RESEARCH ARTICLE

# Joint Modeling of Multivariate Nonparametric Longitudinal Data and Survival Data – A Local Smoothing Approach

Lu You*[1] | Peihua Qiu[2]

[1]Health Informatics Institute, University of South Florida, Florida, U.S.A.

[2]Department of Biostatistics, University of Florida, Florida, U.S.A.

**Correspondence**

*Lu You, Health Informatics Institute, University of South Florida, 3650 Spectrum Boulevard, Tampa, FL 33612, U.S.A. Email: Lu.You@epi.usf.edu

**Summary**

In many clinical studies, evaluating the association between longitudinal and survival outcomes is of primary concern. For analyzing data from such studies, joint modeling of longitudinal and survival data becomes an appealing approach. In some applications, there are multiple longitudinal outcomes whose longitudinal pattern is difficult to describe by a parametric form. For such applications, existing research on joint modeling is limited. In this paper, we develop a novel joint modeling method to fill the gap. In the new method, a local polynomial mixed-effects model is used for describing the nonparametric longitudinal pattern of the multiple longitudinal outcomes. Two model estimation procedures, i.e., the local EM algorithm and the local penalized quasi-likelihood estimation, are explored. Practical guidelines for choosing tuning parameters and for variable selection are provided. The new method is justified by some theoretical arguments and numerical studies.

**KEYWORDS:**

joint modeling; local kernel smoothing; longitudinal data; multiple outcomes; nonparametric mixed-effects model; survival data

## 1 | INTRODUCTION

Evaluating the effect of longitudinal variables on survival outcomes is of particular interest in many clinical trials and observational studies. To make use of the information in all available data, joint modeling of longitudinal and survival data has become a powerful tool. In many applications, there are multiple longitudinal variables whose longitudinal pattern is difficult to describe by a parametric form. This paper aims to develop an effective methodology for such applications.

The framework of joint modeling of longitudinal and survival data is first discussed in Wulfsohn and Tsiatis[1], where a linear mixed-effects model is used for modeling the growth curve of a univariate longitudinal outcome, and the growth curve is used for quantifying the survival rate of a patient. Since then, many methods have been proposed to extend the original joint modeling method to cases with multiple longitudinal outcomes[2,3] and cases with nonlinear longitudinal trajectories[4,5]. Most of these existing methods use regression splines or other parametric forms to generalize the univariate linear mixed-effects model to different nonlinear mixed-effects models with a single or multiple longitudinal outcomes. In these generalizations, the numbers of random-effects terms and model parameters can be quite large. Thus, model estimation often requires a considerable computational effort in evaluating certain intractable integrals related to the random effects. As an alternative, Brown et al[4] proposed a Bayesian framework for joint modeling of multivariate longitudinal data and survival data. In that Bayesian framework, B-spline basis functions were used to model the nonparametric longitudinal trajectories of the longitudinal variables, and

---

[0]**Abbreviations:** EM, expectation-maximization; MCMC, Markov chain Monte Carlo; AIC, Akaike information criterion; PQL, penalized quasi-likelihood; LASSO, least absolute shrinkage and selection operator; MSE, mean squared error; MISE, mean integrated squared error.

the Markov Chain Monte Carlo (MCMC) algorithm was used to evaluate the intractable integrals. However, in cases when the dimensions of the parameter space and the random-effects are high, it is hard to search the entire parameter space using the MCMC algorithm. Another limitation of that method is that it is often challenging to select the number of basis functions and the knots of the B-splines properly. After making a thorough literature review, Yang et al[3] claimed that their paper was the first one on joint modeling of nonlinear multivariate longitudinal data and survival data using the maximum likelihood estimation approach. However, their method was based on the polynomial function approximation, and it was challenging to deal with certain intractable integrals related to the random-effects of high dimensions in cases when either the number of basis functions or the dimension of longitudinal outcomes is large. Because of this challenge, they considered cases with only two longitudinal outcomes and with quadratic polynomial basis functions in their numerical studies. Of course, their assumed polynomial structure may not be adequate in characterizing the real pattern of the longitudinal data.

In this paper, we try to make another effort in joint modeling of multivariate nonparametric longitudinal data and survival data. A novel new method is suggested for solving this challenging problem. Instead of using spline or polynomial basis functions for approximating the multivariate nonparametric longitudinal trajectories of the data, we suggest a multivariate local mixed-effects model, in which the longitudinal trajectories are approximated *locally* by polynomial functions of lower degrees. Thus, the new method is flexible in describing the longitudinal trajectories of the observed data, and its computation is relatively simple as well.

The rest of the article is organized as follows. In Section 2, we will describe the suggested method in detail for joint modeling of multivariate nonparametric longitudinal data and survival data. Its statistical properties will be discussed in Section 3. Simulation studies are presented in Section 4 to evaluate the numerical performance of the proposed method in comparison with some alternative methods. A real-data application is discussed in Section 5. Some remarks conclude the article in Section 6.

## 2 | PROPOSED METHOD FOR JOINT MODELING

We describe our proposed new method in three parts. In Subsection 2.1, model formulation is discussed in detail for joint modeling of multivariate nonparametric longitudinal data and survival data. Then, an existing method based on the maximum likelihood estimation (MLE) and the regression spline function approximation is introduced in Subsection 2.2. Finally, the proposed new method is discussed in Subsection 2.3.

## 2.1 | Model formulation

Suppose we have observations of both longitudinal and survival outcomes of $m$ subjects in a study. For the $i$th subject, let $T_i$ be the observed last follow-up time, $\Delta_i = 1$ denote the event that a "survival" outcome is observed at the last follow-up time $T_i$, and $\Delta_i = 0$ denote its complementary event. The last-follow-up time $T_i$ is assumed to be the minimum of the time to event $D_i$ and the censoring time $C_i$, i.e., $T_i = \min\{D_i, C_i\}$. Besides the survival outcomes, it is assumed that $p$-dimensional longitudinal outcomes $\mathbf{y}$ are observed at times $t_{i1}, \ldots, t_{in_i}$ with $t_{i1} < \ldots < t_{in_i}$, and their observations are denoted as $\mathbf{y}_{i1}, \ldots, \mathbf{y}_{in_i}$. It is further assumed that all observation times $\{t_{ij}\}$ and the last follow-up time $T_i$ are within the study period $[0, \mathcal{T}]$, and the observed longitudinal outcomes follow the following nonparametric mixed-effects model:

$$\mathbf{y}_{ij} = \boldsymbol{\mu}(t_{ij}) + \mathbf{v}_i(t_{ij}) + \boldsymbol{\epsilon}_i(t_{ij}), \text{ for } j = 1, \ldots, n_i, i = 1, \ldots, m, \tag{1}$$

where $\boldsymbol{\mu}(t)$ is the population mean outcomes, $\mathbf{v}_i(t)$ is the random-effects function that models the subject-specific variation from $\boldsymbol{\mu}(t)$, and $\boldsymbol{\epsilon}_i(t)$ is the pure measurement error. The random-effects function $\mathbf{v}_i(t)$ is assumed to have the mean $\mathrm{E}[\mathbf{v}_i(t)] = \mathbf{0}$ and the covariance $\mathrm{Cov}(\mathbf{v}_i(s), \mathbf{v}_i(t)) = \boldsymbol{\Sigma}_v(s, t)$, for any $t, s \in [0, \mathcal{T}]$. The pure measurement errors $\boldsymbol{\epsilon}_i(t)$, for different $i$, are assumed to be independent realizations of a process with mean 0 and variance function $\mathrm{Var}(\boldsymbol{\epsilon}_i(t)) = \boldsymbol{\Sigma}_e(t)$. In model (1), let $\mathbf{m}_i(t) = \boldsymbol{\mu}(t) + \mathbf{v}_i(t)$, which denotes the latent trajectory of the longitudinal outcomes after the pure measurement errors are removed. This latent trajectory is assumed to be associated with the observed survival outcomes through the following proportional hazards model:

$$\lambda_i(t) = \lambda_0(t) \exp\{\boldsymbol{\beta}^\mathsf{T} \mathbf{m}_i(t)\}, \tag{2}$$

where $\lambda_i(t) = \lim_{dt \downarrow 0} P\{D_i \in (t, t+dt) | D_i > t\}/dt$ is the hazard function for the $i$th subject, $\lambda_0(t)$ is the baseline hazard function, and $\boldsymbol{\beta}$ is the coefficient vector. For ease of presentation, we introduce the following notations. Let $\mathcal{Y}_i = \{\mathbf{y}_{ij}, j = 1, \ldots, n_i\}$ and $S_i = \{(\Delta_i, T_i)\}$ denote the observed longitudinal and survival data of the $i$th subject, and $\mathcal{Y} = \{\mathcal{Y}_i, i = 1, \ldots, m\}$ and $S = \{S_i, i = 1, \ldots, m\}$ denote the observed survival and longitudinal data of all subjects. Furthermore, $R(t) = \{i : T_i \geq t\}$

denotes the set of subjects at risk at time $t$, $N_i(t) = \Delta_i I(T_i \geq t)$ is the counting process of disease incidence of the $i$th subject, and $N(t) = \sum_{i=1}^{m} N_i(t)$.

Conditional on $\mathbf{m}_i(t)$, the log joint probability density function of $\mathcal{Y}_i$ is

$$\log f_1(\mathcal{Y}_i | \{\mathbf{m}_i(t)\}; \boldsymbol{\Sigma}_e) = -\frac{n_i}{2} \log \det(2\pi\boldsymbol{\Sigma}_e) - \frac{1}{2} \sum_{j=1}^{n_i} \left[\mathbf{y}_{ij} - \mathbf{m}_i(t_{ij})\right]^{\mathsf{T}} \boldsymbol{\Sigma}_e^{-1} \left[\mathbf{y}_{ij} - \mathbf{m}_i(t_{ij})\right], \tag{3}$$

where $\det(\cdot)$ denotes the determinant of a matrix. Conditional on $\mathbf{m}_i(t)$, the log joint probability density function of $S_i$ is $\Delta_i \log \lambda_i(T_i) - \int_0^{T_i} \lambda_i(u)\,du$. Because in the likelihood function, we only need to know the baseline hazard function $\lambda_0(t)$ at each observed time, the log conditional joint probability density function can also be written as follows [6,7]:

$$\log f_2(S_i | \{\mathbf{m}_i(t)\}; \boldsymbol{\beta}, \lambda_0) = \Delta_i \left\{ \log \lambda_0\{T_i\} + \boldsymbol{\beta}^{\mathsf{T}} \mathbf{m}_i(T_i) \right\} - \sum_{l=1}^{m} \Delta_l \lambda_0\{T_l\} \exp\{\boldsymbol{\beta}^{\mathsf{T}} \mathbf{m}_i(T_l)\} I(T_i \geq T_l), \tag{4}$$

where we have followed the notational convention of Murphy and Van der Vaart [7] that $\lambda_0\{t\}$ denotes the mass of the estimated baseline hazard function at $t$, and $\lambda_0 = \{\lambda_0\{t\}, t = T_1, \ldots, T_n\}$ is a collection of all $\lambda_0\{t\}$. Murphy and Van der Vaart [7] further showed that the log conditional joint probability density function for $\{S_i\}_{i=1}^{m}$ can be rewritten as

$$\sum_{i=1}^{m} \log f_2(S_i | \{\mathbf{m}_i(t)\})$$
$$= \sum_{i=1}^{m} \Delta_i \left\{ \log \lambda_0\{T_i\} + \boldsymbol{\beta}^{\mathsf{T}} \mathbf{m}_i(T_i) \right\} - \sum_{i=1}^{m} \sum_{l=1}^{m} \Delta_i \lambda_0\{T_i\} \exp\{\boldsymbol{\beta}^{\mathsf{T}} \mathbf{m}_l(T_i)\} I(T_l \geq T_i) \tag{5}$$
$$= \sum_{i=1}^{m} \Delta_i \left[ \log \lambda_0\{T_i\} + \boldsymbol{\beta}^{\mathsf{T}} \mathbf{m}_i(T_i) - \lambda_0\{T_i\} \sum_{l=1}^{m} \exp\{\boldsymbol{\beta}^{\mathsf{T}} \mathbf{m}_l(T_i)\} I(T_l \geq T_i) \right].$$

By plugging in the following expression

$$\lambda_0\{T_i\} = \frac{\Delta_i}{\sum_{l=1}^{m} \exp\{\boldsymbol{\beta}^{\mathsf{T}} \mathbf{m}_l(T_i)\} I(T_l \geq T_i)},$$

Expression (5) becomes the familiar Cox's partial likelihood function. [7]

## 2.2 | Maximum likelihood estimation by using regression splines

Many existing methods for joint modeling of longitudinal and survival data approximate the latent longitudinal trajectories $\mathbf{m}_i(t)$ by using regression splines that are linear combinations of some pre-specified B-splines or other types of basis functions. [4,5,3] The major steps in such methods are briefly described below. Let $\phi_1(t), \ldots, \phi_q(t)$ be $q$ basis functions. Then, each component of $\mathbf{m}_i(t) = \boldsymbol{\mu}(t) + \mathbf{v}_i(t) = (m_{i1}(t), \ldots, m_{ip}(t))^{\mathsf{T}}$ can be expressed as $m_{ik}(t) = \sum_{l=1}^{q} c_{kl}\phi_l(t) + \sum_{l=1}^{q} a_{ikl}\phi_l(t) = \sum_{l=1}^{q} b_{ikl}\phi_l(t)$, where $c_{kl}$ are the fixed-effects coefficients, and $a_{ikl}$ are the random-effects coefficients, and $b_{ikl} = c_{kl} + a_{ikl}$. The longitudinal outcomes can be described by the following linear mixed-effects model:

$$\begin{bmatrix} \mathbf{y}_{i1} \\ \vdots \\ \mathbf{y}_{in_i} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Phi}(t_{i1}) \\ \vdots \\ \boldsymbol{\Phi}(t_{in_i}) \end{bmatrix} \mathbf{c} + \begin{bmatrix} \boldsymbol{\Phi}(t_{i1}) \\ \vdots \\ \boldsymbol{\Phi}(t_{in_i}) \end{bmatrix} \mathbf{a}_i + \begin{bmatrix} \boldsymbol{\epsilon}_i(t_{i1}) \\ \vdots \\ \boldsymbol{\epsilon}_i(t_{in_i}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Phi}(t_{i1}) \\ \vdots \\ \boldsymbol{\Phi}(t_{in_i}) \end{bmatrix} \mathbf{b}_i + \begin{bmatrix} \boldsymbol{\epsilon}_i(t_{i1}) \\ \vdots \\ \boldsymbol{\epsilon}_i(t_{in_i}) \end{bmatrix},$$

where

$$\boldsymbol{\Phi}(t_{ij}) = \mathbf{I}_{p \times p} \otimes \left( \phi_1(t_{ij}), \ldots, \phi_q(t_{ij}) \right)$$
$$= \begin{bmatrix} \phi_1(t_{ij}) & \cdots & \phi_q(t_{ij}) & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \phi_1(t_{ij}) & \cdots & \phi_q(t_{ij}) & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 & \cdots & \phi_1(t_{ij}) & \cdots & \phi_q(t_{ij}) \end{bmatrix}$$
$$\mathbf{c} = \left[ (c_{11}, \ldots, c_{1q}) \quad \ldots \quad (c_{p1}, \ldots, c_{pq}) \right]^{\mathsf{T}},$$
$$\mathbf{a}_i = \left[ (a_{i11}, \ldots, a_{i1q}) \quad \ldots \quad (a_{ip1}, \ldots, a_{ipq}) \right]^{\mathsf{T}}, \mathbf{b}_i = \left[ (b_{i11}, \ldots, b_{i1q}) \quad \ldots \quad (b_{ip1}, \ldots, b_{ipq}) \right]^{\mathsf{T}}.$$

In the above model, the fixed-effects part $\boldsymbol{\mu}(t)$, the random-effects part $\mathbf{v}_i(t)$, and the latent trajectory $\mathbf{m}_i(t)$ are modeled by $\boldsymbol{\Phi}(t)\mathbf{c}$, $\boldsymbol{\Phi}(t)\mathbf{a}_i$, and $\boldsymbol{\Phi}(t)\mathbf{b}_i$ respectively. $\mathbf{c}$ is a $pq$-dimensional coefficient vector of the fixed-effects, $\mathbf{a}_i$ is a $pq$-dimensional coefficient vector of the random-effects that is assumed to follow the distribution $N(\mathbf{0}, \boldsymbol{\Sigma}_b)$, $\mathbf{b}_i$ is therefore $pq$-dimensional and $N(\mathbf{c}, \boldsymbol{\Sigma}_b)$ distributed, and the error terms $\{\epsilon_i(t_{ij})\}$ are assumed to be independent of each other and follow the common distribution $N(\mathbf{0}, \boldsymbol{\Sigma}_e)$. Then, conditional on $\mathbf{a}_i$, the log joint probability density function of $\mathcal{Y}_i$ and $S_i$ are given respectively by

$$\log f_1(\mathcal{Y}_i|\mathbf{b}_i; \boldsymbol{\Sigma}_e) = -\frac{n_i}{2} \log \det(2\pi\boldsymbol{\Sigma}_e) - \frac{1}{2} \sum_{j=1}^{n_i} \left[\mathbf{y}_{ij} - \boldsymbol{\Phi}(t_{ij})\mathbf{b}_i\right]^\mathsf{T} \boldsymbol{\Sigma}_e^{-1} \left[\mathbf{y}_{ij} - \boldsymbol{\Phi}(t_{ij})\mathbf{b}_i\right]$$

and

$$\log f_2(S_i|\mathbf{a}_i; \boldsymbol{\beta}, \lambda_0) = \Delta_i \left\{ \log \lambda_0\{T_i\} + \boldsymbol{\beta}^\mathsf{T}\boldsymbol{\Phi}(T_i)\mathbf{b}_i \right\} - \sum_{l=1}^{m} \Delta_l \lambda_0\{T_l\} \exp\{\boldsymbol{\beta}^\mathsf{T}\boldsymbol{\Phi}(T_l)\mathbf{b}_i\} I(T_i \geq T_l).$$

The log probability density function of $\mathbf{b}_i$ is

$$\log f_3(\mathbf{b}_i; \mathbf{c}, \boldsymbol{\Sigma}_b) = -\frac{1}{2} \log \det(2\pi\boldsymbol{\Sigma}_b) - \frac{1}{2}(\mathbf{b}_i - \mathbf{c})^\mathsf{T}\boldsymbol{\Sigma}_b^{-1}(\mathbf{b}_i - \mathbf{c}).$$

So, the likelihood function of the observed longitudinal and survival data can be written as

$$L(\mathbf{c}, \boldsymbol{\Sigma}_b, \boldsymbol{\Sigma}_e, \boldsymbol{\beta}, \lambda_0) = \prod_{i=1}^{m} \int_{\mathbb{R}^{pq}} f_1(\mathcal{Y}_i|\mathbf{b}_i; \boldsymbol{\Sigma}_e) f_2(S_i|\mathbf{b}_i; \boldsymbol{\beta}, \lambda_0) f_3(\mathbf{b}_i|\mathbf{c}, \boldsymbol{\Sigma}_b)\, d\mathbf{a}_i. \tag{6}$$

Then, all model parameters can be estimated by maximizing the above likelihood function. The trajectories $\mathbf{m}_i(t)$ can be estimated by $\boldsymbol{\Phi}(t)(\mathbf{c} + \mathrm{E}[\mathbf{a}_i|\mathcal{Y}_i, S_i])$, which is an analogy of the best linear unbiased predictor in the linear mixed-effects model literature [8]. In (6), the integrals are taken over the $pq$-dimensional space $\mathbb{R}^{pq}$ for the random effects $\mathbf{a}_i$. Most existing methods in the literature used either the Gaussian quadrature or the Monte Carlo method to approximate these integrals. However, a $Q$-point Gaussian quadrature will involve $Q^{pq}$ evaluations of the integrand function, and the Monte Carlo methods with an insufficient number of sampling points will lead to large variability in the estimates. Furthermore, properly selecting of the number of basis functions and knots of splines basis functions is always challenging. Therefore, although such methods are theoretically possible, they are often practically inconvenient to implement.

## 2.3 | A proposed method by using the local polynomial mixed-effects modeling

In this part, we describe our proposed method for joint modeling of longitudinal and survival data. In our proposed method, an alternative way to model the nonparametric longitudinal trajectories $\mathbf{m}_i(t)$ is suggested. In the literature, there have been some existing discussions on nonparametric longitudinal data modeling [9,10,11,12]. Most of these methods are for modeling univariate longitudinal data. Here, we try to develop a multivariate local polynomial mixed-effects model for analyzing multivariate longitudinal data. Our model is similar to the one discussed in Wu and Zhang [11], although the latter is for univariate cases only and it does not take into account any survival data. Let $\boldsymbol{\mu}_c(t) = \mathrm{E}[\mathbf{m}_i(t)|T_i \geq t]$ be the conditional mean of $\mathbf{m}_i(t)$ given the $i$th subject being at risk by the time $t$, and $\mathbf{v}_{c,i}(t) = \mathbf{m}_i(t) - \boldsymbol{\mu}_c(t)$ be the mean variation of the $i$th subject from $\boldsymbol{\mu}_c(t)$. Then, in a small neighborhood $[t - h, t + h]$ of a given time $t \in [0, \mathcal{T}]$, where $h$ is a bandwidth parameter, we have the following Taylor's expansion of order $r$:

$$\boldsymbol{\mu}_c(s) \approx \sum_{l=0}^{r} (s - t)^l \frac{1}{l!} \boldsymbol{\mu}_c^{(l)}(t) = \mathbf{X}(s - t) \begin{bmatrix} \boldsymbol{\mu}_c(t) \\ \vdots \\ \frac{1}{r!}\boldsymbol{\mu}_c^{(r)}(t) \end{bmatrix},$$

$$\mathbf{v}_{c,i}(s) \approx \sum_{l=0}^{r} (s - t)^l \frac{1}{l!} \mathbf{v}_{c,i}^{(l)}(t) = \mathbf{X}(s - t) \begin{bmatrix} \mathbf{v}_{c,i}(t) \\ \vdots \\ \frac{1}{r!}\mathbf{v}_{c,i}^{(r)}(t) \end{bmatrix},$$

where the superscript $(l)$ denotes the $l$th derivative of a function, $\mathbf{X}(t) = \left[\mathbf{I}_{p \times p}, t\mathbf{I}_{p \times p}, \ldots, t^r\mathbf{I}_{p \times p}\right]$, and $\mathbf{I}_{p \times p}$ is the $p \times p$ dimensional identity matrix. By using these local function approximations, model (1) can be approximated by the following multiple linear mixed-effects model in $[t - h, t + h]$

$$\mathbf{y}_{ij} = \mathbf{X}(t_{ij} - t)\left[\mathbf{c}(t) + \mathbf{a}_i(t)\right] + \epsilon_i(t_{ij}) = \mathbf{X}(t_{ij} - t)\mathbf{b}_i(t) + \epsilon_i(t_{ij}), \tag{7}$$

where $\mathbf{c}(t) = \left[\boldsymbol{\mu}_c(t)^\mathsf{T}, \ldots, \frac{1}{r!}\boldsymbol{\mu}_c^{(r)}(t)^\mathsf{T}\right]^\mathsf{T}$ is the fixed-effects term, and $\mathbf{a}_i(t) = \left[\mathbf{v}_{c,i}(t)^\mathsf{T}, \ldots, \frac{1}{r!}\mathbf{v}_{c,i}^{(r)}(t)^\mathsf{T}\right]^\mathsf{T}$ is the random-effects term. Let $\mathbf{b}_i(t) = \mathbf{c}(t) + \mathbf{a}_i(t)$. Then, $\mathbf{X}(s-t)\mathbf{b}_i(t)$ in (7) is an approximation of $\mathbf{m}_i(s)$ in the neighborhood $[t-h, t+h]$. It is assumed that $\mathbf{b}_i(t)$ follows a multivariate normal distribution with the mean $\mathbf{c}(t)$ and the variance-covariance matrix $\boldsymbol{\Sigma}_b(t)$, and that the error term $\epsilon_i(t)$ follows a multivariate distribution with mean $\mathbf{0}$ and the variance-covariance matrix $\boldsymbol{\Sigma}_e(t)$, given that the $i$th subject is at risk at time $t$. To estimate the time-varying parameters $\mathbf{c}(t)$, $\boldsymbol{\Sigma}_b(t)$ and $\boldsymbol{\Sigma}_e(t)$ in Model (7), consider using the local kernel smoothing MLE that is described below. Let $K(s)$ be a symmetric kernel function with the support $[-1, 1]$ and $K(0) = 1$. In this paper, the Epanechnikov kernel function $K(u) = (1 - u^2)I(|u| \leq 1)$ is used because of its good theoretical properties [13]. Assume that the $i$th subject is in $R(t)$ (i.e., it is at risk by the time $t$). Then, conditional on $\mathbf{b}_i(t)$, the locally weighted log joint probability density function of $\mathcal{Y}_i$ is

$$\log f_{1,t}(\mathcal{Y}_i | \mathbf{b}_i(t); \boldsymbol{\Sigma}_e(t)) = -\frac{n_i}{2} \log \det(2\pi\boldsymbol{\Sigma}_e(t)) K_h(t_{ij} - t)$$
$$- \frac{1}{2}\sum_{j=1}^{n_i} \left[\mathbf{y}_{ij} - \mathbf{X}(t_{ij} - t)\mathbf{b}_i(t)\right]^\mathsf{T}\boldsymbol{\Sigma}_e(t)^{-1}\left[\mathbf{y}_{ij} - \mathbf{X}(t_{ij} - t)\mathbf{b}_i(t)\right]K_h(t_{ij} - t),$$

where $K_h(s) = K(s/h)$. Conditional on $\mathbf{b}_i(t)$, the locally weighted log probability density function of $S_i$ is

$$\log f_{2,t}(S_i | \mathbf{b}_i(t); \boldsymbol{\beta}, \lambda_0) = \Delta_i\left\{\log \lambda_0\{T_i\} + \boldsymbol{\beta}^\mathsf{T}\mathbf{X}(T_i - t)\mathbf{b}_i(t)\right\}K_h(T_i - t)$$
$$- \sum_{l=1}^m \Delta_l \lambda_0\{T_l\} \exp\{\boldsymbol{\beta}^\mathsf{T}\mathbf{X}(T_l - t)\mathbf{b}_i(t)\}K_h(T_l - t)I(T_i \geq T_l),$$

and the log probability density function of $\mathbf{b}_i(t)$ is

$$\log f_{3,t}(\mathbf{b}_i(t) | \mathbf{c}(t), \boldsymbol{\Sigma}_b(t)) = -\frac{1}{2}\log \det(2\pi\boldsymbol{\Sigma}_b(t)) - \frac{1}{2}[\mathbf{b}_i(t) - \mathbf{c}(t)]^\mathsf{T}\boldsymbol{\Sigma}_b(t)^{-1}[\mathbf{b}_i(t) - \mathbf{c}(t)].$$

The entire local likelihood function of the observed longitudinal and survival data in the neighborhood $[t-h, t+h]$ is then given by

$$L_t(\mathbf{c}(t), \boldsymbol{\Sigma}_b(t), \boldsymbol{\Sigma}_e(t), \lambda_0) = \prod_{i \in R(t)} \int_{\mathbb{R}^{pr}} f_{1,t}(\mathcal{Y}_i | \mathbf{b}_i(t); \boldsymbol{\Sigma}_e(t)) f_{2,t}(S_i | \mathbf{b}_i(t); \boldsymbol{\beta}, \lambda_0) f_{3,t}(\mathbf{b}_i(t) | \mathbf{c}(t), \boldsymbol{\Sigma}_b(t)) \, d\mathbf{b}_i(t). \tag{8}$$

From (8), it can be seen that when the bandwidth parameter $h$ is larger, the above local likelihood function would be closer to the (global) likelihood function defined in (6). However, these two likelihood functions are still different in several aspects. First, the dimension of random-effects in the local likelihood function depends on the order of polynomials in the local function approximation (cf., (7)), while the dimension of random-effects depends on the number of basis functions used in the (global) likelihood function. Because the order of polynomials considered in (8) is usually just 1, which corresponds to the local linear kernel estimation [14], the dimension of random-effects in (8) would be much smaller than that in (6). Second, the local likelihood function defined in (8) involves the observed data in the neighborhood $[t-h, t+h]$ only, its computation would be relatively simple, compared to the computation for the (global) likelihood function. Although the local likelihood function needs to be evaluated at all observed event times, when the trajectories of the longitudinal outcomes are nonparametric it is still more computationally efficient than the (global) likelihood function in (6) where the conditional expectations of the random-effects also need to be evaluated at all observed event times. Finally, the mixed-effects model that results in the (global) likelihood function (6) assumes time-independent variance-covariance matrices for the random-effects and the pure measurement error, while time-varying variance-covariance matrices are used in (8). Obviously, the latter is much more flexible for describing the data correlation in the observed longitudinal data.

**Model estimation by a local EM algorithm.** The EM algorithm [15] is one of the most popular tools for model estimation in joint modeling of longitudinal and survival data [1]. Because the proposed method is formulated based on the local likelihood function, we need to develop a local version of the EM algorithm. In the literature, the local EM algorithm is first discussed by Betensky et al. for estimating the hazard function in cases with interval-censored data [16]. It has been used for nonparametric density estimation [17] and spatial data analysis [18,19]. However, its application to joint modeling of longitudinal and survival data is still missing. The local EM algorithm is a straightforward solution to maximizing a local likelihood. It is also a computationally efficient method to approximate a complex model locally. See Nguyen et al. [19] for a more comprehensive discussion about its preferable features. Our proposed local EM algorithm for joint modeling of longitudinal and survival data is described below.

Similar to the conventional EM algorithm, the proposed local EM algorithm proceeds by iterating between the expectation step and the maximization step. In the expectation step, the expectation of the local log-likelihood function is evaluated, conditional

on the observed data. Then, in the maximization step, parameter estimates are updated by maximizing the conditional expectation of the local log-likelihood function. Different from the conventional EM algorithm, in the current local EM algorithm, we work with the local log-likelihood function, and thus the conditional expectation in the local EM algorithm should take with respect to the local probability density function.

To proceed with the proposed local EM algorithm, we first need to derive a formula for calculating the expectations of the random-effects $\mathbf{b}_i(t)$ conditional on the observed data. Let $\boldsymbol{\psi}(\mathbf{b}_i(t))$ be some function of $\mathbf{b}_i(t)$, and $\widehat{\boldsymbol{\Omega}}$ be the set of all parameters estimated in the maximization step. For a given $t$, denote the expectation of $\boldsymbol{\psi}(\mathbf{b}_i(t))$ conditional on $(\mathcal{Y}_i, \mathcal{S}_i)$ and $\widehat{\boldsymbol{\Omega}}$ by

$$\widetilde{\mathrm{E}}_{i,t}[\boldsymbol{\psi}(\mathbf{b}_i(t))] = \int_{\mathbb{R}^{pr}} \boldsymbol{\psi}(\mathbf{b}_i(t)) f_t(\mathbf{b}_i(t)|\mathcal{Y}_i, \mathcal{S}_i; \widehat{\boldsymbol{\Omega}}) \, d\mathbf{b}_i(t),$$

where by the Bayes rule,

$$f_t(\mathbf{b}_i(t)|\mathcal{Y}_i, \mathcal{S}_i; \widehat{\boldsymbol{\Omega}}) = \frac{f_{2,t}(\mathcal{S}_i|\mathbf{b}_i(t); \widehat{\boldsymbol{\Omega}}) f_t(\mathbf{b}_i(t)|\mathcal{Y}_i; \widehat{\boldsymbol{\Omega}})}{\int_{\mathbb{R}^{pr}} f_{2,t}(\mathcal{S}_i|\mathbf{b}_i(t); \widehat{\boldsymbol{\Omega}}) f_t(\mathbf{b}_i(t)|\mathcal{Y}_i; \widehat{\boldsymbol{\Omega}}) \, d\mathbf{b}_i(t)}.$$

The quantity $f_t(\mathbf{b}_i(t)|\mathcal{Y}_i)$ in the above expression is the density function of a multivariate normal distribution whose expression is given in Appendix A.1. Similarly, $\widetilde{\mathrm{E}}[\cdot]$ denotes the expectation of a quantity conditional on $(\mathcal{Y}, \mathcal{S})$ and $\widehat{\boldsymbol{\Omega}}$. Then, the expectation of the local log-likelihood function conditional on the observed data is

$$\begin{aligned}
&\widetilde{\mathrm{E}}\big[\ell_t\big(\mathbf{c}(t), \boldsymbol{\Sigma}_b(t), \boldsymbol{\Sigma}_e(t)\big)\big] \\
&= \widetilde{\mathrm{E}}\Bigg[ \sum_{i \in R(t)} \big( \log f_{1,t}(\mathcal{Y}_i|\mathbf{b}_i(t); \boldsymbol{\Sigma}_e(t)) + \log f_{2,t}(\mathcal{S}_i|\mathbf{b}_i(t); \boldsymbol{\beta}, \lambda_0) + \log f_{3,t}(\mathbf{b}_i(t)|\mathbf{c}(t), \boldsymbol{\Sigma}_b(t))\big)\Bigg] \\
&= -\frac{1}{2} \sum_{i=1}^m \sum_{j=1}^{n_i} \widetilde{\mathrm{E}}_{i,t}\Big[ \big[\mathbf{y}_{ij} - \mathbf{X}(t_{ij} - t)\mathbf{b}_i(t)\big]^\mathsf{T} \boldsymbol{\Sigma}_e(t)^{-1} \big[\mathbf{y}_{ij} - \mathbf{X}(t_{ij} - t)\mathbf{b}_i(t)\big]\Big] \\
&\quad \times K_h(t_{ij} - t) I(T_i \geq t) - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^{n_i} \log \det(2\pi\boldsymbol{\Sigma}_e(t)) K_h(t_{ij} - t) I(T_i \geq t) \\
&\quad + \sum_{i=1}^m \Delta_i \Big[ \log \lambda_0\{T_i\} + \boldsymbol{\beta}^\mathsf{T} \mathbf{X}(T_i - t) \widetilde{\mathrm{E}}_{i,t}[\mathbf{b}_i(t)]\Big] K_h(T_i - t) \\
&\quad - \sum_{i=1}^m \sum_{l=1}^m \Delta_l \lambda_0\{T_l\} \widetilde{\mathrm{E}}_{i,t}\Big[ \exp\big\{\boldsymbol{\beta}^\mathsf{T} \mathbf{X}(T_l - t)\mathbf{b}_i(t)\big\}\Big] K_h(T_l - t) I(T_i \geq T_l) \\
&\quad - \frac{1}{2} \sum_{i=1}^m \widetilde{\mathrm{E}}_{i,t}\Big[ \big[\mathbf{b}_i(t) - \mathbf{c}(t)\big]^\mathsf{T} \boldsymbol{\Sigma}_b(t)^{-1} \big[\mathbf{b}_i(t) - \mathbf{c}(t)\big]\Big] I(T_i \geq t) \\
&\quad - \frac{1}{2} \sum_{i=1}^m \log \det(2\pi\boldsymbol{\Sigma}_b(t)) I(T_i \geq t).
\end{aligned}$$

To maximize the above quantity, we can set its derivatives with respect to each time varying parameter $\mathbf{c}(t)$, $\boldsymbol{\Sigma}_b(t)$ and $\boldsymbol{\Sigma}_e(t)$ to zero. The resulting parameter updating formulas in the maximization step are as follows:

$$\mathbf{c}(t) \leftarrow \frac{\sum_{i=1}^m \widetilde{\mathrm{E}}_{i,t}[\mathbf{b}_i(t)] I(T_i \geq t)}{\sum_{i=1}^m I(T_i \geq t)} \tag{9}$$

$$\boldsymbol{\Sigma}_b(t) \leftarrow \frac{\sum_{i=1}^m \widetilde{\mathrm{E}}_{i,t}\big[(\mathbf{b}_i(t) - \mathbf{c}(t))^{\otimes 2}\big] I(T_i \geq t)}{\sum_{i=1}^m I(T_i \geq t)} \tag{10}$$

$$\boldsymbol{\Sigma}_e(t) \leftarrow \frac{\sum_{i=1}^m \sum_{j=1}^{n_i} \widetilde{\mathrm{E}}_{i,t}\big[(\mathbf{y}_{ij} - \mathbf{X}(t_{ij} - t)\mathbf{b}_i(t))^{\otimes 2}\big] K_h(t_{ij} - t) I(T_i \geq t)}{\sum_{i=1}^m \sum_{j=1}^{n_i} K_h(t_{ij} - t) I(T_i \geq t)}, \tag{11}$$

where $\mathbf{x}^{\otimes 2} = \mathbf{x}\mathbf{x}^\mathsf{T}$ denotes the outer product of a vector with itself. To update parameters in the survival model, we only need to focus on the survival part in the likelihood function since the longitudinal part does not involve the parameter $\boldsymbol{\beta}$. It is often easier to work with the Cox log partial likelihood defined in (5), which results in

$$\log \mathrm{pl}(\boldsymbol{\beta}) = \sum_{i=1}^m \Delta_i \Bigg[ \log \lambda_0\{T_i\} + \boldsymbol{\beta}^\mathsf{T} \mathbf{X}(0)\mathbf{b}_i(T_i) - \lambda_0\{T_i\} \sum_{l=1}^m \exp\{\boldsymbol{\beta}^\mathsf{T} \mathbf{X}(0)\mathbf{b}_l(T_i)\} I(T_l \geq T_i) \Bigg],$$

where $\mathbf{m}_i(t)$ in (5) has been replaced by its corresponding local polynomial estimator $\mathbf{X}(0)\mathbf{b}_i(t)$. The conditional expectations of the log partial-likelihood is then

$$\widetilde{\mathrm{E}}[\log \mathrm{pl}(\boldsymbol{\beta})] =$$
$$\sum_{i=1}^{m} \Delta_i \left[ \log \lambda_0\{T_i\} + \boldsymbol{\beta}^{\mathsf{T}} \widetilde{\mathrm{E}}_{i,T_i}[\mathbf{X}(0)\mathbf{b}_i(T_i)] - \lambda_0\{T_i\} \sum_{l=1}^{m} \widetilde{\mathrm{E}}_{l,T_i}\left[ \exp\{\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}(0)\mathbf{b}_l(T_i)\} \right] I(T_l \geq T_i) \right].$$

By maximizing the above expected log-partial likelihood and using the Fisher-Scoring approach, we can update the parameters in the survival model as follows:

$$\lambda_0\{T_i\} \leftarrow \frac{\Delta_i}{\sum_{l=1}^{m} \widetilde{\mathrm{E}}_{l,T_i}[\exp\{\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}(0)\mathbf{b}_l(T_i)\}]I(T_l \geq T_i)} \tag{12}$$

$$\boldsymbol{\beta} \leftarrow \boldsymbol{\beta} + \mathbf{I}(\boldsymbol{\beta})^{-1}\mathbf{S}(\boldsymbol{\beta}), \tag{13}$$

where

$$\mathbf{S}(\boldsymbol{\beta}) = \sum_{i=1}^{m} \Delta_i \left\{ \widetilde{\mathrm{E}}_{i,T_i}[\mathbf{X}(0)\mathbf{b}_i(T_i)] - \lambda_0\{T_i\} \sum_{l=1}^{m} \widetilde{\mathrm{E}}_{l,T_i}\left[ \mathbf{X}(0)\mathbf{b}_l(T_i) \exp\left\{\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}(0)\mathbf{b}_l(T_i)\right\} \right] I(T_l \geq T_i) \right\},$$

$$\mathbf{I}(\boldsymbol{\beta}) = \sum_{i=1}^{m} \Delta_i \left\{ \lambda_0\{T_i\} \sum_{l=1}^{m} \widetilde{\mathrm{E}}_{l,T_i}\left[ [\mathbf{X}(0)\mathbf{b}_l(T_i)]^{\otimes 2} \exp\{\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}(0)\mathbf{b}_l(T_i)\} \right] I(T_l \geq T_i) \right.$$
$$\left. - \left\{ \lambda_0\{T_i\} \sum_{l=1}^{m} \widetilde{\mathrm{E}}_{l,T_i}\left[ \mathbf{X}(0)\mathbf{b}_l(T_i) \exp\{\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}(0)\mathbf{b}_l(T_i)\} \right] I(T_l \geq T_i) \right\}^{\otimes 2} \right\}$$

are the score and information functions of $\boldsymbol{\beta}$, respectively.

To make a simultaneous inference on the time-varying parameters $\mathbf{c}(t)$, $\boldsymbol{\Sigma}_b(t)$, $\boldsymbol{\Sigma}_e(t)$ and the time-independent parameters $\boldsymbol{\beta}$ in the joint model, we use the updating algorithm described below, which can be viewed as a variation of the kernel profile estimation algorithm discussed by Yu and Lin [20]. First, given an estimate of $\boldsymbol{\beta}$, we can update the time-varying parameters $\mathbf{c}(t)$, $\boldsymbol{\Sigma}_b(t)$ and $\boldsymbol{\Sigma}_e(t)$ by following (9)–(11) for each $t$ in a set of time points that contains all the event times $\{T_i : \Delta_i = 1\}$. Then, to update the time-independent parameters $\boldsymbol{\beta}$, we can use the Newton-Raphson procedure given in (12) and (13). To obtain the final parameter estimates of the model, we can iterate between the above two steps until convergence. In this procedure, we can also obtain an estimate of $\mathbf{m}_i(t)$ by calculating $\mathbf{X}(0)\widetilde{\mathrm{E}}_{i,t}[\mathbf{b}_i(t)]$, which is an analogy of the best linear unbiased predictor. [8]

It should be pointed out that besides the local EM algorithm, we can also consider model estimation by using the local penalized quasi-likelihood (PQL) described below. The main idea is to modify the penalized quasi-likelihood approach [21,22] properly to fit the proposed joint model. The penalized quasi-likelihood discussed in the existing literature [23,21,24,22] is one of the most popular method for estimating generalized linear mixed-effects models. Instead of working with the likelihood functions that involve integrations, they proposed to apply the Laplacian approximation to the log-likelihood function, which leads to an additional penalty term. As an approximation method for estimating generalized linear mixed-effects models, the PQL approach will give slightly biased estimates in finite-sample cases [25,22]. But, because of its relatively simple computation, the PQL approach has become a popular model estimation method in the existing statistical software packages (e.g., R-package lme4 [26]; SAS GLIMMIX [27]). Next, we describe this method briefly, and a more detailed description about its implementation can be found in Appendix A.2. Conditional on $\mathbf{b}_i(t)$, the locally weighted log joint probability density functions of $\mathcal{Y}_i$ and $S_i$ are given by $\log f_{1,t}(\mathcal{Y}_i|\mathbf{b}_i(t))$ and $\log f_{2,t}(S_i|\mathbf{b}_i(t))$ (c.f., (3) and (4)). By the PQL method, an extra term needs to be added to $\log f_{3,t}(\mathbf{b}_i(t))$ to approximate the integrated local likelihood as follows:

$$\log f_{3,t}^{p}(\mathbf{b}_i(t)|\mathbf{c}(t), \boldsymbol{\Sigma}_b(t)) = -\frac{1}{2}[\mathbf{b}_i(t) - \mathbf{c}(t)]^{\mathsf{T}}\boldsymbol{\Sigma}_b(t)^{-1}[\mathbf{b}_i(t) - \mathbf{c}(t)]I(T_i \geq t)$$
$$- \log \left| \mathbf{A}_i(t)\boldsymbol{\Sigma}_b(t) + \mathbf{I}_{pr \times pr} \right| I(T_i \geq t),$$

where

$$\mathbf{A}_i(t) = -\frac{\partial^2 \log f_{1,t}(\mathcal{Y}_i|\mathbf{b}_i(t); \boldsymbol{\Sigma}_e(t))}{\partial \mathbf{b}_i(t)^{\mathsf{T}} \partial \mathbf{b}_i(t)} - \frac{\partial^2 \log f_{2,t}(S_i|\mathbf{b}_i(t); \boldsymbol{\beta}, \lambda_0)}{\partial \mathbf{b}_i(t)^{\mathsf{T}} \partial \mathbf{b}_i(t)}.$$

The locally-weighted log penalized quasi-likelihood function is then given by

$$\log L_t^{p}(\mathbf{c}(t), \boldsymbol{\Sigma}_b(t), \boldsymbol{\Sigma}_e(t)) = \sum_{i \in R(t)} \left[ \log f_{1,t}(\mathcal{Y}_i|\mathbf{b}_i(t); \boldsymbol{\Sigma}_e(t)) + \log f_{2,t}(S_i|\mathbf{b}_i(t); \boldsymbol{\beta}, \lambda_0) + \log f_{3,t}^{p}(\mathbf{b}_i(t)|\mathbf{c}(t), \boldsymbol{\Sigma}_b(t)) \right].$$

We can use the Newton-Raphson algorithm to maximize the above local log likelihood function and estimate the time-varying parameters $\mathbf{c}(t)$, $\mathbf{\Sigma}_b(t)$ and $\mathbf{\Sigma}_e(t)$ in the model. The algorithm for finding the estimates are presented in (A.1)–(A.5) in Appendix A.2. To estimate the time-independent parameters $\boldsymbol{\beta}$ in the model, we maximize the log partial-likelihood function $\log \mathrm{pl}(\boldsymbol{\beta})$, which can be achieved by using the Fisher-Scoring procedure, as in (A.6)–(A.7) of Appendix A.2. Similarly, to make a simultaneous inference on the time-varying and time-independent parameters in the joint model, we can update parameters in the following way. First, given an estimate of $\boldsymbol{\beta}$, we update the time-varying parameters $\mathbf{c}(t)$, $\mathbf{\Sigma}_b(t)$, and $\mathbf{\Sigma}_e(t)$ by following (A.1)–(A.5) in Appendix A.2 for each $t$ in a set of time points that includes all the event times $\{T_i : \Delta_i = 1\}$. Second, we follow (A.6)–(A.7) in Appendix A.2 to update the time-independent parameters $\boldsymbol{\beta}$ in the joint model. In this procedure, we can also obtain an estimate of $\mathbf{m}_i(t)$ from $\mathbf{X}(0)\widehat{\mathbf{b}}_i(t)$, where $\widehat{\mathbf{b}}_i(t)$ is the estimated $\mathbf{b}_i(t)$ by the algorithm.

# 3 | THEORETICAL PROPERTIES AND PRACTICAL GUIDELINES

## 3.1 | Asymptotic properties

We derive some asymptotic properties of the proposed joint modeling method. Let $\boldsymbol{\beta}_0$ be the true coefficients in the Cox proportional hazards model (2). Then, the following theorem gives the statistical consistency of the estimated coefficients by the proposed local EM algorithm.

**Theorem 1.** Let $\widehat{\boldsymbol{\beta}}_{\mathrm{EM}}$ be the estimated coefficients in model (2) by the proposed local EM algorithm (cf., (9)–(13)). If the assumptions (C1)–(C7) given in Appendix A.3 are valid, then we have $\widehat{\boldsymbol{\beta}}_{\mathrm{EM}} \xrightarrow{p} \boldsymbol{\beta}_0$.

It is worth mentioning that the above theorem confirms that we do not need to specify the distribution of the longitudinal outcomes to establish the consistency of $\widehat{\boldsymbol{\beta}}_{\mathrm{EM}}$, although the density functions of normal distributions have been used in Section 2 when we derive the local likelihood function of the joint model.

If the joint model is estimated by the local PQL method, as discussed at the end of Section 2, then the following theorem establishes the statistical consistency of the estimated coefficients.

**Theorem 2.** Let $\widehat{\boldsymbol{\beta}}_{\mathrm{PQL}}$ be the estimated coefficients by the proposed local PQL method. Then, under the assumptions (C1)–(C6) and (C7') given in Appendix A.3, we have $\widehat{\boldsymbol{\beta}}_{\mathrm{PQL}} \xrightarrow{p} \boldsymbol{\beta}_0$.

## 3.2 | Bandwidth selection

Several cross-validation methods for selecting bandwidths in local mixed-effects models with univariate longitudinal outcomes have been reviewed in Section 3 of Wu and Zhang [11]. Here, we introduce a multivariate version of the leave-one-point-out cross-validation (PTCV) method. [28,11] Let $t_1^* < \ldots < t_J^*$ be all distinct time points of the set $\{t_{ij} : i = 1, \ldots, m, j = 1, \ldots, n_i\}$. For a given time point $t_j^*$, let $\mathcal{O}(t_j^*)$ be the set of subjects who have observations at $t_j^*$, where the observed $\mathbf{y}$ is denoted as $\mathbf{y}_i(t_j^*)$. Let $\mathbf{m}_i^{(-j)}(t_j^*)$ denote the estimator of $\mathbf{m}_i(t_j^*)$ when all observations at the design point $t_j^*$ are excluded. In the local EM estimation, we let $\mathbf{m}_i^{(-j)}(t_j^*) = \mathbf{X}(0)\widetilde{\mathrm{E}}_{i,t_j^*}[\mathbf{b}_i^{(-j)}(t_j^*)]$, which is defined similarly to the best linear unbiased prediction for $\mathbf{m}_i(t_j^*)$ in the linear mixed-effects modeling where observations at $t_j^*$ are also excluded when taking the conditional expectation. In the local penalized quasi-likelihood estimation, we define $\mathbf{m}_i^{(-j)}(t_j^*) = \mathbf{X}(0)\mathbf{b}_i^{(-j)}(t_j^*)$, where $\mathbf{b}_i^{(-j)}(t_j^*)$ is the fitted value of $\mathbf{b}_i(t_j^*)$ when all observations at $t_j^*$ are excluded. Then, the PTCV score is defined as

$$\mathrm{CV}(h) = \sum_{j=1}^{J} \sum_{i \in \mathcal{O}(t_j^*)} \left\| \mathbf{y}_i(t_j^*) - \mathbf{m}_i^{(-j)}(t_j^*) \right\|^2,$$

where $\|\mathbf{x}\|$ denotes the Euclidean norm of a vector $\mathbf{x}$. The bandwidth $h$ is then selected by minimizing the PTCV score $\mathrm{CV}(h)$. This PTCV procedure requires fitting the joint model as many times as the number of distinct observation times in the observed dataset. One strategy to reduce the computational burden is to approximate $\widetilde{\mathrm{E}}_{i,t_j^*}[\mathbf{b}_i^{(-j)}(t_j^*)]$ by the expectation of $\mathbf{b}_i(t)$ conditional on the longitudinal data $\mathcal{Y}_i$, which has a closed form solution as shown in Appendix A.1. The use of this strategy is justified by some simulation results given in Supporting Information B.4.1. In cases when the scales of the longitudinal outcomes are quite different, we recommend to scale the longitudinal outcomes so that their sample variances are the same before performing the above PTCV procedure.

## 3.3 | Variable selection

In some applications, a large number of longitudinal outcomes are involved, and some of them may not be associated with the survival outcomes. In such cases, variable selection is necessary to identify important longitudinal outcomes for joint modeling. To this end, the least absolute shrinkage and selection operator (LASSO) and its variations have received wide popularity.[29,30,31] To identify important longitudinal outcomes for joint modeling, we can add the following $L^1$ penalty to the survival part of the likelihood function

$$p_\gamma(\boldsymbol{\beta}) = -\gamma \sum_{k=1}^{p} w_k |\beta_k|,$$

where $\gamma \geq 0$ is a regularization parameter, and $\{w_k\}$ are weights. In the local EM algorithm and the local penalized quasi-likelihood estimation procedure, the resulting objective functions for estimating $\boldsymbol{\beta}$ become respectively

$$\widetilde{\mathrm{E}}\big[\log \mathrm{pl}(\boldsymbol{\beta})\big] + p_\gamma(\boldsymbol{\beta}) \quad \text{and} \quad \log \mathrm{pl}(\boldsymbol{\beta}) + p_\gamma(\boldsymbol{\beta}).$$

The coordinate optimization algorithm by Simon et al[32] can be used to find the parameter estimate of $\boldsymbol{\beta}$. It is well-discussed in [31] that the adaptive weights $w_k = 1/|\widehat{\beta}_k|$ can be chosen, where $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_p)^\mathsf{T}$ is the estimate of $\boldsymbol{\beta}$ without using the penalty term. The parameter $\gamma$ in the penalty term can be selected by the cross-validation procedure discussed by van Houwelingen et al[33].

## 4 | SIMULATION STUDIES

In this section, we present some simulation results to evaluate the proposed methods. In the simulation study, we choose $n = 500$, $p = 3$, $r = 2$, and $\mathcal{T} = 1.0$. Longitudinal data are generated under three different scenarios, described below. In scenario (I), the population mean functions $\mathbf{m}_i(t) = (m_{i1}(t), \ldots, m_{ip}(t))^\mathsf{T}$ are assumed to be linear

$$m_{ik}(t) = a_{ik}^\dagger + b_{ik}^\dagger t, \quad (i = 1, \ldots, m, k = 1, \ldots, p)$$

where the intercepts $(a_{i1}^\dagger, \ldots, a_{ip}^\dagger)^\mathsf{T}$ follow a multivariate normal distribution with mean $\mathbf{1}_{p \times 1}$ and covariance matrix

$$\begin{bmatrix} 1.0 & 0.1 & 0.1 \\ 0.1 & 1.0 & 0.1 \\ 0.1 & 0.1 & 1.0 \end{bmatrix}$$

and the slopes $(b_{i1}^\dagger, \ldots, b_{ip}^\dagger)^\mathsf{T}$ follow multivariate normal distribution with mean $\mathbf{1}_{p \times 1}$ and covariance matrix $\mathbf{I}_{p \times p}$. The error covariance matrix in model (1) is specified to be $\boldsymbol{\Sigma}_e(t) = 0.2^2 \times \mathbf{I}_{p \times p}$. Observation times $t_{ij}$ follow the uniform distribution in the interval $[(j-1)/50, j/50]$ until $T_i$. The survival data are generated following Cox's proportional hazards model with $\boldsymbol{\beta} = (0.4, 0.3, 0.2)^\mathsf{T}$, $\lambda_0(t) = 0.05 I(t > 0.1)$, and $C_i = \mathcal{T}$. In scenario (II), the population mean functions $\mathbf{m}_i(t)$ are nonlinear and defined as

$$m_{ik}(t) = a_{ik}^\dagger + b_{ik}^\dagger \exp\big\{ \sin\big[\omega_{ik}^\dagger\big(t - \phi_{ik}^\dagger\big)\pi\big] \big\}, \quad (i = 1, \ldots, n, k = 1, \ldots, p)$$

where $(a_{i1}^\dagger, \ldots, a_{ip}^\dagger)^\mathsf{T}$ are independent and follow a multivariate normal distribution with mean $-2 \times \mathbf{1}_{p \times 1}$ and covariance matrix

$$\begin{bmatrix} 1.0 & 0.1 & 0.1 \\ 0.1 & 1.0 & 0.1 \\ 0.1 & 0.1 & 1.0 \end{bmatrix},$$

$\{b_{ik}^\dagger\}$ are iid with the distribution $N(3, 0.5^2)$, $\{\omega_{ik}^\dagger\}$ are iid with the distribution $N(3, 0.5^2)$, and $\{\phi_{ik}^\dagger\}$ are iid with the uniform distribution in $[0, 1]$. The error covariance matrix in model (1) is specified to be $\boldsymbol{\Sigma}_e(t) = \mathbf{I}_{p \times p}$. In this scenario, the random-effect terms $a_{ik}^\dagger$ are time-independent, and the trajectories of the longitudinal outcomes are periodic with $b_{ik}, \omega_{ik}^\dagger$ and $\phi_{ik}^\dagger$ controlling the amplitude, frequency and phase of the related periodic functions, respectively. Observation times $t_{ij}$ are assumed to follow the uniform distribution in the interval $[(j-1)/100, j/100]$ until $T_i$. The survival data are generated following the Cox proportional hazards model with $\boldsymbol{\beta} = (0.4, 0.3, 0.2)^\mathsf{T}$, $\lambda_0(t) = 0.05 I(t > 0.1)$, and $C_i = \mathcal{T}$. In scenario (III), $\mathbf{m}_i(t)$ has the following more complicated form:

$$m_{ik}(t) = a_{ik1}^\dagger \sin\big[\omega_{ik1}^\dagger\big(t - \phi_{ik1}^\dagger\big)\pi\big] + a_{ik2}^\dagger \log\big\{2 + \sin\big[\omega_{ik2}^\dagger\big(t - \phi_{ik2}^\dagger\big)\pi\big]\big\} + a_{ik3}^\dagger \exp\big\{ \sin\big[\omega_{ik3}^\dagger\big(t - \phi_{ik3}^\dagger\big)\pi\big]\big\},$$

where $\{a_{ikl}^\dagger : i = 1, \ldots, m; \ k = 1, \ldots, p; \ l = 1, 2, 3\}$ are i.i.d. with the common $N(1, 0.2^2)$ distribution, $\{\phi_{ikl}^\dagger : i = 1, \ldots, m; \ k = 1, \ldots, p; \ l = 1, 2, 3\}$ are iid with the uniform distribution in $[0, 1]$, $\omega_{ik1}^\dagger$, $\omega_{ik2}^\dagger$ and $\omega_{ik3}^\dagger$ are iid with the uniform

distributions in the intervals [1, 3], [3, 5] and [5, 7], respectively. In this scenario, the trajectories of the longitudinal outcomes are combinations of several complex periodic functions, and $a_{ik}^{\dagger}$, $\omega_{ik}^{\dagger}$ and $\phi_{ik}^{\dagger}$ control the amplitude, frequency and phase of the related periodic functions, respectively. The error covariance matrix in model (1) is specified to be $\Sigma_e(t) = \mathbf{I}_{p \times p}$. Observation times $t_{ij}$ are assumed to follow the uniform distribution in the interval $[(j-1)/100, j/100]$ until $T_i$. The survival data are generated from the Cox proportional hazards model with $\beta = (0.4, 0.3, 0.2)^{\mathsf{T}}$, $\lambda_0(t) = 0.05 I(t > 0.1)$, and $C_i = \mathcal{T}$.

In all simulation studies, we compare the proposed methods with two existing joint modeling methods that use the linear functions to describe the longitudinal trajectories [1,34], and the linear combinations of cubic B-spline basis functions to describe the longitudinal trajectories [4,5,35]. These two methods are denoted as "Linear" and "B-spline", respectively. For the B-spline method, Brown et al[4] and Yao[5] suggested using evenly-spaced knots for the B-spline basis functions. In such cases, the only tuning parameter in the model is the number of basis functions $q$. When $q \geq 5$, the basis functions are the cubic B-splines with equally spaced knots in $[0, \mathcal{T}]$, and when $q = 2, 3, 4$, the basis functions are linear, quadratic or cubic polynomial basis functions. By following the suggestions in Yao[5] and Park and Qiu[36], the tuning parameter $q$ is selected by minimizing the AIC criterion defined to be $-2 \log L(\mathbf{c}, \Sigma_b, \Sigma_e, \beta) + 2[pq + pq(pq+1)/2 + p(p+1)/2 + p]$, where the quantity in the square brackets is the number of parameters in the related model. The proposed methods by the local EM algorithm and the local PQL procedure are denoted as "Local EM" and "Local PQL", respectively. In these two methods, the bandwidth parameter $h$ is selected by the method described in Section 3.2, where the conditional expectation is taken with respect to the longitudinal data only to reduce the computational burden.

We evaluate the performance of the related methods based on 1,000 replicated simulations. The tuning parameters $q$ in B-spline and $h$ in Local EM and Local PQL are selected based on the first 20 simulated datasets. In Table 1, the mean of the estimated coefficients $\beta = (\beta_1, \ldots, \beta_p)^{\mathsf{T}}$ are presented, along with their corresponding relative biases and standard errors. In Table 2, we present the MSE (i.e., mean squared error) for $\widehat{\beta}$ and the MISE (mean integrated squared error) for both $\widehat{\mathbf{m}}_i(t)$ and $\widehat{\boldsymbol{\mu}}_c(t)$ under the three scenarios. Here, the MSE for $\widehat{\beta}$ is defined by $\mathrm{E}\left[\|\widehat{\beta} - \beta\|^2\right]$, the MISE for $\widehat{\mathbf{m}}_i(t)$ is defined by $\mathrm{E}\left[\sum_{i=1}^{m} \int_0^{T_i} \|\widehat{\mathbf{m}}_i(t) - \mathbf{m}_i(t)\|^2 \, dN(t)\right]$, and the MISE for $\widehat{\boldsymbol{\mu}}_c(t)$ is defined by $\mathrm{E}\left[\int_0^{\mathcal{T}} \|\widehat{\boldsymbol{\mu}}_c(t) - \boldsymbol{\mu}_c(t)\|^2 \, dN(t)\right]$. From the tables, we can have the following conclusions. In Scenario (I), B-spline correctly identifies the linearity of trajectories by tuning the selected parameter $q$ using AIC, and thus yields the same results as those by Linear. For the two proposed methods, a large bandwidth parameter $h$ is selected by the proposed bandwidth selection procedure, reducing the local polynomial mixed-effects model to a global polynomial mixed-effects model. Thus, the MSE values of $\widehat{\beta}$ are all similar for the four methods, while Linear and B-spline are slightly better in terms of MISE for $\widehat{\mathbf{m}}_i(t)$ and $\widehat{\boldsymbol{\mu}}_c(t)$. In Scenarios (II) and (III), Linear performed markedly worse than the other three nonparametric methods, as the assumed linear model is inadequate to describe the nonlinear trajectories of the longitudinal data. In terms of MSE of $\widehat{\beta}$, Local EM performs the best among the four methods, and Local PQL is slightly worse than the other two nonparametric models. The compromised performance of Local PQL models in estimation of $\beta$ can be explained by the estimation bias of the PQL methods, which is a phenomenon that has been discussed in the literature [25,22]. In terms of MISE for $\mathbf{m}_i(t)$ and $\widehat{\boldsymbol{\mu}}_c(t)$, the two local likelihood methods, Local PQL and Local EM outperform the B-spline method. This simulation study shows that the proposed local likelihood methods are more flexible in describing the trajectories of the longitudinal outcomes. For the four methods, we also recorded their computation times, and the results are presented in Figure 1. From the plots in the figure, it can be seen that in Scenario (I) when the longitudinal trajectories are linear, the B-spline method is as computationally efficient as the Linear method, the Local PQL method is nearly as computationally efficient as the Linear method, and the Local EM method tends to be time-consuming. In Scenario (II) and (III) when the true longitudinal trajectories are nonlinear, Linear is still computationally efficient, but its performance is poor, as shown in Table 1. In these two scenarios, B-spline is the most time-consuming among the four methods because it needs more basis functions to describe the longitudinal trajectories reasonably well. In comparison, Local EM and Local PQL take much less computing times. In particular, Local PQL is substantially faster than Local EM and B-spline, and comparable with Linear.

Next, we use a simple example to evaluate the performance of the proposed joint modeling methods with the LASSO variable selection component (cf., Section 3.3). In the example, the dimension of the longitudinal outcomes is set to be $p = 10, 15$, or $20$. Due to the large dimension of the random-effects in the joint model, the Local PQL method is considered in this example, since it is more computationally efficient than the Local EM method, as shown in Figure 1. The true coefficients in the Cox model (2) is set to be $\beta = (0.5, 0.4, 0.3, 0.0, \ldots, 0.0)^{\mathsf{T}}$, which implies that only the first three longitudinal outcomes are associated with the survival outcome. The population mean functions $\mathbf{m}_i(t)$ are assumed to be

$$m_{ik}(t) = b_{ik}^{\dagger} \sin\left[2\pi\left(\omega_{ik}^{\dagger} t - \phi_{ik}^{\dagger}\right)\right],$$
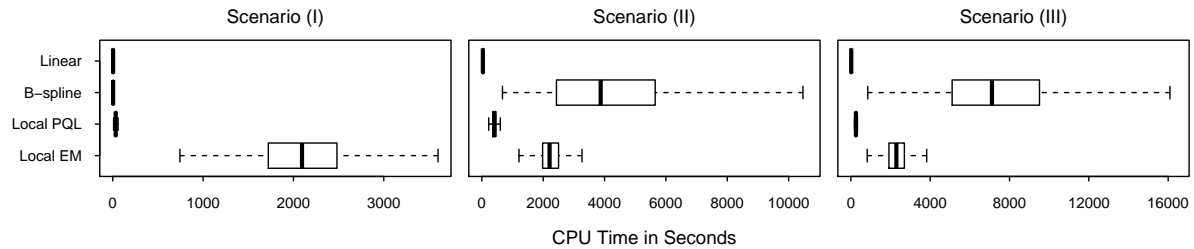
**FIGURE 1** Boxplots of the computing times of the four methods Linear, B-spline, Local PQL, and Local EM in the example of Table 1.

where $b_{ik}^{\dagger}$ follow the $N(3, 0.5^2)$ distribution, $\omega_{ik}$ follow the $N(1, 0.1^2)$ distribution, and $\phi_{ik}^{\dagger}$ follow the uniform distribution in $[0, 1]$. The error covariance matrix is set to be $\Sigma_e(t) = 0.5^2 \times I_{p \times p}$. The observation times $t_{ij}$ are generated from the uniform distribution in the interval $[(j-1)/100, j/100]$ with upper bound of $T_i$. Other settings are the same as those in the previous example. The performance of the proposed method is evaluated by the MSE of the estimated coefficients $\widehat{\beta}$ with and without using the LASSO variable selection. The results are presented in Table 3, where relative efficiency is defined to be the ratio of "MSE without LASSO" and "MSE with LASSO".

When the dimension of the longitudinal outcomes $p$ increases, the proportion of non-zero coefficients in $\beta$ decreases. From Table 3, it can be seen that (i) the relative efficiency values are above 1 in all cases considered, and (ii) the relative efficiency is larger when $p$ is larger. Thus, this example shows that the LASSO variable selection can indeed improve the efficiency of the estimated model, and the improvement is more significant when the proportion of non-zero coefficients in $\beta$ gets smaller.

# 5 | CASE STUDY

In this section, we demonstrate the proposed methods by applying them to a real dataset obtained from the Framingham Heart Study. The data set contained the systolic and diastolic blood pressure readings of $m = 1055$ patients, among which 27 patients had strokes at the last follow-up times. The last follow-up times range between 23 and 30 years, and their average is 26.23 years. In this example, the blood pressure data are used as the longitudinal outcomes and the occurrence of strokes is used as the event to define the survival outcomes. Each patient was followed 7 times, and the blood pressure readings were obtained at each follow-up time. The mean baseline systolic and diastolic blood pressures are 120.13 (SD=14.07) and 77.82 (SD=9.87), respectively. The observation times for different patients are aligned by patients' ages in this study. Since patients entered the study at different ages, we need to modify the definition of the risk set $R(t)$ to account for left truncation in the data analysis. Let $E_i$ be the age at which the $i$th patient entered the study, and we define $R(t) = \{i : E_i \leq t \leq T_i\}$. The original dataset is displayed in Figure 2. Before the data analysis, the systolic and diastolic blood pressure data are scaled to have the sample mean of 0 and the sample variance of 1. To justify the use of the proposed methods, the Cox proportional hazards assumption is tested using the test discussed in Park and Qiu[36], which extended the test statistics by Grønnesby and Borgan[37] and May and Hosmer[38] to a joint modeling setting. To use this testing method, we first define the integrated martingale residual of the $i$th patient by

$$r_i = \Delta_i - \int_{E_i}^{T_i} \widehat{\lambda}_0(t) \exp\{\widehat{\beta}^{\top} \widehat{m}_i(t)\} \, dt,$$

where $\widehat{\beta}$, $\widehat{m}_i(t)$ and $\widehat{\lambda}_0(t)$ are the estimated parameter values of the joint model. Then, all patients are divided into 10 groups of similar sizes based on the ranks of the estimated baseline risks $\{\widehat{\beta}^{\top} \widehat{m}_i(E_i), i = 1, \ldots, m\}$. Let $H_k$ be the sum of $r_i$ in the $k$th group. Then, the following test statistic should approximately follow a $\chi^2$-distribution with 9 degrees of freedom under Cox's proportional hazards assumption

$$T_{\mathrm{PH}} = (H_1, \ldots, H_9) \Sigma_H^{-1} (H_1, \ldots, H_9)^{\top},$$

where $\Sigma_H$ is an estimate of the covariance matrix of $(H_1, \ldots, H_9)$ whose formula can be found in Grønnesby and Borgan[37] and the supplementary materials. In the above expression, $H_{10}$ is omitted since $\sum_{k=1}^{10} H_k = 0$. This test gives a test statistic value of $T_{\mathrm{PH}} = 7.376$ with the p-value of 0.598. So, the Cox proportional hazards assumption cannot be rejected by this test.

**FIGURE 2** Readings of the systolic blood pressure and the diastolic blood pressure in the stroke dataset from the Framingham Heart Study. Gray thin dashed lines represent the observed data of the stroke patients, while gray thin solid lines represent data of the non-stroke patients. Black dashed bold curves and black solid bold curves in each plot are the estimated mean curves of the blood pressure trajectories among stroke and non-stroke patients, respectively.
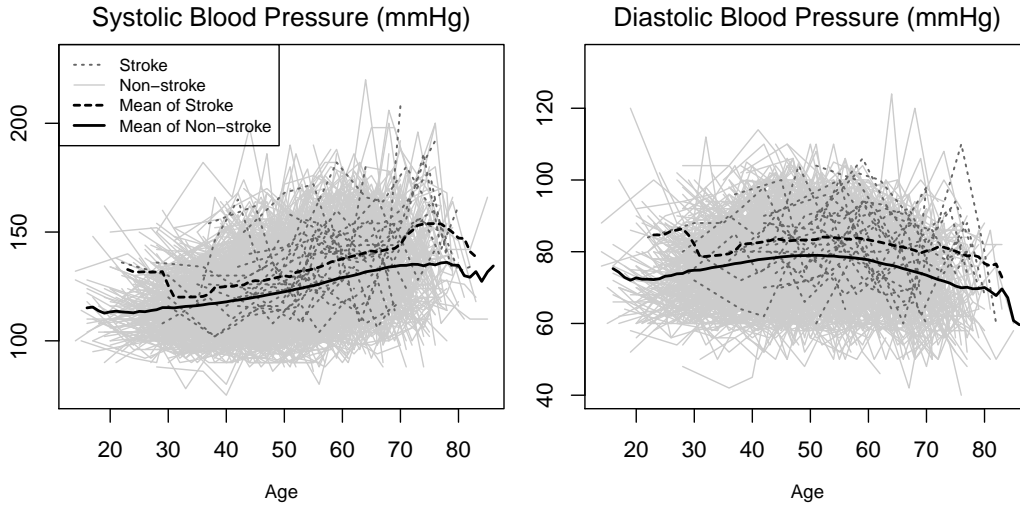


The four different methods Linear, B-spline, Local PQL, and Local EM are then applied to the dataset. Similar to the simulation study, the bandwidth $h$ in the two local approaches is selected by the PTCV criterion where the approximation discussed in Section 3.2 is used to reduce the computational burden, and the number of basis functions $q$ in B-spline is selected by the AIC. The selected $h$ is 21 years and the selected $q$ is 4 in this example. In Figure 3, the fitted longitudinal trajectories of all patients by the four different methods are displayed, along with the fitted trajectories of 8 randomly selected patients. From the plots in the figure, we can see that Linear suffers from the shortcoming of overly simplifying the longitudinal trajectories in certain cases and thus cannot provide a good fit to the observed data. Among the three nonparametric methods B-spline, Local PQL, and Local EM, the last two methods exhibit better flexibility in fitting the longitudinal data. To evaluate the longitudinal fits of the four different methods, we further compared their cross-validated mean squared prediction errors (CVMSPE) of the longitudinal outcomes. To this end, for each $i$ and $j$ ($1 \le i \le m$ and $1 \le j \le n_i$), we define $\widetilde{\mathbf{m}}_i^{-(ij)}(t_{ij})$ to be the predicted values of $\mathbf{m}_i(t_{ij})$ when the observation $\mathbf{y}_{ij}$ is excluded from the model fitting. The CVMSPE value is then defined by

$$\text{CVMSPE} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left\| \mathbf{y}_{ij} - \widetilde{\mathbf{m}}_i^{-(ij)}(t_{ij}) \right\|^2.$$

The CVMSPE values by Linear, B-spline, Local PQL, and Local EM are 1.0694 (0.0192), 1.0304 (0.0189), 1.0236 (0.0190), 1.0071 (0.0186), respectively, where the values in parentheses are the corresponding standard errors. Thus, in terms of CVMSPE, Linear performs worse than the three nonparametric methods. Among the three nonparametric methods, Local EM performs slightly better than the other two methods.

To evaluate the survival fit, we compare the mean squared integrated martingale residuals of the four different methods. To this end, for each $i$, a model is fitted from the data without the $i$th patient. Then, we define $r_i^{-(i)} = \Delta_i - \int_{E_i}^{T_i} \widehat{\lambda}_0^{-(i)}(t) \exp\{\widehat{\boldsymbol{\beta}}^{-(i)\mathsf{T}} \widehat{\mathbf{m}}_i^{-(i)}(t)\} \, dt$ as the integrated martingale residual of the $i$th patient, where $\widehat{\lambda}_0^{-(i)}(t)$, $\widehat{\boldsymbol{\beta}}^{-(i)}$ and $\widehat{\mathbf{m}}_i^{-(i)}(t)$ are the fitted values of $\lambda_0(t)$, $\boldsymbol{\beta}$ and $\mathbf{m}_i(t)$ from the estimated model without using the observations of the $i$th patient. The mean squared integrated martingale residuals $\frac{1}{m} \sum_{i=1}^{m} \{r_i^{-(i)}\}^2$ by Linear, B-spline, Local PQL, and Local EM are 0.02516 (0.00429), 0.02482 (0.00424), 0.02495 (0.00424), and 0.02397 (0.00438), respectively, where the numbers in parentheses are the standard errors. It can be seen that Local EM has the smallest mean squared martingale residuals value, while Linear has the largest value. These results also suggest that Local EM provides a better fit to the observed survival data, compared to the other three methods, though the differences among the four methods are quite small, due to a small number of events in this dataset.

## 6 | DISCUSSION AND CONCLUDING REMARKS

In this paper, we propose a local kernel smoothing method for joint modeling of multivariate nonparametric longitudinal data and survival data, where a multivariate local polynomial mixed-effects model is used for describing the longitudinal data, and the Cox's proportional hazards model is used to describe the association between the longitudinal and the survival outcomes. Two estimation procedures, i.e., the local EM algorithm and the local penalized quasi-likelihood estimation procedure, are suggested to estimate the joint model. Theoretical properties of the estimated models are provided and practical guidelines for implementing the proposed methods are provided. Through some simulation studies and a real data application, the numerical performance of the proposed methods has been evaluated. It has been confirmed by these numerical studies that the proposed methods perform better than some representative existing methods in most cases considered, since the proposed methods are more flexible in model fitting. In this paper, the proposed methods are based on the standard Cox's proportional hazards model (2), which can be generalized in several different ways. For instance, we may want to include some time-independent covariates (e.g., patients' demographic or baseline information in the stroke example discussed in Section 5) in that model. That model can also be generalized to a recurrent-events model or a competing-risk model in cases when there are recurrent survival outcomes or multiple survival outcomes. There are some other possibilities to explore in future research. For instance, this paper focuses on cases with continuous longitudinal outcomes only. In some applications, however, there could be a mixture of continuous numerical, discrete numerical, or even categorical longitudinal outcomes. It requires much future research to accommodate such cases in the proposed methods. Furthermore, dynamic prediction models have been popular recently in the joint modeling research, where the future survival of a subject can be predicted using the history data[39]. Such dynamic prediction models should have important applications for disease screening. However, due to the local nature of the proposed methods, the relationship between longitudinal outcomes at two different time points is not estimated directly, and thus future research is needed to modify the proposed methods properly to predict future trajectories given the history data.

## ACKNOWLEDGMENTS

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

## DATA AVAILABILITY STATEMENT

The data in the case study are available from the corresponding author upon request. The R package for implementing the proposed methods can be found online in the Supporting Information section.

## References

1. Wulfsohn MS, Tsiatis AA. A joint model for survival and longitudinal data measured with error. *Biometrics.* 1997; 67(1): 330–339.

2. Chi YY, Ibrahim JG. Joint models for multivariate longitudinal and multivariate survival data. *Biometrics.* 2006; 62(2): 432–445.

3. Yang L, Yu M, Gao S. Joint models for multiple longitudinal processes and time-to-event outcome. *J Stat Comput Sim.* 2016; 86(18): 3682–3700.

4. Brown ER, Ibrahim JG, DeGruttola V. A flexible B-spline model for multiple longitudinal biomarkers and survival. *Biometrics.* 2005; 61(1): 64–73.

5. Yao F. Functional principal component analysis for longitudinal and survival data. *Stat Sinica.* 2007: 965–983.

6. Andersen PK, Borgan O, Gill RD, Keiding N. *Statistical Models Based on Counting Processes*. Springer Science & Business Media . 1993.

7. Murphy SA, Vaart V. dAW. On profile likelihood. *J Am Stat Assoc.* 2000; 95(450): 449–465.

8. Henderson CR. Best linear unbiased estimation and prediction under a selection model. *Biometrics.* 1975: 423–447.

9. Shi M, Weiss RE, Taylor JM. An analysis of paediatric CD4 counts for acquired immune deficiency syndrome using flexible random curves. *J R Stat Soc Ser C Appl Stat.* 1996; 45(2): 151–163.

10. Brumback BA, Rice JA. Smoothing spline models for the analysis of nested and crossed samples of curves. *J Am Stat Assoc.* 1998; 93(443): 961–976.

11. Wu H, Zhang JT. Local polynomial mixed-effects models for longitudinal data. *J Am Stat Assoc.* 2002; 97(459): 883–897.

12. Xiang D, Qiu P, Pu X. Local polynomial regression analysis of multivariate longitudinal data. *Stat Sinica.* 2013; 23: 769–789.

13. Epanechnikov VA. Non-parametric estimation of a multivariate probability density. *Theor Probab Appl.* 1969; 14: 153–158.

14. Fan J, Gijbels I. Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *J R Stat Soc Ser B Methodol.* 1995; 57(2): 371–394.

15. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Ser B Methodol.* 1977; 39(1): 1–22.

16. Betensky RA, Lindsey JC, Ryan LM, Wand MP. Local EM estimation of the hazard function for interval-censored data. *Biometrics* 1999; 55(1): 238–245.

17. Braun J, Duchesne T, Stafford JE. Local likelihood density estimation for interval censored data. *Can J Stat.* 2005; 33(1): 39–60.

18. Fan CPS, Stafford J, Brown PE. Local-EM and the EMS Algorithm. *J Comp Graph Stat.* 2011; 20(3): 750–766.

19. Nguyen P, Brown P, Stafford J. Mapping cancer risk in southwestern Ontario with changing census boundaries. *Biometrics* 2012; 68(4): 1228–1237.

20. Yu Z, Lin X. Semiparametric regression with time-dependent coefficients for failure time data analysis. *Stat Sinica.* 2010; 20(2): 853–869.

21. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. *J Am Stat Assoc.* 1993; 88(421): 9–25.

22. Lin X, Breslow NE. Bias correction in generalized linear mixed models with multiple components of dispersion. *J Am Stat Assoc.* 1996; 91(435): 1007–1016.

23. Schall R. Estimation in generalized linear models with random effects. *Biometrika.* 1991; 78(4): 719–727.

24. Wolfinger R, O'Connell M. Generalized linear mixed models a pseudo-likelihood approach. *J Stat Comput Sim.* 1993; 48(3): 233–243.

25. Pinheiro JC, Bates DM. Approximations to the log-likelihood function in the nonlinear mixed-effects model. *J Comp Graph Stat.* 1995; 4(1): 12–35.

26. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. *J Stat Softw.* 2015.

27. Schabenger O. *Introducing the GLIMMIX Procedure for Generalized Linear Mixed Models*. Cary, NC: SAS Institute Inc . 2005.

28. Altman NS. Kernel smoothing of data with correlated errors. *J Am Stat Assoc.* 1990; 85(411): 749–759.

29. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B Methodol.* 1996; 58(1): 267–288.

30. Tibshirani R. The lasso method for variable selection in the Cox model. *Stat Med.* 1997; 16(4): 385–395.

31. Zou H. The adaptive lasso and its oracle properties. *J Am Stat Assoc.* 2006; 101(476): 1418–1429.

32. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for Cox's proportional hazards model via coordinate descent. *J Stat Softw.* 2011; 39(5): 1.

33. Van Houwelingen HC, Bruinsma T, Hart AA, Veer vLJ, Wessels LF. Cross-validated Cox regression on microarray gene expression data. *Stat Med.* 2006; 25(18): 3201–3216.

34. Rizopoulos D, Verbeke G, Molenberghs G. Multiple-imputation-based residuals and diagnostic plots for joint models of longitudinal and survival outcomes. *Biometrics.* 2010; 66(1): 20–29.

35. Rizopoulos D, Ghosh P. A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Stat Med.* 2011; 30(12): 1366–1380.

36. Park KY, Qiu P. Model selection and diagnostics for joint modeling of survival and longitudinal data with crossing hazard rate functions. *Stat Med.* 2014; 33(26): 4532–4546.

37. Grønnesby JK, Borgan Ø. A method for checking regression models in survival analysis based on the risk score. *Lifetime Data Anal.* 1996; 2(4): 315–328.

38. May S, Hosmer DW. A simplified method of calculating an overall goodness-of-fit test for the Cox proportional hazards model. *Lifetime Data Anal.* 1998; 4(2): 109–120.

39. Rizopoulos D. Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics.* 2011; 67(3): 819–829.

**TABLE 1** Estimated coefficients, their relative biases, empirical standard errors, and model-based standard errors by the four methods "Linear", "B-spline", "Local PQL" and "Local EM" in Scenarios (I)–(III). The true regression coefficients are $\beta_1 = 0.4$, $\beta_2 = 0.3$, and $\beta_3 = 0.2$.

| | Estimate | | | Relative Bias | | | Monte Carlo SE | | | Model-Based SE | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_1$ | $\beta_2$ | $\beta_3$ |
| **Scenario (I)** | | | | | | | | | | | | |
| Linear | 0.3998 | 0.3041 | 0.1995 | 0.0004 | 0.0137 | 0.0025 | 0.0028 | 0.0027 | 0.0029 | 0.0863 | 0.0861 | 0.0862 |
| B-spline | 0.3998 | 0.3041 | 0.1995 | 0.0004 | 0.0137 | 0.0025 | 0.0028 | 0.0027 | 0.0029 | 0.0863 | 0.0861 | 0.0862 |
| Local PQL | 0.4002 | 0.3044 | 0.1998 | 0.0006 | 0.0147 | 0.0009 | 0.0028 | 0.0027 | 0.0029 | 0.0863 | 0.0862 | 0.0863 |
| Local EM | 0.3998 | 0.3041 | 0.1996 | 0.0006 | 0.0135 | 0.0020 | 0.0028 | 0.0027 | 0.0029 | 0.0863 | 0.0861 | 0.0862 |
| **Scenario (II)** | | | | | | | | | | | | |
| Linear | 0.5575 | 0.4124 | 0.2747 | 0.3937 | 0.3748 | 0.3735 | 0.0018 | 0.0018 | 0.0019 | 0.0381 | 0.0383 | 0.0385 |
| B-spline | 0.4168 | 0.3099 | 0.2056 | 0.0421 | 0.0331 | 0.0282 | 0.0009 | 0.0009 | 0.0008 | 0.0264 | 0.0249 | 0.0241 |
| Local PQL | 0.4523 | 0.3380 | 0.2247 | 0.1307 | 0.1266 | 0.1233 | 0.0011 | 0.0010 | 0.0009 | 0.0281 | 0.0262 | 0.0252 |
| Local EM | 0.4106 | 0.3070 | 0.2039 | 0.0266 | 0.0234 | 0.0194 | 0.0009 | 0.0009 | 0.0008 | 0.0263 | 0.0249 | 0.0242 |
| **Scenario (III)** | | | | | | | | | | | | |
| Linear | 0.5657 | 0.4314 | 0.2835 | 0.4144 | 0.4379 | 0.4173 | 0.0087 | 0.0090 | 0.0090 | 0.1696 | 0.1697 | 0.1699 |
| B-spline | 0.4137 | 0.3113 | 0.2074 | 0.0341 | 0.0376 | 0.0368 | 0.0028 | 0.0027 | 0.0026 | 0.0731 | 0.0729 | 0.0728 |
| Local PQL | 0.4634 | 0.3520 | 0.2328 | 0.1586 | 0.1734 | 0.1638 | 0.0031 | 0.0030 | 0.0028 | 0.0772 | 0.0771 | 0.0771 |
| Local EM | 0.4104 | 0.3117 | 0.2062 | 0.0259 | 0.0389 | 0.0312 | 0.0027 | 0.0026 | 0.0025 | 0.0731 | 0.0731 | 0.0732 |

**TABLE 2** MSE values for $\widehat{\boldsymbol{\beta}}$, and MISE values for $\widehat{\mathbf{m}}_i(t)$ and $\widehat{\boldsymbol{\mu}}_c(t)$ in Scenarios (I)–(III) for the four different methods "Linear", "B-spline", "Local PQL" and "Local EM". Numbers in parentheses are the corresponding standard errors.

| Method | MSE for $\widehat{\boldsymbol{\beta}}$ | MISE for $\widehat{\mathbf{m}}_i(t)$ | MISE for $\widehat{\boldsymbol{\mu}}_c(t)$ |
|---|---|---|---|
| **Scenario (I)** | | | |
| Linear | $2.355 \times 10^{-2}$ ($0.060 \times 10^{-2}$) | $4.780 \times 10^{-3}$ ($0.007 \times 10^{-3}$) | $3.676 \times 10^{-6}$ ($0.067 \times 10^{-6}$) |
| B-spline | $2.355 \times 10^{-2}$ ($0.060 \times 10^{-2}$) | $4.780 \times 10^{-3}$ ($0.007 \times 10^{-3}$) | $3.676 \times 10^{-6}$ ($0.067 \times 10^{-6}$) |
| Local PQL | $2.362 \times 10^{-2}$ ($0.060 \times 10^{-2}$) | $4.908 \times 10^{-3}$ ($0.007 \times 10^{-3}$) | $5.260 \times 10^{-6}$ ($0.087 \times 10^{-6}$) |
| Local EM | $2.354 \times 10^{-2}$ ($0.060 \times 10^{-2}$) | $4.897 \times 10^{-3}$ ($0.007 \times 10^{-3}$) | $5.260 \times 10^{-6}$ ($0.087 \times 10^{-6}$) |
| **Scenario (II)** | | | |
| Linear | $53.053 \times 10^{-3}$ ($0.707 \times 10^{-3}$) | $15.620 \times 10^{0}$ ($0.010 \times 10^{0}$) | $64.357 \times 10^{-3}$ ($0.589 \times 10^{-3}$) |
| B-spline | $2.776 \times 10^{-3}$ ($0.079 \times 10^{-3}$) | $1.009 \times 10^{0}$ ($0.001 \times 10^{0}$) | $1.873 \times 10^{-3}$ ($0.024 \times 10^{-3}$) |
| Local PQL | $7.875 \times 10^{-3}$ ($0.200 \times 10^{-3}$) | $1.041 \times 10^{0}$ ($0.001 \times 10^{0}$) | $1.066 \times 10^{-3}$ ($0.010 \times 10^{-3}$) |
| Local EM | $2.439 \times 10^{-3}$ ($0.067 \times 10^{-3}$) | $0.996 \times 10^{0}$ ($0.001 \times 10^{0}$) | $1.068 \times 10^{-3}$ ($0.010 \times 10^{-3}$) |
| **Scenario (III)** | | | |
| Linear | $28.947 \times 10^{-2}$ ($0.716 \times 10^{-2}$) | $34.945 \times 10^{-1}$ ($0.017 \times 10^{-1}$) | $51.323 \times 10^{-4}$ ($0.499 \times 10^{-4}$) |
| B-spline | $2.124 \times 10^{-2}$ ($0.056 \times 10^{-2}$) | $3.680 \times 10^{-1}$ ($0.002 \times 10^{-1}$) | $7.551 \times 10^{-4}$ ($0.096 \times 10^{-4}$) |
| Local PQL | $3.410 \times 10^{-2}$ ($0.086 \times 10^{-2}$) | $3.819 \times 10^{-1}$ ($0.002 \times 10^{-1}$) | $2.977 \times 10^{-4}$ ($0.023 \times 10^{-4}$) |
| Local EM | $2.057 \times 10^{-2}$ ($0.052 \times 10^{-2}$) | $3.628 \times 10^{-1}$ ($0.002 \times 10^{-1}$) | $3.048 \times 10^{-4}$ ($0.024 \times 10^{-4}$) |

**TABLE 3** MSE and relative efficiency of the penalized versus unpenalized estimates of $\boldsymbol{\beta}$. Numbers in parentheses are the corresponding standard errors.

| $p$ | MSE without LASSO | MSE with LASSO | Relative Efficiency |
|---|---|---|---|
| 10 | 0.0226 (0.0005) | 0.0182 (0.0005) | 1.2424 (0.0065) |
| 15 | 0.0361 (0.0007) | 0.0215 (0.0005) | 1.6777 (0.0122) |
| 20 | 0.0497 (0.0009) | 0.0278 (0.0006) | 1.7898 (0.0105) |

**FIGURE 3** Fitted longitudinal trajectories of all patients (1st column) by Linear, B-spline, Local PQL, and Local EM, along with the fitted longitudinal trajectories of 8 randomly selected patients (columns 2-9). The dark dots in the plots on columns 2-9 are the observed blood pressure readings.