Statistics and Its Interface Volume 11 (2018) 1–18

Estimation and variable selection in generalized partially nonlinear models with nonignorable missing responses

Niansheng Tang* and Lin Tang

Based on the local kernel estimation method and propensity score adjustment method, we develop a penalized likelihood approach to simultaneously select covariates and explanatory variables in the considered parametric respondent model, and estimate parameters and nonparametric functions in generalized partially nonlinear models with nonignorable missing responses. An EM algorithm is proposed to evaluate the penalized likelihood estimations of parameters. The IC_Q criterion is employed to select the optimal penalty parameter. Under some regularity conditions, we show some asymptotic properties of parameter estimators such as oracle property. It can be shown that the proposed local linear kernel estimator of the nonparametric component is an estimator of a least favorable curve. The consistency of the IC_Q-based selection procedure is obtained. Simulation studies are conducted, and a real data set is used to illustrate the proposed methodologies.

Keywords and phrases: Generalized partially nonlinear models, Local kernel estimation, Nonignorable missing responses, Propensity score, Variable selection.

1. INTRODUCTION

Generalized nonlinear model (GNM) is a natural extension of generalized linear models and nonlinear regression models, GNM provides an effective tool for modeling nonnormal data such as count data and nonlinear relationship between mean of response variable and its associated factors. Over the past three decades, many studies have been done on GNMs. For example, Jorgensen [1] discussed asymptotic properties of maximum likelihood estimators of parameters in GNMs. Cordeiro and Paula [2] gave a general Bartlett adjustment formula for the expected likelihood ratio statistics in GNMs. Cox and Ma [3] developed asymptotic confidence bands for a linear combination of parameters in GNMs. Lindsey et al. [4] used GNMs to fit pharmacokinetic data. Kosmidis and Firth [5] presented a more general family of bias-reducing adjusted scores for a class of GNMs. Recently, Turner and Firth [6] developed an R package to make statistical inference on GNMs. However, it is recognized that incorporating a nonparametric function within parametric regression models is important for accommodating a possible inhomogeneity with respect to some covariates of interest and addressing the curse of dimensionality (e.g., see [7, 8]). Therefore, this paper considers a new model that is referred to as a generalized partially nonlinear model (GPNM) by introducing a nonparametric function into a GNM.

GPNMs retain the flexibility of nonparametric models and the ease of interpretation of parametric models, and include a lot of semiparametric regression models such as partially linear models [7, 9], generalized partially linear models [10] and partially nonlinear models [8, 11, 12].

Missing data commonly occurs in many fields such as psychological, educational, economical and biomedical studies [13]. The potential reasons for missing data may include: study drop out, subjects' refusal to answer items on a questionnaire, or failing to attend a scheduled clinic visit. To this end, many methods have been developed to analyze semiparametric regression models with missing data. For example, see [14, 15, 16, 17]. Their works are mainly focused on missing at random (MAR) assumption of missing responses/covariates. However, in many applications, missing data is nonignorable in the sense that the reason for missingness often depends on the missing values themselves [13, 18]. Hence, this paper aims to develop an approach to estimate parameters and unknown functions, and select important explanatory factors for predicting responses in GPNMs with nonignorable missing responses.

Variable selection is an important step in data analysis. Many methods have been proposed to address variable selection issue for parametric, nonparametric and semiparametric models. Traditional variable selection methods include: the stepwise regression and best subset selection associated with the Akaike information criterion (AIC) [19], Bayesian information criterion (BIC) [20] and Deviance information criterion (DIC) [21]. They often suffer from instability and computationally intensive burden [22] when the number of covariates is large. To address the issue, various penalization-based methods have been developed to simultaneously estimate parameters and select important covariates over the past years. For example, see the least absolute shrinkage and selection operator (LASSO) [23], smoothly clipped absolute deviation (SCAD) [24], adaptive LASSO

*Corresponding author.

(ALASSO) [25], least squares approximation [26], and the folded concave penalty method [27]. These methods have received a lot of attention in recent years. For example, see [28] for semiparametric models; [29] for partially linear measurement error models; [30] and [31] for semiparametric varying-coefficient models; [32] for semiparametric mixed models; [33, 34] for regression models with missing data and Cox regression models with covariates MAR; [35] for partially linear single-index models with longitudinal data. However, to the best of our knowledge, there is not work done on automatically and simultaneously selecting variables in GPNMs with nonignorable missing responses.

Motivated by [24] and [25], we here develop an approach to simultaneously estimate parameters and nonparametric functions, and select covariates in a GPNM as well as a respondent model. Our proposed method incorporates the idea of the least-favorable curve [36, 37], local kernel estimation method [38], and propensity score adjustment method for nonignorable nonresponse [39]. The IC_Q criterion [40] is adopted to select the optimal penalty parameter. We also study asymptotic properties of parameter estimators and nonparametric function estimators, and the consistency of the IC_Q-based selection procedure under some regularity conditions. The proposed method has the following merits. First, it allows us to simultaneously maximize the penalized likelihood function and estimate the penalty parameters using the local linear approximation algorithm. Second, compared with the profile approach, the linear approximation approach is computationally less intensive.

The rest of this paper is organized as follows. In Section 2, we propose an estimation procedure for parameter and nonparametric function in GPNMs with nonignorable missing responses. Asymptotic properties of the resulting estimators are studied in Section 2. Section 3 develops an EM algorithm to implement the maximum penalized likelihood (MPL) estimation and select penalty parameters via the IC_Q criterion [40]. Also, asymptotic properties of the resulting MPL estimators are investigated in Section 3. Simulation studies are used to evaluate the finite sample performance of the proposed estimators, and an example is illustrated in Section 4. Some discussions are given in Section 5. Technical details are presented in the Appendix.

2. MODEL AND ESTIMATION METHOD

2.1 Model and notation

Consider a data set $\{(y_i, \boldsymbol{x}_i, t_i) : i = 1, \ldots, n\}$ with observations measured on n independent subjects, where y_i is response variable, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$ $(p < n)$ is a $p \times 1$ vector of covariates, and t_i is the time measured for the ith subject. Let $\boldsymbol{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$, $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^{\mathrm{T}}$ and $\mathcal{T} = (t_1, \ldots, t_n)^{\mathrm{T}}$. It is assumed that given $\boldsymbol{x}_i$ and t_i, y_i follows a one-parameter exponential family, whose probability density function is

$$p(y_i|\boldsymbol{x}_i, t_i; \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)\right\}, \tag{1}$$

where θ_i is the canonical parameter, $b(\cdot)$ and $c(\cdot, \cdot)$ are known continuously differentiable functions, and ϕ is a scale parameter which is known or to be estimated. For simplicity, it is assumed that ϕ is known throughout this paper. Model (1) includes normal distribution, Poisson distribution and Gamma distribution as its special cases. Following McCullagh and Nelder [41], we assume that the systematic part of the model satisfies

$$\eta_i = G(\mu_i) = f(\boldsymbol{x}_i, \boldsymbol{\beta}) + g(t_i), \ i = 1, \ldots, n, \tag{2}$$

where μ_i is the conditional mean of y_i (i.e., $\mu_i = E(y_i|\boldsymbol{x}_i, t_i)$), $G(\cdot)$ is a known strictly monotone differentiable link function, $f(\boldsymbol{x}_i, \boldsymbol{\beta})$ is a known continuously differentiable nonlinear function in $\boldsymbol{\beta}$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is a set of unknown parameters of interest defined in a compact set $\mathcal{B} \subset \mathcal{R}^p$ and associated with covariates $\boldsymbol{x}_i$, and $g(\cdot)$ is a twice continuously differentiable smooth function on some finite interval, for example, $[0, 1]$. The model defined in Equations (1) and (2) is referred to as a GPNM.

Suppose that $\boldsymbol{x}_i$'s and t_i's are fully observed, while y_i's are subject to missingness. Let $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)^{\mathrm{T}}$, where $\delta_i = 1$ if y_i is observed and $\delta_i = 0$ if y_i is missing. It is assumed that δ_i and δ_j are independent for any $i \neq j$, and δ_i depends on y_i, $\boldsymbol{z}_i$ and t_i such that $\pi_i = \pi(y_i, \boldsymbol{z}_i, t_i) \stackrel{\triangle}{=} \Pr(\delta_i = 1|y_i, \boldsymbol{z}_i, t_i)$. Here, $\boldsymbol{z}_i$ is a subset of $\boldsymbol{x}_i$, i.e., $\boldsymbol{x}_i = (\boldsymbol{z}_i^{\mathrm{T}}, \boldsymbol{v}_i^{\mathrm{T}})^{\mathrm{T}}$ in which $\boldsymbol{v}_i$ is regarded as a vector of instrument variables. Thus, the missing mechanism defined above is nonignorable [13], and the respondent model is identifiable. Following [18], we consider the following missingness data mechanism

$$\begin{aligned} p(\boldsymbol{\delta}|\boldsymbol{y}, \boldsymbol{z}, \mathcal{T}; \boldsymbol{\varphi}) &= \textstyle\prod_{i=1}^n p(\delta_i|y_i, \boldsymbol{z}_i, t_i; \boldsymbol{\varphi}) \\ &= \textstyle\prod_{i=1}^n \pi_i^{\delta_i}(1 - \pi_i)^{1-\delta_i}, \end{aligned} \tag{3}$$

where $\boldsymbol{z} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n\}$. Generally, $\pi_i = \pi(y_i, \boldsymbol{z}_i, t_i; \boldsymbol{\varphi})$ can be specified by the following logistic regression model

$$\mathrm{logit}(\pi_i) = \varphi_0 + \boldsymbol{\varphi}_z^{\mathrm{T}}\boldsymbol{z}_i + \varphi_t t_i + \varphi_y y_i, \tag{4}$$

where $\mathrm{logit}(a) = \log\{a/(1 - a)\}$, and $\boldsymbol{\varphi} = (\varphi_0, \boldsymbol{\varphi}_z^{\mathrm{T}}, \varphi_t, \varphi_y)^{\mathrm{T}}$ is a $m \times 1$ $(m < p + 3)$ vector of unknown parameters.

For notational simplicity, let $\boldsymbol{y}_o$ be a vector of the observed response variables, $\boldsymbol{y}_m$ be a vector of missing components of $\boldsymbol{y}$ (i.e., $\boldsymbol{y} = \{\boldsymbol{y}_o, \boldsymbol{y}_m\}$), $\mathbf{D}_c = \{\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\delta}, \mathcal{T}\}$ be the complete data set, $\mathbf{D}_o = \{\boldsymbol{y}_o, \boldsymbol{x}, \boldsymbol{\delta}, \mathcal{T}\}$ be the observed data set, and $\boldsymbol{\gamma} = \{\boldsymbol{\beta}, \boldsymbol{\varphi}\}$ be the unknown parameter set of interest. Then, the complete data likelihood for $\mathbf{D}_c$ is given by

$$\begin{aligned} p(\mathbf{D}_c|\boldsymbol{\gamma}) &= \prod_{i=1}^n p(y_i, \delta_i|\boldsymbol{x}_i, t_i; \boldsymbol{\gamma}) \\ &= \prod_{i=1}^n p(y_i|\boldsymbol{x}_i, t_i; \boldsymbol{\beta}, \phi) p(\delta_i|y_i, \boldsymbol{z}_i, t_i; \boldsymbol{\varphi}). \end{aligned}$$

2.2 Estimations of parameters and nonparametric functions

Denote $\mathcal{L}(\boldsymbol{\beta}, g_\beta(t)) = \sum_{i=1}^n \mathcal{L}_i(\boldsymbol{\beta}, g_\beta(t_i)) = \sum_{i=1}^n \{(y_i\theta_i - b(\theta_i))/\phi + c(y_i, \phi)\}$ in which $\eta_i = \mathrm{G}(\mu_i) = f(\boldsymbol{x}_i, \boldsymbol{\beta}) + g_\beta(t_i)$, where $g_\beta(t) = g(\boldsymbol{\beta}, t)$. Let $\mathcal{L}_{i,\beta}(\boldsymbol{\beta}, g_\beta(t_i)) = \partial\mathcal{L}_i(\boldsymbol{\beta}, g_\beta(t_i))/\partial\beta$ be the first-order partial derivative of the log-likelihood function for the ith subject with respect to $\boldsymbol{\beta}$. Let $\mathcal{L}_{i,g_\beta}(\boldsymbol{\beta}, g_\beta(t_i)) = \partial\mathcal{L}_i(\boldsymbol{\beta}, g_\beta(t_i))/\partial g_\beta$ be the first-order partial derivative of the log-likelihood function for the ith subject with respect to g_β. Their corresponding second order partial derivatives are denoted by $\mathcal{L}_{i,\beta\beta}(\cdot)$, $\mathcal{L}_{i,\beta g_\beta}(\cdot)$ and $\mathcal{L}_{i,g_\beta g_\beta}(\cdot)$, respectively.

It is assumed that the nonparametric component $g(\cdot)$ is an infinite-dimensional nuisance parameter and $t_i \in [0,1]$. Motivated by Severini and Wong [36] and Murphy and van der Vaart [37], we define a curve $\boldsymbol{\beta} \to g(\boldsymbol{\beta}, t)$, which satisfies $g(\boldsymbol{\beta}^*, t) = g^*(t)$, where $\boldsymbol{\beta}^*$ and $g^*(t)$ are the true values of $\boldsymbol{\beta}$ and $g(t)$, respectively. Let $\mathrm{SM}(\boldsymbol{\beta}) = \{g(\boldsymbol{\beta}, t) : g(\cdot,\cdot)$ is twice smooth continuous on $\mathcal{B} \times [0,1]\}$ be a submodel. Clearly, $g_\beta(t) = g(\boldsymbol{\beta}, t) \in \mathrm{SM}(\boldsymbol{\beta})$. Similar to Fan et al. [38] and using the propensity score adjusted (PSA) method of Riddles [39], for any fixed $\boldsymbol{\beta}$, the local kernel estimator $\hat{g}_\beta = \hat{g}_\beta(t)$ of $g_\beta(t)$ and its first derivative $\hat{g}_\beta^{(1)} = \partial\hat{g}_\beta(t)/\partial t$ can be obtained by solving the following equation

$$0 = \frac{1}{n}\sum_{i=1}^n \frac{\delta_i}{\pi_i} K_h(t_i - t)\mathcal{L}_{i,g_\beta}(\boldsymbol{\beta}, \hat{g}_\beta(t_i))(1, \frac{t_i - t}{h})^{\mathrm{T}}, \tag{5}$$

where $\hat{g}_\beta(t_i) = \hat{g}_\beta + \hat{g}_\beta^{(1)}(t_i - t)$, $K_h(\cdot) = K(\cdot/h)/h$ in which $K(\cdot)$ is a kernel function, h is a bandwidth. For the above specified missingness data mechanism, π_i is usually unknown and can be estimated via some proper method. Let $\hat{\boldsymbol{\varphi}}$ be a consistent estimator of $\boldsymbol{\varphi}$. Then, replacing π_i for $\hat{\pi}_i = \pi_i(\hat{\boldsymbol{\varphi}})$ in Equation (5), we can obtain $\hat{g}_\beta$ and $\hat{g}_\beta^{(1)}$ by solving the following equation

$$0 = \frac{1}{n}\sum_{i=1}^n \frac{\delta_i}{\hat{\pi}_i} K_h(t_i - t)\mathcal{L}_{i,g_\beta}(\boldsymbol{\beta}, \hat{g}_\beta(t_i))(1, \frac{t_i - t}{h})^{\mathrm{T}}. \tag{6}$$

Based on the local kernel estimator $\hat{g}_\beta$ of $g_\beta(t)$, the maximum likelihood estimator (MLE) $\hat{\boldsymbol{\gamma}}$ of $\boldsymbol{\gamma}$ can be obtained by maximizing the following log-likelihood function

$$L(\boldsymbol{\gamma}) = \sum_{i=1}^n \{\delta_i \log p(y_i, \delta_i|\boldsymbol{x}_i, t_i; \boldsymbol{\gamma}) + (1-\delta_i)\log\textstyle\int p(y_i, \delta_i|\boldsymbol{x}_i, t_i; \boldsymbol{\gamma})dy_i\}.$$

Generally, it is rather difficult to maximize the above objective function with respect to $\boldsymbol{\gamma}$ due to an intractable integral involved. To address the issue, we can adopt the expectation-maximization (EM) algorithm [42] to evaluate the MLE of $\boldsymbol{\gamma}$. Following [42], the EM algorithm is composed of two steps: one is the expectation step (E-step) and the other is the maximization step (M-step). Given the value $\boldsymbol{\gamma}^{(s)}$ of $\boldsymbol{\gamma}$ at the s-th iteration, the E-step is to evaluate the following Q-function:

$$Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(s)}) = Q_1(\boldsymbol{\beta}|\boldsymbol{\gamma}^{(s)}) + Q_2(\boldsymbol{\varphi}|\boldsymbol{\gamma}^{(s)}), \tag{7}$$

where

$$Q_1(\boldsymbol{\beta}|\boldsymbol{\gamma}^{(s)}) = \sum_{i=1}^n [\delta_i \log p(y_i|\boldsymbol{x}_i, t_i; \boldsymbol{\beta}) + (1-\delta_i)E_{y_i|\cdot}\{\log p(y_i|\boldsymbol{x}_i, t_i; \boldsymbol{\beta})\}]$$

in which $E_{y_i|\cdot}$ represents the conditional expectation taken with respect to the posterior probability density function $p(y_i|\boldsymbol{x}_i, t_i, \delta_i; \boldsymbol{\gamma}^{(s)})$ of y_i given $(\boldsymbol{x}_i, t_i, \delta_i, \boldsymbol{\gamma}^{(s)})$, and

$$Q_2(\boldsymbol{\varphi}|\boldsymbol{\gamma}^{(s)}) = \sum_{i=1}^n [\delta_i \log p(\delta_i|y_i, \boldsymbol{z}_i, t_i; \boldsymbol{\varphi}) + (1-\delta_i)E_{y_i|\cdot}\{\log p(\delta_i|y_i, \boldsymbol{z}_i, t_i; \boldsymbol{\varphi})\}].$$

Clearly, it is not easy to evaluate the Q-function due to the intractable integrals involved. Following Ibrahim, Chen and Lipsitz [43], we can approximate the Q-function via the Markov chain Monte Carlo algorithm. That is, when y_i is missing, we first generate $\mathcal{M}$ observations $\{y_i^{(s,l)} : l = 1, \ldots, \mathcal{M}\}$ from $p(y_i|\boldsymbol{x}_i, t_i, \delta_i; \boldsymbol{\gamma}^{(s)})$ via the Metropolis-Hastings (MH) algorithm for $i = 1, \ldots, n$, and then $Q_1(\boldsymbol{\beta}|\boldsymbol{\gamma}^{(s)})$ and $Q_2(\boldsymbol{\varphi}|\boldsymbol{\gamma}^{(s)})$ can be approximated by

$$Q_1(\boldsymbol{\beta}|\boldsymbol{\gamma}^{(s)}) \approx \frac{1}{\mathcal{M}}\sum_{l=1}^{\mathcal{M}}\sum_{i=1}^n \left\{\frac{y_i^{(s,l)}\theta_i - b(\theta_i)}{\phi} + c(y_i^{(s,l)}, \phi)\right\},$$

$$Q_2(\boldsymbol{\varphi}|\boldsymbol{\gamma}^{(s)}) \approx \frac{1}{\mathcal{M}}\sum_{l=1}^{\mathcal{M}}\sum_{i=1}^n \{\delta_i \boldsymbol{\varphi}_{\omega i}^{(s,l)} - \log(1 + \exp(\boldsymbol{\varphi}_{\omega i}^{(s,l)}))\},$$

respectively, where $\boldsymbol{\varphi}_{\omega i}^{(s,l)} = \boldsymbol{\varphi}^{\mathrm{T}}\boldsymbol{\omega}_i^{(s,l)}$ in which $\boldsymbol{\omega}_i^{(s,l)} = (1, \boldsymbol{z}_i^{\mathrm{T}}, t_i, y_i^{(s,l)})^{\mathrm{T}}$, and $y_i^{(s,l)} = y_i$ when y_i is observed. The details for implementing MH algorithm are given in Appendix. The M-step involves maximizing $Q_1(\boldsymbol{\beta}|\boldsymbol{\gamma}^{(s)})$ and $Q_2(\boldsymbol{\varphi}|\boldsymbol{\gamma}^{(s)})$ with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\varphi}$, respectively. There are not analytic solutions to the following equations: $\partial Q_1(\boldsymbol{\beta}|\boldsymbol{\gamma}^{(s)})/\partial\boldsymbol{\beta} = 0$ and $\partial Q_2(\boldsymbol{\varphi}|\boldsymbol{\gamma}^{(s)})/\partial\boldsymbol{\varphi} = 0$. To this end, the Fisher's scoring algorithm can be employed to obtain their solutions.

The above introduced EM algorithm can be implemented by the following steps.

Step 0. Select an initial value $\hat{\boldsymbol{\gamma}}^{(0)} = (\hat{\boldsymbol{\beta}}^{(0)\mathrm{T}}, \hat{\boldsymbol{\varphi}}^{(0)\mathrm{T}})^{\mathrm{T}}$ of $\boldsymbol{\gamma}$, where $\hat{\boldsymbol{\gamma}}^{(0)}$ is taken to be an estimate of $\boldsymbol{\gamma}$ obtained from the completely observed data set. And set $s = 0$.

Step 1. Given $\hat{\boldsymbol{\gamma}}^{(s)}$ and t, $\hat{g}_{\hat{\beta}^{(s)}}$ and $\hat{g}_{\hat{\beta}^{(s)}}^{(1)}$ are evaluated by solving the following equation:

$$0 = \frac{1}{n}\sum_{i=1}^n \frac{\delta_i}{\hat{\pi}_i^{(s)}} K_h(t_i - t)\mathcal{L}_{i,g_{\hat{\beta}^{(s)}}}(\hat{\boldsymbol{\beta}}^{(s)}, \hat{g}_{\hat{\beta}^{(s)}}(t_i))(1, \frac{t_i - t}{h})^{\mathrm{T}},$$

where $\hat{g}_{\hat{\beta}^{(s)}}(t_i) = \hat{g}_{\hat{\beta}^{(s)}} + \hat{g}^{(1)}_{\hat{\beta}^{(s)}}(t_i - t)$ and $\hat{\pi}_i^{(s)} = \pi_i(\hat{\boldsymbol{\varphi}}^{(s)})$.

Step 2. Given $\hat{g}_{\hat{\beta}^{(s)}} = g_{\hat{\beta}^{(s)}}(t)$, $\hat{\boldsymbol{\beta}}^{(s+1)}$ and $\hat{\boldsymbol{\varphi}}^{(s+1)}$ are computed by solving the following equations:

$$\dot{Q}_1(\boldsymbol{\beta}) = \frac{1}{\mathcal{M}} \sum_{l=1}^{\mathcal{M}} \sum_{i=1}^{n} (y_i^{(s,l)} - \mu_i)\{\dot{G}(\mu_i)\ddot{b}(\theta_i)\}^{-1} \dot{f}_i(\boldsymbol{\beta}) = 0,$$

$$\dot{Q}_2(\boldsymbol{\varphi}) = \frac{1}{\mathcal{M}} \sum_{l=1}^{\mathcal{M}} \sum_{i=1}^{n} \left\{ \delta_i - \frac{\exp(\boldsymbol{\varphi}^{\mathrm{T}} \boldsymbol{\omega}_i^{(s,l)})}{1 + \exp(\boldsymbol{\varphi}^{\mathrm{T}} \boldsymbol{\omega}_i^{(s,l)})} \right\} \boldsymbol{\omega}_i^{(s,l)} = 0,$$

respectively, via the Fisher's scoring algorithm, where $\mu_i = \dot{b}(\theta_i) = db(\theta_i)/d\theta_i$, $\dot{G}(\mu_i) = dG(\mu_i)/d\mu_i$, $\ddot{b}(\theta_i) = d^2b(\theta_i)/d\theta_i^2$, $\dot{f}_i(\boldsymbol{\beta}) = \partial f(\boldsymbol{x}_i, \boldsymbol{\beta})/\partial\beta$, and $\eta_i = G(\mu_i) = f(\boldsymbol{x}_i, \boldsymbol{\beta}) + g_{\hat{\beta}^{(s)}}(t_i)$.

Step 3. Repeat steps 1 and 2 until the convergence of the EM algorithm. The algorithm is monitored by the following stopping rule: if $\max_{j\in\{1,\ldots,p+m\}} |\gamma_j^{(s+1)} - \gamma_j^{(s)}| \le c_0$, we claim the convergence of the EM algorithm; otherwise, we repeat steps 1 and 2, where γ_j is the jth component of $\boldsymbol{\gamma}$ and c_0 is some user-given sufficiently small constant.

2.3 Asymptotic properties

In this subsection, we investigate the consistency of the local kernel estimators of $g_\beta(t)$ and its derivative as well as asymptotic properties of the MLE of $\boldsymbol{\gamma}$. To this end, we first consider the semiparametric efficiency and least-favorable curve when there is not missing data. Following Severini and Wong [36], any curve $g_\beta = g(\beta, t) \in \mathrm{SM}(\boldsymbol{\beta})$ is said to be a least favorable curve if

$$\begin{aligned} &\mathrm{E}\left\{ \tfrac{\partial}{\partial\beta}\mathcal{L}(\beta, g_\beta) \tfrac{\partial}{\partial\beta^{\mathrm{T}}}\mathcal{L}(\beta, g_\beta) \right\}_{\beta=\beta^*} \\ &\le \mathrm{E}\left\{ \tfrac{\partial}{\partial\beta}\mathcal{L}(\beta, g_{1\beta}) \tfrac{\partial}{\partial\beta^{\mathrm{T}}}\mathcal{L}(\beta, g_{1\beta}) \right\}_{\beta=\beta^*} \end{aligned} \tag{8}$$

holds for any other smooth curve $g_{1\beta} \in \mathrm{SM}(\boldsymbol{\beta})$ with $g_{1\beta^*} = g_{\beta^*}$, where $\boldsymbol{\beta}^*$ is the true value of $\boldsymbol{\beta}$. The left term in Equation (8) is referred to as the semiparametric information bound. Under Assumption A given in the Appendix, we have

Lemma 2.1. *A function $g(\beta, t) \in \mathrm{SM}(\boldsymbol{\beta})$ is a least favorable curve if and only if*

$$\frac{\partial g_{\beta^*}}{\partial\beta} = -\frac{E_t[\mathcal{L}_{\beta g_\beta}(\boldsymbol{\beta}^*, g_{\beta^*})]}{E_t[\mathcal{L}_{g_\beta g_\beta}(\boldsymbol{\beta}^*, g_{\beta^*})]},$$

where $E_t[\cdot] = E[\cdot|\mathcal{T} = t]$, $\mathcal{L}_{\beta g_\beta}(\boldsymbol{\beta}, g_\beta) = \sum_{i=1}^n \mathcal{L}_{i,\beta g_\beta}(\boldsymbol{\beta}, g_\beta)$ and $\mathcal{L}_{g_\beta g_\beta}(\boldsymbol{\beta}, g_\beta) = \sum_{i=1}^n \mathcal{L}_{i,g_\beta g_\beta}(\boldsymbol{\beta}, g_\beta)$.

Theorem 2.1. *Under Assumption A given in the Appendix, the estimators $\hat{g}_\beta(t)$ and $\hat{g}^{(1)}_\beta(t)$ obtained by solving equation (6) satisfy*

(i) $\hat{g}_\beta(t) \xrightarrow{a.s.} g(\beta, t)$, $\hat{g}^{(1)}_\beta(t) \xrightarrow{a.s.} \frac{\partial}{\partial t} g(\beta, t)$;

(ii) $\frac{\partial \hat{g}_\beta(t)}{\partial\beta} \xrightarrow{a.s.} \frac{\partial g(\beta,t)}{\partial\beta}$, $\frac{\partial^2 \hat{g}_\beta(t)}{\partial\beta\partial\beta^{\mathrm{T}}} \xrightarrow{a.s.} \frac{\partial^2 g(\beta,t)}{\partial\beta\partial\beta^{\mathrm{T}}}$.

Corollary 2.1. *Suppose that the conditions given in Theorem 2.1 hold. Thus, the proposed estimator $\hat{g}_\beta(t)$ of g_β is an estimator of the least favorable curve when there is not missing data.*

Theorem 2.2. *Under Assumption A given in the Appendix, the asymptotic expansion of $\hat{g}_\beta(t)$ is given by*

$$\begin{aligned} \hat{g}_\beta(t) - g_\beta(t) = {} & \tfrac{h^2}{2}\kappa_2(K) g^{(2)}(\boldsymbol{\beta}, t) \\ & - \tfrac{1}{nf_{\mathbb{T}}(t)\psi(t)} \sum_{i=1}^{n} \tfrac{\delta_i}{\pi_i} K_h(t_i - t) \mathcal{L}_{i,g_\beta}(\boldsymbol{\beta}, g_\beta(t_i)) \\ & + o_p\{h^2 + (nh)^{-1/2}\}, \end{aligned}$$

which leads to

$$\begin{aligned} &(nh)^{1/2}\{\hat{g}_\beta(t) - g_\beta(t) - \tfrac{h^2}{2}\kappa_2(K) g^{(2)}(\boldsymbol{\beta}, t)\} \\ &\xrightarrow{D} N(0, \mu_0 f_{\mathbb{T}}^{-1}(t)\psi^{-1}(t)), \end{aligned}$$

where $\kappa_2(K) = \int t^2 K(t)dt$, $g^{(2)}(\boldsymbol{\beta}, t) = \partial^2 g_\beta(t)/\partial t^2$, $\mu_0(t) = \int K^2(t)dt$, $\psi(t)$ and $f_{\mathbb{T}}(t)$ are defined in the Appendix.

From Theorem 2.2, we can define the bias of local kernel estimator of nonparametric function: $\mathrm{bias}(\hat{g}_\beta) = E\{\hat{g}(\boldsymbol{\beta}, t) - g(\boldsymbol{\beta}, t)\} = \frac{h^2}{2}\kappa_2(K) g^{(2)}(\boldsymbol{\beta}, t) + o_p\{h^2\}$, which shows that the bias of local kernel estimator depends on the bandwidth and kernel function. Particularly, if we replace Assumption A(5) by $nh^2/\log n \to \infty$ and $nh^5 \to 0$, that is, if undersmoothing is used, thus the bias term $h^2\kappa_2(K)g^{(2)}(\beta, t)/2$ vanishes asymptotically [44] and $(nh)^{1/2}\{\hat{g}_\beta(t) - g_\beta(t)\} \xrightarrow{D} N(0, \mu_0 f_{\mathbb{T}}^{-1}(t)\psi^{-1}(t))$.

The bandwidth h should be appropriately selected to obtain an efficient estimator $\hat{g}_\beta$ of $g_\beta(t)$. The commonly used data-driven methods include cross-validation (CV) and generalized cross-validation (GCV). However, these methods are not easily implemented in the presence of missing data. From Theorem 2.2, it is easily seen that the optimal rate is $n^{-1/5}$, we here adopt a simple bandwidth $h = c\hat{\sigma}_{\mathcal{T}} n^{-1/5}$, where c is constant and $\hat{\sigma}_{\mathcal{T}}$ is the standard deviation of the fixed design time points in $\mathcal{T}$.

Theorem 2.3. *Under Assumptions A(1) and B(1)–B(4) given in the Appendix, we have*

$$n^{1/2}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*) \xrightarrow{D} N(\mathbf{0}, A(\boldsymbol{\gamma}^*)^{-1} B(\boldsymbol{\gamma}^*) A(\boldsymbol{\gamma}^*)^{-1}),$$

where $\boldsymbol{\gamma}^*$ is the true value of $\boldsymbol{\gamma}$, $A(\boldsymbol{\gamma}^*)$ and $B(\boldsymbol{\gamma}^*)$ are defined in Assumption B(4).

3. VARIABLE SELECTION

3.1 EM algorithm for maximizing the penalized likelihood

In this subsection, we simultaneously consider variable selection and parameter estimation problem based on some

proper penalized likelihood function. To this end, we consider the following penalized log-likelihood function

$$\begin{aligned}\mathrm{PL}(\boldsymbol{\gamma}|\boldsymbol{\lambda}) = &\sum_{i=1}^{n}\{\delta_i \log p(y_i,\delta_i|\boldsymbol{x}_i,t_i;\boldsymbol{\gamma})\\ &+(1-\delta_i)\log\textstyle\int p(y_i,\delta_i|\boldsymbol{x}_i,t_i;\boldsymbol{\gamma})dy_i\}\\ &-n\sum_{j=1}^{p} p_{\lambda_{\beta,j}}(|\beta_j|) - n\sum_{k=1}^{m} p_{\lambda_{\varphi,k}}(|\varphi_k|),\end{aligned}$$

where $\boldsymbol{\lambda} = (\lambda_{\beta,1},\ldots,\lambda_{\beta,p},\lambda_{\varphi,1},\ldots,\lambda_{\varphi,m})^{\mathrm{T}}$, $\lambda_{\beta,j}$ is the penalty parameter corresponding to the jth coefficient β_j in $\boldsymbol{\beta}$ for $j = 1,\ldots,p$, while $\lambda_{\varphi,k}$ represents the penalty parameter corresponding to the kth coefficient φ_k in $\boldsymbol{\varphi}$ for $k = 1,\ldots,m$, and $p_{\lambda_{\beta,j}}(\cdot)$ and $p_{\lambda_{\varphi,k}}(\cdot)$ are user-specified penalty functions, which are nonnegative, nondecreasing and differentiable on the interval $(0,\infty)$ [24, 25]. Generally, one can take the penalty function to be the LASSO penalty, SCAD penalty [24, 25] and MC penalty [45]. It is rather difficult to simultaneously select variables and estimate parameters based on the above penalized log-likelihood function $\mathrm{PL}(\boldsymbol{\gamma}|\boldsymbol{\lambda})$ due to an intractable integral involved. In this case, a Monte Carlo EM algorithm is employed to evaluate the maximum penalized likelihood estimation (MPLE) (denoted as $\hat{\boldsymbol{\gamma}}_\lambda$) of $\boldsymbol{\gamma}$. Following the idea of EM algorithm, given the value $\boldsymbol{\gamma}^{(s)}$ of $\boldsymbol{\gamma}$ at the sth iteration, the E-step is to evaluate the following penalized Q-function

$$\begin{aligned}&Q_\lambda(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(s)})\\ &= Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(s)}) - n\sum_{j=1}^{p} p_{\lambda_{\beta,j}}(|\beta_j|) - n\sum_{k=1}^{m} p_{\lambda_{\varphi,k}}(|\varphi_k|)\\ &= Q_1(\boldsymbol{\beta}|\boldsymbol{\gamma}^{(s)}) + Q_2(\boldsymbol{\varphi}|\boldsymbol{\gamma}^{(s)})\\ &\quad -n\sum_{j=1}^{p} p_{\lambda_{\beta,j}}(|\beta_j|) - n\sum_{k=1}^{m} p_{\lambda_{\varphi,k}}(|\varphi_k|)\\ &\triangleq Q_{1,\lambda}(\boldsymbol{\beta}|\boldsymbol{\gamma}^{(s)}) + Q_{2,\lambda}(\boldsymbol{\varphi}|\boldsymbol{\gamma}^{(s)}),\end{aligned} \tag{9}$$

where

$$Q_{1,\lambda}(\boldsymbol{\beta}|\boldsymbol{\gamma}^{(s)}) = Q_1(\boldsymbol{\beta}|\boldsymbol{\gamma}^{(s)}) - n\sum_{j=1}^{p} p_{\lambda_{\beta,j}}(|\beta_j|)$$

and

$$Q_{2,\lambda}(\boldsymbol{\varphi}|\boldsymbol{\gamma}^{(s)}) = Q_2(\boldsymbol{\varphi}|\boldsymbol{\gamma}^{(s)}) - n\sum_{k=1}^{m} p_{\lambda_{\varphi,k}}(|\varphi_k|)$$

in which $Q_1(\boldsymbol{\beta}|\boldsymbol{\gamma}^{(s)})$ and $Q_2(\boldsymbol{\varphi}|\boldsymbol{\gamma}^{(s)})$ are defined in Equation (7).

The M-step is to maximize $Q_\lambda(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(s)})$, which is a rather difficult task because $Q_\lambda(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(s)})$ is a non-differentiable and nonconcave function of $\boldsymbol{\gamma}$. Following Fan and Li [24], this issue can be addressed by maximizing the second-order Taylor expansions of $Q_{1,\lambda}(\boldsymbol{\beta}|\boldsymbol{\gamma}^{(s)})$ and $Q_{2,\lambda}(\boldsymbol{\varphi}|\boldsymbol{\gamma}^{(s)})$ at $\boldsymbol{\beta}^{(s)}$ and $\boldsymbol{\varphi}^{(s)}$, respectively. Thus, the problem of maximizing $Q_{1,\lambda}(\boldsymbol{\beta}|\boldsymbol{\gamma}^{(s)})$ and $Q_{2,\lambda}(\boldsymbol{\varphi}|\boldsymbol{\gamma}^{(s)})$ with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\varphi}$ reduces to an optimization problem of the penalized weighted least squares regression, which can be implemented via some appropriate optimization algorithm such as the local quadratic approximation algorithm [24], local linear approximation algorithm [46] and best convex minorization-maximization algorithm [47].

Let $\boldsymbol{\beta}^{(s+1)} = \mathrm{argmax}_\beta Q_{1,\lambda}(\boldsymbol{\beta},|\boldsymbol{\gamma}^{(s)})$ and $\boldsymbol{\varphi}^{(s+1)} = \mathrm{argmax}_\varphi Q_{2,\lambda}(\boldsymbol{\varphi}|\boldsymbol{\gamma}^{(s)})$. Note that $\boldsymbol{\gamma}^{(s+1)}$ is evaluated by maximizing the second-order Taylor expansions of $Q_{1,\lambda}(\boldsymbol{\beta}|\boldsymbol{\gamma}^{(s)})$ and $Q_{2,\lambda}(\boldsymbol{\varphi}|\boldsymbol{\gamma}^{(s)})$, respectively, but it is not the maximizer of $Q_\lambda(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(s)})$ with respect to $\boldsymbol{\gamma}$. To this end, a hybrid algorithm combining the local linear approximation algorithm of Zou and Li [46] and the expectation conditional maximization algorithm [48] is developed to find $\boldsymbol{\gamma}^{(s+1)}$ such that $Q_\lambda(\boldsymbol{\gamma}^{(s+1)}|\boldsymbol{\gamma}^{(s)}) > Q_\lambda(\boldsymbol{\gamma}^{(s)}|\boldsymbol{\gamma}^{(s)})$. Iterating the above procedure until the convergence of the hybrid algorithm yields the desirable maximum penalized likelihood estimation $\hat{\boldsymbol{\gamma}}_\lambda$ of $\boldsymbol{\gamma}$.

3.2 Penalty parameter selection

To ensure that the resultant estimator $\hat{\boldsymbol{\gamma}}_\lambda$ of $\boldsymbol{\gamma}$ has the well-known oracle property, it is necessary to appropriately select the penalty parameter $\boldsymbol{\lambda}$. The commonly used criterion for selecting the penalty parameter includes the generalized cross-validation (GCV) and Bayesian information criterion (BIC), which are the data-driven methods. These criteria are not easily implemented in the presence of missing data because the observed data likelihood function involves an intractable integral. In addition, the GCV method may lead to a significant overfitting even in linear models [14]. Hence, we here adopt the IC_Q criterion [40] to select the penalty parameter $\boldsymbol{\lambda}$. Following [40], the optimal penalty parameter $\boldsymbol{\gamma}$ can be obtained by minimizing

$$\mathrm{IC}_{\mathrm{Q}}(\boldsymbol{\lambda}) = -2Q(\hat{\boldsymbol{\gamma}}_\lambda|\hat{\boldsymbol{\gamma}}) + c_n(\hat{\boldsymbol{\gamma}}_\lambda), \tag{10}$$

where $\hat{\boldsymbol{\gamma}}$ is the MLE of $\boldsymbol{\gamma}$ introduced in Section 2, and $c_n(\boldsymbol{\gamma})$ is a function of the data and the fitted model. Different selection of $c_n(\boldsymbol{\gamma})$ leads to different criterion. For example, when $c_n(\boldsymbol{\gamma}) = 2d_n$, where d_n is the total number of unknown parameters, the above defined IC_Q criterion reduces to the AIC; when $c_n(\boldsymbol{\gamma}) = d_n\log(n)$, the IC_Q criterion becomes the BIC. For a given $\boldsymbol{\lambda}$, it is easy to implement $\mathrm{IC}_Q(\boldsymbol{\lambda})$ because the Q-function is a direct byproduct of the above introduced hybrid algorithm output.

3.3 Theoretical properties

In this subsection, we establish asymptotic properties of the penalized likelihood estimators and the consistency of the penalty parameter selection procedure based on the IC_Q criterion.

Let $\mathcal{S}_\beta = \{j : \beta_j \neq 0\}$ be the index set of nonzero components of the true value $\boldsymbol{\beta}^*$ of $\boldsymbol{\beta}$, and $\mathcal{S}_\varphi = \{k : \varphi_k \neq 0\}$ be the index set of nonzero components of the true value $\boldsymbol{\varphi}^*$ of $\boldsymbol{\varphi}$. Denote the cardinalities of $\mathcal{S}_\beta$ and $\mathcal{S}_\varphi$ as $p_1 = |\mathcal{S}_\beta|$ and

$q_1 = |\mathcal{S}_\varphi|$, respectively, which are usually unknown. Then, $\mathcal{S}_I = \mathcal{S}_\beta \bigcup \mathcal{S}_\varphi$ is the index set of the true model. Without loss of generality, we assume $\boldsymbol{\beta} = (\boldsymbol{\beta}_{(1)}^{\mathrm{T}}, \boldsymbol{\beta}_{(2)}^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{\beta}_{(1)}$ and $\boldsymbol{\beta}_{(2)}$ correspond to the nonzero and zero components of $\boldsymbol{\beta}$ with the dimensions being p_1 and $p_2 = p - p_1$, respectively, which indicates that $\boldsymbol{\beta}^*$ has the following decomposition $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_{(1)}^{*\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}$. The corresponding decomposition of $\hat{\boldsymbol{\beta}}_\lambda$ can be written as $\hat{\boldsymbol{\beta}}_\lambda = (\hat{\boldsymbol{\beta}}_{(1)\lambda}^{\mathrm{T}}, \hat{\boldsymbol{\beta}}_{(2)\lambda}^{\mathrm{T}})^{\mathrm{T}}$. Similarly, we assume that $\boldsymbol{\varphi}$ has the following decomposition: $\boldsymbol{\varphi} = (\boldsymbol{\varphi}_{(1)}^{\mathrm{T}}, \boldsymbol{\varphi}_{(2)}^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{\varphi}_{(1)}$ and $\boldsymbol{\varphi}_{(2)}$ correspond to the nonzero and zero components of $\boldsymbol{\varphi}$ with the dimensions being q_1 and $q_2 = m - q_1$, respectively, which shows that $\boldsymbol{\varphi}^*$ has the following form $\boldsymbol{\varphi}^* = (\boldsymbol{\varphi}_{(1)}^{*\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}$, and the corresponding decomposition of $\hat{\boldsymbol{\varphi}}_\lambda$ can be written as $\hat{\boldsymbol{\varphi}}_\lambda = (\hat{\boldsymbol{\varphi}}_{(1)\lambda}^{\mathrm{T}}, \hat{\boldsymbol{\varphi}}_{(2)\lambda}^{\mathrm{T}})^{\mathrm{T}}$. Let $\boldsymbol{\vartheta} = (\boldsymbol{\beta}_{(1)}^{\mathrm{T}}, \boldsymbol{\varphi}_{(1)}^{\mathrm{T}})^{\mathrm{T}}$, and its corresponding penalized likelihood estimator and true value are denoted as $\hat{\boldsymbol{\vartheta}}_\lambda = (\hat{\boldsymbol{\beta}}_{(1)\lambda}^{\mathrm{T}}, \hat{\boldsymbol{\varphi}}_{(1)\lambda}^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{\vartheta}^* = (\boldsymbol{\beta}_{(1)}^{\mathrm{T}*}, \boldsymbol{\varphi}_{(1)}^{\mathrm{T}*})^{\mathrm{T}}$, respectively.

Theorem 3.1. *Under Assumptions A(1) and B given in the Appendix, we have*

(i) (Consistency) $\hat{\boldsymbol{\gamma}}_\lambda - \boldsymbol{\gamma}^* = O_p(n^{-1/2})$ *as* $n \to \infty$*;*

(ii) (Sparsity) $\Pr(\hat{\boldsymbol{\beta}}_{(2)\lambda} = \mathbf{0}, \hat{\boldsymbol{\varphi}}_{(2)\lambda} = \mathbf{0}) \to 1$*;*

(iii) (Asymptotic normality) $n^{1/2}\{\hat{\boldsymbol{\vartheta}}_\lambda - \boldsymbol{\vartheta}^* + (\tilde{\mathbf{A}}(\boldsymbol{\vartheta}^*) + \mathbf{J}(\boldsymbol{\vartheta}^*))^{-1}\mathbf{h}(\boldsymbol{\vartheta}^*)\} \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\vartheta}^*))$*, where* $\tilde{\mathbf{A}}(\boldsymbol{\vartheta}^*)$*,* $\mathbf{J}(\boldsymbol{\vartheta}^*)$*,* $\mathbf{h}(\boldsymbol{\vartheta}^*)$ *and* $\boldsymbol{\Sigma}(\boldsymbol{\vartheta}^*)$ *are defined in the Appendix.*

Theorem 3.1 indicates that (i) $\hat{\boldsymbol{\gamma}}_\lambda$ is a root-n consistent estimator of $\boldsymbol{\gamma}$ if the penalty parameter vector $\boldsymbol{\lambda}$ is appropriately selected; (ii) $\hat{\boldsymbol{\gamma}}_\lambda$ possesses the sparsity property, i.e., $\hat{\boldsymbol{\beta}}_{(2)\lambda} = \mathbf{0}$ and $\hat{\boldsymbol{\varphi}}_{(2)\lambda} = \mathbf{0}$ with probability tending to 1 as $n \to \infty$; (iii) $(\hat{\boldsymbol{\beta}}_{(1)\lambda}^{\mathrm{T}}, \hat{\boldsymbol{\varphi}}_{(1)\lambda}^{\mathrm{T}})^{\mathrm{T}}$ is asymptotically distributed as the normal distribution.

To investigate whether the $\mathrm{IC}_Q(\boldsymbol{\lambda})$ criterion can consistently select the correct model, we define the candidate model as $\mathcal{S}_\lambda = \{j : \hat{\beta}_{\lambda j} \neq 0\} \bigcup \{k : \hat{\varphi}_{\lambda k} \neq 0\}$ based on the MPLE $\hat{\boldsymbol{\gamma}}_\lambda$ of $\boldsymbol{\gamma}$ for a given $\boldsymbol{\lambda} \in \mathcal{R}^{p+m}$. Thus, $\mathcal{S}_\lambda$ might be either an underfitted model or an overfitted model or a correctly specified model, which correspond to the following three disjoint regions: $\mathcal{R}_u = \{\boldsymbol{\lambda} \in \mathcal{R}^{p+m} : \mathcal{S}_\lambda \not\supset \mathcal{S}_I\}$, $\mathcal{R}_o = \{\boldsymbol{\lambda} \in \mathcal{R}^{p+m} : \mathcal{S}_\lambda \supset \mathcal{S}_I \text{ and } \mathcal{S}_\lambda \neq \mathcal{S}_I\}$ and $\mathcal{R}_c = \{\boldsymbol{\lambda} \in \mathcal{R}^{p+m} : \mathcal{S}_\lambda = \mathcal{S}_I\}$, respectively. We can always choose a reference penalty parameter sequence $\{\boldsymbol{\lambda}_n \in \mathcal{R}^{p+m}\}_{n=1}^{\infty}$ satisfying the conditions given in Theorem 3.1 so that $\mathcal{S}_{\lambda_n} = \mathcal{S}_I$ a.s. [33]. Following Ibrahim, Zhu and Tang [40], we can use $\mathrm{dIC}_{\mathrm{Q}}(\boldsymbol{\lambda}_2, \boldsymbol{\lambda}_1) = \mathrm{IC}_Q(\boldsymbol{\lambda}_2) - \mathrm{IC}_{\mathrm{Q}}(\boldsymbol{\lambda}_1) = 2Q(\hat{\boldsymbol{\gamma}}_{\lambda_1}|\hat{\boldsymbol{\gamma}}) - 2Q(\hat{\boldsymbol{\gamma}}_{\lambda_2}|\hat{\boldsymbol{\gamma}}) + c_n(\hat{\boldsymbol{\gamma}}_{\lambda_2}) - c_n(\hat{\boldsymbol{\gamma}}_{\lambda_1})$ to select the better model in terms of the following criterion: under the assumption $\mathcal{S}_{\lambda_2} \supset \mathcal{S}_{\lambda_1}$, if $\mathrm{dIC}_Q(\boldsymbol{\lambda}_2, \boldsymbol{\lambda}_1) > 0$, we select the penalty parameter $\boldsymbol{\lambda}_1$, otherwise $\boldsymbol{\lambda}_2$ is selected.

Define $\delta_Q(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) = E\{Q(\boldsymbol{\gamma}_{\mathcal{S}_{\lambda_1}}^*|\boldsymbol{\gamma}^*)\} - E\{Q(\boldsymbol{\gamma}_{\mathcal{S}_{\lambda_2}}^*|\boldsymbol{\gamma}^*)\}$ and $\delta_c(\boldsymbol{\lambda}_2, \boldsymbol{\lambda}_1) = c_n(\hat{\boldsymbol{\gamma}}_{\lambda_2}) - c_n(\hat{\boldsymbol{\gamma}}_{\lambda_1})$ in which $\boldsymbol{\gamma}_{\mathcal{S}}^* = \mathrm{argsup}_{\boldsymbol{\gamma}:\gamma_j \neq 0, j \in \mathcal{S}} E\{\mathrm{Q}(\boldsymbol{\gamma}|\boldsymbol{\gamma}^*)\}$.

Theorem 3.2. *Suppose that* $\mathcal{S}_{\lambda_1}$ *is a subset of* $\mathcal{S}_{\lambda_2}$*. Under Assumptions A(1) and B in the Appendix, we have the following results.*

(i) If for all $\mathcal{S}_\lambda \not\supset \mathcal{S}_I$*,* $\mathrm{liminf}_n \delta_Q(\boldsymbol{\lambda}, \mathbf{0})/n > 0$ *and* $\delta_c(\boldsymbol{\lambda}, \mathbf{0}) = o_p(n)$*, thus* $\mathrm{dIC}_Q(\boldsymbol{\lambda}, \mathbf{0}) > 0$ *in probability for all* $\mathcal{S}_\lambda \not\supset \mathcal{S}_I$*;*

(ii) If $E\{Q(\boldsymbol{\gamma}_{\mathcal{S}_{\lambda_1}}^*|\hat{\boldsymbol{\gamma}})\} - E\{Q(\boldsymbol{\gamma}_{\mathcal{S}_{\lambda_2}}^*|\hat{\boldsymbol{\gamma}})\} = O_p(n^{1/2})$ *and* $Q(\hat{\boldsymbol{\gamma}}_{\lambda_r}|\hat{\boldsymbol{\gamma}}) - E\{Q(\boldsymbol{\gamma}_{\mathcal{S}_{\lambda_r}}^*|\hat{\boldsymbol{\gamma}})\} = O_p(n^{1/2})$ *for* $r = 1, 2$*, thus* $\mathrm{dIC}_Q(\boldsymbol{\lambda}_2, \boldsymbol{\lambda}_1) > 0$ *in probability as* $n^{-1/2}\delta_c(\boldsymbol{\lambda}_2, \boldsymbol{\lambda}_1) \xrightarrow{p} \infty$*;*

(iii) If $Q(\hat{\boldsymbol{\gamma}}_{\lambda_1}|\hat{\boldsymbol{\gamma}}) - Q(\hat{\boldsymbol{\gamma}}_{\lambda_2}|\hat{\boldsymbol{\gamma}}) = O_p(1)$*, thus* $\mathrm{dIC}_Q(\boldsymbol{\lambda}_2, \boldsymbol{\lambda}_1) > 0$ *in probability as* $\delta_c(\boldsymbol{\lambda}_2, \boldsymbol{\lambda}_1) \xrightarrow{p} \infty$*.*

Theorem 3.2(i) indicates that $\mathrm{IC}_{\mathrm{Q}}(\boldsymbol{\lambda})$ selects all the significant covariates with probability 1 for any $\mathcal{S}_\lambda \not\supset \mathcal{S}_I$. Generally, the widely used criterion such as the BIC criterion $\hat{c}_n(\boldsymbol{\gamma}) = \dim(\boldsymbol{\gamma})\log(n)$ and AIC criterion $\hat{c}_n(\boldsymbol{\gamma}) = 2\dim(\boldsymbol{\gamma})$ satisfy the condition $\delta_c(\boldsymbol{\lambda}, \mathbf{0}) = o_p(n)$. The condition $\mathrm{liminf}_n \delta_Q(\boldsymbol{\lambda}, \mathbf{0})/n > 0$ is used to elucidate the effect of the underfitted model [49] and to ensure that the IC_Q criterion can select a better model with large $E\{Q(\boldsymbol{\gamma}_{\mathcal{S}}^*|\boldsymbol{\gamma}^*)\}$.

By Theorem 3.2(ii) and (iii), if $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ have the same average $n^{-1}E[Q(\boldsymbol{\gamma}_{\mathcal{S}_\lambda}^*|\boldsymbol{\gamma}^*)]$, then the IC_{Q} criterion selects the optimal model $\mathcal{S}_{\lambda_1}$ when $\delta_c(\boldsymbol{\lambda}_2, \boldsymbol{\lambda}_1)$ increases to ∞ at a certain rate. For example, when $\mathcal{S}_{\lambda_1} \subset \mathcal{S}_{\lambda_2}$, since the BIC criterion $\delta_c(\boldsymbol{\lambda}_2, \boldsymbol{\lambda}_1) = \{\dim(\hat{\boldsymbol{\gamma}}_{\mathcal{S}_{\lambda_2}}) - \dim(\hat{\boldsymbol{\gamma}}_{\mathcal{S}_{\lambda_1}})\}\log(n)$ increases to ∞ at a rate $\log(n)$, $\mathcal{S}_{\lambda_1}$ is thus selected. However, the AIC criterion $\delta_c(\boldsymbol{\lambda}_2, \boldsymbol{\lambda}_1) = 2\{\dim(\hat{\boldsymbol{\gamma}}_{\mathcal{S}_{\lambda_2}}) - \dim(\hat{\boldsymbol{\gamma}}_{\mathcal{S}_{\lambda_1}})\}$ does not satisfy the above mentioned condition, then the model selected by the AIC criterion tends to be an overfitted model. Thus, we extend the results given in Garcia, Ibrahim and Zhu [33, 34] and Ibrahim et al. [50] to our considered GPNMs with nonignorable missing responses.

4. NUMERICAL EXAMPLES

In this section, simulation studies were conducted to investigate the finite sample performance of the above proposed methodologies, and an example from the AIDS Clinical Trials Group was used to illustrate the preceding proposed methodologies.

4.1 Simulation studies

In the first simulation study, for $i = 1, \ldots, n$, covariates x_{ij}'s were independently generated from the standard normal distribution $\mathrm{N}(0, 1)$ for $j = 1, \ldots, 8$, t_i's were independently simulated from the uniform distribution $\mathrm{U}(0, 1)$, y_i's were independently drawn from the normal distribution $\mathrm{N}(\mu_i, \sigma^2)$ with $\mu_i = \exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) + g(t_i)$ and $\sigma^2 = 1$, where $g(t) = \cos(3\pi t)$ and $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{i8})^{\mathrm{T}}$. Clearly, the above generated data set was from a GPNM. Here, the true value of $\boldsymbol{\beta}$ was set to be $\boldsymbol{\beta}^* = (0.5, 0.5, 0, 0, 0.5, 0, 0, 0)^{\mathrm{T}}$, which indicated that there were three non-zero coefficients and five zero coefficients in $\boldsymbol{\beta}$. To create missing data,

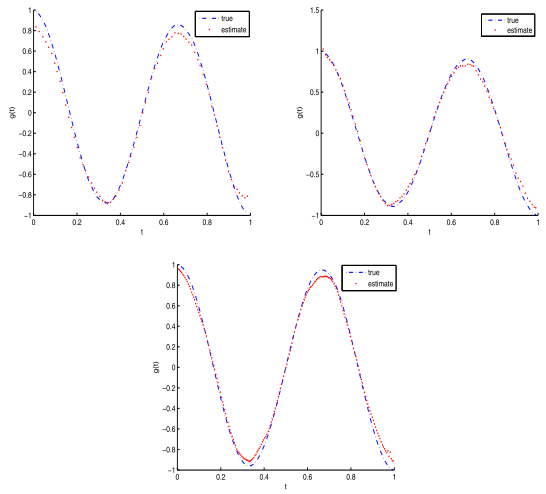

Figure 1. True curve of $g(t)$ against its estimated curve for $n = 75$ (top left panel), 100 (top right panel) and 200 (bottom panel) in the first simulation study.

we considered the following missingness data mechanism: $\text{logit}(\pi_i) = \varphi_0 + \varphi_1 x_{i3} + \varphi_2 x_{i4} + \varphi_3 y_i$, and took the true value of $\boldsymbol{\varphi} = (\varphi_0, \varphi_1, \varphi_2, \varphi_3)^{\mathrm{T}}$ to be $\boldsymbol{\varphi}^* = (1.2, 0, 0, 0.5)^{\mathrm{T}}$, which implied that there were two zero coefficients and two non-zero coefficients in $\boldsymbol{\varphi}$, where $\pi_i = \Pr(\delta_i = 1|y_i, \boldsymbol{z}_i)$ in which δ_i is the missing indicator for y_i, and $\boldsymbol{z}_i = (x_{i3}, x_{i4})^{\mathrm{T}}$. Here, we considered three different numbers of observations (e.g., $n = 75$, 100 and 200). The average missing proportion was about 15.7%.

For each of 100 data sets generated above, we used the above introduced EM algorithm together with the MH algorithm to evaluate estimates of unknown parameters in $\boldsymbol{\beta}$ and nonparametric function $g(t)$. To approximate Q function in implementing the E-step of EM algorithm, we generated 1000 observations (i.e., $\mathcal{M} = 1000$) from the conditional distribution $p(y_i|\boldsymbol{x}_i, t_i, \delta_i; \boldsymbol{\beta}, \boldsymbol{\varphi})$ of missing y_i via the MH algorithm. For the MH algorithm, we took a normal proposal distribution with $\sigma_y^2 = 16$, giving an average acceptance rate 0.303. To estimate nonparametric function, we took the kernel function to be $K(u) = (2\pi)^{-1/2}\exp(-u^2/2)$ and set the bandwidth h to be $c\hat{\sigma}_{\mathcal{T}} n^{-1/5}$, where c was taken to be 0.2, and $\hat{\sigma}_{\mathcal{T}}$ was the standard deviation of observations $\{t_i : i = 1, \ldots, n\}$.

Figure 1 plotted the true value of $g(t)$ against its estimated value for $n = 75$, 100 and 200. Examination of Figure 1 showed that the above proposed nonparametric estimation procedure was efficient in the sense that the estimated value of $g(t)$ fitted its true value well. Parameter estimations were presented in Table 1, where 'Bias' was the absolute difference between the true value and the mean of the estimates based on 100 replications. 'RMS' was the root mean square between the estimates based on 100 replications and its true value, and 'SD' was the standard deviation of the estimates based on 100 replications. Inspection of Table 1 showed that

Table 1. Performance of MLEs of parameters in the first simulation study

	$n = 75$			$n = 100$			$n = 200$		
Par.	Bias	RMS	SD	Bias	RMS	SD	Bias	RMS	SD
β_1	0.011	0.131	0.130	0.004	0.070	0.070	0.001	0.038	0.038
β_2	0.005	0.118	0.118	0.009	0.072	0.072	0.005	0.042	0.041
β_3	0.006	0.110	0.110	0.004	0.077	0.077	0.003	0.040	0.040
β_4	0.011	0.104	0.103	0.007	0.080	0.079	0.003	0.050	0.050
β_5	0.007	0.088	0.088	0.014	0.069	0.068	0.004	0.043	0.043
β_6	0.002	0.103	0.103	0.009	0.080	0.080	0.002	0.044	0.044
β_7	0.003	0.111	0.111	0.009	0.086	0.085	0.005	0.047	0.047
β_8	0.015	0.106	0.106	0.008	0.072	0.071	0.002	0.043	0.043
φ_0	0.168	0.459	0.427	0.153	0.409	0.379	0.092	0.211	0.190
φ_1	0.006	0.353	0.353	0.001	0.351	0.351	0.026	0.213	0.211
φ_2	0.036	0.380	0.378	0.017	0.371	0.378	0.001	0.225	0.225
φ_3	0.010	0.372	0.371	0.035	0.343	0.341	0.054	0.165	0.156

(i) MLEs of $\boldsymbol{\beta}$ and $\boldsymbol{\varphi}$ were reasonably accurate in the sense that almost all the Bias values of parameters were less than 0.1, and the RMS values of parameters were quite close to their corresponding SD values; (ii) increasing sample size improved the accuracy of parameter estimation as expected.

Also, for each of 100 data sets generated above, the above introduced EM algorithm and variable selection procedure together with (i) the SCAD penalty function (denoted as EM-SCAD method) and (ii) the ALASSO penalty function (denoted as EM-ALASSO method) were used to evaluate the MPL estimates of parameters in $\boldsymbol{\beta}$ and $\boldsymbol{\varphi}$ and select important covariates. Following Fan and Li [24], we took the SCAD penalty of the form: $\dot{p}_\lambda(|\gamma|) = \lambda I(|\lambda| \leq \lambda) + \frac{(a\lambda - |\gamma|)_+}{(a-1)} I(|\gamma| > \lambda)$ for $|\gamma| > 0$, where $\dot{p}_\lambda(|\gamma|) = dp_\lambda(\gamma)/d\gamma$, $I(\cdot)$ was an indicator function, $f_+ = \max\{f, 0\}$ and a was taken to be 3.7. For the EM-SCAD method, we set $\lambda_{\beta,j} = \lambda_{01}$ for $j = 1, \ldots, 8$ and $\lambda_{\varphi,k} = \lambda_{02}$ for $k = 1, \ldots, 4$. Following Zou [25], the ALASSO penalty functions were taken to be $p_{\lambda_{\beta,j}}(|\beta_j|) = \lambda_{01}|\beta_j|/|\hat{\beta}_j|^\tau$ and $p_{\lambda_{\varphi,k}}(|\varphi_k|) = \lambda_{02}|\varphi_k|/|\hat{\varphi}_k|^\tau$, where $\hat{\beta}_j$ and $\hat{\varphi}_k$ were MLEs of β_j and φ_k, respectively, and $\tau > 0$ was set to be 1. To evaluate IC_Q, the BIC criterion $c_n(\boldsymbol{\gamma}) = \dim(\boldsymbol{\gamma})\log(n)$ was here used. Table 2 presented the average number of zero coefficients correctly identified to be zero (i.e., the column labeled 'Correct' in Table 2) and the average number of nonzero coefficients incorrectly detected to be zero (i.e., the column labeled 'Incorrect' in Table 2). For comparison, Table 2 also depicted the performance of the oracle estimators of parameters.

To investigate the finite sample performance of the proposed MPL estimators, we calculated model error $\text{ME}(\hat{\boldsymbol{\beta}}_\lambda) = (\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*)^{\mathrm{T}} E(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}})(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*)$ for each of 100 MPL estimates $\hat{\boldsymbol{\beta}}_\lambda$. Since there was not a closed form of model error for $\hat{\boldsymbol{\varphi}}_\lambda$, we approximated model error of $\hat{\boldsymbol{\varphi}}_\lambda$ via Monte Carlo samples. To compare the performance of the proposed MPL estimator and MLE, we calculated the relative model error of MPL estimator to MLE for parameter vector $\boldsymbol{\gamma}$ via

Table 2. Simulation results for variable selection in the first simulation study

		$\hat{\boldsymbol{\beta}}_\lambda$ with NMAR(Complete Case)			$\hat{\boldsymbol{\varphi}}_\lambda$		
		MRME	# of 0 coeff.		MRME	# of 0 coeff.	
n	Meth.	(%)	C	IC	(%)	C	IC
75	MS	98.78(89.60)	4.49(4.55)	0.13(0.09)	79.16	1.80	0.05
	MA	41.99(37.71)	4.82(4.93)	0.08(0.04)	78.03	1.66	0.21
	MO	23.23(22.65)	5.00(5.00)	0.00(0.00)	57.22	2.00	0.00
100	MS	93.20(96.22)	4.46(4.46)	0.01(0.00)	75.10	1.73	0.03
	MA	42.38(36.59)	4.90(4.92)	0.01(0.00)	83.10	1.76	0.18
	MO	16.43(16.15)	5.00(5.00)	0.00(0.00)	64.53	2.00	0.00
200	MS	85.52(85.31)	4.81(4.81)	0.00(0.00)	91.69	1.85	0.00
	MA	27.35(29.32)	4.93(4.97)	0.00(0.00)	80.15	1.80	0.03
	MO	23.29(23.76)	5.00(5.00)	0.00(0.00)	68.36	2.00	0.00

Note: 'MS' denotes the SCAD method, 'MA' represents the ALASSO method, 'MO' represents the Oracle method. 'C' represents the average number of zero coefficients correctly identified to be zero for 100 replications, 'IC' denotes the average number of nonzero coefficients incorrectly detected to be zero for 100 replications.

RME=ME($\hat{\boldsymbol{\gamma}}_\lambda$)/ME($\hat{\boldsymbol{\gamma}}$). The median of the relative model errors (MRME) for 100 simulated datasets for the SCAD and ALASSO penalty functions were given in Table 2.

Examination of Table 2 showed that (i) the proposed MPL estimator performed better than the MLE regardless of sample sizes and the adopted penalty functions because all the MRME values were less than 1; (ii) for $\hat{\boldsymbol{\beta}}_\lambda$, the MPL estimators obtained under the complete case assumption performed as good as those obtained under nonignorable missing assumption, the EM-ALASSO method behaved better than the EM-SCAD method regardless of the nonignorable missing or complete case assumptions, and the former significantly reduced model error regardless of sample sizes; (iii) for $\hat{\boldsymbol{\varphi}}_\lambda$, the SCAD method performed as similar as the ALASSO when sample size is 75 and 100; while the SCAD method outperformed the ALASSO method when sample size is 200; (iv) the performance of the ALASSO method was expected to be as good as that of the oracle estimator as sample size n increases (e.g., $n = 200$).

In the second simulation study, 100 data sets $\{y_i : i = 1, \ldots, n\}$ were generated from the Poisson distribution Poisson(μ_i) with the conditional mean satisfying $\log(\mu_i) = \exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) + g(t_i)$, where $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{i8})^{\mathrm{T}}$ and $g(t) = \cos(4\pi t)$. For $j = 1, \ldots, 8$, covariates x_{ij}'s were independently generated from the standard normal distribution N(0, 1). For $i = 1, \ldots, n$, t_i's were independently simulated from the uniform distribution U(0, 1). The missingness data mechanism was $\mathrm{logit}(\pi_i) = \varphi_0 + \varphi_1 x_{i3} + \varphi_2 x_{i4} + \varphi_3 x_{i5} + \varphi_4 y_i$, where $\pi_i = \Pr(\delta_i = 1 | y_i, \boldsymbol{z}_i)$ in which $\boldsymbol{z}_i = (x_{i3}, x_{i4}, x_{i5})^{\mathrm{T}}$. The true values of $\boldsymbol{\beta}$ and $\boldsymbol{\varphi} = (\varphi_0, \varphi_1, \varphi_2, \varphi_3, \varphi_4)^{\mathrm{T}}$ were taken to be $\boldsymbol{\beta}^* = (0.5, 0.5, 0, 0, 0.5, 0, 0, 0)^{\mathrm{T}}$ and $\boldsymbol{\varphi}^* = (1, 0.7, 0, 0, 0.3)^{\mathrm{T}}$, respectively, which indicated that there were five zero coefficients in $\boldsymbol{\beta}^*$ and two zero coefficients in $\boldsymbol{\varphi}^*$. We considered $n = 150$, 200 and 300. The average missing proportion was about 13.94%.

Similarly, for each of the above generated 100 data sets, the above proposed EM algorithm and variable selection procedure were used to evaluate the MLEs and MPL estimates of parameters in $\boldsymbol{\beta}$ and $\boldsymbol{\varphi}$, and select the important covariates. The settings given in the first simulation study were used except for $\sigma_y^2 = 8^2$ in the proposal distribution for generating missing values of y_i's, which leaded to the average acceptance rate 0.257, and $c = 0.15$ in setting the bandwidth.

Table 3. Performance of MLEs of parameters in the second simulation study

	$n = 150$			$n = 200$			$n = 300$		
Par.	Bias	RMS	SD	Bias	RMS	SD	Bias	RMS	SD
β_1	0.007	0.015	0.014	0.009	0.015	0.012	0.010	0.014	0.010
β_2	0.012	0.021	0.017	0.009	0.015	0.012	0.010	0.013	0.009
β_3	0.002	0.017	0.017	0.003	0.013	0.013	0.002	0.012	0.011
β_4	0.002	0.018	0.018	0.002	0.014	0.013	0.003	0.013	0.013
β_5	0.010	0.017	0.014	0.009	0.014	0.011	0.009	0.013	0.010
β_6	0.002	0.017	0.017	0.001	0.015	0.015	0.004	0.011	0.010
β_7	0.003	0.018	0.017	0.003	0.015	0.014	0.005	0.011	0.010
β_8	0.002	0.018	0.018	0.004	0.011	0.010	0.002	0.011	0.011
φ_0	0.023	0.496	0.495	0.103	0.355	0.340	0.027	0.303	0.302
φ_1	0.040	0.356	0.353	0.082	0.286	0.274	0.007	0.233	0.233
φ_2	0.012	0.337	0.336	0.038	0.259	0.256	0.007	0.185	0.185
φ_3	0.008	0.344	0.344	0.049	0.258	0.254	0.012	0.223	0.223
φ_4	0.031	0.169	0.166	0.009	0.127	0.127	0.010	0.084	0.084

Figure 2 plotted the true curve of $g(t)$ against its estimated curve for 100 replications. From Figure 2, we observed that the estimated value of $g(t)$ fitted its true value well. The MLEs of parameters were presented in Table 3. Examination of Table 3 indicated that (i) MLEs of $\boldsymbol{\beta}$ and $\boldsymbol{\varphi}$ were reasonably accurate in the sense that their corresponding Bias values were less than 0.1, and RMS values were quite close to their corresponding SD values for $\boldsymbol{\beta}$ and $\boldsymbol{\varphi}$; (ii) increasing sample size improved the accuracy of parameter estimation as expected.

Table 4 presented the median of the relative model er-

Table 4. Simulation results for variable selection in the second simulation study

		$\hat{\boldsymbol{\beta}}_\lambda$ with NMAR(Complete Case)			$\hat{\boldsymbol{\varphi}}_\lambda$		
		MRME	# of 0 coeff.		MRME	# of 0 coeff.	
n	Meth	(%)	C	IC	(%)	C	IC
150	MS	88.5(80.8)	4.06(4.03)	0.03(0.03)	98.2	1.10	0.11
	MA	69.2(55.0)	4.80(4.78)	0.01(0.00)	93.3	1.55	0.20
	MO	22.4(22.2)	5.00(5.00)	0.00(0.00)	65.2	2.00	0.00
200	MS	94.0(98.7)	4.15(4.14)	0.01(0.01)	113.6	1.49	0.03
	MA	83.2(51.0)	4.94(4.84)	0.00(0.00)	99.9	1.80	0.08
	MO	28.5(30.6)	5.00(5.00)	0.00(0.00)	66.8	2.00	0.00
300	MS	100.0(100.0)	3.65(3.70)	0.00(0.00)	124.0	1.30	0.01
	MA	81.2(92.3)	4.66(4.71)	0.00(0.00)	106.8	1.76	0.04
	MO	39.7(41.4)	5.00(5.00)	0.00(0.00)	74.3	2.00	0.00

Note: 'MS' denotes the SCAD method, 'MA' represents the ALASSO method, 'MO' represents the Oracle method.
'C' represents the average number of zero coefficients correctly identified to be zero for 100 replications,
'IC' denotes the average number of nonzero coefficients incorrectly detected to be zero for 100 replications.

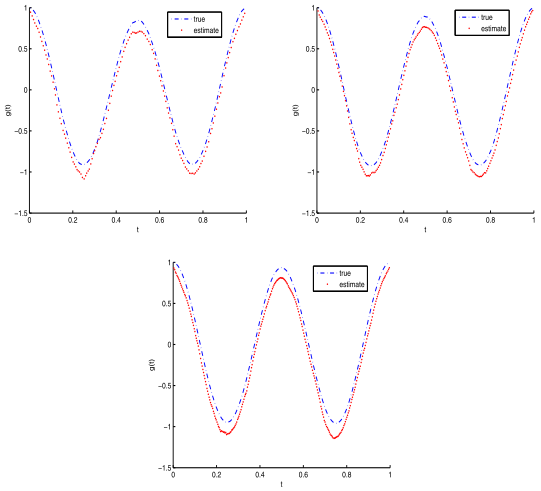

Figure 2. True curve of $g(t)$ against its estimated curve for $n = 150$ (top left panel), 200 (top right panel) and 300 (bottom panel) in the second simulation study.

rors (MRME) for 100 replications, and the average numbers of zero coefficients correctly identified to be zero and nonzero coefficients incorrectly detected to be zero. Examination of Table 4 showed that (i) for $\hat{\boldsymbol{\beta}}_\lambda$, the MPL estimates under nonignorable missing assumption outperformed those under the complete case assumption, the EM-ALASSO estimates performed better than the MLEs regardless of sample sizes, the complete case assumptions, and the adopted penalty functions because all the MRME values were less than 1; the EM-ALASSO estimates outperformed the MLEs for $n = 150$ and 200, while the EM-SCAD estimates performed the same good as the MLEs for $n = 300$; (ii) for $\hat{\boldsymbol{\varphi}}_\lambda$, the MPL estimate performed better than the MLE for $n = 150$ regardless of the adopted penalty functions, the EM-ALASSO estimate performed the same good as the MLE for $n = 200$, the MLE outperformed the EM-

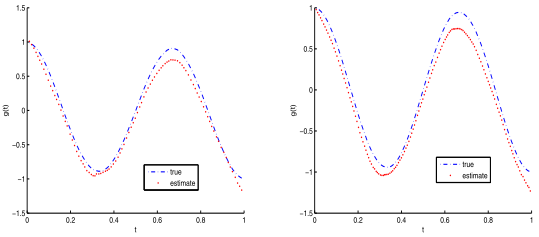

Figure 3. True curve of $g(t)$ against its estimated curve for $n = 100$ (left panel), and 150 (right panel) in the third simulation study.

SCAD estimate for $n = 200$, the MLE outperformed the MPL estimate regardless of the adopted penalty functions for $n = 300$; (iii) for both $\hat{\boldsymbol{\beta}}_\lambda$ and $\hat{\boldsymbol{\varphi}}_\lambda$, the ALASSO method behaved better than the SCAD procedure regardless of sample sizes.

To investigate the effect of missing proportion, we conducted the third simulation study. In this simulation study, 100 data sets $\{(y_i, \boldsymbol{x}_i, t_i, \delta_i) : i = 1, \ldots, n\}$ were generated as in the second simulation study except for $g(t) = \cos(3\pi t)$, and $\text{logit}(\pi_i) = \varphi_0 + \varphi_1 x_{i3} + \varphi_2 x_{i4} + \varphi_3 y_i$, where $\pi_i = \Pr(\delta_i = 1 | y_i, \boldsymbol{z}_i)$ and $\boldsymbol{z}_i = (x_{i3}, x_{i4})^{\mathrm{T}}$. The true values of $\boldsymbol{\beta}$ and $\boldsymbol{\varphi} = (\varphi_0, \varphi_1, \varphi_2, \varphi_3)^{\mathrm{T}}$ were taken to be $\boldsymbol{\beta}^* = (0.5, 0.5, 0, 0, 0.5, 0, 0, 0)^{\mathrm{T}}$ and $\boldsymbol{\varphi}^* = (-1.2, 0, 0, 0.5)^{\mathrm{T}}$, respectively. Here, we considered $n = 100$ and $n = 150$. The average missing proportion was about 36.7%. We calculated the MLEs of parameters via EM algorithm with the same settings as in the first simulation study except for $\sigma_y^2 = 6^2$, giving an average acceptance rate of 0.313, and $c = 0.3$ in selecting the bandwidth. Figure 3 plotted the true curve of $g(t)$ against its estimated curve for 100 replications. Examination of Figure 3 indicated that the estimated curve of $g(t)$ fitted its true curve well. The MLEs of $\boldsymbol{\beta}$ and $\boldsymbol{\varphi}$ were given in Table 5. Comparing Table 3 and Table 5, we obtained the same observations, which showed that there was not effect

Table 5. Performance of MLEs of parameters in the third simulation study

	$n = 100$			$n = 150$		
Par.	Bias	RMS	SD	Bias	RMS	SD
β_1	0.008	0.029	0.027	0.009	0.023	0.021
β_2	0.008	0.026	0.025	0.010	0.019	0.016
β_3	0.001	0.030	0.030	0.003	0.020	0.020
β_4	0.001	0.028	0.028	0.004	0.018	0.018
β_5	0.007	0.027	0.025	0.011	0.024	0.021
β_6	0.005	0.022	0.022	0.002	0.020	0.020
β_7	0.007	0.026	0.025	0.002	0.015	0.015
β_8	0.004	0.026	0.026	0.002	0.017	0.016
φ_0	0.130	0.614	0.600	0.046	0.400	0.398
φ_1	0.006	0.312	0.312	0.002	0.269	0.269
φ_2	0.003	0.315	0.315	0.000	0.228	0.228
φ_3	0.048	0.228	0.222	0.006	0.142	0.142

of missing proportions on the performance of the MLEs of parameters.

Table 6 presented the results of variable selection corresponding to Table 4. Examination of Table 6 showed that (i) for $\hat{\boldsymbol{\beta}}_\lambda$, the MPL estimators performed better than the MLEs regardless of sample sizes and the adopted penalty functions because all the MRME values were less than 1; (ii) for $\hat{\boldsymbol{\varphi}}_\lambda$, the EM-ALASSO estimators outperformed the MLEs regardless of sample sizes, but the EM-SCAD estimation procedure performed as good as the MLE method for $n = 100$, and the MLE method outperformed the EM-SCAD estimation procedure for $n = 150$; (iii) for $\hat{\boldsymbol{\beta}}_\lambda$ and $\hat{\boldsymbol{\varphi}}_\lambda$, the ALASSO method performed better than the SCAD method. The above results indicated that the proposed estimation method and variable selection procedure can be used to the situation where the nonignorable missing proportion was relatively high.

4.2 A real data

In this subsection, a data set from AIDS Clinical Trials Group Protocol (ACTG175) [51] was used to illustrate the proposed methodologies. In this clinical trail, 2139 HIV-infected patients were randomized into four groups to receive monotherapy (ZDV) or combined therapy (ADV+didanosine, ZDV+zalcitabine, and didanosine). The data set has been analyzed by Ding and Wang [52] when comparing the treatment effect of monotherapy and combined therapy for male patients. Inspired by Ding and Wang [52], we only used the data set from the monotherapy treatment for 100 female patients to illustrate the proposed parameter estimation procedure and covariate selection method. To wit, our main objective is to simultaneously estimate parameters in the considered model and select important factors leading to missingness of responses and important explanatory factors having significant effects on the CD4 cell count at 96±5 weeks, whose decrease means the potential to develop the acquired immunodeficiency syndrome (AIDS). To this end, similar to Ding and Wang [52], we took the CD4 cell count at 96±5 weeks (CD496, y) to be response variable, regarded age as the time (i.e, variable t) measured, and took the following five characteristics: weight (x_1), CD4 cell counts at baseline (CD40, x_2), CD4 cell counts at 20 ± 5 weeks (CD420, x_3), CD8 cell counts at baseline (CD80, x_4) and CD8 cell counts at 20 ± 5 weeks (CD820, x_5) as five covariates. Due to some reasons, response variable y was subject to missingness, while five covariates were completely observed. The missing proportion of responses was 42%.

To use the proposed method to select covariates and missing data mechanism, we considered the following GPNM: $y_i \sim \text{IG}(\mu_i, \phi)$ with $\log(\mu_i) = \exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) + g(t_i)$ in which $\boldsymbol{x}_i = (1, x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5})^{\mathrm{T}}$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_5)^{\mathrm{T}}$, and the following missingness data mechanism: $\text{logit}(\pi_i) = \varphi_0 + \varphi_1 x_{i4} + \varphi_2 x_{i5} + \varphi_t t_i + \varphi_y y_i$, where $\pi_i = \Pr(\delta_i = 1 | y_i, \boldsymbol{z}_i, t_i)$ and $\boldsymbol{z}_i = (x_{i4}, x_{i5})^{\mathrm{T}}$. $\text{IG}(\alpha_1, \alpha_2)$ represents the inverse Gamma distribution with parameters α_1 and α_2.

Based on the aforementioned specifications, we used the above proposed EM algorithm to evaluate the MPL estimates of unknown parameters $\boldsymbol{\beta}$ and $\boldsymbol{\varphi} = (\varphi_0, \boldsymbol{\varphi}_z^{\mathrm{T}}, \varphi_t, \varphi_y)^{\mathrm{T}}$, where $\boldsymbol{\varphi}_z = (\varphi_1, \varphi_2)^{\mathrm{T}}$. Similar to simulation studies, we took the kernel function to be $K(u) = (2\pi)^{-1/2}\exp(-u^2/2)$ and set the bandwidth h to be $c\hat{\sigma}_{\mathcal{T}} n^{-1/5}$, where $c = 0.6$

Table 6. Simulation results for variable selection in the third simulation study

		$\hat{\boldsymbol{\beta}}_\lambda$ with NMAR(Complete Case)			$\hat{\boldsymbol{\varphi}}_\lambda$		
		MRME	# of 0 coeff.		MRME	# of 0 coeff.	
n	Meth	(%)	C	IC	(%)	C	IC
100	MS	53.2(82.9)	3.98(4.28)	0.00(0.02)	99.8	1.61	0.13
	MA	57.7(90.8)	4.87(4.87)	0.00(0.03)	82.5	1.81	0.14
	MO	13.7(14.9)	5.00(5.00)	0.00(0.00)	80.9	2.00	0.00
150	MS	71.7(88.2)	4.04(4.11)	0.00(0.01)	111.7	1.74	0.04
	MA	69.7(71.8)	4.88(4.88)	0.01(0.01)	91.4	1.89	0.05
	MO	24.9(29.3)	5.00(5.00)	0.00(0.00)	86.7	2.00	0.00

Note: 'MS' denotes the SCAD method, 'MA' represents the ALASSO method, 'MO' represents the Oracle method.
'C' represents the average number of zero coefficients correctly identified to be zero for 100 replications,
'IC' denotes the average number of nonzero coefficients incorrectly detected to be zero for 100 replications.

Table 7. Estimates (Est) and standard deviations (SD) in the ACTG175 data

Model	Cov.	SCAD		ALASSO		MLE	
		Est	SD	Est	SD	Est	SD
GPNM	Const.	1.52	0.66	1.62	0.59	5.58	1.79
	Weight	0.00	0.00	0.00	0.11	-2.67	1.28
	CD40	0.10	0.11	0.00	0.05	-0.43	0.46
	CD420	-0.52	0.32	-0.41	0.25	-1.51	0.45
	CD80	0.00	0.04	0.00	0.00	0.24	0.10
	CD820	0.00	0.05	0.00	0.01	-0.27	0.13
Missing Mechanism	Const.	0.09	0.55	0.07	1.32	0.34	1.57
	CD80	0.08	0.10	0.07	0.12	0.07	0.14
	CD820	0.00	0.09	0.00	0.10	-0.02	0.12
	AGE	0.00	1.38	0.00	2.96	0.45	3.59
	CD496	-0.12	0.06	-0.10	0.07	-0.15	0.07

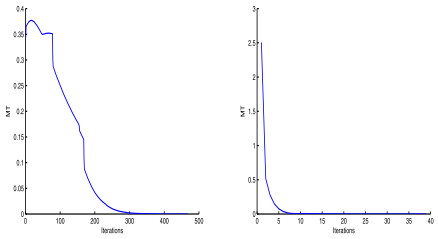

Figure 4. MT value against iteration for $\hat{\beta}_\lambda$ (left panel) and $\hat{\varphi}_\lambda$ (right panel) with the ALASSO penalty in the ACTG175 data. A small constant $c_0 = 10^{-5}$ was used to monitor the convergence of the algorithm.

and $\hat{\sigma}_{\mathcal{T}}$ was the standard deviation of t_i's. In the MH algorithm, we set $\sigma_y^2 = 20^2$ leading to an average acceptance rate 0.379. In the E-step of EM algorithm, we sampled 1000 observations (i.e., $\mathcal{M} = 1000$) from the conditional distribution $p(y_i|\boldsymbol{x}_i, t_i, \delta_i; \boldsymbol{\beta}, \boldsymbol{\varphi})$ after 200 burn-in iterations. To monitor the convergence of the proposed EM algorithm, we computed the value of statistic: $\mathrm{MT}^{(s+1)} = \max_{j\in\{1,\ldots,p+m\}}|\gamma_j^{(s+1)} - \gamma_j^{(s)}|$ for $s = 0, 1, \ldots$. To save space, we only presented index plot of $\mathrm{MT}^{(s+1)}$ for the ALASSO procedure in Figure 4. Inspection of Figure 4 showed that the proposed EM algorithm converges after about 400 iterations for $\hat{\boldsymbol{\beta}}_\lambda$ and 15 iterations for $\hat{\boldsymbol{\varphi}}_\lambda$ in terms of the convergence criterion given in Section 2.2. Hence, we took the iteration value of $\boldsymbol{\beta}$ at the 400th iteration to be its MPL estimate and set the iteration value of $\boldsymbol{\varphi}$ at the 15th iteration to be its MPL estimate. Parameter estimates were presented in Table 7, where the standard deviation (SD) was calculated via the bootstrapping resampling method for 100 replications.

Examination of Table 7 indicated that (i) the SCAD method has the same performance as the ALASSO method because they identified CD420 as the most negatively significant predictor that was not detected by the ML method, and weight as no significant predictor that was detected by the ML method, which implied that the penalized methods performed better than the ML method; (ii) the SCAD and ALASSO methods detected CD496 as the significant covariate, which implied that the missing data mechanism was nonignorable, while the ML method detected the missing data mechanism as ignorable. Also, we presented results corresponding to NMAR and MAR assumptions in Table 8, which indicated that parameter estimates were sensitive to missing data mechanism. Figure 5 depicted the estimated curve of nonparametric function $g(t)$. From Figure 5, we observed that CD496 changed as age increased.

Table 8. Estimates (Est) and standard deviations (SD) in the ACTG175 data with responses missing at random (MAR) and not missing at random (NMAR)

Model	Covariate	MAR		NMAR	
		Est	SD	Est	SD
GPNM	Constant	4.22	0.96	5.58	1.79
	Weight	-1.82	0.67	-2.67	1.28
	CD40	-0.35	0.26	-0.43	0.46
	CD420	-0.98	0.24	-1.51	0.45
	CD80	0.09	0.05	0.24	0.10
	CD820	-0.07	0.07	-0.27	0.13
Missing Mechanism	Constant	-0.69	1.19	0.34	1.57
	CD80	0.08	0.10	0.07	0.14
	CD820	-0.04	0.10	-0.02	0.12
	AGE	1.79	2.79	0.45	3.59
	CD496	—	—	-0.15	0.07

5. CONCLUDING REMARKS

This paper first discusses the estimation problem of parameters and nonparametric function in a GPNM with nonignorable missing responses by combining the local kernel estimation method and propensity score adjustment method for nonignorable nonresponse. An EM algorithm is developed to evaluate the MLEs of parameters and nonparametric function by combining the EM algorithm and the Metropolis-Hastings algorithm within the Gibbs sampler.

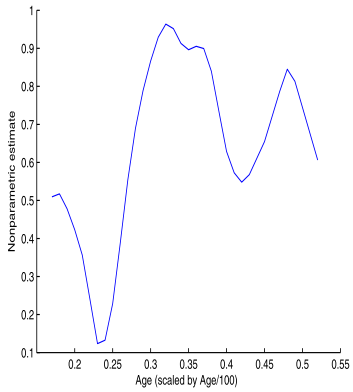

Figure 5. The estimated curve of $g(t)$ in the ACTG175 data.

Under some regularity conditions, we obtain the consistency and asymptotic normality of the proposed MLEs for parameters and nonparametric function. Simulation studies are conducted to investigate the finite sample performance of the proposed estimation procedure and MLEs. Empirical results evidence that the proposed estimation method behaves well in terms of Bias, RMS and SD.

Next, this paper considers the problem of simultaneously estimating parameters and nonparametric functions and selecting important covariates in a GPNM with nonignorable missing responses. Although variable selection procedure in the presence of missing responses/covariates has been investigated, to the best of our knowledge, there are not theories and methods developed to simultaneously select important explanatory variables in GPNMs and missingness data mechanism models. To this end, we propose a double penalized likelihood approach by imposing two nonconcave shrinkage penalties on nonlinear coefficients in a GPNM and linear coefficient in a nonignorable missingness data mechanism model to achieve model sparsity based on the SCAD and ALASSO penalty functions. We present a computationally feasible algorithm for simultaneously optimizing the penalized likelihood function and estimating the penalty parameters. Particularly, we present an IC_Q criterion to select the penalty parameters. Under some regularity conditions, we show that the proposed variable selection procedure based on IC_Q consistently selects the significant covariates in a GPNM or a nonignorable missingness data mechanism model. Simulation studies show that the proposed maximum penalized likelihood method performs better than the maximum likelihood method in terms of the median of relative model errors, the average numbers of zero coefficients correctly identified to be zero and nonzero coefficients incorrectly detected to be zero.

Although this paper only considers the situation where responses are subject to missingness, the proposed maximum penalized likelihood method can be easily extended to the case that responses and covariates are subject to missingness.

It is also interesting to consider robust estimate procedure in a GPNM with nonignorable missing data when the missingness data mechanism model is misspecified or the link function in a GPNM is misspecified.

APPENDIX

A.1 Sampling missing data via MH algorithm

It can be shown that the conditional distribution $p(y_i|\boldsymbol{x}_i, t_i, \delta_i; \boldsymbol{\beta}, \boldsymbol{\varphi})$ is proportional to

(11)
$$\begin{aligned} p(y_i|\boldsymbol{x}_i, t_i, \delta_i; \boldsymbol{\beta}, \boldsymbol{\varphi}) &\propto p(y_i|\boldsymbol{x}_i, t_i; \boldsymbol{\beta}) p(\delta_i|y_i, \boldsymbol{z}_i, t_i; \boldsymbol{\varphi}) \\ &\propto \exp\{\tfrac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi) \\ &\quad + \delta_i \boldsymbol{\varphi}^{\mathrm{T}} \boldsymbol{\omega}_i - \log(1 + \exp(\boldsymbol{\varphi}^{\mathrm{T}} \boldsymbol{\omega}_i))\} \end{aligned}$$

where $\boldsymbol{\omega}_i = (1, \boldsymbol{z}_i^{\mathrm{T}}, t_i, y_i)^{\mathrm{T}}$.

To simulate observations from the conditional distribution $p(y_i|\boldsymbol{x}_i, t_i, \delta_i; \boldsymbol{\beta}, \boldsymbol{\varphi})$, we denote $\ddot{c}(0, \phi) = \partial^2 c(y_i, \phi)/\partial y_i^2|_{y_i=0}$ and

$$\Omega_y^{-1} = \frac{\exp(\varphi_0 + \varphi_z^{\mathrm{T}} \boldsymbol{z}_i + \varphi_t t_i)}{\{1 + \exp(\varphi_0 + \varphi_z^{\mathrm{T}} \boldsymbol{z}_i + \varphi_t t_i)\}^2} \varphi_y^2 - \ddot{c}(0, \phi).$$

Then, the MH algorithm for sampling observations from (11) at the $(s+1)$th iteration with a current value $y_i^{(s)}$ is implemented as follows.

Step 1. Sample a new candidate y_i^* from $\mathrm{N}(y_i^{(s)}, \sigma_y^2 \Omega_y)$, and independently sample κ from the uniform distribution $\mathrm{U}(0, 1)$;

Step 2. if

$$\kappa \le \min\left\{1, \frac{p(y_i^*|\boldsymbol{x}_i, t_i, \delta_i; \boldsymbol{\beta}, \boldsymbol{\varphi})}{p(y_i^{(s)}|\boldsymbol{x}_i, t_i, \delta_i; \boldsymbol{\beta}, \boldsymbol{\varphi})}\right\},$$

we let $y_i^{(s+1)} = y_i^*$, otherwise let $y_i^{(s+1)} = y_i^{(s)}$. The variance σ_y^2 is chosen such that the average acceptance rate is about 0.25 or more [53].

A.2 Assumptions and proofs

To obtain asymptotic properties of the proposed estimators, we require the following assumptions.

Assumption A. Regularity conditions:

(1) The true value $\boldsymbol{\gamma}^*$ of $\boldsymbol{\gamma}$ is unique, it lies in the interior of parameter space, and the true function $g^*(t)$ of $g(t)$ is twice smooth in the interval $[0, 1]$.
(2) $f(\boldsymbol{x}_i, \boldsymbol{\beta})$ and $g(\boldsymbol{\beta}, t_i)$ are thrice continuously differentiable with respect to $\boldsymbol{\beta} \in \mathcal{B}$.
(3) Integration and differentiation with respect to $g_\beta = g(\boldsymbol{\beta}, t)$ can be interchanged in $E\{\mathcal{L}_{g_\beta}(\boldsymbol{\beta}, g_\beta)\}$.
(4) The kernel function $K(u)$ is symmetry and continuously differentiable in the interval $[-1, 1]$, and satis-

fies $\int_{-1}^{1} uK(u)du = 0$ and Lipschitz condition, i.e., $|K(u_1) - K(u_2)| \leq \alpha_0|u_1 - u_2|^{\alpha_1}$ with $\alpha_0 > 0$ and $0 < \alpha_1 \leq 1$.

(5) As $n \to \infty$, the bandwidth satisfies $h \to 0$, $nh/\log(n) \to \infty$, $h \geq \{\log(n)/n\}^{1-2/\alpha}$ with $\alpha > 2$. Define $f_{\mathbb{T}}(t)$ as the probability density of $\{t_i\}$, which is bounded, positive and continuous.

(6) Let $\mathbb{I}(t) = E\{\mathcal{L}_{g_\beta}(\boldsymbol{\beta}, g(\boldsymbol{\beta}, t))\mathcal{L}_{g_\beta}^{\mathrm{T}}(\boldsymbol{\beta}, g(\boldsymbol{\beta}, t))\}$ be the local Fisher information matrix, then $\mathbb{I}'(t)$ is bounded and continuous and $\inf_{t\in\mathcal{T}} \min\{f_{\mathbb{T}}(t), \mathbb{I}(t)\} > 0$.

(7) For any one in $\{(y_i, \boldsymbol{x}_i, t_i) : i = 1, \ldots, n\}$, $(\partial^{r+s}/\partial\beta^r\partial g_\beta^s)\mathcal{L}(\boldsymbol{\beta}, g_\beta)$ exists for $0 \leq r, s \leq 4$, $r + s \leq 4$ and $E\{\sup_\beta \sup_{g_\beta} |(\partial^{r+s}/\partial\beta^r\partial g_\beta^s)\mathcal{L}(\boldsymbol{\beta}, g_\beta)|^2\} < \infty$.

(8) There exists a neighborhood $\mathcal{N}(\boldsymbol{\beta}^*, g^*(t))$ satisfying

$$\max_{k=1,2} \sup_{t\in\mathcal{T}} \left\| \sup_{(\beta, g_\beta)\in\mathcal{N}(\beta^*, g^*(t))} \left| \frac{\partial^k}{\partial g_\beta^k} \mathcal{L}(\boldsymbol{\beta}, g_\beta) \right| \right\|_{\alpha,t} < \infty$$

for $\alpha \in (2, \infty]$, where $\|\cdot\|_{\alpha,t}$ is the L^α-norm conditioned on $\mathcal{T} = t$. Furthermore,

$$\sup_{t\in\mathcal{T}} E_t \left\{ \sup_{(\beta, g_\beta)\in\mathcal{N}(\beta^*, g^*(t))} \left| \frac{\partial^3}{\partial g_\beta^3} \mathcal{L}(\boldsymbol{\beta}, g_\beta) \right| \right\} < \infty,$$

where $E_t(\cdot) = E(\cdot|\mathcal{T} = t)$.

(9) Let $\hat{\boldsymbol{\varphi}}$ be the maximum likelihood estimator of $\boldsymbol{\varphi}$. The respondent probability $\pi(\omega, \boldsymbol{\varphi})$ is positive on the support $(Y, \boldsymbol{Z}, \mathcal{T})$ and is thrice continuously differentiable in $\boldsymbol{\varphi}$ and

$$E\left\{\frac{\partial \log\pi(\omega, \boldsymbol{\varphi})}{\partial\varphi}\right\} = 0, \quad E\left\{\frac{\partial^2 \log\pi(\omega, \boldsymbol{\varphi})}{\partial\varphi\partial\varphi^{\mathrm{T}}}\right\} = -\mathbb{J}(\boldsymbol{\varphi}),$$

where $\omega = (\mathbf{1}_n, \boldsymbol{Z}, \mathcal{T}, Y)^{\mathrm{T}}$.

Conditions A(1) and A(2) are commonly used in many literatures such as Chen [54]. Condition A(3) holds for exponential family nonlinear models, generalized linear models and nonlinear regression models. Assumptions A(4) and A(5) are quite common in nonparametric regression models, and assumption A(4) is the standard condition of the kernel function and assumption A(5) is needed to ensure the strong consistency of local kernel estimator of nonparametric function. Assumption A(6) ensures that the nonparametric function can be reasonably estimated. Assumptions A(7) and A(8) are the moment requirements imposed on the log-likelihood function, and are extensions of conditions required in deriving theories of maximum likelihood estimator in parametric models. Assumption A(9) is needed to ensure the consistency of local kernel estimator of nonparametric function when nonignorable missing data exist.

The observed data log-likelihood is given by $l(\boldsymbol{\gamma}) = \sum_{i=1}^n l_i(\boldsymbol{\gamma}) = \sum_{i=1}^n \{\delta_i \log p(y_i, \delta_i|\boldsymbol{x}_i, t_i, \boldsymbol{\gamma}) + (1 - \delta_i)\log \int p(y_i, \delta_i|\boldsymbol{x}_i, t_i, \boldsymbol{\gamma})dy_i\}$. According to White [55], even though the model is misspecified, MLE $\hat{\boldsymbol{\gamma}}$ of $\boldsymbol{\gamma}$ converges to its pseudo true value $\boldsymbol{\gamma}^*$. It is assumed that $\boldsymbol{\gamma}_n^* = \mathrm{argsup}_\gamma E\{l(\boldsymbol{\gamma})\}$. Then, without loss of generality, we assume that $E\{\partial_\gamma l_i(\boldsymbol{\gamma})\}|_{\gamma=\gamma_n^*} = 0$ holds for $\forall n$ and $\forall i$. We define $\boldsymbol{\gamma}_{\mathcal{S}}^* = \mathrm{argsup}_{\gamma:\ \gamma_j\neq 0, j\in\mathcal{S}} E\{Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^*)\}$, where the expectation $E\{\cdot\}$ is taken with respect to the probability density of the observed random variables. To derive asymptotic properties for variable selection procedure, we require the following assumptions.

Assumption B. Conditions for variable selection procedure:

(1) $\hat{\boldsymbol{\gamma}} \xrightarrow{p} \boldsymbol{\gamma}^*$, where $\hat{\boldsymbol{\gamma}}$ is the MLE of $\boldsymbol{\gamma}$.

(2) For $i = 1, \ldots, n$, the likelihood function $l_i(\boldsymbol{\gamma})$ is thrice continuously differentiable with respect to $\boldsymbol{\gamma}$ on parameter space of $\boldsymbol{\gamma}$. Furthermore, there exist functions $\mathrm{B}_i(\mathrm{D}_{o,i})$ for $i = 1, \ldots, n$ such that $l_i(\boldsymbol{\gamma})$, $|\partial_j l_i(\boldsymbol{\gamma})|^2$ and $|\partial_j\partial_k\partial_l l_i(\boldsymbol{\gamma})|$ are dominated by $\mathrm{B}_i(\mathrm{D}_{o,i})$ for all $j, k, l = 1, \ldots, p+m$, where $\mathrm{D}_{o,i}$ is the subset of $\mathbf{D}_o$ corresponding to the ith subject. The same smoothness condition holds for $E\{\log p(y_{m,i}|\mathrm{D}_{o,i}; \boldsymbol{\gamma})|\mathrm{D}_{o,i}; \boldsymbol{\gamma}\}$, where $y_{m,i} = y_i$ if $\delta_i = 1$.

(3) For any $\epsilon > 0$, there exists a finite constant K such that $\sup_{n\geq 1} n^{-1} \sum_{i=1}^n E\{\mathrm{B}_i(\mathrm{D}_{o,i})\ \mathbf{1}_{\mathrm{B}_i(\mathrm{D}_{o,i})>K}\} < \epsilon$ holds.

(4) There are positive definite matrices $\boldsymbol{A}(\boldsymbol{\gamma}^*)$ and $\boldsymbol{B}(\boldsymbol{\gamma}^*)$ such that

$$\lim_{n\to\infty} -\frac{1}{n}\sum_{i=1}^n \partial_\gamma^2 l_i(\boldsymbol{\gamma}^*) = \mathbf{A}(\boldsymbol{\gamma}^*),$$

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \partial_\gamma l_i(\boldsymbol{\gamma}^*)\partial_\gamma l_i(\boldsymbol{\gamma}^*)^{\mathrm{T}} = \mathbf{B}(\boldsymbol{\gamma}^*).$$

(5) Denote $a_n = \max(\max_{j=1,\ldots,p}\{p'_{\lambda_{\beta,j}}(|\beta_j^*|) : \beta_j^* \neq 0\}, \max_{k=1,\ldots,m}\{p'_{\lambda_{\varphi,k}}(|\varphi_k^*|) : \varphi_k^* \neq 0\})$, and $b_n = \max(\max_{j=1,\ldots,p}\{p''_{\lambda_{\beta,j}}(|\beta_j^*|) : \beta_j^* \neq 0\}, \max_{k=1,\ldots,m}\{p''_{\lambda_{\varphi,k}}(|\varphi_k^*|) : \varphi_k^* \neq 0\})$. It is assumed that

(i) $\max(\max_{j=1,\ldots,p}\{\lambda_{\beta,j} : \beta_j^* \neq 0\}, \max_{k=1,\ldots,m}\{\lambda_{\varphi,k} : \varphi_k^* \neq 0\}) = o_p(1)$;

(ii) $a_n = O_p(n^{-1/2})$;

(iii) $b_n = o_p(1)$.

(6) Denote $d_n = \min(\min_{j=1,\ldots,p}\{\lambda_{\beta,j} : \beta_j^* = 0\}, \min_{k=1,\ldots,m}\{\lambda_{\varphi,k} : \varphi_k^* = 0\})$. It is assumed that

(i) for any $j \in \{j : \beta_j^* = 0\}$, $\lim_{n\to\infty}\lambda_{\beta,j}^{-1}\mathrm{liminf}_{\epsilon\to 0_+} p'_{\lambda_{\beta,j}}(\epsilon) > 0$ holds in probability;

(ii) for any $k \in \{k : \varphi_k^* = 0\}$, $\lim_{n\to\infty}\lambda_{\varphi,k}^{-1}\mathrm{liminf}_{\epsilon\to 0_+} p'_{\lambda_{\varphi,k}}(\epsilon) > 0$ holds in probability;

(iii) $n^{1/2}d_n \xrightarrow{p} \infty$.

Proof of Lemma 2.1. According to [36], inequality (8) is equivalent to

$$E\left\{\frac{\partial}{\partial\beta}\mathcal{L}(\boldsymbol{\beta},g_\beta)\frac{\partial}{\partial\beta^{\mathrm{T}}}\mathcal{L}(\boldsymbol{\beta},g_\beta)\right\}\Big|_{\beta=\beta^*}$$
$$=\inf_g E\left\{\frac{\partial}{\partial\beta}\mathcal{L}(\boldsymbol{\beta},g_{1\beta})\frac{\partial}{\partial\beta^{\mathrm{T}}}\mathcal{L}(\boldsymbol{\beta},g_{1\beta})\right\}\Big|_{\beta=\beta^*}$$

for any other smooth curve $g_{1\beta}=g(1\beta,t)\in \mathrm{SM}(\boldsymbol{\beta})$, which indicates that $-\mathcal{L}_{g_\beta}(\boldsymbol{\beta}^*,g_{\beta^*})\partial g_{\beta^*}/\partial\beta$ is the projection of $\mathcal{L}_\beta(\boldsymbol{\beta}^*,g_{\beta^*})$ onto the span $\{\mathcal{L}_{g_\beta}(\boldsymbol{\beta}^*,g_{\beta^*})\partial g_{1\beta}/\partial\beta:\ g_{1\beta}\in \mathrm{SM}(\boldsymbol{\beta})\}$. Then, we have

$$E\left\{\left(\mathcal{L}_\beta(\boldsymbol{\beta}^*,g_{\beta^*})+\mathcal{L}_{g_\beta}(\boldsymbol{\beta}^*,g_{\beta^*})\frac{\partial g_{\beta^*}}{\partial\beta}\right)\mathcal{L}_{g_\beta}(\boldsymbol{\beta}^*,g_{\beta^*})\frac{\partial g_{1\beta}}{\partial\beta}\right\}=0,$$

that is,

$$E\left\{\mathcal{L}_{\beta g_\beta}(\boldsymbol{\beta}^*,g_{\beta^*})\frac{\partial g_{1\beta}}{\partial\beta^{\mathrm{T}}}+\mathcal{L}_{g_\beta g_\beta}(\boldsymbol{\beta}^*,g_{\beta^*})\frac{\partial g_{\beta^*}}{\partial\beta}\frac{\partial g_{1\beta}}{\partial\beta^{\mathrm{T}}}\right\}=0,$$

where $\mathcal{L}_{\beta g_\beta}(\boldsymbol{\beta},g_\beta)=\sum_{i=1}^n\mathcal{L}_{i,\beta g_\beta}(\boldsymbol{\beta},g_\beta)$ and $\mathcal{L}_{g_\beta g_\beta}(\boldsymbol{\beta},g_\beta)=\sum_{i=1}^n\mathcal{L}_{i,g_\beta g_\beta}(\boldsymbol{\beta},g_\beta)$.

Thus, we obtain

$$E\left\{\mathcal{L}_{\beta g_\beta}(\boldsymbol{\beta}^*,g_{\beta^*})\big|\mathcal{T}=t\right\}\frac{\partial g_{1\beta}}{\partial\beta^{\mathrm{T}}}$$
$$+E\left\{\mathcal{L}_{g_\beta g_\beta}(\boldsymbol{\beta}^*,g_{\beta^*})\big|\mathcal{T}=t\right\}\frac{\partial g_{\beta^*}}{\partial\beta}\frac{\partial g_{1\beta}}{\partial\beta^{\mathrm{T}}}=0,$$

for any continuous smooth curve $g_{1\beta}$ on $\mathcal{B}\times[0,1]$, the above equation yields

$$\frac{\partial g_{\beta^*}}{\partial\beta}=-\frac{E\left\{\mathcal{L}_{\beta g_\beta}(\boldsymbol{\beta}^*,g_{\beta^*})\big|\mathcal{T}=t\right\}}{E\left\{\mathcal{L}_{g_\beta g_\beta}(\boldsymbol{\beta}^*,g_{\beta^*})\big|\mathcal{T}=t\right\}},$$

which shows that Lemma 2.1 holds. □

Proof of Corollary 2.1. According to equation (6), the estimators $\hat{g}_\beta$ and $\hat{g}_\beta^{(1)}$ satisfy

$$\frac{1}{n}\sum_{i=1}^n\frac{\delta_i}{\hat{\pi}_i}K_h(t_i-t)\mathcal{L}_{i,g_\beta}(\boldsymbol{\beta},\hat{g}_\beta(t_i))(1,\frac{t_i-t}{h})=0,$$

where $\hat{g}_\beta(t_i)=\hat{g}_\beta+\hat{g}_\beta^{(1)}(t_i-t)$. By taking a partial derivative with respect to $\boldsymbol{\beta}$ on both sides of the above equation and taking the first element of the resulting vector, we have

$$\tfrac{1}{n}\sum_{i=1}^n\tfrac{\delta_i}{\hat{\pi}_i}K_h(t_i-t)\left\{\mathcal{L}_{i,\beta g_\beta}(\boldsymbol{\beta}^*,\hat{g}_{\beta^*}(t_i))\right.$$
$$\left.+\,\mathcal{L}_{i,g_\beta g_\beta}(\boldsymbol{\beta}^*,\hat{g}_{\beta^*}(t_i))\tfrac{\partial\hat{g}_{\beta^*}}{\partial\beta}\right\}=0.$$

Therefore,

$$\frac{\partial\hat{g}_{\beta^*}}{\partial\beta}=-\frac{\frac{1}{n}\sum_{i=1}^n\frac{\delta_i}{\hat{\pi}_i}K_h(t_i-t)\mathcal{L}_{i,\beta g_\beta}(\boldsymbol{\beta}^*,\hat{g}_{\beta^*}(t_i))}{\frac{1}{n}\sum_{i=1}^n\frac{\delta_i}{\hat{\pi}_i}K_h(t_i-t)\mathcal{L}_{i,g_\beta g_\beta}(\boldsymbol{\beta}^*,\hat{g}_{\beta^*}(t_i))}.$$

According to condition A(9), we obtain $||\hat{\pi}_i-\pi_i||=O_p(n^{-1/2})=o_p(1)$, and

$$\begin{aligned}&\tfrac{1}{n}\sum_{i=1}^n\tfrac{\delta_i}{\hat{\pi}_i}K_h(t_i-t)\mathcal{L}_{i,\beta g_\beta}(\boldsymbol{\beta}^*,\hat{g}_{\beta^*}(t_i))\\&=\tfrac{1}{n}\sum_{i=1}^n\tfrac{\delta_i}{\pi_i}K_h(t_i-t)\mathcal{L}_{i,\beta g_\beta}(\boldsymbol{\beta}^*,\hat{g}_{\beta^*}(t_i))+o_p(1)\\&=E\big[\tfrac{\delta}{\pi}\mathcal{L}_{\beta g_\beta}(\boldsymbol{\beta}^*,\hat{g}_{\beta^*})\big|\mathcal{T}=t\big]f_{\mathbb{T}}(t)+o_p(1)\\&=E_t\big[\mathcal{L}_{\beta g_\beta}(\boldsymbol{\beta}^*,\hat{g}_{\beta^*})\big]f_{\mathbb{T}}(t)+o_p(1),\end{aligned}\tag{12}$$

where $f_{\mathbb{T}}(t)$ is the probability density function of series $\{t_i\}$ defined in condition A(5), $E_t[\cdot]=E[\cdot|\mathcal{T}=t]$. The last equality is obtained by using the iterated expectation.

Similarly, we have

$$\begin{aligned}&\tfrac{1}{n}\sum_{i=1}^n\tfrac{\delta_i}{\hat{\pi}_i}K_h(t_i-t)\mathcal{L}_{i,g_\beta g_\beta}(\boldsymbol{\beta}^*,\hat{g}_{\beta^*}(t_i))\\&=E_t\big[\mathcal{L}_{g_\beta g_\beta}(\boldsymbol{\beta}^*,\hat{g}_{\beta^*})\big]f_{\mathbb{T}}(t)+o_p(1).\end{aligned}\tag{13}$$

From Theorem 2.1, for any $\boldsymbol{\beta}\in\mathcal{B}$, $\hat{g}_\beta(t)\xrightarrow{a.s.}g(\boldsymbol{\beta},t)$ holds. Then, from the above equations (12) and (13), we have

$$\begin{aligned}&\tfrac{1}{n}\sum_{i=1}^n\tfrac{\delta_i}{\hat{\pi}_i}K_h(t_i-t)\mathcal{L}_{i,\beta g_\beta}(\boldsymbol{\beta}^*,\hat{g}_{\beta^*}(t_i))\\&=E_t\big[\mathcal{L}_{\beta g_\beta}(\boldsymbol{\beta}^*,g_{\beta^*})\big]f_{\mathbb{T}}(t)+o_p(1),\end{aligned}\tag{14}$$

$$\begin{aligned}&\tfrac{1}{n}\sum_{i=1}^n\tfrac{\delta_i}{\hat{\pi}_i}K_h(t_i-t)\mathcal{L}_{i,g_\beta g_\beta}(\boldsymbol{\beta}^*,\hat{g}_{\beta^*}(t_i))\\&=E_t\big[\mathcal{L}_{g_\beta g_\beta}(\boldsymbol{\beta}^*,g_{\beta^*})\big]f_{\mathbb{T}}(t)+o_p(1).\end{aligned}\tag{15}$$

In addition, by condition A(5), $f_{\mathbb{T}}(t)$ is positive, bounded and continuous, we have

$$\frac{\partial\hat{g}_{\beta^*}}{\partial\beta}\longrightarrow-\frac{E_t[\mathcal{L}_{\beta g_\beta}(\boldsymbol{\beta}^*,g_{\beta^*})]}{E_t[\mathcal{L}_{g_\beta g_\beta}(\boldsymbol{\beta}^*,g_{\beta^*})]}.$$

Thus, Corollary 2.1 holds. □

Proof of Theorem 2.1. For any $\boldsymbol{\beta}\in\mathcal{B}$, $g(\boldsymbol{\beta},t_i)=g_\beta(t_i)$ can be approximated by a linear function within the neighborhood of t via the Taylor's expansion: $g_\beta(t_i)\approx g_\beta(t)+g_\beta^{(1)}(t)(t_i-t)$. Then, the local linear estimators $\hat{g}_\beta$ and $\hat{g}_\beta^{(1)}$ of g_β and $g_\beta^{(1)}$ can be obtained by solving the following equation:

$$n^{-1}\sum_{i=1}^n\frac{\delta_i}{\hat{\pi}_i}\mathcal{L}_{i,g_\beta}(\boldsymbol{\beta},\hat{g}_\beta(t_i))K_h(t_i-t)\left(1,\frac{t_i-t}{h}\right)^{\mathrm{T}}=0.$$

Since $\hat{\boldsymbol{\varphi}}$ is the MLE of $\boldsymbol{\varphi}$ and $\pi_i=\pi_i(\boldsymbol{\varphi})$ is a continuous function of $\boldsymbol{\varphi}$, $||\hat{\pi}_i-\pi_i||=O_p(n^{-1/2})$ holds. By the iterated expectation, we have

$$E\left\{\frac{1}{n}\sum_{i=1}^n\frac{\delta_i}{\hat{\pi}_i}\mathcal{L}_{i,g_\beta}(\boldsymbol{\beta},\hat{g}_\beta(t_i))K_h(t_i-t)\left(1,\frac{t_i-t}{h}\right)^{\mathrm{T}}\right\}=0,$$

where $\hat{g}_\beta(t_i)=\hat{g}_\beta+\hat{g}_\beta^{(1)}(t_i-t)$. Under conditions A(1)–A(8), similar to the argument of Lemma 4.1 given in Zhao [56],

we obtain

$$\sup_t \left| n^{-1} \sum_{i=1}^n \tfrac{\delta_i}{\hat\pi_i} \mathcal{L}_{i,g_\beta}(\boldsymbol{\beta}, \hat g_\beta(t_i)) K_h(t_i - t) \left(\tfrac{t_i - t}{h}\right)^j \right| = O\left(\left\{ \tfrac{\log(n)}{nh} \right\}^{\frac{1}{2}} + h^2 \right) \ (a.s.),$$

for $j = 0$ and 1. Hence, it follows from Theorem 2.2 of Zhao [56] that the local estimators $(\hat g_\beta, \hat g_\beta^{(1)})$ exist and satisfy

$$\sup_t |\hat g_\beta(t) - g(\beta, t)| = O\left(\left\{ \frac{\log(n)}{nh} \right\}^{\frac{1}{2}} + h^2 \right) \ (a.s.),$$

$$\sup_t |\hat g_\beta^{(1)}(t) - \frac{\partial}{\partial t} g(\beta, t)| = O\left(\left\{ \frac{\log(n)}{nh} \right\}^{\frac{1}{2}} + h^2 \right) \ (a.s.),$$

which imply that $\hat g_\beta(t) \to g(\beta, t)$ $(a.s.)$ and $\hat g_\beta^{(1)}(t) \to \partial g(\beta, t)/\partial t$ $(a.s.)$.

For any fixed t, we denote

$$u(\boldsymbol{\beta}, a_0) = n^{-1} \sum_{i=1}^n \frac{\delta_i}{\hat\pi_i} \partial_{a_0} \mathcal{L}_i(\boldsymbol{\beta}, a_0 + \hat g_\beta^{(1)}(t_i - t)) K_h(t_i - t),$$

where $\partial_{a_0} = \partial/\partial a_0$. Since $\hat g_\beta^{(1)}(t) \to \partial g(\beta, t)/\partial t$ and $\hat g_\beta$ is the local estimator of g_β, it follows from conditions A(2)–A(3) that

$$\begin{aligned} 0 &= \tfrac{\partial}{\partial \beta} u(\boldsymbol{\beta}, \hat g(\boldsymbol{\beta}, t)) \\ &= \tfrac{\partial}{\partial \beta} u(\boldsymbol{\beta}, \hat g(\boldsymbol{\beta}, t)) + \tfrac{\partial}{\partial g_\beta} u(\boldsymbol{\beta}, \hat g(\boldsymbol{\beta}, t)) \tfrac{\partial}{\partial \beta} \hat g(\boldsymbol{\beta}, t). \end{aligned}$$

Hence, we have $\partial \hat g(\boldsymbol{\beta}, t)/\partial \beta = -\mathcal{L}_{n,g_\beta g_\beta}^{-1} \mathcal{L}_{n,\beta g_\beta}$, where $\mathcal{L}_{n,g_\beta g_\beta} = n^{-1} \sum_{i=1}^n \delta_i \mathcal{L}_{i,g_\beta g_\beta}(\boldsymbol{\beta}, \hat g_\beta + \hat g_\beta^{(1)}(t_i - t)) K_h(t_i - t)/\hat\pi_i$ and $\mathcal{L}_{n,\beta g_\beta} = n^{-1} \sum_{i=1}^n \delta_i \mathcal{L}_{i,\beta g_\beta}(\boldsymbol{\beta}, \hat g_\beta + \hat g_\beta^{(1)}(t_i - t)) K_h(t_i - t)/\hat\pi_i$. Similar to the proof of Corollary 2.1 of Claeskens and Van Keilegom [57], we have

$$\begin{aligned} &\sup_t \left| \mathcal{L}_{n,g_\beta g_\beta} - E(\mathcal{L}_{g_\beta g_\beta} | \mathcal{T} = t) f_{\mathbb{T}}(t) \right| \\ &\quad = O\left(\left\{ \frac{\log(n)}{nh} \right\}^{\frac{1}{2}} + h^2 \right), \\ &\sup_t \left| \mathcal{L}_{n,\beta g_\beta} - E(\mathcal{L}_{\beta g_\beta} | \mathcal{T} = t) f_{\mathbb{T}}(t) \right| \\ &\quad = O\left(\left\{ \frac{\log(n)}{nh} \right\}^{\frac{1}{2}} + h^2 \right). \end{aligned}$$

Combining the above equations yields $\partial \hat g(\boldsymbol{\beta}, t)/\partial \beta \xrightarrow{a.s.} \partial g(\boldsymbol{\beta}, t)/\partial \beta$. Following the similar argument, we can obtain $\partial^2 \hat g_\beta(t)/\partial \beta \partial \beta^{\mathrm{T}} \xrightarrow{a.s.} \partial^2 g(\boldsymbol{\beta}, t)/\partial \beta \partial \beta^{\mathrm{T}}$. □

Proof of Theorem 2.2. Under Assumptions A(4) and A(5), following the argument of Theorem 1 of Carroll et al. [58], the first-order Taylor's expansion of Equation (6) yields

$$\begin{aligned} 0 &= n^{-1} \sum_{i=1}^n \tfrac{\delta_i}{\hat\pi_i} K_h(t_i - t) \mathcal{L}_{i,g_\beta}(\boldsymbol{\beta}, g_\beta(t_i)) H_i \\ &\quad + n^{-1} \sum_{i=1}^n \tfrac{\delta_i}{\hat\pi_i} K_h(t_i - t) \mathcal{L}_{i,g_\beta g_\beta}(\boldsymbol{\beta}, g_\beta(t_i)) H_i H_i^{\mathrm{T}} \Delta \\ &\quad + o_p(1). \end{aligned} \tag{16}$$

where $H_i = (1, (t_i - t)/h)^{\mathrm{T}}$, $\Delta = (\hat g_\beta - g_\beta,\ \hat g_\beta^{(1)} - g_\beta^{(1)})^{\mathrm{T}}$.

Let $\xi_i = \delta_i K_h(t_i - t)/\hat\pi_i$. Then, according to assumption A(9), we have

$$\sqrt{n}(\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}) = \mathbb{J}^{-1}(\boldsymbol{\varphi}) \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\partial \log \pi(\omega_i, \boldsymbol{\varphi})}{\partial \varphi} + o_p(1), \tag{17}$$

where $\omega_i = (1, \boldsymbol{z}_i^{\mathrm{T}}, t_i, y_i)^{\mathrm{T}}$. Taking the first-order Taylor's expansion of $\pi_i(\hat{\boldsymbol{\varphi}})$ yields

$$\pi_i(\hat{\boldsymbol{\varphi}}) = \pi_i(\boldsymbol{\varphi}) + (\frac{\partial \pi_i}{\partial \boldsymbol{\varphi}})^{\mathrm{T}} (\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}) + o_p(1). \tag{18}$$

Under condition A(9) and Equations (17) and (18), we obtain

$$\begin{aligned} &\tfrac{1}{n} \sum_{i=1}^n \tfrac{\xi_i (t_i - t)}{h} \\ &= \tfrac{1}{n} \sum_{i=1}^n \tfrac{\delta_i}{\pi_i} K_h(t_i - t) \tfrac{t_i - t}{h} - \tfrac{1}{n} \sum_{i=1}^n \tfrac{\hat\pi_i - \pi_i}{\pi_i^2} \delta_i K_h(t_i - t) \tfrac{t_i - t}{h} \\ &\quad + o_p(1) \\ &= -\tfrac{1}{n} \sum_{i=1}^n \tfrac{\delta_i}{\pi_i^2} (\tfrac{\partial \pi_i}{\partial \varphi})^{\mathrm{T}} \left\{ V(\boldsymbol{\varphi}) + o_p(n^{-1/2}) \right\} K_h(t_i - t) \tfrac{t_i - t}{h} \\ &\quad + o_p(1) \\ &= -V(\boldsymbol{\varphi}) \tfrac{1}{n} \sum_{i=1}^n \tfrac{\delta_i}{\pi_i^2} (\tfrac{\partial \pi_i}{\partial \varphi})^{\mathrm{T}} K_h(t_i - t) \tfrac{t_i - t}{h} + o_p(n^{-1/2}) \\ &= -V(\boldsymbol{\varphi}) \mathrm{E} \left\{ \tfrac{1}{\pi} (\tfrac{\partial \pi}{\partial \varphi})^{\mathrm{T}} \tfrac{\mathcal{T} - t}{h} | \mathcal{T} = t \right\} f_{\mathbb{T}}(t) + o_p(n^{-1/2}) \\ &= o_p(n^{-1/2}), \end{aligned}$$

where $V(\boldsymbol{\varphi}) = \mathbb{J}^{-1}(\boldsymbol{\varphi}) \frac{1}{n} \sum_{j=1}^n \partial \log \pi(\omega_j, \boldsymbol{\varphi})/\partial \boldsymbol{\varphi}$. The fourth equality is obtained by using the law of iterated expectation and the definition of kernel function.

Similarly, we have

$$\begin{aligned} &\tfrac{1}{n} \sum_{i=1}^n \xi_i \mathcal{L}_{i,g_\beta g_\beta}(\boldsymbol{\beta}, g_\beta(t_i)) \\ &= \tfrac{1}{n} \sum_{i=1}^n \tfrac{\delta_i}{\pi_i} K_h(t_i - t) \mathcal{L}_{i,g_\beta g_\beta}(\boldsymbol{\beta}, g_\beta(t_i)) \\ &\quad - \tfrac{1}{n} \sum_{i=1}^n \tfrac{\hat\pi_i - \pi_i}{\pi_i^2} \delta_i K_h(t_i - t) \mathcal{L}_{i,g_\beta g_\beta}(\boldsymbol{\beta}, g_\beta(t_i)) + o_p(1) \\ &= E\left\{ \tfrac{\delta}{\pi} \mathcal{L}_{g_\beta g_\beta}(\boldsymbol{\beta}, g_\beta(t)) | \mathcal{T} = t \right\} f_{\mathbb{T}}(t) \\ &\quad - V(\boldsymbol{\varphi}) E\left\{ \tfrac{1}{\pi} (\tfrac{\partial \pi}{\partial \varphi})^{\mathrm{T}} \mathcal{L}_{g_\beta g_\beta} | \mathcal{T} = t \right\} f_{\mathbb{T}}(t) + o_p(n^{-1/2}) \\ &= \varpi(t) f_{\mathbb{T}}(t) - V(\boldsymbol{\varphi}) \tilde\omega(t) f_{\mathbb{T}}(t) + o_p(n^{-1/2}) \\ &= \psi(t) f_{\mathbb{T}}(t) + o_p(n^{-1/2}), \end{aligned}$$

where $\varpi(t) = E\{\mathcal{L}_{g_\beta g_\beta}(\boldsymbol{\beta}, g_\beta(t)) | \mathcal{T} = t\}$, $\tilde\omega(t) = E\{(\partial \pi/\partial \varphi)^{\mathrm{T}} \mathcal{L}_{g_\beta g_\beta}/\pi | \mathcal{T} = t\}$, $\psi(t) = \varpi(t) - V(\boldsymbol{\varphi}) \tilde\omega(t)$, $\varpi(t)$ and $\tilde\omega(t)$ are obtained by using the iterated expectation. Then, it follows from Equation (16) and the proof of Theorem 1 of Carroll et al. [58] that

$$\begin{aligned} &-\psi(t) f_{\mathbb{T}}(t) (\hat g_\beta - g_\beta) \\ &\quad = \tfrac{1}{n} \sum_{i=1}^n \xi_i \mathcal{L}_{i,g_\beta}(\boldsymbol{\beta}, g_\beta(t_i)) \end{aligned}$$

$$
\begin{aligned}
&- \tfrac{h^2}{2} g^{(2)}(\boldsymbol{\beta}, t) \tfrac{1}{n} \sum_{i=1}^{n} \xi_i \left(\tfrac{t_i - t}{h}\right)^2 \mathcal{L}_{i, g_\beta g_\beta}(\boldsymbol{\beta}, g_\beta(t_i)) \\
&+ o_p\{h^2 + (nh)^{-1/2}\} \\
&= \tfrac{1}{n} \sum_{i=1}^{n} \xi_i \mathcal{L}_{i, g_\beta}(\boldsymbol{\beta}, g_\beta(t_i)) - \tfrac{h^2}{2} g^{(2)}(\boldsymbol{\beta}, t) \mu_2(K) \psi(t) f_{\mathbb{T}}(t) \\
&+ o_p\{h^2 + (nh)^{-1/2}\},
\end{aligned}
$$

which leads to asymptotic expansion of $\hat{g}_\beta(t)$. Combining the above equations yields the asymptotic bias of nonparametric estimator. □

Proof of Theorem 2.3. Under Assumptions A(1) and B(1)–B(4), it follows from White [55] that

$$
\begin{aligned}
&n^{-1/2} \sum_{i=1}^{n} \partial_\gamma l_i(\boldsymbol{\gamma}^*) \xrightarrow{D} N(\mathbf{0}, \mathbf{B}), \\
&\sqrt{n}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*) \xrightarrow{D} N(\mathbf{0}, \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}),
\end{aligned} \tag{19}
$$

where $\mathbf{A}$ and $\mathbf{B}$ are evaluated at $\boldsymbol{\gamma}^*$. □

Proof of Theorem 3.1. Let $\boldsymbol{u} = (\boldsymbol{u}_1^{\mathrm{T}}, \boldsymbol{u}_2^{\mathrm{T}}, \boldsymbol{u}_3^{\mathrm{T}})^{\mathrm{T}}$ be a $(p+m) \times 1$ vector corresponding to parameter vector $\boldsymbol{\gamma}$, where $\boldsymbol{u}_1$ is a $p_1 \times 1$ subvector corresponding to nonzero components in $\boldsymbol{\beta}^*$, $\boldsymbol{u}_2$ is a $q_1 \times 1$ subvector corresponding to nonzero components in $\boldsymbol{\varphi}^*$, and $\boldsymbol{u}_3$ is a $(p+m-p_1-q_1) \times 1$ subvector corresponding to zero components in $\boldsymbol{\beta}^*$ and $\boldsymbol{\varphi}^*$. To show that $\hat{\boldsymbol{\gamma}}_\lambda$ is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\gamma}^*$, it is sufficient to show that for an enough large $\mathbb{C}$, as $n \to \infty$, we have

$$
\begin{aligned}
p \Bigg(&\sup_{\|\boldsymbol{u}\| = \mathbb{C}} \Bigg\{ l(\tilde{\boldsymbol{\gamma}}_u) - n \sum_{j=1}^{p} p_{\lambda_{\beta,j}}(|\tilde{\beta}_{uj}|) \\
&- n \sum_{k=1}^{m} p_{\lambda_{\varphi,k}}(|\tilde{\varphi}_{uk}|) \Bigg\} - l(\boldsymbol{\gamma}^*) \\
&+ n \sum_{j=1}^{p} p_{\lambda_{\beta,j}}(|\beta_j^*|) + n \sum_{k=1}^{m} p_{\lambda_{\varphi,k}}(|\varphi_k^*|) < 0 \Bigg) \to 1,
\end{aligned}
$$

which shows that there exists a local maximizer $\hat{\boldsymbol{\gamma}}_\lambda$ of $\mathrm{PL}(\boldsymbol{\gamma}|\boldsymbol{\lambda})$ in the ball $\{\boldsymbol{\gamma} = \boldsymbol{\gamma}^* + n^{-1/2}\boldsymbol{u} : \|\boldsymbol{u}\| \le \mathbb{C}\}$ such that $\|\hat{\boldsymbol{\gamma}}_\lambda - \boldsymbol{\gamma}^*\| = O_p(n^{-1/2})$, where $\tilde{\boldsymbol{\gamma}}_u = \boldsymbol{\gamma}^* + n^{-1/2}\boldsymbol{u}$, $\tilde{\beta}_{uj} = \beta_j^* + n^{-1/2}u_j$ and $\tilde{\varphi}_{uk} = \varphi_k^* + n^{-1/2}u_{p+k}$. The second-order Taylor's expansion of $\mathrm{PL}(\boldsymbol{\gamma}|\boldsymbol{\lambda})$ yields

$$
\begin{aligned}
&l(\tilde{\boldsymbol{\gamma}}_u) - n \sum_{j=1}^{p} p_{\lambda_{\beta,j}}(|\tilde{\beta}_{uj}|) - n \sum_{k=1}^{m} p_{\lambda_{\varphi,k}}(|\tilde{\varphi}_{uk}|) \\
&\quad - l(\boldsymbol{\gamma}^*) + n \sum_{j=1}^{p} p_{\lambda_{\beta,j}}(|\beta_j^*|) + n \sum_{k=1}^{m} p_{\lambda_{\varphi,k}}(|\varphi_k^*|) \\
&\le l(\tilde{\boldsymbol{\gamma}}_u) - n \sum_{j=1}^{p_1} p_{\lambda_{\beta,j}}(|\tilde{\beta}_{uj}|) - n \sum_{k=1}^{q_1} p_{\lambda_{\varphi,k}}(|\tilde{\varphi}_{uk}|) \\
&\quad - l(\boldsymbol{\gamma}^*) + n \sum_{j=1}^{p_1} p_{\lambda_{\beta,j}}(|\beta_j^*|) + n \sum_{k=1}^{q_1} p_{\lambda_{\varphi,k}}(|\varphi_k^*|) \\
&= n^{-1/2} \boldsymbol{u}^{\mathrm{T}} \partial_\gamma l(\boldsymbol{\gamma}^*) - \tfrac{1}{2} \boldsymbol{u}^{\mathrm{T}} \{-\tfrac{1}{n} \partial_\gamma^2 l(\boldsymbol{\gamma}^*)\} \\
&\quad - n^{1/2} \sum_{j=1}^{p_1} \{p'_{\lambda_{\beta,j}}(|\beta_j^*|) \mathrm{sgn}(\beta_j^*) u_j\}
\end{aligned} \tag{20}
$$

$$
\begin{aligned}
&- n^{1/2} \sum_{k=1}^{q_1} \{p'_{\lambda_{\varphi,k}}(|\varphi_k^*|) \mathrm{sgn}(\varphi_k^*) u_k\} \\
&- \tfrac{1}{2} \sum_{j=1}^{p_1} \{p''_{\lambda_{\beta,j}}(|\beta_j^*|) u_j^2\} \\
&- \tfrac{1}{2} \sum_{k=1}^{q_1} \{p''_{\lambda_{\varphi,k}}(|\varphi_k^*|) u_k^2\} + o_p(1) \\
&\le n^{-1/2} \boldsymbol{u}^{\mathrm{T}} \partial_\gamma l(\boldsymbol{\gamma}^*) - \tfrac{1}{2} \boldsymbol{u}^{\mathrm{T}} \mathbf{A}(\boldsymbol{\gamma}^*) \boldsymbol{u} \\
&\quad + (p_1 n)^{1/2} a_n \|\boldsymbol{u}_1\| + (q_1 n)^{1/2} a_n \|\boldsymbol{u}_2\| \\
&\quad - \tfrac{1}{2} |b_n| \|\boldsymbol{u}_1\|^2 - \tfrac{1}{2} |b_n| \|\boldsymbol{u}_2\|^2 + o_p(1) \\
&\le n^{-1/2} \boldsymbol{u}^{\mathrm{T}} \partial_\gamma l(\boldsymbol{\gamma}^*) - \tfrac{1}{2} \boldsymbol{u}^{\mathrm{T}} \mathbf{A}(\boldsymbol{\gamma}^*) \boldsymbol{u} \\
&\quad + (p_1 n)^{1/2} a_n \|\boldsymbol{u}_1\| + (q_1 n)^{1/2} a_n \|\boldsymbol{u}_2\| + o_p(1).
\end{aligned}
$$

The first inequality in Equation (20) holds because of $p_\lambda(0) = 0$ and $p_\lambda(\cdot) > 0$ for the SCAD and ALASSO penalty functions. The second inequality in Equation (20) is obtained from Equation (19) and the second-order Taylor's expansion of the penalty function. Condition B(4) and $\sum_{j=1}^{p_1} |u_j| \le (p_1 \sum_{j=1}^{p_1} u_j^2)^{1/2}$ yields the third inequality in Equation (20). Equation (19) and Assumptions B(2)–B(5) indicate $n^{-1/2} \boldsymbol{u}^{\mathrm{T}} \partial_\gamma l(\boldsymbol{\gamma}^*) = O_p(1)$. Note that $\boldsymbol{u}^{\mathrm{T}} \mathbf{A}(\boldsymbol{\gamma}^*) \boldsymbol{u}$ is bounded below by $\mathbb{E}_{\min}\{\mathbf{A}(\boldsymbol{\gamma}^*)\} \|\boldsymbol{u}\|^2$, where $\mathbb{E}_{\min}\{\mathbf{A}(\boldsymbol{\gamma}^*)\}$ is the smallest eigenvalue of matrix $\mathbf{A}(\boldsymbol{\gamma}^*)$. Thus, the second term in the last inequality of Equation (20) dominates other four terms, and the last inequality in Equation (20) is negative by selecting an enough large $\mathbb{C}$. The above argument shows that Theorem 3.1(i) holds.

Now, we prove Theorem 3.1(ii). From the above argument, we obtain that $\hat{\boldsymbol{\gamma}}_\lambda$ is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\gamma}^*$ and satisfies $\|\hat{\boldsymbol{\gamma}}_\lambda - \boldsymbol{\gamma}^*\| = O_p(n^{-1/2})$ and $\|\hat{\boldsymbol{\beta}}_{(2)\lambda}\| = \|\hat{\boldsymbol{\varphi}}_{(2)\lambda}\| = O_p(n^{-1/2}) = o_p(1)$. Let $\tilde{p}_{\lambda_\beta} = \sum_{j=1}^{p} p_{\lambda_{\beta,j}}(|\beta_j|)$ and $\tilde{p}_{\lambda_\varphi} = \sum_{k=1}^{q} p_{\lambda_{\varphi,k}}(|\varphi_k|)$. Then, we have

$$
\begin{aligned}
\mathbf{0} &= n^{-1/2} \{\partial_\gamma l(\hat{\boldsymbol{\gamma}}_\lambda) - n \partial_\gamma \tilde{p}_{\lambda_\beta}|_{\gamma = \hat{\gamma}_\lambda} - n \partial_\gamma \tilde{p}_{\lambda_\varphi}|_{\gamma = \hat{\gamma}_\lambda}\} \\
&= n^{-1/2} \partial_\gamma l(\boldsymbol{\gamma}^*) - n^{1/2} (\hat{\boldsymbol{\gamma}}_\lambda - \boldsymbol{\gamma}^*)^{\mathrm{T}} \{-\tfrac{1}{n} \partial_\gamma^2 l(\boldsymbol{\gamma}^*)\} \\
&\quad + o_p(1) - n^{1/2} \partial_\gamma \tilde{p}_{\lambda_\beta}|_{\gamma = \hat{\gamma}_\lambda} - n^{1/2} \partial_\gamma \tilde{p}_{\lambda_\varphi}|_{\gamma = \hat{\gamma}_\lambda} \\
&= O_p(1) - n^{1/2} \partial_\gamma \tilde{p}_{\lambda_\beta}|_{\gamma = \hat{\gamma}_\lambda} - n^{1/2} \partial_\gamma \tilde{p}_{\lambda_\varphi}|_{\gamma = \hat{\gamma}_\lambda}.
\end{aligned} \tag{21}
$$

The second equality is obtained from the Taylor's expansion, and the last equality holds because of $n^{-1/2} \partial_\gamma l(\boldsymbol{\gamma}^*) = n^{1/2} (\hat{\boldsymbol{\gamma}}_\lambda - \boldsymbol{\gamma}^*)^{\mathrm{T}} \{-\partial_\gamma^2 l(\boldsymbol{\gamma}^*)/n\} = O_p(1)$. Note that for $j = 1, \ldots, p$, we have $-n^{1/2} \partial_{\beta_j} \tilde{p}_{\lambda_\beta}|_{\gamma = \hat{\gamma}_\lambda} = -\mathrm{sgn}(\hat{\beta}_j) n^{1/2} \lambda_{\beta,j} \{\lambda_{\beta,j}^{-1} p'_{\lambda_{\beta,j}}(|\hat{\beta}_j|)\}$. Since $\|\hat{\boldsymbol{\beta}}_{(2)\lambda}\| = o_p(1)$, Assumption B(6i) implies that $\lambda_{\beta,j}^{-1} p'_{\lambda_{\beta,j}}(|\hat{\beta}_j|) > 0$ holds for $j = 1, \ldots, p$. Then, $-\mathrm{sgn}(\hat{\beta}_j) n^{1/2} d_n$ dominates the second term of the last equality of (21). Similarly, for $k = 1, \ldots, m$, we have $-n^{1/2} \partial_{\varphi_k} \tilde{p}_{\lambda_\varphi}|_{\gamma = \hat{\gamma}_\lambda} = -\mathrm{sgn}(\hat{\varphi}_j) n^{1/2} \lambda_{\varphi,k} \{\lambda_{\varphi,k}^{-1} p'_{\lambda_{\varphi,k}}(|\hat{\varphi}_k|)\}$. Again, since $\|\hat{\boldsymbol{\varphi}}_{(2)\lambda}\| = o_p(1)$, Assumption B(6ii) implies that $\lambda_{\varphi,k}^{-1} p'_{\lambda_{\varphi,k}}(|\hat{\varphi}_k|) > 0$ holds for $k = 1, \ldots, m$. Thus, the term $-\mathrm{sgn}(\hat{\varphi}_k) n^{1/2} d_n$ dominates the third term of the last equality of (21). It

follows from Assumption B(6iii) that $n^{1/2}d_n \xrightarrow{p} \infty$, which shows that $\hat{\beta}_j$ must be zero for $j = p_1+1,\ldots,p$ and $\hat{\varphi}_k$ must be zero for $k = q_1+1,\ldots,m$, otherwise the absolute values of the gradients of the second and third terms in the last equality of (21) could be large so that the last equality of (21) is not equal to zero.

Next, we prove Theorem 3.1(iii). From the above arguments, it is easily seen that under Assumptions A(1) and B(1)–B(6), there is a consistent MPL estimator $\hat{\boldsymbol{\gamma}}_\lambda = (\hat{\boldsymbol{\beta}}_{(1)\lambda}^{\mathrm{T}}, \hat{\boldsymbol{\beta}}_{(2)\lambda}^{\mathrm{T}}, \hat{\boldsymbol{\varphi}}_{(1)\lambda}^{\mathrm{T}}, \hat{\boldsymbol{\varphi}}_{(2)\lambda}^{\mathrm{T}})^{\mathrm{T}}$ of $\boldsymbol{\gamma} = (\hat{\boldsymbol{\beta}}_{(1)}^{\mathrm{T}}, \hat{\boldsymbol{\beta}}_{(2)}^{\mathrm{T}}, \hat{\boldsymbol{\varphi}}_{(1)}^{\mathrm{T}}, \hat{\boldsymbol{\varphi}}_{(2)}^{\mathrm{T}})^{\mathrm{T}}$ satisfying $\hat{\boldsymbol{\beta}}_{(2)\lambda} = \hat{\boldsymbol{\varphi}}_{(2)\lambda} = \mathbf{0}$. Let $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_{(1)}^{*\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\varphi}^* = (\boldsymbol{\varphi}_{(1)}^{*\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\gamma}_{(1)}^* = (\boldsymbol{\beta}_{(1)}^{*\mathrm{T}}, \boldsymbol{\varphi}_{(1)}^{*\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\gamma}_{(1)} = (\boldsymbol{\beta}_{(1)}^{\mathrm{T}}, \boldsymbol{\varphi}_{(1)}^{\mathrm{T}})^{\mathrm{T}}$, $\hat{\boldsymbol{\gamma}}_{(1)\lambda} = (\hat{\boldsymbol{\beta}}_{(1)\lambda}^{\mathrm{T}}, \hat{\boldsymbol{\varphi}}_{(1)\lambda}^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\gamma}^* = (\boldsymbol{\beta}^{*\mathrm{T}}, \boldsymbol{\varphi}^{*\mathrm{T}})^{\mathrm{T}}$, and $\tilde{l}(\boldsymbol{\gamma}) = l((\boldsymbol{\beta}_{(1)}^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}}, \boldsymbol{\varphi}_{(1)}^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}}))$. Let $\tilde{\mathbf{A}}(\boldsymbol{\gamma})$ be an sub-matrix of $\mathbf{A}(\boldsymbol{\gamma}^*)$ obtained by deleting the $p_1+1,\ldots,p$ and $p+q_1+1,\ldots,p+m$ rows and columns of $\mathbf{A}(\boldsymbol{\gamma}^*)$, and $\tilde{\mathbf{B}}(\boldsymbol{\gamma})$ is similarly defined. Denote

$$\mathbf{h}_1(\boldsymbol{\gamma}) = \Big\{p'_{\lambda_{\beta,1}}(|\beta_1|)\mathrm{sgn}(|\beta_1|),\ldots,p'_{\lambda_{\beta,p_1}}(|\beta_{p_1}|)\mathrm{sgn}(|\beta_{p_1}|)\,, p'_{\lambda_{\varphi,1}}(|\varphi_1|)\mathrm{sgn}(|\varphi_1|),\ldots,p'_{\lambda_{\varphi,q_1}}(|\varphi_{q_1}|)\mathrm{sgn}(|\varphi_{q_1}|)\Big\}^{\mathrm{T}},$$

$$\mathbf{J}_1(\boldsymbol{\gamma}) = \mathrm{diag}\left\{p''_{\lambda_{\beta,1}}(|\beta_1|),\ldots,p''_{\lambda_{\beta,p_1}}(|\beta_{p_1}|)\,, p''_{\lambda_{\varphi,1}}(|\varphi_1|),\ldots,p''_{\lambda_{\varphi,q_1}}(|\varphi_{q_1}|)\right\},$$

$$\mathbf{h}(\boldsymbol{\gamma}^*) = \begin{pmatrix}\mathbf{h}_1(\boldsymbol{\gamma}^*)\\ \mathbf{0}\end{pmatrix},\quad \mathbf{J}(\boldsymbol{\gamma}^*) = \begin{pmatrix}\mathbf{J}_1(\boldsymbol{\gamma}^*) & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}$$

$$\boldsymbol{\Sigma}(\boldsymbol{\gamma}^*) = \{\tilde{\mathbf{A}}(\boldsymbol{\gamma}^*) + \mathbf{J}(\boldsymbol{\gamma}^*)\}^{-1}\tilde{\mathbf{B}}(\boldsymbol{\gamma}^*)\{\tilde{\mathbf{A}}(\boldsymbol{\gamma}^*) + \mathbf{J}(\boldsymbol{\gamma}^*)\}^{-1}.$$

Then, taking the second-order Taylor's expansion of the penalized likelihood at $\boldsymbol{\gamma}^*$ yields

$$\begin{aligned}
0 &= \partial_\gamma \tilde{l}(\hat{\boldsymbol{\gamma}}_\lambda) - n\partial_\gamma\left\{\sum_{j=1}^{p} p_{\lambda_{\beta,j}}(|\beta_{\lambda_j}|)\right\}|_{\gamma=\hat{\gamma}_\lambda}\\
&\quad - n\partial_\gamma\left\{\sum_{k=1}^{m} p_{\lambda_{\varphi,k}}(|\varphi_{\lambda_k}|)\right\}|_{\gamma=\hat{\gamma}_\lambda}\\
&= \partial_\gamma \tilde{l}(\boldsymbol{\gamma}^*) - n\mathbf{h}(\boldsymbol{\gamma}^*)\\
&\quad - n(\hat{\boldsymbol{\gamma}}_\lambda - \boldsymbol{\gamma}^*)^{\mathrm{T}}\{-\tfrac{1}{n}\partial_\gamma^2\tilde{l}(\boldsymbol{\gamma}^*) + \mathbf{J}(\boldsymbol{\gamma}^*)\} + o_p(1)\\
&= n^{-1/2}\partial_\gamma \tilde{l}(\boldsymbol{\gamma}^*) - n^{1/2}\mathbf{h}(\boldsymbol{\gamma}^*)\\
&\quad - n^{-1/2}(\hat{\boldsymbol{\gamma}}_\lambda - \boldsymbol{\gamma}^*)^{\mathrm{T}}\{\tilde{\mathbf{A}}(\boldsymbol{\gamma}^*) + \mathbf{J}(\boldsymbol{\gamma}^*)\} + o_p(1),
\end{aligned}$$

which implies $n^{1/2}\{\hat{\boldsymbol{\gamma}}_\lambda - \boldsymbol{\gamma}^* + [\tilde{\mathbf{A}}(\boldsymbol{\gamma}^*) + \mathbf{J}(\boldsymbol{\gamma}^*)]^{-1}\mathbf{h}(\boldsymbol{\gamma}^*)\} \xrightarrow{D} n^{-1/2}\{\tilde{\mathbf{A}}(\boldsymbol{\gamma}^*) + \mathbf{J}(\boldsymbol{\gamma}^*)\}^{-1}\partial_\gamma l(\boldsymbol{\gamma}^*)$. Therefore, it follows from Theorem 2.3 that Theorem 3.1(iii) holds. □

Proof of Theorem 3.2. It follows from Garcia, Ibrahim and Zhu [33], Ibrahim et al. [50] and Garcia, Ibrahim and Zhu [34] that Theorem 3.2 holds. □

ACKNOWLEDGEMENTS

The authors are grateful to the Editor, an Associate Editor, and two referee for their valuable suggestions and comments that greatly improved the manuscript. This work was supported by grants from the National Natural Science Foundation of China (No. 11671349).

Received 9 June 2016

REFERENCES

[1] Jorgensen, B. (1983). Maximum likelihood estimation and large-sample inference for generalized linear and nonlinear regression models. *Biometrika* **70** 19–28. MR0742972

[2] Cordeiro, G. M. and Paula, G. A. (1989). Improved likelihood ratio statistics for exponential family nonlinear models. *Biometrika* **76** 93–100.

[3] Cox, C. and Ma, G. (1995). Asymptotic confidence bands for generalized nonlinear regression models. *Biometrics* **51** 142–150. MR1341232

[4] Lindsey, J., Byrom, W., Wang, J., Jarvis, P. and Jones, B. (2000). Generalized nonlinear models for pharmacokinetic data. *Biometrics* **56** 81–88.

[5] Kosmidis, L. and Firth, D. (2009). Bias reduction in exponential family nonlinear models. *Biometrika* **96** 793–804. MR2564491

[6] Turner, H. and Firth, D. (2012). *Generalized nonlinear models in R: an overview of the gnm package*, R package version 1.0-6. http://CRAN.R-project.org/package=gnm.

[7] Engle, R. F., Granger, C. W., Rice, J. and Weiss, A. (1986). Semiparametric estimate of the relation between weather and electricity sales. *Journal of the American Statistical Association* **81** 310–320.

[8] Zhu, Z. Y., Tang, N. S. and Wei, B. C. (2000). On confidence regions of semiparametric nonlinear regression models (a geometric approach). *Acta Mathematica Scientia* **20** 68–75. MR1770397

[9] Hardel, W., Liang, H. and Gao, J. (2000). *Partially linear models*. Springer-Verlag, New York.

[10] Severini, T. A. and Staniswalis, J. G. (1994). Quasi-likelihood estimation in semi-parametric models. *Journal of the American Statistical Association* **89** 501–512.

[11] Li, R. and Nie, L. (2008). Efficient statistical inference procedures for partially nonlinear models and their applications. *Biometrics* **64** 904–911.

[12] Wang, Y. and Ke, C. (2009). Smoothing spline semiparametric nonlinear regression models. *Journal of Computational and Graphical Statistics* **18** 165–183. MR2649643

[13] Little, R. J. A. and Rubin, D. B. (2002). *Statistical analysis with missing data*, 2nd ed. Wiley, New York.

[14] Wang, Q. H., Lindon, Q. and Hardle, W. (2004). Semiparametric regression analysis with missing response at random. *Journal of the American Statistical Association* **99** 334–345.

[15] Liang, H., Wang, S. J. and Carroll, R. J. (2007). Partially linear models with missing response variables and error-prone covariates. *Biometrics* **94** 185–198.

[16] Liang, H. and Qin, Y. S. (2008). Empirical likelihood based inference for partially linear models with missing covariates. *Australian New Zealand Journal of Statistics* **50** 347–359.

[17] Tang, N. S. and Zhao, P. Y. (2013). Empirical likelihood semiparametric nonlinear regression analysis for longitudinal data with responses missing at random. *Annals of the Institute of Statistical Mathematics* **65** 639–665.

[18] Lee, S. Y. and Tang, N. S. (2006). Bayesian analysis of nonlinear structural equation models with nonignorable missing data. *Psychometrika* **71** 541–564. MR2272542

[19] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, in *Proceedings of the 2nd International Symposium Information Theory*, B. N. Petrov and F. Csaki (Eds.), pp. 267–281, Akademia Kiado, Budapest, Hungary, 1973.

[20] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6** 461–464.

[21] Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Seies B* **64** 583–639. MR1979380
[22] Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Annals of Statistics* **24** 2350–2383.
[23] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58** 267–288. MR1379242
[24] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96** 1348–1360.
[25] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101** 1418–1429.
[26] Wang, H. and Leng, C. (2007). Unified lasso estimation via least squares approximation. *Journal of the American Statistical Association* **101** 1039–1048. MR2411663
[27] Lv, J. and Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Annals of Statistics* **37** 3498–3528.
[28] Fan, J. and Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *Journal of the American Statistical Association* **99** 710–723.
[29] Liang, H. and Li, R. (2009). Variable selection for partially linear models with measurement errors. *Journal of the American Statistical Association* **104** 234–248.
[30] Li, R. and Liang, H. (2008). Variable selecion in semiparametric regression modeling. *Annals of Statistics* **36** 261–286.
[31] Kai, Bo, Li, R. and Zou, H. (2011). New efficient estimation and variable selection methods for semiparametric varying-coefficient partially linear models. *Annals of Statistics* **39** 305–332.
[32] Ni, X., Zhang, D. and Zhang, H. (2010). Variable selection for semiparametric mixed models in longitudinal studies. *Biometrics* **66** 79–88.
[33] Garcia, R. I., Ibrahim, J. G. and Zhu, H. T. (2010a). Variable selection for regression models with missing data. *Statistica Sinica* **20** 149–165.
[34] Garcia, R. I., Ibrahim, J. G. and Zhu, H. T. (2010b). Variable selection in the Cox regression model with covariates missing at random. *Biometrics* **66** 97–104. MR2756695
[35] Li, G., Lai, P. and Lian, H. (2015). Variable selection and estimation for partially linear single-index models with longitudinal data. *Statistics and Computing* **25** 579–593.
[36] Severini, T. A. and Wong, H. W. (1992). Profile likelihood and conditionally parametric models. *Annals of Statistics* **20** 1768–1802.
[37] Murphy, S. A. and van der Vaart, A. W. (2000). On profile likelihood. *Journal of the American Statistical Association* **95** 449–485.
[38] Fan, J. Q., Farmen, M. and Gijbels, I. (1998). Local maximum likelihood estimation and inference. *Journal of the Royal Statistical Society, Series B* **60** 591–608.
[39] Riddles, M. K. (2013). Propensity score adjusted method for missing data (Doctoral dissertation). Iowa State University.
[40] Ibrahim, J. G., Zhu, H. and Tang, N. (2008). Model selection criteria for missing data problem via the EM algorithm. *Journal of the American Statistical Association* **103** 1648–1658. MR2510293
[41] McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*, 2nd edition. Chapman and Hall, London.
[42] Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood for incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39** 1–38.
[43] Ibrahim, J. G., Chen, M. H. and Lipsitz, S. R. (1999). Monte carlo EM for missing covariates in parametric regression models. *Biometrics* **55** 591–596.
[44] Staniswalis, J. G. (1989). The kernel estimate of a regression function in likelihood-based models. *Journal of the American Statistical Association* **84** 276–283.
[45] Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics* **38** 894–932.
[46] Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood modles. *Statistica Sinica* **36** 1509–1533.
[47] Hunter, D. R. and Li, R. (2005). Variable selection using MM algorithm. *Annals of Statistics* **33** 1617–1642.
[48] Meng, X. L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80** 267–278.
[49] Wang, H., Li, R. and Tsai, C. L. (2007). Tuning parameter selector for the smoothly clipped absolute deviation method. *Biometrika* **80** 267–278.
[50] Ibrahim, J. G., Zhu, H. T., Garcia, R. I. and Guo, R. X. (2011). Fixed and random effects selection in mixed effects models. *Biometrics* **67** 495–503.
[51] Hammer, S. M., et al. (1996). Trial comparing nucleotide monotherapy with combined therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine* **335** 1081–1090.
[52] Ding, X. B. and Wang, Q. H. (2011). Fusion-refinement procedure for dimension reduction with missing response at random. *Journal of the American Statistical Association* **106** 1193–1207.
[53] Gelman, A., Roberts, G. O. and Gilks, W. R. (1996). Efficient Metropolis jumping rules, in *Bayesian Statistics* **5**, J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith (Eds), pp. 599–607. Oxford University Press, Oxford.
[54] Chen, H. (1995). Asymptotically efficient estimation in semiparametric generalized linear models. *Annals of Statistics* **23** 1102–1129.
[55] White, H. (1994). *Estimation, Inference and Specification Analysis*. Cambridge University Press, New York.
[56] Zhao, P. L. (1994). Asymptotics of kernel estimator based on local maximum likelihood. *Journal of Nonparametric Statistics* **4** 79–90.
[57] Claeskens, G. and Van Keilegom, I. (2003). Bootstrap confidence bands for regression curves and their derivatives. *Annals of Statistics* **31** 1852–1884.
[58] Carroll, R. J., Gijbels, I. and Wand, M. P. (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association* **92** 477–489.

Niansheng Tang
Yunnan Provincial Key Laboratory of Statistical Modeling and Data Analysis
Yunnan University
Kunming
China
E-mail address: nstang@ynu.edu.cn

Lin Tang
Yunnan Provincial Key Laboratory of Statistical Modeling and Data Analysis
Yunnan University
Kunming
China
E-mail address: totoroxyz@163.com