Joint Likelihood Estimation for Joint Modeling Survival and Multiple Longitudinal Processes

Runze Li, Xiaoyu Liu and Megan E. Piper

October 12, 2016

Abstract

Motivated from an empirical analysis of data collected by a smoking cession study, we propose a joint model (JM) of survival data and multiple longitudinal covariate processes, develop an estimation procedure for this model using likelihood-based approach, and further establish the consistency and asymptotic normality of the resulting estimate. Computation for the proposed likelihood-based approach in the joint modeling is particularly challenging since the estimation procedure involves numerical integration over multi-dimensional space for the random effects in the JM. Existing numerical integration methods become ineffective or infeasible for the JM. We introduce a numerical integration method based on computer experimental designs for the JM. We conduct Monte Carlo simulation to examine the finite sample performance of the procedure procedure and compare the new numerical integration method with existing ones. We further illustrate the proposed procedure via an empirical study of smoking cession data.

Abbreviated Title: JM survival and multiple longitudinal processes
Key Words and phrases: Cox's model, mixed effect models, partial likelihood.

*Runze Li is Distinguished Professor, Department of Statistics and the Methodology Center, the Pennsylvania State University, University Park, PA 16802-2111, USA. Email: rzli@psu.edu. Xiaoyu Liu is a Ph.D. student, Department of Statistics, the Pennsylvania State University, University Park, PA 16802-2111, USA. Email: xol5086@psu.edu. Megan Piper is Assistant Professor, The Center for Tobacco Research and Intervention, University of Wisconsin, Madison, WI 53711-2027, USA. email MEP@ctri.wisc.edu. Li and Liu's research was supported by NIDA, NIH grants P50 DA10075 and P50 DA036107. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIDA or the NIH.

1 Introduction

In the past two decades, driven by the needs to explore the relationship between longitudinal covariate process and time to event in biomedical and public health research, statisticians have developed and modified joint modeling approach to simultaneously analyze the two types of processes via their shared information (Wulfsohn and Tsiatis, 1997; Song et al., 2002b; Hsieh et al., 2006; Faucett and Thomas, 1996; Faucett et al., 1998; Henderson et al., 2000; Tsiatis and Davidian, 2001). Longitudinal and survival data are originally specified by mixed-effects models and hazard models, respectively. However, such naive separate estimation has been proved to yield great bias in regression coefficients due to the ignorance of measurement errors and missing observation at event time (Prentice, 1982; Tsiatis et al., 1995). To address such issue, the earliest joint modeling approach was developed to examining whether CD4 counts serve as a good biomarker for survival time of HIV patients (Tsiatis et al., 1995), and the later improvement in modeling techniques has been applied to more medical and public health studies (Wang and Taylor, 2001; Yu et al., 2004; Liu, 2009; Yu and Ghosh, 2010).

Basically, there are two broad types of modeling techniques for joint modeling problems. The first type of methods links the two models by modifying partial likelihood equations of Cox's model to incorporate longitudinal predictors. Such strategies, including early regression calibration approach (Tsiatis et al., 1995) and conditional score approach (Tsiatis and Davidian, 2001; Song et al., 2002b), mainly focus on the regression coefficients in Cox's models and do not concern too much about the parameters in longitudinal models. Regression calibration is a "two-step" estimation approach which estimates the conditional expectations of true longitudinal predictors given the observed measurements in the first step, and substitute them into the partial likelihood as predictors to solve for the regression coefficients in the second step. No large sample property has been developed for this approach and it is con-

sidered highly dependent on the model assumption of the longitudinal processes (Wulfsohn and Tsiatis, 1997). Conditional score approach (Tsiatis and Davidian, 2001; Song et al., 2002b) treats the random effects in the mixed-effects models as nuisance parameters and estimates survival regression coefficients from a modified partial likelihood which conditions on a complete and sufficient estimator of the random effects. The estimates from conditional score approach, though possessing asymptotic properties, have not been proved efficient.

The second broad type of methods links the two models through their joint likelihood, and the parameters are estimated by maximizing the likelihood function. Since the likelihood function involves integration over the random effects, it is challenging to find the maximizer. One solution is to use Bayesian techniques (Faucett and Thomas, 1996; Xu and Zeger, 2001b; Wang and Taylor, 2001; Brown and Ibrahim, 2003). Based on the joint likelihood, Bayesian approach makes assumptions about prior distributions of all the parameters, and update the parameter estimates using Gibbs sampling from the full conditional distributions of each parameter given the observed data and current estimates of all the other parameters. The merit of Bayesian approaches is that they are not constrained by the dimension random effects, whereas the drawback is that the procedure itself is quite computationally intensive. An alternative is maximum joint likelihood approach (Wulfsohn and Tsiatis, 1997), which maximizes the joint likelihood equation via Expectation-Maximization (EM) algorithm, treating the unobservable random effects as missing values. Up to now, It has been proposed for joint model with only a single longitudinal process. This method is dimension-sensitive, but has good theoretical properties. Zeng and Cai (2005) proves that in the single-covariate setting, the maximum likelihood estimators and efficient and enjoy large sample property.

While most joint modeling literatures focus on the setting with a single longitudinal predictor in the survival model, fewer authors consider the model with multiple longitudinal predictors, which is a more flexible and useful model setting (Xu and Zeger, 2001a; Song et al., 2002a; Huang et al., 2001; Ibrahim et al., 2004; Brown et al., 2005; Chi and Ibrahim,

2006; Albert and Shih, 2010; Hatfield et al., 2011).

The challenges for joint modeling with multiple longitudinal predictors are twofold. First, the number of random effects grows as the number of longitudinal predictors increases. This results in higher dimension integrals in the objective functions, which are already time-consuming enough to optimize even in the single-covariate case. Second, the correlations among the multiple longitudinal processes over time have to be considered in the model. This complicates the theoretical development.

Most of the related studies (Xu and Zeger, 2001a; Ibrahim et al., 2004; Brown et al., 2005; Chi and Ibrahim, 2006; Hatfield et al., 2011) resort to Bayesian approach to handle the complex computation issue. Song et al. (2002a) and Albert and Shih (2010) avoid it by not considering the likelihood-based approaches, they instead use the first type of methods mentioned above, and applied conditional score and modified regression calibration techniques, respectively, to solve the problem. Huang et al. (2001) is the only literature we found that uses EM algorithm to maximize the complicated joint likelihood equation. However, the model setting considered in that study is specifically designed for the data of interest, and quite different from the general joint modeling framework. For example, it adopts discrete latent variables in the model and thus avoids the high-dimensional integral problems. Among all the studies, Song et al. (2002a) extends the asymptotic properties of the conditional score estimators (Tsiatis and Davidian, 2001) to the multiple longitudinal-covariate setting. Most of the studies focus more on the application of the methodology than theory establishment.

Motivated by an empirical analysis of smoking cession data, we proposed a joint model setting with multiple longitudinal covariate process. We developed an estimation procedure for the proposed joint model based on joint likelihood approach. We systematically studied the asymptotic property of the proposed estimation procedure. We establish the consistency and asymptotic normality of the resulting estimate by using the formulation of Zeng and

Cai (2005), in which the authors established the sampling property of maximum likelihood estimate for joint model with a single longitudinal covariate process. The theoretical development for joint model with multiple longitudinal processes is much more challenging than that for the one with a single longitudinal covariate process since we have to deal with the covariance among the multiple longitudinal covariates rather than variance for the single longitudinal covariate.

As mentioned before, it is computational challenging in carrying out the estimation procedure based on the maximum joint likelihood approach because the use of EM algorithm to optimize the objective function involves numerical integration over a high-dimensional space for the shared random effect in the E-step of the EM algorithm. We proposed an approach to carrying out the numerical integration based on design of experiment interpolation technique (DoIt, Joseph 2012). We conducted Monte Carlo simulation to compare the proposed numerical integration method with existing ones under the setting of joint models with single covariate process. Our numerical results indicate that the proposed method performs very well in terms of computing numerical and statistical estimation accuracy. We further examine the performance of the proposed numerical method for joint model with multiple covariate processes Our numerical results implies that the proposed numerical integration method works very well.

Another challenge of joint modeling is to estimate standard errors of the resulting estimate. We propose an estimation method for the standard error by bootstrap method. We conducted Monte Carlo simulation to examine finite sample performance of the proposed estimation procedures including estimation of the parameters and estimation of their standard errors. Our numerical results indicate that the proposed estimation procedure performs very well with moderate sample size. We further applied the proposed estimation method to a smoking cessation study (Piper et al., 2009) to assess the relationship between time to lapse and multiple longitudinal measurements of withdrawal symptoms. We find that the results

of joint models with multiple longitudinal covariates offer deeper insights into the applied study than the model with single longitudinal covariate.

The rest of this paper is organized as follows. We describes the model setting, introduces the estimation approach, explains the computing techniques, and presents the asymptotic properties of the proposed estimators in Section 2. Two simulation examples and a real data example are presented in Section 3. Conclusion and discussion are given in Section 4. Technical proofs are given in Section 5.

2 Joint likelihood approach

In this section, we first present our model settings, and then propose the maximum likelihood approach to estimating the model parameters. A new numerical integration technique is introduced to deal with computational issues in maximizing the joint likelihood with integrals, and the asymptotic properties of the resulting estimates are established.

2.1 Model settings

Suppose that a random sample consists of observations of survival time to an event of interest, multiple longitudinal processes, and time-invariant covariates from n subjects. The multiple longitudinal processes of ith subject are observed at $t_{i1}, \ldots, t_{iN_i}$. The data that we are interested in modeling would be

$$(T_i, \mathbf{Z}_i, \mathcal{X}_i),\ i = 1, 2, \ldots, n, \tag{2.1}$$

where T_i is the time to the event, and $\mathbf{Z}_i = (Z_{i1}, \cdots, Z_{iq})^T$ is a vector consisting of q-dimensional time-independent covariates. Denote by $\mathbf{X}_i(t) = (X_{i1}(t), \ldots, X_{ip}(t))^T$ the p-dimensional longitudinal covariates recorded at time t, and $\mathcal{X}_i = (\mathbf{X}_i(t_{i1})^T, \ldots, \mathbf{X}_i(t_{i,N_i})^T)^T$ is a $N_i \times p$ matrix representing the longitudinal covariates for the ith subject at all the

observational time points.

The major goal of joint modeling is to elucidate the relationship between the survival time T_i and the covariates $\{\mathbf{Z}_i, \mathcal{X}_i\}$. In practice, T_i and $\mathcal{X}_i$ may not be observable due to censoring and measurement error, respectively. In survival data, let C_i be the censoring time for the ith subject. Denote $V_i = \min(T_i, C_i)$ and $\Delta_i = I(T_i \leq C_i)$, the censoring indicator. Under the right-censoring scheme, the observed survival data for the ith subject is (V_i, Δ_i) instead of T_i. In longitudinal data, instead of the true longitudinal process $\mathcal{X}_i$, what can be observed are the realizations of the error-contaminated processes at times $t_{i1}, \ldots, t_{iN_i}$, which are denoted by $\mathcal{W}_i$ as follows

$$\mathcal{W}_i = (\mathbf{W}_i^T(t_{i1}), \ldots, \mathbf{W}_i^T(t_{iN_i}))^T, \tag{2.2}$$

where $\mathbf{W}_i(t) = (W_{i1}(t), \ldots, W_{ip}(t))^T$.

As a result, the observed data from the i-subject becomes

$$D_o = (V_i, \Delta_i, \mathbf{Z}_i, \mathcal{W}_i, \mathbf{t}_i), \tag{2.3}$$

where $t_i = (t_{ij} : 1 \leq j \leq N_i \text{ and } t_{ij} \leq V_i)$ is the observed time points for the ith individual. Since the longitudinal process is observed until the event or censoring happens, the observed covariate process is also truncated at V_i, i.e., $\{\mathbf{W}_i(t_{ij}) : t_{ij} \leq V_i\}$.

To take into account the measurement error or biological variation, we specify the observed longitudinal covariate $W_{ik}(t)$ by the following linear mixed-effects model

$$\begin{aligned} W_{ik}(t) &= X_{ik}(t) + e_{ik}(t), \\ X_{ik}(t) &= \tilde{\boldsymbol{\rho}}_k(t)^T \boldsymbol{\mu}_k + \boldsymbol{\rho}_k(t)^T \mathbf{b}_{ik}, \quad k = 1, \ldots, p, \end{aligned} \tag{2.4}$$

where $e_{ik}(t)$ is a random error with mean zero, and $X_{ik}(t)$ is the unobservable underlying longitudinal process composed of a fixed part, $\tilde{\boldsymbol{\rho}}_k(t)^T \boldsymbol{\mu}_k$, and a random part, $\boldsymbol{\rho}_k(t)^T \mathbf{b}_{ik}$. $\boldsymbol{\rho}_k(t)$ and $\tilde{\boldsymbol{\rho}}_k(t)$ are two sets of functions of time t, including polynomial basis functions as a special case. For example, $\boldsymbol{\rho}(t) = (1, t)^T$ yields the simplest linear function of time. The

form of the function is flexible and may vary across different longitudinal processes, and their corresponding dimensions (i.e., $\dim(\boldsymbol{\rho}_k(t)) = d_k$ and $\dim(\ \tilde{\boldsymbol{\rho}}_k(t)) = \tilde{d}_k$) would vary accordingly. $\boldsymbol{\mu}_k$ is a $\tilde{d}_k \times 1$ vector of fixed effects, and $\mathbf{b}_{ik}$ is a $d_k \times 1$ vector of random effects accounting for the within-subject variation.

In practice, it is typical to assume that the observed longitudinal covariates $\mathcal{W}_i$ are independent for different individuals, but are correlated across time and across different covariates at the same time for the same person. In order to take into account these independence and correlation structures, we assume $\mathbf{b}_{ik}$ are independent across i and k, $\mathbf{e}_i(t) = (e_{i1}(t), \ldots, e_{ip}(t))^T$ are independent across i, and the random effects $\mathbf{b}$ and the measurement errors $\mathbf{e}$ are independent. Thus the correlations of $W_{ik}(t)$ over time are modeled via the variance of $\mathbf{b}_{ik}$, and the within-subject correlations across the p covariates are modeled via the variance covariance structure of $\mathbf{e}_{i.}(t)$. We further assume that $\mathbf{b}_{ik} \sim N_{d_k}(\mathbf{0}, \boldsymbol{\Sigma}_{bk})$, and $\mathbf{e}_i(t) \sim N_p(\mathbf{0}, \boldsymbol{\Sigma}_e)$.

If we write equation (2.4) in matrix form, it becomes

$$\mathbf{W}_i(t) = \tilde{\boldsymbol{\rho}}(t)^T \boldsymbol{\mu} + \boldsymbol{\rho}(t)^T \mathbf{b}_i + \mathbf{e}_i(t), \tag{2.5}$$

where $\boldsymbol{\rho}(t)$ and $\tilde{\boldsymbol{\rho}}(t)$ are $d \times p$- and $\tilde{d} \times p$-dimensional matrix containing p basis functions of time, respectively, i.e.,

$$\boldsymbol{\rho}(t) = \begin{pmatrix} \boldsymbol{\rho}_1(t)^T & & \\ & \ddots & \\ & & \boldsymbol{\rho}_p(t)^T \end{pmatrix}^T, \quad \tilde{\boldsymbol{\rho}}(t) = \begin{pmatrix} \tilde{\boldsymbol{\rho}}_1(t)^T & & \\ & \ddots & \\ & & \tilde{\boldsymbol{\rho}}_p(t)^T \end{pmatrix}^T.$$

$\boldsymbol{\mu} = (\boldsymbol{\mu}_1^T, \ldots, \boldsymbol{\mu}_p^T)^T$ is the $\tilde{d}$-dimensional vector consists of the p fixed effects, and $\mathbf{b}_i = (\mathbf{b}_{i1}^T, \ldots, \mathbf{b}_{ip}^T)^T$ is a d-dimensional vector consists of the p random effects. Note that $d = \sum_{k=1}^p d_k$ and $\tilde{d} = \sum_{k=1}^p \tilde{d}_k$, and $\mathbf{b} \sim N_d(\mathbf{0}, \boldsymbol{\Sigma}_b)$, where $\boldsymbol{\Sigma}_b = \mathrm{diag}(\boldsymbol{\Sigma}_{b1}, \ldots, \boldsymbol{\Sigma}_{bp})$.

For the survival data, conditional on $\mathbf{t}_i(t) = \{t_{ij} : t_{ij} \leq t\}$, $\mathbf{e}_{i..}(t) = \{e_{ik}(t_{ij}) : t_{ij} \leq t, k = 1, \ldots, p\}$, $T_i \geq t$, and other covariates, the relationship between the time to event T_i and the

covariates is assumed to follow a Cox model with the hazard function at t being

$$\begin{aligned} h_i(t) &= \lim_{h\to 0} h^{-1} P\{t \le T_i < t+h | T_i \ge u, C_i, \mathbf{Z}_i, \mathbf{b}_i, \mathbf{t}_i(t), \mathbf{e}_{i\cdot\cdot}(t)\} \\ &= \lambda(t)\exp\{\boldsymbol{\beta}^T(\boldsymbol{\rho}(t)^T\mathbf{b}_i) + \boldsymbol{\eta}^T\mathbf{Z}_i\}, \end{aligned} \tag{2.6}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\boldsymbol{\eta} \in \mathbb{R}^q$ are the regression coefficient of the longitudinal processes and time-independent covariates, respectively, and $\lambda(u)$ is an unspecified baseline hazard function. Model (2.6) implies the assumption that censoring time C_i, observation time points t_{ij}'s and the error $\mathbf{e}_{i\cdot\cdot}$ are non-informative in predicting the time to event T_i.

Let $\boldsymbol{\Omega} = (\Lambda(t), \boldsymbol{\theta})$ be the parameters in joint models of (2.5) and (2.6), where $\boldsymbol{\theta}$ is the set of parameters in the parametric part. Specifically

$$\boldsymbol{\theta} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_p, \mathrm{Vec}(\boldsymbol{\Sigma}_e), \mathrm{Vec}(\boldsymbol{\Sigma}_b), \boldsymbol{\beta}, \boldsymbol{\eta}),$$

where $\mathrm{Vec}(\boldsymbol{\Sigma})$ is the vector consisting of all the elements in the upper triangular part of $\boldsymbol{\Sigma}$. $\Lambda(t)$ is the cumulative baseline hazard defined as $\Lambda(t) = \int_0^t \lambda(u)du$.

2.2 Joint Likelihood Method and EM algorithm

We use maximum joint likelihood approach to estimate $\boldsymbol{\Omega}$ for the joint models with multiple longitudinal covariates. Based on the models (2.5) and (2.6) and their associated assumptions, we can write out the density of $\mathbf{b}_i$, as well as the conditional densities of $(V_i, \Delta_i|\mathbf{b}_i)$ and $(\mathcal{W}_i|\mathbf{b}_i)$. They are of the following form

$$\begin{aligned} f(V_i, \Delta_i|\mathbf{b}_i) &= \left\{\lambda(V_i)e^{\boldsymbol{\beta}^T(\boldsymbol{\rho}(V_i)^T\mathbf{b}_i)+\boldsymbol{\eta}^T\mathbf{z}_i}\right\}^{\Delta_i} \exp\left\{-\int_0^{V_i}\lambda(u)e^{\boldsymbol{\beta}^T(\boldsymbol{\rho}(u)^T\mathbf{b}_i)+\boldsymbol{\eta}^T\mathbf{z}_i}du\right\}, \\ f(\mathcal{W}_i|\mathbf{b}_i) &= \{(2\pi)^p|\boldsymbol{\Sigma}_e|\}^{-\frac{N_i}{2}} \exp\left\{-\frac{1}{2}\sum_{j=1}^{N_i}(\mathbf{W}_{ij} - \tilde{\boldsymbol{\rho}}_{ij}^T\boldsymbol{\mu} - \boldsymbol{\rho}_{ij}^T\mathbf{b}_i)^T\boldsymbol{\Sigma}_e^{-1}(\mathbf{W}_{ij} - \tilde{\boldsymbol{\rho}}_{ij}^T\boldsymbol{\mu} - \boldsymbol{\rho}_{ij}^T\mathbf{b}_i)\right\}, \\ f(\mathbf{b}_i) &= (2\pi)^{-d/2}|\boldsymbol{\Sigma}_b|^{-1/2}\exp\{-\frac{1}{2}\mathbf{b}_i^T\boldsymbol{\Sigma}_b^{-1}\mathbf{b}_i\}, \end{aligned} \tag{2.7}$$

where $\mathbf{W}_{ij}$, $\tilde{\boldsymbol{\rho}}_{ij}$ and $\boldsymbol{\rho}_{ij}$ are the vector of $\mathbf{W}_i(t) = (W_{i1}(t), \ldots, W_{ip}(t))^T$ and the matrices of $\tilde{\boldsymbol{\rho}}(t)$ and $\boldsymbol{\rho}(t)$ taking values at time t_{ij}, respectively.

Assume that the observed longitudinal processes $\mathcal{W}_i$ are independent of the observed survival process $\{V_i, \Delta_i\}$ given the random effects $\mathbf{b}_i$. By (2.7) the joint likelihood of the observed data $\mathbf{D}_o$ can be written as

$$L(\boldsymbol{\Omega}) = \prod_{i=1}^{n} L_i(\boldsymbol{\Omega}) = \prod_{i=1}^{n} \int f(V_i, \Delta_i | \mathbf{b}_i) \cdot f(\mathcal{W}_i | \mathbf{b}_i) \cdot f(\mathbf{b}_i) d\mathbf{b}_i. \tag{2.8}$$

In order to obtain the maximum likelihood estimates (MLE) of Λ, we use Breslow's estimator (Breslow, 1974) and let $\lambda(t)$ take mass only at each event time T_i for which $\Delta_i = 1$. Thus the dimension of $\lambda(t)$ reduces from infinity to a finite value $\sum_{i=1}^{n} \Delta_i$. The MLE of all the parameters in $\boldsymbol{\theta}$ and $\lambda(t)$ at each event time point are obtained by maximizing a modified version of joint likelihood (2.8), where $f(\mathcal{W}_i|\mathbf{b}_i)$ and $f(\mathbf{b}_i)$ stay the same and $f(V_i, \Delta_i|\mathbf{b}_i)$ becomes

$$\begin{aligned} f(V_i, \Delta_i | \mathbf{b}_i) &= \left\{ \lambda(V_i) e^{\boldsymbol{\beta}^T (\boldsymbol{\rho}(V_i)^T \mathbf{b}_i) + \boldsymbol{\eta}^T \mathbf{z}_i} \right\}^{\Delta_i} \\ &\quad \times \exp\left\{ - \sum_{j=1}^{n} \lambda(V_j) e^{\boldsymbol{\beta}^T (\boldsymbol{\rho}(V_j)^T \mathbf{b}_i) + \boldsymbol{\eta}^T \mathbf{z}_i} I(V_i \geq V_j, \Delta_j = 1) \right\}. \end{aligned} \tag{2.9}$$

Maximizing the joint likelihood function (2.8) is challenging. Wulfsohn and Tsiatis (1997) propose to use expectation-maximization (EM) algorithm (Dempster et al., 1977) to maximize the joint likelihood (2.8) with $p = 1$. In the EM algorithm, the unobserved random effects are treated as missing data, and the parameter estimates are updated iteratively between the expectation and maximization steps until the algorithm converges. This method has been extensively applied and demonstrated feasible and robust in the later related studies (Henderson et al., 2000; Tsiatis and Davidian, 2001; Song et al., 2002b; Hsieh et al., 2006), but all for single longitudinal process setting. Here we extend the EM method for joint likelihood with multiple longitudinal processes (i.e., $p > 1$).

Denote the logarithm of the joint likelihood contributed by the ith subject to be

$$l_i(\mathbf{\Omega}) = \log(L_i(\mathbf{\Omega})) = \log \int f(V_i, \Delta_i|\mathbf{b}_i) \cdot f(\mathcal{W}_i|\mathbf{b}_i) \cdot f(\mathbf{b}_i) d\mathbf{b}_i = \log\{f(V_i, \Delta_i, \mathcal{W}_i)\}.$$

Let θ denote a generic element in $\mathbf{\Omega}$. Take derivative of $l_i(\mathbf{\Omega})$ with respect to θ and assume the derivative and integral are interchangeable under certain conditions, we obtain

$$S_i(\theta) = \frac{\partial l_i(\mathbf{\Omega})}{\partial \theta} = \frac{\partial}{\partial \theta} \left\{ E(l_{1i}(\mathbf{b}_i)) + E(l_{2i}(\mathbf{b}_i)) + E(l_{3i}(\mathbf{b}_i)) \right\}, \tag{2.10}$$

where $l_{1i}(\mathbf{b}_{i.}) = \log\{f(V_i, \Delta_i|\mathbf{b}_{i.})\}$, $l_{2i}(\mathbf{b}_{i.}) = \log\{f(\mathbf{W}_{i..}|\mathbf{b}_{i.})\}$, $l_{3i}(\mathbf{b}_{i.}) = \log\{f(\mathbf{b}_{i.})\}$, and $E(\cdot)$ is conditional expectations of $\boldsymbol{b}_i$ given observed data D_o defined in (2.3) and the updated parameter estimates $\widehat{\mathbf{\Omega}}$ in EM algorithm. By definition of l_{ki}, their conditional expectations in (2.10) can be written out as the following expressions

$$\begin{aligned}
E\{l_{i1}(\mathbf{b}_i)\} =& \Delta_i \log\{\lambda(V_i)\} + \Delta_i \boldsymbol{\beta}^T E\left\{\boldsymbol{\rho}_{\cdot}(V_i)^T \mathbf{b}_i\right\} + \boldsymbol{\eta}^T \mathbf{Z}_i \\
& - \sum_{j=1}^{n} \lambda(V_j) E\left\{ e^{\boldsymbol{\beta}^T (\boldsymbol{\rho}_{\cdot}(V_j)^T \mathbf{b}_i) + \boldsymbol{\eta}^T \mathbf{z}_i} I(V_i \geq V_j, \Delta_j = 1) \right\}, \qquad (2.11) \\
E\{l_{i2}(\mathbf{b}_i)\} =& -\frac{pN_i}{2} \log(2\pi) - \frac{N_i}{2} \log(|\mathbf{\Sigma}_e|) \\
& - \frac{1}{2} \sum_{j=1}^{N_i} E(\mathbf{W}_{ij} - \tilde{\boldsymbol{\rho}}_{ij}^T \boldsymbol{\mu} - \boldsymbol{\rho}_{ij}^T \mathbf{b}_i)^T \mathbf{\Sigma}_e^{-1} (\mathbf{W}_{ij} - \tilde{\boldsymbol{\rho}}_{ij}^T \boldsymbol{\mu} - \boldsymbol{\rho}_{ij}^T \mathbf{b}_i), \qquad (2.12) \\
E\{l_{i3}(\mathbf{b}_i)\} =& -\frac{p}{2} \log(2\pi) - \frac{1}{2} \log(|\mathbf{\Sigma}_b|) + E(\mathbf{b}_i^T \mathbf{\Sigma}_b^{-1} \mathbf{b}_i). \qquad (2.13)
\end{aligned}$$

In order to obtain the MLE's by setting $\sum_{i=1}^{n} S_i(\theta)$ equal to 0 and solving the equations, the EM algorithm first calculates the conditional expectations of the functions of $\mathbf{b}_i$ in (2.11), (2.12) and (2.13) in the E-step. Then in the M-step, the log likelihood is maximized by taking partial derivative of $E\{l_{i1}(\mathbf{b}_i)\}, E\{l_{i2}(\mathbf{b}_i)\}, E\{l_{i3}(\mathbf{b}_i)\}$ with respect to their corresponding parameters in $\mathbf{\Omega}$. Let $\mathbf{W}_{i.} = (\mathbf{W}_{i1}^T, \ldots, \mathbf{W}_{ip}^T)^T$ be the $N_i p \times 1$ vector of observed longitudinal covariates for the ith subject, and $\mathbf{R}_i = (\boldsymbol{\rho}_{i1}, \ldots, \boldsymbol{\rho}_{iN_i})$, $\tilde{\mathbf{R}}_i = (\tilde{\boldsymbol{\rho}}_{i1}, \ldots, \tilde{\boldsymbol{\rho}}_{iN_i})$ be the $d \times N_i p$ and $\tilde{d} \times N_i p$ observed time matrices of the ith subject. It can be easily derived that all the

parameters except $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ have the following closed-form MLE's:

$$\widehat{\boldsymbol{\mu}} = (\sum_{i=1}^{n} \tilde{\mathbf{R}}_i \tilde{\mathbf{R}}_i^T)^{-1} \{\sum_{i=1}^{n} \tilde{\mathbf{R}}_i E(\mathbf{W}_{i\cdot} - \mathbf{R}_i^T \mathbf{b}_i)\}, \tag{2.14}$$

$$\widehat{\boldsymbol{\Sigma}}_b = \frac{1}{n} \sum_{i=1}^{n} E\mathbf{b}_i \mathbf{b}_i^T, \tag{2.15}$$

$$\widehat{\boldsymbol{\Sigma}}_e = \frac{1}{\sum_{i=1}^{n} N_i} \sum_{i=1}^{n} \sum_{j=1}^{N_i} E(\mathbf{W}_{ij} - \tilde{\boldsymbol{\rho}}_{ij}^T \widehat{\boldsymbol{\mu}} - \boldsymbol{\rho}_{ij}^T \mathbf{b}_i)(\mathbf{W}_{ij} - \tilde{\boldsymbol{\rho}}_{ij}^T \widehat{\boldsymbol{\mu}} - \boldsymbol{\rho}_{ij}^T \mathbf{b}_i)^T, \tag{2.16}$$

$$\widehat{\lambda}(u) = \frac{\sum_{i=1}^{n} \Delta_i I(V_i = u)}{\sum_{j=1}^{n} E e^{\widehat{\boldsymbol{\beta}}^T (\boldsymbol{\rho}(u)^T \mathbf{b}_j) + \widehat{\boldsymbol{\eta}}^T \mathbf{z}_j} I(V_j \geq u)}, \tag{2.17}$$

where u only takes value at the event points. For other time points, $\widehat{\lambda}(u) = 0$.

The MLE's of the regression coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ in the Cox model are obtained by applying Newton-Raphson algorithm to the profile likelihood of $l_{i1}(\mathbf{b}_i)$ after plugging in $\widehat{\lambda}(u)$. Accordingly, the lth dimension of $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ are estimated by

$$\widehat{\boldsymbol{\beta}}_l^{(k)} = \widehat{\boldsymbol{\beta}}_l^{(k-1)} + I_{\beta_l}^{-1}(\widehat{\boldsymbol{\beta}}^{(k-1)}) S_{\beta_l}(\widehat{\boldsymbol{\beta}}^{(k-1)}), \tag{2.18}$$

$$\widehat{\boldsymbol{\eta}}_l^{(k)} = \widehat{\boldsymbol{\eta}}_l^{(k-1)} + I_{\eta_l}^{-1}(\widehat{\boldsymbol{\eta}}^{(k-1)}) S_{\eta_l}(\widehat{\boldsymbol{\eta}}^{(k-1)}), \tag{2.19}$$

where $S_{\beta_l}(\widehat{\boldsymbol{\beta}}^{(k-1)})$, $S_{\eta_l}(\widehat{\boldsymbol{\eta}}^{(k-1)})$ and $I_{\beta_l}(\widehat{\boldsymbol{\beta}}^{(k-1)})$, $I_{\eta_l}(\widehat{\boldsymbol{\eta}}^{(k-1)})$ are the score functions and information matrices of β_l and η_l, respectively, taking values at the $(k-1)$th updated estimates.

The score and the information matrices of the lth element of $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ are given below

$$\begin{aligned} S_{\beta_l}(\boldsymbol{\beta}) =& \sum_{i=1}^{n} \frac{\partial E\{l_{i1}(\mathbf{b}_i)\}}{\partial \beta_l} \\ =& \sum_{i=1}^{n} \Delta_i \left\{ E(\boldsymbol{\rho}_l(V_i)^T \mathbf{b}_{il}) - \frac{\sum_{j=1}^{n} E(\boldsymbol{\rho}_l(V_i)^T \mathbf{b}_{jl}) e^{\boldsymbol{\beta}^T(\boldsymbol{\rho}(V_i)^T \mathbf{b}_j) + \boldsymbol{\eta}^T \mathbf{z}_j} I(V_j \geq V_i)}{\sum_{j=1}^{n} E e^{\boldsymbol{\beta}^T(\boldsymbol{\rho}(V_i)^T \mathbf{b}_j) + \boldsymbol{\eta}^T \mathbf{z}_j} I(V_j \geq V_i)} \right\}, \end{aligned} \tag{2.20}$$

$$\begin{aligned} I_{\beta_l}(\boldsymbol{\beta}) =& -\frac{\partial S(\boldsymbol{\beta}_l)}{\partial \beta_l} \\ =& \sum_{i=1}^{n} \Delta_i \left\{ \frac{\sum_{j=1}^{n} E(\mathbf{b}_{jl}^T \boldsymbol{\rho}_l(V_i))^2 e^{\boldsymbol{\beta}^T(\boldsymbol{\rho}(V_i)^T \mathbf{b}_j) + \boldsymbol{\eta}^T \mathbf{z}_j} I(V_j \geq V_i)}{\sum_{j=1}^{n} E e^{\boldsymbol{\beta}^T(\boldsymbol{\rho}(V_i)^T \mathbf{b}_j) + \boldsymbol{\eta}^T \mathbf{z}_j} I(V_j \geq V_i)} \right. \\ & \left. - \left[\frac{\sum_{j=1}^{n} E(\mathbf{b}_{jl}^T \boldsymbol{\rho}_l(V_i)) e^{\boldsymbol{\beta}^T(\boldsymbol{\rho}(V_i)^T \mathbf{b}_j) + \boldsymbol{\eta}^T \mathbf{z}_j} I(V_j \geq V_i)}{\sum_{j=1}^{n} E e^{\boldsymbol{\beta}^T(\boldsymbol{\rho}(V_i)^T \mathbf{b}_j) + \boldsymbol{\eta}^T \mathbf{z}_j} I(V_j \geq V_i)} \right]^2 \right\}; \end{aligned} \tag{2.21}$$

$$\begin{aligned} S_{\eta_l}(\boldsymbol{\eta}) =& \sum_{i=1}^{n} \frac{\partial E(l_{i1}(\mathbf{b}_{i\cdot}))}{\partial \eta_l} \\ =& \sum_{i=1}^{n} \Delta_i \left\{ Z_{il} - \frac{\sum_{j=1}^{n} Z_{il} E e^{\boldsymbol{\beta}^T(\mathbf{b}_{j\cdot}^T \boldsymbol{\rho}_{\cdot}(V_i)) + \boldsymbol{\eta}^T \mathbf{z}_j} I(V_j \geq V_i)}{\sum_{j=1}^{n} E e^{\boldsymbol{\beta}^T(\mathbf{b}_{j\cdot}^T \boldsymbol{\rho}_{\cdot}(V_i)) + \boldsymbol{\eta}^T \mathbf{z}_j} I(V_j \geq V_i)} \right\}, \end{aligned} \tag{2.22}$$

$$\begin{aligned} I_{\eta_l}(\boldsymbol{\eta}) =& -\frac{\partial S(\boldsymbol{\eta}_l)}{\partial \eta_l} \\ =& \sum_{i=1}^{n} \Delta_i \left\{ \frac{\sum_{j=1}^{n} Z_{il}^2 E e^{\boldsymbol{\beta}^T(\mathbf{b}_{j\cdot}^T \boldsymbol{\rho}_{\cdot}(V_i)) + \boldsymbol{\eta}^T \mathbf{z}_j} I(V_j \geq V_i)}{\sum_{j=1}^{n} E e^{\boldsymbol{\beta}^T(\mathbf{b}_{j\cdot}^T \boldsymbol{\rho}_{\cdot}(V_i)) + \boldsymbol{\eta}^T \mathbf{z}_j} I(V_j \geq V_i)} \right. \\ & \left. - \left[\frac{\sum_{j=1}^{n} Z_{il} E e^{\boldsymbol{\beta}^T(\mathbf{b}_{j\cdot}^T \boldsymbol{\rho}_{\cdot}(V_i)) + \boldsymbol{\eta}^T \mathbf{z}_j} I(V_j \geq V_i)}{\sum_{j=1}^{n} E e^{\boldsymbol{\beta}^T(\mathbf{b}_{j\cdot}^T \boldsymbol{\rho}_{\cdot}(V_i)) + \boldsymbol{\eta}^T \mathbf{z}_j} I(V_j \geq V_i)} \right]^2 \right\}. \end{aligned} \tag{2.23}$$

In EM algorithm, the E-step and M-step are calculated iteratively until the algorithm converges. By (2.14) through (2.19), in each iteration, conditional expectations need to be

evaluated for the following six functions of the random effects for all the subjects:

$$\begin{aligned}
g_1(\mathbf{b}_i) &= \mathbf{b}_i, \\
g_2(\mathbf{b}_i) &= \mathbf{b}_i\mathbf{b}_i^T, \\
g_3(\mathbf{b}_i) &= \boldsymbol{\rho}(V_i)^T\mathbf{b}_i, \\
g_4(\mathbf{b}_j) &= e^{\boldsymbol{\beta}^T(\boldsymbol{\rho}_{(V_i)^T}\mathbf{b}_j)}, \text{ for } j = 1, \ldots, n, \text{ and } V_j \geq V_i \\
g_5(\mathbf{b}_j) &= \{\boldsymbol{\rho}(V_i)^T\mathbf{b}_j\}e^{\boldsymbol{\beta}^T(\boldsymbol{\rho}_{(V_i)^T}\mathbf{b}_j)}, \text{ for } j = 1, \ldots, n, \text{ and } V_j \geq V_i, \\
g_6(\mathbf{b}_j) &= \{\boldsymbol{\rho}(V_i)^T\mathbf{b}_j\}^2e^{\boldsymbol{\beta}^T(\boldsymbol{\rho}_{.(V_i)^T}\mathbf{b}_j)}, \text{ for } j = 1, \ldots, n, \text{ and } V_j \geq V_i.
\end{aligned} \tag{2.24}$$

Since multi-dimensional integrals are involved in the conditional expectations $E\{g(\mathbf{b}_i)\}$ in the E-step, numerical integration techniques such as Gaussian-Hermite Quadrature (Wulfsohn and Tsiatis, 1997), Markov Chain Monte Carlo (Henderson et al., 2000; Tseng et al., 2005; Ding and Wang, 2008) and fully exponential Laplace approaches (Rizopoulos et al., 2009) have been applied to approximate the target expectations. Since most of these techniques have been shown to be insufficient in estimating the joint models with large dimensions of random effects, we propose to approximate the conditional expectations using design of experiment based interpolation techniques (DoIt; Joseph 2012).

Another challenge of joint modeling is to obtain standard errors (SE's) for the MLEs. Louis (1982) suggested that the accurate variance estimation in EM algorithm would require the calculation of the observed Fisher information matrix for the entire parameter set. However, this approach is impractical for our case considering the high dimensionality of $\boldsymbol{\Omega}$ mainly caused by $\lambda(t)$. On the other hand, if we use the profile likelihood $pl(\boldsymbol{\theta})$(i.e., substituting the estimate of $\lambda(t)$ into $l(\boldsymbol{\Omega})$) to solve for the SE's for $\boldsymbol{\theta}$ has been proved to yield biased results (Hsieh et al., 2006). Due to these limitations, in this study, we propose to estimate the SE's using bootstrap technique, the procedure of which will be explained in detail in simulation studies in Section 3.

2.3 Implementation

It has always been a challenge to approximate the large-dimensional integral numerically. In order to approximate the conditional expectations of $E\{g(\mathbf{b}_i)\}$ in (2.14) through (2.19), several techniques have been proposed in the previous literature. Gaussian-Hermite quadrature method (Wulfsohn and Tsiatis, 1997) evaluates the function values at M different quadrature points and approximates the expectations using the weighted sums. This method is usually accurate for low dimensional integrals but the number of quadrature points increases exponentially with the dimension. Markov Chain Monte Carlo method (Henderson et al., 2000; Tseng et al., 2005; Ding and Wang, 2008) treats the posterior densities $f(\mathbf{b}_i|D_o, \widehat{\mathbf{\Omega}})$ as transition probabilities of a Markov Chain from which the samples are drawn and updated. By the law of large number, the sample means of $g(\mathbf{b}_i)$ converge to the expectations with a the rate of $O(M^{-1/2})$ in probability, where M is the number of random samples. Fully exponential Laplace (Tierney and Kadane, 1986; Tierney et al., 1989) does not rely on evaluation or sample points. Instead, it approximates the posterior distributions of $\mathbf{b}_{ik}$ by normal distributions with the posterior modes of $f(\mathbf{b}_i|D_o, \widehat{\mathbf{\Omega}})$ as the means and the inverse Fisher information matrices as the variances. The approximation accuracy depends only on the sample size of each individual and was proved to be of the order $O(n_i^{-2})$ (Rizopoulos et al., 2009). All of these methods, though work well for low-dimensional cases, become either time-consuming or difficult to implement as the dimensions of the random effects $\mathbf{b}_i$ increase. In order to extend joint models to the more flexible settings, a more efficient and stable computing technique is needed to better implement the EM algorithm.

The design of experiment based interpolation techniques (DoIt) was recently proposed by Joseph (2012) as a method to approximate the "expensive conditional densities" in Bayesian computation. DoIt method borrows and extends the idea of Laplace approximation which approximates the posterior densities of interest via normal distributions. But instead of using

a single normal distribution, DoIt also incorporates the idea of quadrature-based methods and approximates $f(\mathbf{b}_i|D_o, \widehat{\boldsymbol{\Omega}})$ using the weighted sum of a sequence of normal densities with the means at M pre-specified evaluation points $(\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_M)$, i.e.,

$$f_{\mathbf{b}_i|D_o,\widehat{\boldsymbol{\Omega}}}(\mathbf{b}_i) \approx \frac{1}{\sum_{l=1}^{M} c_l} \sum_{l=1}^{M} c_l \phi_l(\mathbf{b}_i), \tag{2.25}$$

where $\phi_l(\mathbf{b}_i) = \phi(\mathbf{b}_i; \boldsymbol{\nu}_l, \mathbf{D}_i^{-1})$ denote the normal density of $\mathbf{b}_i$ with mean $\boldsymbol{\nu}_{kl}$ and variance $\mathbf{D}_i^{-1}$, the inverse of Fisher information matrix of $f(\mathbf{b}_i|D_o, \widehat{\boldsymbol{\Omega}})$ evaluated at the mode. The c_l's are the weights associated with $\phi_l(\cdot)$ and can be calculated by solving the linear equations

$$\boldsymbol{Q}\boldsymbol{c} = \mathbf{h},$$

where $h(\mathbf{b}_i) \propto f(\mathcal{W}_i, V_i, \Delta_i|\mathbf{b}_i) f(\mathbf{b}_i)$ is the unnormalized posterior densities of $\mathbf{b}_i$, and $\mathbf{h} = (h(\boldsymbol{\nu}_1), \ldots, h(\boldsymbol{\nu}_M))$ is a vector of $h(\cdot)$ taking values at the M evaluation points. $\mathbf{Q}$ is an $M \times M$ matrix with ijth element being the unnormalized normal density

$$q(\boldsymbol{\nu}_i; \boldsymbol{\nu}_j, \mathbf{D}_i^{-1}) = \exp\{-\frac{1}{2}(\boldsymbol{\nu}_i - \boldsymbol{\nu}_j)^T \mathbf{D}_i (\boldsymbol{\nu}_i - \boldsymbol{\nu}_j)\}.$$

Since $q(\boldsymbol{\nu}; \boldsymbol{u}, \mathbf{D}_i^{-1})$ is a positive definite function, $\mathbf{Q}^{-1}$ exists, provided $\boldsymbol{\nu}_i \neq \boldsymbol{\nu}_j$ for all i and j. Thus a unique solution of $\widehat{\boldsymbol{c}} = \mathbf{Q}^{-1}\mathbf{h}$ is guaranteed.

With the capability to approximate the expensive posterior densities, DoIt can also be used to approximate the conditional expectations of real value functions of $\mathbf{b}_i$, $g(\mathbf{b}_i)$, in the E-step of EM algorithm:

$$E\{g(\mathbf{b}_i)\} \approx \frac{1}{\sum_{l=1}^{M} c_l} \sum_{l=1}^{M} c_l E_l\{g(\mathbf{b}_i)\}, \tag{2.26}$$

where $E_l\{g(\mathbf{b}_i)\}$ is the expectation of $g(\mathbf{b}_i)$ with respect to the normal distribution $N(\boldsymbol{\nu}_l, \mathbf{D}_i^{-1})$. By introducing the idea of DoIt into the computation of joint modeling, we find it produces good parameter estimates, and thus can be taken as an alternative to the existing computing techniques for joint models.

The locations of $(\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_M)$ are determined by the pre-specified space-filling design points transformed to the parameter space of posterior distribution via $\widehat{\mathbf{b}}_i$ and $\mathbf{D}_i^{-1}$. Similar to other methods that depend on evaluation points, DoIt has the nice feature of achieving the desired level of accuracy by adding more points, i.e., increasing M (Joseph, 2012). However, DoIt method does not suffer from "curse of dimensionality" like Gaussian quadrature method because the number of M grows much slower when the dimension increases. A common rule of thumb in the literature of computer experiment is to use $M = 10 \times d_b$ (Loeppky et al., 2009), where d_b is the dimension of the integral. With good space-filling design points, the order of the convergence rate can reach $O(M^{-1} \log^{d_b} M)$ if the integrand is a function with bounded total variation (Fang and Wang, 1994). Although various space-filling designs can be considered to specify the locations of the deterministic points $(\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_M)$, in this study we apply minimax Latin Hypercube Design (MmLHD; Joseph 2012) as the design scheme for DoIt implementation with the existing R package *LHS* which automatically generates MmLHD samples.

2.4 Sampling properties

Let $(\widehat{\boldsymbol{\theta}}, \widehat{\Lambda})$ be the maximum likelihood estimators maximizing the likelihood equation given in (2.8). Zeng and Cai (2005) established the consistency and asymptotic normality of the joint maximum likelihood estimate for joint model with a single covariate process. In this section, we adopt their statistical formulation. It is worth pointing out that since we have multivariate response in the longitudinal model, the covariance structure for the multiple longitudinal processes is specified by the covariance matrix $\boldsymbol{\Sigma}_e$ instead of the scalar σ_e^2 for the single longitudinal response in Zeng and Cai (2005). This makes the theoretical proof much more challenging compared with the single covariate case. For completeness, we state the theorems below and provide the proof details of Theorem 1 and 2 that follow and extend

the work of Zeng and Cai (2005). We eliminate the proof of Theorem 3 as it follows exactly the same arguments as in Zeng and Cai (2005). All the other proofs and the required technical conditions are specified in Section 5.

Theorem 1 Under assumption (A.1) - (A.9), the maximum likelihood estimator $(\widehat{\boldsymbol{\theta}}, \widehat{\Lambda})$ is strongly consistent under the product metric of the Euclidean norm and the supremum norm on $[0, \tau]$; that is,

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| + \sup_{t\in[0,\tau]} |\widehat{\Lambda}(t) - \Lambda_0(t)| \to 0 \qquad a.s.$$

Theorem 2 Under assumption (A.1) - (A.9), $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0, \widehat{\Lambda} - \Lambda_0)$ weakly converges to a Gaussian random element in $\mathbb{R}^{d_\theta} \times l^\infty[0, \tau]$, where d_θ is the dimension of $\boldsymbol{\theta}$ and $l^\infty[0, \tau]$ is the metric space of all bounded functions in $[0, \tau]$. Furthermore, $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ weakly converges to a multivariate normal distribution with mean zero and its asymptotic variance attains the semiparametric efficiency bound for $\boldsymbol{\theta}_0$.

Theorem 3 Under assumptions (A.1) - (A.9), $2\{pl_n(\widehat{\boldsymbol{\theta}}) - pl_n(\boldsymbol{\theta}_0)\}$ weakly converges to a chi-square distribution with d_θ degrees of freedom and, moreover,

$$-\frac{pl_n(\widehat{\boldsymbol{\theta}} + h_n\mathbf{e}) - 2pl_n(\widehat{\boldsymbol{\theta}}) + pl_n(\widehat{\boldsymbol{\theta}} - h_n\mathbf{e})}{nh_n^2} \xrightarrow{p} \mathbf{e}^T \boldsymbol{I}\mathbf{e},$$

where $pl_n(\boldsymbol{\theta})$ is the profile likelihood function of $\boldsymbol{\theta}$ defined as $pl_n(\boldsymbol{\theta}) = \max_{\Lambda\in\mathcal{V}_n} l_n(\boldsymbol{\theta}, \Lambda)$. $h_n = O_p(n^{-1/2})$, $\mathbf{e}$ is any vector in $\mathbb{R}^{d_\theta}$ with unit norm, and $\boldsymbol{I}$ is the efficient information matrix for $\boldsymbol{\theta}_0$.

3 Numerical studies

In this section we conduct simulation studies and real data analysis to examine the estimation performance of the proposed approach.

3.1 Simulation studies

We consider joint models with both single and multiple longitudinal covariates. Under the framework of joint likelihood approach with EM algorithm, we compare different integral approximation techniques, using the existing R package *JM* (Rizopoulos, 2010) as the benchmark. Below are the implementation details of each method.

(i) *JM* package (JM): we use the option "method = Cox-PH-GH" in the JM() function, in which an adaptive Gaussian-Hermite quadrature method with 35 quadrature points on each dimension of $\mathbf{b}_{ik}$ is used to calculate the conditional expectations in the E-step of the EM algorithm.

(ii) Gaussian-Hermite quadrature method (GHQ): we take 5 quadrature points on each dimension of $\mathbf{b}_{ik}$, and the total number of evaluation points are $M = 5^{d_b}$.

(iii) MCMC method: we use the R package *MHadaptive* to generate random samples from the unnormalized posterior density of $\mathbf{b}_{ik}$. For each individual 1000 MCMC samples are generated, with 100 of them as "burn-in" samples (i.e., M=900). Since the computing process is extremely slow with this method, a more relaxed EM stopping is used (i.e., $|\epsilon| < 10^{-2}$ for MCMC; $|\epsilon| < 10^{-4}$ for other methods, where $\epsilon = |\widehat{\boldsymbol{\theta}}^{(k)} - \widehat{\boldsymbol{\theta}}^{(k+1)}|_\infty$ is the maximum dimension of the difference between the parameter estimates in the current and the previous iterations).

(iv) Fully exponential Laplace method (FEL): we adopt the technique described in Rizopoulos et al. (2009).

(v) DoIt method: the R package *lhs* is used to generate design points from MmLHD. The suggested number of design points $M = 10 \times d_b$ is used.

All the results are based on $N = 100$ replicates with $n = 100$ subjects in each data set.

Example 1. We consider a simple joint model with a single longitudinal covariate:

$$W_i(t_{ij}) = X_i(t_{ij}) + e_i(t_{ij}) = b_{0i} + b_{1i}t_{ij} + e_i(t_{ij}),$$

$$\lambda_i(t) = \lambda(t) \exp\{\beta X_i(t) + \eta Z_i\}$$

with the assumption

$$\mathbf{b}_i = \begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} \overset{i.i.d}{\sim} N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}),\ e_i(t_{ij}) \overset{i.i.d}{\sim} N(0, \sigma^2),$$

and $\boldsymbol{b}_i$ are independent of $e_i(t_{ij})$. This is a low-dimensional setting with $\dim(\mathbf{b}_i) = 2$, and $\boldsymbol{\rho}(t) = (1, t)^T$. The true values of the parameters are based on Hsieh et al. (2006). For the longitudinal process, we set $\boldsymbol{\mu} = (-4.9078, 0.500)^T$, $(\sigma_{11}, \sigma_{12}, \sigma_{22}) = (0.5, -0.001, 0.04)$, $\sigma^2 = 0.1$, where $(\sigma_{11}, \sigma_{12}, \sigma_{22})$ are the upper diagonal elements of $\boldsymbol{\Sigma}$. The observational time $\mathbf{t}_i$ is generated first by $\mathbf{t}_i = seq(0, 12, 38)$ in R (i.e., 38 equally spaced time points between 0 and 12), and then truncated by the survival time of different individuals. For the event process, we assume the baseline hazard function is constant over time, i.e., $\lambda(t) = 1$. The true regression coefficients are $\beta = 1$ and $\eta = -1$. The censoring time for each subject is generated from an exponential distribution with mean 25, resulting in about 30% censoring. The survival time T_i is generated from the inverse survival function derived from the cumulative hazard $\Lambda_i(t)$ with the constant baseline hazard λ:

$$\begin{aligned} \Lambda_i(t) &= \int_0^t \lambda(u) \exp\{\beta X_i(u) + \eta Z_i\} \\ &= e^{\eta Z_i} \lambda \int_0^t e^{\beta(b_{0i} + b_{1i}u)} du \\ &= \lambda \frac{1}{\beta b_{1i}} e^{\beta b_{0i} + \eta Z_i} \left(e^{\beta b_{1i} t} - 1 \right). \end{aligned}$$

Since the survival probability at time t is $S_i(t) = 1 - F_i(t) = \exp\{-\Lambda_i(t)\}$, $S_i(t)$ can be generated using the Monte Carlo samples $U_i \sim U(0,1)$, the uniform distribution, and $\exp\{-\Lambda_i(t)\} = U_i$. Thus,

$$T_i = \frac{1}{\beta_X b_{1i}} \log\{1 - \frac{\beta_X b_{1i} \log U_i}{\lambda \exp(\beta_X b_{0i} + \beta_Z Z_i)}\}. \tag{3.1}$$

Note that since the logarithm is involved in (3.1), we need to monitor the sign of the expression inside log. Since U_i is between 0 and 1, $\log U_i$ is negative. Thus we need $b_{1i} > 0$ to guarantee the positiveness of $-\beta_X b_{1i} \log U_i$. In practice, most of the random samples of b_{1i} is positive with the setting of $\mu_2 = 0.500$ and $\Sigma_{22} = 0.04$. However, negative expression can occur with some specific samples of b_{1i} and U_i, for which the above equation to generate T_i is not well-defined. Such cases are treated as censoring in the simulation. However, these cases are very rare (usually less that 2%) under our parameter setting, and they hardly affect the distribution assumptions of the simulation.

After truncated by the observed event time V_i, the average number of longitudinal observations is $\bar{n}_\cdot = 20$.

Table 1: Simulation results for Example 1 with $\bar{n}_\cdot = 20$

		JM	GHQ	MCMC	FEL	DoIt (M=10)
Time(s)	Median	14.60	4.67	584.10	9.97	9.41
	Bias	-0.0330	0.0022	-0.0012	-0.0195	-0.0043
$\beta = 1.0000$	SD	0.1270	0.1288	0.1276	0.1187	0.1278
	RMSE	0.1312	0.1288	0.1276	0.1203	0.1279
	Bias	0.0310	0.0106	0.0616	0.0345	0.0157
$\eta = -1.0000$	SD	0.3198	0.3187	0.2677	0.3106	0.3167
	RMSE	0.3213	0.3189	0.2747	0.3125	0.3171
	Bias	0.0063	-0.0034	-0.0034	-0.0006	-0.0004
$\mu_1 = -4.9078$	SD	0.1990	0.0749	0.0604	0.0729	0.0730
	RMSE	0.1991	0.0750	0.0605	0.0729	0.0730
	Bias	-0.0027	0.0028	0.0031	0.0000	-0.0001
$\mu_2 = 0.5000$	SD	0.0395	0.0231	0.0201	0.0224	0.0229
	RMSE	0.0400	0.0233	0.0203	0.0224	0.0229
	Bias	0.0967	0.0066	-0.0441	-0.0100	-0.0112
$\sigma_{11} = 0.5000$	SD	0.0965	0.0725	0.0732	0.0721	0.0719
	RMSE	0.1366	0.0728	0.0855	0.0728	0.0728
	Bias	-0.0134	-0.0069	0.0070	-0.0007	-0.0008
$\sigma_{12} = -0.0010$	SD	0.0208	0.0159	0.0145	0.0155	0.0156
	RMSE	0.0247	0.0173	0.0161	0.0170	0.0156
	Bias	0.0009	0.0048	-0.0033	-0.0008	-0.0009
$\sigma_{22} = 0.0400$	SD	0.0073	0.0067	0.0064	0.0061	0.0062
	RMSE	0.0074	0.0082	0.0072	0.0062	0.0063
	Bias	0.2731	0.0092	0.0166	-0.0049	-0.0006
$\sigma^2 = 0.1000$	SD	0.0096	0.0037	0.0099	0.0032	0.0033
	RMSE	0.2733	0.0099	0.0193	0.0059	0.0034

In Table 1, GHQ with 25 evaluation points takes the shortest computing time, showing its superiority over other methods when the dimension of random effects is small. DoIt with 10 evaluation points is the second fastest technique, followed by FEL, both faster than the standard *JM* package. MCMC is extremely slow, mostly because it requires a large number of random samples to be drawn for each individual in each iteration.

All the methods yield comparable estimates in terms of accuracy (i.e., Bias) and efficiency (i.e., RMSE) in Table 1. The only exception is the JM estimates of σ^2, which has great bias compared with other methods. The reason for this is that JM package only focuses on estimating regression coefficients β and η in survival models using the joint information of the two processes. As for the parameters in the longitudinal model, it directly uses the linear mixed-effects models' estimates from the R function lme(), which does not consider the joint information from the survival part, and thus may lead to bias and inefficiency. For all the other methods, the estimates from lme() functions are only taken as initial values for EM algorithm. DoIt method with $M = 10$ evaluation points provides good estimating results in the table, and such good performance is consistent for all the parameters.

In our simulation studies, we use bootstrap technique to obtain the standard errors (SE's) of the maximum likelihood estimates. For each dataset replication, we randomly sample n_1 subjects with replacement from the uncensored group of individuals, and randomly sample n_2 subjects with replacement from the censored group. $n_2 = n - n_1$ is the number of censoring in the original replicate dataset. The sampling with replacement is conducted within each cluster so as to remain the same censoring rate as the original data. The two groups of random samples are then combined into a complete dataset and the parameter estimates are obtained using the maximum joint likelihood (MJL) method. Such estimating procedure is repeated $B = 100$ times for each replicate, and the standard deviations of these $B = 100$ parameter estimates are recorded as the bootstrap SE's of the replicate dataset.

Table 2 shows the performance of the bootstrap SE's of the $N = 100$ replicates for

MJL with DoIt algorithm. The second column, Mean Est, is the mean of the MLEs of the 100 replicates; the third column, SD_{Est}, is the standard deviations of the MLEs of the 100 replicates; the fourth column, SE_{Boot}, is the average of the bootstrap SE's across the 100 replicates; and the last column, $\text{SD}_{SE.Boot}$, is the standard deviations of the bootstrap SE's across the 100 replicates. The results suggest that the average bootstrap SE's are very similar to the SD's for all the parameters. More specifically, all the SD's fall in the range of $\text{SE}_{Boot} \pm 2\text{SD}_{SE.Boot}$. This indicates that the bootstrap SE's are reliable.

Table 2: Performance of Bootstrap Standard Errors of Example 1

	True	Mean Est	SD_{Est}	SE_{Boot}	$\text{SD}_{SE.Boot}$
β	1.0000	0.9957	0.1278	0.1343	0.0238
η	-1.0000	-0.9843	0.3167	0.2894	0.0295
μ_1	-4.9078	-4.9082	0.0730	0.0697	0.0073
μ_2	0.5000	0.4999	0.0229	0.0202	0.0021
σ_{11}	0.5000	0.4888	0.0719	0.0692	0.0147
σ_{12}	-0.0010	-0.0018	0.0156	0.0149	0.0023
σ_{22}	0.0400	0.0391	0.0062	0.0058	0.0010
σ^2	0.1000	0.0994	0.0033	0.0031	0.0003

Example 2. In this example we consider a joint model with two longitudinal covariates.

$$W_{i1}(t_{ij}) = X_{i1}(t_{ij}) + e_{i1}(t_{ij}) = b_{10i} + b_{11i}t_{ij} + e_{i1}(t_{ij}),$$

$$W_{i2}(t_{ij}) = X_{i2}(t_{ij}) + e_{i2}(t_{ij}) = b_{20i} + b_{21i}t_{ij} + e_{i2}(t_{ij}),$$

$$\lambda_i(t) = \lambda(t)\exp\{\beta_1 X_{1i}(t) + \beta_2 X_{2i}(t) + \eta Z_i\}$$

with the assumption

$$\mathbf{b}_{ik} = \begin{pmatrix} b_{k0i} \\ b_{k1i} \end{pmatrix} \overset{i.i.d}{\sim} N_2(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_{bk}),\ e_{ik}(t_{ij}) \overset{i.i.d}{\sim} N(0, \sigma_k^2),$$

where the random effects $\mathbf{b}_{ik}$ are independent of the measurement errors $e_{ik}(t_{ij})$, and the two longitudinal processes are independent of each other. The true values of the parameters are $\boldsymbol{\mu}_1 = (-5, 0.5)^T$, $\boldsymbol{\mu}_2 = (-2, 1)^T$, $Vec(\boldsymbol{\Sigma}_1) = (1, -0.001, 0.04)$, $Vec(\boldsymbol{\Sigma}_2) = (0.5, -0.001, 0.09)$, $\sigma_1^2 = \sigma_2^2 = 1$, $\beta_1 = 1$, $\beta_2 = 2$, $\eta_1 = -1$. They are specified in details in the output tables.

In this example, the dimension of the random effects becomes $\dim(\mathbf{b}_{i\cdot}) = 4$. In the event process, we take $\lambda(t) = 1$. The average censoring rate is around 15%. Using similar techniques as in example 1, we generate survival time as follows:

$$T = \frac{1}{(\beta_1 b_{11i} + \beta_2 b_{21i})} \log\left\{1 - \frac{(\beta_1 b_{11i} + \beta_2 b_{21i}) \log U_i}{\lambda \exp(\beta_1 b_{10i} + \beta_2 b_{20i} + \eta Z_i)}\right\}.$$

Since *JM* package is not available with multiple longitudinal covariates and MCMC falls out of scope due to its slow speed, we consider only GHQ, FEL and DoIt method for this example.

In Table 3, GHQ becomes much slower than the other two methods because the number of evaluation points increases dramatically to $M = 5^4 = 625$. DoIt with $M = 10 \times 4 = 40$ uses the shortest computing time. The estimates of the three methods are of similar bias levels, indicating they are similar in terms of estimating accuracy. However, the estimates from GHQ method, especially those of the longitudinal process, have much larger SD and RMSE than the corresponding estimates from FEL and DoIt. This suggests FEL and DoIt are more stable in estimation and provide more efficient estimates. Although FEL is comparable with DoIt in computing time and estimating performance, it involves the complex tensor computation (Rizopoulos et al., 2009). This makes it less appealing than DoIt for larger-dimensional problems.

Similar to Table 2, the results of Table 4 shows that the bootstrap SE's closely resemble the standard deviations in the second column, and thus can be regarded as valid standard error estimates for MLE's.

Table 3: Simulation results for Example 2 with $\bar{n}_\cdot = 20$

		GHQ	FEL	DoIt ($M = 10d$)
Time(s)	Median	440.8	122.6	118.5
	Bias (SD)	-0.0161(0.1769)	-0.0120 (0.1466)	0.0032 (0.1302)
$\beta_1 = 1.0000$	RMSE	0.1776	0.1471	0.1303
	Bias (SD)	0.0121(0.3119)	0.0304 (0.2461)	-0.0119 (0.1972)
$\beta_2 = 2.0000$	RMSE	0.3121	0.2480	0.1975
	Bias (SD)	0.0176 (0.2979)	0.0161 (0.3076)	-0.0032 (0.2715)
$\eta = -1.0000$	RMSE	0.2984	0.3080	0.2715
	Bias (SD)	0.0309 (0.5114)	-0.0174 (0.0951)	-0.0091 (0.1031)
$\mu_{11} = -5.0000$	MSE	0.5123	0.0967	0.1035
	Bias (SD)	-0.0025 (0.0537)	0.0002 (0.0185)	-0.0010 (0.0201)
$\mu_{12} = 0.5000$	RMSE	0.0538	0.0185	0.0201
	Bias (SD)	0.0173 (0.2136)	0.0008 (0.0699)	0.0132 (0.0665)
$\mu_{21} = -2.0000$	RMSE	0.2143	0.0699	0.0678
	Bias (SD)	-0.0109 (0.1050)	-0.0062 (0.0306)	-0.0006 (0.0276)
$\mu_{22} = 1.0000$	RMSE	0.1056	0.0312	0.0277
	Bias (SD)	-0.0057 (0.1809)	-0.0164 (0.1480)	-0.0323 (0.1297)
$\sigma_{111} = 1.0000$	RMSE	0.1810	0.1489	0.1337
	Bias (SD)	-0.0067 (0.0210)	0.0013 (0.0216)	-0.0012 (0.0189)
$\sigma_{112} = -0.0010$	RMSE	0.0220	0.0216	0.0189
	Bias (SD)	0.0055 (0.0082)	-0.0007 (0.0066)	-0.0004 (0.0064)
$\sigma_{122} = 0.0400$	RMSE	0.0099	0.0067	0.0064
	Bias (SD)	-0.0022 (0.0814)	-0.0153 (0.0616)	-0.0109 (0.0661)
$\sigma_{211} = 0.5000$	RMSE	0.0814	0.0635	0.0670
	Bias (SD)	-0.0088 (0.0208)	-0.0004 (0.0215)	-0.0020 (0.0220)
$\sigma_{212} = -0.0010$	RMSE	0.0226	0.0215	0.0220
	Bias (SD)	0.0070 (0.0177)	-0.0007 (0.0141)	-0.0010 (0.0143)
$\sigma_{222} = 0.0900$	RMSE	0.0190	0.0141	0.0144
	Bias (SD)	0.0088 (0.0115)	-0.0054 (0.0029)	0.0001 (0.0029)
$\sigma_1^2 = 0.1000$	RMSE	0.0145	0.0062	0.0030
	Bias (SD)	0.0087 (0.0115)	-0.0043 (0.0032)	0.0003 (0.0027)
$\sigma_2^2 = 0.1000$	RMSE	0.0144	0.0054	0.0027

Table 4: Performance of Bootstrap Standard Errors of Example 2

	True	Mean Est	SD_{Est}	SE_{Boot}	$\text{SD}_{SE.Boot}$
β_1	1.0000	1.0032	0.1302	0.1372	0.0309
β_2	2.0000	1.9881	0.1972	0.1980	0.0431
η	-1.0000	-1.0032	0.2715	0.2734	0.0501
μ_{11}	-5.0000	-5.0091	0.1031	0.0988	0.0171
μ_{12}	0.5000	0.4990	0.0201	0.0211	0.0039
μ_{21}	-2.0000	-1.9868	0.0665	0.0696	0.0114
μ_{22}	1.0000	0.9994	0.0276	0.0306	0.0053
σ_{111}	1.0000	0.9677	0.1297	0.1359	0.0366
σ_{112}	-0.0010	-0.0022	0.0189	0.0211	0.0039
σ_{122}	0.0400	0.0396	0.0064	0.0065	0.0016
σ_{211}	0.5000	0.4891	0.0661	0.0704	0.0159
σ_{212}	-0.0010	-0.0030	0.0220	0.0222	0.0044
σ_{222}	0.0900	0.0890	0.0143	0.0131	0.0034
σ_1^2	0.1000	0.1001	0.0029	0.0032	0.0005
σ_2^2	0.1000	0.1003	0.0027	0.0032	0.0006

3.2 A real data example

In this section, we illustrate the proposed model and estimation procedure by an empirical analysis of the data collected from a smoking cession study. Specifically, this data set was collected from a randomized, placebo-controlled clinical trial (N=1504) of five active smoking-cessation pharmacotherapies, in which daily smokers who were highly motivated to quit were recruited (Piper et al., 2009). The focus of this example is to examine the effects of multiple smoking withdrawal symptoms on the time to lapse (T_L, first smoking after quit) and time to relapse (T_RL, smoke for 4 consecutive days after quit) in a two-week post-quit study period (i.e. $t \in (0, 14)$). The withdrawal symptoms of interest include craving for

smoking (Crav), negative affect (NA) and fatigue of quitting smoking (Fatig). All of these symptoms are supposed to exhaust the self-control resources that prevent the participants from smoking when they attempt to quit. All the items are self-reported by the participants four times a day (i.e., morning, night and 2 random times). The five treatments are divided into three types: placebo, mono therapy and combined therapy, and are treated as time-independent covariates in the joint model. We build two working data sets for analysis: one with T_L as the response of interest (i.e., Lapse Data), and the other with T_RL as the response of interest (i.e., Relapse Data). After data cleaning, N=794 subjects are used for the analysis with T_L as the response, and N=811 subjects are used for then analysis with T_RL as the response.

Participants without a record for the longitudinal covariates after the actual quit date or with the survival response equal to 0 are removed out of the study. In addition, we only use the records between 0 and $\min(T\_L, 14)$ for Lapse Data and between 0 and $\min(T\_RL, 14)$ for Relapse Data. As a result, there are 794 subjects left in Lapse Data (83 placebo, 412 monotherapy and 299 combined therapy) and the censor rate is 59.69%; 811 subjects are left in Relapse Data (86 placebo, 422 monotherapy and 303 combined therapy) and the consor rate is 91.24%.

In the previous study (Liu et al., 2013) we modeled the three longitudinal processes nonparametrically and found all of them decrease linearly with time (see Figure 1). Thus in joint modeling framework we specify them using linear mixed-effects model of the following forms:

$$W_{\text{Cravi}}(t_{ij}) = X_{\text{Cravi}}(t_{ij}) + e_{ic}(t_{ij}) = b_{c0i} + b_{c1i}t_{ij} + e_{ci}(t_{ij}),$$

$$W_{\text{NA}i}(t_{ij}) = X_{\text{NA}i}(t_{ij}) + e_{in}(t_{ij}) = b_{n0i} + b_{n1i}t_{ij} + e_{ni}(t_{ij}),$$

$$W_{\text{Fatig}i}(t_{ij}) = X_{\text{Fatig}i}(t_{ij}) + e_{if}(t_{ij}) = b_{f0i} + b_{f1i}t_{ij} + e_{fi}(t_{ij}),$$

where

$$\mathbf{b}_{ci} = (b_{c0i}, b_{c1i})^T \sim N_2(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_{bc}),$$

$$\mathbf{b}_{ni} = (b_{n0i}, b_{n1i})^T \sim N_2(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_{bn}),$$

$$\mathbf{b}_{fi} = (b_{f0i}, b_{f1i})^T \sim N_2(\boldsymbol{\mu}_f, \boldsymbol{\Sigma}_{bf}),$$

and

$$e_{ci}(t_{ij}) \sim N(0, \sigma_c^2),$$

$$e_{ni}(t_{ij}) \sim N(0, \sigma_n^2),$$

$$e_{fi}(t_{ij}) \sim N(0, \sigma_f^2),$$

and all these random terms are assumed independent.

In this analysis we model the survival response using Cox's models with both single covariate process and multiple covariates processes. The single-covariate models are of the following form:

$$\lambda_i(t) = \lambda(t) \exp\{\beta_c X_{\text{Cravi}}(t_{ij}) + \eta Z_i\},$$

$$\lambda_i(t) = \lambda(t) \exp\{\beta_n X_{\text{NAi}}(t_{ij}) + \eta Z_i\},$$

$$\lambda_i(t) = \lambda(t) \exp\{\beta_f X_{\text{Fatigi}}(t_{ij}) + \eta Z_i\},$$

and the multiple-covariates model is of the following form

$$\lambda_i(t) = \lambda(t) \exp\{\beta_c X_{\text{Cravi}}(t_{ij}) + \beta_n X_{\text{NAi}}(t_{ij}) + \beta_f X_{\text{Fatigi}}(t_{ij}) + \eta Z_i\},$$

where Z_i is the treatment variable with three levels (i.e., 2 dummy variables in practice).

The goals of the analysis are: (a) to compare maximum joint likelihood method (with DoIt algorithm) with the naive separate estimation, and (b) to compare the single-covariate-process model with the multiple-covariates-process model. The standard errors of maximum joint likelihood approach are calculated using the bootstrap techniques as discussed in simulation studies. The estimating results are presented in the following table.

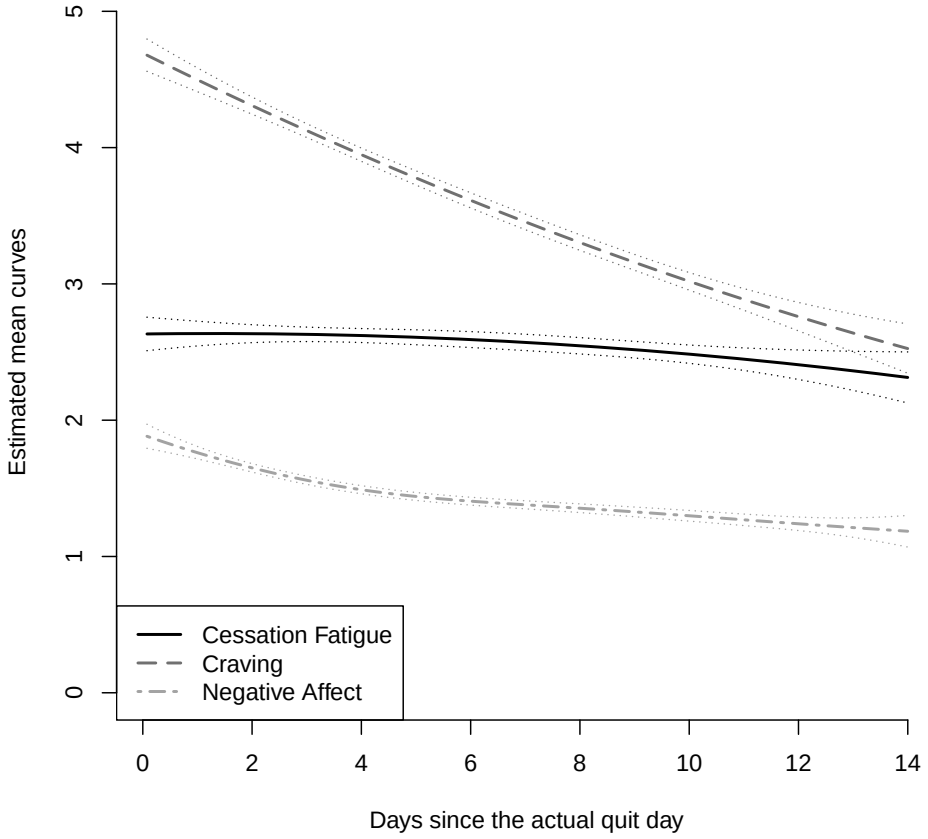

Figure 1: Nonparametric estimations of the three covariate processes

Table 5: Estimation for Lapse Data with Single Covariate Process

	Craving				Negative Affect				Fatigue			
	Separate Estimation		Joint Likelihood		Separate Estimation		Joint Likelihood		Separate Estimation		Joint Likelihood	
	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE
η_1	-0.08	0.12	-0.22	0.16	-0.08	0.12	-0.22	0.16	-0.05	0.12	-0.20	0.17
η_2	-0.08	0.12	-0.45*	0.16	-0.10	0.12	-0.48*	0.16	-0.08	0.13	-0.48*	0.18
β	0.06*	0.01	0.13*	0.02	0.10*	0.02	0.15*	0.04	0.03*	0.01	0.03	0.02
μ_1	4.63*	0.11	4.64*	0.10	1.83*	0.05	1.84*	0.05	2.72*	0.11	2.72*	0.29
μ_2	-0.17*	0.11	-0.16*	0.01	1.83*	0.05	1.84*	0.05	-0.006	0.01	-0.005	0.01
Σ_{11}	7.86	-	7.89	0.33	2.00	-	2.00	0.14	8.96	-	8.95	1.05
Σ_{12}	-0.24	-	-0.23	0.03	-0.08	-	-0.08	0.01	-0.18	-	-0.18	0.07
Σ_{22}	0.04	-	0.04	0.005	0.01	-	0.01	0.001	0.10	-	0.10	0.02
σ^2	2.04	-	4.16	0.19	0.95	-	0.91	0.04	1.52	-	2.30	0.25

Table 5 represents the estimation results for the single-process models with days to lapse as the survival response and craving, negative affect, and fatigue as the single predictors in the three models, respectively. The estimation results indicate that the separation estimation and maximum joint likelihood method yield quit similar estimates for all the parameters in longitudinal models except σ^2, the variance of the random error. However, the parameter estimates for the survival models are quit different among the two models. Separate estimation suggests neither of the two active treatments are effective compared with placebo, whereas the MJL method suggests the combined therapy has significant effect in reducing the risk of lapse. This is consistent with the knowledge that the estimates of separate estimation tends to bias towards the null (Prentice, 1982). The separate estimation also shows that all the three longitudinal covariates are significantly positively associated with the risk of lapse, whereas according to MJL method, only Craving and Negative Affect are significant. Note that the absolute values of the coefficient estimates of the significant β's from MJL is much more larger than those from separate estimation.

*represents statistically significant at 0.05 level.

Table 6: Estimation for Lapse Data with Multiple Covariate Processes

		Separate Estimation		Joint Likelihood	
		Estimate	SE	Estimate	Bootstrap SE
	η_1	-0.07	0.12	-0.22	0.17
	η_2	-0.07	0.13	-0.45*	0.19
	β_{crav}	0.05*	0.01	0.13*	0.02
	β_{na}	0.05*	0.025	-0.004	0.05
	β_{fatig}	0.008	0.01	-0.001	0.02
Crav	μ_1	4.63*	0.10	4.64*	0.09
	μ_2	-0.17*	0.01	-0.16*	0.01
	Σ_{11}	7.86	-	7.89	0.36
	Σ_{12}	-0.24	-	-0.23	0.04
	Σ_{22}	0.04	-	0.04	0.01
	σ^2	2.04	-	4.16	0.16
NA	μ_1	1.83*	0.05	1.83*	0.05
	μ_2	-0.06*	0.004	-0.06*	0.005
	Σ_{11}	2.00	-	2.00	0.14
	Σ_{12}	-0.08	-	-0.08	0.01
	Σ_{22}	0.01	-	0.01	0.001
	σ^2	0.95	-	0.91	0.04
Fatig	μ_1	2.72*	0.11	2.72*	0.11
	μ_2	-0.005	0.01	-0.006	0.01
	Σ_{11}	8.96	-	8.98	0.55
	Σ_{12}	-0.18	-	-0.18	0.06
	Σ_{22}	0.10	-	0.10	0.02
	σ^2	1.52	-	2.19	0.12

Table 6 summarizes the results of joint models with three longitudinal covariates modeled simultaneously. Similar to the single-covariate models, the estimating results are close across the two methods for longitudinal parameters but show substantial difference for survival coefficients. The two active treatments are again estimated to be non-significant by the separate estimation, whereas MJL method implies the combined therapy has significant effect in reducing the risk of lapse. Separate estimation shows both Craving and Negative Affect have significant positive effects on lapse, whereas MJL indicates that Craving is the only significant longitudinal factor associated with lapse when the three longitudinal covariates are modeled together.

Compared with the results in Table 5, we found that the process of Negative Affect is identified to be a significant covariate for lapse in the single-covariate model, but becomes nonsignificant in the multiple-covariates model. The latter inference is consistent with the conjecture of the smoking cessation study that Negative Affect exerts influence on patients mainly through the feeling of Craving for smoking. Thus, using multiple-covariates model instead of single-covariate models does provide more reasonable insight into the analysis in this case.

Table 7: Estimation for Relapse Data with Single Covariate Process

	Craving				Negative Affect				Fatigue			
	Separate Estimation		Joint Likelihood		Separate Estimation		Joint Likelihood		Separate Estimation		Joint Likelihood	
	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE
η_1	-0.04	0.12	-0.57	0.32	-0.03	0.12	-0.56	0.30	-0.02	0.12	-0.57	0.32
η_2	-0.05	0.12	-0.66*	0.33	-0.04	0.11	-0.63	0.32	-0.04	0.12	-0.70*	0.33
β	0.04*	0.01	0.19*	0.04	0.08*	0.02	0.41*	0.07	0.002	0.01	0.06*	0.03
μ_1	4.63*	0.10	4.64*	0.11	1.82*	0.05	1.82*	0.05	2.74*	0.10	2.74*	0.12
μ_2	-0.15*	0.01	-0.15*	0.01	-0.05*	0.004	-0.05*	0.005	-0.003	0.01	-0.002	0.01
Σ_{11}	7.81	-	7.84	0.29	1.94	-	1.95	0.14	7.82	-	8.91	0.50
Σ_{12}	-0.22	-	-0.22	0.03	-0.08	-	-0.08	0.01	-0.22	-	-0.15	0.04
Σ_{22}	0.04	-	0.04	0.004	0.01	-	0.01	0.001	0.04	-	0.07	0.01
σ^2	2.11	-	4.45	0.15	0.98	-	0.97	0.04	2.10	-	2.40	0.13

Table 8: Estimation for Relapse Data with Multiple Covariate Processes

		Separate Estimation		Joint Likelihood	
		Estimate	SE	Estimate	Bootstrap SE
	η_1	-0.03	0.12	-0.55	0.33
	η_2	-0.03	0.12	-0.61	0.35
	β_{crav}	0.02*	0.01	0.11*	0.05
	β_{na}	0.05	0.03	0.31*	0.10
	β_{fatig}	0.01	0.01	-0.01	0.04
Crav	μ_1	4.64*	0.11	4.64*	0.10
	μ_2	-0.15*	0.01	-0.15*	0.01
	Σ_{11}	7.82	-	7.82	0.37
	Σ_{12}	-0.22	-	-0.22	0.03
	Σ_{22}	0.04	-	0.04	0.004
	σ^2	2.11	-	4.16	0.17
NA	μ_1	1.82*	0.05	1.83*	0.05
	μ_2	-0.05*	0.004	-0.06*	0.005
	Σ_{11}	1.94	-	1.94	0.14
	Σ_{12}	-0.08	-	-0.08	0.01
	Σ_{22}	0.01	-	0.01	0.001
	σ^2	0.98	-	0.97	0.04
Fatig	μ_1	2.74*	0.11	2.74*	0.12
	μ_2	-0.003	0.01	-0.003	0.01
	Σ_{11}	8.92	-	8.93	0.50
	Σ_{12}	-0.15	-	-0.16	0.04
	Σ_{22}	0.07	-	0.07	0.01
	σ^2	1.55	-	2.30	0.11

Table 7 and Table 8 present the estimation results for the models with time to relapse as the the survival response. The results of the single-covariate models are summarized in Table 7, where all the three covariate processes are identified to be significantly positively associated with relapse by MJL method. The combined treatment is shown to have significant effect on reducing the risk of relapse by MJL method when modeled with Craving or Fatigue. The two treatments are still nonsignificant in separate estimation.

In Table 8 both the combined treatment and the process of Fatigue are no longer significant by MJL method, even though the absolute values of the coefficient estimates increase dramatically compared with those of the separate estimation. Note that in the Relapse Data we have an extremely high censoring rate (91.24%), which means the information might be insufficient to make accurate inference on the survival coefficients. However, the fact that the estimation results are different between the single-covariate models and the multiple-covariate model still indicates the necessity of applying them separately to different questions according to specific needs.

4 Discussions

We proposed a likelihood-based method to estimate the joint models with multiple longitudinal covariates for the time to event response. Measurement errors and the covariance structure among the multiple longitudinal covariates were considered in the model. The asymptotic properties were established for the related maximum likelihood estimators. To circumvent the computational complexity involved in evaluating the large-dimensional integral in the likelihood function, we introduced the DoIt algorithm, which was implemented in R and combined with EM algorithm to speed up the convergence. In real data analysis, we found that the models with multiple biomarkers (i.e., longitudinal covariates) generated different results from the models with single biomarker, and the former offered a more com-

prehensive interpretation of the data. Although in this data example we only demonstrated the feasibility of the proposed approach with three biomarkers, the algorithm can accommodate the data with a larger number of longitudinal predictors of interest. The computing feasibility makes it possible to consider more complex joint model settings in future work. For example, one may consider nonparametric mixed-effects model for the longitudinal processes, or generalized mixed-effects model for the non-continuous longitudinal processes. However, a limitation of our and the similar approach is that there was no explicit form of the standard errors for the estimators, the numerical solution was difficult, and we had to bootstrap the standard errors at the cost of computing time.

5 Technical conditions and technical proofs

The joint model (2.5) and (2.6) resemble the setting of Zeng and Cai (2005) except that (2.4) extends the longitudinal model from a single longitudinal covariate scenario to a case with multivariate longitudinal covariates. Therefore, the same techniques in Zeng and Cai (2005) can be borrowed and extended to justify the asymptotic properties of the MLE's in our case. We state the assumptions for the theorems in the following paragraphs. The assumptions are made for any given subject in the study. The notations, if not respecified, follow the ones defined in section 2.

(A.1) Denote τ to be the end of the study time, $\mathcal{T}$ to be the longitudinal observation time, and $\mathcal{Z}$ to be the baseline covariates. In the interval $[0, \tau]$ and given the random effects $\mathbf{b}$, $\mathcal{T}$ and $\mathcal{Z}$ are conditionally independent of all the random variables in model (2.5) and (2.6).

(A.2) With probability one, every dimension of the functional vector $\boldsymbol{\rho}(t)$ and $\tilde{\boldsymbol{\rho}}(t)$ is continuously differentiable in $[0, \tau]$, and $\max_{t\in[0,\tau]} \|\boldsymbol{\rho}'(t)\| < \infty$, $\max_{t\in[0,\tau]} \|\tilde{\boldsymbol{\rho}}'(t)\| < \infty$. In

addition, the baseline covariates $\mathcal{Z}$ are bounded with probability one.

(A.3) Conditional on $\mathcal{T}$ and $\mathcal{Z}$, the censoring time C is non-informative for the joint model (i.e., given $\mathcal{T}$ and $\mathcal{Z}$, C is independent of T, $\mathcal{W}$ and $\mathbf{b}$).

(A.4) Let N be the number of longitudinal observations. There exists an integer n_0 such that $P(N \leq n_0) = 1$. In addition, $P(N > 2d|\mathcal{T}, \mathcal{Z}, T) > 0$ and $P(N > \tilde{d}|\mathcal{T}, \mathcal{Z}, T) > 0$ with probability one, where d is the total dimension of the random effects $\mathbf{b}$, and $\tilde{d}$ is the total dimension of the $(\tilde{\boldsymbol{\rho}}_1, \ldots, \tilde{\boldsymbol{\rho}}_p)$ for $(\boldsymbol{\mu}_1(t), \ldots, \boldsymbol{\mu}_p(t))$.

(A.5) The maximal right-censoring time is equal to τ.

(A.6) At time t_j, denote the value of $\boldsymbol{\rho}(t_j)$ and $\tilde{\boldsymbol{\rho}}(t_j)$ by $\boldsymbol{\rho}_j$ and $\tilde{\boldsymbol{\rho}}_j$, respectively. Remember that $\boldsymbol{\rho}_j$ is a $d \times p$ matrix, and $\tilde{\boldsymbol{\rho}}_j$ is a $\tilde{d} \times p$ matrix, where $d \geq p$ and $\tilde{d} \geq p$. Both $P(\boldsymbol{\rho}_j$ is full rank) and $P(\tilde{\boldsymbol{\rho}}_j$ is full rank) are positive for all $j = 1, \ldots, N$. In addition, define the $d \times pN$ matrix $\mathbf{R} = (\boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_N)$, then $P(\mathbf{R}\mathbf{R}^T$ is full rank) is also positive.

(A.7) If there exists constant vectors $\boldsymbol{\beta}_c$ and $\boldsymbol{\eta}_c$ (of the same dimensions with $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$, respectively) such that, with positive probability, $\boldsymbol{\beta}_c^T(\boldsymbol{\rho}(t)^T\mathbf{1}) = \beta_{c1}(\boldsymbol{\rho}_1(t)^T\mathbf{1}) + \cdots + \beta_{cp}(\boldsymbol{\rho}_p(t)^T\mathbf{1}) = g(t)$ and $\boldsymbol{\eta}_c^T\mathbf{Z} = 0$ for a deterministic function $g(t)$ for all $t \in [0, \tau]$, then $\boldsymbol{\beta}_c = \mathbf{0}$, $\boldsymbol{\eta}_c = \mathbf{0}$ and $g(t) = 0$.

(A.8) Let $\Theta \subseteq \mathbb{R}^{d_\theta}$ be the domain of $\boldsymbol{\theta}$, where d_θ is the dimension of $\boldsymbol{\theta}$. For any $\boldsymbol{\theta} \in \Theta$, assume $\|\boldsymbol{\theta}\| \leq M_0$, $\min_{\|\mathbf{e}\|=1} \mathbf{e}^T\boldsymbol{\Sigma}_e\mathbf{e} > M_0^{-1}, \min_{\|\mathbf{e}\|=1} \mathbf{e}^T\boldsymbol{\Sigma}_b\mathbf{e} > M_0^{-1}$ for a known positive constant M_0.

(A.9) The true baseline hazard function, denoted by $\lambda_0(t)$, is bounded and positive in $[0, \tau]$.

Remark Assumption (A.4) implies that for all the subjects, the number of longitudinal observations is bounded from above by n_0, and for at least some subjects the number of

longitudinal observations is bounded from below by the $2d$. Assumption (A.8) indicates that Θ is a compact set. (A.8) and (A.9) together indicate the true hazard function is bounded and positive in $[0, \tau]$. Combined with assumption (A.2), this implies that $P(T > \tau|\mathcal{T}, \mathcal{Z}) > c_0$ for some positive constant c_0. Since (A.5) assumes that all the subjects surviving after τ censor at τ (i.e., $C = \tau$), this implies that $P(C \geq \tau|\mathcal{T}, \mathcal{Z}) = P(C = \tau|\mathcal{T}, \mathcal{Z}) > c_0$.

Recall that $\boldsymbol{\theta} = (\boldsymbol{\mu}^T, \text{Vec}(\boldsymbol{\Sigma}_e), \text{Vec}(\boldsymbol{\Sigma}_b), \boldsymbol{\beta}^T, \boldsymbol{\eta}^T)^T$ is the constant parameter set and $\Lambda(t) = \int_0^t \lambda(s)ds$ is the functional parameter of the likelihood. The observed likelihood function of $(\boldsymbol{\theta}, \Lambda)$ is

$$
\begin{aligned}
L(\boldsymbol{\theta}, \Lambda) = \prod_{i=1}^{n} \int_{\mathbf{b}} \Bigg\{ & (2\pi)^{-pN_i/2} |\boldsymbol{\Sigma}_e|^{-N_i/2} \exp\{-\sum_{j=1}^{N_i} (\mathbf{W}_{ij} - \tilde{\boldsymbol{\rho}}_{ij}^T \boldsymbol{\mu} - \boldsymbol{\rho}_{ij}^T \mathbf{b})^T \boldsymbol{\Sigma}_e^{-1} (\mathbf{W}_{ij} - \tilde{\boldsymbol{\rho}}_{ij}^T \boldsymbol{\mu} - \boldsymbol{\rho}_{ij}^T \mathbf{b})\} \\
& \times \lambda(V_i)^{\Delta_i} \exp\left[\Delta_i \{\boldsymbol{\beta}^T (\boldsymbol{\rho}_i(V_i)^T \mathbf{b}) + \boldsymbol{\eta}^T \mathbf{Z}_i\} - \int_0^{V_i} e^{\boldsymbol{\beta}^T (\boldsymbol{\rho}_i(s)^T \mathbf{b}) + \boldsymbol{\eta}^T \mathbf{z}_i} d\Lambda(s) \right] \\
& \times (2\pi)^{-d/2} |\boldsymbol{\Sigma}_b|^{-1/2} \exp\{-\frac{1}{2} \mathbf{b}^T \boldsymbol{\Sigma}_b^{-1} \mathbf{b}\} \Bigg\} d\mathbf{b}. \qquad (5.1)
\end{aligned}
$$

In order to obtain the maximum likelihood estimator $(\widehat{\boldsymbol{\theta}}, \widehat{\Lambda})$, we let $\lambda(t)$ take mass only at the the event time V_i for which $\Delta_i = 1$. Thus $\Lambda(t)$ becomes an increasing and right-continuous step function with jumps only at V_i, and the baseline hazard $\lambda(t)$ in (5.1) is replaced by $\Lambda\{V_i\}$, the jump size of $\Lambda(\cdot)$ at V_i. The domain of $\Lambda(t)$ is denoted by $\mathcal{V}_n$, which consists of all the right-continuous step functions with jump at V_i for which $\Delta_i = 1$. The domain depends on n because the number of the jumps of $\Lambda(t)$ equals n. Denote the logarithm of the modified likelihood function by $l_n(\boldsymbol{\theta}, \Lambda)$.

Below are the proofs of Theorem 1 and 2 based on the above conditions.

Proof of Theorem 1

The proof of Theorem 1 is completed by verifying the following statements (i) - (iv). Note that all the statements are made and hold for a fixed ω in the probability space, except for some null sets.

(i) The maximizer of $l_n(\boldsymbol{\theta}, \Lambda)$, $(\widehat{\boldsymbol{\theta}}, \widehat{\Lambda})$ exists for each n.

(ii) $\widehat{\Lambda}(\tau)$ is bounded when n goes to infinity.

(iii) There exist a constant vector $\boldsymbol{\theta}^*$ and a right-continuous monotone function $\Lambda^*(t)$ such that $\widehat{\boldsymbol{\theta}} \to \boldsymbol{\theta}^*$ and $\widehat{\Lambda}(t)$ weakly converges to $\Lambda^*(t)$ for $t \in [0, \tau]$.

(iv) $\boldsymbol{\theta}^* = \boldsymbol{\theta}_0$ and $\Lambda^* = \Lambda_0$.

PROOF OF (i). Recall that by replacing $\lambda(V_i)$ with $\Lambda\{V_i\}$ in (5.1), the objective function becomes

$$
\begin{aligned}
l_n(\boldsymbol{\theta}, \Lambda) = \sum_{i=1}^{n} \log \int_{\mathbf{b}} \Bigg\{ & (2\pi)^{-pN_i/2} |\boldsymbol{\Sigma}_e|^{-N_i/2} \exp\{-\frac{1}{2} \sum_{j=1}^{N_i} (\mathbf{W}_{ij} - \tilde{\boldsymbol{\rho}}_{ij}^T \boldsymbol{\mu} - \boldsymbol{\rho}_{ij}^T \mathbf{b})^T \boldsymbol{\Sigma}_e^{-1} (\mathbf{W}_{ij} - \tilde{\boldsymbol{\rho}}_{ij}^T \boldsymbol{\mu} - \boldsymbol{\rho}_{ij}^T \mathbf{b})\} \\
& \times \Lambda\{V_i\}^{\Delta_i} \exp\left[\Delta_i \{\boldsymbol{\beta}^T (\boldsymbol{\rho}_i(V_i)^T \mathbf{b}) + \boldsymbol{\eta}^T \mathbf{Z}_i\} - \sum_{j=1}^{n} I(V_j \le V_i) \Lambda\{V_j\} e^{\boldsymbol{\beta}^T (\boldsymbol{\rho}_i(V_j)^T \mathbf{b}) + \boldsymbol{\eta}^T \mathbf{Z}_i} \right] \\
& \times (2\pi)^{-d/2} |\boldsymbol{\Sigma}_b|^{-1/2} \exp\{-\frac{1}{2} \mathbf{b}^T \boldsymbol{\Sigma}_b^{-1} \mathbf{b}\} \Bigg\} d\mathbf{b}, \qquad (5.2)
\end{aligned}
$$

and $(\widehat{\boldsymbol{\theta}}, \widehat{\Lambda})$ maximizes $l_n(\boldsymbol{\theta}, \Lambda)$ over the set $\{(\boldsymbol{\theta}, \Lambda) : \boldsymbol{\theta} \in \Theta, \Lambda \in \mathcal{V}_n\}$. Since it is easy to verify that $l_n(\boldsymbol{\theta}, \Lambda)$ is concave, the existence of $\widehat{\boldsymbol{\theta}}$ holds because its domain Θ is compact. Thus we only need to verify the existence of $\widehat{\Lambda}$, which is satisfied when $\mathcal{V}_n$ is compact. Hence it is suffice to prove that the jump size of Λ at V_i for which $\Delta_i = 1$ is finite.

Since for any $x > 0$, $e^x \geq 1 + x$, and $e^{-x} \leq (1 + x)^{-1}$, for each i we have

$$\begin{aligned}
&\exp\left[-\sum_{j=1}^{n} I(V_j \leq V_i)\Lambda\{V_j\}e^{\boldsymbol{\beta}^T(\boldsymbol{\rho}_i(V_j)^T\mathbf{b})+\boldsymbol{\eta}^T\mathbf{z}_i}\right] \\
\leq &\left[1+\sum_{j=1}^{n} I(V_j \leq V_i)\Lambda\{V_j\}e^{\boldsymbol{\beta}^T(\boldsymbol{\rho}_i(V_j)^T\mathbf{b})+\boldsymbol{\eta}^T\mathbf{z}_i}\right]^{-1} \\
\leq &\left[\sum_{j=1}^{n} I(V_j \leq V_i)\Lambda\{V_j\}e^{\boldsymbol{\beta}^T(\boldsymbol{\rho}_i(V_j)^T\mathbf{b})+\boldsymbol{\eta}^T\mathbf{z}_i}\right]^{-1} \\
\leq &\Lambda\{V_{j0}^{(i)}\}^{-1}e^{-\{\boldsymbol{\beta}^T(\boldsymbol{\rho}_i(V_{j0}^{(i)})^T\mathbf{b})+\boldsymbol{\eta}^T\mathbf{z}_i\}},
\end{aligned}$$

where $V_{j0}^{(i)}$ is any observed event time in the set $\{V_j : V_j \leq V_i, \Delta_j = 1, j = 1, \ldots, n\}$.

Hence, for any i such that $\Delta_i = 1$, take $V_{j0}^{(i)} = V_i$, the second line of the likelihood (5.2) satisfies

$$\begin{aligned}
&\Lambda\{V_i\}^{\Delta_i}\exp\left[\Delta_i\{\boldsymbol{\beta}^T(\boldsymbol{\rho}_i(V_i)^T\mathbf{b})+\boldsymbol{\eta}^T\mathbf{Z}_i\}-\sum_{j=1}^{n} I(V_j \leq V_i)\Lambda\{V_j\}e^{\boldsymbol{\beta}^T(\boldsymbol{\rho}_i(V_j)^T\mathbf{b})+\boldsymbol{\eta}^T\mathbf{z}_i}\right] \\
\leq &\Lambda\{V_i\}\exp\left[\{\boldsymbol{\beta}^T(\boldsymbol{\rho}_i(V_i)^T\mathbf{b})+\boldsymbol{\eta}^T\mathbf{Z}_i\}\right]\times\Lambda\{V_i\}^{-1}\exp\left[-\{\boldsymbol{\beta}^T(\boldsymbol{\rho}_i(V_i)^T\mathbf{b})+\boldsymbol{\eta}^T\mathbf{Z}_i\}\right]=1.
\end{aligned}$$

For those i with $\Delta_i = 0$, take any $V_{j0}^{(i)} \in \{V_j : V_j \leq V_i, \Delta_j = 1, j = 1, \ldots, n\}$, the second line of the likelihood (5.2) satisfies

$$\begin{aligned}
&\Lambda\{V_i\}^{\Delta_i}\exp\left[\Delta_i\{\boldsymbol{\beta}^T(\boldsymbol{\rho}_i(V_i)^T\mathbf{b})+\boldsymbol{\eta}^T\mathbf{Z}_i\}-\sum_{j=1}^{n} I(V_j \leq V_i)\Lambda\{V_j\}e^{\boldsymbol{\beta}^T(\boldsymbol{\rho}_i(V_j)^T\mathbf{b})+\boldsymbol{\eta}^T\mathbf{z}_i}\right] \\
\leq &\Lambda\{V_{j0}^{(i)}\}^{-1}\exp\left[-\{\boldsymbol{\beta}^T(\boldsymbol{\rho}_i(V_{j0}^{(i)})^T\mathbf{b})+\boldsymbol{\eta}^T\mathbf{Z}_i\}\right].
\end{aligned}$$

Accordingly, the likelihood (5.2) satisfies

$$
\begin{aligned}
l_n(\boldsymbol{\theta},\Lambda) \le & \sum_{i=1}^{n} I(\Delta_i=1)\log\int_{\mathbf{b}}\big\{(2\pi)^{-pN_i/2}|\boldsymbol{\Sigma}_e|^{-N_i/2} \\
& \times\exp\{-\sum_{j=1}^{N_i}(\mathbf{W}_{ij}-\tilde{\boldsymbol{\rho}}_{ij}^T\boldsymbol{\mu}-\boldsymbol{\rho}_{ij}^T\mathbf{b})^T\boldsymbol{\Sigma}_e^{-1}(\mathbf{W}_{ij}-\tilde{\boldsymbol{\rho}}_{ij}^T\boldsymbol{\mu}-\boldsymbol{\rho}_{ij}^T\mathbf{b})\} \\
& \times(2\pi)^{-d/2}|\boldsymbol{\Sigma}_b|^{-1/2}\exp\{-\frac{1}{2}\mathbf{b}^T\boldsymbol{\Sigma}_b^{-1}\mathbf{b}\}\Big\}\,d\mathbf{b}. \\
& +\sum_{i=1}^{n} I(\Delta_i=0)\log\int_{\mathbf{b}}\big\{(2\pi)^{-pN_i/2}|\boldsymbol{\Sigma}_e|^{-N_i/2} \\
& \times\exp\{-\sum_{j=1}^{N_i}(\mathbf{W}_{ij}-\tilde{\boldsymbol{\rho}}_{ij}^T\boldsymbol{\mu}-\boldsymbol{\rho}_{ij}^T\mathbf{b})^T\boldsymbol{\Sigma}_e^{-1}(\mathbf{W}_{ij}-\tilde{\boldsymbol{\rho}}_{ij}^T\boldsymbol{\mu}-\boldsymbol{\rho}_{ij}^T\mathbf{b})\} \\
& \times\Lambda\{V_{j0}^{(i)}\}^{-\Delta_i}\exp\Big\{-(\boldsymbol{\beta}^T(\boldsymbol{\rho}_i(V_{j0}^{(i)})^T\mathbf{b})+\boldsymbol{\eta}^T\mathbf{Z}_i)\Big\} \\
& \times(2\pi)^{-d/2}|\boldsymbol{\Sigma}_b|^{-1/2}\exp\{-\frac{1}{2}\mathbf{b}^T\boldsymbol{\Sigma}_b^{-1}\mathbf{b}\}\Big\}\,d\mathbf{b}.
\end{aligned}
$$

Thus if $\Lambda\{V_{j0}^{(i)}\}\to\infty$ for some i and j_0, it follows that $l_n(\boldsymbol{\theta},\Lambda)\to-\infty$. We conclude the jump size of Λ is finite. Therefore, the maximum likelihood estimate $(\widehat{\boldsymbol{\theta}},\widehat{\Lambda})$ exists. □

PROOF OF (ii). Define $\widehat{\xi}=\log\widehat{\Lambda}(\tau)$ and $\tilde{\Lambda}=\widehat{\Lambda}/e^{\widehat{\xi}}$. Thus $\tilde{\Lambda}(\tau)=1$. Note that after rescaling, $\tilde{\Lambda}(t)=\widehat{\Lambda}(t)/\widehat{\Lambda}(\tau)$ is the ratio of jump size at time t relative to that at time τ, and the range of $\tilde{\Lambda}$ is $[0,1]$. To prove (ii), it is sufficient to show that $\widehat{\xi}$ is bounded when n goes to infinity. By applying some algebra to (5.1), for any $\Lambda\in\mathcal{V}_n$, the average of the likelihood at $\widehat{\boldsymbol{\theta}}$ over the n subjects is given by

$$
\begin{aligned}
\frac{1}{n}l_n(\widehat{\boldsymbol{\theta}},\Lambda) = & -\frac{\sum_{i=1}^n N_i}{n}\log\{(2\pi)^{-p/2}|\widehat{\boldsymbol{\Sigma}}_e|^{-1/2}\}-\log\{(2\pi)^{-d/2}|\widehat{\boldsymbol{\Sigma}}_b|^{-1/2}\} \\
& +\frac{1}{n}\sum_{i=1}^{n}\left[\Delta_i(\log\Lambda\{V_i\}+\widehat{\boldsymbol{\eta}}^T\mathbf{Z}_i)-\sum_{j=1}^{N_i}\frac{1}{2}(\mathbf{W}_{ij}-\tilde{\boldsymbol{\rho}}_{ij}^T\widehat{\boldsymbol{\mu}})^T\boldsymbol{\Sigma}_e^{-1}(\mathbf{W}_{ij}-\tilde{\boldsymbol{\rho}}_{ij}^T\widehat{\boldsymbol{\mu}})\right] \\
& +\frac{1}{n}\sum_{i=1}^{n}\log\int_{\mathbf{b}}\exp\left\{-\frac{1}{2}\mathbf{b}^T\boldsymbol{G}_i^{-1}\mathbf{b}+\boldsymbol{h}_i^T\mathbf{b}-\int_0^{V_i}e^{Q_{1i}(t,b,\widehat{\theta})}d\Lambda(t)\right\}d\mathbf{b}, \qquad (5.3)
\end{aligned}
$$

where

$$\boldsymbol{G}_i^{-1} = \widehat{\boldsymbol{\Sigma}}_b^{-1} + \sum_{j=1}^{N_i} \boldsymbol{\rho}_{ij} \widehat{\boldsymbol{\Sigma}}_e^{-1} \boldsymbol{\rho}_{ij}^T,$$

$$\boldsymbol{h}_i^T = \Delta_i \left\{ \widehat{\boldsymbol{\beta}} \boldsymbol{\rho}_i(V_i)^T \right\} + \sum_{j=1}^{N_i} (\mathbf{W}_{ij} - \tilde{\boldsymbol{\rho}}_{ij}^T \widehat{\boldsymbol{\mu}})^T \widehat{\boldsymbol{\Sigma}}_e^{-1} \boldsymbol{\rho}_{ij}^T,$$

$$Q_{1i}(t, b, \widehat{\theta}) = \widehat{\boldsymbol{\beta}} \left\{ \boldsymbol{\rho}_i(V_i)^T \mathbf{b} \right\} + \widehat{\boldsymbol{\eta}}^T \mathbf{Z}_i.$$

Let $\tilde{\mathbf{b}} = \boldsymbol{G}_i^{-1/2}(\mathbf{b} - \boldsymbol{G}_i \boldsymbol{h}_i)$, thus the third part of the right-hand side of equation (5.3) becomes

$$\frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{1}{2} \log |\boldsymbol{G}_i| + \frac{1}{2} \boldsymbol{h}_i^T \boldsymbol{G}_i \boldsymbol{h}_i + \log \int_{\tilde{\mathbf{b}}} \exp\{ -\frac{1}{2} \tilde{\mathbf{b}}^T \tilde{\mathbf{b}} - \int_0^{V_i} e^{\tilde{Q}_{1i}(t, \tilde{b}, \widehat{\theta})} d\Lambda(t) \} d\tilde{\mathbf{b}} \right\},$$

where

$$\tilde{Q}_{1i}(t, \tilde{b}, \widehat{\theta}) = \{ \boldsymbol{\beta} \boldsymbol{\rho}_i(t)^T \} \boldsymbol{G}_i^{1/2} \tilde{\mathbf{b}} + \{ \boldsymbol{\beta} \boldsymbol{\rho}_i(t)^T \} \boldsymbol{G}_i \boldsymbol{h}_i + \boldsymbol{\eta}^T \mathbf{Z}_i.$$

Since $\widehat{\xi}$ maximizes the log-likelihood $l_n(\widehat{\boldsymbol{\theta}}, e^{\xi} \tilde{\Lambda})$, we have $l_n(\widehat{\boldsymbol{\theta}}, e^{\widehat{\xi}} \tilde{\Lambda}) \geq l_n(\widehat{\boldsymbol{\theta}}, e^{0} \tilde{\Lambda})$. It follows that

$$\begin{aligned} 0 \leq & n^{-1} l_n(\widehat{\boldsymbol{\theta}}, e^{\widehat{\xi}} \tilde{\Lambda}) - n^{-1} l_n(\widehat{\boldsymbol{\theta}}, \tilde{\Lambda}) \\ = & \frac{1}{n} \sum_{i=1}^{n} \left[\Delta_i \widehat{\xi} + \log \int_{\tilde{\mathbf{b}}} \exp \left\{ -\frac{\tilde{\mathbf{b}}^T \tilde{\mathbf{b}}}{2} - e^{\widehat{\xi}} \int_0^{V_i} e^{\tilde{Q}_{1i}(t, \tilde{b}, \widehat{\theta})} d\tilde{\Lambda}(t) \right\} d\tilde{\mathbf{b}} \right. \\ & \left. - \log \int_{\tilde{\mathbf{b}}} \exp \left\{ -\frac{\tilde{\mathbf{b}}^T \tilde{\mathbf{b}}}{2} - \int_0^{V_i} e^{\tilde{Q}_{1i}(t, \tilde{b}, \widehat{\theta})} d\tilde{\Lambda}(t) \right\} d\tilde{\mathbf{b}} \right] \end{aligned} \tag{5.4}$$

Note that according to assumption (A.2), (A.4) and the boundedness of $\boldsymbol{\theta}$, there exist some positive constants C_1, C_2, and C_3 such that

$$|\tilde{Q}_{i1}(t, \tilde{b}, \widehat{\theta})| \leq C_1 \|\tilde{\mathbf{b}}\| + C_2 \sum_{j=1}^{N_i} \|\mathbf{W}_{ij}\| + C_3 \leq C_1 \|\tilde{\mathbf{b}}\| + C_2 n_0 \|\mathbf{W}_{ij}\| + C_3. \tag{5.5}$$

Since $\tilde{\mathbf{b}}$ is of standard multivariate normal distribution, applying the above inequality (5.5), we obtain

$$\begin{aligned} & - \log \int_{\tilde{\mathbf{b}}} \exp \left\{ -\frac{\tilde{\mathbf{b}}^T \tilde{\mathbf{b}}}{2} - \int_0^{V_i} e^{\tilde{Q}_{1i}(t, \tilde{b}, \widehat{\theta})} d\tilde{\Lambda}(t) \right\} d\tilde{\mathbf{b}} \\ & \quad = (2\pi)^{d/2} \log E_{\tilde{b}} \exp \left\{ - \int_0^{V_i} e^{\tilde{Q}_{1i}(t, \tilde{b}, \widehat{\theta})} d\tilde{\Lambda}(t) \right\} \\ & \quad \geq (2\pi)^{d/2} \log E_{\tilde{b}} \exp \left\{ -e^{C_1 \|\tilde{\mathbf{b}}\| + C_2 n_0 \|\mathbf{W}_{ij}\| + C_3} \right\} \end{aligned} \tag{5.6}$$

Using Jensen's inequality, the above expression satisfies

$$\begin{aligned}
&(2\pi)^{d/2}\log E_{\tilde{b}}\exp\left\{-e^{C_1\|\tilde{\mathbf{b}}\|+C_2n_0\|\mathbf{W}_{ij}\|+C_3}\right\}\\
\geq&(2\pi)^{d/2}E_{\tilde{b}}\left\{-e^{C_1\|\tilde{\mathbf{b}}\|+C_2n_0\|\mathbf{W}_{ij}\|+C_3}\right\}\\
=&-e^{C_2n_0\|\mathbf{W}_{ij}\|+C_4}
\end{aligned}\tag{5.7}$$

for some constant C_4. Since $\mathbf{W}_{ij}$ is normally distributed, by the strong law of large numbers, there exist some positive constant C_5 such that

$$\frac{1}{n}\sum_{i=1}^{n}e^{C_2n_0\|\mathbf{W}_{ij}\|+C_4}\to Ee^{C_2n_0\|\mathbf{W}_{1j}\|+C_4}=C_5\qquad a.s.$$

Thus

$$-\frac{1}{n}\sum_{i=1}^{n}\log\int_{\tilde{\mathbf{b}}}\exp\left\{-\frac{\tilde{\mathbf{b}}^T\tilde{\mathbf{b}}}{2}-\int_0^{V_i}e^{\tilde{Q}_{1i}(t,\tilde{b},\widehat{\theta})}d\tilde{\Lambda}(t)\right\}d\tilde{\mathbf{b}}$$

is bounded by C_5 from above when n goes to infinity. Then (5.4) becomes

$$\begin{aligned}
0\leq&\frac{1}{n}\sum_{i=1}^{n}\Delta_i\widehat{\xi}+\frac{1}{n}\sum_{i=1}^{n}\log\int_{\tilde{\mathbf{b}}}\exp\left\{-\frac{\tilde{\mathbf{b}}^T\tilde{\mathbf{b}}}{2}-e^{\widehat{\xi}}\int_0^{V_i}e^{\tilde{Q}_{1i}(t,\tilde{b},\widehat{\theta})}d\tilde{\Lambda}(t)\right\}d\tilde{\mathbf{b}}+C_5\\
\leq&\frac{1}{n}\sum_{i=1}^{n}\Delta_i\widehat{\xi}+\frac{1}{n}\sum_{i=1}^{n}I(V_i=\tau)\log\int_{\tilde{\mathbf{b}}}\exp\left\{-\frac{\tilde{\mathbf{b}}^T\tilde{\mathbf{b}}}{2}-e^{\widehat{\xi}}\int_0^{\tau}e^{\tilde{Q}_{1i}(t,\tilde{b},\widehat{\theta})}d\tilde{\Lambda}(t)\right\}d\tilde{\mathbf{b}}\\
&+\frac{1}{n}\sum_{i=1}^{n}I(V_i\neq\tau)\log\int_{\tilde{\mathbf{b}}}\exp\left\{-\frac{\tilde{\mathbf{b}}^T\tilde{\mathbf{b}}}{2}\right\}d\tilde{\mathbf{b}}+C_5\\
\leq&\frac{1}{n}\sum_{i=1}^{n}\Delta_i\widehat{\xi}+\frac{1}{n}\sum_{i=1}^{n}I(V_i=\tau)\log\int_{\tilde{\mathbf{b}}}\exp\left\{-\frac{\tilde{\mathbf{b}}^T\tilde{\mathbf{b}}}{2}-e^{\widehat{\xi}}\int_0^{\tau}e^{\tilde{Q}_{1i}(t,\tilde{b},\widehat{\theta})}d\tilde{\Lambda}(t)\right\}d\tilde{\mathbf{b}}+C_6,
\end{aligned}\tag{5.8}$$

for some constant C_6. The last inequality follows by the integral of the unnormalized standard normal density of $\tilde{\mathbf{b}}$. Since for any $\Gamma\geq 0$, $x\geq 0$, we have $e^{x/\Gamma}\geq(1+x/\Gamma)$, it follows that $e^{-x}\leq(1+x/\Gamma)^{-\Gamma}$. Using the similar arguments as in (5.6) and (5.7), the following inequality

holds

$$
\begin{aligned}
&\int_{\tilde{\mathbf{b}}} \exp\left\{-\frac{\tilde{\mathbf{b}}^T\tilde{\mathbf{b}}}{2} - e^{\widehat{\xi}}\int_0^\tau e^{\tilde{Q}_{1i}(t,\tilde{b},\widehat{\theta})}d\tilde{\Lambda}(t)\right\}d\tilde{\mathbf{b}} \\
\leq &\int_{\tilde{\mathbf{b}}} \exp\left\{-\frac{\tilde{\mathbf{b}}^T\tilde{\mathbf{b}}}{2}\right\}\left\{1+\frac{e^{\widehat{\xi}}}{\Gamma}\int_0^\tau e^{\tilde{Q}_{1i}(t,\tilde{b},\widehat{\theta})}d\tilde{\Lambda}(t)\right\}^{-\Gamma} d\tilde{\mathbf{b}} \\
\leq &\left(\frac{\Gamma}{e^{\widehat{\xi}}}\right)^\Gamma \int_{\tilde{\mathbf{b}}} \exp\left\{-\frac{\tilde{\mathbf{b}}^T\tilde{\mathbf{b}}}{2}\right\}\left\{\int_0^\tau e^{\tilde{Q}_{1i}(t,\tilde{b},\widehat{\theta})}d\tilde{\Lambda}(t)\right\}^{-\Gamma} d\tilde{\mathbf{b}} \\
\leq &\left(\frac{\Gamma}{e^{\widehat{\xi}}}\right)^\Gamma \int_{\tilde{\mathbf{b}}} \exp\left\{-\frac{\tilde{\mathbf{b}}^T\tilde{\mathbf{b}}}{2}\right\}\left\{e^{-C_1\|\tilde{\mathbf{b}}\|-C_2n_0\|\mathbf{W}_{ij}\|-C_3}\right\}^{-\Gamma} d\tilde{\mathbf{b}} \\
= &\left(\frac{\Gamma}{e^{\widehat{\xi}}}\right)^\Gamma (2\pi)^{d/2}E_{\tilde{b}}\left\{e^{\Gamma C_1\|\tilde{\mathbf{b}}\|+\Gamma C_2n_0\|\mathbf{W}_{ij}\|+\Gamma C_3}\right\} \\
= &\left(\frac{\Gamma}{e^{\widehat{\xi}}}\right)^\Gamma e^{\Gamma C_2n_0\|\mathbf{W}_{ij}\|+C_7(\Gamma)},
\end{aligned}
$$

where $C_7(\Gamma)$ is a deterministic function of Γ. Thus by strong law of large numbers, (5.8) becomes

$$
\begin{aligned}
0 \leq &\frac{1}{n}\sum_{i=1}^n \Delta_i\widehat{\xi} + \frac{1}{n}\sum_{i=1}^n I(V_i=\tau)\Gamma\left(\log\Gamma - \widehat{\xi} + C_2n_0\|\mathbf{W}_{ij}\| + C_7(\Gamma)\right) + C_6 \\
\leq &\frac{1}{n}\sum_{i=1}^n \Delta_i\widehat{\xi} - \frac{\Gamma}{n}\sum_{i=1}^n I(V_i=\tau)\widehat{\xi} + C_2n_0\frac{\Gamma}{n}\sum_{i=1}^n\|\mathbf{W}_{ij}\| + C_8(\Gamma) \\
\to &\frac{1}{n}\sum_{i=1}^n \Delta_i\widehat{\xi} - \frac{\Gamma}{n}\sum_{i=1}^n I(V_i=\tau)\widehat{\xi} + C_9(\Gamma), \qquad a.s.
\end{aligned} \tag{5.9}
$$

where both $C_8(\Gamma)$ and $C_9(\Gamma)$ are deterministic functions of Γ. Take Γ large enough such that $\frac{1}{n}\sum_{i=1}^n \Delta_i \leq \frac{\Gamma}{2n}\sum_{i=1}^n I(V_i=\tau)$, then by assumption (A.5) and the strong law of large numbers

$$
\begin{aligned}
&0 \leq C_9(\Gamma) - \frac{\Gamma}{2n}\sum_{i=1}^n I(V_i=\tau)\widehat{\xi} \\
&\widehat{\xi} \leq 2C_9(\Gamma)/\frac{\Gamma}{n}\sum_{i=1}^n I(V_i=\tau) \to 2C_9(\Gamma)/\Gamma P(V=\tau) = B_0 \qquad a.s.
\end{aligned} \tag{5.10}
$$

Thus $\widehat{\xi}$ is bounded by some constant B_0. Since the above statement holds for every ω in the sample space except the null set, we conclude that with probability 1, $\widehat{\Lambda}(\tau)$ is bounded for

any n. $\square$

PROOF OF (iii). By the assumption (A.8) that $\Theta \in \mathbb{R}^{d_\theta}$ is a compact set, there exists a subsequence of $\widehat{\boldsymbol{\theta}}_n$ and a constant vector $\boldsymbol{\theta}^* \in \Theta$ such that the subsequence converges to $\boldsymbol{\theta}^*$. If we can show in (iv) that $\boldsymbol{\theta}^* = \boldsymbol{\theta}_0$, the unique true parameter, let $\widehat{\boldsymbol{\theta}}_{n_m}$ be a subsequence of $\widehat{\boldsymbol{\theta}}_n$ that does not converge to $\boldsymbol{\theta}^*$. Thus, $\exists\ \delta_0 > 0$ and some large M s.t. $\|\widehat{\boldsymbol{\theta}}_{n_m} - \boldsymbol{\theta}^*\| > \delta_0$ for all the $m > M$. However, since $\widehat{\boldsymbol{\theta}}_{m_n} \in \Theta$ and the limit $\boldsymbol{\theta}^*$ is unique, there exists a sub-subsequence $\widehat{\boldsymbol{\theta}}_{m_{n_k}}$ s.t. $\widehat{\boldsymbol{\theta}}_{m_{n_k}} \to \boldsymbol{\theta}^*$ as $k \to \infty$. This contradict with the previous statement of the non-convergence of $\widehat{\boldsymbol{\theta}}_{m_n}$. Thus all the subsequence of $\widehat{\boldsymbol{\theta}}_n$ coverage to $\boldsymbol{\theta}^*$. We conclude that $\widehat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}^*$ as $n \to \infty$.

Now consider $\widehat{\Lambda}$. By Helly's Selection Theorem, there exists a subsequence of $\widehat{\Lambda}_n(t)$ that weakly converges to some right-continuous monotone function $\Lambda^*(t)$ for each $t \in [0, \tau]$. Using the similar argument as for $\widehat{\boldsymbol{\theta}}$, if we can show in (iv) that $\Lambda^* = \Lambda_0$, the unique true functional parameter, then $\widehat{\Lambda}_n$ itself converges to Λ^* as n goes to infinity. Note that since the convergence holds for every ω in the sample space except the null sets, $\widehat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}^*$ and $\widehat{\Lambda}_n \to \Lambda^*$ with probability 1. Thus the only thing left to prove is $\boldsymbol{\theta}^* = \boldsymbol{\theta}_0$ and $\Lambda^* = \Lambda_0$. $\square$

PROOF OF (iv). For a given subject, let $\mathbf{O} = \{\mathbf{W}_j, \boldsymbol{\rho}(s), \tilde{\boldsymbol{\rho}}_j, V, \Delta, \mathbf{Z}, j = 1, \ldots, N, 0 \leq s \leq t\}$ be the observed data. Denote

$$\begin{aligned} G(\mathbf{b}, \mathbf{O}; \boldsymbol{\theta}, \Lambda) =& (2\pi)^{-pN/2} |\boldsymbol{\Sigma}_e|^{-N/2} (2\pi)^{d/2} |\boldsymbol{\Sigma}_b|^{-1/2} \\ & \times \exp\left\{ -\frac{1}{2} \sum_{j=1}^{N} (\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T \boldsymbol{\mu} - \boldsymbol{\rho}_j^T \mathbf{b})^T \boldsymbol{\Sigma}_e^{-1} (\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T \boldsymbol{\mu} - \boldsymbol{\rho}_j^T \mathbf{b}) - \frac{\mathbf{b}^T \boldsymbol{\Sigma}_b^{-1} \mathbf{b}}{2} \right. \\ & \left. + \Delta \{ \boldsymbol{\beta}^T (\boldsymbol{\rho}(V)^T \mathbf{b}) + \boldsymbol{\eta}^T \mathbf{Z} \} - \int_0^V e^{\boldsymbol{\beta}^T (\boldsymbol{\rho}(s)^T \mathbf{b}) + \boldsymbol{\eta}^T \mathbf{Z}} d\Lambda(s) \right\}. \end{aligned}$$

In addition, define

$$Q(V, \mathbf{O}; \boldsymbol{\theta}, \Lambda) = \frac{\int_{\mathbf{b}} G(\mathbf{b}, \mathbf{O}; \boldsymbol{\theta}, \Lambda) \exp\{\boldsymbol{\beta}^T (\boldsymbol{\rho}(V)^T \mathbf{b}) + \boldsymbol{\eta}^T \mathbf{Z}\} d\mathbf{b}}{\int_{\mathbf{b}} G(\mathbf{b}, \mathbf{O}; \boldsymbol{\theta}, \Lambda) d\mathbf{b}}$$

to be the posterior expectation of $\exp\{\boldsymbol{\beta}^T(\boldsymbol{\rho}(V)^T\mathbf{b})+\boldsymbol{\eta}^T\mathbf{Z}\}$ with respect to $\mathbf{b}$ given $(\mathbf{O};\boldsymbol{\theta},\Lambda)$. For any measurable function $f(\mathbf{O})$, we use operator notation to define

$$\mathbf{P}_n f = \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{O}_i), \quad \text{and}$$

$$\mathbf{P} f = \int f d\mathbf{P} = E[f(\mathbf{O})].$$

Thus $\sqrt{n}(\mathbf{P}_n - \mathbf{P})$ is the associated empirical process based on $\mathbf{O}$. Define the class $\mathcal{F} = \{Q(V,\mathbf{O};\boldsymbol{\theta},\Lambda) : V \in [0,\tau], \boldsymbol{\theta} \in \Theta, \Lambda \in \mathcal{V}, \Lambda(0) = 0, \Lambda(\tau) \leq B_0\}$, where B_0 is the upper bound of $\Lambda(\tau)$ given in the proof of (ii), and $\mathcal{V}$ contains all the right-continuous monotone functions in $[0,\tau]$. Using the same techniques as in the Appendix A.1 of Zeng and Cai (2005), it is easy to verify that $\mathcal{F}$ is P-Donsker.

Differentiating $l_n(\boldsymbol{\theta},\Lambda)$ in (5.2) with respect to $\Lambda\{V_k\}$ and set the equation to zero, we obtain

$$\widehat{\Lambda}\{V_k\} = \frac{\Delta_k}{n\mathbf{P}_n\{I(V \geq v)Q(v,\mathbf{O};\widehat{\boldsymbol{\theta}},\widehat{\Lambda})\}|_{v=V_k}}. \tag{5.11}$$

Use the same structure, define

$$\bar{\Lambda}\{V_k\} = \frac{\Delta_k}{n\mathbf{P}_n\{I(V \geq v)Q(v,\mathbf{O};\boldsymbol{\theta}_0,\Lambda_0)\}|_{v=V_k}}. \tag{5.12}$$

Thus

$$\bar{\Lambda}(t) = \sum_{k=1}^{n} I(V_k \leq t)\bar{\Lambda}\{V_k\} = \frac{1}{n}\sum_{k=1}^{n} \frac{I(V_k \leq t)\Delta_k}{\mathbf{P}_n\{I(V \geq v)Q(v,\mathbf{O};\boldsymbol{\theta}_0,\Lambda_0)\}|_{v=V_k}}.$$

Since $(\boldsymbol{\theta}_0,\Lambda_0)$ maximizes $E[l_n(\boldsymbol{\theta},\Lambda)]$, by taking derivative of $E[l_n(\boldsymbol{\theta},\Lambda)]$ with respect to Λ and set the equation to 0, it can be verified that

$$\Lambda_0(t) = E\left[\frac{I(V_k \leq t)\Delta_k}{\mathbf{P}\{I(V \geq v)Q(v,\mathbf{O};\boldsymbol{\theta}_0,\Lambda_0)\}|_{v=V_k}}\right].$$

$$
\begin{aligned}
&\sup_{t\in[0,\tau]} |\bar{\Lambda}(t) - \Lambda_0(t)| \\
\leq &\sup_{t\in[0,\tau]} \left| \frac{1}{n}\sum_{k=1}^{n} I(V_k \leq t)\Delta_k \left[\frac{1}{\mathbf{P}_n\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\}} - \frac{1}{\mathbf{P}\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\}} \right] \right|_{v=V_k} \\
&+ \sup_{t\in[0,\tau]} \left| (\mathbf{P}_n - \mathbf{P}) \left[\frac{I(V_k \leq t)\Delta_k}{\mathbf{P}\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\}|_{v=V_k}} \right] \right| \\
\leq &\sup_{v\in[0,\tau]} \left| \frac{1}{\mathbf{P}_n\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\}} - \frac{1}{\mathbf{P}\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\}} \right| \\
&+ \sup_{t\in[0,\tau]} \left| (\mathbf{P}_n - \mathbf{P}) \left[\frac{I(V_k \leq t)\Delta_k}{\mathbf{P}\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\}|_{v=V_k}} \right] \right|
\end{aligned}
\tag{5.13}
$$

By Zeng and Cai (2005), $\mathbf{P}_n\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\}$ and $\mathbf{P}\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\}$ are bounded from below. Thus the first part of the above inequality satisfies

$$
\begin{aligned}
&\sup_{v\in[0,\tau]} \left| \frac{1}{\mathbf{P}_n\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\}} - \frac{1}{\mathbf{P}\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\}} \right| \\
= &\sup_{v\in[0,\tau]} \left| \frac{\mathbf{P}_n\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\} - \mathbf{P}\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\}}{\mathbf{P}_n\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\}\mathbf{P}\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\}} \right| \\
\leq &C_{10} \sup_{v\in[0,\tau]} |\mathbf{P}_n\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\} - \mathbf{P}\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\}| \,.
\end{aligned}
\tag{5.14}
$$

for some constant C_{10}.

Using the same argument as in Appendix A.1 of Zeng and Cai (2005), it can be verified that $\{Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0) : v \in [0, \tau]\}$ is a bounded Glivenko-Cantelli class. Since $\{I(V \geq v) : v \in [0, \tau]\}$ is also a Glivenko-Cantelli class, and the functional $(f, g) \mapsto fg$ for any two bounded functions f and g is Lipschitz continuous, $\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0) : v \in [0, \tau]\}$ is a bounded Glivenko-Cantelli class. Thus

$$
\sup_{v\in[0,\tau]} |\mathbf{P}_n\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\} - \mathbf{P}\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\}| \longrightarrow 0, \quad n \to \infty.
\tag{5.15}
$$

Hence the right-hand side of the above inequality (5.14) converge to 0 as n goes to infinity, and the first part of (5.13) disappears.

Similarly, since the class $\{I(V \le t)/\mathbf{P}\{I(V \ge v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\}|_{v=V} : t \in [0, \tau]\}$ is also a Glivenko-Cantelli class, the second part of the above inequality also converges to zero as n goes to infinity. Therefore by inequality (5.13) we conclude that

$$\sup_{t\in[0,\tau]} \|\bar{\Lambda}(t) - \Lambda_0(t)\| \to 0,$$

that is, $\bar{\Lambda}$ uniformly converges to Λ_0 in $[0, \tau]$.

Using the expressions of $\widehat{\Lambda}$ and $\bar{\Lambda}$ in (5.11) and (5.12), we have

$$\frac{\widehat{\Lambda}\{v\}}{\bar{\Lambda}\{v\}} = \frac{\mathbf{P}_n\{I(V \ge v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\}}{\mathbf{P}_n\{I(V \ge v)Q(v, \mathbf{O}; \widehat{\boldsymbol{\theta}}, \widehat{\Lambda})\}},$$

and accordingly,

$$\widehat{\Lambda}(t) = \int_0^t \frac{\mathbf{P}_n\{I(V \ge v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\}}{\mathbf{P}_n\{I(V \ge v)Q(v, \mathbf{O}; \widehat{\boldsymbol{\theta}}, \widehat{\Lambda})\}} d\bar{\Lambda}(v). \tag{5.16}$$

This implies that $\widehat{\Lambda}(t)$ is absolutely continuous with respect to $\bar{\Lambda}(t)$.

Since $\{I(V \ge v) : v \in [0, \tau]\}$ and $\mathcal{F}$ are both Glivenko-Cantelli classes, $\{I(V \ge v)Q(v, \mathbf{O}; \boldsymbol{\theta}, \Lambda) : v \in [0, \tau], \boldsymbol{\theta} \in \Theta, \Lambda \in \mathcal{V}, \Lambda_0(0) = 0, \Lambda(\tau) \le B_0\}$ is also a Glivenko-Cantelli class. Thus,

$$\sup_{v\in[0,\tau]} |(\mathbf{P}_n - \mathbf{P})\{I(V \ge v)Q(v, \mathbf{O}; \widehat{\boldsymbol{\theta}}, \widehat{\Lambda})\}| \to 0 \qquad a.s. \tag{5.17}$$

Since $\widehat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}^*$ and $\widehat{\Lambda}$ weakly converges to Λ^*, by bounded convergence theorem, for each $v \in [0, \tau]$, when n goes to infinity,

$$\mathbf{P}\{I(V \ge v)Q(v, \mathbf{O}; \widehat{\boldsymbol{\theta}}, \widehat{\Lambda})\} \to \mathbf{P}\{I(V \ge v)Q(v, \mathbf{O}; \boldsymbol{\theta}^*, \Lambda^*)\}. \tag{5.18}$$

Using assumption (A.2), it is easy to check that the derivative of $\mathbf{P}\{I(V \ge v)Q(v, \mathbf{O}; \boldsymbol{\theta}^*, \Lambda^*)\}$ with respect to v is uniformly bounded in $[0, \tau]$. Thus $\mathbf{P}\{I(V \ge v)Q(v, \mathbf{O}; \boldsymbol{\theta}^*, \Lambda^*)\}$ is equi-continuous. By Arzela-Ascoli theorem which states that a bounded and equi-continuous

functional sequence has uniformly convergent subsequence, we strengthen the conclusion in (5.18) to uniform convergence:

$$\sup_{v\in[0,\tau]} |\mathbf{P}\{I(V \geq v)Q(v, \mathbf{O}; \widehat{\boldsymbol{\theta}}, \widehat{\Lambda})\} - \mathbf{P}\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}^*, \Lambda^*)\}| \to 0, \quad n \to \infty.$$

This conclusion, together with (5.17), implies that

$$\sup_{v\in[0,\tau]} |\mathbf{P}_n\{I(V \geq v)Q(v, \mathbf{O}; \widehat{\boldsymbol{\theta}}, \widehat{\Lambda})\} - \mathbf{P}\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}^*, \Lambda^*)\}| \to 0. \tag{5.19}$$

By the conclusion of (5.19) and (5.15), it follows that

$$\frac{\widehat{\Lambda}\{v\}}{\bar{\Lambda}\{v\}} = \frac{\mathbf{P}_n\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\}}{\mathbf{P}_n\{I(V \geq v)Q(v, \mathbf{O}; \widehat{\boldsymbol{\theta}}, \widehat{\Lambda})\}} \to \frac{\mathbf{P}\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\}}{\mathbf{P}\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}^*, \Lambda^*)\}}, \tag{5.20}$$

uniformly in $[0, \tau]$.

Taking limit with respect to n on both sides of (5.16), and applying the conclusion of (5.20), we obtain

$$\Lambda^*(t) = \int_0^t \frac{\mathbf{P}\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\}}{\mathbf{P}\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}^*, \Lambda^*)\}} d\Lambda_0(v). \tag{5.21}$$

The above equation (5.21) indicates that both $\Lambda_0(t)$ and $\Lambda^*(t)$ are differentiable with respect to the Lebesgue measure. Denote $\lambda^*(t)$ to be the derivative of $\Lambda^*(t)$, and $\lambda_0(t)$ to be the derivative of $\Lambda_0(t)$. It follows from (5.21) that

$$\frac{\lambda^*(v)}{\lambda_0(v)} = \frac{\mathbf{P}\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\}}{\mathbf{P}\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}^*, \Lambda^*)\}}.$$

Thus (5.20) implies that

$$\frac{\widehat{\Lambda}\{V\}}{\bar{\Lambda}\{V\}} \to \frac{\lambda^*(V)}{\lambda_0(V)} \tag{5.22}$$

uniformly in $[0, \tau]$. Integrate the denominator and numerator over V in $[0, v]$ on both sides in (5.22), we obtain that $\widehat{\Lambda}(v)/\bar{\Lambda}(v)$ uniformly converges to $\Lambda^*(v)/\Lambda_0(v)$ in $[0, \tau]$. Since by (5.17) we already know that $\bar{\Lambda}(v)$ uniformly converges to $\Lambda_0(v)$ in $[0, \tau]$ for the denominators

in (5.5), we conclude that $\widehat{\Lambda}(v)$ uniformly converges to $\Lambda^*(v)$ in $[0, \tau]$.

Since $(\widehat{\boldsymbol{\theta}}, \widehat{\Lambda})$ maximizes $l_n(\boldsymbol{\theta}, \Lambda)$, it follows that

$$0 \leq \frac{1}{n} l_n(\widehat{\boldsymbol{\theta}}, \widehat{\Lambda}) - \frac{1}{n} l_n(\boldsymbol{\theta}_0, \bar{\Lambda}) = \mathbf{P}_n \left[\Delta \log \frac{\widehat{\Lambda}\{V\}}{\bar{\Lambda}\{V\}} \right] + \mathbf{P}_n \left[\log \frac{\int_{\mathbf{b}} G(\mathbf{b}, \mathbf{O}; \widehat{\boldsymbol{\theta}}, \widehat{\Lambda}) d\mathbf{b}}{\int_{\mathbf{b}} G(\mathbf{b}, \mathbf{O}; \boldsymbol{\theta}_0, \bar{\Lambda}) d\mathbf{b}} \right]. \quad (5.23)$$

Similar as Zeng and Cai (2005), it is easy to verify that $\Delta \log[\widehat{\Lambda}\{V\}/\bar{\Lambda}\{V\}]$ and $\log[\int_{\mathbf{b}} G(\mathbf{b}, \mathbf{O}; \widehat{\boldsymbol{\theta}}, \widehat{\Lambda}) d\mathbf{b} / \int_{\mathbf{b}} G(\mathbf{b}, \mathbf{O}; \boldsymbol{\theta}_0, \bar{\Lambda}) d\mathbf{b}]$ are both Glivenko-Cantelli classes. Thus

$$\begin{aligned} \sup_{V \in [0,\tau]} \left| (\mathbf{P}_n - \mathbf{P}) \left[\Delta \log \frac{\widehat{\Lambda}\{V\}}{\bar{\Lambda}\{V\}} \right] \right| &\to 0, \\ \sup_{V \in [0,\tau]} \left| (\mathbf{P}_n - \mathbf{P}) \left[\log \frac{\int_{\mathbf{b}} G(\mathbf{b}, \mathbf{O}; \widehat{\boldsymbol{\theta}}, \widehat{\Lambda}) d\mathbf{b}}{\int_{\mathbf{b}} G(\mathbf{b}, \mathbf{O}; \boldsymbol{\theta}_0, \bar{\Lambda}) d\mathbf{b}} \right] \right| &\to 0. \end{aligned} \quad (5.24)$$

Consider the first part on the right-hand side of equation (5.23). By (5.22) we know that $\widehat{\Lambda}\{V\}/\bar{\Lambda}\{V\}$ uniformly converges to $\lambda^*(V)/\lambda_0(V)$, thus by applying the bounded convergence theorem, we obtain

$$\mathbf{P} \left[\Delta \log \frac{\widehat{\Lambda}\{V\}}{\bar{\Lambda}\{V\}} \right] \to \mathbf{P} \left[\Delta \log \frac{\lambda^*(V)}{\lambda_0(V)} \right] \quad (5.25)$$

Similarly, for the second part on the right-hand side of (5.23), since $\widehat{\boldsymbol{\theta}} \to \boldsymbol{\theta}^*$, $\widehat{\Lambda}$ uniformly converges to Λ^* and $\bar{\Lambda}$ uniformly converges to Λ_0, applying the bounded convergence theorem again, we obtain

$$\mathbf{P} \left[\log \frac{\int_{\mathbf{b}} G(\mathbf{b}, \mathbf{O}; \widehat{\boldsymbol{\theta}}, \widehat{\Lambda}) d\mathbf{b}}{\int_{\mathbf{b}} G(\mathbf{b}, \mathbf{O}; \boldsymbol{\theta}_0, \bar{\Lambda}) d\mathbf{b}} \right] \to \mathbf{P} \left[\log \frac{\int_{\mathbf{b}} G(\mathbf{b}, \mathbf{O}; \boldsymbol{\theta}^*, \Lambda^*) d\mathbf{b}}{\int_{\mathbf{b}} G(\mathbf{b}, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0) d\mathbf{b}} \right]. \quad (5.26)$$

Combining the conclusions of (5.24), (5.25) and (5.26), taking limit with respect to n on both sides of (5.23), we obtain

$$\mathbf{P} \left[\log \left\{ \frac{\lambda^*(V)^{\Delta} \int_{\mathbf{b}} G(\mathbf{b}, \mathbf{O}; \boldsymbol{\theta}^*, \Lambda^*) d\mathbf{b}}{\lambda_0(V)^{\Delta} \int_{\mathbf{b}} G(\mathbf{b}, \mathbf{O}; \theta_0, \Lambda_0) d\mathbf{b}} \right\} \right] \geq 0.$$

Since the measure $\mathbf{P}$ is with respect to the distribution with true parameter $(\boldsymbol{\theta}_0, \Lambda_0)$, the left-hand side of the above inequality is the negative Kullback-Leibler information. Then it

follows that, with probability one,

$$\lambda^*(V)^{\Delta}\int_{\mathbf{b}} G(\mathbf{b},\mathbf{O};\boldsymbol{\theta}^*,\Lambda^*)d\mathbf{b} = \lambda_0(V)^{\Delta}\int_{\mathbf{b}} G(\mathbf{b},\mathbf{O};\boldsymbol{\theta}_0,\Lambda_0)d\mathbf{b}. \tag{5.27}$$

According to assumption (A.8) and (A.9), $P(V \le \tau, \Delta = 1|\mathcal{T},\mathcal{Z}) > 0$ with probability one. Thus equation (5.27) holds for the set $\{(V,\Delta) : V \in [0,\tau], \Delta = 1\}$. By assumption (A.5), $P(V = \tau, \Delta = 0|\mathcal{T},\mathcal{Z}) = P(C = \tau, \Delta = 0|\mathcal{T},\mathcal{Z}) > 0$ with probability one. Thus equation (5.27) also holds for the set $\{V = \tau, \Delta = 0\}$. However, since assumption (A.5), (A.8) and (A.9) imply that $P(C \ge \tau|\mathcal{T},\mathcal{Z}) > c_0$ with probability one for some positive constant c_0, $P(V < \tau, \Delta = 0|\mathcal{T},\mathcal{Z}) = P(C < \tau, \Delta = 0|\mathcal{T},\mathcal{Z}) < P(C < \tau|\mathcal{T},\mathcal{Z}) < 1 - c_0$ with probability one. Thus the set $\{(V,\Delta) : V \in [0,\tau), \Delta = 0\}$ might be a null set for which equation (5.27) may not hold. Zeng and Cai (2005) proved that equation (5.27) also holds for $\{(V,\Delta) : V \in [0,\tau), \Delta = 0\}$.

Let $\Delta = 0$ and $V = 0$ in equation (5.27). Since the expressions inside the integrals on both sides of (5.27) are quadratic functions of $\mathbf{b}$, after integrating over $\mathbf{b}$, we obtain that the following explicit equation holds with probability one:

$$\begin{aligned}
&|\boldsymbol{\Sigma}_e^*|^{-N/2}|\boldsymbol{\Sigma}_b^*|^{-1/2}|\boldsymbol{\Sigma}_b^{*-1} + \sum_{j=1}^N \boldsymbol{\rho}_j\boldsymbol{\Sigma}_e^{*-1}\boldsymbol{\rho}_j|^{-1/2}\exp\left\{-\frac{1}{2}\sum_{j=1}^N(\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}^*)^T\boldsymbol{\Sigma}_e^{*-1}(\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}^*)\right.\\
&\left.+\frac{1}{2}\sum_{j=1}^N\left[(\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}^*)^T\boldsymbol{\Sigma}_e^{*-1}\boldsymbol{\rho}_j^T\right](\boldsymbol{\Sigma}_b^{*-1} + \sum_{j=1}^N\boldsymbol{\rho}_j^T\boldsymbol{\Sigma}_e^{*-1}\boldsymbol{\rho}_j)^{-1}\sum_{j=1}^N\left[\boldsymbol{\rho}_j\boldsymbol{\Sigma}_e^{*-1}(\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}^*)\right]\right\}\\
=&|\boldsymbol{\Sigma}_{0e}|^{-N/2}|\boldsymbol{\Sigma}_{0b}|^{-1/2}|\boldsymbol{\Sigma}_{0b}^{-1} + \sum_{j=1}^N \boldsymbol{\rho}_j\boldsymbol{\Sigma}_{0e}^{-1}\boldsymbol{\rho}_j|^{-1/2}\exp\left\{-\frac{1}{2}\sum_{j=1}^N(\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_{0e}^{-1}(\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)\right.\\
&\left.+\frac{1}{2}\sum_{j=1}^N\left[(\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu})^T\boldsymbol{\Sigma}_{0e}^{-1}\boldsymbol{\rho}_j^T\right](\boldsymbol{\Sigma}_{0b}^{-1} + \sum_{j=1}^N\boldsymbol{\rho}_j^T\boldsymbol{\Sigma}_{0e}^{-1}\boldsymbol{\rho}_j)^{-1}\sum_{j=1}^N\left[\boldsymbol{\rho}_j\boldsymbol{\Sigma}_{0e}^{-1}(\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)\right]\right\}
\end{aligned} \tag{5.28}$$

Let $\mathbf{D} = (\boldsymbol{\Sigma}_b^{-1} + \sum_{j=1}^N \boldsymbol{\rho}_j^T\boldsymbol{\Sigma}_e^{-1}\boldsymbol{\rho}_j)^{-1}$. The quadratic and linear terms of $\mathbf{W}_j$ in the expo-

nential part yield

$$\begin{aligned}&\sum_{j=1}^{N}\mathbf{W}_j^T\boldsymbol{\Sigma}_e^{*-1}\mathbf{W}_j-\left[\sum_{j=1}^{N}\mathbf{W}_j^T\boldsymbol{\Sigma}_e^{*-1}\boldsymbol{\rho}_j^T\right]\mathbf{D}^*\left[\sum_{j=1}^{N}\boldsymbol{\rho}_j\boldsymbol{\Sigma}_e^{*-1}\mathbf{W}_j\right]\\=&\sum_{j=1}^{N}\mathbf{W}_j^T\boldsymbol{\Sigma}_{0e}^{-1}\mathbf{W}_j-\left[\sum_{j=1}^{N}\mathbf{W}_j^T\boldsymbol{\Sigma}_{0e}^{-1}\boldsymbol{\rho}_j^T\right]\mathbf{D}_0\left[\sum_{j=1}^{N}\boldsymbol{\rho}_j\boldsymbol{\Sigma}_{0e}^{-1}\mathbf{W}_j\right],\end{aligned}\tag{5.29}$$

and

$$\begin{aligned}&\sum_{j=1}^{N}(\tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}^*)^T\boldsymbol{\Sigma}_e^{*-1}\mathbf{W}_j-\left[\sum_{j=1}^{N}(\tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}^*)^T\boldsymbol{\Sigma}_e^{*-1}\boldsymbol{\rho}_j^T\right]\mathbf{D}^*\left[\sum_{j=1}^{N}\boldsymbol{\rho}_j\boldsymbol{\Sigma}_e^{*-1}\mathbf{W}_j\right]\\=&\sum_{j=1}^{N}(\tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_{0e}^{-1}\mathbf{W}_j-\left[\sum_{j=1}^{N}(\tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_{0e}^{-1}\boldsymbol{\rho}_j^T\right]\mathbf{D}_0\left[\sum_{j=1}^{N}\boldsymbol{\rho}_j\boldsymbol{\Sigma}_{0e}^{-1}\mathbf{W}_j\right]\end{aligned}\tag{5.30}$$

Let $\mathbf{W}_1\neq 0$ and $\mathbf{W}_j=0$ for all $j=2,\ldots,N$, (5.29) and (5.30) become

$$\mathbf{W}_1^T(\boldsymbol{\Sigma}_e^{*-1}-\boldsymbol{\Sigma}_e^{*-1}\boldsymbol{\rho}_1^T\mathbf{D}^*\boldsymbol{\rho}_1\boldsymbol{\Sigma}_e^{*-1})\mathbf{W}_1=\mathbf{W}_1^T(\boldsymbol{\Sigma}_{0e}^{-1}-\boldsymbol{\Sigma}_{0e}^{-1}\boldsymbol{\rho}_1^T\mathbf{D}_0\boldsymbol{\rho}_1\boldsymbol{\Sigma}_{0e}^{-1})\mathbf{W}_1\tag{5.31}$$

$$(\tilde{\boldsymbol{\rho}}_1^T\boldsymbol{\mu}^*)^T(\boldsymbol{\Sigma}_e^{*-1}-\boldsymbol{\Sigma}_e^{*-1}\boldsymbol{\rho}_1^T\mathbf{D}^*\boldsymbol{\rho}_1\boldsymbol{\Sigma}_e^{*-1})\mathbf{W}_1=(\tilde{\boldsymbol{\rho}}_1^T\boldsymbol{\mu}_0)^T(\boldsymbol{\Sigma}_{0e}^{-1}-\boldsymbol{\Sigma}_{0e}^{-1}\boldsymbol{\rho}_1^T\mathbf{D}_0\boldsymbol{\rho}_1\boldsymbol{\Sigma}_{0e}^{-1})\mathbf{W}_1\tag{5.32}$$

Since $\mathbf{W}_1$ is arbitrary, (5.31) yields

$$\boldsymbol{\Sigma}_e^{*-1}-\boldsymbol{\Sigma}_e^{*-1}\boldsymbol{\rho}_1^T\mathbf{D}^*\boldsymbol{\rho}_1\boldsymbol{\Sigma}_e^{*-1}=\boldsymbol{\Sigma}_{0e}^{-1}-\boldsymbol{\Sigma}_{0e}^{-1}\boldsymbol{\rho}_1^T\mathbf{D}_0\boldsymbol{\rho}_1\boldsymbol{\Sigma}_{0e}^{-1}.\tag{5.33}$$

Combine this with (5.32), and use full rank assumption of $\tilde{\boldsymbol{\rho}}_j$ in (A.6), we obtain that $\boldsymbol{\mu}^*=\boldsymbol{\mu}_0$.

In order to prove $\boldsymbol{\Sigma}_e^*=\boldsymbol{\Sigma}_{0e}$ and $\boldsymbol{\Sigma}_b^*=\boldsymbol{\Sigma}_{0b}$, we reexamine equality (5.29). Let $\mathbf{W}=(\mathbf{W}_1^T,\mathbf{W}_2^T,\ldots,\mathbf{W}_N^T)^T$, $\boldsymbol{\Sigma}_E^{*-1}=\mathbf{I}_N\otimes\boldsymbol{\Sigma}_e^{*-1}$ and $\boldsymbol{\Sigma}_{0E}^{-1}=\mathbf{I}_N\otimes\boldsymbol{\Sigma}_{0e}^{-1}$. Follow the assumption (A.4) and let $\mathbf{R}=(\boldsymbol{\rho}_1,\ldots,\boldsymbol{\rho}_p)$. Thus equality (5.29) can be rewritten as

$$\mathbf{W}^T\boldsymbol{\Sigma}_E^{*-1}\mathbf{W}-\mathbf{W}^T\boldsymbol{\Sigma}_E^{*-1}\mathbf{R}^T\mathbf{D}^*\mathbf{R}\boldsymbol{\Sigma}_E^{*-1}\mathbf{W}=\mathbf{W}^T\boldsymbol{\Sigma}_{0E}^{-1}\mathbf{W}-\mathbf{W}^T\boldsymbol{\Sigma}_{0E}^{-1}\mathbf{R}^T\mathbf{D}_0\mathbf{R}\boldsymbol{\Sigma}_{0E}^{-1}\mathbf{W}.$$

Since both sides of the equality are quadratic forms of $\mathbf{W}$ and $\mathbf{W}$ is arbitrary in the space $\mathbb{R}^{pN}$, the equality can be reduces to the linear form

$$\boldsymbol{\Sigma}_E^{*-1}\mathbf{W}-\boldsymbol{\Sigma}_E^{*-1}\mathbf{R}^T\mathbf{D}^*\mathbf{R}\boldsymbol{\Sigma}_E^{*-1}\mathbf{W}=\boldsymbol{\Sigma}_{0E}^{-1}\mathbf{W}-\boldsymbol{\Sigma}_{0E}^{-1}\mathbf{R}^T\mathbf{D}_0\mathbf{R}\boldsymbol{\Sigma}_{0E}^{-1}\mathbf{W}.\tag{5.34}$$

Define two subspaces

$$\mathrm{Ker}^* = \{\mathbf{W} \in \mathbb{R}^{pN} : \mathbf{R}\boldsymbol{\Sigma}_E^{*-1}\mathbf{W} = \mathbf{0}\},$$

$$\mathrm{Ker}_0 = \{\mathbf{W} \in \mathbb{R}^{pN} : \mathbf{R}\boldsymbol{\Sigma}_{0E}^{-1}\mathbf{W} = \mathbf{0}\},$$

and the associated degrees of freedom are $\dim(\mathrm{Ker}^*) = \dim(\mathrm{Ker}_0) = pN - d$. According to assumption (A.4) that $N > 2d$, since $p \geq 1$, we have $pN > 2d$, thus by the inequality

$$\dim(\mathrm{Ker}^*) + \dim(\mathrm{Ker}_0) - \dim(\mathrm{Ker}^* \cap \mathrm{Ker}_0) = \dim(\mathrm{Ker}^* + \mathrm{Ker}_0) \leq pN,$$

we obtain that

$$\dim(\mathrm{Ker}^* \cap \mathrm{Ker}_0) \geq 2(pN - d) - pN = pN - 2d.$$

For any $\mathbf{W} \in \mathrm{Ker}^* \cap \mathrm{Ker}_0$, equation (5.34) reduces to

$$(\boldsymbol{\Sigma}_E^{*-1} - \boldsymbol{\Sigma}_{0E}^{-1})\mathbf{W} = \mathbf{0}.$$

Let $\mathbf{A} = \boldsymbol{\Sigma}_E^{*-1} - \boldsymbol{\Sigma}_{0E}^{-1}$ and $\mathbf{A}_0 = \boldsymbol{\Sigma}_e^{*-1} - \boldsymbol{\Sigma}_{0e}^{-1}$. Our goal is to prove that $\mathbf{A}_0 = \mathbf{0}$. Denote the kernel of operator $\mathbf{A}$ as

$$\mathrm{Ker}(\mathbf{A}) = \{\mathbf{W} \in \mathbb{R}^{pN} : \mathbf{A}\mathbf{W} = \mathbf{0}\}.$$

Since $\mathrm{Ker}^* \cap \mathrm{Ker}_0 \subset \mathrm{Ker}(\mathbf{A})$, we have

$$\dim(\mathrm{Ker}(\mathbf{A})) \geq \dim(\mathrm{Ker}^* \cap \mathrm{Ker}_0) \geq pN - 2d. \tag{5.35}$$

Denote the ranges of the operator $\mathbf{A}$ and $\mathbf{A}_0$ as

$$\mathrm{Ran}(\mathbf{A}) = \{\mathbf{A}\mathbf{W} : \mathbf{W} \in \mathbb{R}^{pN}\}, \quad \text{and}$$

$$\mathrm{Ran}(\mathbf{A}_0) = \{\mathbf{A}_0\mathbf{W}_0 : \mathbf{W}_0 \in \mathbb{R}^{p}\},$$

respectively. Since $\mathbf{A} = \mathbf{I}_N \otimes \mathbf{A}_0$, it is easy to verify that

$$\dim(\mathrm{Ran}(\mathbf{A})) = N\dim(\mathrm{Ran}(\mathbf{A}_0)). \tag{5.36}$$

In addition, since for operator $\mathbf{A}$ we have $\dim(\mathrm{Ker}(\mathbf{A})) + \dim(\mathrm{Ran}(\mathbf{A})) = pN$ (citation to add), it follows from (5.35) and (5.36) that

$$N\dim(\mathrm{Ran}(\mathbf{A}_0)) = \dim(\mathrm{Ran}(\mathbf{A})) \leq pN - (pN - 2d) = 2d.$$

According to assumption (A.4) that $N > 2d$, we conclude that $\dim(\mathrm{Ran}(\mathbf{A}_0)) = 0$. Therefore, $\mathbf{A}_0 = \mathbf{0}$, i.e., $\boldsymbol{\Sigma}_e^* = \boldsymbol{\Sigma}_{0e}$. Moreover, according to assumption (A.6) that $\mathbf{R}\mathbf{R}^T$ is full rank, it is easy to verify that $\mathbf{D}^* = \mathbf{D}_0$, which directly yields $\boldsymbol{\Sigma}_b^* = \boldsymbol{\Sigma}_{0b}$.

Next, let $\Delta = 0$ in equation (5.27), since all the density parts related to $\mathbf{W}_i$ cancel out by the above conclusions, we obtain

$$E_{\mathbf{b}}\left[\exp\left\{-\int_0^V e^{\boldsymbol{\beta}^{*T}(\boldsymbol{\rho}(t)^T\mathbf{b})+\boldsymbol{\eta}^{*T}\mathbf{z}}d\Lambda^*(t)\right\}\right] = E_{\mathbf{b}}\left[\exp\left\{-\int_0^V e^{\boldsymbol{\beta}_0^T(\boldsymbol{\rho}(t)^T\mathbf{b})+\boldsymbol{\eta}_0^T\mathbf{z}}d\Lambda_0(t)\right\}\right],$$

where $\mathbf{b} \sim N_d(\tilde{\mu}_b, \tilde{\boldsymbol{\Sigma}}_b)$ with $\tilde{\mu}_b = (\boldsymbol{\Sigma}_{0b}^{-1} + \sum_{j=1}^N \boldsymbol{\rho}_j^T\boldsymbol{\Sigma}_{0e}^{-1}\boldsymbol{\rho}_j)^{-1}\left[\sum_{j=1}^N \boldsymbol{\rho}_j\boldsymbol{\Sigma}_{0e}^{-1}(\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)\right]$ and $\tilde{\boldsymbol{\Sigma}}_b = (\boldsymbol{\Sigma}_{0b}^{-1} + \sum_{j=1}^N \boldsymbol{\rho}_j^T\boldsymbol{\Sigma}_{0e}^{-1}\boldsymbol{\rho}_j)^{-1}$.

Treat $\sum_{j=1}^N \boldsymbol{\rho}_j^T\boldsymbol{\Sigma}_{0e}^{-1}\boldsymbol{\rho}_j$ and $\sum_{j=1}^N \boldsymbol{\rho}_j^T\boldsymbol{\Sigma}_{0e}^{-1}\tilde{\boldsymbol{\rho}}_j$ as fixed and $\sum_{j=1}^N \boldsymbol{\rho}_j^T\boldsymbol{\Sigma}_{0e}^{-1}\mathbf{W}_j$ as parameter, $\mathbf{b}$ is complete sufficient statistics for $\sum_{j=1}^N \boldsymbol{\rho}_j^T\boldsymbol{\Sigma}_{0e}^{-1}\mathbf{W}_j$. Thus

$$\exp\left\{-\int_0^V e^{\boldsymbol{\beta}^{*T}(\boldsymbol{\rho}(t)^T\mathbf{b})+\boldsymbol{\eta}^{*T}\mathbf{z}}d\Lambda^*(t)\right\} = \exp\left\{-\int_0^V e^{\boldsymbol{\beta}_0^T(\boldsymbol{\rho}(t)^T\mathbf{b})+\boldsymbol{\eta}_0^T\mathbf{z}}d\Lambda_0(t)\right\}.$$

Equivalently,

$$\lambda^*(t)e^{\boldsymbol{\beta}^{*T}(\boldsymbol{\rho}(t)^T\mathbf{b})+\boldsymbol{\eta}^{*T}\mathbf{z}} = \lambda_0(t)e^{\boldsymbol{\beta}_0^T(\boldsymbol{\rho}(t)^T\mathbf{b})+\boldsymbol{\eta}_0^T\mathbf{z}}.$$

Take logarithm on both sides of the equation and rearrange the terms, we obtain that there exits some function of time $\tilde{g}(t)$ such that

$$\tilde{g}(t) = \log\lambda^*(t) - \log\lambda_0(t) = (\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*)^T(\boldsymbol{\rho}(t)^T\mathbf{b}) + (\boldsymbol{\eta}_0 - \boldsymbol{\eta}^*)^T\mathbf{Z},$$

for any $\mathbf{b}$. According to assumption (A.7) and (A.11) we conclude that $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0$, $\boldsymbol{\eta}^* = \boldsymbol{\eta}_0$ and $\Lambda^* = \Lambda_0$. $\square$

Proof of Theorem 2

Let $\boldsymbol{\psi} = (\boldsymbol{\theta}, \Lambda) \in \Psi = \{(\boldsymbol{\theta}, \Lambda) : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| + \sup_{t\in[0,\tau]} |\Lambda(t) - \Lambda_0(t)| \leq \delta\}$ for a fixed small δ. Note that Ψ is a convex set. Define a set

$$\mathcal{H} = \{(\mathbf{h}_1, h_2) : \|\mathbf{h}_1\| \leq 1, \|h_2\|_V \leq 1\},$$

where $\|h_2\|_V$ is the total variation of h_2 in $[0, \tau]$ defined as

$$\sup_{0=t_0\leq t_1\leq\cdots\leq t_N=\tau} \sum_{j=1}^{N} |h_2(t_j) - h_2(t_{j-1})|.$$

Recall that $l(\mathbf{O}; \boldsymbol{\theta}, \Lambda)$ is the log likelihood of a single subject. The associated Fréchet derivative is given by

$$f_{\psi,h} = l_\theta(\boldsymbol{\theta}, \Lambda)^T \mathbf{h}_1 + l_\Lambda(\boldsymbol{\theta}, \Lambda)[h_2], \quad (\boldsymbol{\theta}, \Lambda) \in \Psi, (\mathbf{h}_1, h_2) \in \mathcal{H}, \tag{5.37}$$

where $l_\theta(\boldsymbol{\theta}, \Lambda) = \partial l(\mathbf{O}; \boldsymbol{\theta}, \Lambda)/\partial\boldsymbol{\theta}$, and $l_\Lambda(\boldsymbol{\theta}, \Lambda)$ is the derivative of $l(\mathbf{O}; \boldsymbol{\theta}, \Lambda_\epsilon)$ with respect to ϵ at $\epsilon = 0$, where $\Lambda_\epsilon(t) = \int_0^t (1 + \epsilon h_2(s)) d\Lambda_0(s)$. Define empirical processes

$$S_n(\boldsymbol{\psi})(\mathbf{h}_1, h_2) = \mathbf{P}_n f_{\psi,h},$$

$$S(\boldsymbol{\psi})(\mathbf{h}_1, h_2) = \mathbf{P} f_{\psi,h}.$$

By the definition of $f_{\psi,h}$, S_n and S are both maps from Ψ to $l^\infty(\mathcal{H})$ (i.e., the collection of all bounded functions from $\mathcal{H}$ to $\mathbb{R}$).

The asymptotic normality of $\widehat{\boldsymbol{\psi}} = (\widehat{\boldsymbol{\theta}}, \widehat{\Lambda})$ is established by checking the four conditions of the following theorem (Van Der Vaart and Wellner, 1996, Theorem 3.3.1):

Theorem Let Ψ be a subset of a Banach space that contains the true parameter $\boldsymbol{\psi}_0$. Let S be a fixed map and S_n be a series of random maps, both of which map from Ψ to a Banach space such that

(a) $\sqrt{n}(S_n - S)(\widehat{\boldsymbol{\psi}}_n) - \sqrt{n}(S_n - S)(\boldsymbol{\psi}_0) = o_p^*(1 + \sqrt{n}\|\widehat{\boldsymbol{\psi}}_n - \boldsymbol{\psi}_0\|)$;

(b) The sequence $\sqrt{n}(S_n - S)(\boldsymbol{\psi}_0)$ converges in distribution to a tight random element $\mathbf{Z}_0$;

(c) The function $\boldsymbol{\psi} \to S(\boldsymbol{\psi})$ is Fréchet differentiable at $\boldsymbol{\psi}_0$ with a continuously invertible derivative ∇S_{ψ_0} on its range;

(d) $S(\psi_0) = 0$. $\widehat{\boldsymbol{\psi}}_n$ satisfies $S_n(\widehat{\boldsymbol{\psi}}_n) = o_p^*(n^{-1/2})$ and $\widehat{\boldsymbol{\psi}}_n$ converges in outer probability to $\boldsymbol{\psi}_0$.

Then, $\sqrt{n}(\widehat{\boldsymbol{\psi}}_n - \boldsymbol{\psi}_0) \Rightarrow -\nabla S^{-1}_{\boldsymbol{\psi}_0} \mathbf{Z}_0$.

We first check condition (a). According to Van Der Vaart and Wellner (1996), when the observations are independent and identical (iid), the theorem can be applied with $S_n(\boldsymbol{\psi})\mathbf{h} = \mathbf{P}_n f_{\psi,h}$ and $S(\boldsymbol{\psi})\mathbf{h} = \mathbf{P} f_{\psi,h}$, where $f_{\psi,h}$ is a measurable function indexed by Ψ and $\mathcal{H}$. In this case, for given $\boldsymbol{\psi} \in \Psi$,

$$\sqrt{n}(S_n - S)(\boldsymbol{\psi})\mathbf{h} = \sqrt{n}(\mathbf{P}_n - \mathbf{P}) f_{\psi,h} \stackrel{\Delta}{=} \{G_n f_{\psi,h} : \mathbf{h} = (\mathbf{h}_1, h_2) \in \mathcal{H}\}$$

is an empirical process indexed by the class $\{f_{\psi,h} : \mathbf{h} \in \mathcal{H}\}$. Thus, for the iid case, condition (a) in the above theorem becomes

$$\|G_n(f_{\widehat{\psi}_n,h} - f_{\psi_0,h})\| = o_p^*(1 + \sqrt{n}\|\widehat{\boldsymbol{\psi}}_n - \boldsymbol{\psi}_0\|). \tag{5.38}$$

Therefore, using the measurable function $f_{\psi,h}$ defined in (5.37), we only need to verify (5.38) instead of condition (a).

Lemma 3.3.5 in Van Der Varrt and Wellner (1996) provides sufficient conditions for (5.38). It claims that (5.38) holds if $\widehat{\boldsymbol{\psi}}_n$ converges in outer probability to $\boldsymbol{\psi}_0$, and the following tow conditions are satisfied

(a.1) $\{f_{\psi,h} - f_{\psi_0,h} : \boldsymbol{\psi} \in \Psi, \mathbf{h} \in \mathcal{H}\}$ is a P-Donsker,

(a.2) $\sup_{h\in\mathcal{H}} P(f_{\widehat{\psi}_n,h} - f_{\psi_0,h})^2 \to 0$ as $\widehat{\boldsymbol{\psi}} \to \psi_0$.

Since the convergence of $\widehat{\boldsymbol{\psi}}_n$ to $\boldsymbol{\psi}_0$ is justified by Theorem 1, and both (a.1) and (a.2) are verified in the Appendix A.2 of Zeng and Cai (2005), equation (5.38) holds accordingly. Thus condition (a) is satisfied.

For condition (b), since it is easy to verify that the class $\{f_{\psi_0,h} : \mathbf{h} \in \mathcal{H}\}$ is P-Donsker (see (Zeng and Cai, 2005), Appendix A.2, for more details), there exists a tight Gaussian process $\mathbf{Z}_0 \in l^\infty(\mathcal{H})$ such that the empirical process $\sqrt{n}(S_n - S)(\boldsymbol{\psi}_0) = \sqrt{n}(\mathbf{P}_n - \mathbf{P})f_{\psi_0,h}$ converges to $\mathbf{Z}_0$ in distribution.

For condition (d), $S(\boldsymbol{\psi}_0) = 0$ and $S(\widehat{\boldsymbol{\psi}}_n) = 0$ because $(\boldsymbol{\theta}_0, \Lambda_0)$ maximizes $\mathbf{P}l(\mathbf{O}; \boldsymbol{\theta}, \Lambda)$, and $(\widehat{\boldsymbol{\theta}}, \widehat{\Lambda})$ maximizes $\mathbf{P}_n(\mathbf{O}; \boldsymbol{\theta}, \Lambda)$. Thus with the consistency result of theorem 1, condition (d) is also satisfied.

It remains to verify condition (c). By the definition of Fréchet differentiable, $S(\boldsymbol{\psi})$ is Fréchet differentiable at $\boldsymbol{\psi}_0$ if there exists a linear operator $A_{\boldsymbol{\psi}_0} : \Psi \mapsto l^\infty(\mathcal{H})$ such that

$$\begin{aligned} &S(\boldsymbol{\psi})(\mathbf{h}_1, h_2) - S(\boldsymbol{\psi}_0)(\mathbf{h}_1, h_2) \\ &= A_{\boldsymbol{\psi}_0}(\boldsymbol{\psi} - \boldsymbol{\psi}_0)(\mathbf{h}_1, h_2) + o(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| + \sup_{t\in[0,\tau]} |\Lambda(t) - \Lambda_0(t)|)(\|\mathbf{h}_1\| + \|h_2\|_V) \end{aligned} \tag{5.39}$$

for any $(\mathbf{h}_1, h_2) \in \mathcal{H}$. The existence of $A_{\boldsymbol{\psi}_0}$ is proved below.

Let $(\mathbf{h}_1^e, \mathbf{h}_1^b, \mathbf{h}_1^\mu, \mathbf{h}_1^\beta, \mathbf{h}_1^\eta)$ be the components of $\mathbf{h}_1$ corresponding to each of the parameters $(\boldsymbol{\Sigma}_e, \boldsymbol{\Sigma}_b, \boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\eta})$. Thus $f_{\psi,h} = l_\theta(\boldsymbol{\theta}, \Lambda)^T\mathbf{h}_1 + l_\Lambda(\boldsymbol{\theta}, \Lambda)[h_2]$ can be written out explicitly in the following expression

$$g_1(\mathbf{O}; \boldsymbol{\theta}, \Lambda)^T\mathbf{h}_1 - \int_0^V g_2(t, \mathbf{O}; \boldsymbol{\theta}, \Lambda)^T\mathbf{h}_1 d\Lambda(t) + \Delta h_2(V) - \int_0^V g_3(t, \mathbf{O}; \boldsymbol{\theta}, \Lambda)h_2(t)d\Lambda(t), \tag{5.40}$$

where

$$
\begin{aligned}
g_1(\mathbf{O};\boldsymbol{\theta},\Lambda)^T\mathbf{h}_1 = &\left\{\int_{\mathbf{b}} G(\mathbf{b},\mathbf{O};\boldsymbol{\theta},\Lambda)d\mathbf{b}\right\}^{-1} \\
&\times \int_{\mathbf{b}} G(\mathbf{b},\mathbf{O};\boldsymbol{\theta},\Lambda)\left[\frac{1}{2}\mathbf{b}^T\Sigma_b^{-1}\mathcal{D}_b\Sigma_b^{-1}\mathbf{b} - \frac{1}{2}tr(\boldsymbol{\Sigma}_b^{-1}\mathcal{D}_b)\right. \\
&+ \frac{1}{2}\sum_{j=1}^{N}(\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu} - \boldsymbol{\rho}_j^T\mathbf{b})^T\boldsymbol{\Sigma}_e^{-1}\mathcal{D}_e\boldsymbol{\Sigma}_e^{-1}(\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu} - \boldsymbol{\rho}_j^T\mathbf{b}) \\
&- \frac{N}{2}tr(\boldsymbol{\Sigma}_e^{-1}\mathcal{D}_e) \\
&\left.+ \sum_{j=1}^{N}(\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu} - \boldsymbol{\rho}_j^T\mathbf{b})^T\Sigma_e^{-1}\tilde{\boldsymbol{\rho}}_j^T\mathbf{h}_1^{\mu} + \Delta\{\mathbf{h}_1^{\beta T}(\boldsymbol{\rho}(V)^T\mathbf{b}) + \mathbf{h}_1^{\eta T}\mathbf{Z}\}\right] d\mathbf{b},
\end{aligned}
$$

$$
\begin{aligned}
g_2(t,\mathbf{O};\boldsymbol{\theta},\Lambda)^T\mathbf{h}_1 = &\left\{\int_{\mathbf{b}} G(\mathbf{b},\mathbf{O};\boldsymbol{\theta},\Lambda)d\mathbf{b}\right\}^{-1} \\
&\times \int_{\mathbf{b}} G(\mathbf{b},\mathbf{O};\boldsymbol{\theta},\Lambda) \times \left[\{\mathbf{h}_1^{\beta T}(\boldsymbol{\rho}(t)^T\mathbf{b}) + \mathbf{h}_1^{\eta T}\mathbf{Z}\}e^{\boldsymbol{\beta}^T(\boldsymbol{\rho}(t)^T\mathbf{b})+\boldsymbol{\eta}^T\mathbf{z}}\right] d\mathbf{b},
\end{aligned}
$$

$$
g_3(t,\mathbf{O};\boldsymbol{\theta},\Lambda) = \left\{\int_{\mathbf{b}} G(\mathbf{b},\mathbf{O};\boldsymbol{\theta},\Lambda)d\mathbf{b}\right\}^{-1} \int_{\mathbf{b}} G(\mathbf{b},\mathbf{O};\boldsymbol{\theta},\Lambda) \times e^{\boldsymbol{\beta}^T(\boldsymbol{\rho}(t)^T\mathbf{b})+\boldsymbol{\eta}^T\mathbf{z}}d\mathbf{b}.
$$

Here, $\mathcal{D}_b$ and $\mathcal{D}_e$ are symmetric matrices such that $\mathrm{Vec}(\mathcal{D}_b) = \mathbf{h}_1^b$ and $\mathrm{Vec}(\mathcal{D}_e) = \mathbf{h}_1^e$, respectively.

For $j = 1, 2, 3$, we denote $\nabla_\theta g_j$ to be the derivative of g_j with respect to $\boldsymbol{\theta}$, and denote $\nabla_\Lambda g_j[\delta\Lambda]$ to be the derivative of g_j with respect to Λ along the path $\Lambda + \epsilon\delta\Lambda$. It is easy to check that for $j = 1, 2, 3$, the derivative of g_j with respect to Λ along the path $\Lambda + \epsilon\delta\Lambda$ can be expressed as $\nabla_\Lambda g_j[\delta\Lambda] = \int_0^t g_{j+3}(s,\mathbf{O};\boldsymbol{\theta},\Lambda)d\delta\Lambda(s)$ for some $g_k(s,\mathbf{O};\boldsymbol{\theta},\Lambda)$, $k = 4, 5, 6$. Thus, by the mean value theorem, for any $(\boldsymbol{\theta},\Lambda,\mathbf{h}_1,h_2)$ in $\Psi\times\mathcal{H}$,

$$
\begin{aligned}
& l_\theta(\boldsymbol{\theta},\Lambda)^T\mathbf{h}_1 + l_\Lambda(\boldsymbol{\theta},\Lambda)[h_2] - l_\theta(\boldsymbol{\theta}_0,\Lambda_0)^T\mathbf{h}_1 - l_\Lambda(\boldsymbol{\theta}_0,\Lambda_0)[h_2] \\
=&(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T\left\{\nabla_\theta g_1(\mathbf{O};\tilde{\boldsymbol{\theta}},\tilde{\Lambda}) - \int_0^V \nabla_\theta g_2(t,\mathbf{O};\tilde{\boldsymbol{\theta}},\tilde{\Lambda})d\Lambda_0(t)\right\}\mathbf{h}_1 \\
&+\mathbf{h}_1^T\int_0^\tau I(t\le V)\Big\{g_4(t,\mathbf{O};\tilde{\boldsymbol{\theta}},\tilde{\Lambda}) - g_2(t,\mathbf{O};\tilde{\boldsymbol{\theta}},\tilde{\Lambda}) \\
&-\ g_5(t,\mathbf{O};\tilde{\boldsymbol{\theta}},\tilde{\Lambda})\int_t^V d\Lambda_0(s)\Big\}d(\Lambda-\Lambda_0)(t) \\
&-(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T\int_0^\tau I(t\le V)\nabla_\theta g_3(t,\mathbf{O};\tilde{\boldsymbol{\theta}},\tilde{\Lambda})h_2(t)d\Lambda_0(t) \\
&-\int_0^\tau\Big\{I(t\le V)g_6(t,\mathbf{O};\tilde{\boldsymbol{\theta}},\tilde{\Lambda})\int_t^V h_2(s)d\Lambda_0(s) \\
&+\ I(t\le V)g_3(t,\mathbf{O};\tilde{\boldsymbol{\theta}},\tilde{\Lambda})h_2(t)\Big\}d(\Lambda-\Lambda_0)(t),
\end{aligned}
$$

where $(\tilde{\boldsymbol{\theta}},\tilde{\Lambda}) = \xi^*(\boldsymbol{\theta},\Lambda) + (1-\xi^*)(\boldsymbol{\theta}_0,\Lambda_0)$ for some $\xi^*\in[0,1]$. Thus by the definition of $S(\boldsymbol{\psi})(\mathbf{h}_1,h_2)$, it follows that

$$
\begin{aligned}
& S(\boldsymbol{\psi})(\mathbf{h}_1,h_2) - S(\boldsymbol{\psi}_0)(\mathbf{h}_1,h_2) \\
=&(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T\mathbf{P}\left\{\nabla_\theta g_1(\mathbf{O};\tilde{\boldsymbol{\theta}},\tilde{\Lambda}) - \int_0^V \nabla_\theta g_2(t,\mathbf{O};\tilde{\boldsymbol{\theta}},\tilde{\Lambda})d\Lambda_0(t)\right\}\mathbf{h}_1 \\
&+\mathbf{h}_1^T\int_0^\tau \mathbf{P}\Big[I(t\le V)\Big\{g_4(t,\mathbf{O};\tilde{\boldsymbol{\theta}},\tilde{\Lambda}) - g_2(t,\mathbf{O};\tilde{\boldsymbol{\theta}},\tilde{\Lambda}) \\
&-\ g_5(t,\mathbf{O};\tilde{\boldsymbol{\theta}},\tilde{\Lambda})\int_t^V d\Lambda_0(s)\Big\}\Big]d(\Lambda-\Lambda_0)(t) \\
&-(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T\int_0^\tau \mathbf{P}I(t\le V)\nabla_\theta g_3(t,\mathbf{O};\tilde{\boldsymbol{\theta}},\tilde{\Lambda})h_2(t)d\Lambda_0(t) \\
&-\int_0^\tau \mathbf{P}\Big\{I(t\le V)g_6(t,\mathbf{O};\tilde{\boldsymbol{\theta}},\tilde{\Lambda})\int_t^V h_2(s)d\Lambda_0(s) \\
&+\ I(t\le V)g_3(t,\mathbf{O};\tilde{\boldsymbol{\theta}},\tilde{\Lambda})h_2(t)\Big\}d(\Lambda-\Lambda_0)(t).
\end{aligned}
$$

Follow the above equation, define $A_{\boldsymbol{\psi}_0}(\boldsymbol{\psi}-\boldsymbol{\psi}_0)$ by the following expression

$$
\begin{aligned}
&A_{\boldsymbol{\psi}_0}(\boldsymbol{\psi}-\boldsymbol{\psi}_0)(\mathbf{h}_1,h_2)\\
=&(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T\mathbf{P}\left\{\nabla_\theta g_1(\mathbf{O};\boldsymbol{\theta}_0,\Lambda_0)-\int_0^V \nabla_\theta g_2(t,\mathbf{O};\boldsymbol{\theta}_0,\Lambda_0)d\Lambda_0(t)\right\}\mathbf{h}_1\\
&+\mathbf{h}_1^T\int_0^\tau \mathbf{P}\left[I(t\le V)\left\{g_4(t,\mathbf{O};\boldsymbol{\theta}_0,\Lambda_0)-g_2(t,\mathbf{O};\boldsymbol{\theta}_0,\Lambda_0)\right.\right.\\
&\left.\left.-\ g_5(t,\mathbf{O};\boldsymbol{\theta}_0,\Lambda_0)\int_t^V d\Lambda_0(s)\right\}\right]d(\Lambda-\Lambda_0)(t) \qquad (5.41)\\
&-(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T\int_0^\tau \mathbf{P}I(t\le V)\nabla_\theta g_3(t,\mathbf{O};\boldsymbol{\theta}_0,\Lambda_0)h_2(t)d\Lambda_0(t)\\
&-\int_0^\tau \mathbf{P}\left\{I(t\le V)g_6(t,\mathbf{O};\boldsymbol{\theta}_0,\Lambda_0)\int_t^V h_2(s)d\Lambda_0(s)\right.\\
&\left.+\ I(t\le V)g_3(t,\mathbf{O};\boldsymbol{\theta}_0,\Lambda_0)h_2(t)\right\}d(\Lambda-\Lambda_0)(t).
\end{aligned}
$$

According to the appendix A.3 of Zeng and Cai (2005), for $j=1,\ldots,6$, the following inequalities hold for some constant r_1 and r_2

$$
\begin{aligned}
&\sup_{t\in[0,\tau]}\|g_j(t,\mathbf{O};\tilde{\boldsymbol{\theta}},\tilde{\Lambda})-g_j(t,\mathbf{O};\boldsymbol{\theta}_0,\Lambda_0)\|\\
&\le e^{r_1+r_2\sum_{j=1}^N\|W_j\|}\left\{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|+\sup_{t\in[0,\tau]}|\Lambda(t)-\Lambda_0(t)|\right\}.
\end{aligned}
$$

Using this inequality, it is convenient to check that equation (5.39) holds for $A_{\boldsymbol{\psi}_0}(\boldsymbol{\psi}-\boldsymbol{\psi}_0)(\mathbf{h}_1,h_2)$ defined in (5.41). Therefore, $S(\boldsymbol{\psi}_0)$ is Fréchet differentiable at $\boldsymbol{\psi}_0$, and we can denote

$$\nabla S_{\psi_0}(\boldsymbol{\psi}-\boldsymbol{\psi}_0)(\mathbf{h}_1,h_2)=A_{\boldsymbol{\psi}_0}(\boldsymbol{\psi}-\boldsymbol{\psi}_0)(\mathbf{h}_1,h_2)$$

to be the derivative of $S(\boldsymbol{\psi}_0)$ at $\boldsymbol{\psi}_0$. Note that similar to $S(\boldsymbol{\psi})$, ∇S_{ψ_0} is a function mapping from Ψ to $l^\infty(\mathcal{H})$. It remains to show that ∇S_{ψ_0} is continuously invertible on its range in $l^\infty(\mathcal{H})$.

From the definition of $A_{\boldsymbol{\psi}_0}(\boldsymbol{\psi}-\boldsymbol{\psi}_0)(\mathbf{h}_1,h_2)$ in (5.41), it is clear that ∇S_{ψ_0} can be rewritten

into

$$\begin{aligned}&\nabla S_{\psi_0}(\boldsymbol{\theta}_1-\boldsymbol{\theta}_2,\Lambda_1-\Lambda_2)(\mathbf{h}_1,h_2)\\&=(\boldsymbol{\theta}_1-\boldsymbol{\theta}_2)^T\Omega_1[\mathbf{h}_1,h_2]+\int_0^\tau\Omega_2[\mathbf{h}_1,h_2]d(\Lambda_1-\Lambda_2)(t),\end{aligned}\tag{5.42}$$

where

$$\begin{aligned}\Omega_1[\mathbf{h}_1,h_2]=&\mathbf{h}_1^T\mathbf{P}\left\{\nabla_\theta g_1(\mathbf{O};\boldsymbol{\theta}_0,\Lambda_0)-\int_0^V\nabla_\theta g_2(t,\mathbf{O};\boldsymbol{\theta}_0,\Lambda_0)d\Lambda_0(t)\right\}\\&-\int_0^\tau\mathbf{P}I(t\le V)\nabla_\theta g_3(t,\mathbf{O};\boldsymbol{\theta}_0,\Lambda_0)h_2(t)d\Lambda_0(t),\end{aligned}$$

$$\begin{aligned}\Omega_2[\mathbf{h}_1,h_2]=&\mathbf{h}_1^T\int_0^\tau\mathbf{P}\left[I(t\le V)\left\{g_4(t,\mathbf{O};\boldsymbol{\theta}_0,\Lambda_0)-g_2(t,\mathbf{O};\boldsymbol{\theta}_0,\Lambda_0)\right.\right.\\&\left.\left.-\ g_5(t,\mathbf{O};\boldsymbol{\theta}_0,\Lambda_0)\int_t^V d\Lambda_0(s)\right\}\right]\\&-\int_0^\tau\mathbf{P}\left\{I(t\le V)g_6(t,\mathbf{O};\boldsymbol{\theta}_0,\Lambda_0)\int_t^V h_2(s)d\Lambda_0(s)\right\}\\&-\mathbf{P}\{I(t\le V)g_3(t,\mathbf{O};\boldsymbol{\theta}_0,\Lambda_0)\}h_2(t).\end{aligned}$$

From the above definitions, the operator $\Omega=(\Omega_1,\Omega_2)$ can be taken as a linear operator that maps from $\mathcal{H}\subset\mathbb{R}^d\times BV[0,\tau]$ to itself, where $BV[0,\tau]$ contains all the functions with finite total variation in $[0,\tau]$.

From equation (5.42), the operator $(\boldsymbol{\theta}_1-\boldsymbol{\theta}_2,\Lambda_1-\Lambda_2)$ can be treated as a functional element in $l^\infty(\mathcal{H})$ via the following definition

$$(\boldsymbol{\theta}_1-\boldsymbol{\theta}_2,\Lambda_1-\Lambda_2)(\mathbf{h}_1,h_2)=(\boldsymbol{\theta}_1-\boldsymbol{\theta}_2)^T\mathbf{h}_1+\int_0^\tau h_2(t)d(\Lambda_1-\Lambda_2)(t)$$

for any $(\mathbf{h}_1,h_2)\in\mathbb{R}^d\times BV[0,\tau]$. Thus by (5.42), the function ∇S_{ψ_0} can be regarded as a linear operator from $l^\infty(\mathcal{H})$ to itself, and for any $(\delta\boldsymbol{\theta},\delta\Lambda)\in l^\infty(\mathcal{H})$ the norm of ∇S_{ψ_0} is given by

$$
\begin{aligned}
\|\nabla S_{\psi_0}(\delta\boldsymbol{\theta}, \delta\Lambda)\|_{l^\infty(\mathcal{H})} &= \sup_{(h_1,h_2)\in\mathcal{H}} |\nabla S_{\psi_0}(\delta\boldsymbol{\theta}, \delta\Lambda)(\mathbf{h}_1, h_2)| \\
&= \sup_{(h_1,h_2)\in\mathcal{H}} \left| \delta\boldsymbol{\theta}^T \Omega_1[\mathbf{h}_1, h_2] + \int_0^\tau \Omega_2[\mathbf{h}_1, h_2] d\delta\Lambda(t) \right| \\
&= \sup_{\Omega([h_1,h_2])\in\Omega(\mathcal{H})} |(\delta\boldsymbol{\theta}, \delta\Lambda)\Omega[\mathbf{h}_1, h_2]| \\
&= \|(\delta\boldsymbol{\theta}, \delta\Lambda)\|_{l^\infty(\Omega(\mathcal{H}))}.
\end{aligned}
$$

Thus if we can find some positive constant ε such that $\varepsilon\mathcal{H} \subset \Omega(\mathcal{H})$, it follows that

$$
\|(\delta\boldsymbol{\theta}, \delta\Lambda)\|_{l^\infty(\Omega(\mathcal{H}))} \geq \varepsilon \|(\delta\boldsymbol{\theta}, \delta\Lambda)\|_{l^\infty(\mathcal{H})},
$$

and ∇S_{ψ_0} is hence continuously invertible.

According to Zeng and Cai (2005), in order to show that $\varepsilon\mathcal{H} \subset \Omega(\mathcal{H})$ for some ε (i.e., Ω is invertible), it is sufficient to verify that Ω is one to one. Since Ω is linear, it is left to prove that if $\Omega[\mathbf{h}_1, h_2] = 0$, then $\mathbf{h}_1 = 0$ and $h_2 = 0$.

If $\Omega[\mathbf{h}_1, h_2] = 0$, by choosing $\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 = \tilde{\varepsilon}\mathbf{h}_1$ and $\Lambda_1 - \Lambda_2 = \tilde{\varepsilon}\int h_2 d\Lambda_0$ in (5.42) for a small constant $\tilde{\varepsilon}$, we obtain that $\nabla S_{\psi_0}(\mathbf{h}_1, \int h_2 d\Lambda_0)[\mathbf{h}_1, h_2] = 0$. Note that the left-hand side is the negative information matrix in the submodel with parameter $(\boldsymbol{\theta}_0 + \tilde{\varepsilon}\mathbf{h}_1, \Lambda_0 + \tilde{\varepsilon}\int h_2 d\Lambda_0)$. The corresponding score equation should also equal 0. That is, $l_\theta(\boldsymbol{\theta}_0, \Lambda_0)^T\mathbf{h}_1 + l_\Lambda(\boldsymbol{\theta}_0, \Lambda_0)[h_2] = 0$. Thus, using the notation $(\mathbf{h}_1^e, \mathbf{h}_1^b, \mathbf{h}_1^\mu, \mathbf{h}_1^\beta, \mathbf{h}_1^\eta), \mathcal{D}_b, \mathcal{D}_e$ defined above, together with the expression of (5), the following equation holds with probability one

$$
\begin{aligned}
0 = \int_{\mathbf{b}} & G(\mathbf{b}, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0) \\
& \Bigg[\frac{1}{2}\mathbf{b}^T \boldsymbol{\Sigma}_{0b}^{-1} \mathcal{D}_b \boldsymbol{\Sigma}_{0b}^{-1} \mathbf{b} - \frac{1}{2} tr(\boldsymbol{\Sigma}_{0b}^{-1} \mathcal{D}_b) \\
& \quad \frac{1}{2} \sum_{j=1}^{N} (\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T \boldsymbol{\mu}_0 - \boldsymbol{\rho}_j^T \mathbf{b})^T \boldsymbol{\Sigma}_{0e}^{-1} \mathcal{D}_e \boldsymbol{\Sigma}_{0e}^{-1} (\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T \boldsymbol{\mu}_0 - \boldsymbol{\rho}_j^T \mathbf{b}) - \frac{N}{2} tr(\boldsymbol{\Sigma}_{0e}^{-1} \mathcal{D}_e) \\
& \quad + \sum_{j=1}^{N} (\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T \boldsymbol{\mu}_0 - \boldsymbol{\rho}_j^T \mathbf{b})^T \boldsymbol{\Sigma}_{0e}^{-1} \tilde{\boldsymbol{\rho}}_j^T \mathbf{h}_1^{\mu} + \Delta \{ (\boldsymbol{\rho}(V)^T \mathbf{b})^T \mathbf{h}_1^{\beta} + \mathbf{Z}^T \mathbf{h}_1^{\eta} \} \\
& \quad - \int_0^V \{ (\boldsymbol{\rho}(t)^T \mathbf{b})^T \mathbf{h}_1^{\beta} + \mathbf{Z}^T \mathbf{h}_1^{\eta} \} e^{\boldsymbol{\beta}_0^T (\boldsymbol{\rho}(t)^T \mathbf{b}) + \boldsymbol{\eta}_0^T \mathbf{z}} d\Lambda_0(t) \Bigg] d\mathbf{b} \\
& + \int_{\mathbf{b}} G(\mathbf{b}, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0) \left[\Delta h_2(V) - \int_0^V h_2(t) e^{\boldsymbol{\beta}_0^T (\boldsymbol{\rho}(t)^T \mathbf{b}) + \boldsymbol{\eta}_0^T \mathbf{z}} d\Lambda_0 \right] d\mathbf{b}.
\end{aligned}
\tag{5.43}
$$

Using the same argument as for equation (5.31) in the proof for consistency, we obtain that (5.43) holds for all $\{(V, \Delta) : \Delta = 0, V \in [0, \tau]\}$. Let $\Delta = 0$ and $V = 0$, then (5.43) becomes

$$
\begin{aligned}
0 = \int_{\mathbf{b}} & G_0(\mathbf{b}, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0) \\
& \Bigg\{ \frac{1}{2}\mathbf{b}^T \boldsymbol{\Sigma}_{0b}^{-1} \mathcal{D}_b \boldsymbol{\Sigma}_{0b}^{-1} \mathbf{b} - \frac{1}{2} tr(\boldsymbol{\Sigma}_{0b}^{-1} \mathcal{D}_b) \\
& \quad \frac{1}{2} \sum_{j=1}^{N} (\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T \boldsymbol{\mu}_0 - \boldsymbol{\rho}_j^T \mathbf{b})^T \boldsymbol{\Sigma}_{0e}^{-1} \mathcal{D}_e \boldsymbol{\Sigma}_{0e}^{-1} (\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T \boldsymbol{\mu}_0 - \boldsymbol{\rho}_j^T \mathbf{b}) - \frac{N}{2} tr(\boldsymbol{\Sigma}_{0e}^{-1} \mathcal{D}_e) \\
& \quad + \sum_{j=1}^{N} (\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T \boldsymbol{\mu}_0 - \boldsymbol{\rho}_j^T \mathbf{b})^T \boldsymbol{\Sigma}_{0e}^{-1} \tilde{\boldsymbol{\rho}}_j^T \mathbf{h}_1^{\mu} + \Delta \{ (\boldsymbol{\rho}(V)^T \mathbf{b})^T \mathbf{h}_1^{\beta} + \mathbf{Z}^T \mathbf{h}_1^{\eta} \} \Bigg\} d\mathbf{b},
\end{aligned}
\tag{5.44}
$$

where

$$
\begin{aligned}
G_0(\mathbf{b},\mathbf{O};\boldsymbol{\theta}_0,\Lambda_0) =& (2\pi)^{(pN+d)/2}|\boldsymbol{\Sigma}_{0e}|^{-N/2}|\boldsymbol{\Sigma}_{0b}|^{-1/2} \\
& \times \exp\left\{-\frac{1}{2}\mathbf{b}^T(\boldsymbol{\Sigma}_{0b}^{-1}+\sum_{j=1}^{N}\boldsymbol{\rho}_j\boldsymbol{\Sigma}_{0e}^{-1}\boldsymbol{\rho}_j)\mathbf{b}+\sum_{j=1}^{N}(\mathbf{W}_j-\tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_{0e}^{-1}\boldsymbol{\rho}_j^T\mathbf{b}\right\} \\
& \times \exp\left\{-\frac{1}{2}\sum_{j=1}^{N}(\mathbf{W}_j-\tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_{0e}^{-1}(\mathbf{W}_j-\tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)\right\}.
\end{aligned}
\tag{5.45}
$$

Thus, using the same technique as in the proof of consistency, $\mathbf{b}$ can be treated as a random vector from the normal distribution $N_d(\boldsymbol{\nu},\boldsymbol{\Gamma})$, where $\boldsymbol{\Gamma}=(\boldsymbol{\Sigma}_{0b}^{-1}+\sum_{j=1}^{N}\boldsymbol{\rho}_j^T\boldsymbol{\Sigma}_{0e}^{-1}\boldsymbol{\rho}_j)^{-1}$ and $\boldsymbol{\nu}=\boldsymbol{\Gamma}\left[\sum_{j=1}^{N}\boldsymbol{\rho}_j\boldsymbol{\Sigma}_{0e}^{-1}(\mathbf{W}_j-\tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)\right]$. Hence equation (5.44) can be treated as the expectation of a quadratic function of $\mathbf{b}$, and thus having the following explicit form

$$
\begin{aligned}
0 =& (2\pi)^{-pN/2}|\mathbf{D}|^{1/2}|\boldsymbol{\Sigma}_{0e}|^{N/2}|\boldsymbol{\Sigma}_{0b}|^{-1/2} \\
& \times \exp\left\{\frac{1}{2}\left[\sum_{j=1}^{N}(\mathbf{W}_j-\tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_{0e}^{-1}\boldsymbol{\rho}_j^T\right]\boldsymbol{\Gamma}\left[\sum_{j=1}^{N}\boldsymbol{\rho}_j\boldsymbol{\Sigma}_{0e}^{-1}(\mathbf{W}_j-\tilde{\boldsymbol{\rho}}_j\boldsymbol{\mu}_0)^T\right]\right. \\
& \left.-\frac{1}{2}\sum_{j=1}^{N}(\mathbf{W}_j-\tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_{0e}^{-1}(\mathbf{W}_j-\tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)\right\} \\
& \times \left\{\frac{1}{2}tr\left[(\boldsymbol{\Sigma}_{0b}^{-1}\mathcal{D}_b\boldsymbol{\Sigma}_{0b}^{-1}+\sum_{j=1}^{N}\boldsymbol{\rho}_j\boldsymbol{\Sigma}_{0e}^{-1}\mathcal{D}_e\boldsymbol{\Sigma}_{0e}^{-1}\boldsymbol{\rho}_j^T)\boldsymbol{\Gamma}\right]-\frac{1}{2}tr(\boldsymbol{\Sigma}_{0b}^{-1}\mathcal{D}_b)-\frac{N}{2}tr(\boldsymbol{\Sigma}_{0e}^{-1}\mathcal{D}_e)\right. \\
& +\frac{1}{2}\sum_{j=1}^{N}(\mathbf{W}_j-\tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_{0e}^{-1}\mathcal{D}_e\boldsymbol{\Sigma}_{0e}^{-1}(\mathbf{W}_j-\tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)+\boldsymbol{\nu}^T\boldsymbol{\Sigma}_{0b}\mathcal{D}_b\boldsymbol{\Sigma}_{0b}\boldsymbol{\nu} \\
& +\frac{1}{2}\boldsymbol{\nu}^T\left[\sum_{j=1}^{N}\boldsymbol{\rho}_j\boldsymbol{\Sigma}_{0e}^{-1}\mathcal{D}_e\boldsymbol{\Sigma}_{0e}^{-1}\boldsymbol{\rho}_j^T\right]\boldsymbol{\nu}-\sum_{j=1}^{N}(\mathbf{W}_j-\tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_{0e}^{-1}\mathcal{D}_e\boldsymbol{\Sigma}_{0e}^{-1}\boldsymbol{\rho}_j^T\boldsymbol{\nu} \\
& \left.+\sum_{j=1}^{N}(\mathbf{W}_j-\tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_{0e}^{-1}\tilde{\boldsymbol{\rho}}_j^T\mathbf{h}_1^{\mu}-\boldsymbol{\nu}^T\left[\sum_{j=1}^{N}\boldsymbol{\rho}_j\boldsymbol{\Sigma}_{0e}^{-1}\tilde{\boldsymbol{\rho}}_j^T\right]\mathbf{h}_1^{\mu}\right\}
\end{aligned}
$$

Rearranging the above equation and canceling the non-negative multipliers, we obtain

$$\begin{aligned}
0 =& tr(\boldsymbol{\Sigma}_{0b}^{-1}\mathcal{D}_b\boldsymbol{\Sigma}_{0b}^{-1}\boldsymbol{\Gamma}) + tr(\sum_{j=1}^{N}\boldsymbol{\rho}_j\boldsymbol{\Sigma}_{0e}^{-1}\mathcal{D}_e\boldsymbol{\Sigma}_{0e}^{-1}\boldsymbol{\rho}_j^T\boldsymbol{\Gamma}) - tr(\boldsymbol{\Sigma}_{0b}^{-1}\mathcal{D}_b) - Ntr(\boldsymbol{\Sigma}_{0e}^{-1}\mathcal{D}_e) \\
&+ \sum_{j=1}^{N}(\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_{0e}^{-1}\mathcal{D}_e\boldsymbol{\Sigma}_{0e}^{-1}(\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0) + \boldsymbol{\nu}^T\boldsymbol{\Sigma}_{0b}\mathcal{D}_b\boldsymbol{\Sigma}_{0b}\boldsymbol{\nu} \\
&+ \boldsymbol{\nu}^T\left[\sum_{j=1}^{N}\boldsymbol{\rho}_j\boldsymbol{\Sigma}_{0e}^{-1}\mathcal{D}_e\boldsymbol{\Sigma}_{0e}^{-1}\boldsymbol{\rho}_j^T\right]\boldsymbol{\nu} - 2\sum_{j=1}^{N}(\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_{0e}^{-1}\mathcal{D}_e\boldsymbol{\Sigma}_{0e}^{-1}\boldsymbol{\rho}_j^T\boldsymbol{\nu} \\
&+ 2\sum_{j=1}^{N}(\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_{0e}^{-1}\tilde{\boldsymbol{\rho}}_j^T\mathbf{h}_1^\mu - 2\boldsymbol{\nu}^T\left[\sum_{j=1}^{N}\boldsymbol{\rho}_j\boldsymbol{\Sigma}_{0e}^{-1}\tilde{\boldsymbol{\rho}}_j^T\right]\mathbf{h}_1^\mu
\end{aligned} \tag{5.46}$$

Since the right-hand side of (5.46) is a second-order polynomial of $(\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)$ and $\mathbf{W}_j$ is arbitrary for all $j = 1, \ldots, N$, the first and second order terms are zero, respectively. Let $\mathbf{x}_j = \mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0$.

We first check the first-order terms $\mathbf{x}_j$. Let $\tilde{\mathbf{E}} = \sum_{j=1}^{N}\boldsymbol{\rho}_j\boldsymbol{\Sigma}_{0e}^{-1}\tilde{\boldsymbol{\rho}}_j^T$, we obtain

$$\begin{aligned}
0 &= \sum_{j=1}^{N}(\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_{0e}^{-1}\tilde{\boldsymbol{\rho}}_j^T\mathbf{h}_1^\mu - \sum_{j=1}^{N}(\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_{0e}^{-1}\boldsymbol{\rho}_j^T\tilde{\mathbf{E}}\mathbf{h}_1^\mu \\
&= \sum_{j=1}^{N}(\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_{0e}^{-1}(\tilde{\boldsymbol{\rho}}_j - \tilde{\mathbf{E}}\boldsymbol{\rho}_j)^T\mathbf{h}_1^\mu
\end{aligned} \tag{5.47}$$

Since $(\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)$ is arbitrary in $\mathbb{R}^p$, it follows that

$$\boldsymbol{\Sigma}_{0e}^{-1}(\tilde{\boldsymbol{\rho}}_j^T - \boldsymbol{\rho}_j^T\tilde{\mathbf{E}})\mathbf{h}_1^\mu = \mathbf{0}, \quad \text{for } j = 1, \ldots, N.$$

Define the $pN \times pN$ matrix $\mathbf{A} = \boldsymbol{\Sigma}_{0e}^{-1} \otimes I_N$, and the $pN \times \tilde{d}$ matrix $\mathbf{B} = (\tilde{\boldsymbol{\rho}}_1^T - \boldsymbol{\rho}_1^T\tilde{\mathbf{E}}, \ldots, \tilde{\boldsymbol{\rho}}_N^T - \boldsymbol{\rho}_N^T\tilde{\mathbf{E}})^T$, the above equation can be rewritten into

$$\mathbf{A}\mathbf{B}\mathbf{h}_1^\mu = \mathbf{0}.$$

Multiply both sides of the above equation by $\mathbf{A}^{-1}$, we obtain

$$\mathbf{B}\mathbf{h}_1^\mu = \mathbf{0}.$$

Then multiply both sides of the above equation by $\mathbf{B}^T$, we obtain

$$\mathbf{B}^T\mathbf{B}\mathbf{h}_1^\mu = \mathbf{0}.$$

By assumption (A.4), since $N > \tilde{d}$ and $p \geq 1$, it follows that $pN > \tilde{d}$. Then the $\tilde{d} \times \tilde{d}$ matrix $\mathbf{B}^T\mathbf{B}$ is of full rank. Thus we conclude that $\mathbf{h}_1^\mu = \mathbf{0}$.

Next, and check the second-order terms of $\mathbf{x}_j$ in (5.46), we obtain that

$$\begin{aligned}
0 = &\left(\sum_{j=1}^N \mathbf{x}_j^T\boldsymbol{\Sigma}_{0e}^{-1}\boldsymbol{\rho}_j^T\right)\boldsymbol{\Gamma}\boldsymbol{\Sigma}_{0b}^{-1}\mathcal{D}_b\boldsymbol{\Sigma}_{0b}^{-1}\boldsymbol{\Gamma}\left(\sum_{j=1}^N \boldsymbol{\rho}_j\boldsymbol{\Sigma}_{0e}^{-1}\mathbf{x}_j\right) \\
&+ \sum_{j=1}^N \mathbf{x}_j^T\boldsymbol{\Sigma}_{0e}^{-1}\mathcal{D}_e\boldsymbol{\Sigma}_{0e}^{-1}\mathbf{x}_j \\
&+ \left(\sum_{j=1}^N \mathbf{x}_j^T\boldsymbol{\Sigma}_{0e}^{-1}\boldsymbol{\rho}_j^T\right)\boldsymbol{\Gamma}\left(\sum_{j=1}^N \boldsymbol{\rho}_j\boldsymbol{\Sigma}_{0e}^{-1}\mathcal{D}_e\boldsymbol{\Sigma}_{0e}^{-1}\boldsymbol{\rho}_j^T\right)\boldsymbol{\Gamma}\left(\sum_{j=1}^N \boldsymbol{\rho}_j\boldsymbol{\Sigma}_{0e}^{-1}\mathbf{x}_j\right) \\
&- 2\left(\sum_{j=1}^N \mathbf{x}_j^T\boldsymbol{\Sigma}_{0e}^{-1}\mathcal{D}_e\boldsymbol{\Sigma}_{0e}^{-1}\boldsymbol{\rho}_j^T\right)\boldsymbol{\Gamma}\left(\sum_{j=1}^N \boldsymbol{\rho}_j\boldsymbol{\Sigma}_{0e}^{-1}\mathbf{x}_j\right).
\end{aligned} \tag{5.48}$$

Let $\mathbf{S} = \sum_{j=1}^N \boldsymbol{\rho}_j\boldsymbol{\Sigma}_{0e}^{-1}\mathbf{x}_j$ and $\mathbf{E} = \boldsymbol{\Sigma}_{0e}^{-1}\mathcal{D}_e\boldsymbol{\Sigma}_{0e}^{-1}$. The above equation becomes

$$\begin{aligned}
0 &= \mathbf{S}^T\boldsymbol{\Gamma}\boldsymbol{\Sigma}_{0b}^{-1}\mathcal{D}_b\boldsymbol{\Sigma}_{0b}^{-1}\boldsymbol{\Gamma}\mathbf{S} + \sum_{j=1}^N \mathbf{x}_j^T\mathbf{E}\mathbf{x}_j + \mathbf{S}^T\boldsymbol{\Gamma}\left(\sum_{j=1}^N \boldsymbol{\rho}_j\mathbf{E}\boldsymbol{\rho}_j^T\right)\boldsymbol{\Gamma}\mathbf{S} - 2\left(\sum_{j=1}^N \mathbf{x}_j^T\mathbf{E}\boldsymbol{\rho}_j^T\right)\boldsymbol{\Gamma}\mathbf{S} \\
&= \mathbf{S}^T\boldsymbol{\Gamma}\boldsymbol{\Sigma}_{0b}^{-1}\mathcal{D}_b\boldsymbol{\Sigma}_{0b}^{-1}\boldsymbol{\Gamma}\mathbf{S} + \sum_{j=1}^N \left(\mathbf{x}_j^T\mathbf{E}\mathbf{x}_j + \mathbf{S}^T\boldsymbol{\Gamma}\boldsymbol{\rho}_j\mathbf{E}\boldsymbol{\rho}_j^T\boldsymbol{\Gamma}\mathbf{S} - 2\mathbf{x}_j^T\mathbf{E}\boldsymbol{\rho}_j^T\boldsymbol{\Gamma}\mathbf{S}\right) \\
&= \mathbf{S}^T\boldsymbol{\Gamma}\boldsymbol{\Sigma}_{0b}^{-1}\mathcal{D}_b\boldsymbol{\Sigma}_{0b}^{-1}\boldsymbol{\Gamma}\mathbf{S} + \sum_{j=1}^N \left(\mathbf{x}_j - \boldsymbol{\rho}_j^T\boldsymbol{\Gamma}\mathbf{S}\right)^T\mathbf{E}\left(\mathbf{x}_j - \boldsymbol{\rho}_j^T\boldsymbol{\Gamma}\mathbf{S}\right)
\end{aligned} \tag{5.49}$$

By definition of $\mathbf{S}$, in the above equation, $\mathbf{S}$ is an arbitrary vector in $\mathbb{R}^d$ because $\mathbf{x}_j$ is arbitrary in $\mathbb{R}^p$ and $\boldsymbol{\rho}_j$ is full rank by assumption (A.6). Thus, the right-hand side of (5.49) is the sum of $N+1$ quadratic terms. Since both $\mathbf{E}$ and $\boldsymbol{\Gamma}\boldsymbol{\Sigma}_{0b}^{-1}\mathcal{D}_b\boldsymbol{\Sigma}_{0b}^{-1}\boldsymbol{\Gamma}$ are symmetric and non-negative definite, it follows that $\mathbf{E} = \boldsymbol{\Sigma}_{0e}^{-1}\mathcal{D}_e\boldsymbol{\Sigma}_{0e}^{-1} = \mathbf{0}$ and $\boldsymbol{\Gamma}\boldsymbol{\Sigma}_{0b}^{-1}\mathcal{D}_b\boldsymbol{\Sigma}_{0b}^{-1}\boldsymbol{\Gamma} = \mathbf{0}$ Since $\boldsymbol{\Gamma}$ is invertible, we conclude that $\mathcal{D}_e = \mathbf{0}$ and $\mathcal{D}_b = \mathbf{0}$.

Next, let $\Delta = 0$ in (5.43). Based on the above conclusions that $\mathbf{h}_1^\mu = \mathbf{0}$, $\mathcal{D}_e = \mathbf{0}$ and $\mathcal{D}_b = \mathbf{0}$, we obtain that

$$\begin{aligned} E_{\mathbf{b}} \Bigg[\exp\Big\{ - \int_0^V e^{\boldsymbol{\beta}_0^T(\boldsymbol{\rho}(t)^T\mathbf{b}) + \boldsymbol{\eta}_0^T\mathbf{z}} d\Lambda_0(t) \Big\} \\ \times \int_0^V e^{\boldsymbol{\beta}_0^T(\boldsymbol{\rho}(t)^T\mathbf{b}) + \boldsymbol{\eta}_0^T\mathbf{z}} \{(\boldsymbol{\rho}(t)^T\mathbf{b})^T\mathbf{h}_1^\beta + \mathbf{Z}^T\mathbf{h}_1^\eta + h_2(t)\} d\Lambda_0(t) \Bigg] = 0, \end{aligned} \tag{5.50}$$

where $\mathbf{b}$ is from $N_d(\boldsymbol{\nu}, \boldsymbol{\Gamma})$ as in (5.44), with

$$\boldsymbol{\Gamma} = (\boldsymbol{\Sigma}_{0b}^{-1} + \sum_{j=1}^N \boldsymbol{\rho}_j^T \boldsymbol{\Sigma}_{0e}^{-1} \boldsymbol{\rho}_j)^{-1},$$

and

$$\boldsymbol{\nu} = \boldsymbol{\Gamma} \left[\sum_{j=1}^N \boldsymbol{\rho}_j \boldsymbol{\Sigma}_{0e}^{-1} (\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T \boldsymbol{\mu}_0) \right].$$

Since $\mathbf{b}$ is complete and sufficient statistic for $\boldsymbol{\nu}$, it follows that

$$\int_0^V e^{\boldsymbol{\beta}_0^T(\boldsymbol{\rho}(t)^T\mathbf{b}) + \boldsymbol{\eta}_0^T\mathbf{z}} \{(\boldsymbol{\rho}(t)^T\mathbf{b})^T\mathbf{h}_1^\beta + \mathbf{Z}^T\mathbf{h}_1^\eta + h_2(t)\} d\Lambda_0(t) = 0,$$

which yields

$$(\boldsymbol{\rho}(t)^T\mathbf{b})^T\mathbf{h}_1^\beta + \mathbf{Z}^T\mathbf{h}_1^\eta + h_2(t) = 0,$$

with an arbitrary $\mathbf{b}$. Thus by assumption (A.7), we conclude that $\mathbf{h}_1^\beta = \mathbf{0}$, $\mathbf{h}_1^\eta = \mathbf{0}$ and $h_2(t) \equiv 0$.

Now that we have verified that condition (a)-(d) of the theorem are satisfied, $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0, \widehat{\Lambda} - \Lambda_0)$ weakly converges to a tight random element in $l^\infty(\mathcal{H})$. Using the same argument as in Zeng and Cai (2005), we conclude that $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0, \widehat{\Lambda} - \Lambda_0)$ weakly converges to a Gaussian process in $l^\infty(\mathcal{H})$, and $\widehat{\boldsymbol{\theta}}$ is an efficient estimator for $\boldsymbol{\theta}_0$. □

References

Albert, P. S. and J. H. Shih (2010). An approach for jointly modeling multivariate longitudinal measurements and discrete time-to-event data. *The Annals of Applied Statistics 4*(3), 1517.

Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics 30*(1), 89–99.

Brown, E. and J. Ibrahim (2003). A bayesian semiparametric joint hierarchical model for longitudinal and survival data. *Biometrics 59*(2), 221–228.

Brown, E. R., J. G. Ibrahim, and V. DeGruttola (2005). A flexible b-spline model for multiple longitudinal biomarkers and survival. *Biometrics 61*(1), 64–73.

Chi, Y.-Y. and J. G. Ibrahim (2006). Joint models for multivariate longitudinal and multivariate survival data. *Biometrics 62*(2), 432–445.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological) 39*(1), 1–38.

Ding, J. and J.-L. Wang (2008). Modeling longitudinal data with nonparametric multiplicative random effects jointly with survival data. *Biometrics 64*(2), 546–556.

Fang, K.-T. and Y. Wang (1994). *Number theoretic methods in statistics*, Volume 51. CRC Press.

Faucett, C. L., N. Schenker, and R. M. Elashoff (1998). Analysis of censored survival data with intermittently observed time-dependent binary covariates. *Journal of the American Statistical Association 93*(442), 427–437.

Faucett, C. L. and D. C. Thomas (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: a gibbs sampling approach. *Statistics in Medicine 15*(15), 1663–1685.

Hatfield, L. A., M. E. Boye, and B. P. Carlin (2011). Joint modeling of multiple longitudinal patient-reported outcomes and survival. *Journal of Biopharmaceutical Statistics 21*(5), 971–991.

Henderson, R., P. Diggle, and A. Dobson (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics 1*(4), 465–480.

Hsieh, F., Y.-K. Tseng, and J.-L. Wang (2006). Joint modeling of survival and longitudinal data: likelihood approach revisited. *Biometrics 62*(4), 1037–1043.

Huang, W., S. L. Zeger, J. C. Anthony, and E. Garrett (2001). Latent variable model for joint analysis of multiple repeated measures and bivariate event times. *Journal of the American Statistical Association 96*(455), 906–914.

Ibrahim, J. G., M.-H. Chen, and D. Sinha (2004). Bayesian methods for joint modeling of longitudinal and survival data with applications to cancer vaccine trials. *Statistica Sinica 14*(3), 863–884.

Joseph, V. R. (2012). Bayesian computation using design of experiments-based interpolation technique. *Technometrics 54*(3), 209–225.

Liu, L. (2009). Joint modeling longitudinal semi-continuous data and survival, with application to longitudinal medical cost data. *Statistics in Medicine 28*(6), 972–986.

Loeppky, J. L., J. Sacks, and W. J. Welch (2009). Choosing the sample size of a computer experiment: A practical guide. *Technometrics 51*(4), 366–376.

Louis, T. A. (1982). Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological) 44*(2), 226–233.

Piper, M. E., S. S. Smith, T. R. Schlam, M. C. Fiore, D. E. Jorenby, D. Fraser, and T. B. Baker (2009). A randomized placebo-controlled clinical trial of 5 smoking cessation pharmacotherapies. *Archives of General Psychiatry 66*(11), 1253–1262.

Prentice, R. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika 69*(2), 331–342.

Rizopoulos, D. (2010). Jm: An r package for the joint modeling of longitudinal and time-to-event data. *Journal of Statistical Software 35*(9), 1–33.

Rizopoulos, D., G. Verbeke, and E. Lesaffre (2009). Fully exponential laplace approximations for the joint modeling of survival and longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71*(3), 637–654.

Song, X., M. Davidian, and A. A. Tsiatis (2002a). An estimator for the proportional hazards model with multiple longitudinal covariates measured with error. *Biostatistics 3*(4), 511–528.

Song, X., M. Davidian, and A. A. Tsiatis (2002b). A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data. *Biometrics 58*(4), 742–753.

Tierney, L. and J. B. Kadane (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association 81*(393), 82–86.

Tierney, L., R. E. Kass, and J. B. Kadane (1989). Fully exponential laplace approximations to expectations and variances of nonpositive functions. *Journal of the American Statistical Association 84*(407), 710–716.

Tseng, Y.-K., F. Hsieh, and J.-L. Wang (2005). Joint modelling of accelerated failure time and longitudinal data. *Biometrika 92*(3), 587–603.

Tsiatis, A., V. Degruttola, and M. Wulfsohn (1995). Modeling the relationship of survival to longitudinal data measured with error. applications to survival and cd4 counts in patients with aids. *Journal of the American Statistical Association 90*(429), 27–37.

Tsiatis, A. A. and M. Davidian (2001). A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika 88*(2), 447–458.

Wang, Y. and J. M. G. Taylor (2001). Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. *Journal of the American Statistical Association 96*(455), 895–905.

Wulfsohn, M. S. and A. A. Tsiatis (1997). A joint model for survival and longitudinal data measured with error. *Biometrics 53*(1), 330–339.

Xu, J. and S. L. Zeger (2001a). The evaluation of multiple surrogate endpoints. *Biometrics 57*(1), 81–87.

Xu, J. and S. L. Zeger (2001b). Joint analysis of longitudinal data comprising repeated measures and times to events. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 50*(3), 375–387.

Yu, B. and P. Ghosh (2010). Joint modeling for cognitive trajectory and risk of dementia in the presence of death. *Biometrics 66*(1), 294–300.

Yu, M., N. J. Law, J. M. Taylor, and H. M. Sandler (2004). Joint longitudinal-survival-cure models and their application to prostate cancer. *Statistica Sinica 14*(3), 835–862.

Zeng, D. and J. Cai (2005). Asymptotic results for maximum likelihood estimators in joint analysis of repeated measurements and survival time. *The Annals of Statistics 33*(5), 2132–2163.