Consider a longitudinal study in which the response process of interest is observed only at some discrete sampling time points. For each subject i, $i = 1, \cdots, n$, let $N_i(t)$ be the observation process, which gives the cumulative number of observation times up to time t. In practice, one observes $\widetilde{N}_i(t) = N_i(t \wedge C_i)$ where $a \wedge b = min(a, b)$ and C_i denotes the censoring or follow-up time. Let $Y_i(t)$ denote the response process, which gives the response of interest at time t but is observed only at m_i discrete observation times $\{T_{i,1}, \cdots, T_{i,m_i}\}$ when $\widetilde{N}_i(t)$ has jumps. Suppose that there exists a p-dimensional vector of covariates denoted by $\mathbf{Z}_i$, which will be assumed to be time-independent.

In the following, we model the correlation between $Y_i(t)$, $N_i(t)$ and C_i through an unobserved random vector $\mathbf{b}_i(t) = (b_{1i}(t), b_{2i}(t), b_{3i}(t))'$, which could be time-dependent. Define $\mathscr{B}_{it} = \{\mathbf{b}_i(s), s \le t\}$. It will be assumed that the $\mathbf{b}_i(t)$'s are independent and identically distributed, $\mathscr{B}_{it}$ is independent of $\mathbf{Z}_i$, and given $\mathbf{Z}_i$ and $\mathscr{B}_{it}$, C_i, $N_i(t)$ and $Y_i(t)$ are mutually independent. To be specific, the mean function of $Y_i(t)$ is assumed to follow the proportional means model

$$E\{Y_i(t)|\mathbf{Z}_i, \mathbf{b}_i(t)\} = \Lambda_0(t) \exp\{\beta'\mathbf{Z}_i + b_{1i}(t)\}, \tag{1}$$

where $\Lambda_0(t)$ is an unknown baseline mean function and β denotes a vector of p-dimensional regression coefficients. When $b_{1i}(t) = 0$ meaning that $Y_i(t)$ is independent of both $N_i(t)$ of C_i, model (1) has been considered extensively by Cheng and Wei (2000), Sun and Wei (2000), Zhang (2002) and Hu et al. (2003) among others. When $b_{1i}(t)$ is time-independent, model (1) is equivalent to model (3) considered in Zhao et al. (2013). In general, $b_{1i}(t)$ is unknown and may follow an arbitrary distribution.

The observation process $N_i(t)$ follows the proportional rates model

$$E\{dN_i(t)|\mathbf{Z}_i, \mathbf{b}_i(t)\} = \exp\{\gamma'\mathbf{Z}_i + b_{2i}(t)\}d\mu_0(t)\,, \tag{2}$$

where γ is a vector of unknown parameters and $d\mu_0(t)$ is an unknown baseline rate function. For the $C_i's$, motivated by the additive hazards models that have been commonly used in survival analysis (Kalbfleisch and Prentice 2002; Lin and Ying 2001; Zhang et al. 2005), we consider the additive hazards model. That is, the hazard $\lambda_i(t|\mathbf{Z}_i, \mathbf{b}_i(t))$ of C_i, defined as the rate of observing C_i at time t provided that C_i is no larger than t, is given by

$$\lambda_i(t|\mathbf{Z}_i, \mathbf{b}_i(t)) = \lambda_0(t) + \xi'\mathbf{Z}_i + b_{3i}(t)\,. \tag{3}$$

Here $\lambda_0(t)$ is an unknown baseline hazard function and ξ denotes the effect of covariates on the hazard function of $C_i's$. Note that instead of model (3), one may consider the proportional hazards model. As pointed out by Lin et al. (1998) and others, the additive model (3) can be more plausible than the proportional hazards model in many applications. Related applications and model-checking techniques of model (3) can be found in Yuen and Burke (1997), Kim and Lee (1998), Ghosh (2003) and Gandy and Jensen (2005) among others.

In the above, models (1)–(3) can be viewed as natural generalizations of some existing and commonly used models. In fact, when any of the $b_{ki}(t)$'s ($k = 1, 2, 3$) is zero or independent from other $b_{ji}(t)$'s ($j = 1, 2, 3$ and $j \neq k$), the corresponding process is independent from the others. Therefore, the proposed joint model also applies to special cases when either the observation or censoring times are noninformative. In general, since the form or distribution of $\mathbf{b}_i(t)$ is arbitrary and completely unspecified, the joint model described above is quite flexible compared to many existing procedures.

Note that in models (1)–(3), for simplicity, we have assumed that the set of covariates that may affect $Y_i(t)$, $N_i(t)$ and C_i is the same. In practice, it is apparent that this may not be the case and actually the estimation procedure proposed below still applies as long as one replaces $\mathbf{Z}_i$ by appropriate covariates. As an alternative, one can define a single and big covariate vector by combining all different covariates together. In the following, we will focus on estimation of regression parameters β along with γ and ξ. For this, it is easy to see that the use of the existing procedures that assume independence could give biased or even misleading results.

3 Estimation Procedure

In this section, we will present an inference procedure for estimation of β which is usually of the primary interest. For this, first note that the counting process $\widetilde{N}_i(t) = N_i(t \wedge C_i)$ jumps by one at time t if and only if $C_i \geq t$ and $dN_i(t) = 1$. Also we have

$$\begin{aligned}
E\{d\widetilde{N}_i(t)|\mathbf{Z}_i\} &= E\{I(t \leq C_i)dN_i(t)|\mathbf{Z}_i\} \\
&= E\left[E\{I(t \leq C_i)dN_i(t)|\mathbf{Z}_i, \mathscr{B}_{it}\}\middle|\mathbf{Z}_i\right] \\
&= E\left[E\{I(t \leq C_i)|\mathbf{Z}_i, \mathscr{B}_{it}\}E\{dN_i(t)|\mathbf{Z}_i, \mathscr{B}_{it}\}\middle|\mathbf{Z}_i\right] \\
&= E\left[exp\{-\Lambda_0^*(t) - B_i(t) - \xi'\mathbf{Z}_i^*(t)\}\exp\{\gamma'\mathbf{Z}_i + b_{2i}(t)\}d\mu_0(t)\middle|\mathbf{Z}_i\right] \\
&= \exp\{\gamma'\mathbf{Z}_i - \xi'\mathbf{Z}_i^*(t)\}d\Lambda_1^*(t), \qquad (4)
\end{aligned}$$

where

$$\Lambda_0^*(t) = \int_0^t \lambda_0(s)ds, \quad B_i(t) = \int_0^t b_{3i}(s)ds, \quad \mathbf{Z}_i^*(t) = \int_0^t \mathbf{Z}_i ds$$

and

$$d\Lambda_1^*(t) = \exp\{-\Lambda_0^*(t)\}E[exp\{b_{2i}(t) - B_i(t)\}]d\mu_0(t).$$

Define

$$dM_i^*(t;\eta) = d\widetilde{N}_i(t) - e^{\eta'\mathbf{X}_i(t)} d\Lambda_1^*(t)$$

and $dM_i^*(t) = dM_i^*(t;\eta_0)$, where $\eta = (\gamma, \quad \xi)'$, $\mathbf{X}_i(t) = (\mathbf{Z}_i, \quad -\mathbf{Z}_i^*(t))'$ and η_0 denotes the true value of η. It can be shown that $M_i^*(t)$ is a mean-zero stochastic process. It follows that the estimators of η and $d\Lambda_1^*(t)$ can be obtained by solving the following two estimating equations

$$U_\eta(\eta) = \sum_{i=1}^n \int_0^\tau \left\{ \mathbf{X}_i(t) - \bar{X}(t;\eta) \right\} d\widetilde{N}_i(t) = 0 \tag{5}$$

and

$$\sum_{i=1}^n \left[d\widetilde{N}_i(t) - e^{\eta'\mathbf{X}_i(t)} d\Lambda_1^*(t) \right] = 0. \tag{6}$$

In the above, τ is the longest follow-up time, $\bar{X}(t;\eta) = S^{(1)}(t;\eta)/S^{(0)}(t;\eta)$ and $S^{(k)}(t;\eta) = n^{-1}\sum_{i=1}^n e^{\eta'\mathbf{X}_i(t)}\mathbf{X}_i(t)^{\otimes k}$ with $a^{\otimes 0} = 1$, $a^{\otimes 1} = a$, $\bar{x}(t) = lim_{n\to\infty}\bar{X}(t;\eta_0)$ and $s^{(k)}(t) = lim_{n\to\infty}S^{(k)}(t;\eta_0)$, $k = 0, 1$.

To estimate β, consider

$$\begin{aligned}
&E\{Y_i(t)d\widetilde{N}_i(t)|\mathbf{Z}_i, \mathscr{B}_{it}\}\\
&\quad = E\{I(t \le C_i)Y_i(t)dN_i(t)|\mathbf{Z}_i, \mathscr{B}_{it}\}\\
&\quad = E\{I(t \le C_i)|\mathbf{Z}_i, \mathscr{B}_{it}\}E\{Y_i(t)|\mathbf{Z}_i, \mathscr{B}_{it}\}E\{dN_i(t)|\mathbf{Z}_i, \mathscr{B}_{it}\}\\
&\quad = exp\{-\Lambda_0^*(t) - B_i(t) - \xi'\mathbf{Z}_i^*(t)\}\\
&\Lambda_0(t)\exp\{\beta'\mathbf{Z}_i + b_{1i}(t)\}\exp\{\gamma'\mathbf{Z}_i + b_{2i}(t)\}d\mu_0(t)\\
&\quad = \exp\{(\beta+\gamma)'\mathbf{Z}_i - \xi'\mathbf{Z}_i^*(t)\}\\
&\qquad \exp\{-\Lambda_0^*(t) + b_{1i}(t) + b_{2i}(t) - B_i(t)\}\Lambda_0(t)d\mu_0(t),
\end{aligned}$$

and therefore

$$E\{Y_i(t)d\widetilde{N}_i(t)|\mathbf{Z}_i\} = \exp\{\beta'\mathbf{Z}_i + \eta'\mathbf{X}_i(t)\}d\Lambda_2^*(t), \tag{7}$$

where

$$d\Lambda_2^*(t) = \exp\{-\Lambda_0^*(t)\}\Lambda_0(t)E[exp\{b_{1i}(t) + b_{2i}(t) - B_i(t)\}]d\mu_0(t).$$

Define

$$dM_i(t;\beta,\eta) = Y_i(t)d\widetilde{N}_i(t) - \exp\{\beta'\mathbf{Z}_i + \eta'\mathbf{X}_i(t)\}d\Lambda_2^*(t)$$

and $dM_i(t) = dM_i(t; \beta_0, \eta_0)$, where β_0 denotes the true value of β. Then $M_i(t)$ is a mean-zero stochastic process. This naturally suggests the following estimating equations to estimate β and $d\Lambda_2^*(t)$:

$$U_\beta(\beta; \hat{\eta}) = \sum_{i=1}^{n} \int_0^\tau W(t)\mathbf{Z}_i \Big[Y_i(t)d\widetilde{N}_i(t) - e^{\beta'\mathbf{Z}_i + \hat{\eta}'\mathbf{X}_i(t)} d\Lambda_2^*(t) \Big] = 0, \qquad (8)$$

and

$$\sum_{i=1}^{n} \Big[Y_i(t)d\widetilde{N}_i(t) - e^{\beta'\mathbf{Z}_i + \hat{\eta}'\mathbf{X}_i(t)} d\Lambda_2^*(t) \Big] = 0, \;\; 0 \le t \le \tau, \qquad (9)$$

where $\hat{\eta} = (\hat{\gamma}, \quad \hat{\xi})'$ and $d\widehat{\Lambda}_1^*(t)$ are the estimators of η and $d\Lambda_1^*(t)$, respectively, solved from (5) and (6), and $W(t)$ is a possibly data-dependent weight function. We denote the estimates of β and $d\Lambda_2^*(t)$ by $\hat{\beta}$ and $d\widehat{\Lambda}_2^*(t)$, respectively, solved from (8) and (9).

To establish the asymptotic properties of $\hat{\beta}$ and $\hat{\eta}$, define

$$\widehat{M}_i^*(t) = \widetilde{N}_i(t) - \int_0^t e^{\hat{\eta}'\mathbf{X}_i(s)} d\widehat{\Lambda}_1^*(s; \hat{\eta}),$$

$$\widehat{M}_i(t) = \int_0^t Y_i(s)d\widetilde{N}_i(s) - \int_0^t e^{\hat{\beta}'\mathbf{Z}_i + \hat{\eta}'\mathbf{X}_i(s)} d\widehat{\Lambda}_2^*(s; \hat{\beta}, \hat{\eta}),$$

$$\widehat{E}_Z(t; \beta, \eta) = \frac{\sum_{i=1}^n \mathbf{Z}_i e^{\beta'\mathbf{Z}_i + \eta'\mathbf{X}_i(t)}}{\sum_{i=1}^n e^{\beta'\mathbf{Z}_i + \eta'\mathbf{X}_i(t)}} \text{ and } e_z(t) = lim_{n\to\infty}\widehat{E}_Z(t; \beta_0, \eta_0).$$

The following theorem gives the consistency and asymptotic normality of $\hat{\beta}$ and $\hat{\eta}$.

Theorem 1. *Assume that the conditions (C1)–(C5) given in the Appendix hold. Then $\hat{\eta}$ and $\hat{\beta}$ are consistent estimators of η_0 and β_0, respectively. The distributions of $n^{1/2}(\hat{\eta} - \eta_0)$ and $n^{1/2}(\hat{\beta} - \beta_0)$ can be asymptotically approximated by the normal distributions with mean zero and covariance matrices $\widehat{\Sigma}_\eta = \widehat{\Omega}_\eta^{-1}\widehat{\Psi}\widehat{\Omega}_\eta^{-1}$ and $\widehat{\Sigma}_\beta = \widehat{A}_\beta^{-1}\widehat{\Sigma}\widehat{A}_\beta^{-1}$, respectively, where $a^{\otimes 2} = aa'$, $\widehat{\Psi} = n^{-1}\sum_{i=1}^n \hat{u}_i^{\otimes 2}$, $\widehat{\Sigma} = n^{-1}\sum_{i=1}^n (\hat{v}_{1i} - \hat{v}_{2i})^{\otimes 2}$,*

$$\hat{u}_i = \int_0^\tau \Big(\mathbf{X}_i(t) - \bar{X}(t; \hat{\eta})\Big) d\widehat{M}_i^*(t)\,,$$

$$\hat{v}_{1i} = \int_0^\tau W(t)\Big(\mathbf{Z}_i - \widehat{E}_Z(t; \hat{\beta}, \hat{\eta})\Big) d\widehat{M}_i(t)\,,$$

$$\hat{v}_{2i} = \int_0^\tau \widehat{A}_\eta \widehat{\Omega}_\eta^{-1}\Big(\mathbf{X}_i(t) - \bar{X}(t; \hat{\eta})\Big) d\widehat{M}_i^*(t)\,,$$

$$\widehat{A}_\beta = n^{-1}\sum_{i=1}^{n}\int_0^\tau W(t)e^{\hat{\beta}'\mathbf{Z}_i+\hat{\eta}'\mathbf{X}_i(t)}\Big(\mathbf{Z}_i-\widehat{E}_Z(t;\hat{\beta},\hat{\eta})\Big)^{\otimes 2} d\widehat{\Lambda}_2^*(t;\hat{\beta},\hat{\eta}),$$

$$\widehat{A}_\eta = n^{-1}\sum_{i=1}^{n}\int_0^\tau W(t)e^{\hat{\beta}'\mathbf{Z}_i+\hat{\eta}'\mathbf{X}_i(t)}\Big(\mathbf{Z}_i-\widehat{E}_Z(t;\hat{\beta},\hat{\eta})\Big)X_i'(t)d\widehat{\Lambda}_2^*(t;\hat{\beta},\hat{\eta})$$

and

$$\widehat{\Omega}_\eta = n^{-1}\sum_{i=1}^{n}\int_0^\tau \{\mathbf{X}_i(t)-\bar{X}(t;\hat{\eta})\}^{\otimes 2}e^{\hat{\eta}'\mathbf{X}_i(t)}d\widehat{\Lambda}_1^*(t;\hat{\eta}).$$

4 A Simulation Study

In this section, we report some results obtained from a simulation study conducted to assess the finite sample behavior of the estimation procedure proposed in the previous sections. For each subject i, the covariate $\mathbf{Z}_i$ was assumed to be a Bernoulli random variable with the probability of success being 0.5. Given $\mathbf{Z}_i$ and some unobserved random effects $\mathbf{b}_i(t) = (b_{1i}(t), b_{2i}(t), b_{3i}(t))'$, the hazard function of the censoring time C_i was assumed to have the form

$$\lambda_i(t|\mathbf{Z}_i, \mathscr{B}_{it}) = \lambda_0 + \xi\mathbf{Z}_i + b_{3i}(t), \tag{10}$$

with the largest follow-up time $\tau = 1$. The number of observations $\widetilde{N}_i(t)$ was assumed to follow a Poisson process on $(0, C_i)$ with the mean function

$$E\{N_i(t)|\mathbf{Z}_i, \mathscr{B}_{it}\} = \int_0^t \exp\{\gamma\mathbf{Z}_i + b_{2i}(s)\}d\mu_0(s)\,. \tag{11}$$

In practice, the exact time of C_i may not be observable and $d\widetilde{N}_i(t)$ is observed instead of $dN_i(t)$, thus we considered $E\{\widetilde{N}_i(t)|\mathscr{B}_{it}\}$ for the observation process. From (10) and (11),

$$E\{d\widetilde{N}_i(t)|\mathbf{Z}_i, \mathscr{B}_{it}\} = \exp\{\gamma\mathbf{Z}_i - \xi\mathbf{Z}_i t\}d\Lambda_1^*(t),$$

where $d\Lambda_1^*(t) = \exp\{-\lambda_0 t + b_{2i}(t) - B_i(t)\}d\mu_0(t)$ and $B_i(t) = \int_0^t b_{3i}(s)ds$. Given $\mathbf{Z}_i$ and $\mathscr{B}_{it}$, $\widetilde{N}_i(t)$ was assumed to follow a nonhomogeneous Poisson process and the total number of observation times m_i was generated with mean $E\{m_i\} = E\{\widetilde{N}_i(\tau)|Z_i, \mathscr{B}_{i\tau}\}$. Then the observation times $\{T_{i,1}, \ldots, T_{i,m_i}\}$ were taken as m_i order statistics from the density function

$$f_{\widetilde{N}}(t) = \frac{\exp\{\gamma\mathbf{Z}_i - \xi\mathbf{Z}_i t\}d\Lambda_1^*(t)}{\int_0^\tau \exp\{\gamma\mathbf{Z}_i - \xi\mathbf{Z}_i t\}d\Lambda_1^*(t)}.$$

The longitudinal response $Y_i(t)$ was generated from a mixed Poisson process with the mean function

$$E\{Y_i(t)|\mathbf{Z}_i, \mathscr{B}_{it}\} = Q_i \Lambda_0(t) \exp\{-\beta \mathbf{Z}_i + b_{1i}(t)\}, \tag{12}$$

where Q_i was generated independently from a gamma distribution with mean 1 and variance 0.5. The results given below are based on the sample size of 100 or 200 with 1000 replications and $W(t) = W_i = 1$.

Table 1 shows the estimation results on β for the situation when b_{1i}, b_{2i} and b_{3i} are time-independent. Note that here $\xi_0 = 0$ or $\gamma_0 = 0$ represents the cases when either censoring or the observation times is independent of covariates, respectively. For the random effects, we took $b_{1i} = b_{2i} = b_{3i} = b_i$, where the $b_i's$ were generated from the uniform distribution over (−0.5, 0.5). It can be seen that the proposed estimates seem unbiased and the estimated standard errors (SEE) are close to the sample standard errors (SSE). Also the empirical 95 % coverage probabilities (CP) are quite accurate.The same conclusions are also obtained for the situation when b_{1i}, b_{2i} and b_{3i} are time-dependent, for which the results are presented in Table 2. Here we took $b_{1i}(t) = b_i t^{1/3}$, $b_{2i}(t) = b_i t^{1/2}$ and $b_{3i} = b_i$ with the same b_i generated as for Table 1. We also considered other set-ups such as using different baselines and with Z_i being a continuous variable and obtained similar results.

Table 1 Estimation results with $\lambda_0 = 2$, $\mu_0(t) = 20t$, $\Lambda_0(t) = 5t$, $b_{1i} = b_{2i} = b_{3i}$

	$n = 100$			$n = 200$		
β_0	0	0.2	0.5	0	0.2	0.5
$(\gamma_0, \xi_0) = (0, 0)$						
Bias	0.007	0.012	0.000	−0.009	−0.005	−0.003
SEE	0.177	0.177	0.179	0.127	0.128	0.129
SSE	0.194	0.188	0.199	0.134	0.129	0.132
CP	0.924	0.934	0.905	0.934	0.946	0.934
$(\gamma_0, \xi_0) = (0, 0.2)$						
Bias	0.036	0.035	0.042	0.036	0.036	0.042
SEE	0.178	0.180	0.182	0.127	0.128	0.130
SSE	0.192	0.186	0.197	0.133	0.134	0.138
CP	0.922	0.937	0.921	0.922	0.932	0.923
$(\gamma_0, \xi_0) = (0.5, 0)$						
Bias	0.006	−0.005	0.004	0.004	−0.003	0.002
SEE	0.173	0.174	0.174	0.123	0.125	0.125
SSE	0.177	0.179	0.183	0.126	0.130	0.130
CP	0.938	0.939	0.937	0.934	0.943	0.927
$(\gamma_0, \xi_0) = (0.5, 0.2)$						
Bias	0.047	0.043	0.035	0.042	0.037	0.041
SEE	0.174	0.173	0.176	0.125	0.125	0.126
SSE	0.181	0.184	0.182	0.128	0.131	0.134
CP	0.918	0.922	0.936	0.929	0.931	0.923

Table 2 Estimation results with $\lambda_0 = 2$, $\mu_0(t) = 20t$, $\Lambda_0(t) = 5t$, $b_{1i}(t) = b_i t^{1/3}$, $b_{2i}(t) = b_i\sqrt{t}$ and $b_{3i}(t) = b_i$

	$n = 100$			$n = 200$		
β_0	0	0.2	0.5	0	0.2	0.5
$(\gamma_0, \xi_0) = (0, 0)$						
Bias	0.003	−0.005	−0.006	−0.003	−0.001	−0.004
SEE	0.172	0.171	0.173	0.123	0.123	0.125
SSE	0.182	0.181	0.181	0.127	0.128	0.130
CP	0.940	0.928	0.933	0.940	0.944	0.942
$(\gamma_0, \xi_0) = (0, 0.2)$						
Bias	0.045	0.038	0.040	0.036	0.044	0.042
SEE	0.173	0.173	0.175	0.123	0.125	0.127
SSE	0.183	0.186	0.185	0.129	0.132	0.133
CP	0.921	0.923	0.927	0.927	0.918	0.926
$(\gamma_0, \xi_0) = (0.5, 0)$						
Bias	0.006	−0.004	−0.002	−0.006	0.006	0.002
SEE	0.168	0.168	0.169	0.120	0.120	0.121
SSE	0.178	0.181	0.173	0.129	0.127	0.122
CP	0.939	0.933	0.944	0.939	0.928	0.944
$(\gamma_0, \xi_0) = (0.5, 0.2)$						
Bias	0.051	0.043	0.035	0.037	0.044	0.036
SEE	0.166	0.169	0.171	0.120	0.120	0.122
SSE	0.182	0.179	0.169	0.126	0.123	0.128
CP	0.911	0.921	0.939	0.922	0.914	0.925

To further investigate the performance of the proposed estimators of β in comparison with those proposed by He et al. (2009) and Sun et al. (2012), we carried out a simulation study and estimated β using all four methods. Note that unlike the proposed estimation procedures, the latter two methods require observing the exact time of a censoring or terminal event C_i. For this, we used the subjects' last observation times as commonly done in practice. With respect to the method given by Sun et al. (2012), we applied it by using C_i as its original terminal event time D_i and τ as its C_i. Note that as mentioned earlier, both He et al. (2009) and Sun et al. (2012) considered the distribution-based random effects for possible correlations. For the comparison, we focus on the performances of their procedures when the random effects follow various distributions besides those assumed. However, since both of them involve covariate effects in forms different from those considered by our proposed models, we fix $\beta_0 = 0$ and $\xi_0 = 0$ in order to avoid unfair comparisons caused by the misspecification of covariate effects. The estimation results are given in Table 3 with three set-ups. In the first set-up, referred to as M_1, we considered the situation as used for Table 1 except $\mu_0(t) = 10t$ and $b_{1i} = -b_{2i} = b_{3i}$. In the second and third set-ups called M_2 and M_3, we generated $b_{1i}(t)$, $b_{2i}(t)$ and $b_{3i}(t)$ from various distributions such that the assumptions required by either Sun et al. (2012)

Table 3 Estimation results on β based on the proposed procedure and the procedures given in Sun et al. (2012) and He et al. (2009) with $\beta_0 = \xi_0 = \gamma_0 = 0$

	Proposed	Sun et al. (2012)	He et al. (2009)
M_1, $n = 100$			
Bias	−0.003	−0.004	0.009
SSE	0.162	0.261	0.206
M_1, $n = 200$			
Bias	−0.003	−0.003	0.007
SSE	0.116	0.184	0.154
M_2, $n = 100$			
Bias	0.004	0.004	0.003
SSE	0.123	0.306	0.184
M_2, $n = 200$			
Bias	−0.001	−0.003	0.011
SSE	0.089	0.227	0.145
M_3, $n = 100$			
Bias	0.001	−0.010	0.000
SSE	0.074	0.221	0.071
M_3, $n = 200$			
Bias	0.002	0.000	−0.003
SSE	0.055	0.150	0.051

Set-up M_1: $\mu_0(t) = 10t$, $\lambda_0 = 2$, $\Lambda_0(t) = 5t$, $b_{1i} = -b_{2i} = b_{3i} = b_i$, where b_i followed a uniform distribution on $(-0.5, 0.5)$

Set-up M_2: $\mu_0(t) = 10t$, $\lambda_0 = 0$, $\Lambda_0(t) = 5t$, $b_{1i} = -b_{2i} = b_i$, where b_i followed a uniform distribution on $(-0.5, 0.5)$ and b_{3i} followed an extreme value distribution with distribution function $F(t) = 1 - \exp\{-\exp(t)\}$ Set-up M_3: $\mu_0(t) = 4t$, $\lambda_0 = 0$, $\Lambda_0(t) = 5t$, $b_{1i} = 0.2b_{2i} + 0.2b_{2i}$, $b_{2i} = \log(b_{2i}^*)$ and $b_{3i} = \exp(v_i)$, where v_i and b_{2i}^* were generated, respectively, from a normal distribution with mean 0 and standard deviation 0.5 and gamma distribution with mean 4 and variance 8

or He et al. (2009) are satisfied. For example, we took $\lambda_0(t) = 0$ and generated $b_{3i}(t)$ from an extreme-value distribution as assumed by Sun et al. (2012). We also generated $b_{1i}(t)$, $b_{2i}(t)$ and $b_{3i}(t)$ from the assumed distributions required by He et al. (2009).

Note that in all set-ups considered above, our proposed models are correctly specified because there are no assumed distributions on $b_{1i}(t)$, $b_{2i}(t)$ or $b_{3i}(t)$. In contrast, the models from either of He et al. (2009) or Sun et al. (2012) are only correctly specified in one of the set-ups. On the other hand, since there are no covariate effects in all set-ups, we do not expect that the point estimates of β given by He et al. (2009) or Sun et al. (2012) are much biased even if the imposed distributions are misspecified in the estimation. For their variance estimates, we expect that SEE and SSE agree for both, because the former applied bootstrap resampling and the latter did not involve any assumed distribution of random effects in their variance estimation. Therefore, we only compare bias and SSE. It can be seen that all estimation procedures gave comparably small bias as expected. However, it appears that the proposed estimators are more efficient for all cases in general. In comparison, the method given by He et al. (2009) is comparably efficient to the proposed estimators only under M_3 when all its distribution assumptions are satisfied. For the method given by Sun et al. (2012), it is worth noting that when D_i is substituted by the last observation time C_i from subject i, it gives relatively large SSE, especially when C_i's vary much, regardless of whether the assumption about $b_{3i}(t)$ is satisfied (for M_2) or not (for M_3).

5 Concluding Remarks

We proposed a joint model for analyzing longitudinal data with informative censoring and observation times. The mutual correlations are characterized via a shared vector of time-dependent random effects. As mentioned earlier, several procedures have been developed in the literature for longitudinal data when either censoring or observation process is informative. However when both of them are informative, there is limited work that can apply except those given in He et al. (2009) and Sun et al. (2012). In addition, all the existing procedures assumed time-independent or specifically distributed correlation structures. The proposed joint model is flexible in that the shared vector of random effects can be time-dependent and neither of its structure nor distribution are specified. For the parameter estimation, the proposed procedure is simple and easy to implement.

There exist several directions for future research. One is that as mentioned above, one may want to consider other models rather than models (1)–(3) and develop similar estimation procedures. Of course, a related problem is model selection and one may want to develop some model selection techniques to choose the optimal model among several candidate models (Tong et al. 2009; Wang et al. 2014). Note that in the proposed method, we have employed a weight function $W(t)$ and it would be desirable to develop some procedures for the selection of an optimal $W(t)$. As in most similar situations, this is clearly a difficult problem as it requires the specification of the covariance function of $Y_i(t)$ and $\widetilde{N}_i(t)$ (Sun et al. 2012). Finally in the above, we have focused on regression analysis of $Y_i(t)$ with time-independent covariates. Sometimes one may face time-dependent covariates and thus it would be helpful to generalize the proposed method to this latter situation. Also sometimes nonparametric estimation of $Y_i(t)$ or the baseline functions may be of interest. For those purposes, some constraints should be imposed on $\mathbf{b}_i(t)$ for identifiability, for example, $E\{\mathbf{b}_i(t)\} = \mathbf{0}$. When panel count data arise (Sun and Zhao, 2013), the generalization of existing nonparametric estimation procedures to cases with informative observation or censoring times is a challenging direction for future work too.

Appendix

Proof of Theorem 1

To derive the asymptotic properties of the proposed estimators $\hat{\beta}$ and $\hat{\eta}$, we need the following regularity conditions:

(C1) $\{\widetilde{N}_i(\cdot), Y_i(\cdot), C_i, \mathbf{Z}_i\}_{i=1}^n$ are independent and identically distributed.
(C2) There exists a $\tau > 0$ such that $P(C_i \geq \tau) > 0$.
(C3) Both $\widetilde{N}_i(t)$ and $Y_i(t)$ $(0 \leq t \leq \tau, i = 1, \ldots, n)$ are bounded.
(C4) $W(t)$ and $\mathbf{Z}_i, i = 1, \ldots, n$, have bounded variations and $W(t)$ converges almost surely to a deterministic function $w(t)$ uniformly in $t \in [0, \tau]$.
(C5) $A_\beta = E\{\int_0^\tau w(t) e^{\beta_0'\mathbf{Z}_i+\eta_0'\mathbf{X}_i(t)}[\mathbf{Z}_i - e_z(t)]^{\otimes 2} d\Lambda_2^*(t)\}$ and $\Omega_\eta = E\Big[\int_0^\tau \{\mathbf{X}_i(t) - \bar{x}(t)\}^{\otimes 2} e^{\eta_0'\mathbf{X}_i(t)} d\Lambda_1^*(t)\Big]$ are both positive definite.

Under condition (C2), we define

$$U_1(\beta; \hat{\eta}) = \sum_{i=1}^n \int_0^\tau W(t)\mathbf{Z}_i \Big[Y_i(t) d\widetilde{N}_i(t) - e^{\beta'\mathbf{Z}_i + \hat{\eta}'\mathbf{X}_i(t)} d\widehat{\Lambda}_2^*(t) \Big],$$

which is integrable under conditions (C3) and (C4). Also note that $d\widehat{\Lambda}_2^*(t)$ satisfies

$$\sum_{i=1}^n \Big[Y_i(t) d\widetilde{N}_i(t) - e^{\beta'\mathbf{Z}_i + \hat{\eta}'\mathbf{X}_i(t)} d\widehat{\Lambda}_2^*(t) \Big] = 0, \ \ 0 \leq t \leq \tau. \tag{13}$$

Let

$$\widehat{A}_\beta(\beta) = -n^{-1}\partial U_1(\beta, \hat{\eta})/\partial\beta', \widehat{A}_\eta(\eta) = -n^{-1}\partial U_1(\beta_0, \eta)/\partial\eta',$$

and under (C1), let

$$A_\beta = \lim_{n\to\infty} \widehat{A}_\beta(\beta_0), \ A_\eta = \lim_{n\to\infty} \widehat{A}_\eta(\eta_0).$$

The consistency of $\hat{\beta}$ and $\hat{\eta}$ follows from the facts that $U_1(\beta_0; \hat{\eta})$ and $U_\eta(\eta_0)$ both tend to 0 in probability as $n \to \infty$, and that under condition (C5), $\widehat{A}_\beta(\beta)$ and $-n^{-1}\partial U_\eta(\eta)/\partial\eta'$ both converge uniformly to the positive definite matrices A_β and Ω_η over β and η, respectively, in neighborhoods around the true values β_0 and η_0. Then the Taylor series expansions of $U_1(\hat{\beta}; \hat{\eta})$ at $(\beta_0; \hat{\eta})$ and (β_0, η_0) yield $n^{1/2}(\hat{\beta} - \beta_0) = A_\beta^{-1} n^{-1/2} U_1(\beta_0; \hat{\eta}) + o_p(1) = A_\beta^{-1}\Big\{n^{-1/2}U_1(\beta_0; \eta_0) - A_\eta n^{1/2}(\hat{\eta} - \eta_0)\Big\} + o_p(1)$.

The proof of Theorem 1 is sketched as follows:

(1) First, using some derivation operation to $U_1(\beta;\hat{\eta})$ and (13), we can get

$$\widehat{A}_\beta(\beta) = n^{-1}\sum_{i=1}^{n}\int_0^\tau W(t)\{\mathbf{Z}_i - \widehat{E}_Z(t;\beta,\hat{\eta})\}^{\otimes 2} e^{\beta'\mathbf{Z}_i+\hat{\eta}'\mathbf{X}_i(t)} d\widehat{\Lambda}_2^*(t;\beta,\hat{\eta}).$$

(2) Solving $d\widehat{\Lambda}_2^*(t;\beta_0,\eta_0)$ from (13) and applying to $U_1(\beta_0;\eta_0)$ yields

$$U_1(\beta_0;\eta_0) = \sum_{i=1}^{n}\int_0^\tau w(t)\Big(\mathbf{Z}_i - e_z(t)\Big)dM_i(t) + o_p(n^{1/2}),$$

where $e_z(t) = lim_{n\to\infty}\widehat{E}_Z(t;\beta_0,\eta_0)$ as defined earlier in Sect. 3 and $w(t)$ is a deterministic function defined under (C5).

(3) Differentiation of $U_1(\beta_0,\eta)$ and (13) with respect to η yields

$$\widehat{A}_\eta(\eta) = n^{-1}\sum_{i=1}^{n}\int_0^\tau W(t)\big[\mathbf{Z}_i - \widehat{E}_Z(t;\beta_0,\eta)\big] e^{\beta_0'\mathbf{Z}_i+\eta'\mathbf{X}_i(t)} X_i'(t) d\widehat{\Lambda}_2^*(t;\beta_0,\eta)\,.$$

(4) According to Eq. (5) and by using the asymptotic results in Lin et al. (2000) (A.5), one can show that

$$n^{1/2}\{\hat{\eta}-\eta_0\} = \Omega_\eta^{-1} n^{-1/2}\sum_{i=1}^{n}\Bigg[\int_0^\tau \Big(\mathbf{X}_i(t) - \frac{s^{(1)}(t)}{s^{(0)}(t)}\Big)dM_i^*(t)\Bigg] + o_p(1),$$

where $\Omega_\eta = E\Big[\int_0^\tau \{\mathbf{X}_i(t) - \bar{x}(t)\}^{\otimes 2} e^{\eta_0'\mathbf{X}_i(t)} d\Lambda_1^*(t)\Big]$, which is invertible under (C5).

Combining the results in steps (1)–(4), we have

$$\begin{aligned} U_1(\beta_0;\hat{\eta}) = &\sum_{i=1}^{n}\Bigg[\int_0^\tau w(t)\big\{\mathbf{Z}_i - e_z(t)\big\}dM_i(t)\Bigg] \\ &-A_\eta \Omega_\eta^{-1}\sum_{i=1}^{n}\Bigg[\int_0^\tau \{\mathbf{X}_i(t) - \bar{x}(t)\}dM_i^*(t)\Bigg] + o_p(n^{1/2}). \end{aligned}$$

Since A_β is also invertible under (C5), it then follows from the multivariate central limit theorem that the conclusions hold.

References

Cheng, S. C., & Wei, L. J. (2000). Inferences for a semiparametric model with panel data. *Biometrika, 87*, 89–97.

Gandy, A., & Jensen, U. (2005). Checking a semiparametric additive hazards model. *Lifetime Data Analysis, 11*, 451–472.

Ghosh, D. (2003). Goodness-of-fit methods for additive-risk models in tumorignenicity experiments. *Biometrics, 59*, 721–726.

He, X., Tong, X., & Sun, J. (2009). Semiparametric analysis of panel count data with correlated observation and follow-up times. *Lifetime Data Analysis, 15*, 177–196.

Hu, X. J., Sun J., & Wei, L. J. (2003). Regression parameter estimation from panel counts. *Scandinavian Journal of Statistics, 30*, 25–43.

Huang, C. Y., Wang, M. C., & Zhang, Y. (2006). Analysing panel count data with informative observation times. *Biometrika, 93*, 763–775.

Kalbfleisch, J. D., & Prentice, R. L. (2002). *The statistical analysis of failure time data*. New York: Wiley.

Kim, J., & Lee, S.Y. (1998). Two-sample goodness-of-fit tests for additive risk models with censored observations. *Biometrika, 85*, 593–603.

Kim, S., Zeng, D., Chambless, L., & Li, Y. (2012). Joint models of longitudinal data and recurrent events with informative terminal event. *Statistics in Biosciences, 4*, 262–281.

Li, N., Zhao, H., & Sun, J. (2013). Semiparametric transformation models for panel count data with correlated observation and follow-up times. *Statistics in Medicine, 32*(17), 3039–3054.

Lin, D. Y., Oaks, D., & Ying, Z. (1998). Additive hazards regression with current status data. *Biometrika, 85*(2), 289–298.

Lin, D. Y., & Ying, Z. (2001). Semiparametric and Nonparametric Regression Analysis of Longitudinal Data (with discussion). *Journal of the American Statistical Association, 96*(453), 103–113.

Sun, J., & Kalbfleisch, J. D. (1995). Estimation of the mean function of point processes based on panel count data. *Statistica Sinica, 5*, 279–289.

Sun, J., Tong, X., & He, X. (2007). Regression analysis of panel count data with dependent observation times. *Biometrics, 63*, 1053–1059.

Sun, J., & Wei, L. J. (2000). Regression analysis of panel count data with covariate-dependent observation and censoring times. *Journal of the Royal Statistical Society, Series B, 62*, 293–302.

Sun, J., & Zhao, X., (2013). *The statistical analysis of panel count data*. New York: Springer.

Sun, L., Song, X., Zhou, J., & Liu, L. (2012). Joint analysis of longitudinal data with informative observation times and a dependent terminal event. *Journal of the American Statistical Association, 107*(498), 688–700.

Tong, X., Sun, L., He, X., & Sun, J. (2009). Variable selection for panel count data via non-concave penalized estimating function. *Scandinavian Journal of Statistics, 36*, 620–635.

Wang, H., Li, Y., & Sun, J. (2014). Focused and model average estimation for regression analysis of panel count data. *Scandinavian Journal of Statistics*. doi:10.1002/sjos.12133.

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

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

Yuen, K.C., & Burke, M.D. (1997). A test of fit for a semiparametric additive risk model. *Biometrika, 84*, 631–639.

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

Zhang, Z., Sun, J., & Sun, L. (2005). Statistical analysis of current status data with informative observation times. *Statistics in Medicine, 24*, 1399–1407.

Zhao, X., & Tong, X. (2011). Semiparametric regression analysis of panel count data with informative observation times. *Computational Statistics and Data Analysis, 55*(1), 291–300.

Zhao, X., Tong, X., & Sun, J. (2013). Robust estimation for panel count data with informative observation times. *Computational Statistics and Data Analysis, 57*, 33–40.

Zhou, J., Zhao, X., & Sun, L. (2013). A new inference approach for joint models of longitudinal data with informative observation and censoring times. *Statistica Sinica, 23*, 571–593.