Computational Statistics and Data Analysis 107 (2017) 120–130

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Semiparametric regression analysis of multivariate longitudinal data with informative observation times

Shirong Deng[a], Kin-yat Liu[b], Xingqiu Zhao[b,c,*]

[a] School of Mathematics and Statistics, Wuhan University, Wuhan, Hubei, China
[b] Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong
[c] The Hong Kong Polytechnic University, Shenzhen Research Institute, Shenzhen, China

ARTICLE INFO

Article history:
Received 22 July 2015
Received in revised form 25 March 2016
Accepted 10 October 2016
Available online 27 October 2016

Keywords:
Estimating equation
Informative observation times
Latent variable
Model checking
Multivariate longitudinal data
Semiparametric regression

ABSTRACT

Multivariate longitudinal data arises when subjects under study may experience several possible related response outcomes. This article proposed a new class of flexible semiparametric models for multivariate longitudinal data with informative observation times through latent variables and completely unspecified link functions, which allows for any functional forms of covariate effects on the intensity functions for the observation processes. A novel estimating equation approach that does not rely on forms of link functions and distributions of frailties is developed. The asymptotic properties for the resulting estimators and the model checking technique for the overall fit of the proposed models are established. The simulation results show that the proposed approach works well. The analysis of skin cancer chemoprevention trial data is provided for illustration.

1. Introduction

In many longitudinal studies, multivariate longitudinal data arise when subjects under study may experience several related events repeatedly at distinct time points during a relatively long follow-up period. An example of multivariate longitudinal data that motivated this work is a skin cancer chemoprevention trial conducted by the University of Wisconsin Comprehensive Cancer Center in Madison, Wisconsin (Lee, 2008; Li, 2011). It was a 5-year randomized, double-blinded, and placebo-controlled Phase III clinical trial. The primary objective of this trial was to evaluate the effectiveness of 0.5 g/m^2/day PO difluoromethylornithine (DFMO) in preventing new skin cancers in a population of individuals with a history of non-melanoma skin cancers: basal cell carcinoma or squamous cell carcinoma. The subjects missed scheduled visits or visited clinic on unscheduled dates. At each visit, the number of occurrences of both basal cell carcinoma and squamous cell carcinoma since the previous visit were recorded.

In the irregularly observed longitudinal data analysis, there are two important processes involved: the response process and the observation process. A basic assumption behind the usual methods is that the observation times are independent of response variable, completely or given covariates, i.e., the observation process is noninformative (e.g., Lin and Ying, 2001; Welsh et al., 2002). However, this assumption may be violated in many applications. Such as the skin cancer study, Li et al. (2011) and Zhang et al. (2013) have verified that the clinical visit times contain some relevant information about the recurrence processes of two cancers. We call these response-dependent visit times as informative observation times.

[*] Corresponding author at: Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong.
E-mail address: xingqiu.zhao@polyu.edu.hk (X. Zhao).

http://dx.doi.org/10.1016/j.csda.2016.10.006

Thus it is very necessary to incorporate the relationship between the response process and the observation process into longitudinal data models.

For the univariate longitudinal data analysis with informative observation times, two methods have been developed. One is the conditional modeling approach (Sun et al., 2005), which obviously characterized the dependence of the response process and the observation times. Another one is the frailty-based approach proposed by Sun et al. (2007), Liang et al. (2009), Zhao et al. (2012), and Zhou et al. (2013) among others. For example, Sun et al. (2007) used a shared latent variable or frailty to characterize the correlation between the response process and the observation times with informative censoring times. Liang et al. (2009) modeled the longitudinal data with informative observation times via two latent variables that satisfied a linear relationship where the distributional assumption for a latent variable is required. Zhao et al. (2012) considered more general joint models using a completely unspecified link function and a latent variable to characterize the correlations between the response process and the observation process, and developed estimating equation approaches. Zhou et al. (2013) considered a semiparametric mixed random effect model for the response process in the presence of informative observation and censoring times.

For the multivariate longitudinal data with informative observation times, one additional issue involved is the correlation among different types of the response processes. For the analysis of such complex data, the existing research mainly focuses on its special case where each longitudinal response process is a counting process. For example, Li et al. (2011), Zhang et al. (2013), Zhao et al. (2013) and Li et al. (2015) proposed different semiparametric regression models that allow the recurrent event process and the observation process to be correlated, by leaving the dependence structures for related types of panel count processes completely unspecified.

In this paper, motivated by Zhao et al. (2012), we propose a semiparametric marginal modeling approach for the multivariate longitudinal data with informative observation times through latent variables and different completely unspecified link functions to characterize different correlations between each type of response process and the corresponding observation process. An important advantage for the modeling is the nonrestrictive condition on the correlation between different types of response process and different correlations between each type of response process and the corresponding observation process.

The remainder of this paper is organized as follows. We begin in Section 2 by introducing notation and describing models for multivariate longitudinal data with informative observation times. In Section 3 an estimating equation approach is developed to estimate the regression parameters involved in the proposed models, and the asymptotic properties for the resulting estimators are given in this section. In Section 4, we discuss the model checking technique for goodness of fit of our models. The simulation results are presented in Section 5 to assess the finite-sample performance of the proposed inference procedure, and the analysis of skin cancer chemoprevention trial data is provided to illustrate the proposed method in Section 6. Some concluding remarks are made in Section 7.

2. Statistical model

Consider a longitudinal study that consists of n independent subjects and suppose that each subject may experience K different types of longitudinal outcomes. For subject i, let $Y_{ik}(t)$ denote the longitudinal response process with type k and suppose that $Y_{ik}(t)$ is observed at distinct time points $0 < T_{ik,1} < T_{ik,2} < \cdots < T_{ik,m_{ik}}$, where m_{ik} is the potential or scheduled number of observations on subject i with respect to the kth longitudinal response variable. Let $\mathbf{X}_i$ be a p-dimensional vector of covariates and C_i the follow-up or censoring time for subject i, $i = 1, \dots, n$. Note that here for the simplicity of presentation, we assume that $\mathbf{X}_i$ and C_i are the same for different types of longitudinal response. The inference approach proposed below can be easily extended to the situation where there exists different covariates and follow-up or censoring times for different responses. Define $N_{ik}(t) = \sum_{j=1}^{m_{ik}} I\left(T_{ik,j} \le t\right)$, where $I(\cdot)$ is the indicator function. Then $\tilde{N}_{ik}(t) = N_{ik}(t \wedge C_i)$ is a counting process characterizing the number of observation times on subject i with respect to the kth longitudinal response variable up to time t. Then the process $Y_{ik}(t)$ is observed only at the time points where $\tilde{N}_{ik}(t)$ jumps.

For the analysis, suppose that $\mathbf{Z}_i = (Z_{i1}, \dots, Z_{iK})'$ is an unobserved random vector independent of $\mathbf{X}_i$ with Z_{ik} being positive, and assume that given $\mathbf{X}_i$ and $\mathbf{Z}_i$, $Y_{ik}(t)$ follows the marginal model

$$E\{Y_{ik}(t)|\mathbf{X}_i, \mathbf{Z}_i\} = \mu_{0k}(t) + \beta'\mathbf{X}_i + h_k(Z_{ik}), \tag{1}$$

where $\mu_{0k}(t)$ is an unknown baseline mean function, β is a p-dimensional vector of unknown regression parameters, and $h_k(\cdot)$ is a completely unspecified function with $E\{h_k(Z_{ik})\} = 0$ for identifiability. The condition $E\{h_k(Z_{ik})\} = 0$ yields that $E\{Y_{ik}(t)|\mathbf{X}_i\} = \mu_{0k}(t) + \beta'\mathbf{X}_i$ such that the uniqueness of $\mu_{0k}(t)$ and β can be ensured. Model (1) assumes that the baseline mean functions can be different for different types of longitudinal responses, however, the effects of covariates on different types of longitudinal responses are the same for the simplicity of presentation. The correlations among the K longitudinal response processes are characterized by a K-dimensional vector of unobserved frailties, where distributions of frailties are free. So, the correlation structure of longitudinal response processes is unspecified. The goal here is to estimate regression parameter β.

Give $\mathbf{X}_i$ and $\mathbf{Z}_i$, we assume that $N_{ik}(t)$ satisfies the following rate model

$$E\{dN_{ik}(t) \mid \mathbf{X}_i, \mathbf{Z}_i\} = Z_{ik} g_k(\mathbf{X}_i) d\Lambda_{0k}(t), \tag{2}$$

where $g_k(\cdot)$ is a completely unspecified positive function and $\Lambda_{0k}(t)$ is a completely unknown continuous baseline function. In model (2), it is assumed that the observation process is affected by the covariate $\mathbf{X}_i$ in a flexible way, while Zhao et al. (2012) assumed that the observation process is a mixed Poisson model with intensity function satisfying their model (2.2) where the form of covariate effect is specified. Under models (1) and (2), it is obvious that for each type k, given covariates, the longitudinal response process may be related to the observation process through the unobserved frailty Z_{ik} and the unspecified link function h_k.

In the following, we assume that conditional on $\mathbf{X}_i$ and $\mathbf{Z}_i$, the two processes N_{ik} and Y_{ik} are independent. Also assume that C_i is independent of $\{Y_{ik}(t), N_{ik}(t), \mathbf{X}_i, \mathbf{Z}_i, 0 \le t \le \tau\}$ where τ denotes the length of the study.

3. Estimation procedure and asymptotic results

For estimation of β, define

$$\bar{Y}_{ik} = \int_0^\tau Y_{ik}(t) d\tilde{N}_{ik}(t).$$

Since

$$E\{\bar{Y}_{ik}|\mathbf{X}_i, C_i, \mathbf{Z}_i\} = \int_0^\tau \xi_i(t)\{\mu_{0k}(t) + \beta'\mathbf{X}_i + h_k(Z_{ik})\}Z_{ik}g_k(\mathbf{X}_i)d\Lambda_{0k}(t),$$

where $\xi_i(t) = I(C_i \ge t)$, then

$$E\{\bar{Y}_{ik}|\mathbf{X}_i, \mathbf{Z}_i\} = \int_0^\tau P(C_i \ge t)\{\mu_{0k}(t) + \beta'\mathbf{X}_i + h_k(Z_{ik})\}Z_{ik}g_k(\mathbf{X}_i)d\Lambda_{0k}(t).$$

Thus

$$\begin{aligned}
E\{\bar{Y}_{ik}|\mathbf{X}_i\} &= \int_0^\tau P(C_i \ge t)E(Z_{ik}|\mathbf{X}_i)\{\mu_{0k}(t) + \beta'\mathbf{X}_i\}g_k(\mathbf{X}_i)d\Lambda_{0k}(t) \\
&\quad + \int_0^\tau P(C_i \ge t)E\{h_k(Z_{ik})Z_{ik}|\mathbf{X}_i\}g_k(\mathbf{X}_i)d\Lambda_{0k}(t) \\
&= E(Z_{ik})\int_0^\tau P(C_i \ge t)d\Lambda_{0k}(t)g_k(\mathbf{X}_i)\beta'\mathbf{X}_i \\
&\quad + \int_0^\tau P(C_i \ge t)[\mu_{0k}(t)E(Z_{ik}) + E\{h_k(Z_{ik})Z_{ik}\}]g_k(\mathbf{X}_i)d\Lambda_{0k}(t).
\end{aligned}$$

From model (2), we have

$$E\{m_{ik}|\mathbf{X}_i\} = E(Z_{ik})E[\Lambda_{0k}(C_i)]g_k(\mathbf{X}_i).$$

Define

$$\alpha_k \triangleq \int_0^\tau \frac{P(C_1 \ge t)}{E[\Lambda_{0k}(C_1)]}\left[\mu_{0k}(t) + \frac{E\{h_k(Z_{1k})Z_{1k}\}}{E(Z_{1k})}\right]d\Lambda_{0k}(t), \quad k = 1, \ldots, K.$$

Then, we have

$$E\{\bar{Y}_{ik}|\mathbf{X}_i\} = E\{m_{ik}|\mathbf{X}_i\}\beta'\mathbf{X}_i + \alpha_k E\{m_{ik}|\mathbf{X}_i\},$$

that is

$$E\left\{\bar{Y}_{ik} - m_{ik}(\beta'\mathbf{X}_i + \alpha_k)\middle|\mathbf{X}_i\right\} = 0. \tag{3}$$

Let $\alpha = (\alpha_1, \ldots, \alpha_K)'$ and $\theta = (\beta', \alpha')'$. Also let e_k denote the K-dimensional vector of zeros except its kth entry equal to 1 and $\bar{\mathbf{X}}_{ik} = (\mathbf{X}_i', e_k')'$. To estimate θ, motivated by Eq. (3), we propose to use the following estimating equation

$$U(\theta) = n^{-1}\sum_{i=1}^n \sum_{k=1}^K W_i\bar{\mathbf{X}}_{ik}\{\bar{Y}_{ik} - m_{ik}\theta'\bar{\mathbf{X}}_{ik}\} = 0, \tag{4}$$

where W_i's are weights that could depend on the $\mathbf{X}_i$'s and C_i's. Let $\hat{\theta}$ denote the solution to $U(\theta) = 0$. Then

$$\hat{\theta} = \left\{\sum_{i=1}^n \sum_{k=1}^K W_i m_{ik}\bar{\mathbf{X}}_{ik}^{\otimes 2}\right\}^{-1}\left\{\sum_{i=1}^n \sum_{k=1}^K W_i\bar{\mathbf{X}}_{ik}\bar{Y}_{ik}\right\}, \tag{5}$$

where $a^{\otimes 2} = aa'$ for a vector a.

Let θ_0 be the true values of θ. Then it can be shown that $\hat{\theta}$ is consistent under the regularity conditions given in the Appendix. Define

$$A = E\left[\sum_{k=1}^{K} W_1 \bar{\mathbf{X}}_{1k}^{\otimes 2} m_{1k}\right].$$

Then as we show in the Appendix, under some regularity conditions, $n^{1/2}(\hat{\theta} - \theta_0)$ converges in distribution to a random normal variable with mean zero and a covariance matrix $A^{-1}\Sigma A^{-1}$, where $\Sigma = E(\phi_1 \phi_1')$ with

$$\phi_1 = \sum_{k=1}^{K} W_1 \bar{\mathbf{X}}_{1k}\{\bar{Y}_{1k} - m_{1k}\theta' \bar{\mathbf{X}}_{1k}\}.$$

In addition, the covariance matrix given above can be consistently estimated by $\hat{A}^{-1}\hat{\Sigma}\hat{A}^{-1}$, where

$$\hat{A} = n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K} W_i \bar{\mathbf{X}}_{ik}^{\otimes 2} m_{ik},$$

and

$$\hat{\Sigma} = n^{-1}\sum_{i=1}^{n} \hat{\phi}_i \hat{\phi}_i',$$

with $\hat{\phi}_i = \sum_{k=1}^{K} W_i \bar{\mathbf{X}}_{ik}\{\bar{Y}_{ik} - m_{ik}\hat{\theta}' \bar{\mathbf{X}}_{ik}\}$.

4. Model diagnostics

In practice, in addition to the estimation of β, one may also be interested in checking the adequacy of models (1) and (2) given the observed data. To develop a procedure for this, we note that

$$E\left\{\int_0^t Y_{ik}(u) d\tilde{N}_{ik}(u) | \mathbf{X}_i\right\} = E\{m_{ik}|\mathbf{X}_i\}\mathcal{A}_{0k}(t) + \beta_0' \mathbf{X}_i E\{\tilde{N}_{ik}(t)|\mathbf{X}_i\},$$

where

$$\mathcal{A}_{0k}(t) = \int_0^t \frac{P(C_1 \geq u)}{E[\Lambda_{0k}(C_1)]}\left[\mu_{0k}(u) + \frac{E\{h_k(Z_{1k})Z_{1k}\}}{E(Z_{1k})}\right] d\Lambda_{0k}(u),$$

that can be estimated by

$$\hat{\mathcal{A}}_{0k}(t) = n^{-1}\sum_{i=1}^{n}\int_0^t \frac{Y_{ik}(u) - \hat{\beta}'\mathbf{X}_i}{n^{-1}\sum\limits_{i=1}^{n} m_{ik}} d\tilde{N}_{ik}(u), \quad k = 1, \ldots, K.$$

For each i and k, following Lin et al. (2000), define the residual

$$\hat{R}_{ik}(t) = \int_0^t [Y_{ik}(u) - \hat{\beta}'\mathbf{X}_i] d\tilde{N}_{ik}(u) - m_{ik}\hat{\mathcal{A}}_{0k}(t).$$

It can be seen that $\hat{R}_{ik}(t)$ represents the difference between the observed and model-predicted values of the kth type of longitudinal response experienced by subject i up to time t. To test the goodness-of-fit of models (1) and (2), we propose to apply the statistic

$$\Phi(t, \mathbf{x}) = n^{-1/2}\sum_{i=1}^{n}\sum_{k=1}^{K} I(\mathbf{X}_i \leq \mathbf{x})\hat{R}_{ik}(t),$$

where the event $I(\mathbf{X}_i \leq \mathbf{x})$ means that each of the components of $\mathbf{X}_i$ is not larger than the corresponding component of $\mathbf{x}$. It is easy to see that $\Phi(t, \mathbf{x})$ is the cumulative sum of $\hat{R}_{ik}(t)$ over the values of $\mathbf{X}_i$'s. Define

$$\begin{aligned}
S_{k0} &= n^{-1}\sum_{i=1}^{n} m_{ik}, \\
S_k(\mathbf{x}) &= n^{-1}\sum_{i=1}^{n} I(\mathbf{X}_i \leq \mathbf{x}) m_{ik}, \\
B(t, \mathbf{x}) &= n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K}\int_0^{\tau}\left\{I(\mathbf{X}_i \leq \mathbf{x}) - \frac{S_k(\mathbf{x})}{S_{k0}}\right\}\mathbf{X}_i d\tilde{N}_{ik}(u).
\end{aligned}$$

In the Appendix, we will show that the null distribution of $\Phi(t, \mathbf{x})$ can be approximated by the zero mean Gaussian process

$$\tilde{\Phi}(t, \mathbf{x}) = n^{-1/2} \sum_{i=1}^{n} \sum_{k=1}^{K} \int_0^t \left\{ I(\mathbf{X}_i \le \mathbf{x}) - \frac{S_k(\mathbf{x})}{S_{k0}} \right\} d\hat{R}_{ik}(u) - B(t, \mathbf{x})' n^{-1/2} \sum_{i=1}^{n} \hat{a}_i, \qquad (6)$$

where $\hat{a}_i$ is the vector $\hat{A}^{-1}\hat{\phi}_i$ without the last K entries. To make inference for the goodness-of-fit test on models (1) and (2) based on the observed residual, we need to evaluate the distribution of the supremum of the goodness-of-fit process $\Phi(t, \mathbf{x})$ when models (1) and (2) hold. However, it is impossible to evaluate this distribution analytically because the limiting process of $\Phi(t, \mathbf{x})$ does not have an independent increments structure. For this, we propose to use the simulation procedure discussed in Cheng et al. (1997) and Lin et al. (2000). Let $(G_1, \ldots, G_n)$ be independent standard normal variables independent of the observed data. Then it can be shown that the distribution of the process $\Phi(t, \mathbf{x})$ can be approximated by that of the zero mean Gaussian process

$$\hat{\Phi}(t, \mathbf{x}) = n^{-1/2} \sum_{i=1}^{n} \sum_{k=1}^{K} \int_0^t \left\{ I(\mathbf{X}_i \le \mathbf{x}) - \frac{S_k(\mathbf{x})}{S_{k0}} \right\} d\hat{R}_{ik}(u) G_i - B(t, \mathbf{x})' n^{-1/2} \sum_{i=1}^{n} \hat{a}_i G_i. \qquad (7)$$

To perform the goodness-of-fit test on models (1) and (2), based on (6) and (7), one can first repeatedly generate the standard normal random sample $(G_1, \ldots, G_n)$ given the observed data, and then obtain a large number of realizations from $\hat{\Phi}(t, \mathbf{x})$. More formally, we can apply the supremum test statistic $\sup_{t,\mathbf{x}} |\Phi(t, \mathbf{x})|$, where the p-value can be obtained by comparing the observed value of $\sup_{t,\mathbf{x}} |\Phi(t, \mathbf{x})|$ to a large number, say 1000, of realizations of $\sup_{t,\mathbf{x}} |\hat{\Phi}(t, \mathbf{x})|$.

5. Simulation study

In this section, we conducted Monte Carlo simulation studies to evaluate the finite sample properties of the proposed estimators. In this study, we assumed that there exist two related types of longitudinal response variables, that is, $K = 2$. To generate the simulated data, we first generate the covariates $\mathbf{X}_i = (X_{1i}, X_{2i})'$s with X_{1i} from a Bernoulli distribution with success probability 0.5 and X_{2i} from a uniform distribution over $(0, 1]$. The latent variable Z_{i1} was generated from the gamma distribution with mean 20 and variance 40, and we let $Z_{i2} = \rho_z Z_{i1} + S_i$ where S_i was generated from the gamma distribution with mean 60 and variance 120. The follow-up time C_i was generated from the uniform distribution over interval $(\tau/2, \tau)$.

For the longitudinal response processes, we generated them from the following models:

$$Y_{ik}(t) = \mu_{0k}(t) + \beta_1 X_{1i} + \beta_2 X_{2i} + h_k(Z_{ik}) + \varepsilon_i(t), \quad k = 1, 2,$$

where

$$\mu_{01}(t) = \log(1 + t), \qquad \mu_{02}(t) = \sin t,$$
$$h_1(Z_{i1}) = \rho(Z_{i1} - 20)/\sqrt{40}, \qquad h_2(Z_{i2}) = Z_{i2}^{\rho} - E(Z_{i2}^{\rho}),$$

with $\rho = -0.5$, 0, and 0.5, and $\varepsilon_i(t)$'s are independent standard normal variables. It can be seen that the correlation between the longitudinal response processes $Y_{1i}(t)$ and $Y_{2i}(t)$ is characterized by ρ_z. When $\rho_z > 0$ or $\rho_z < 0$, the two processes are positively or negatively correlated; when $\rho_z = 0$, the two processes have no correlation given the covariates. Here, three situations with $\rho_z = 0.5$, 0, and 0.5 were considered.

For the generation of the observation process $N_{ik}(t)$, $k = 1$ or 2, we considered a homogeneous Poisson process (HPP) with $\lambda_{0k}(t) = 1/\tau$ or a nonhomogeneous Poisson process (NHPP) with $\lambda_{0k}(t) = (t + 1)/\{\tau(\tau/2 + 1)\}$.

For the homogeneous Poisson process, given $\mathbf{X}_i$, C_i, and $\mathbf{Z}_i$, m_{ik} was generated from the Poisson distribution with mean

$$Z_{ik} g_k(\mathbf{X}_i) \Lambda_{0k}(C_i) = \frac{Z_{ik} \exp(\mathbf{X}_i'\gamma) C_i}{\tau},$$

where $g_k(\mathbf{x}) = \exp(\mathbf{x}'\gamma)$ and $\gamma = (1, 1)'$. Given m_{ik}, the observation times $(T_{ik,1}, \ldots, T_{ik,m_{ik}})$ were taken to be the order statistics of the random sample of size m_{ik} from the uniform distribution over $(0, C_i)$.

For the nonhomogeneous Poisson process, given $\mathbf{X}_i$, C_i, $\mathbf{Z}_i$, m_{ik} was generated from the Poisson distribution with mean

$$Z_{ik} g_k(\mathbf{X}_i) \Lambda_{0k}(C_i) = \frac{Z_{ik} (\mathbf{X}_i'\gamma)^2 (C_i^2/2 + C_i)}{\tau^2/2 + \tau},$$

with $g_k(\mathbf{x}) = (\mathbf{x}'\gamma)^2$. Given m_{ik}, the observation times $(T_{ik,1}, \ldots, T_{ik,m_{ik}})$ were the order statistics of a random sample of size m_{ik} from the density function

$$\frac{t + 1}{C_i^2/2 + C_i} I\{0 \le t \le C_i\}.$$

We took $\beta_0 = (\beta_{10}, \beta_{20})'$ as $(-1, 1)$, representing the effects of the covariates $\mathbf{X}$ on the response variable. For each case, we considered $n = 100$ and 200. All the results reported here were based on 1000 replications.

Table 1
Simulation results for β under (HPP, HPP) observation process with $\tau = 18$.

n	ρ	ρ_z	$\hat{\beta}_1$				$\hat{\beta}_2$			
			BIAS	SSE	ESE	CP	BIAS	SSE	ESE	CP
100	−0.5	−0.5	−0.0017	0.1351	0.1222	0.9180	−0.0053	0.0875	0.0818	0.9300
		0	0.0031	0.1164	0.1111	0.9250	0.0001	0.0809	0.0766	0.9350
		0.5	0.0001	0.1130	0.1042	0.9190	0.0009	0.0752	0.0718	0.9370
	0	−0.5	−0.0026	0.0402	0.0398	0.9410	0.0000	0.0276	0.0270	0.9430
		0	−0.0010	0.0369	0.0376	0.9450	−0.0012	0.0271	0.0257	0.9310
		0.5	−0.0000	0.0372	0.0353	0.9190	0.0006	0.0249	0.0243	0.9390
	0.5	−0.5	−0.0025	0.2146	0.2022	0.9360	0.0065	0.1039	0.1014	0.9400
		0	−0.0055	0.2214	0.2105	0.9280	0.0047	0.1095	0.1077	0.9410
		0.5	−0.0025	0.2511	0.2313	0.9320	−0.0026	0.1218	0.1190	0.9440
200	−0.5	−0.5	0.0024	0.0908	0.0886	0.9410	0.0002	0.0603	0.0590	0.9390
		0	−0.0025	0.0861	0.0830	0.9420	−0.0021	0.0580	0.0560	0.9400
		0.5	0.0055	0.0782	0.0759	0.9450	0.0030	0.0529	0.0525	0.9430
	0	−0.5	0.0002	0.0293	0.0286	0.9420	−0.0003	0.0189	0.0192	0.9500
		0	−0.0014	0.0262	0.0268	0.9470	−0.0003	0.0182	0.0183	0.9520
		0.5	−0.0004	0.0263	0.0254	0.9430	0.0002	0.0173	0.0174	0.9540
	0.5	−0.5	0.0086	0.1493	0.1482	0.9480	0.0003	0.0769	0.0733	0.9410
		0	−0.0042	0.1548	0.1553	0.9530	0.0024	0.0777	0.0782	0.9580
		0.5	0.0070	0.1779	0.1694	0.9350	0.0007	0.0857	0.0854	0.9530

Table 2
Simulation results for β under (HPP, NHPP) observation process with $\tau = 18$.

n	ρ	ρ_z	$\hat{\beta}_1$				$\hat{\beta}_2$			
			BIAS	SSE	ESE	CP	BIAS	SSE	ESE	CP
100	−0.5	−0.5	−0.0048	0.1468	0.1349	0.9260	−0.0017	0.0892	0.0886	0.9410
		0	0.0035	0.1332	0.1247	0.9340	0.0066	0.0875	0.0837	0.9370
		0.5	−0.0054	0.1234	0.1183	0.9420	−0.0029	0.0829	0.0796	0.9390
	0	−0.5	−0.0008	0.0438	0.0446	0.9450	0.0010	0.0304	0.0294	0.9410
		0	−0.0007	0.0422	0.0414	0.9350	−0.0008	0.0286	0.0277	0.9340
		0.5	−0.0001	0.0402	0.0397	0.9500	0.0004	0.0269	0.0269	0.9490
	0.5	−0.5	−0.0038	0.2159	0.1948	0.9080	−0.0018	0.1087	0.1023	0.9350
		0	0.0027	0.2242	0.2059	0.9130	0.0017	0.1138	0.1075	0.9320
		0.5	−0.0026	0.2475	0.2287	0.9190	−0.0120	0.1217	0.1185	0.9420
200	−0.5	−0.5	−0.0033	0.1028	0.0972	0.9350	0.0050	0.0628	0.0638	0.9530
		0	−0.0026	0.0954	0.0919	0.9390	0.0015	0.0618	0.0607	0.9370
		0.5	−0.0008	0.0881	0.0851	0.9490	−0.0033	0.0583	0.0570	0.9430
	0	−0.5	−0.0000	0.0319	0.0317	0.9380	0.0001	0.0211	0.0208	0.9450
		0	−0.0006	0.0304	0.0298	0.9430	−0.0003	0.0206	0.0198	0.9370
		0.5	0.0000	0.0292	0.0283	0.9300	−0.0001	0.0188	0.0190	0.9530
	0.5	−0.5	−0.0064	0.1511	0.1403	0.9180	−0.0028	0.0742	0.0729	0.9420
		0	0.0013	0.1549	0.1516	0.9450	−0.0015	0.0843	0.0783	0.9280
		0.5	−0.0054	0.1730	0.1665	0.9360	−0.0039	0.0883	0.0853	0.9390

For comparison, we consider three combinations of the observation process $(N_{i1}(t), N_{i2}(t))$: (i) (HPP, HPP), (ii) (HPP, NHPP), and (iii) (NHPP, NHPP). Tables 1–3 present the obtained simulation results on estimation of β_1 and β_2 for the three situations with sample size $n = 100$, or 200, the true values of $\beta_0 = (-1, 1)$, $\rho_z = 0.5$, 0, or -0.5, $\rho = 0.5$, 0, or -0.5, and $\tau = 18$. Table 4 shows the obtained simulation results for the (HPP, HPP) case with $\tau = 10$, and similar results can be obtained for the other two cases. The tables include the estimated bias (BIAS) given by the average of proposed estimates of β minus the true value, the sample standard error (SSE) of the proposed estimates, the mean of the estimated standard error (ESE), and the empirical 95% coverage probabilities (CP). These results indicate that the proposed estimate seems to be unbiased and the proposed variance estimation procedure provides reasonable estimates. Also the results on the empirical coverage probabilities indicate that the normal approximation seems to be appropriate when the sample size increases, as expected.

6. Application

To illustrate the proposed methodology, we consider a skin cancer chemoprevention trial as mentioned in Section 1. It is a double-blinded and placebo-controlled randomized phase III clinical trial. The object of the study is a population of the individuals with a history of two related types of non-melanoma skin cancers, basal cell carcinoma and squamous cell

Table 3
Simulation results for β under (NHPP, NHPP) observation process with $\tau = 18$.

n	ρ	ρ_z	$\hat{\beta}_1$				$\hat{\beta}_2$			
			BIAS	SSE	ESE	CP	BIAS	SSE	ESE	CP
100	−0.5	−0.5	−0.0041	0.1386	0.1309	0.9200	−0.0003	0.0933	0.0882	0.9400
		0	0.0010	0.1266	0.1217	0.9440	0.0043	0.0840	0.0825	0.9450
		0.5	0.0055	0.1238	0.1118	0.9220	0.0003	0.0816	0.0778	0.9320
	0	−0.5	0.0036	0.0646	0.0603	0.9190	0.0001	0.0411	0.0397	0.9410
		0	0.0021	0.0558	0.0558	0.9430	−0.0001	0.0387	0.0377	0.9360
		0.5	−0.0020	0.0537	0.0523	0.9320	−0.0004	0.0356	0.0357	0.9520
	0.5	−0.5	−0.0055	0.2297	0.2130	0.9160	−0.0002	0.1132	0.1100	0.9410
		0	−0.0042	0.2424	0.2235	0.9230	−0.0030	0.1185	0.1160	0.9450
		0.5	−0.0223	0.2599	0.2419	0.9240	−0.0010	0.1276	0.1254	0.9410
200	−0.5	−0.5	−0.0046	0.0959	0.0954	0.9480	0.0015	0.0627	0.0632	0.9500
		0	−0.0008	0.0921	0.0880	0.9380	0.0010	0.0604	0.0595	0.9490
		0.5	−0.0021	0.0832	0.0815	0.9440	0.0004	0.0561	0.0559	0.9460
	0	−0.5	−0.0031	0.0437	0.0430	0.9320	0.0014	0.0307	0.0283	0.9270
		0	−0.0014	0.0418	0.0399	0.9430	−0.0006	0.0267	0.0267	0.9480
		0.5	−0.0020	0.0375	0.0373	0.9440	−0.0017	0.0255	0.0254	0.9470
	0.5	−0.5	−0.0073	0.1624	0.1547	0.9350	−0.0030	0.0801	0.0793	0.9520
		0	−0.0070	0.1731	0.1627	0.9410	0.0027	0.0848	0.0840	0.9460
		0.5	−0.0065	0.1834	0.1778	0.9310	−0.0011	0.0923	0.0904	0.9480

Table 4
Simulation results for β under (HPP, HPP) observation process with $\tau = 10$.

n	ρ	ρ_z	$\hat{\beta}_1$				$\hat{\beta}_2$			
			BIAS	SSE	ESE	CP	BIAS	SSE	ESE	CP
100	−0.5	−0.5	0.0006	0.1341	0.1262	0.9350	−0.0022	0.0882	0.0830	0.9320
		0	−0.0014	0.1250	0.1173	0.9280	0.0001	0.0798	0.0784	0.9360
		0.5	0.0041	0.1177	0.1101	0.9310	0.0012	0.0756	0.0741	0.9440
	0	−0.5	−0.0027	0.0525	0.0492	0.9230	−0.0005	0.0308	0.0292	0.9290
		0	−0.0048	0.0491	0.0488	0.9380	0.0004	0.0293	0.0287	0.9350
		0.5	−0.0007	0.0509	0.0475	0.9180	−0.0007	0.0300	0.0282	0.9350
	0.5	−0.5	−0.0141	0.2180	0.2015	0.9200	−0.0033	0.1057	0.1016	0.9350
		0	−0.0090	0.2278	0.2148	0.9320	0.0020	0.1107	0.1085	0.9430
		0.5	−0.0003	0.2490	0.2342	0.9360	0.0008	0.1229	0.1192	0.9440
200	−0.5	−0.5	−0.0012	0.0935	0.0919	0.9490	−0.0016	0.0621	0.0602	0.9460
		0	−0.0053	0.0844	0.0850	0.9570	−0.0012	0.0576	0.0564	0.9360
		0.5	−0.0016	0.0808	0.0794	0.9450	0.0031	0.0525	0.0536	0.9620
	0	−0.5	−0.0002	0.0367	0.0354	0.9310	0.0003	0.0208	0.0209	0.9510
		0	−0.0010	0.0306	0.0347	0.9330	0.0004	0.0202	0.0205	0.9460
		0.5	−0.0021	0.0353	0.0342	0.9340	−0.0003	0.0203	0.0201	0.9440
	0.5	−0.5	−0.0014	0.1494	0.1466	0.9390	0.0001	0.0760	0.0734	0.9450
		0	−0.0030	0.1581	0.1553	0.9400	−0.0056	0.0837	0.0787	0.9310
		0.5	−0.0057	0.1745	0.1697	0.9330	−0.0055	0.0860	0.0859	0.9460

carcinoma denoted by Type 1 and Type 2 cancers, respectively. The purpose of the trial is to assess the effectiveness of 0.5 $\text{g}/\text{m}^2/\text{day}$ PO difluoromethylornithine (DFMO) in reducing the recurrence rates of both types of new skin cancers. The data set can be founded from data set III of Chapter 9 in Sun and Zhao (2013).

The data set of the study actually involves 290 patients with effective information. They were randomly assigned to either the placebo group (147) or the DFMO group (143). The patients were scheduled to be assessed or observed every 6 months for the recurrence of one of the two non-melanoma skin cancers. However, the real observation times differ from individual to individual and so as the follow-up times.

For the analysis, for patient i, we took $Y_{ik}(t)$ as the natural logarithm of the number of observed Type k cancers at time t on the patient plus 1 to avoid 0, $i = 1, \ldots, 290$, $k = 1, 2$. The same covariate $\mathbf{X}_i$ was considered for these two different types of skin cancers. We set the first and second component of $\mathbf{X}_i$ to be the number of prior skin cancers from the first diagnosis to randomization, and the age, respectively; the third component to be 1 if the ith patient is a male and 0 otherwise; and the fourth component to be 1 if the ith patient was given the DFMO and 0 for placebo. We got a set of bivariate longitudinal data.

First, we analyzed this data set by the proposed models with the same covariate effect on different types of longitudinal responses. The estimation results for the covariate effect $\hat{\beta}$ obtained by the proposed estimation procedure are given in Table 5. These results suggested that the DFMO treatment was not effective on reducing the occurrence rate of the two skin

Table 5
Application results for the skin cancer chemoprevention trial.

Model I	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$				
Estimate	0.0097	−0.0001	0.0159	−0.0036				
ESE	0.0015	0.0005	0.0090	0.0091				
p-value	0.0000	0.7983	0.0775	0.6937				
Model II	Basal cell carcinoma				Squamous cell carcinoma			
	$\hat{\beta}_{11}$	$\hat{\beta}_{21}$	$\hat{\beta}_{31}$	$\hat{\beta}_{41}$	$\hat{\beta}_{12}$	$\hat{\beta}_{22}$	$\hat{\beta}_{32}$	$\hat{\beta}_{42}$
Estimate	0.0118	−0.0017	0.0139	−0.0131	0.0077	0.0014	0.0180	0.0061
ESE	0.0017	0.0009	0.0152	0.0143	0.0023	0.0005	0.0113	0.0116
p-value	0.0000	0.9763	0.1787	0.8207	0.0004	0.0015	0.0561	0.3004

Model I: the proposed models with the same covariate effect; Model II: the proposed models with different covariate effects.

cancers. Moreover, the age did not have a significant relationship with the occurrence rate. The occurrence rate, however, was positively and significantly affected by the number of the prior skin cancers as suggested by the *p*-value. These results are consistent with those in Li et al. (2011) and Zhang et al. (2013). In addition, by our approach, we can concluded that the recurrence rate of skin cancer for male patients was slightly higher than that for female patients at significance level of 0.08. To check the overall fit of our proposed model, we found out that the *p*-value for $\sup_{t,\mathbf{x}} |\Phi(t,\mathbf{x})|$ was 0.139 based on 1000 realizations. This suggests that these models seem to be appropriate for the skin cancer chemoprevention data considered here.

Second, we analyzed this data set by the proposed models with different covariate effects on different types of longitudinal responses, that is,

$$E\{Y_{ik}(t)|\mathbf{X}_i, \mathbf{Z}_i\} = \mu_{0k}(t) + \beta_k'\mathbf{X}_i + h_k(Z_{ik}), \quad k = 1, 2,$$

where β_k denotes the vector of regression parameters for Type k. Then the proposed procedure yielded the estimation results in the second part of Table 5. From the table, we can see that the DFMO treatment did not have a significant effect on reducing the recurrence rates of both types of new skin cancers, while the numbers of the prior skin cancers had a significant positive effect on the recurrence rates of both types of new skin cancers. The age of patients did not have a significant effect on the occurrence rate of basal cell carcinoma cancer, but it had a significant positive effect on the recurrence rate of squamous cell carcinoma cancer; the gender of patients did not have a significant effect on the recurrence rate of basal cell carcinoma cancer, but the recurrence rate of squamous cell carcinoma cancer for male patients was slightly higher than that for female patients at significance level of 0.06. Furthermore, the *p*-value for $\sup_{t,\mathbf{x}} |\Phi(t,\mathbf{x})|$ for the proposed models was 0.314 based on 1000 realizations. These analysis results indicate that the proposed models with different covariate effects seem more plausible for these data.

7. Concluding remarks

In this paper, we have proposed flexible joint models for the analysis of the multivariate longitudinal outcomes with irregular and informative observation processes, where no restrictive condition is made on both the correlation between different types of response process and the different correlations between each type of response process and the corresponding observation process. For estimation of the covariate effect on the outcomes, we have proposed a novel estimating equation-based inference procedure, which depends on neither the form of the link of the frailty nor the distribution of the frailty. The resulting estimators have explicit expressions and so the proposed inference procedures are easy to implement. The asymptotic properties of the proposed estimators are established. The finite-sample properties of the proposed estimates are evaluated through simulation studies and a set of multivariate longitudinal data from a skin cancer chemoprevention trial is analyzed.

In particular, we have developed a robust and easy to implement method since our estimation approach does not involve estimation of unknown parameters in the observation process model, but estimating unknown parameters for the observation process model are required by Zhao et al. (2012)'s estimation procedure.

Note that for the simplicity of presentation, our proposed estimation approach assume that the covariate effects on different types of response processes are the same. Actually, this may not be true in practice. Thus, instead of model (1), one may consider

$$E\{Y_{ik}(t)|\mathbf{X}_{ik}, \mathbf{Z}_i\} = \mu_{0k}(t) + \beta_k'\mathbf{X}_{ik} + h_k(Z_{ik}),$$

where β_k denotes the vector of regression parameters, $k = 1, \ldots, K$. It is straightforward to generalize the proposed estimation approach to this situation as we have considered in the application section. Furthermore, the covariates and covariate effects may be time-dependent in some applications. For this problem, one may consider the following model for the response process $Y_{ik}(t)$,

$$E\{Y_{ik}(t)|\mathbf{X}_i, \mathbf{Z}_i\} = \mu_{0k}(t) + \beta(t)'\mathbf{X}_i(t) + h_k(Z_{ik})$$

where $\beta(t)$ and $\mathbf{X}_i(t)$ are defined as before except being time-dependent.

In addition, it is also interesting to consider the generalized linear mixed model (GLMM) for longitudinal response processes as follows,

$$g(E\{Y_{ik}|\mathbf{X}_i,\mathbf{Z}_i\}) = \mu_{0k}(t) + \beta'\mathbf{X} + h_k(Z_{ik}),$$

where $g(\cdot)$ is a GLM link function. However, it is not straightforward to extend our proposed estimating equation approach to this model. A new estimation method needs to be developed for the further work.

Acknowledgments

The authors would like to thank the Editor, Professor Jae Chang Lee, the Associate Editor and the two reviewers for their constructive and insightful comments and suggestions that greatly improved the paper. Deng's research is partly supported by the Natural Science Foundation of China (No. 11401443) and the Fundamental Research Funds for the Central Universities (No. 2042014kf0075), and Zhao's research is partly supported by the Research Grant Council of Hong Kong (503513), the Natural Science Foundation of China (No. 11371299), and The Hong Kong Polytechnic University.

Appendix. Proofs

To study the asymptotic properties of the proposed estimators, we need the following regularity conditions.

C1 $P(C \geq \tau) > 0$ and $E(Z_{1k}^2) < \infty, k = 1, 2.$
C2 $\mathbf{X}$ is bounded, $Y_{ik}(\tau), N_{ik}(\tau)$ are bounded almost surely, $k = 1, \ldots, K, i = 1, \ldots, n.$
C3 $A \triangleq E\left[\sum_{k=1}^{K} W_1 \bar{\mathbf{X}}_{1k}^{\otimes 2} m_{1k}\right]$ is nonsingular.

Proof of the consistency of $\hat{\theta}$

Note that

$$\begin{aligned}
U(\theta_0) &= n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K} W_i\bar{\mathbf{X}}_{ik}\{\bar{N}_{ik} - m_{ik}\theta_0'\bar{\mathbf{X}}_{ik}\} \\
&= E\left\{\sum_{k=1}^{K} W_1\bar{\mathbf{X}}_{1k}(\bar{N}_{1k} - m_{1k}\theta_0'\bar{\mathbf{X}}_{1k})\right\} + o_p(1) \\
&= E\left\{\sum_{k=1}^{K} W_1\bar{\mathbf{X}}_{1k}\left[E\{\bar{N}_{1k}|\mathbf{X}_1\} - E\{m_{1k}|\mathbf{X}_1\}\theta_0'\bar{\mathbf{X}}_{1k}\right]\right\} + o_p(1) \\
&= E\left\{\sum_{k=1}^{K} W_1\bar{\mathbf{X}}_{1k}E\{m_{1k}|\mathbf{X}_1\}\left[\beta_0'\mathbf{X}_1 + \alpha_{0k} - \theta_0'\bar{\mathbf{X}}_{1k}\right]\right\} + o_p(1) \\
&= o_p(1),
\end{aligned}$$

and

$$\hat{A}(\theta) = -\frac{\partial U(\theta)}{\partial \theta} = n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K} W_i\bar{\mathbf{X}}_{ik}^{\otimes 2} m_{ik}$$

converges uniformly to a positive definite matrix A over θ. Thus the solution $\hat{\theta}$ of the estimating equation $U(\theta) = 0$ is unique and consistent for θ_0.

Proof of the asymptotic normality of $\hat{\theta}$

In this part, we will derive the asymptotic normality of $\sqrt{n}(\hat{\theta} - \theta_0)$. Note that by Taylor expansion, we have

$$\begin{aligned}
\sqrt{n}(\hat{\theta} - \theta_0) &= \left[-\frac{\partial U(\theta)}{\partial \theta}\bigg|_{\theta=\theta_0}\right]^{-1}\sqrt{n}U(\theta_0) + o_p(1) \\
&= A^{-1}n^{-1/2}\sum_{i=1}^{n}\phi_i + o_p(1).
\end{aligned}$$

It follows from the central limit theorem and Slutsky theorem that the asymptotic normality of $\hat{\beta}$ holds as stated in Section 3.

Proof of the asymptotic properties of $\Phi(t,\mathbf{x})$.

We will prove the weak convergence of $\Phi(t,\mathbf{x})$ under models (1) and (2). Note that

$$\begin{aligned}\hat{\mathcal{A}}_{0k}(t)-\mathcal{A}_{0k}(t) &= n^{-1}\sum_{i=1}^{n}\int_0^t \frac{Y_{ik}(u)-\hat\beta'\mathbf{X}_i}{n^{-1}\sum\limits_{i=1}^{n} m_{ik}}d\tilde N_{ik}(u)-\mathcal{A}_{0k}(t)\\ &= -\left[n^{-1}\sum_{i=1}^{n}\int_0^t \frac{\mathbf{X}_i}{n^{-1}\sum\limits_{i=1}^{n} m_{ik}}d\tilde N_{ik}(u)\right]'(\hat\beta-\beta_0)\\ &\quad + n^{-1}\sum_{i=1}^{n}\int_0^t \frac{Y_{ik}(u)-\beta_0'\mathbf{X}_i}{n^{-1}\sum\limits_{i=1}^{n} m_{ik}}d\tilde N_{ik}(u)-\mathcal{A}_{0k}(t)\\ &= -\left[n^{-1}\sum_{i=1}^{n}\int_0^t \frac{\mathbf{X}_i}{n^{-1}\sum\limits_{i=1}^{n} m_{ik}}d\tilde N_{ik}(u)\right]'(\hat\beta-\beta_0)+n^{-1}\sum_{i=1}^{n}\int_0^t\frac{dR_{ik}(t)}{n^{-1}\sum\limits_{i=1}^{n} m_{ik}},\end{aligned}$$

where

$$R_{ik}(t)=\int_0^t [Y_{ik}(t)-\beta_0'\mathbf{X}_i]d\tilde N_{ik}(u)-m_{ik}\mathcal{A}_{0k}(t).$$

Thus,

$$\begin{aligned}\Phi(t,\mathbf{x}) &= n^{-1/2}\sum_{i=1}^{n}\sum_{k=1}^{K} I(\mathbf{X}_i\le\mathbf{x})R_{ik}(t)-\left[n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K}\int_0^t I(\mathbf{X}_i\le\mathbf{x})\mathbf{X}_i d\tilde N_{ik}(u)\right]'\sqrt{n}(\hat\beta-\beta_0)\\ &\quad - n^{-1/2}\sum_{i=1}^{n}\sum_{k=1}^{K}\int_0^t I(\mathbf{X}_i\le\mathbf{x})m_{ik}d\{\hat{\mathcal{A}}_{0k}(u)-\mathcal{A}_{0k}(u)\}\\ &= n^{-1/2}\sum_{i=1}^{n}\sum_{k=1}^{K}\int_0^t\left\{I(\mathbf{X}_i\le\mathbf{x})-\frac{S_k(\mathbf{x})}{S_{k0}}\right\}dR_{ik}(u)-B(t,\mathbf{x})'\sqrt{n}(\hat\beta-\beta_0)\\ &= n^{-1/2}\sum_{i=1}^{n}\sum_{k=1}^{K}\int_0^t\left\{I(\mathbf{X}_i\le\mathbf{x})-\frac{s_k(\mathbf{x})}{s_{k0}}\right\}dR_{ik}(u)-b(t,\mathbf{x})'\sqrt{n}(\hat\beta-\beta_0)+o_p(1), \qquad (8)\end{aligned}$$

where s_{k0}, $s_k(\mathbf{x})$ and $b(t,\mathbf{x})$ are the limits for S_{k0}, $S_k(\mathbf{x})$ and $B(t,\mathbf{x})$ respectively. The tightness of the first term on the right hand side of (8) follows directly from the arguments in Appendix A.5 of Lin et al. (2000). The second term is also tight because that $\sqrt{n}(\hat\beta-\beta_0)$ converges in distribution and $b(t,\mathbf{x})$ is some deterministic function. Thus $\Phi(t,\mathbf{x})$ is tight.

Based on (8), we can rewrite $\Phi(t,\mathbf{x})$ as

$$\Phi(t,\mathbf{x})=n^{-1/2}\sum_{i=1}^{n}\Psi_i(t,\mathbf{x})+o_p(1),$$

with

$$\Psi_i(t,\mathbf{x})=\sum_{k=1}^{K}\int_0^t\left\{I(\mathbf{X}_i\le\mathbf{x})-\frac{s_k(\mathbf{x})}{s_{k0}}\right\}dR_{ik}(u)-b(t,\mathbf{x})'a_i+o_p(1),$$

where a_i is the vector $\hat A^{-1}\phi_i$ without the last K entries. The multivariate central limit theorem, together with the tightness of Φ implies that $\Phi(t,\mathbf{x})$ convergences weakly to a zero-mean Gaussian process which can be approximated by the zero mean Gaussian process $\tilde\Phi(t,\mathbf{x})$ defined in (6).

References

Cheng, S.C., Wei, L.J., Ying, Z., 1997. Predicting survival probabilities with semiparametric transformation models. J. Amer. Statist. Assoc. 92, 227–235.
Lee, L.-Y., 2008. Nonparametric and semiparametric models for multivariate panel count data (Ph.D. dissertation), University of Wisconsin-Madison.
Li, N., 2011. Semiparametric transformation models for panel count data (Ph.D. Dissertation), University of Missouri, Columbia.

Li, Y., He, X., Wang, H., Zhang, B., Sun, J., 2015. Semiparametric regression of multivariate panel count data with informative observation times. J. Multivariate Anal. 140, 209–219.

Li, N., Park, D.H., Sun, J., Kim, K., 2011. Semiparametric transformation models for multivariate panel count data with dependent observation process. Canad. J. Statist. 39, 458–474.

Liang, Y., Lu, W., Ying, Z., 2009. Joint modeling and analysis of longitudinal data with informative observation times. Biometrics 65, 377–384.

Lin, D.Y., Wei, L.J., Yang, I., Ying, Z., 2000. Semiparametric regression for the mean and rate functions of recurrent events. J. R. Stat. Soc. Ser. B 62, 711–730.

Lin, D.Y., Ying, Z., 2001. Semiparametric and nonparametric regression analysis of longitudinal data. J. Amer. Statist. Assoc. 96, 103–126.

Sun, J., Park, D., Sun, L., Zhao, X., 2005. Semiparametric regression analysis of longitudinal data with informative observation times. J. Amer. Statist. Assoc. 100, 882–889.

Sun, J., Sun, L., Liu, D., 2007. Regression analysis of longitudinal data in the presence of informative observation and censoring times. J. Amer. Statist. Assoc. 102, 1397–1406.

Sun, J., Zhao, X., 2013. Statistical Analysis of Panel Count Data. Springer.

Welsh, A.H., Lin, X., Carroll, R.J., 2002. Marginal longitudinal nonparametric regression: locality and efficiency of spline and Kernel Methods. J. Amer. Statist. Assoc. 97, 482–493.

Zhang, H., Zhao, H., Sun, J., Wang, D., Kim, K., 2013. Regression analysis of multivariate panel count data with an informative observation process. J. Multivariate Anal. 119, 71–80.

Zhao, H., Li, L., Sun, J., 2013. Semiparametric analysis of multivariate panel count data with dependent observation processes and a terminal event. J. Nonparametr. Stat. 25, 379–394.

Zhao, X., Tong, X., Sun, L., 2012. Joint analysis of longitudinal data with dependent observation times. Statist. Sinica 22, 317–336.

Zhou, J., Zhao, X., Sun, L., 2013. A new inference approach for joint models of longitudinal data with informative observation and censoring Times. Statist. Sinica 23, 571–593.