Statistica Sinica **23** (2013), 1181-1213
doi:http://dx.doi.org/10.5705/ss.2012.118

METHOD OF SIEVES TO JOINTLY MODEL SURVIVAL AND LONGITUDINAL DATA

Fushing Hsieh[1], Jimin Ding[2] and Jane-Ling Wang[1]

[1] University of California, Davis, and [2] Washington University at St. Louis

Abstract: In biomedical studies, longitudinal covariates are often used to monitor the progress of a disease as well as survival time. However, a sparse covariate history, possibly in combination with measurement error, adds complications to the survival analysis. Moreover, marginal analysis of the longitudinal covariates may incur biases due to informative dropout of the longitudinal processes when death is the endpoint for survival time. Joint modeling of survival and longitudinal data can gain information from both components, and has proved as an effective way to model their relationship. A common approach is the semiparametric joint likelihood approach of Wulfsohn and Tsiatis (1997). However, it suffers from computational instability due to the large number of parameters involved in the likelihood and difficulties with standard error estimation. In this article, we propose the method of sieves and establish asymptotic consistency and the rate of convergence of the resulting sieve maximum-likelihood estimate (SMLE), including the estimate for the baseline hazard function. Results from numerical studies support this approach. The proposed SMLE is applied to a liver cirrhosis study for further illustration.

Key words and phrases: Asymptotic theory, EM algorithm, joint likelihood, missing data, Monte Carlo integration, nonparametric maximum likelihood method.

1. Introduction

In biomedical research, subjects are monitored throughout a study, and key concerns are their progression toward some event of interest as well as the dynamics of some longitudinal biomarker processes. From the statistical perspective, the former involves survival analysis, while the latter involves longitudinal data analysis. Recently, more attention has been drawn to jointly modeling the survival and longitudinal data to better understand the entire dynamic system.

Several issues arise when longitudinal data are collected with survival data. The first is that survival models with time-dependent covariates need the complete history of the longitudinal process from entry time to event-time. This is not always feasible, because subjects are only measured intermittently and measurement times can vary among subjects even when the original schedule is regular. Although many imputation techniques, such as the Last-Value-Carry-Forward (LVCF) employed in standard statistical packages, provide convenient ways to fill in the missing covariate values, most of them introduce biases in subsequent survival analysis. A second issue is the termination of the longitudinal observations by the event of interest, which occurs when a subject is either dead or censored. The loss of longitudinal measurements caused by death triggers informative drop-out of the longitudinal process and induces bias in the analysis unless proper care is taken. A survey article by Tsiatis and Davidian (2004) reviews several existing approaches to reduce or correct both biases simultaneously. Guo, Ratcliffe and Ten Have (2004) provided a practical implementation of the Bayesian joint modeling approach suggested by Henderson, Diggle and Dobson (2000). A more recent review on joint models can be found in Verbeke and Davidian (2008). The semiparametric joint likelihood approach to modeling both types of data together emerges as the most satisfactory method. This approach, first proposed by Wulfsohn and Tsiatis (1997), is based on nonparametric maximum likelihood estimation (NPMLE) of the baseline hazard function, which assigns point masses to all uncensored event-times for the baseline hazard function. This leads to a situation where the number of parameters is of the order of the sample size and causes both computational and theoretical challenges. Elegant asymptotic theory was established in Zeng and Cai (2005) and Dupuy, Grama and Mesbah (2006) for finite dimensional parameters and the cumulative baseline hazard function, but the resulting point mass baseline hazard estimator is not consistent and the computational challenges remain unresolved, as described in Section 2.

We propose here a more convenient estimator found according to the method of sieves of Grenander (1981). The idea is to by-pass the high-dimensional MLE approach by first restricting the parameter space to a low-dimensional "sieve space", which grows with the sample size and eventually densely fills the entire parameter space. By suitably choosing the sieve space, we can reduce the number of parameters to close to $O(n^{1/3})$, where n is the sample size. This produces a consistent baseline hazard estimation, which NPMLE does not. The dimension reduction facilitates not only asymptotic theory but also computational stability of the estimates and their corresponding asymptotic covariance matrix, as illustrated by simulations and a case study.

This paper is organized as follows. In Section 2, we set up the joint model and explain the computational and theoretical challenges of the NPMLE that motivates the use of the method of sieves. In Section 3, we explain the method of sieves and establish the asymptotic properties of the proposed sieve estimators. Proofs are relegated to the Appendix. Several simulation studies in Section 4 demonstrate the validity of the asymptotic variance estimates of the sieve MLE, and show that it circumvents the instability of the NPMLE. The method is applied to the Primary Biliary Cirrhosis (PBC) data in Section 5.

2. Model and Motivation

Let T_i be the time to the event of interest for the ith subject, which is subject to a censoring time C_i. We assume that the T_i and C_i are independent in each risk set at t, conditioning on the covariate history to time t. The observed survival data from the i-th individual is (V_i, Δ_i) with $V_i = \min(T_i, C_i)$ and $\Delta_i = 1_{(T_i \le C_i)}$. Often, the survival time relates to some covariates. We assume for simplicity that there is only one time-dependent process $X_i(t)$ and that the hazard function follows the Cox model

$$\lambda(t|\bar{X}_i(t)) = \lambda(t)\exp\{\beta X_i(t)\}, \tag{2.1}$$

where $\lambda(t)$ is the baseline hazard function shared by all subjects and $\bar{X}_i(t) = \{X_i(s) : 0 \le s \le t\}$ indicates the entire history of the covariate process for the i-th subject up to time t. However, $X_i(t)$ is not observed, and instead we observe $\mathbf{W}_i = (W_{i1}, \ldots, W_{in_i})'$ at intermittently scheduled time points $\mathbf{t}_i = (t_{i1}, t_{i2}, \ldots, t_{in_i})'$ through

$$W_{ij} = W_i(t_{ij}) = X_i(t_{ij}) + e_i(t_{ij}), \tag{2.2}$$

where the measurement errors $e_{ij} = e(t_{ij}) \overset{\text{iid}}{\sim} N(0, \sigma_e^2)$ and independent of the time-varying covariate process $X_i(t)$. The observed $\mathbf{W}_i$ is only available up to V_i, hence $t_{in_i} \le V_i$ and the number of observed measurements n_i for each subject is random.

The classical partial likelihood approach is not applicable here because it requires the complete history $\bar{X}_i(t)$. As a resolution, a linear mixed effect model is assumed for the unobservable covariate process $X_i(t)$,

$$X_i(t) \triangleq X_i(t; \mathbf{b}_i) = \mathbf{b}_i'\psi(t), \tag{2.3}$$

where $\psi(t) = (\psi_0(t), \psi_1(t), \ldots, \psi_{q-1}(t))'$ are q linearly independent basis functions, and the vector of random coefficients $\mathbf{b}_i = (b_{i0}, b_{i1}, \ldots, b_{iq-1})$ is independent across subjects, and with normal distribution $\pi(\cdot\,; \alpha, \Sigma_b)$. A common choice of the basis functions in the literature is $\{\psi_0(t) \equiv 1, \psi_1(t) = t\}$, aiming at capturing the overall linear trend of the longitudinal process.

Combining the Cox model (2.1) for survival data and the longitudinal measurements (2.2) with the random effect structure (2.3), the marginal likelihood function of the observed data $\{O_i = (V_i, \Delta_i, \mathbf{W}_i, \mathbf{t}_i, n_i);\quad i = 1, \ldots, n\}$ is

$$\mathcal{L}(\theta) = \prod_{i=1}^{n}\Big[\int_{R^q}\Big\{\prod_{j=1}^{n_i} f(W_{ij}|\mathbf{b}_i, \sigma_e^2)\Big\}\pi(\mathbf{b}_i; \alpha, \Sigma_b) f(V_i, \Delta_i|\mathbf{b}_i, \beta, \lambda)\mathrm{d}\mathbf{b}_i\Big], \tag{2.4}$$

where $\theta = \{\alpha, \Sigma_b, \sigma_e^2, \beta, \lambda(\cdot)\}$ is the model parameter and f denotes the corresponding density functions for the longitudinal and survival data:

$$f(W_{ij}|\mathbf{b}_i, \sigma_e^2) = (2\pi\sigma_e^2)^{-1/2} \exp\Big\{ - \frac{[W_{ij} - X_i(t_{ij}; \mathbf{b}_i)]^2}{2\sigma_e^2} \Big\},$$
$$f(V_i, \Delta_i|\mathbf{b}_i, \beta, \lambda(\cdot)) = [\lambda(V_i) \exp\{\beta X_i(V_i; \mathbf{b}_i)\}]^{\Delta_i} \exp\left[- \int_0^{V_i} \lambda(t) e^{\beta X_i(t;\mathbf{b}_i)} \mathrm{d}t \right].$$

If we treat the random effect $\boldsymbol{b}_i$ as missing data, we can write the complete likelihood for data $\{O_i, \boldsymbol{b}_i\}$ from the ith subject as $L_i^{(c)}(\theta) = \prod_{j=1}^{n_i} f(W_{ij}|\mathbf{b}_i, \sigma_e^2)$ $\pi(\mathbf{b}_i; \alpha, \Sigma_b) f(V_i, \Delta_i|\mathbf{b}_i, \beta, \lambda)$. Then the likelihood contributed by the ith subject can be written as $L_i = \int_{R^k} L_i^{(c)}(\theta) \mathrm{d}\mathbf{b}_i$ and the posterior density of random effects is

$$h(\mathbf{b}_i|O_i, \theta) = \frac{L_i^{(c)}(\theta)}{\int_{R^k} L_i^{(c)}(\theta) \mathrm{d}\mathbf{b}_i}. \tag{2.5}$$

Following the classic Cox model Johansen (1983), one can take the derivative of the logarithm of the observed likelihood (2.4) and derive self-consistent equations for the parameters as

$$\alpha = \sum_{i=1}^n \frac{E_{i,\theta}(\mathbf{b}_i)}{n},$$
$$\Sigma_b = \sum_{i=1}^n \frac{E_{i,\theta}[(\mathbf{b}_i - \hat\alpha)(\mathbf{b}_i - \hat\alpha)']}{n},$$
$$\sigma_e^2 = \frac{\sum_{i=1}^n \sum_{j=1}^{m_i} E_{i,\theta}[(W_{ij} - X_i(t_{ij}))^2]}{\sum_{i=1}^n m_i},$$
$$\lambda(t) = \sum_{i=1}^n \frac{\Delta_i 1_{(V_i = t)}}{\sum_{j=1}^n E_{j,\theta}[\exp\{\beta X_j(t)\}] Y_j(t)},$$

where $Y_j(t) = 1_{(V_j \geq t)}$ is an indicator at risk set and $E_{i,\theta}[\cdot]$ is taken with respect to the posterior density. As for the solution of β, these expressions can be plugged into the likelihood $\mathcal{L}(\theta)$ to obtain the implicit pseudo profiled score equation

$$S^{ip}(\beta) = \sum_{i=1}^n \Delta_i \left[E_{i,\theta}[X_i(V_i)] - \frac{\sum_{j=1}^n E_{j,\theta}[X_j(V_i) \exp\{\beta X_j(V_i)\}] Y_j(V_i)}{\sum_{j=1}^n E_{j,\theta}[\exp\{\beta X_j(V_i)\}] Y_j(V_i)} \right]. \tag{2.6}$$

We note that this is not the conventional profile equation studied in Murphy and van der Vaart (2000), because the pseduo profile score equation (2.6) does not profile out the baseline hazard function in the sense that the conditional expectation $E_{i,\theta}$ involves λ through (2.5). Therefore, a direct application of Newton-Raphson method to estimate β is not feasible and a proxy version through the expectation conditional maximization (ECM) algorithm by Meng and Rubin (1993) is employed instead. Alternatively, one can consider a more sophisticatedly modified version of Newton-Raphson algorithm such as the reweighted iteration procedure of Chen (2009).

This nonparametric maximum likelihood approach Kiefer and Wolfowitz (1956) has been successfully applied in the classic Cox model due to its special structure of proportional hazards. This approach continues to hold in the joint modeling setting as carried out in Wulfsohn and Tsiatis (1997). Although the NPMLE approach leads to satisfactory estimation of the parametric components via the EM-algorithm, it is computationally unstable and faces challenges in estimating the standard errors. The computational challenge is attributable to the mixture structure of the joint likelihood with the random effects that appear in both the longitudinal and survival model. The posterior expectation in the score functions is a nonlinear function of all parameters, so the score equations cannot be solved explicitly, rendering no tractable profile likelihood for the survival regression parameters. Thus, unlike the partial likelihood approach in the classic Cox model, the survival regression parameter now depends on the baseline hazard function due to the presence of the random effects.

Another complication is that the sample Fisher information $\hat{i}_{\lambda\lambda}$ may not be semi-positive definite unlike in the classic Cox Model. A positive probability that $\hat{i}_{\lambda\lambda}$ is not semi-positive definite leads to the possibility of multiple local maxima of the likelihood function. Thus the NPMLE for λ may not be unique in the joint modeling setting and the proxy version of the EM algorithm may fail to converge. The method of sieves provides a potential remedy by reducing the dimensionality of the baseline hazard function: some subjects are pooled to obtain the information for λ so that $i_{\lambda\lambda}$ is become semi-positive definite. This makes standard error estimation possible.

For the baseline hazard function $\lambda(t)$, the corresponding NPMLE is a point mass function with masses at all uncensored event-time points. This baseline hazard estimation is inconsistent and, due to the large number of parameters introduced by $\lambda(t)$, the estimating algorithm is usually computationally unstable. We propose to estimate λ within piecewise constant sieve space, as discussed in details in next section. By reducing the number of parameters, we obtain computational stability and standard error estimation, as well as consistent estimation of the baseline hazard function.

Remark. The current model can be easily extended to include additional time-independent covariates in either survival or longitudinal submodels or both. More complicated models add computational costs. But, under some regularity conditions, the theoretical conclusions are similar.

3. The Method of Sieves

The method of sieves as proposed in Grenander (1981) uses finite-dimensional spaces to approximate an infinite-dimensional space. It is an attractive approach that exploits the flexibility of a semi-parametric model while retaining the simplicity of a parametric approach. There are two difficulties to be overcome: the choice of the sieve space is crucial (Geman and Hwang (1982); Shen and Wong (1994)) and varies from case to case; there is a sieve bias that may be hard to nail down. Some recent work in survival analysis has successfully employed the method of sieves, including Xue, Lam and Li (2004). To overcome the difficulties in our semi-parametric setting, we need to reduce the dimension of the nonparametric part to the extent that the estimation bias induced by the approximation is negligible. This requires the asymptotic sieve bias to be of a smaller order than the variance. As it happens the special structure of the Cox model allows the term $\prod_{i=1}^{n}[\lambda(t)]^{\Delta_i}$ to be taken out of the integration in the joint likelihood (2.4), so that the conditional expectation $E_{i,\theta}[\cdot]$ depends on λ only through $\int_0^{V_i}\lambda(t)\exp\{\beta X_i(t;b_i)\}\mathrm{d}t$. Taking advantage of this, we show that simple piecewise constant hazard functions provide a simple way to track the sieve biases and are therefore a natural choice for the sieve space.

Let the parameter space of λ be $\mathbb{S} = \{\lambda(\cdot)|\lambda \in C^{1,d}, 0 < c_0 \leq \lambda(t), \forall t \in [0,\infty)\}$ for some small constant c_0 and $0 < d \leq 1$. Here the space $C^{k,d}$ consists of those functions which have derivatives up to order k, where k is a nonnegative integer, and such that the kth derivatives are Hölder continuous with exponent d, where $0 < d \leq 1$. The technical assumption on the lower bound c_0 could be relaxed and all arguments still hold if λ is bounded from below after some fixed time $\tau: \quad 0 < c_0 \leq \lambda_0(t), \forall t > \tau$. Under careful analysis, c_0 may be allowed go to 0 at a very slow rate, but we forgo such technical refinements in favor of simplicity of presentation.

For the sieve spaces, we use a series of step functions to approximate λ. To be precise, let $0 = t_{(1)} < t_{(2)} < \cdots < t_{m_n} = K_1 \log n$ be a chosen partition on $[0, K_1 \log n]$ for some constant $K_1 > 1/c_0$, and let $\bar{\Delta}_{m_n} = \sup_j |t_{(j-1)} - t_{(j)}|$ be the mesh size. Note that

$$\exp\{-\Lambda(K_1 \log n)\} \leq \exp\Big\{-\int_0^{\log n/c_0} c_0 ds\Big\} = n^{-1},$$

which implies that the baseline survival probability beyond $K_1 \log n$ is bounded by n^{-1}. We consider the sequence of sieve spaces

$$\mathbb{S}_{m_n} = \Big\{\lambda|\lambda(\cdot) = \sum_{j=1}^{m_n} c_j D_j(\cdot), c_0 \leq c_j \leq K_0 m_n, D_j(t) = 1_{(t_{(j-1)},t_{(j)}]}(t), t \in [0, K_1 \log n]\Big\}, \tag{3.1}$$

where K_0 is some constant and m_n goes to infinity as n goes to infinity. It is easy to see that $\mathbb{S}_{m_n} \subseteq \mathbb{S}_{m_n+1}$ and $\mathbb{S} \subseteq \overline{\bigcup_{m_n=1} \mathbb{S}_{m_n}}$ (under the topology introduced by the L^2 norm).

To simplify the notation, we hereafter suppress the subscript n in m_n. We use K and K' for positive constants and ε for small positive constants. Recall that $\theta = \{\alpha, \Sigma_b, \sigma_e^2, \beta, \lambda(\cdot)\}$ denotes all parameters in the model in Section 2. Let $\Theta_{-\{\lambda\}}$ be the finite dimensional parameter space excluding the baseline hazard function and $\theta_{-\{\lambda\}} = \{\alpha, \Sigma_b, \sigma_e^2, \beta\} \in \Theta_{-\{\lambda\}}$. We take the sieve space $\Theta_m = \Theta_{-\{\lambda\}} \times \mathbb{S}_m$ and denote the parameters restricted in the sieve space by $\theta_m = \{\alpha_m, \Sigma_{b,m}, \sigma_{e,m}^2, \beta_m, \lambda_m(\cdot)\}$.

The sieve MLE $\hat{\theta}_m$ is defined as the point which maximizes the observed likelihood within the sieve space,

$$\hat{\theta}_m = \arg\max_{\theta_m \in \Theta_m} \frac{1}{n} \sum_{i=1}^{n} \log L_i(\theta_m) = \arg\inf_{\theta_m \in \Theta_m} \frac{1}{n} \sum_{i=1}^{n} \log \frac{L_i(\theta)}{L_i(\theta_m)}.$$

Note that the true λ may not be in $\mathbb{S}_m$ for any m. Therefore, for a given n and m, the maximum likelihood estimate that maximizes $\log \mathcal{L}(\theta)$ over Θ_m is not targeting θ but rather a parameter value $\theta_m^* \in \Theta_m$ that is the minimizer of the Kullback-Leibler (KL) divergence:

$$\theta_m^* = \arg\inf_{\theta_m \in \Theta_m} E_\theta \Big[\log \frac{L_i(\theta)}{L_i(\theta_m)} \Big] = \arg\inf_{\theta_m \in \Theta_m} \quad KL(\theta, \theta_m). \tag{3.2}$$

Thus $\hat{\theta}_m$ minimizes $\widehat{KL}(\theta, \theta_m)$, where $\widehat{KL}(\theta, \theta_m) = (1/n) \sum_{i=1}^{n} \log L_i(\theta)/L_i(\theta_m)$.

Since θ_m^* may be outside the original parameter space Θ, the distance between θ_m^* and θ, the sieve bias, depends on the choice of m and should be calculated and controlled carefully. Typically, a larger sieve space Θ_m yields to a smaller distance to θ and thus a smaller sieve bias. And, on the other hand, a larger sieve space leads to a larger "modulus of continuity" of the centered processes $\sqrt{n}(\widehat{KL} - KL)$, which measures the variation between $\hat{\theta}_m$ and θ_m^* (van der Vaart and Wellner (1996, p.323)). Thus, the "bias" and "variance" should be balanced to achieve a good rate of convergence. We show that the "sieve bias" in our case can be reduced to zero to obtain $\sqrt{n}$ consistency for the finite-dimensional parameters $\theta_{-\{\lambda\}}$.

Another issue here is that, even if Θ_m is convex, the KL divergence is not a metric distance but only a premetric that measures the statistical departure between two distributions. Therefore, θ_m^* may not be unique and one needs to carefully select the sieve space. A practical suggestion is provided in Section 4.

3.1. Sieve bias

To look at the sieve bias for the sieve space (3.1), we define a metric distance $\rho(\theta, \theta_m)$ between two parameter values, θ and θ_m as

$$\rho(\theta, \theta_m)^2 = \int |\lambda_m(t) - \lambda(t)|^2 \mathrm{d}t + \|\theta_{m-\{\lambda\}} - \theta_{-\{\lambda\}}\|^2,$$

where $\|\cdot\|$ denotes the Euclidian norm. The sieve bias is thus equal to $\rho(\theta, \theta_m^*)$.

To better compute the bias, we first turn to a similar but simpler situation under the Cox regression model, in which the longitudinal covariate is time-invariant and known. With $\theta_m = (\beta_m, \lambda_m(\cdot))$, the corresponding KL divergence is:

$$\begin{aligned} KL(\theta, \theta_m) &= E_{X,\theta}[\log \frac{f(T|\theta)}{f(T|\theta_m)}], \\ &\approx \frac{1}{2} E_{X,\theta}[(\frac{\lambda_m(T)}{\lambda(T)} - 1) + (\beta_m - \beta)X(T)]^2. \end{aligned}$$

For piecewise constant sieve space $\mathbb{S}_m$, the minimizer of $KL(\theta, \theta_m)$ is the solution to

$$\begin{aligned} c_k &= \frac{E_\theta[\lambda^{-1}(T)D_k] - (\beta_m - \beta)E_{X,\theta}[X(T)\lambda^{-1}(T)D_k]}{E_\theta[\lambda^{-2}(T)D_k]}, \quad k = 1, \ldots, m; \\ \beta_m &= \beta - E_X\{\frac{\sum_{k=1}^m E_\theta[X(T)(c_k\lambda^{-1}(T) - 1)D_k]}{E_\theta[X^2(T)]}\}. \end{aligned}$$

Let $\tilde{\theta}_m = (\tilde{\beta}_m, \tilde{c}_1, \ldots, \tilde{c}_m)$ be the unique solution of this system of equations,

$$\begin{aligned} \tilde{\beta}_m &= \beta + \frac{E_{X,\theta}[X(T)] - \sum_{k=1}^m \frac{E_\theta[\lambda^{-1}(T)D_k(T)]E_{X,\theta}[\lambda^{-1}(T)D_k(T)X(T)]}{E_\theta[\lambda^{-2}(T)D_k(T)]}}{E_{X,\theta}[X^2(T)] - \sum_{k=1}^m \frac{E_{X,\theta}^2[\lambda^{-1}(T)D_k(T)X(T)]}{E_\theta[\lambda^{-2}(T)D_k(T)]}}, \\ \tilde{c}_k &= \frac{E_\theta[\lambda^{-1}(T)D_k(T)]}{E_\theta[\lambda^{-2}(T)D_k(T)]} - (\tilde{\beta}_m - \beta)\frac{E_{X,\theta}[\lambda^{-1}(T)D_k(T)X(T)]}{E_\theta[\lambda^{-2}(T)D_k(T)]}. \end{aligned}$$

Given $\lambda \in C^{(1,d)}$, since the covariate process $X(t)$ is bounded and smooth enough, it can be shown that $\tilde{\beta}_m - \beta = O(\bar{\Delta}_m^{1+d})$ and $[\tilde{c}_k D_k(t)/\lambda(t)] - 1 = O(\bar{\Delta}_m)$. Therefore, we have

$$KL(\theta, \tilde{\theta}_m) \approx \frac{1}{2} E_\theta\Big[\sum_{k=1}^m \frac{\tilde{c}_k D_k(T)}{\lambda(T)} - 1\Big]^2 = O(\bar{\Delta}_m^2), \tag{3.3}$$

Thus the sieve bias essentially depends on the size of the mesh $\bar{\Delta}_m$, and the minimum KL divergence from the sieve space Θ_m to the true value θ is dominated by the bias of the baseline hazard function $\tilde{c}_k D_k(t)/\lambda(t) - 1$.

Similar results hold in our joint modeling setting. In particular, when all other parameters are known except for λ, the KL divergence can be calculated as

$$KL(\lambda, \lambda_m) \approx \frac{1}{2} E_\theta \Big[\log \frac{\lambda(T)}{\lambda_m(T)} - 1 + \frac{\lambda_m(T)}{\lambda(T)} \Big] \approx \frac{1}{2} E_\theta \Big[\frac{\lambda_m(T)}{\lambda(T)} - 1 \Big]^2,$$

so the order of the KL divergence is the same as the order of the quadratic term $E_\theta[\lambda_m(T)/\lambda(T) - 1]^2$.

More generally, given an uncensored datum, we approximate the KL divergence via the Taylor expansion

$$\begin{aligned}
KL(\theta, \theta_m) &= E_\theta[\log \frac{L_i(\theta)}{L_i(\theta_m)}] \\
&= E_\theta[-\frac{\partial}{\partial \theta} \log L_i(\theta)(\theta_m - \theta) \\
&\quad -\frac{1}{2}(\theta_m - \theta)' \frac{\partial^2}{\partial \theta^2} \log L_i(\theta)(\theta_m - \theta) + o(\|\theta_m - \theta\|)^2] \\
&\approx \frac{1}{2} E_\theta \{ (\lambda_m(T) - \lambda(T), \beta_m - \beta, \theta_{m-\{\beta,\lambda\}} - \theta_{-\{\beta,\lambda\}}) \\
&\quad [i^m_{\theta,\theta}](\lambda_m(T) - \lambda(T), \beta_m - \beta, \theta_{m-\{\beta,\lambda\}} - \theta_{-\{\beta,\lambda\}})' \}, \qquad (3.4)
\end{aligned}$$

where $\theta_{-\{\beta,\lambda\}} = (\alpha, \Sigma_b, \sigma_e^2)$ and $i^m_{\theta,\theta}$ denote the sample Fisher information matrix in the sieve space Θ_m evaluated at the true parameters, assumed to be uniformly bounded away from 0 in the neighborhood of θ. (Here only one subject is considered and contributes to $i^m_{\theta,\theta}$.) Furthermore, the sieve bias can be calculated through the minimizer

$$\begin{aligned}
&(\tilde{\lambda}_m(\cdot) - \lambda(\cdot), \tilde{\beta}_m - \beta, \tilde{\alpha}_m - \alpha, \tilde{\Sigma}_{b,m} - \Sigma_b, \tilde{\sigma^2}_m - \sigma^2) \\
&\quad = \arg\inf_{\theta_m \in \Theta_m} E_\theta \{ (\lambda_m(T) - \lambda(T), \beta_m - \beta, \theta_{m-\{\beta,\lambda\}} - \theta_{0-\{\beta,\lambda\}}) \\
&\quad [i^m_{\theta,\theta}](\lambda_m(T) - \lambda(T), \beta_m - \beta, \theta_{m-\{\beta,\lambda\}} - \theta_{0-\{\beta,\lambda\}})' \}. \qquad (3.5)
\end{aligned}$$

Theorem 1 (Sieve bias). *Under the joint modeling setting, for the sieve space $\mathbb{S}_m$ of* (3.1)*, the sieve biases are*

$$\frac{c_k^* D_k(t)}{\lambda(t)} - 1 = O(\bar{\Delta}_m), \quad k = 1, \ldots, m, \qquad (3.6)$$

$$\beta_m^* - \beta = O(\bar{\Delta}_m^{1+d}), \qquad (3.7)$$

$$\theta^*_{m-\{\beta,\lambda\}} - \theta_{-\{\beta,\lambda\}} = O(\bar{\Delta}_m^{1+d}), \qquad (3.8)$$

for each t and

$$KL(\theta, \theta_m^*) = \inf_{\theta_m \in \Theta_m} E_\theta \Big[\log \frac{L_i(\theta)}{L_i(\theta_m)} \Big],$$

$$\approx \frac{1}{2}E_{X,\theta}\Big[\sum_{k=1}^{m}\frac{c_k^*D_k(T)}{\lambda(T)}-1\Big]^2. \tag{3.9}$$

3.2. Strong consistency and convergence rates of the sieve MLEs

As noted, the KL divergence is not a metric and hence the sieve target θ_m^* may not be unique. Further, the sieve MLE $\hat{\theta}_m \in \Theta_m$ involves high-dimensional optimization when m is large and there may be multiple sieve MLEs. Therefore, for a sample of size n, we define the collection of such maximum likelihood estimators in Θ_m as:

$$M_m^n = \Big\{\hat{\theta}_m \in \Theta_m | \mathcal{L}(\hat{\theta}_m) = \sup_{\Theta_m} \mathcal{L}(\theta_m)\Big\}.$$

The notation $a_n \sim b_n$ means that there are constants $0 < K < K' < \infty$ such that $K \leq a_n/b_n \leq K'$ for all n. Following Theorem 1 and related arguments in Geman and Hwang (1982), we now show the strong consistency of the sieve MLE and its rate of convergence in the joint modeling setting.

Theorem 2 (Strong Consistency)**.** *For the likelihood function $\mathcal{L}(\theta)$ as* (2.4) *and a sequence of piecewise constant sieve spaces $\{\mathbb{S}_m\}$ on $[0, K_1 \log n]$ as equation* (3.1)*. If $m < n$ and $m \sim n^{1/3-\varepsilon}$ for some small $\varepsilon > 0$, then*

(i) $\sup_{\hat{\theta}_m \in M_m^n} \rho(\hat{\theta}_m, \theta) \to 0$ *almost surely.*

(ii) *In particular, if $m \sim n^{1/4-\varepsilon}$ for some small $\varepsilon > 0$, then*

$$KL(\theta, M_m^n) = \sup_{\hat{\theta}_m \in M_m^n} KL(\theta, \hat{\theta}_m) = O(n^{-1/4+2\varepsilon}),$$

and any sequence of $\hat{\theta}_m \in M_m^n$ converges to θ at the rate $O(n^{-1/8+\varepsilon})$ almost surely.

This consistency of the sieve MLE utilizes the Borel-Cantelli's Lemma and a crude version of the maximum inequality employed in Geman and Hwang (1982). We note that the rate of convergence $O(n^{-1/8+\varepsilon})$ is too slow to be useful in the derivation of the asymptotic distribution. However, the theorem says that eventually the set of sieve MLE approaches the true value θ. This strong consistency allows us to limit our search of sieve MLEs to a compact set $\mathbb{B}_m \times \mathbb{S}_m$, where $\mathbb{B}_m$ is a bounded subset of $\mathbb{R}^Q$ containing $\theta_{-\{\lambda\}}$.

Remark. The rate of convergence can be significantly improved if a better approximation of $\phi''(t)$ can be obtained (see the Appendix). It might also be possible to improve the rate of convergence by applying inequalities for the KL divergence as in Wong and Shen (1995). We show in Theorem 3 and 4 that indeed the optimal rate of convergence for the parametric component is $\sqrt{n}$.

3.3. Consistency of the sieve MLE in a finer sieve space

Even though strong consistency of the sieve MLE can be obtained from Theorem 2, a finer sieve space and a refined maximum inequality from empirical process theory (Pollard (1984,1990)) are needed to obtain a better rate of convergence by restricting the maximum likelihood estimation on $\mathbb{B}_m \times \mathbb{S}_m$.

Theorem 3. *For $\mathbb{S}_m$ as* (3.1)*, suppose that the sieve target θ_m^* is an interior point of a bounded set $\mathbb{B}_m \times \mathbb{S}_m$. For $m \sim n^p$ and $\varepsilon \le p \le 1-\varepsilon$ for some small $\varepsilon > 0$,*

$$\widehat{KL}(\theta,\theta_m) - KL(\theta,\theta_m) = O_p(n^{-(1-p)/2}(\log n)^{3/2}), \tag{3.10}$$

uniformly for $\theta_m \in \mathbb{B}_m \times \mathbb{S}_m$.

For a fixed m, the sieve MLE is simply an M-estimator that maximizes $-KL(\theta,\theta_m)$ over $\mathbb{B}_m \times \mathbb{S}_m$. The uniform convergence of the maximizing function guarantees the convergence of the M-estimator, provided the maximum point is unique and isolated. When m increases with n at a rate faster than n^{ε}, then $\hat\theta_m$ is consistent.

Corollary 1 (Consistency)**.** *Suppose θ is the isolated unique minimizer of the Kullback-Leibler divergence, that is,*

$$\inf_{\theta^*:\ \rho(\theta,\theta^*)\ge\delta} KL(\theta,\theta^*) > 0.$$

Under the assumptions of Theorem 3*,*

$$\hat\theta_m \to \theta \quad \textit{in probability},$$

and consequently by Theorem 1 *that $\hat\theta_m - \theta_m^* \to 0$ in probability.*

Proof. We have $\widehat{KL}(\theta,\hat\theta_m) \le \widehat{KL}(\theta,\theta_m^*) + o(1)$ by the definition of the sieve MLE for all m and n, and $\widehat{KL}(\theta,\theta_m^*) = KL(\theta,\theta_m^*) + o_p(1)$ by choosing θ_m^* in Theorem 3. Hence $\widehat{KL}(\theta,\hat\theta_m) \le KL(\theta,\theta_m^*) + o_p(1)$, and

$$\begin{aligned} 0 &\le KL(\theta,\hat\theta_m) - KL(\theta,\theta_m^*) \le KL(\theta,\hat\theta_m) - \widehat{KL}(\theta,\hat\theta_m) + o_p(1) \\ &\le \sup_{\theta_m} |KL(\theta,\theta_m) - \widehat{KL}(\theta,\theta_m)| + o_p(1) = o_p(1), \end{aligned}$$

by (3.10). Hence, $KL(\theta,\hat\theta_m) - KL(\theta,\theta_m^*) \to 0$ in probability. Furthermore, Theorem 1 implies $KL(\theta,\theta_m^*) = O(\bar\Delta_m^2) = o(1)$ when $mn^{-\varepsilon} \to +\infty$. Hence, $KL(\theta,\hat\theta_m) = o_p(1)$. By the assumptions, for every $\epsilon > 0$, there exists a number $\eta > 0$ such that $KL(\theta,\theta^*) > \eta$ for every θ^* with $\rho(\theta,\theta^*) \ge \epsilon$. Thus the event $\{\rho(\hat\theta_m,\theta) \ge \epsilon\}$ is contained in the event $\{KL(\theta,\hat\theta_m) > \eta\}$, the probability of which converges to 0.

3.4. Convergence rate of the sieve MLE

We now establish the convergence rate of the sieve estimates for each component of $\theta = (\theta_{-\{\lambda\}}, \lambda)$. We show that the conventional $\sqrt{n}$-rate can be achieved for estimates of the finite dimensional parameters in $\theta_{-\{\lambda\}}$ but not for $\lambda(\cdot)$, the sieve estimates of which converge at the slower rate $n^{(1-p)/2}$ for $p > 1/3$. This slower rate of convergence is due to the unbounded domain $[0, K_1 \log n]$ of the sieve spaces $\mathbb{S}_m$.

Consider the score equations and the sample Fisher information matrix derived from the observed log likelihood at $\theta_m \in \Theta_m$:

$$\begin{aligned} S_n^*(\theta_m) &= \sum_{i=1}^{n} \frac{\partial}{\partial \theta_m} \log L_i(\theta_m), \\ i^m_{\theta_m,\theta_m} &= -\sum_{i=1}^{n} \frac{\partial^2}{\partial \theta_m^2} \log L_i(\theta_m). \end{aligned}$$

Under classical regularity conditions, Corollary 3.1 implies that the score equation evaluated at the sieve MLE can be expanded as

$$\begin{aligned} 0 \equiv S_n^*(\hat{\theta}_m) &= S_n^*(\theta_m^*) - i^m_{\theta_m^*,\theta_m^*}(\hat{\theta}_m - \theta_m^*) + o_p(i^m_{\theta_m^*,\theta_m^*}(\hat{\theta}_m - \theta_m^*)), \\ &\Rightarrow \hat{\theta}_m - \theta_m^* = (i^m_{\theta_m^*,\theta_m^*} + o_p(i^m_{\theta_m^*,\theta_m^*}))^{-1} S_n^*(\theta_m^*). \end{aligned} \tag{3.11}$$

With the chosen sieve space $\mathbb{S}_m$, the scores in (3.11) can be written as

$$\begin{aligned} &S_n^*(\theta_m) = (S_n^*(\alpha|\theta_m^*), S_n^*(\Sigma_b|\theta_m^*), S_n^*(\sigma_e^2|\theta_m^*), S_n^*(\lambda_m|\theta_m^*), S_n^*(\beta|\theta_m^*))^T, \\ &\text{where } S_n^*(\lambda_m|\theta_m^*) = (S_n^*(c_1|\theta_m^*), \ldots, S_n^*(c_m|\theta_m^*)), \\ &\qquad S_n^*(\Sigma_b|\theta_m^*) = (S_n^*(u_1|\theta_m^*), \ldots, S_n^*(u_{q(q+1)/2}|\theta_m^*)), \end{aligned}$$

and $u = (u_1, \ldots, u_{q(q+1)/2})^T$ is the parameter involved in Σ_b. More specifically, letting $E_i^*[\cdot]$ denote the conditional expectation w.r.t. the posterior density $h(\mathbf{b}_i|O_i, \theta_m^*)$ with θ_m^* in the sieve space Θ_m, we have

$$\begin{aligned} S_n^*(\alpha|\theta_m^*) &= \Sigma_b^{-1} \sum_{i=1}^{n} E_i^*(\mathbf{b}_i) - n\Sigma_b^{-1}\alpha, \\ S_n^*(u|\theta_m^*) &= -\frac{n}{2}\mathrm{tr}(\Sigma_b^{-1}\frac{\partial \Sigma_b}{\partial u}) + \frac{1}{2}\sum_{i=1}^{n} E_i^*[(\mathbf{b}_i - \alpha)^T \Sigma_b^{-1} \frac{\partial \Sigma_b}{\partial u} \Sigma_b^{-1}(\mathbf{b}_i - \alpha)], \\ S_n^*(\sigma_e^2|\theta_m^*) &= -\frac{N}{2}\sigma_e^2 + \sum_{i=1}^{n}\sum_{j=1}^{m_i} \frac{E_i^*[(W_{ij} - X_{ij})^2]}{2\sigma_e^4}, \\ S_n^*(c_k|\theta_m^*) &= \sum_{i=1}^{n} \left\{ \frac{\Delta_i D_k(V_i)}{c_k} - E_i^*\Big[\int_0^{V_i} D_k(s) \exp\{\beta X_i(s)\} \mathrm{d}s\Big] \right\}, \end{aligned} \tag{3.12}$$

$$k = 1, \ldots, m,$$

$$S_n^*(\beta|\theta_m^*) = \sum_{i=1}^{n} \Big\{ \Delta_i E_i^*[X_i(V_i)] - \sum_{k=1}^{m} c_j E_i^* \Big[\int_0^{V_i} D_k(s) X_i(s) \exp\{\beta X_i(s)\} \mathrm{d}s \Big] \Big\}. \tag{3.13}$$

Note that the expectation of each score function is 0 only under θ_m^*, but not under θ. That is: $E_\theta[S_n^*(\theta_m^*)] \neq 0$, when the sieve does not contain the true parameter θ. The difficulty is the fact that the score S_n^* is derived at θ_m^*, while the expectation is taken with respect to θ. In order to derive the asymptotic properties of the sieve MLE $\hat{\theta}_m$ based on these equations, we need to show that $E_\theta[S_n^*(\theta_m^*)]$ approaches 0 at a fast enough rate for each parameter. The convergence rates for each component of $E_\theta(S_n^*(\theta_m^*))$ to 0, and the componentwise bounds of variances for the $S_n^*(\theta_m^*)$ are given as follows.

Theorem 4. *Suppose the true baseline function satisfies $\lambda(t) \in C^{(1,d)}$ with $d \geq 1/2$ and $\theta_{-\{\lambda\}}$ is an interior point in $\Theta_{-\{\lambda\}}$. For $\mathbb{S}_m$ as* (3.1)*, if $m \sim n^p$ with $1/3 < p < 1$, then for every finite-dimensional parameter ω in $(\alpha, \Sigma_b, \sigma_e^2, \beta)$ and all c_k,*

(i) $E_\theta[n^{-1/2} S_n^*(\omega|\theta_m^*)] = O(n^{1/2} \bar{\Delta}_m^{1+d}) = O(n^{1/2-p(1+d)} (\log n)^{1+d}),$

$E_\theta[n^{-(1-p)/2} S_n^*(c_k|\theta_m^*)] = O(n^{(1-p)/2} \bar{\Delta}_m) = O(n^{(1-3p)/2} \log n).$

(ii) $\mathrm{Var}_\theta[n^{-1/2} S_n^*(\omega|\theta_m^*)] = O(1),$

$\mathrm{Var}_\theta[n^{-(1-p)/2} S_n^*(c_k|\theta_m^*)] = O(1).$

(iii) *Hence,* $n^{1/2}(\hat{\theta}_{m-\{\lambda\}} - \theta_{-\{\lambda\}}) \to O_p(1),$

$n^{(1-p)/2}(\hat{\lambda}_m(t) - \lambda(t)) \to O_P(1),$

for a fixed time point t.

Remark. Heuristically, for each c_k there are only n/m subjects falling in $(t_{k-1}, t_k]$ to affect c_k, so the variance of effective subjects should be $(n/m)^{-1} = n^{-(1-p)}$.

3.5. Implementation of the method of sieves

The practical estimation procedure of the sieve MLE is implemented through an ECM algorithm. In the E-step, we calculate the conditional expectations in the scores equations using Monte Carlo expectations. The random effects, $\boldsymbol{b}_i$, are generated from the posterior density (2.5). In the M-step, we solve the scores (3.12) and (3.13) given the conditional expectations and the parameter estimates from the previous iterations. The E-step and M-step are then iterated until the parameter estimates converge. Finally, the observed Fisher information matrix is inverted to provide a standard error estimation.

Remarks:

1. Smaller Monte Carlo samples are used at the beginning of the iterations to speed up the computation and larger samples are used at the later iterations to protect the accuracy of the integrations. Finally, the observed Fisher information matrix is inverted to provide a standard error estimation.

2. The initial estimates for ECM algorithm can be selected subjectively. We use the two-stage procedure, described in the simulation section.

3. Theoretically, the number of sieves, m, should grow slowly as the sample size increases, while a larger m leads to a smaller sieve bias. The best rate, according to Theorem 4, is to allow $n^{1/3}m^{-1}$ to approach zero as slowly as possible. In practice, we recommend partitioning the sieve space to uniformly place uncensored even-times in each piece and trying a few different m until a clear shape emerges for the baseline hazard function. A good starting point is $m = Cn^{1/3}$ with some initial choices of C. If the estimated baseline hazard function is roughly a constant, then a smaller number of sieves need to be considered. A larger number of sieves is often necessary. The choice of m is a challenging model selection problem that deserves futher research.

4. Simulation Results

We set up three simulations to demonstrate the proposed method of sieves and compare it with the NPMLE. The results based on 100 Monte Carlo samples are reported for four different methods. The "ideal" method used the simulated longitudinal profile $X(t)$, unobservable in practice, and serves as the benchmark. In the two-stage ("TS") procedure, the longitudinal component was modeled through a linear mixed-effects model at the first stage and the resulting prediction for the longitudinal covariate was then employed at the next stage to model the survival component through the partial likelihood method for the Cox model. This two-stage procedure was included to demonstrate the bias when survival and longitudinal components are modelled marginally. The NPMLE is similar to the proposal in Wulfsohn and Tsiatis (1997), but we used the two-stage estimates as the initial values and Monte Carlo integration in the E-step. The method of sieves with different numbers of sieves and different partitions on the time interval was investigated. The time interval was partitioned into m pieces with equal numbers of event-times in each piece. Here, "SMLE1" used uncensored event times for the partition rule; while "SMLE2" used all observed times (including censored times) to partition the sieve space. In addition to reporting the Monte Carlo standard deviation (SD), bias, and mean square (MSE) for each procedure, we

Table 1. Estimation for β in Simulation 1: Constant Baseline with Exponential Censoring. Five methods are compared with IDEAL using the simulated, but unobservable, profile $X(t)$; TS denotes the two-stage method; NPMLE denotes the nonparametric maximum likelihood estimate; and SMLE1 and SMLE2 are two different partition rules for the sieves method.

$\beta = -1$	IDEAL	TS	NPMLE*	**SMLE1**			SMLE2		
				m=3	$m = 5$	$m = 7$	$m = 3$	$m = 5$	$m = 7$
Bias	-0.008	0.196	0.095/0.059	-0.009	-0.035	-0.044	-0.006	-0.024	-0.053
SD	0.180	0.230	0.285/0.291	0.170	0.208	0.257	0.207	0.248	0.287
SE	—	—	—/—	0.181	0.205	0.243	0.203	0.229	0.270
MSE	0.033	0.091	0.091/0.089	0.029	0.044	0.068	0.043	0.062	0.085

*: The first column uses the same rule as the other procedures in this table, with 1,000 Monte Carlo samples to do integration and the procedure stops if the relative difference is less than 0.01; the second column uses 10,000 Monte Carlo samples with a stopping rule of 0.001.

also report the standard error (SE) of β for the sieves estimates using the sample Fisher information matrix. The results reported in Tables 1, 3, and 5 confirm that these SE estimates are close to the SD, providing evidence of the advantage of the method of sieves to produce reliable SE estimates.

The sample sizes in the three simulations were 100, 100, and 400, respectively. The longitudinal measurements for the ith subject were uniformly placed at n_i randomly selected time points on [0,5], where n_i was randomly generated from 1 to 11. (All subjects were measured at time 0.) These time points were further truncated by the observed event-time $V_i = \min(T_i, C_i)$ and this resulted in an average of four longitudinal measurements per subject. Observed individual longitudinal profiles followed $W_i(t_{ij}) = b_{i0} + b_{i1}t_{ij} + e_{ij}$, where $[b_{i0}, b_{i1}]$ were independently normal with mean α and covariance $(\sigma_{11}, \sigma_{12}, \sigma_{22}) = (1, -0.3, 0.2)$. The stochastic noise e_{ij} was independent of the random effects and normal with mean 0 and variance 0.6. In the three simulations, the means were $\alpha = (5, -2)', (5, -1.5)'$, and $(4, -1)'$, respectively. In addition, the survival times were generated from the Cox model with $X(t) = b_{i0} + b_{i1}t$, $\beta = -1$ and $\lambda(t) = 1, \quad 1, \quad \exp(-4 + t)$, respectively, in the three simulations. Observed event-times were subject to random censoring with three distributions: (i) exponential distribution with mean 10, which led to 14% censoring rate; (ii) uniform censoring distribution on $[3, 7]$, which led to 16% censoring rate; (iii) $\min(5, U_i)$, where U_i was uniformly distributed on $[3.7, 5.7]$, which led to 46% censoring rate.

The true baseline hazard $\lambda(t)$ was constant in the first two simulations, so there was no bias in the sieve estimation, and the variance expected to decrease as the number of sieves decreased. This was confirmed in Tables 1 and 3. The choice of the number and location of sieves depended on the data, similar to the choice of smoothing parameters in nonparametric smoothing methods. Theoretical results provide the rate for the number of sieves, but the practical choice under finite

Table 2. Estimations for Longitudinal Components in Simulation 1: Constant Baseline with Exponential Censoring. The methods, TS, NPMLE, SMLE1, and SMLE2, as described in Table 1, are compared.

	$\alpha = E[(b_{i0}, b_{i1})] = [5, -2]$	$\Sigma_b = (\sigma_{11}, \sigma_{12}, \sigma_{13}) = (1, -0.3, 0.2)$	$\sigma_e^2 = 0.6$
TS	[4.971(0.114), -1.922(0.073)]	[1.001(0.178), -0.322(0.095), 0.212(0.067)]	0.605(0.066)
NPMLE	[4.998(0.113), -1.988(0.070)]	[0.992(0.163), -0.305(0.079), 0.202(0.058)]	0.598(0.063)
$\mathbf{m = 3}$ **S1**	[4.999(0.113), -2.001(0.070)]	[0.991(0.177), -0.307(0.083), 0.209(0.057)]	0.601(0.064)
$m = 7$ S1	[4.999(0.113), -2.001(0.071)]	[0.990(0.176), -0.305(0.082), 0.208(0.057)]	0.601(0.064)
$m = 3$ S2	[4.999(0.113), -2.000(0.070)]	[0.992(0.177), -0.306(0.083), 0.207(0.056)]	0.601(0.063)
$m = 7$ S2	[4.999(0.113), -2.001(0.072)]	[0.990(0.178), -0.305(0.083), 0.207(0.058)]	0.601(0.063)

S1=SMLE1; S2=SMLE2.

Table 3. Estimation for β in Simulation 2: Constant Baseline with Uniform Censoring. The methods, IDEAL, TS, NPMLE, SMLE1, and SMLE2, as described in Table 1, are compared.

$\beta = -1$	IDEAL	TS	NPMLE*	SMLE1			**SMLE2**		
				$m = 3$	$m = 5$	$m = 7$	$\mathbf{m = 3}$	$m = 5$	$m = 7$
Bias	-0.014	0.181	0.101/0.026	-0.040	-0.048	-0.061	-0.019	-0.044	-0.065
SD	0.136	0.171	0.284/0.221	0.172	0.195	0.215	0.164	0.179	0.200
SE	—	—	— /—	0.168	0.183	0.209	0.158	0.175	0.204
MSE	0.019	0.062	0.091/0.050	0.031	0.040	0.050	0.027	0.034	0.044

*: The first column uses the same rule as the other procedures in this table, with 1,000 Monte Carlo samples to do integration and the procedure stops if the relative difference is less than 0.01; the second column uses 10,000 Monte Carlo samples with a stopping rule of 0.001.

Table 4. Estimations for Longitudinal Components in Simulation 2: Constant Baseline with Uniform Censoring. The methods, TS, NPMLE, SMLE1, and SMLE2, as described in Table 1, are compared.

	$\alpha = E[(b_{i0}, b_{i1})] = [5, -1.5]$	$\Sigma_b = (\sigma_{11}, \sigma_{12}, \sigma_{13}) = (1, -0.3, 0.2)$	$\sigma_e^2 = 0.6$
TS	[4.973(0.110), -1.438(0.060)]	[1.005(0.186), -0.311(0.087), 0.196(0.053)]	0.613(0.056)
NPMLE	[5.007(0.109), -1.499(0.054)]	[0.988(0.182), -0.299(0.079), 0.197(0.045)]	0.600(0.053)
$m = 3$ S1	[5.008(0.109), -1.510(0.057)]	[0.995(0.183), -0.302(0.081), 0.200(0.047)]	0.608(0.054)
$m = 7$ S1	[5.008(0.109), -1.512(0.057)]	[0.994(0.182), -0.302(0.080), 0.201(0.047)]	0.607(0.054)
$\mathbf{m = 3}$ **S2**	[5.007(0.109), -1.509(0.057)]	[0.995(0.183), -0.301(0.081), 0.200(0.048)]	0.607(0.054)
$m = 7$ S2	[5.008(0.109), -1.511(0.057)]	[0.994(0.182), -0.301(0.081), 0.200(0.048)]	0.608(0.054)

S1=SMLE1; S2=SMLE2

sample size needs further investigation. Different values for the number of sieve spaces, m, were tried, though only $m = 3, 5$, and 7 are presented here to illustrate the pattern. In both tables, the SD of the sieve estimator is well estimated by the SE and, as expected, both SD and SE increasd as the number of sieves increased. The sieve approach was computationally more stable than the NPMLE due to the reduced parameter space, and had the smallest Mean Square Error (MSE). Furthermore, different censoring patterns were associated with different better sieve choices. In the first simulation, when fewer subjects were censored at the

Table 5. Estimation for β in Simulation 3: Gompertz Baseline with Uniform Censoring. The methods, IDEAL, TS, NPMLE, and SMLE1, as described in Table 1, are compared.

$\beta = -1$	IDEAL	TS	NPMLE	SMLE1				
				$m = 7$	$m = 13$	$m = 19$	$m = 30$	$\mathbf{m = 40}$
Bias	0.008	0.182	0.021	-0.153	-0.078	0.049	-0.027	-0.017
SD	0.082	0.085	0.124	0.146	0.141	0.138	0.136	0.125
SE	—	—	—	0.124	0.125	0.124	0.124	0.123
MSE	0.007	0.040	0.016	0.045	0.026	0.021	0.019	0.016

Table 6. Estimations for Longitudinal Components in Simulation 3: Gompertz Baseline with Uniform Censoring. The methods, TS, NPMLE, and SMLE1, as described in Table 1, are compared.

	$\alpha = E[(b_{i0}, b_{i1})] = [4, -1]$	$\Sigma_b = (\sigma_{11}, \sigma_{12}, \sigma_{13}) = (1, -0.3, 0.2)$	$\sigma_e^2 = 0.6$
TS	[4.001(0.054), -0.982(0.027)]	[0.973(0.095), -0.278(0.040), 0.177(0.021)]	0.602(0.025)
NPMLE	[4.027(0.055), -1.016(0.028)]	[0.973(0.096), -0.277(0.039), 0.181(0.019)]	0.598(0.024)
$m = 7$ S1	[4.032(0.054), -1.021(0.026)]	[0.966(0.093), -0.277(0.039), 0.181(0.020)]	0.602(0.025)
$\mathbf{m = 40}$ S1	[4.026(0.054), -1.015(0.026)]	[0.971(0.093), -0.279(0.040), 0.183(0.020)]	0.599(0.025)

S1=SMLE1.

end of the study, "SMLE1", which used only uncensored event-times to partition the sieve space, performed better than "SMLE2". This was reversed in the second simulation, which had heavy censoring at the tail and partitioning that used all observed times (including censored ones) as in "SMLE2" was preferred. Generally speaking, more uncensored events at the tail provides more information in estimating baseline hazards in the tail, and lead to more accurate estimate for the regression coefficient β in the Cox model. The results for the longitudinal parameters were similar for all methods, so only $m = 3$ and 7 for the method of sieves are presented in the Tables 2 and 4.

The baseline hazard function in the third simulation followed a Gompertz distribution function. The baseline hazard then increased at an exponential rate, so we expect the empirical bias of the sieve estimates to decrease as the number of sieves m increases. This is reflected in the simulation results reported in Tables 5 and 6. Only "SMLE1" with $m = 7, 13, 19, 30$, and 40 are presented to illustrate the pattern of the estimators. Note that, for this setting with nonconstant hazard, the sieve bias was nonzero and only negligible asymptotically. Hence for a small number of sieves, we observed a notable difference between the SDs and SEs. However, as the sieve biases decreased with increasing numbers of sieves, the gap between the SD and SE got smaller. By using only one-fifth of the parameters to model the baseline hazard function, the method of sieves with $m = 40$ yielded much smaller bias than the NPMLE. Even at the cost of increasing variance, the MSEs of the method of sieves were still comparable with those from the NPMLE. Moreover the method of sieves was much more stable, and had good standard error estimates. The longitudinal parameters were also well estimated for both NPMLE and sieve estimates (for all choices of m). In the left panel of Figure 1, we show the averaged estimates of the cumulative hazard functions. One can see that SMLEs with much fewer number of parameters to model the cumulative hazard function performed as well as NPMLE. The bias of the estimated hazard functions, shown in the right panel of Figure 1, was reduced by increasing m as expected.

Remark.

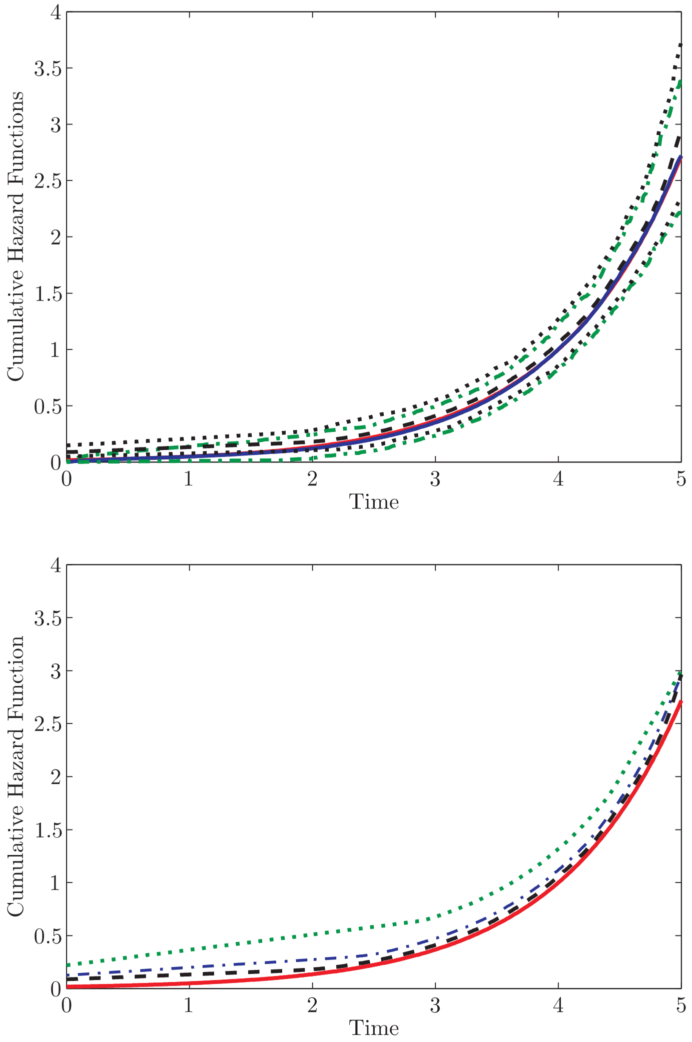
1. The NPMLE is more sensitive to the MC integration random errors than the sieve estimates. When the algorithm was stopped when the relative differences was less than 0.01 and the MC sample size was 1,000, the NPMLE showed bigger bias than the SMLEs. This was improved by increasing the accuracy of Monte Carlo integration to sample size 10^4, using the stricter stopping rule of 0.001. This reduced the biases of the NPMLE to the same range of the SMLEs. This observation confirms that the NPMLE can be trapped at local maxima much more frequently than the SMLEs.
2. The idea of the method of sieves is to pool subjects into each sieve piece to stabilize the estimation of the hazard function. Ironically, the NPMLE is more stable when the censoring rate is higher, since fewer parameters are involved in estimating hazard function. This explains why, in the setting of Simulation 3, the NPMLE performed better than in the first two simulations.

5. Case Study: Primary Biliary Cirrhosis (PBC) Data

PBC is a rare but fatal chronic liver disease. The Mayo Clinic conducted a double-blinded randomized trial from January, 1974 to May, 1984 to compare the drug D-penicillamine with a placebo. The data set we used is from a follow-up longitudinal study that had follow-up laboratory data for each patient in the original study. In the primary PBC data set, 17 baseline measurements for 312 patients were recorded, as well as their survival information. In the survival analysis literature using the PBC data (Murtaugh et al. (1994); Fleming and Harrington (1991)), only age, albumin, bilirubin, edema score, and prothrombin time have been detected as significant predictors. There was no significant treatment effect. For illustration purposes, we consider only the most significant longitudinal covariate, albumin.

We excluded four subjects with extreme measurement values as outliers. The number of measurements per patient ranges from 1 to 16 with a median of 5. The profiles of ten randomly selected subjects are plotted in Figure 2, where the observed albumin are connected by solid lines and death (censoring

Figure 1. Cumulative Hazard Functions in Simulation 3.



Left panel: Red: true cumulative hazard function; solid green curve: NPMLE, dot-dash green curve: 95% Confidence Band for NPMLE; dash black curve: SMLE (with $m = 40$), dotted black curve: 95% Confidence Band for SMLE;
Right panel: SMLE of cumulative hazard functions with $m = 7$(dotted green line), $m = 19$(dot-dash blue line) and $m = 40$(dash black line) as well as the true cumulative hazard function (solid red line).

time) is indicated by a cross (diamond) on the x-axis. Albumin was centered by subtracting its mean to avoid numerical instability. The centered albumin was modeled by a linear mixed effect model as described in Section 2. Observed survival times were from 0.12 to 12 years. About 55% of the patients were right censored or still alive at the end of the study.

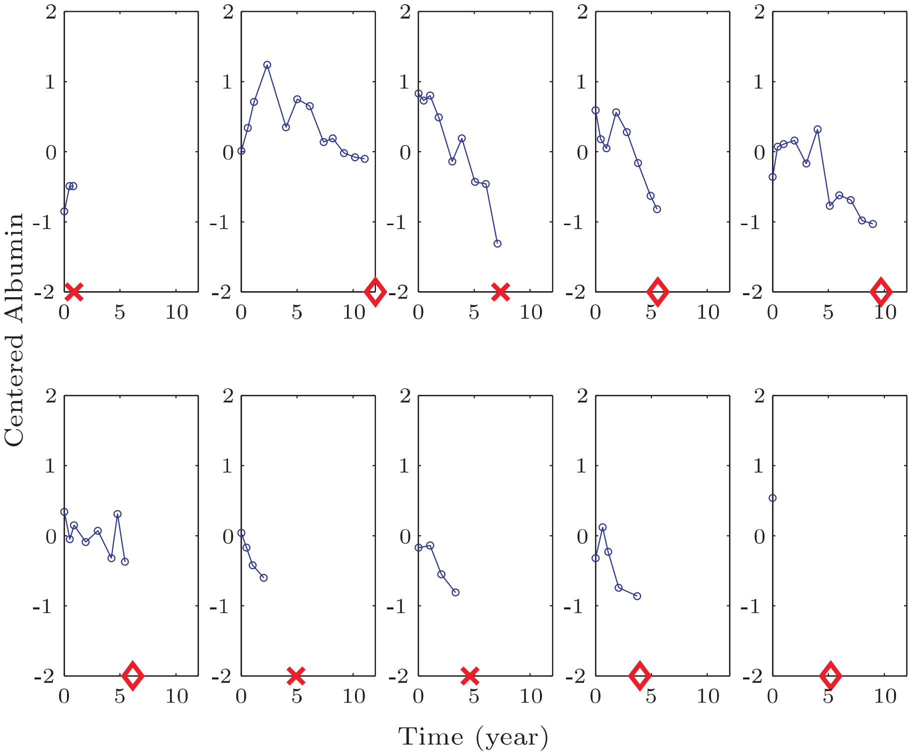


Figure 2. 10 Random Selected Observed Albumin Profiles.

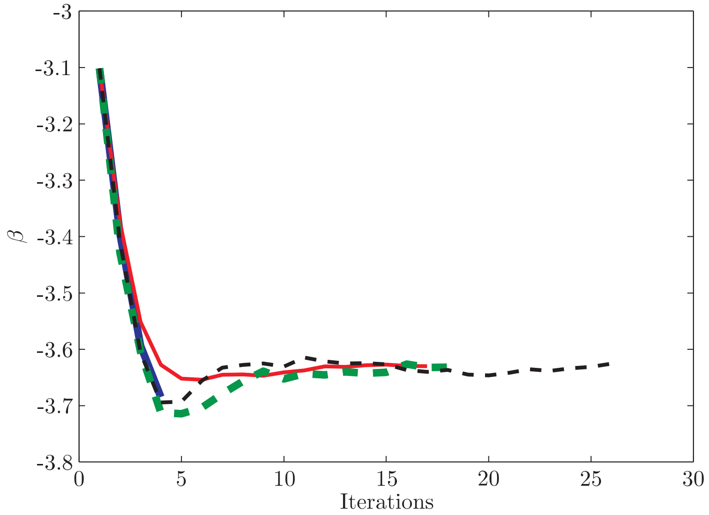
Table 7. Estimation of Survival Regression Coefficient β for four methods: LVCF denotes the last-value-carry-forward method, and TS, NPMLE, and SMLE1 are described in Table 1.

	LVCF	TS	NPMLE*	SMLE1					
				$m=5$	$m=6$	$m=7$	$m=8$	$m=15$	$m=30$
$\hat{\beta}$	-2.70	-3.10	-3.70/-3.63	-3.69	-3.66	-3.64	-3.63	-3.74	-3.74
SE	—	—	—/—	0.33	0.33	0.32	0.32	0.34	0.34

*: The first column uses the same rule as the other procedures in this table, with 1,000 Monte Carlo samples to do integration and the procedure stops if the relative difference is less than 0.01; the second column uses 10,000 Monte Carlo samples with a stopping rule of 0.001.

We compared the sieve estimation with the LVCF (last value carry forward) procedure, two-stage estimation, and the NPMLE. Both NPMLE and sieve estimation used two-stage estimates as the initial values. Results for the regression coefficients are reported in Table 7 with SE (standard errors) for the sieve estimates. The LVCF and two-stage procedures had substantial smaller covariate effects compared the sieve estimates. This was also observed in the simulations where the LVCF and two-stage estimates were biased toward 0. Further, Figure 3 reveals that the NPMLE led to different estimates, -3.70 and -3.63, under

Figure 3. Iterations of NPMLE and SMLE.



Solid lines are iterations for NPMLE and dash lines are for SMLE when $m = 8$. The thicker (solid blue and dash green) lines stem from 1,000 Monte Carlo samples for integration and the convergence criteria of a relative difference of 0.01; the thinner lines stem from 10,000 Monte Carlo samples for integration with iteration stopped when the relative difference is less than 0.001.

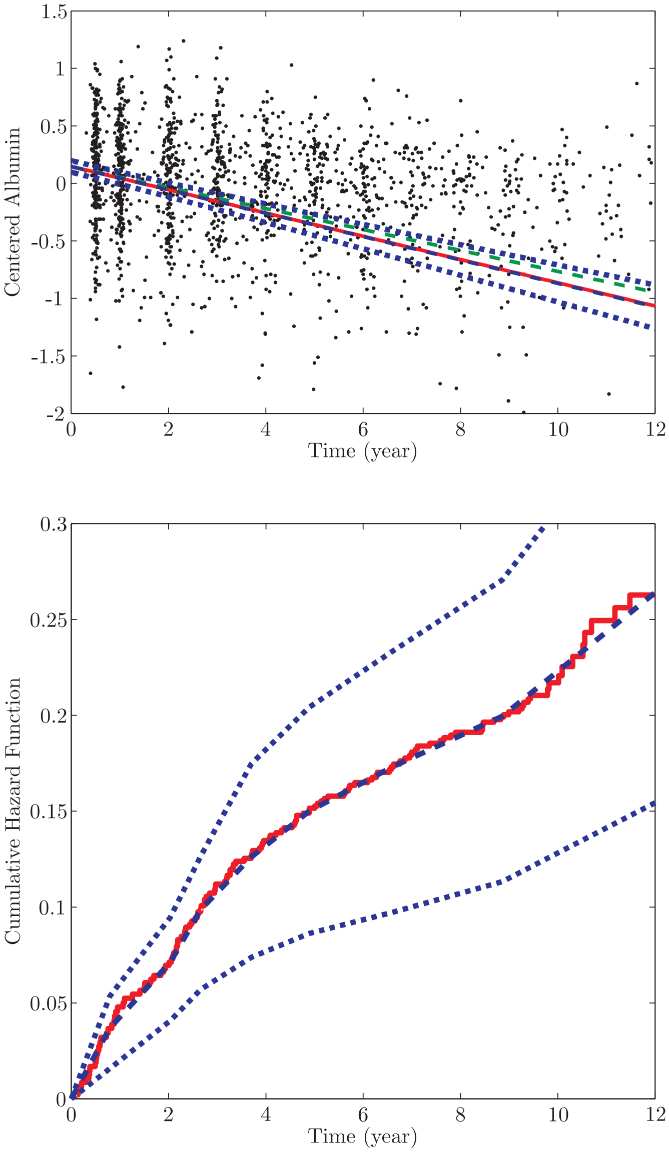
convergence criteria 0.01 and 0.001, respectively, but the SMLE led to essentially the same estimate, -3.64. This illustrates the vulnerability of the NPMLE to get trapped in a local maximum. On the other hand, the NPMLE took fewer iterations than the SMLE to converge.

For the method of sieves, we used uncensored event-time to partition the sieve space (SMLE1) because less censoring happened at the end of the study, and the results were similar if one adopted SMLE2. Several different choices of sieves were investigated. The results were similar for the longitudinal component and we report the longitudinal mean fit and cumulative hazard function estimation in Figure 4 with $m = 8$. We conclude that higher albumin is significantly associated with lower mortality rate, with the mortality rate decrease $\exp^{-3.63} = 0.027$ (95% CI $[0.014, 0.050]$) when albumin increases one unit.

6. Discussion

We propose the method of sieves for practical applications of the joint model and established asymptotic consistency and normality for the proposed SMLE. Compared with the traditional NPMLE, the proposed sieve estimator is computationally stable, provides a consistent baseline hazard estimation, and leads to good standard error estimation.

Figure 4. Longitudinal Mean and Cumulative Hazard Function Estimations.



Left panel: Observed longitudinal measurements (dots) with estimated longitudinal means by two-stage method(green dot-dash line), NPMLE (solid red line) and SMLE for $m = 8$ (thick blue dash line) with the 95% confidence band (blue dotted line).

Right panel: Estimated cumulative hazard functions from NPMLE (solid red line) and SMLE for $m = 8$ (thick blue dash line) with the 95% confidence band (blue dotted line).

Although we have focused on one time-dependent covariate process, the model can be extended to include additional time-independent or time-dependent covariates at the cost of additional computation. The sieve score equations need to be derived for the additional covariates in the same spirit of equations (3.12) and (3.13). The expectations may involve more random coefficients if additional time-dependent covariates are used. In summary, this generation does not affect the theoretical development much, but the notation is messier. In principle, the method of sieves can be applied to any semiparametric survival model, but the justification of the asymptotic properties depends on the characteristics of the likelihood functions (or loss functions) and the choice of sieve spaces. Therefore, the derivations and proofs often vary case by case when the method of sieves is used.

Acknowledgement

This work of all three authors is partially supported by an NIH grant 1R01AG025218-01.

Appendix

Notation List

- Survival observations for the ith subject:
  T_i: event-time; C_i: Censoring time.
  $V_i = \min(T_i, C_i)$: observed event-time; $\Delta_i = 1_{(T_i \leq C_i)}$: censoring indicator;
- Longitudinal observations for the ith subject:
  $\mathbf{W}_i$: Observed longitudinal measurements
  $X_i(t; \mathbf{b}_i) = X_i(t)$: unobserved longitudinal process
  $\mathbf{b}_i$: q-dimensional vector to model individual random effects
  n_i: the number of observed longitudinal measurements.
- $O_i = (V_i, \Delta_i, \mathbf{W}_i, \mathbf{t}_i, n_i)$: Observed data for the ith subject
- $\Theta_{-\{\lambda\}} \subset \mathbb{R}^Q$: finite dimensional parameter space excluding the baseline hazard function
- $\Theta = \Theta_{-\{\lambda\}} \times \mathbb{S}$: true parameter space
- $\mathbb{S}_m$: sieve space for baseline hazard functions
- $\Theta_m = \Theta_{-\{\lambda\}} \times \mathbb{S}_m$: sieve parameter space
- $\theta = \{\beta, \lambda(\cdot), \alpha, \Sigma_b, \sigma_e^2\} = \{\theta_{-\{\lambda\}}, \lambda(\cdot)\}$: true parameters
- $\theta_m^* = \{\beta_m^*, \lambda_m^*(\cdot), \alpha_m^*, \Sigma_{b,m}^*, \sigma_{e,m}^{2*}\}$: target of the sieve estimate which is the projection of the true parameters onto the sieve space $\mathbb{S}_m$
- $\hat{\theta}_m$: sieve MLE in Θ_m
- $\theta_m = \{\beta_m, \lambda_m(\cdot), \alpha_m, \Sigma_{b,m}, \sigma_{e,m}^2\}$: an arbitrary parameter in the sieve space Θ_m
- $\tilde{\theta}_m$: the minimizer of the quadratic approximation of KL divergence

- $i^m_{\theta,\theta}$: the sample Fisher information matrix in the sieve space Θ_m evaluated at the true parameters
- $i^m_{\theta^*_m,\theta^*_m}$: the sample Fisher information matrix in the sieve space Θ_m evaluated at the sieve target θ^*_m.
- $i_{\theta,\theta}$: the sample Fisher information matrix in the space of NPMLE evaluated at the true parameter θ.

Proof of Theorem 1. With the approximation (3.4) of KL divergence in the joint modeling setting, we have

$$\begin{aligned}\frac{\partial}{\partial c^*_k} KL(\theta,\theta^*) &= \frac{\partial}{\partial c^*_k} E_\theta\Big[\log\frac{L_i(\theta)}{L_i(\theta^*)}\Big];\\ &\approx E_{X,\theta}\{(D_k(t),0,0,\ldots,0)[i^m_{\theta,\theta}](\lambda^*_m(t)-\lambda(t),\beta^*_m-\beta,\theta^*_{m-\{\beta,\lambda\}}-\theta_{-\{\beta,\lambda\}})'\}.\end{aligned}$$

The sample Fisher information $i^m_{\theta,\theta}$ in the sieve space Θ_m, contributed by subject i, is

$$\begin{aligned} i_{\beta,\beta} &= E_{i,\theta_m}\Big(\int_0^{V_i}\sum_{j=1}^m c_j D_j(t) X_i^2(t)\exp\{\beta X_i(t)\}\mathrm{d}t\Big)\\ &\quad -\mathrm{Var}_{i,\theta_m}\Big(\Delta_i X_i(V_i) - \int_0^{V_i}\sum_{j=1}^m c_j D_j(t) X_i(t)\exp\{\beta X_i(t)\}\mathrm{d}t\Big),\\ i_{\beta,c_k} &= E_{i,\theta_m}\Big(\int_0^{V_i} D_k(t) X_i(t)\exp\{\beta X_i(t)\}\mathrm{d}t\Big)\\ &\quad +\mathrm{Cov}_{i,\theta_m}\Big(\int_0^{V_i} D_k(t)\exp\{\beta X_i(t)\}\mathrm{d}t,\\ &\qquad\qquad \Delta_i X_i(V_i) - \int_0^{V_i}\sum_{j=1}^m c_j D_j(t) X_i(t)\exp\{\beta X_i(t)\}\mathrm{d}t\Big),\\ i_{\lambda_m,\lambda_m} &= \mathrm{Diag}\Big(\frac{\Delta_i D_k(V_i)}{c_k^2}\Big) - \Big(a_i(h,k)\Big), \end{aligned}$$

where $a_i(h,k) = \mathrm{Cov}_{i,\theta_m}\Big(\int_0^{V_i} D_h(t)\exp\{\beta X_i(t)\}\mathrm{d}t, \int_0^{V_i} D_k(t)\exp\{\beta X_i(t)\}\mathrm{d}t\Big)$. Hence we have

$$\begin{aligned}\tilde{c}_k = \frac{1}{E_\theta\,(i_{c_k,c_k}D_k(t))}\Big\{& E_\theta\,(i_{c_k,c_k}\lambda(t)D_k(t))\\ & -E_{X,\theta}\Big(i_{c_k,\theta_{-\{c_k\}}}D_k(t)\Big)\,(\tilde{\beta}_m-\beta,\tilde{\theta}_{m-\{\beta,\lambda\}}-\theta_{-\{\beta,\lambda\}})'\Big\},\quad k=1,\ldots,m.\end{aligned}$$

Similarly we can derive the remaining equations for each component of $(\tilde{\beta}_m-\beta,\tilde{\theta}_{m-\{\beta,\lambda\}}-\theta_{-\{\beta,\lambda\}})$. By solving them, we find the order of sieve bias:

$$\frac{\tilde{c}_{k,m}D_k(t)}{\lambda(t)} - 1 = O(\bar{\Delta}_m),\quad k=1,\ldots,m,$$

$$\begin{aligned}\tilde{\beta}_m - \beta &= O(\bar{\Delta}_m^{1+d}),\\ \tilde{\theta}_{m-\{\beta,\lambda\}} - \theta_{-\{\beta,\lambda\}} &= O(\bar{\Delta}_m^{1+d}),\end{aligned}$$

and approximate KL by

$$\begin{aligned}KL(\theta,\tilde{\theta}_m) &= \inf_{\mathbb{S}_m\times\mathbb{R}^d} E_\theta\Big(\log\frac{L_i(\theta)}{L_i(\theta_m)}\Big),\\ &\approx \frac{1}{2}E_{X,\theta}\Big(\sum_{k=1}^m \frac{\tilde{c}_{k,m}D_k(T)}{\lambda(T)}-1\Big)^2.\end{aligned}$$

Now we find the bound on the difference between θ_m^* and $\tilde{\theta}_m$, componentwise. Note that $KL(\theta,\theta_m^*) = KL(\theta,\tilde{\theta}_m) + \frac{\mathrm{d}}{\mathrm{d}\theta_m}KL(\theta,\theta_m)|_{\theta_m=\theta^*}(\theta_m^*-\tilde{\theta}_m)$, where $\theta^* = \tilde{\theta}_m + t(\theta_m^*-\tilde{\theta}_m)$ with $t\in[0,1]$. Hence,

$$\begin{aligned}\Big|\theta_m^*-\tilde{\theta}_m\Big| &= \Big|\frac{\mathrm{d}}{\mathrm{d}\theta_m}KL(\theta,\theta_m)|_{\theta_m=\theta^*}\Big|^{-1}\Big|KL(\theta,\theta_m^*)-KL(\theta,\tilde{\theta}_m)\Big|\\ &\le a\Big(|KL(\theta,\theta_m^*)|+|KL(\theta,\tilde{\theta}_m)|\Big)\\ &\le 2a|KL(\theta,\tilde{\theta}_m)| = O(\bar{\Delta}_m^2),\\ \Rightarrow \Big|\theta_m^*-\theta\Big| &\le \Big|\theta_m^*-\tilde{\theta}_m\Big| + \Big|\tilde{\theta}_m-\theta\Big| = O(\tilde{\theta}_m-\theta),\end{aligned}$$

when $d<1$ and we assume $\frac{\mathrm{d}}{\mathrm{d}\theta_m}KL(\theta,\theta_m)$ is uniformly bounded away from 0. This concludes the proof of Theorem 1.

Proof of Theorem 2. The proof follows the arguments of Theorem 2 in Geman and Hwang (1982). We first check the two assumptions used there:

C1 For every m and n, the set of sieve MLE M_m^n is almost surely nonempty.

C2 (a) If for some sequence $\theta_m\in\Theta_m$, $KL(\theta,\theta_m)\to 0$, then $\theta_m\to\theta$.
(b) There is a sequence $\theta_m'\in\Theta_m$ such that $KL(\theta,\theta_m')\to 0$.

Note that the maximum of the function $\log c_k^* - c_k^*\bar{\Delta}_m$ occurs at $c_k^* = 1/\bar{\Delta}_m = O((\log n)m^{-1})$. Moreover, given the observed data, the likelihood function is a continuous function on the compact set

$$\{\theta_m = \{\theta_{m-\{\lambda\}}, c_1,\ldots,c_m\}|\ \ c_0\le c_k\le K_0m, k=1,\ldots,m; \|\ \theta_{m-\{\lambda\}}\ \|\le C_0\},$$

where $c_0>0$ is the lower bound of $\lambda(t)$. Therefore C1 is satisfied, and C2 follows from (3.4) and the fact that the sequence of sieve spaces is dense in the original parameter space.

For each $\delta>0$ and m, define

$$\mathcal{D}_m = \{\theta_m = \{\alpha_m,\Sigma_{b,m},\sigma_{e,m}^2,\beta_m,\lambda_m(\cdot)\}|\lambda_m\in\mathbb{S}_m, KL(\theta,\theta_m)\ge KL(\theta,\theta_m')+\delta\},$$

where θ'_m is the sequence in C2(b). Our goal is to find a finite cover of $\mathcal{D}_m$. That is to find $\mathcal{O}_1^m, \ldots, \mathcal{O}_{l_m}^m$ such that $\mathcal{D}_m \subseteq \cup_{k=1}^{l_m} O_k^m$ and $\sum_{n=1}^{\infty} l_m \cdot (\rho_m)^n < \infty$ with

$$\rho_m = \sup_k \inf_{t \geq 0} E_\theta \left(\exp \left\{ t \log \frac{f(O, \mathcal{O}_k^m)}{f(O, \theta'_m)} \right\} \right),$$

where O denotes a generic copy of the observed data $O_i = (V_i, \Delta_i, \boldsymbol{W}_i, \boldsymbol{t}_i)$, and $f(O, \theta'_m)$ is the likelihood function evaluated at θ'_m. The notation $f(O, \mathcal{O}_k^m)$ denotes $\sup_{\theta^* \in \mathcal{O}_k^m} f(O, \theta^*)$. Theorem 2 then follows from the Borel-Cantelli lemma as shown in Theorem 2 of Geman and Hwang (1982). It thus suffices to construct the suitable covering sets and show $\sum_{n=1}^{\infty} l_m (\rho_m)^n < \infty$, where the $(\rho_m)^n$ is an upper bound of the probability of one sieve MLE being in $\mathcal{D}_m$.

Consider the set of functions

$$\Upsilon_m = \{\lambda^*(t) = \sum_{k=1}^{m} c_k^* D_k(t) | \ \ c_k^* = c_0 + \frac{p}{m}, p = 0, 1, \ldots, K_0 m^2, t \in [0, \log \frac{n}{c_0}]\}.$$

Note that $\Upsilon_m \subseteq \mathbb{S}_m$ and has cardinality $(K_0 m^2)^m \leq m^{am}$ for some positive constant a. Since any function λ_m in $\mathbb{S}_m$ is bounded by $K_0 m$, there exists a λ^* in Υ_m such that $\sup_{0<t<K_1 \log n} |\lambda_m(t) - \lambda^*(t)| < 1/m$. Thus, the collection of open balls $\tilde{\mathcal{O}}_1^m, \ldots, \tilde{\mathcal{O}}_{l^m}^m$ with centers at the $\lambda^*(t)$ and radius $1/m$ covers $\mathbb{S}_m$. Similarly, for finite-dimensional parameters, we consider the set of points $\Xi_m = \{\theta^*_{-\{\lambda\}} = \pm p/m, p = 0, 1, \ldots, C_0 m^2\}$, which has cardinality m^a for some positive constant a, and the open covering balls $\tilde{\mathcal{P}}_1^m, \ldots, \tilde{\mathcal{P}}_{p^m}^m$ with centers at $\theta^*_{-\{\lambda\}}$'s and radius $1/m$. Now take $\mathcal{O}_k^m = (\tilde{\mathcal{O}}_i^m \times \tilde{\mathcal{P}}_j^m) \cap \mathcal{D}_m$. It is not difficult to see that we can cover $\mathcal{D}_m$ with at most m^{am} such $\mathcal{O}_k^m$'s for some positive constant a. Hence $l_m \leq m^{am}$.

For each k and a fixed $\theta^{**} \in \mathcal{O}_k^m$, the approximation of KL divergence in (3.4) implies

$$\begin{aligned} E \left(\log \frac{f(O, \mathcal{O}_k^m)}{f(O, \theta^{**})} \right) &\approx \frac{1}{2} E\{ \sup_{\theta^* \in \mathcal{O}_k^m} [(\theta^{**} - \theta^*)(i_{\theta^*, \theta^*})(\theta^{**} - \theta^*)^T]\}, \\ &< \frac{a' \log m}{m}, \end{aligned}$$

since $|\theta^{**} - \theta^*| < 1/m$. By the definition of $\mathcal{D}_m$, and for θ'_m satisfying C2,

$$\begin{aligned} E \left(\log \frac{f(O, \mathcal{O}_k^m)}{f(O, \theta'_m)} \right) &\leq KL(\theta, \theta^{**}) - KL(\theta, \theta'_m) + \frac{a' \log m}{m}, \\ &< \frac{a' \log m}{m} - \delta, \end{aligned}$$

for all $k = 1, 2, \ldots, l_m$ and all m.

Again for any fixed k, let

$$\phi(t) = E\Big\{ \exp\Big[t \log \frac{f(O, \mathcal{O}_k^m)}{f(O, \theta_m')} \Big] \Big\}, \quad t > 0.$$

Then $\phi(0) = 1$ and $\phi^{'}(0) \le (a' \log m)/m - \delta$. Furthermore, we have the bound $\phi^{''}(t) < (m \log m)^2$ for small t.

$$\text{Hence} \qquad \phi(t) < 1 + (\frac{a' \log m}{m} - \delta)t + \frac{1}{2}(m \log m)^2 t^2$$
$$= 1 - t\left(\delta - \frac{a' \log m}{m} - (m \log m)^2 t\right). \tag{5.1}$$

By choosing $t = (m \log m)^{-2}(\log m)^{-q}$ and q, m sufficiently large, we have $\phi(t) < 1 - \delta/[2(m \log m)^2 (\log m)^q]$ for all $k = 1, \ldots, l_m$. That is, $\rho_m \le 1 - \delta/[m^2(\log m)^{2+q}]$. Finally,

$$\sum_{n=1}^{\infty} l_m(\rho_m)^n \le \sum_{n=1}^{\infty} m^{am}\Big\{1 - \frac{\delta}{m^2(\log m)^{2+q}}\Big\}^n$$
$$= \exp\Big\{am \log m\Big(1 - \frac{\delta}{an^{-1}m^3(\log m)^{3+q}}\Big)\Big\}, \tag{5.2}$$

which is finite when $m \to +\infty$ and $mn^{-1/3+\varepsilon} < +\infty$. Part (i) of Theorem 2 follows from Theorem 2 of Geman and Hwang (1982) by using the Borel-Cantelli lemma.

As for part (ii), from the construction of $\mathcal{D}_m$, we know that the set, M_m^n, of the sieve MLE is within the $\delta + KL(\theta, \theta_m')$ ball of θ in term of KL divergence. Hence if we also allow $\delta \to 0$, we find a convergence rate for the sieve MLE, since KL divergence is $O(\rho(\hat{\theta}_m^{(s)}, \theta)^2)$. However, δ has to tend to 0 slower than $n^{-1}m^3(\log m)^{3+q}$, so that (5.2) is still finite, and slower than $(a' \log m)/m$ in order to stay as the leading term in (5.1). To get the best convergence rate, we choose $m = Kn^{1/4-\varepsilon}$ and $\delta = Kn^{-1/4+2\varepsilon}$ for some sufficient small $\varepsilon > 0$ and constant K. The sum in (5.2) is still finite. Hence the "global" rate of convergence of the sieve MLE, especially for $\hat{\lambda}_m$, is $O(n^{-1/8+\varepsilon})$.

Proof of Theorem 3. Note that

$$\widehat{KL}(\theta, \theta_m) - KL(\theta, \theta_m)$$
$$= \Big(\frac{1}{n}\sum_{i=1}^{n} \log L_i(\theta) - E_\theta(\log L_i(\theta))\Big) - \Big(\frac{1}{n}\sum_{i=1}^{n} \log L_i(\theta_m) - E_\theta(\log L_i(\theta_m))\Big),$$

where the first term is $O_p(n^{-1/2})$ and does not depend on m. We show the uniform (w.r.t. m) rate of the second term by calculating the covering entropy and following the maximum inequalities in Pollard (1990). Let $\log \mathbf{f}(\theta) = (\log L_1, \ldots, \log L_n)$ be the $n-$ vector of individual log-likelihood values corresponding to n copies of observed data. Consider a bounded set $\mathbb{B}_m \in \mathbb{R}^Q$, so the Euclidian norm of the envelope function is

$$\sup_{\theta_m \in \mathbb{B}_m \times \mathbb{S}_m} \| \log \mathbf{f}(\theta_m) \| \leq K n^{1/2} \log m \leq K n^{1/2} \log n.$$

Similar to the proof of Theorem 2, we set up Υ_ϵ and Ξ_ϵ for every given $\epsilon > 0$, m and n as

$$\Upsilon_\epsilon = \Big\{ \lambda^*(t) = \sum_{k=1}^{m} c_k^* D_k(t) \,\Big|\, c_k^* = c_0 + \frac{\epsilon p}{\sqrt{n}}, p = 0, 1, \ldots, \frac{Km\sqrt{n}}{\epsilon} \Big\};$$

$$\Xi_\epsilon = \Big\{ \theta^*_{-\{\lambda\}} = \pm \frac{\epsilon p}{\sqrt{n}}, p = 0, 1, \ldots, \frac{C_0 m \sqrt{n}}{\epsilon} \Big\}.$$

For every $\theta_m \in \mathbb{B}_m \times \mathbb{S}_m$ there exists a $\theta^* \in \Upsilon_\epsilon \times \Xi_\epsilon$ such that$\| \log \mathbf{f}(\theta_m) - \log \mathbf{f}(\theta^*) \| \leq \sqrt{n}(\epsilon/\sqrt{n}) = \epsilon$, hence the $\epsilon-$balls with centers at $\Upsilon_\epsilon \times \Xi_\epsilon$ cover the set $\{\log \mathbf{f}(\theta_m) | \theta_m \in \mathbb{B}_m \times \mathbb{S}_m\}$ with covering number $(Km\sqrt{n}/\epsilon)^{am}$. The integrated squared-root of the covering entropy is hence calculated as

$$\begin{aligned}
&\int_0^{n^{1/2}\log n} \sqrt{am \log(\frac{Km\sqrt{n}}{\epsilon})} \ \ \mathrm{d}\epsilon \\
&\quad = a\sqrt{m}\sqrt{n} \ \ m \int_{\log(Km/\log n)}^{+\infty} \sqrt{x} e^{-x} \ \ \mathrm{d}x \\
&\quad \leq a\sqrt{m}\sqrt{n} \ \ m \Big[-\sqrt{x} e^{-x} \Big|_{\log(Km/\log n)}^{+\infty} + \int_{\log(Km/\log n)}^{+\infty} e^{-x} \ \ \mathrm{d}x \Big] \\
&\quad \leq K' n^{1/2} m^{1/2} (\log n)^{3/2},
\end{aligned}$$

for certain positive constant K'.

The Ψ-Orlicz norm, $\|Z\|_\Psi = \inf\{C > 0 : E\Psi(|Z|/C) \leq 1\}$, is often used to control the tail probability. By the maximum inequality in Pollard (1990, p. 3), we have a bound for the Orlicz norm $\|Z\|_\Psi$,

$$\Big\| \max_{\mathbb{B}_m \times \mathbb{S}_m} \Big| \mathbf{u} \cdot \log \mathbf{f}(\theta_m) \Big| \Big\|_\Psi \leq K' n^{1/2} m^{1/2} (\log n)^{3/2}.$$

Equivalently we have

$$E \left(\frac{1}{5} \exp \Big\{ \frac{\max_{\mathbb{B}_m \times \mathbb{S}_m} |\mathbf{u} \cdot \log \mathbf{f}(\theta_m)|}{n^{1/2} m^{1/2} (\log n)^{3/2}} \Big\} \right) \leq 1,$$

where $\Psi(x) = \exp(x^2)/5$ and $\mathbf{u} = \{u_1, \ldots, u_n\}$ is a n-vector of *i.i.d.* binary random variables with equal probability 1/2 to be ±1. This finiteness of the Ψ-Orlicz norm places a constraint on the rate of the tail probability: $P(|Z| \geq t) \leq 5\exp(-t^2/\|Z\|_\Psi^2)$, for any random variable Z. This and the theorem of symmetrization and conditioning in Pollard (1990, p. 7) imply

$$\begin{aligned}
&P\Big(\max_{\mathbb{B}_m\times\mathbb{S}_m} \Big|\sum_1^n \log f_i(\theta_m) - E[\log f_i(\theta_m)]\Big| \geq t\Big)\\
&\quad\leq P\Big(\max_{\mathbb{B}_m\times\mathbb{S}_m} |\mathbf{u}\cdot\log\mathbf{f}(\theta_m)| \geq t\Big)\\
&\quad\leq 5\exp\Big\{-\frac{t^2}{(K'n^{1/2}m^{1/2}(\log n)^{3/2})^2}\Big\}.
\end{aligned}$$

Hence,

$$\max_{\mathbb{B}_m\times\mathbb{S}_m} \Big|\sum_1^n \log f_i(\theta_m) - E(\log f_i(\theta_m))\Big| = O(n^{1/2}m^{1/2}(\log n)^{3/2}),$$

and the uniform convergence in Theorem 3 is proved.

Proof of Theorem 4. (i) First, we discuss the case without censoring: $\Delta_i = 1, \forall i$. The score of c_k contributed by the ith individual is:

$$\begin{aligned}
&E_\theta\left(\frac{D_k(V_i)}{c_k} - E_i^*\Big(\int_0^{V_i} D_k(s)\exp\{\beta X_i(s)\}\mathrm{d}s\Big)\right)\\
&\quad= \int_{R^{m_i}}\int_{R^+}\left(\frac{D_k(t)}{c_k} - E_i^*\Big(\int_0^t D_k(s)\exp\{\beta X_i(s)\}\mathrm{d}s\Big)\right)L_i(O_i;\theta)\mathrm{d}t\mathrm{d}w. \quad (5.3)
\end{aligned}$$

Let $f_{V_i}(t|\mathbf{b}_i, \beta, \lambda_m^*)$ and $f_{\mathbf{W}_i}(w|\mathbf{b}_i, \sigma_e^2)$ be the densities for survival and longitudinal measurements, given random effects and parameters, respectively. Note that, when $\Delta_i = 1$,

$$\frac{L_i(O_i;\theta|\Delta_i = 1)}{L_i(O_i;\theta_m^*|\Delta_i = 1)} = 1 + O(\bar{\Delta}_m).$$

By Taylor expansion, changing the order of integration, and then applying integration-by-parts, we have

$$\begin{aligned}
&\int_{R^{m_i}}\int_{R^+}\Big(E_i^*\Big[\int_0^t D_k(s)\exp\{\beta X_i(s)\}\mathrm{d}s\Big]\Big)L_i(O_i;\theta)\mathrm{d}t\mathrm{d}w\\
=&\int_{R^{m_i}}\int_{R^+}\Big(\int_{R^q}\Big[\int_0^t D_k(s)\exp\{\beta X_i(s)\}\mathrm{d}s\Big]\frac{L_i^{(c)}(O_i,\boldsymbol{b}_i;\theta_m^*)}{L_i(O_i;\theta_m^*)}d\mathbf{b}_i\Big)L_i(O_i;\theta)\mathrm{d}t\mathrm{d}w\\
=&\int_{R^{m_i}}\int_{R^q}\Big(\int_{R^+}\Big[\int_0^t D_k(s)\exp\{\beta X_i(s)\}\mathrm{d}s\Big]L_i^{(c)}(O_i,\boldsymbol{b}_i;\theta_m^*)\frac{L_i(O_i;\theta)}{L_i(O_i;\theta_m^*)}\mathrm{d}t\Big)\mathrm{d}\mathbf{b}_i\mathrm{d}w
\end{aligned}$$

$$
\begin{aligned}
=&\int_{R^{m_i}}\int_{R^q}\Big(\int_{R^+}\Big[\int_0^t D_k(s)\exp\{\beta X_i(s)\}\mathrm{d}s\Big] f_{V_i}(t|\mathbf{b}_i,\beta,\lambda_m^*)\mathrm{d}t\Big)\\
&\qquad\qquad\qquad\qquad\{f_{\mathbf{W}_i}(w|\mathbf{b}_i,\sigma_e^2)\pi(\mathbf{b}_i;\alpha,\Sigma_b)\}(1+O(\bar{\Delta}_m))\mathrm{d}\mathbf{b}_i\mathrm{d}w\\
=&\int_{R^{m_i}}\int_{R^q}\Big(\int_{R^+}\frac{D_k(t)}{c_k} f_{V_i}(t|\mathbf{b}_i,\beta,\lambda_m^*)\mathrm{d}t\Big) f_{\mathbf{W}_i}(w|\boldsymbol{b}_i,\sigma_e^2)\pi(\mathbf{b}_i;\alpha,\Sigma_b)\\
&\qquad\qquad\qquad\qquad\times(1+O(\bar{\Delta}_m))\mathrm{d}\mathbf{b}_i\mathrm{d}w\\
=&\int_{R^{m_i}}\int_{R^+}\Big(\frac{D_k(t)}{c_k}\Big)\Big(\int_{R^q} f_{V_i}(t|\mathbf{b}_i,\beta,\lambda_m^{(s)})\pi(\mathbf{b}_i;\alpha,\Sigma_b) f_{\mathbf{W}_i}(w|\mathbf{b}_i,\sigma_e^2)\mathrm{d}\mathbf{b}_i\Big)\\
&\qquad\qquad\qquad\qquad\times(1+O(\bar{\Delta}_m)dtdw\\
=&\int_{R^{m_i}}\int_{R^+}\Big(\frac{D_k(t)}{c_k}\Big)\Big(\int_{R^q} f_{V_i}(t|\mathbf{b}_i,\beta,\lambda)\pi(\mathbf{b}_i;\alpha,\Sigma_b) f_{\mathbf{W}_i}(w|\mathbf{b}_i,\sigma_e^2)\mathrm{d}\mathbf{b}_i\Big)\\
&\qquad\qquad\qquad\qquad\times(1+O(\bar{\Delta}_m))\mathrm{d}t\mathrm{d}w\\
=&\int_{R^{m_i}}\int_{R^+}\Big(\frac{D_k(t)}{c_k}\Big)L_i(O_i;\theta)(1+O(\bar{\Delta}_m))\mathrm{d}t\mathrm{d}w. \qquad (5.4)
\end{aligned}
$$

The last approximation implies

$$
\begin{aligned}
E_\theta\left(\frac{D_k(V_i)}{c_k}-E_i^*\Big[\int_0^{V_i} D_k(s)\exp\{\beta X_i(s)\}\mathrm{d}s\Big]\right)&=O(\bar{\Delta}_m)E_\theta\left(\frac{D_k(V_i)}{c_k}\right)\\
&=O(n^{-p}\bar{\Delta}_m),
\end{aligned}
$$

for each individual, since the probability that the survival time of this individual falls in D_k is $O(m^{-1})=O(n^{-p})$. Therefore, $E_\theta[S_n^*(\lambda_m^*|\theta_m^*)]=O(n^{1-p}\bar{\Delta}_m)$.

For the case with censoring, we need to replace $D_k(V_i)$ by $\Delta_i D_k(V_i)$ in (5.3). This means the expectation in (5.3) now consists of two parts:

$$
\begin{aligned}
&E_\theta\left(\frac{\Delta_i D_k(V_i)}{c_k}-E_i^*\Big[\int_0^{V_i} D_k(s)\exp\{\beta X_i(s)\}\mathrm{d}s\Big]\right)\\
&\quad=P(\Delta_i=1)\int_{R^{m_i}}\int_{R^+}\left(\frac{D_k(t)}{c_k}-E_i^*\Big(\int_0^t D_k(s)\exp\{\beta X_i(s)\}\mathrm{d}s|\Delta_i=1\Big)\right)\\
&\qquad\qquad\qquad\qquad\qquad\qquad L_i(O_i;\theta|\Delta_i=1)\mathrm{d}t\mathrm{d}w\\
&\qquad+P(\Delta_i=0)\int_{R^{m_i}}\int_{R^+}-E_i^*\left(\int_0^t D_k(s)\exp\{\beta X_i(s)\}\mathrm{d}s|\Delta_i=0\right)\\
&\qquad\qquad\qquad\qquad\qquad\qquad L_i(O_i;\theta|\Delta_i=0)\mathrm{d}t\mathrm{d}w.
\end{aligned}
$$

The first term from $\Delta_i=1$ can be handled by previous arguments, while the second term from $\Delta_i=0$ is of smaller order; this is because the likelihood ratio of the censoring data $L_i(O_i,\theta|\Delta_i=0)/L_i(O_i,\theta_m^*|\Delta_i=0)$ only contains λ in the format of cumulative hazard function.

Next, we consider the expectation of the score function for β:

$$S_n^*(\beta|\theta_m^*) = \sum_{i=1}^n \Big\{ \Delta_i E_i^*(X_i(V_i)) - \sum_{k=1}^m c_k E_i^* \Big(\int_0^{V_i} D_k(s) X_i(s) \exp\{\beta X_i(s)\} \mathrm{d}s \Big) \Big\}.$$

Similar to the previous argument, the portion contributed by censored data is negligible compared to the portion contributed by uncensored data. We thus have covered the case where all $\Delta_i = 1$.

Following similar arguments as in (5.4), with changing the order of integration and integration-by-parts, we obtain

$$\begin{aligned}
&E_\theta \Big\{ E_i^*(X_i(V_i)) - \sum_{k=1}^m c_k E_i^* \Big(\int_0^{V_i} D_k(s) X_i(s) \exp\{\beta X_i(s)\} \mathrm{d}s \Big) \Big\} \\
&= E_\theta\{E_i^*(X_i(V_i))\} - E_{\theta_m}\{E_i^*(X_i(V_i))\} \\
&\quad + E_{\theta_m} \Big\{ \sum_{k=1}^m c_k E_i^* \Big(\int_0^{V_i} D_k(s) X_i(s) e^{\beta X_i(s)} \mathrm{d}s \Big) \Big\} \\
&\quad - E_\theta \Big\{ \sum_{k=1}^m c_k E_i^* \Big(\int_0^{V_i} D_k(s) X_i(s) e^{\beta X_i(s)} \mathrm{d}s \Big) \Big\} \\
&= \iiint X_i(t) \Big(\frac{L_i(O_i;\theta)}{L_i(O_i;\theta_m^*)} - 1 \Big) L_i^c(O_i, \boldsymbol{b}_i; \theta_m^*) \mathrm{d}\mathbf{b}_i \mathrm{d}t \mathrm{d}w \\
&\quad - \iiint \Big(\int_0^t \sum_{k=1}^m c_k D_k(s) X_i(s) e^{\beta X_i(s)} \mathrm{d}s \Big) \Big(\frac{L_i(O_i;\theta)}{L_i(O_i;\theta_m^*)} - 1 \Big) L_i^c(O_i, \boldsymbol{b}_i; \theta_m^*) \mathrm{d}\mathbf{b}_i \mathrm{d}t \mathrm{d}w \\
&\doteq Q_1 - Q_2.
\end{aligned}$$

Since

$$\begin{aligned}
\frac{L_i(O_i;\theta)}{L_i(O_i;\theta_m^*)} - 1 &= \frac{\int_{R^q} f_{\mathbf{W}_i}(w|\mathbf{b}_i, \sigma_e^2) f_{V_i}(t|\mathbf{b}_i, \beta, \lambda) \pi(\mathbf{b}_i; \alpha, \Sigma_b) \mathrm{d}\mathbf{b}_i}{\int_{R^q} f_{\mathbf{W}_i}(w|\mathbf{b}_i, \sigma_e^2) f_{V_i}(t|\mathbf{b}_i, \beta, \lambda^*) \pi(\mathbf{b}_i; \alpha, \Sigma_b) \mathrm{d}\mathbf{b}_i} - 1 \\
&\approx \Big(\frac{\lambda(V_i)}{\lambda_m^*(V_i)} - 1 \Big) + O(\bar{\Delta}_m^{1+d}),
\end{aligned}$$

and the deviation $(\lambda(V_i)/\lambda_m^*(V_i) - 1)$ is roughly stochastically independent of covariate $X_i(V_i)$ and its finite integral $\int_0^{V_i} \lambda_m^*(s) X_i(s) \exp\{\beta X_i(s)\} \mathrm{d}s$, the difference between Q_1 and Q_2 is of the order $O(\bar{\Delta}_m^{1+d})$. Then we have $E_\theta(S_n^*(\beta|\theta_m^*)) = O(n\bar{\Delta}_m^{1+d}) = O(n^{1-p(1+d)}(\log n)^{1+d})$. The expectation of score functions for other finite dimensional parameters can be derived in a similar fashion.

(ii) We now consider the order of the variance of score functions. Again, $D_k(V_i)/c_k$ is a binomial random variable, so $\mathrm{Var}_\theta(D_k(V_i)/c_k) = O(P(D_k(V_i) = 1)) = O(m^{-1})$. From (3.12), the second term in $S_n^*(c_k|\theta_m^*)$ is

$$\mathrm{Var}_\theta \Big(E_i^* \big(\int_0^{V_i} D_k(s) \exp\{\beta X_i(s)\} \mathrm{d}s \big) \Big) = O(\bar{\Delta}_m \mathrm{Var}_\theta(1_{[V_i > t_{k-1}]}))$$

$$= O\Big(\bar{\Delta}_m \frac{k}{m}(1 - \frac{k}{m})\Big).$$

Hence, $\mathrm{Var}_\theta(D_k(V_i)/c_k - E_i^*\big(\int_0^{V_i} D_k(s)\exp\{\beta X_i(s)\}\mathrm{d}s\big))$=$O(m^{-1})$, and

$$\mathrm{Var}_\theta\big(S_n^*(c_k|\theta_m^*)\big) = O(nm^{-1}) = O(n^{1-p}).$$

Similarly, assuming the covariate process is bounded as in A1, we have

$$\mathrm{Var}_\theta\Big\{\Delta_i E_i^*(X_i(V_i)) - \sum_{k=1}^{m} c_k E_i^*\Big(\int_0^{V_i} D_k(s)X_i(s)\exp\{\beta X_i(s)\}\mathrm{d}s\Big)\Big\} = O(1),$$

and $\mathrm{Var}_\theta\big(S_n^*(\beta|\theta_m^*)\big) = O(n)$ from (3.13). The variance bounds for the remaining finite dimensional parameters $\{\alpha, \Sigma_b, \sigma_e^2\}$ in Theorem 4 are straightforward.

(iii) Combining the results in (i) and (ii), we have $n^{-1/2}S_n^*(\omega|\theta_m^*) = O_p(1)$ for every finite-dimentional parameter ω in $(\alpha, \Sigma_b, \sigma_e^2, \beta)$, and $n^{-(1-p)/2}S_n^*(c_k|\theta_m^*) = O_p(1)$ for all c_k's in λ. From the expansion in (3.11) and Slusky's Theorem, the distance between the sieve estimate and sieve projection can be calculated as follows. For finite-dimensional parameters,

$$n^{1/2}(\hat{\theta}_{m-\{\lambda\}} - \theta_{m-\{\lambda\}}) = O_p(n^{1/2}(i^m_{\theta_m^*,\theta_m^*})^{-1}S_n^*(\hat{\theta}_{m-\{\lambda\}}|\theta_m^*)) = O_p(1),$$

and for the baseline hazard function λ,

$$n^{(1-p)/2}(\hat{\lambda}_m(t) - \lambda_m^*(t)) = O_p(n^{(1-p)/2}(i^m_{\theta_m^*,\theta_m^*})^{-1}S_n^*(\hat{\lambda}_m|\theta_m^*) = O_p(1).$$

The proof of Theorem 4 is now complete, since both terms dominate the sieve biases calculated in Theorem 1.

References

Chen, Y.-H. (2009). Weighted breslow-type and maximum likelihood estimation in semiparametric transformation models. *Biometrika* **96** 591-600.

Dupuy, J., Grama, I. and Mesbah, M. (2006). Asymptotic theory for the Cox model with missing time-depedent covariate. *Ann. Statist.* **34** 903-924.

Fleming, T. R. and Harringtion, D. P. (1991). *Counting Process and Survival Analysis.* Wiley, New York.

Geman, S. and Hwang, C.-R. (1982). Nonparametric maximum likelihood estimation by the method of sieves. *Ann. Statist.* **10** 401-414.

Grenander, U. (1981). *Abstract Inference.* Wiley, New York.

Guo, W., Ratcliffe, S. J. and Ten Have, T. T. (2004). A random pattern-mixture model for longitudinal data with dropouts. *J. Amer. Statist. Assoc.* **99** 929-937.

Henderson, R., Diggle, P. and Dobson, A. (2000). Identification and efficacy of longitudinal markers for survival. *Biostatistics* **3** 33-50.

Johansen, S. (1983). Functional data analysis for sparse longitudinal analysis. *International Statistics Review* **51** 165-174.

Kiefer, J. and Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function for vector chance variables. *Ann. Math. Statistics* **30** 463-489.

Meng, X. and Rubin, D. (1993). Maximum likelihood via the ECM algorithm: A general framework. *Biometrika* **80** 267-278.

Murphy, S. A. and van der Vaart, A. W. (2000). On profile likelihood. *J. Amer. Statist. Assoc.*, **95**, 449-465.

Murtaugh, P. A., Dickson, E. R., Vandam, G. M. and et al. (1994). Primary biliary-cirrhosis - prediction of short-term survival based on repeated patient visits. *Hepatology* **20** 126-134.

Pollard, D. (1984). *Convergence of Stochastic Processes.* Springer-Verlag, New York.

Pollard, D. (1990). *Empirical Processes: Theory and Applications. NSF-CBMS Regional Conference Series in Probability and Statistics* 2. Institute of Mathematical Statistics, Hayward, California.

Shen, X. and Wong, W. H. (1994). Convergence rate of sieve estimates. *Ann. Statist.* **22** 580-615.

Tsiatis, A. A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: An overview. *Statist. Sinica* **14** 809-834.

van der Vaart, A. and Wellner, J. (1996). *Weak convergence and Empirical Processes.* Springer-Verlag, New York.

Verbeke, G. and Davidian, M. (2008). Joint models for longitudinal data: Introduction and overview. In *Longitudinal Data Analysis: A Handbook of Modern Statistical Methods* (Edited by G. Fitzmaurice, M. Davidian, G. Verbeke and G. Molenberghs). Chapman & Hall/CRC, Boca Raton, Florida, 809-834.

Wong, W. H. and Shen, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *Ann. Statist.* **23** 339-362.

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

Xue, H., Lam, K. and Li, G. (2004). Sieve maximum likelihood estimator for semiparametric regression models with current status data. *J. Amer. Statist. Assoc.* **99** 346-356.

Zeng, D. and Cai, J. (2005). Asymptotic results for maximum likelihood estimators in joint analysis of repeated measurements and survival time. *Ann. Statist.* **33** 2132-2163.

Department of Statistics, University of California, Davis, CA 95616, USA.

E-mail: fhsieh@ucdavis.edu

Mathematics Department, Washington University at St. Louis, Saint Louis, MO 63130, USA.

E-mail: jmding@math.wustl.edu

Department of Statistics, University of California, Davis, CA 95616, USA.

E-mail: janelwang@ucdavis.edu

(Received April 2012; accepted October 2012)