Statistical Papers (2021) 62:2627–2659
https://doi.org/10.1007/s00362-020-01205-0

REGULAR ARTICLE

Model pursuit and variable selection in the additive accelerated failure time model

Li Liu[1] · **Hao Wang**[1] · **Yanyan Liu**[1] · **Jian Huang**[2]

Received: 12 January 2020 / Revised: 29 August 2020 / Published online: 12 October 2020

Abstract

In this paper, we propose a new semiparametric method to simultaneously select important variables, identify the model structure and estimate covariate effects in the additive AFT model, for which the dimension of covariates is allowed to increase with sample size. Instead of directly approximating the non-parametric effects as in most existing studies, we take a linear effect out to weak the condition required for model identifiability. To compute the proposed estimates numerically, we use an alternating direction method of multipliers algorithm so that it can be implemented easily and achieve fast convergence rate. Our method is proved to be selection consistent and possess an asymptotic oracle property. The performance of the proposed methods is illustrated through simulations and the real data analysis.

Keywords Additive AFT model · Model pursuit · Variable selection · Penalization · ADMM algorithm

Mathematics Subject Classification 62B10 · 62G20 · 62N01

✉ Yanyan Liu
liuyy@whu.edu.cn

Li Liu
lliu.math@whu.edu.cn

Hao Wang
2015202010071@whu.edu.cn

Jian Huang
jian-huang@uiowa.edu

[1] School of Mathematics and Statistics, Wuhan University, Wuhan, China

[2] Department of Statistics and Actuarial Science, University of Iowa, Iowa City, USA

1 Introduction

The rapid development of technology and information drives high dimensional data collection in practical study areas such as genomic and health sciences. Under the sparsity assumption, various variable selection methods have been proposed to improve the accuracy of the estimation. Among them, penalized methods have been studied extensively for uncensored response. Examples include LASSO (Tibshirani 1996), adaptive LASSO (Zou 2006), SCAD (Fan and Li 2001), the Dantizg selector (Candes and Tao 2007) and MCP (Zhang 2010). Some of these methods have been adapted to analyze censored data based on classical Cox model. For example, Tibshirani (1997) and Fan and Li (2002) extended the LASSO and nonconcave penalized likelihood methods to the Cox model, respectively. Zhang and Lu (2007) developed the adaptive LASSO for Cox model. In the literature of survival analysis, a useful alternative to the Cox model is the accelerated failure time (AFT) model (Wei 1992), which assumes the linear relationship between the logarithm of survival time and covariates of interest. Compared with the Cox model, estimated parameters in the AFT model can be easily interpreted in practice. Inference procedures for the AFT model include the inverse probability weighting (IPW) method (Stute 1993, 1996), Buckley–James iterative method (BJ) and rank-based method (Buckley and James 1979; Zeng and Lin 2007). Some researchers developed variable selection methods for fitting semiparametric AFT models in high dimensional data settings (e.g. Wang et al. 2008; Huang and Ma 2010 etc.). These studies are based on the assumption that the underlying covariate structure is linear. In practice, it is unclear about the adequacy of this linear structure assumption and the impact of model misspecification on the analysis. Therefore, some authors considered more flexible models where nonlinear structure of covariate was considered, e. g. Chen et al. (2005) and Antoniadis et al. (2014). All of these works pre-specified which covariate effects are linear and which are nonlinear.

However, generally it is unknown which covariates have linear effects and which have nonlinear effects in advance for real data. This has driven studies on simultaneously identify and estimate the linear and nonlinear components, which is referred as model pursuit problem. For instance, Zhang et al. (2011) developed a two-step regularization method under the framework of partial linear model. However, they did not prove the selection consistency and their method is difficult to realize. Huang et al. (2012) transformed the model pursuit problem into a group variable selection problem and proposed an easy implement approach. Another important issue in model construction is variable selection especially when the number of covariates is large. Simultaneous model pursuit and variable selection has gained popularity in recent years. Specifically, Wang and Lin (2019) proposed a robust and efficient method to simultaneous identify model structure and select important variables for generalized partial linear varying coefficient models with longitudinal data. This problem was considered for high dimensional data under different models, e.g. additive models (Wu and Stefanski 2015), varying-coefficient models (Chen et al. 2018).

For censored failure time data, the corresponding studies on simultaneously model pursuit and variable selection are limited. Cao et al. (2016) proposed a semiparametric pursuit method based on B-spline expansions through a penalized group selection method with concave penalties. Lian et al. (2013) utilized a double-penalized method

to identify the model structure, select the relevant covariates and estimate the covariate effects for censored data under Cox models with varying coefficients. However, The methods mentioned above are not designed for high dimensional censored data. Our goal is to develop a new approach for high-dimensional right censored data not only to select the important variables, but also to identify which variable have nonlinear effects based on AFT models. We first embed the partial linear AFT model into a nonparametric additive AFT model as existing methods. Our method is different from those exiting methods in the way of decomposing the additive component $g(X)$ to reflect the effect of covariate X. In the existing method (e.g. Zhang et al. 2011, Huang et al. 2012), $g(X)$ is assumed to be the sum of linear and nonlinear components, i.e.

$$g(X) = \beta_0 + \beta_1 X + \phi(X), \tag{1}$$

where $\phi(x)$ is nonparametric function which can be approximated by a linear combination of some basis functions, i.e. $\phi(x) = \sum_{k=1}^{q_n} \theta_k \boldsymbol{\psi}_k(x)$. Then the estimation of $g(X)$ is transformed into estimation of β_0, β_1 and $\theta_k (k = 1, \ldots, q_n)$. To ensure the parameter identifiability, additional assumptions is needed for the basis functions (such as the linear function can not be included as basis function). Instead of assuming $g(X)$ to be the sum of linear and nonlinear components directly as did in many existing researches, we take the linear effect out, i.e. write the covariate effect as $g(X) = X\phi(X)$. Then the important variables can be selected out if and only if $\phi(\cdot) \not\equiv 0$, and the linearity of $g(X)$ can be also identified according to whether the derivative of $\phi(\cdot)$ to be zero or not. We adopt B-spline expansion techniques to approximate the nonparametric function $\phi(\cdot)$ and no additional assumptions are needed to the basis functions. We then simultaneously determine the linear and nonlinear components and select important variables with a double penalized approach. We show that, under the assumption of non-informative censoring and other suitable conditions, the proposed approach is both model pursuit and variable selection consistent, meaning that it can correctly identify the linear and nonlinear components and select important covariates with high probability. We also show that the proposed estimators enjoy an asymptotic oracle property.

The merits of our approach mainly concentrate on the following points. Firstly, the model is identifiable under a mild condition. Secondly, our approach is easy-to-implement. We apply the alternative direction method of multipliers (ADMM) algorithm (Boyd et al. 2011) to overcome the computing difficulties caused by the inseparability of unknown parameters. Simulation studies show that the proposed method can select the true model with high probability, while linear and nonlinear discoverer (LAND) suggested by Zhang et al. (2011) identifies sparser models. Finally, as the considerable difficulties brought by the high dimensional data, we restrict our researches on the case that the covariate dimension diverses with sample size, i.e., $d_n = O(n^{1/4})$, and draw the theoretical conclusions of the model pursuit and variable selection consistency. By combining the SIS techniques (e.g. Zhang et al. 2018), the two-step selection procedure adopted in Neykov et al. (2014) and Ma et al. (2006) makes it feasible analyzing the ultra-high dimensional data, i.e., $d_n = o(\exp(n^{\alpha}))$ with some $0 < \alpha < 1$.

The remainder of the paper is organized as follows. In Sect. 2 we propose a double-penalized procedure for simultaneously selecting variables and model structure. ADMM algorithm for computation is presented in Sect. 3. We evaluate the performance of proposed procedure through simulations studies in Sect. 4. In Sect. 5 we apply the proposed procedure to analyze the real data set. Some remarks are provided in Sect. 6. The proofs of the theoretical results are relegated to the Appendix.

2 Estimation procedures

Suppose that there are n i.i.d. observations in the study. Let $T_i, i = 1, \ldots, n$ be the logarithm of the survival time following an additive AFT model

$$T_i = \sum_{j \in M_1} \tilde{g}_{0j}(X_{ij}) + \sum_{j \in M_2} \beta_{0j} X_{ij} + \sum_{j \in M_3} 0 \cdot X_{ij} + \varepsilon_i, \tag{2}$$

where $\tilde{g}_{0j}$ is a nonzero non-linear function and X_{ij} is the observation for the j-th covariate in the i-th subject. The effects of the covariates to the survival time in model (2) are split into three types: nonzero non-linear effect, nonzero linear effect and null effect, whose index sets correspond to M_1, M_2 and M_3 respectively. We suppose that the density function of the covariate $\boldsymbol{X} = (X_1, \ldots, X_{d_n})^T$ has a positive support on $[\alpha_1, \alpha_2]^{d_n}$, where d_n is the dimension of $\boldsymbol{X}$ being allowed to increase with sample size n and α_1 and α_2 are two finite real numbers. Let $C_i (i = 1, \ldots, n)$ be the logarithm of the censoring time for subject i. With censoring, one observes $Y_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \leq C_i)$, where $I(\cdot)$ is the indicator function. The observed data consists of $\{(Y_i, \delta_i, \boldsymbol{X}_i) : i = 1, 2, \ldots, n\}$. However, it is usually unknown in advance to which part the covariates belong to for real data. To specify the structure of the covariate effects and select important variables, or equivalently, to identify the index sets M_1, M_2 and M_3, we note that Eq (2) is a special case that

$$T_i = \sum_{j=1}^{d_n} g_{0j}(X_{ij}) + \varepsilon_i \tag{3}$$

with $g_{0j}(x_j) = \beta_{0j} x_j + \tilde{g}_{0j}(x_j)$. Thus, we define in model (3) that $g_{0j}(x_j) = x_j \phi_{0j}(x_j)$, and suppose that $\tilde{g}_{0j}(x_j) = x_j \tilde{\phi}_{0j}(x_j)$ for $j = 1, \ldots, d_n$, where ϕ_{0j} is an unknown function with the convention that $0/0 = 1$ and $\tilde{\phi}_{0j}$ is a non-constant function. This kind of model is an economic way for modeling the structure of the covariate effects in the context of partially linear models. It is also a convenient way to achieve the goal of model structure identification and variable selection simultaneously. In fact, the $j-$th covariate X_j is an unimportant variable if and only if the nonparametric part ϕ_{0j} takes 0. Otherwise, the corresponding covariate X_j has a linear effect or nonlinear effect according to the derivative of ϕ_{0j} takes 0 or not. Define $\mathcal{H} = L^2[\alpha_1, \alpha_2]$, $\tilde{\mathcal{H}} = \{\tilde{\phi}_j : \int_{\alpha_1}^{\alpha_2} \tilde{\phi}_j(x)dx = 0, \tilde{\phi}_j \in \mathcal{H}\}$ for some constant C,

$\mathcal{H}^{d_n} = \underbrace{\mathcal{H} \times \ldots \times \mathcal{H}}_{d_n}$ and $\tilde{\mathcal{H}}^{d_n} = \underbrace{\tilde{\mathcal{H}} \times \ldots \times \tilde{\mathcal{H}}}_{d_n}$. This definition shows that $\tilde{\phi}_j$ is either a non-constant function or zero-valued function. The identifiability of model (3) is guaranteed by the following proposition.

Proposition 1 (Identifiability) *We suppose that the covariates are not linearly dependent. For d_n-dimensional function vector $(\phi_1, \ldots, \phi_{d_n})' \in \mathcal{H}^{d_n}$, there exists unique $(\beta_1, \ldots, \beta_{d_n})' \in \mathbb{R}^{d_n}$ and $(\tilde{\phi}_1, \ldots, \tilde{\phi}_{d_n})' \in \tilde{\mathcal{H}}^{d_n}$ such that*

$$\sum_{j=1}^{d_n} x_j \phi_j(x_j) = \sum_{j=1}^{d_n} \beta_j x_j + \sum_{j=1}^{d_n} x_j \tilde{\phi}_j(x_j) \tag{4}$$

with $E(X_j \phi_j(X_j)) = E(\beta_j X_j + X_j \tilde{\phi}_j(X_j))$ for each $j = 1, \ldots, d_n$.

The proposition shows that $\sum_{j=1}^{d_n} g_{0j}(X_j)$ in model (3) can be uniquely written as a sum of non-linear component and linear component as

$$\sum_{j=1}^{d_n} g_{0j}(X_j) = \sum_{j \in M_1} X_j \tilde{\phi}_{0j}(X_j) + \sum_{j \in M_2} X_j \beta_{0j}. \tag{5}$$

In the following, we further assume that $E(\delta_i g_{0j}(X_{ij})) = 0$. As pointed by Huang (1999), centering $E(\delta_i g_{0j}(X_{ij}))$ instead of $E(g_{0j}(X_{ij}))$ simplifies information calculation and asymptotic analysis. For simplicity, we write $g_0(x) = \sum_{j=1}^{d_n} g_{0j}(x)$, $\boldsymbol{\phi}_0 = (\phi_{0j}, j = 1, \ldots, d_n)^T$, $\tilde{\boldsymbol{\phi}}_0 = (\tilde{\phi}_{0j}, j \in M_1)^T$ and $\boldsymbol{\beta}_0 = (\beta_{0j}, j \in M_2)^T$.

We then establish a regularization procedure by constructing a penalized loss function. Let k be a nonnegative integer, and some $\alpha \in (0, 1]$ such that $p = k + \alpha > 0.5$. Define $\phi_j^{(k)}$ as the $k-$th derivative of function ϕ_j and let

$$\mathcal{G} = \{\phi_j : |\phi_j(x_1) - \phi_j(x_2)| \le C|x_1 - x_2|^{\alpha}, x_1, x_2 \in [\alpha_1, \alpha_2], \phi_j^{(k)} \in L^2[\alpha_1, \alpha_2]\} \subset \mathcal{H}.$$

We suppose that the first derivative ϕ_{0j}' exists and $\boldsymbol{\phi}_0 \in \mathcal{G}^{d_n}$.

As a start, we use the B-splines to approximate the unknown nonparametric function $\phi_{0j}, j = 1, \ldots, d_n$ in (3). The interval $[\alpha_1, \alpha_2]$ is split into K_n subintervals $I_{K_n t} = [\xi_t, \xi_{t+1}), \ t = 0, 1, \ldots, K_n - 2$ and $I_{K_n K_n - 1} = [\xi_{K_n - 1}, \xi_{K_n}]$, where $\alpha_1 = \xi_0 < \xi_1 < \cdots < \xi_{K_n} = \alpha_2$ and $K_n = O(n^{\nu})$ with $0 < \nu < 0.5$ being a positive integer such that $\max_{1 \le j \le K_n} |\xi_j - \xi_{j-1}| = O(n^{-\nu})$. Let Ω_n be the space of polynomial splines of order $m \ge 1$ consisting of functions h's, where h is a polynomial of order m on interval $I_{K_n t}$ for $t = 0, 1, \ldots, K_n - 1$, and h is m_2 times continuously differentiable on $[\alpha_1, \alpha_2]$ for $m \ge 2$ and $0 \le m_2 \le m - 2$. According to Schumaker (1981), there exists a local basis $\{\psi_{k,m}, \ k = 1, 2, \ldots, q_n\}$ with the basis number $q_n = K_n + m$ for Ω_n. Thus, the function $\phi_{nj}(\cdot)$ in Ω_n can be approximated by a basis expansion as

$$\phi_{nj}(x) = \boldsymbol{\theta}_j^T \boldsymbol{\psi}_{q_n, m}(x), \ \ j = 1, 2, \ldots, d_n. \tag{6}$$

where $\boldsymbol{\theta}_j = (\theta_{j1}, \ldots, \theta_{jq_n})^T$ and $\boldsymbol{\psi}_{q_n,m}(\cdot) = (\psi_{1,m}(\cdot), \ldots, \psi_{q_n,m}(\cdot))^T$.

In the sequel, we assume that censoring is completely non-informative. Let w_i $(i = 1, \ldots, n)$ be the Kaplan-Meier weight computed as $\omega_1 = \frac{\delta_{(1)}}{n}$ and $\omega_i = \frac{\delta_{(i)}}{n-i+1} \prod_{j=1}^{i-1} \left(\frac{n-j}{n-j+1} \right)^{\delta_{(j)}}$ for $i = 2, \ldots, n$, where $Y_{(1)}, \ldots, Y_{(n)}$ are the order statistics of Y_i's and $\delta_{(1)}, \ldots, \delta_{(n)}$ are the associated censoring indicators. Similarly, let $X_{(1)}, \ldots, X_{(n)}$ be the associated covariates of the ordered Y_i's. Inspired by Stute (1993), we introduce the weighted least square loss function

$$\ell_n(\boldsymbol{\theta}) = \frac{1}{2} \sum_{i=1}^{n} w_i \left[Y_{(i)} - \sum_{j=1}^{d_n} X_{(i)j} \sum_{k=1}^{q_n} \theta_{jk} \psi_{k,m}(X_{(i)j}) \right]^2,$$

for $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{d_n})^T$ with $\boldsymbol{\theta}_j$'s are defined as below (6). Define

$$\overline{g}_{jw}(\phi_j, X_j) = \frac{\sum_{i=1}^{n} w_i X_{(i)j} \phi_j(X_{(i)j})}{\sum_{i=1}^{n} w_i}, \quad \overline{Y}_w = \frac{\sum_{i=1}^{n} w_i Y_{(i)}}{\sum_{i=1}^{n} w_i},$$

$Y^* = (Y^*_{(1)}, \ldots, Y^*_{(n)})^T$ and $g^*_j = (g^*_{(1)j}, \ldots, g^*_{(n)j})^T$, where

$$g^*_{(i)j}(\phi_j, X_j) = (nw_i)^{1/2}(X_{(i)j}\phi_j(X_{(i)j}) - \overline{g}_{jw}(\phi_j, X_j)), \quad Y^*_{(i)} = (nw_i)^{1/2}(Y_{(i)} - \overline{Y}_w).$$

Then the weighted least squares loss function is equivalent to

$$\ell_n(\boldsymbol{\theta}) = \frac{1}{2n} \left\| Y^* - \sum_{j=1}^{d_n} g^*_j(\phi_{nj}, X_j) \right\|^2,$$

where $\| \cdot \|$ is L_2 norm.

We then impose two penalty functions to the loss function to solve a regulariztion problem

$$\min_{\boldsymbol{\theta}} Q_n(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} \ell_n(\boldsymbol{\theta}) + P^1_{\lambda_1}(\boldsymbol{\theta}) + P^2_{\lambda_2}(\boldsymbol{\theta}),$$

where the penalty $P^1_{\lambda_1}(\boldsymbol{\theta})$ aims to specify the model structure and $P^2_{\lambda_2}(\boldsymbol{\theta})$ is for variable selection with tuning parameters λ_1 and λ_2. It is convenient to take $P^2_{\lambda_2}(\boldsymbol{\theta})$ as $\sum_{j=1}^{d_n} P_2(\|\boldsymbol{\theta}_j\|; \lambda_2)$ with $P_2(\cdot)$ being a penalty function. To decide $P^1_{\lambda_1}(\boldsymbol{\theta})$, we note by de Boor (1978) that

$$\phi'_{nj}(x_j) = (\boldsymbol{C}_{\boldsymbol{\xi}} \boldsymbol{\theta}_j)^T \boldsymbol{\psi}_{q_{n-1},m-1}(x_j),$$

where $\boldsymbol{C}_{\boldsymbol{\xi}}$ is defined as in Appendix. The problem of structure identification is then transformed into that of group selection. As a result, we can take

$$P_{\lambda_1}^1(\boldsymbol{\theta}) = \sum_{j=1}^{d_n} P_1(\|\boldsymbol{C}_{\boldsymbol{\xi}}\boldsymbol{\theta}_j\|; \lambda_1)$$

with $P_1(\cdot)$ being a penalty function. Thus, we obtain $\widehat{\boldsymbol{\theta}}_n$, the estimator of $\boldsymbol{\theta}$, by minimizing the following objective function for suitable selected tuning parameters $\lambda_1,\ \lambda_2 \geq 0$,

$$Q_n(\boldsymbol{\theta}) = \ell_n(\boldsymbol{\theta}) + \sum_{j=1}^{d_n} P_1(\|\boldsymbol{C}_{\boldsymbol{\xi}}\boldsymbol{\theta}_j\|; \lambda_1) + \sum_{j=1}^{d_n} P_2(\|\boldsymbol{\theta}_j\|; \lambda_2). \tag{7}$$

Remark 1 Instead of using RKHS norm in penalty function as in Zhang et al. (2011), we transform the regularization problem into double group selection procedure through approximating the nonparametric part by spline expansions, which makes the calculation implement easier.

In our implement, we consider the following penalties:

(1) group SCAD penalty ($\gamma = 3.7$)

$$P(\|\boldsymbol{\theta}_j\|; \lambda) = \begin{cases} \lambda\|\boldsymbol{\theta}_j\|, & \|\boldsymbol{\theta}_j\| \leq \lambda, \\ \frac{2\gamma\lambda\|\boldsymbol{\theta}_j\| - \|\boldsymbol{\theta}_j\|^2 - \lambda^2}{2(\gamma-1)}, & \lambda < \|\boldsymbol{\theta}_j\| \leq \gamma\lambda, \\ (\gamma^2-1)\lambda^2/(2(\gamma-1)), & \|\boldsymbol{\theta}_j\| > \gamma\lambda. \end{cases}$$

(2) group MCP penalty ($\gamma = \frac{2}{1-\max\limits_{i\neq j} x_i^T x_j/n}$)

$$P(\|\boldsymbol{\theta}_j\|; \lambda) = \begin{cases} \lambda\|\boldsymbol{\theta}_j\| - \frac{\|\boldsymbol{\theta}_j\|^2}{2\gamma}, & \|\boldsymbol{\theta}_j\| \leq \gamma\lambda, \\ \frac{\lambda 2\gamma}{2}, & \|\boldsymbol{\theta}_j\| > \gamma\lambda. \end{cases}$$

3 Computational algorithm

3.1 ADMM algorithm

Noting that the regularization problem is equivalent to the following constrained optimization problem:

$$\begin{aligned} \min\ & \ell_n(\boldsymbol{\theta}) + \sum_{j=1}^{d_n} P_1(\|\boldsymbol{\kappa}_j\|; \lambda_{1n}) + \sum_{j=1}^{d_n} P_2(\|\boldsymbol{\tau}_j\|; \lambda_2), \\ \text{subject to}\quad & \boldsymbol{\kappa}_j - \boldsymbol{C}_{\boldsymbol{\xi}}\boldsymbol{\theta}_j = 0, \\ & \boldsymbol{\tau}_j - \boldsymbol{\theta}_j = 0 \quad \text{for any} \quad j = 1, \ldots, d_n, \end{aligned}$$

ADMM algorithm is a natural way to easy the calculation complexity. In the ADMM algorithm, the original regularization problem is transformed to minimize the following augmented Lagrange function with respect to $(\boldsymbol{\theta}, \boldsymbol{\kappa}, \boldsymbol{\tau})$ for given $(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2)$.

$$L_\varrho(\boldsymbol{\theta}, \boldsymbol{\kappa}, \boldsymbol{\tau}; \boldsymbol{\pi}_1, \boldsymbol{\pi}_2) = f_1(\boldsymbol{\theta}; \boldsymbol{\kappa}, \boldsymbol{\tau}, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2) + \sum_{j=1}^{d_n} P_1(\|\boldsymbol{\kappa}_j\|; \lambda_1) + \sum_{j=1}^{d_n} P_2(\|\boldsymbol{\tau}_j\|; \lambda_2),$$

where ϱ is a given constant and

$$\begin{aligned} f_1(\boldsymbol{\theta}; \boldsymbol{\kappa}, \boldsymbol{\tau}, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2) = \ell_n(\boldsymbol{\theta}) + \textstyle\sum_{j=1}^{d_n} \frac{\varrho}{2}\|\boldsymbol{\kappa}_j - \boldsymbol{C}_{\boldsymbol{\xi}}\boldsymbol{\theta}_j\|^2 + \sum_{j=1}^{d_n} \boldsymbol{\pi}_{1j}^T(\boldsymbol{\kappa}_j - \boldsymbol{C}_{\boldsymbol{\xi}}\boldsymbol{\theta}_j) \\ + \sum_{j=1}^{d_n} \frac{\varrho}{2}\|\boldsymbol{\tau}_j - \boldsymbol{\theta}_j\|^2 + \sum_{j=1}^{d_n} \boldsymbol{\pi}_{2j}^T(\boldsymbol{\tau}_j - \boldsymbol{\theta}_j). \end{aligned}$$

The ADMM algorithm is described as follows.

Step 1 Initialize $\boldsymbol{\theta}^{(0)}$, $\boldsymbol{\kappa}^{(0)}$, $\boldsymbol{\tau}^{(0)}$, $\boldsymbol{\pi}_1^{(0)}$, and $\boldsymbol{\pi}_2^{(0)}$.

Step 2 Compute the parameter values at the $(k+1)$th step. For $j = 1, \ldots, d_n$,

(i) take the ridge solution as the initial point and use Newton-Raphson algorithm to solve the following optimization problem.

$$\boldsymbol{\theta}_j^{(k+1)} = \underset{\boldsymbol{\theta}_j}{\operatorname{argmin}}\, f_1(\boldsymbol{\theta}; \boldsymbol{\kappa}^{(k)}, \boldsymbol{\tau}^{(k)}, \boldsymbol{\pi}_1^{(k)}, \boldsymbol{\pi}_2^{(k)}).$$

(ii) update $\boldsymbol{\kappa}_j^{(k+1)}$ and $\boldsymbol{\tau}_j^{(k+1)}$ by

$$\begin{aligned} \boldsymbol{\kappa}_j^{(k+1)} &= \underset{\boldsymbol{\kappa}_j}{\operatorname{argmin}} \left\{ P_1(\|\boldsymbol{\kappa}_j\|; \lambda_1) + \frac{\varrho}{2}\|\boldsymbol{\kappa}_j - \boldsymbol{C}_{\boldsymbol{\xi}}\boldsymbol{\theta}_j^{(k+1)} + \frac{1}{\varrho}\boldsymbol{\pi}_{1j}^{(k)}\|^2 \right\}, \\ \boldsymbol{\tau}_j^{(k+1)} &= \underset{\boldsymbol{\tau}_j}{\operatorname{argmin}} \left\{ P_2(\|\boldsymbol{\tau}_j\|; \lambda_2) + \frac{\varrho}{2}\|\boldsymbol{\tau}_j - \boldsymbol{\theta}_j^{(k+1)} + \frac{1}{\varrho}\boldsymbol{\pi}_{2j}^{(k)}\|^2 \right\}. \end{aligned}$$

(iii) update $\boldsymbol{\pi}_{1j}^{(k+1)}$ and $\boldsymbol{\pi}_{2j}^{(k+1)}$ with

$$\begin{aligned} \boldsymbol{\pi}_{1j}^{(k+1)} &= \boldsymbol{\pi}_{1j}^{(k)} + \varrho(\boldsymbol{\kappa}_j^{(k+1)} - \boldsymbol{C}_{\boldsymbol{\xi}}\boldsymbol{\theta}_j^{(k+1)}), \\ \boldsymbol{\pi}_{2j}^{(k+1)} &= \boldsymbol{\pi}_{2j}^{(k)} + \varrho(\boldsymbol{\tau}_j^{(k+1)} - \boldsymbol{\theta}_j^{(k+1)}). \end{aligned}$$

Step 3 Repeat Step 2 until $\|\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{(k)}\|_\infty$ is small enough, where $\|\cdot\|_\infty$ is the supremum norm.

This ADMM algorithm is efficient and easily implemented. In fact, for some widely used penalties such as the group MCP and the group SCAD, $\boldsymbol{\kappa}_j^{(k+1)}$ and $\boldsymbol{\tau}_j^{(k+1)}$ have

closed forms. We list them as follows with the notations $\boldsymbol{d_j} = \boldsymbol{C_\xi}\boldsymbol{\theta}_j^{(k+1)} - \boldsymbol{\pi}_{1j}^{(k)}/\varrho$, $S(\boldsymbol{d_j}; \lambda_1) = (1 - \lambda_1/\|\boldsymbol{d_j}\|)_+\boldsymbol{d_j}$ and $\boldsymbol{e_j} = \boldsymbol{\theta}_j^{(k+1)} - \boldsymbol{\pi}_{2j}^{(k)}/\varrho$.

(1) For the group SCAD penalty,

$$\boldsymbol{\kappa}_j^{(k+1)} = \begin{cases} S(\boldsymbol{d_j}; \lambda_1/\varrho), & \|\boldsymbol{d_j}\| \le \lambda_1 + \lambda_1/\varrho, \\ \frac{\left(\varrho(\gamma-1) - \gamma\lambda_1/\|\boldsymbol{d_j}\|\right)\boldsymbol{d_j}}{\varrho\gamma - \varrho - 1}, & \lambda_1 + \lambda_1/\varrho < \|\boldsymbol{d_j}\| \le \gamma\lambda_1, \\ \boldsymbol{d_j}, & \|\boldsymbol{d_j}\| > \gamma\lambda_1. \end{cases}$$

(2) For the group MCP penalty,

$$\boldsymbol{\kappa}_j^{(k+1)} = \begin{cases} S(\frac{\varrho\boldsymbol{d_j}}{\varrho - 1/\gamma}; \frac{\lambda_1}{\varrho - 1/\gamma}), & \|\boldsymbol{d_j}\| \le \gamma\lambda_1, \\ \boldsymbol{d_j}, & \|\boldsymbol{d_j}\| > \gamma\lambda_1. \end{cases}$$

For both penalties, $\boldsymbol{\tau}_j^{(k+1)}$ can be updated using the same closed form as $\boldsymbol{\kappa}_j^{(k+1)}$ except replacing $\boldsymbol{d_j}$ by $\boldsymbol{e_j}$.

3.2 Tuning parameter selection

The selection of the tuning parameters has great influence on the performance of the algorithm. Commonly used selection criteria include AIC (Akaike 1973), BIC (Schwartz 1978) and GCV (Craven and Wahba 1979). As our proposed method has double penalty functions, we adopt the generalized cross validation (GCV) criterion combining the ideas suggested by Robert and Gray (1992) to calculate the degree of freedom. The GCV value is defined as

$$\mathrm{GCV}(\lambda_1, \lambda_2) = \frac{\ell_n(\boldsymbol{\theta})}{\{1 - d(\lambda_1, \lambda_2)/n\}^2}, \tag{8}$$

where the degree of freedom $d(\lambda_1, \lambda_2)$ includes two tuning parameters. By Robert and Gray (1992), $d(\lambda_1, \lambda_2)$ is taken as

$$d(\lambda_1, \lambda_2) = (q_n - 2)\sum_{j=1}^{d_n} \|\boldsymbol{C_\xi}\widehat{\boldsymbol{\theta}}_{nj}\|_0 + \sum_{j=1}^{d_n} \|\widehat{\boldsymbol{\theta}}_{nj}\|_0,$$

where $\|\cdot\|_0$ means 0−norm. And the optimal value of (λ_1, λ_2) is given by

$$(\widehat{\lambda}_1, \widehat{\lambda}_2) = \underset{(\lambda_1, \lambda_2)}{\operatorname{argmin}}\, \mathrm{GCV}(\lambda_1, \lambda_2).$$

4 Asymptotic properties

To describe the theoretical results, we first introduce some notations. Let H and G be the distribution functions of Y and C. Define τ_Y, τ_T and τ_C to be the end points of the

support of Y, T and C. Denote

$$\tilde{F}^0(\boldsymbol{x}, t) = \begin{cases} F^0(\boldsymbol{x}, t), & t < \tau_Y, \\ F^0(\boldsymbol{x}, \tau_Y -) + F^0(\boldsymbol{x}, \{\tau_Y\}) I(\tau_Y \in A), & t \geq \tau_Y, \end{cases}$$

where F^0 is the joint distribution of $(\boldsymbol{X}, T)$ and A represents the set of atoms of H. We use the notation X_M to represent a vector or a matrix consisting of $(X_j, j \in M)$ with M being an index set. Let $\Lambda_{\min}(\cdot)$ and $\Lambda_{\max}(\cdot)$ be the minimal and maximal eigenvalues of a matrix respectively.

The following conditions are required technically in the process of theoretical derivation.

(C1) The covariate $\boldsymbol{X}$ is bounded and the true regression parameter $\boldsymbol{\beta}_0$ belongs to an open subset of $\mathbb{R}^{|M_2|}$.

(C2) $E(\varepsilon|\boldsymbol{X}) = 0$ and $E(T^2) < \infty$;

(C3) T and C are conditional independent given $\boldsymbol{X}$ and $P(T \leq C|T, \boldsymbol{X}) = P(T \leq C|T)$;

(C4) There exists a small positive constant ϵ such that $P(\delta = 1|\boldsymbol{X}) > \epsilon$ and $P(\tau_C > \tau_T|\boldsymbol{X}) > \epsilon$ almost surely with respect to the probability measure of $\boldsymbol{X}$;

(C5) For all $(t, \boldsymbol{x}) \in [0, \tau_T] \times [\alpha_1, \alpha_2]^{d_n}$, the joint density $f(t, \boldsymbol{x}, \delta)$ of $(T, \boldsymbol{X}, \delta = 1)$ satisfies $c_1 \leq f(t, \boldsymbol{x}, \delta) \leq c_2$, where $0 < c_1 < c_2 < \infty$ are two constants;

(C6) Let $\Sigma_i = E[\boldsymbol{X}_{M_i}^T \boldsymbol{X}_{M_i}]$ and define $\rho_i = \Lambda_{\min}(\Sigma_i)$ for $i = 1, 2$. There exists a positive constant $r > 0$ such that $0 < r < \min\{\rho_1, \rho_2\} \leq \max\{\rho_1, \rho_2\} = O(1)$;

(C7) Denote that $\rho_n^* = \Lambda_{\max}(E(\boldsymbol{X}^T \boldsymbol{X}))$. We suppose that the concave penalty functions $P_1(\|\boldsymbol{\theta}\|; \lambda_1)$ and $P_2(\|\boldsymbol{\theta}\|; \lambda_2)$ satisfy $-P_1^{''}(\|\boldsymbol{\theta}\|; \lambda_1) \geq c\lambda_1^a \|\boldsymbol{\theta}\|^{-b}$ and $-P_2^{''}(\|\boldsymbol{\theta}\|; \lambda_1) \geq c\lambda_2^a \|\boldsymbol{\theta}\|^{-b}$ near original point for some given constants $a \in \mathbb{R}$, $0 \leq b \leq 2$ and $c > 0$, $P_1'(0+; \lambda_1) = O(\lambda_1)$, $P_2'(0+; \lambda_1) = O(\lambda_2)$, and

$$\frac{\lambda_1^a}{\rho_n^*(d_n^2(n^{-(1-\nu)} + n^{-2\nu p}))^{b/2}} \to \infty, \quad \frac{\lambda_2^a}{\rho_n^*(d_n^2(n^{-(1-\nu)} + n^{-2\nu p}))^{b/2}} \to \infty.$$

(C8) Let $e(y, \boldsymbol{x}) = y - g_0(\boldsymbol{x})$ for $\boldsymbol{x} = (x_1, \ldots, x_{d_n})$. Suppose that $E[\delta e^2(Y, \boldsymbol{X}) \boldsymbol{X}_{M_2}^T \boldsymbol{X}_{M_2}] < \infty$ and $\int |e(y, \boldsymbol{x}) x_j| \, [C(y)]^{1/2} \tilde{F}^0(d\boldsymbol{x}, dy) < \infty$ for $j \in M_2$, where $C(y) = \int_0^{y-} [(1 - H(z))(1 - G(z))]^{-1} G(dz)$.

Condition (C1) gives the support set of the true parameter. (C2) allows the distribution of ε to depend on convariates and allows the heteroscedastic error terms, which is weaker than Buckley–James method (Buckley and James 1979) and the rank based method (Jin et al., 2003). (C3) shows that the censoring indicator δ is conditionally independent of the covariate $\boldsymbol{X}$ given the failure time T and the censoring time is considered noninformative, which is the same as that for the Kanplan-Meier estimtor. (C4) and (C5) are regularity conditions for additive survival models. (C6) and (C8) are quite mild for theoretical justification of the Stute estimator. It deserves to note that when the dimension of non-zero effect covariate is finite, Condition (C6) holds trivially if $\{X_j, j \in M_1 \cup M_2\}$ is linearly uncorrelated. Condition (C7) determines the order of tuning parameters λ_1 and λ_2. In addition, (C7) makes the assumption

on the maximum eigenvalue of the correlation matrix $E(\boldsymbol{X}^T\boldsymbol{X})$. This condition has been used in many literatures (e.g. Joseph 2013) to ensure some near-orthogonality between important and non-important covariates. The requirement for penalties in (C7) is satisfied by most general used penalty functions, such as the group SCAD and group MCP penalty.

For example, for the group SCAD penalty $P(\|\boldsymbol{\theta}\|;\lambda_n)$,

$$-\|\boldsymbol{\theta}\|^b P^{''}(\|\boldsymbol{\theta}\|;\lambda_n) = \frac{\|\boldsymbol{\theta}\|^b}{\gamma-1} I_{\{\lambda_n<\|\boldsymbol{\theta}\|\le\gamma\lambda_1\}} > \frac{\lambda_n^b}{\gamma-1}.$$

Then there exists constant c such that $-P^{''}(\|\boldsymbol{\theta}\|;\lambda_n) \ge c\lambda_n^a\|\boldsymbol{\theta}\|^{-b}$ if $\lambda_n^{b-a}\to\infty$; for the group MCP penalty $P(\|\boldsymbol{\theta}\|;\lambda_n)$, we have

$$n - \|\boldsymbol{\theta}\|^b P^{''}(\|\boldsymbol{\theta}\|;\lambda_n) = \frac{\|\boldsymbol{\theta}\|^b}{\gamma} I_{\{\|\boldsymbol{\theta}\|\le\gamma\lambda_n\}}.$$

So there exists constant c such that $-P^{''}(\|\boldsymbol{\theta}\|;\lambda_n) \ge c\lambda_n^a\|\boldsymbol{\theta}\|^{-b}$ as long as $\gamma\lambda_n^a\to 0$ and $\gamma\lambda_n\to\infty$.

Define $\widehat{\phi}_{nj}(x_j) = \widehat{\boldsymbol{\theta}}_{nj}^T\boldsymbol{\psi}_{q_n,m}(x_j)$ to be the estimator of $\phi_{0j}(x_j)$, and we write the corresponding estimators for $\tilde{\phi}_{0j}$ and non-zero valued β_{0j} in (5) as $\widehat{\boldsymbol{\phi}}_n = (\widehat{\phi}_{nj}, j\in\widehat{M_1})^T$ and $\widehat{\boldsymbol{\beta}}_n = (\widehat{\phi}_{nj}, j\in\widehat{M_2})^T$ respectively. In the main results, we consider the case that $d_n^4/n\to 0$. Recalling that $p>0.5$ is a smoothness parameter for ϕ_{0j}, Theorem 1 presents the convergence rate of the estimators $\widehat{\boldsymbol{\beta}}_n$ and $\widehat{\boldsymbol{\phi}}_n$.

Theorem 1 *Suppose Conditions (C1)–(C7) hold. If the tuning parameters* $\lambda_1 = o(d_n n^{-\nu})$ *and* $\lambda_2 = o(d_n n^{-\nu})$ *with* $0.25/p < \nu < 0.5$, *then* $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|^2 = O_p(d_n^2(n^{-(1-\nu)} + n^{-2\nu p}))$ *and* $\|\widehat{\boldsymbol{\phi}}_n - \tilde{\boldsymbol{\phi}}_0\|^2 = O_p(d_n^2(n^{-(1-\nu)} + n^{-2\nu p}))$.

Theorem 2 *Suppose Conditions (C1)–(C8) hold. If the tuning parameters satisfy that* $\lambda_1 = o(d_n n^{-\nu})$ *and* $\lambda_2 = o(d_n n^{-\nu})$, *then for* $0.25/p < \nu < 0.5$ *and* $\nu(p+q) > 0.5$,

(i) $\lim_{n\to\infty} P(\widehat{\boldsymbol{\theta}}_{nj} = 0 : j\in M_3) = 1$ *and* $\lim_{n\to\infty} P(\boldsymbol{C}_{\boldsymbol{\xi}}\widehat{\boldsymbol{\theta}}_{nj} = 0 : j\in M_2) = 1$;

(ii) For any $\boldsymbol{u}\in\mathbb{R}^{|M_2|}$ *with* $\|\boldsymbol{u}\|_2 = 1$, $\sqrt{n}\boldsymbol{u}^T\Sigma^{-1/2}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)\xrightarrow{d} N(0,1)$, *where* $\Sigma = \Sigma_2^{-1}\Sigma_3\Sigma_2^{-1}$, $\Sigma_3 = Var(\delta\gamma_0(Y)(Y-g_0(\boldsymbol{X}))\boldsymbol{X}_{M_2} + (1-\delta)\gamma_1(Y) - \gamma_2(Y))$ *with* $\gamma_i(y)$*'s* $i = 0,1,2$ *defined as in Appendix.*

Theorem 2 shows that the proposed estimators enjoy an asymptotic oracle property. Specifically, result (i) implies that the estimators are consistent in terms of variable selection and structure identification, i.e., they can select important variables and identify the important non-parametric components simultaneously with high probability; result (ii) states that the estimator of regression coefficient for important linear variables is asymptotically distributed as normal with mean zero and variance-covariance matrix as discribed in the theorem.

5 Simulation studies

In this section, we conduct some simulation studies to evaluate the performance of our proposed method by using different penalties including group SCAD and group MCP. The parameter γ is set to 3.7 in group SCAD and $\gamma = 2(1 - \max_{i \neq j} x_i^T x_j/n)^{-1}$ in group MCP according to Fan and Li (2001) and Zhang (2010) , respectively. The tuning parameters λ_1 and λ_2 are selected using the GCV criterion given by (9). The sample size is taken as $n = 400$ and all the simulation results are based on 100 replications using R software. We compare the proposed method with two competeing approaches in terms of the model selection accuracy as well as estimation accuracy. The first one is the LAND method by Zhang et al. (2011) and the second is the oracle estimate which is obtained by assuming the true model structure is known.

To evaluate the estimation accuracy of the estimator $\widehat{\mu}$, we report the relative model error defined by Fan and Li (2002). We first define the model error of the estimator $\widehat{\mu}(\boldsymbol{x})$ as

$$\mathrm{ME}(\widehat{\mu}) = E\{\widehat{\mu}(\boldsymbol{x}) - \mu(\boldsymbol{x})\}^2 = E\{\exp(\widehat{g}(\boldsymbol{x})) - \exp(g_0(\boldsymbol{x}))\}^2,$$

where $\mu(\boldsymbol{x}) = E(Y|\boldsymbol{x})$ and $\widehat{g}$ is the proposed estimator of g_0. The relative model error (RME) of the proposed model to the linear AFT model is defined as

$$\mathrm{RME}(\widehat{\mu}, \widetilde{\mu}) = \frac{\mathrm{ME}(\widehat{\mu})}{\mathrm{ME}(\widetilde{\mu})},$$

where $\mathrm{ME}(\widetilde{\mu})$ represents the model error under the linear AFT model. We present the median of the relative model error (MRME) over 100 replications. To evaluate performance of the compared methods in structure selection, we report the true positive rate that an important variable is correctly selected.

The survival time is generated from AFT model (3). The covariates are generated as follows. We first generate the covariates $X_1, \ldots, X_p$ from an AR(1) structure model with $X_1 \sim N(0, 1)$ and $\mathrm{Cov}(X_{j_1}, X_{j_2}) = 0.4^{|j_1 - j_2|}$ for $j_1, \ j_2 = 1, \ldots, p$, and then we trim $\boldsymbol{X}$ to the range $[-1, 1]$. The error ε is distributed by $N(0, 1)$. The censoring times are independently generated from the uniform distribution $U[0, c]$, where c is chosen to yield about 20% censoring rate. We use cubic splines to select the important variables and to identify the structure.

Example 1 We set $p = 15$ and g_0 takes the form

$$g_0(\boldsymbol{x}) = 2x_1 - 2x_2 + 2f_1(x_3),$$

where $f_1(x) = 2\sin(2\pi x)$. In this model, the first three variables are important, among which the first two have linear effect and the third has nonlinear effect.

Example 2 We set $p = 15$ and g_0 takes the form

$$g_0(\boldsymbol{x}) = x_1 + 1.5x_2 - 0.8x_3 + 0.5f_1(x_4) + 2f_2(x_5) - 0.3f_3(x_6),$$

where $f_1(x) = 9x^2 - 6x$, $f_2(x) = \sin(2\pi x)$, $f_3(x) = 2\exp(2x) - 3\log(2 + x)$. In this model, the first six variables are important with the first three having linear effect and others having non-linear effect.

Tables 1, 2, 3, and 4 summarize the simulation results. Tables 1 and 3 report the rate of of each component being selected as an important variable and the MRME over 100 replications in Examples 1 and 2, respectively. It can be seen that the proposed methods performs better than LAND in terms of correctly selecting the important variables and smaller MRME. The MRME's are less than 1 for all the considered methods, which suggests that the proposed methods and LAND perform better than the classical AFT regression method which omits the nonlinear effect. Tables 2 and 4 report the rates of each important variable being identified as having a nonlinear effect in Examples 1 and 2, respectively. It can be seen that the proposed method can correctly identify the nonlinear effects with higher probability than LAND method. Tables 1-4 reveal that LAND rules out unimportant variables with higher rate and selects important variables with less rates. Hence, LAND selects sparser model compared with proposed method, while the selected model by using the latter method includes true important linear effect and non-linear effect variables with higher rate.

Figure 1 displays the functional estimates of $g_j(x_j)$, $j = 1, 2, 3$ and their 95% confidence intervals in Example 1 by using group SCAD. The pointwise standard errors are calculated based on 200 bootstrap replications. Figure 2 shows the fitted functions of $g_j(x_j)$, $j = 1, \ldots, 6$ in Example 2 by using group SCAD selector. In these figures, the fitted functions are close to the real ones, and the pointwise confidence intervals cover the real value of the functions perfectly. So our proposed method by using two kinds of comparable penalties can efficiently estimate the regression coefficients of linear effect covariates, the nonparametric functions of non-linear effects and their pointwise standard errors.

Our proposed method works for high-dimensional data. However, it is often that the dimension of covariates is ultrahigh. The two-step selection procedure for ultrahigh dimensional data has been extensively used in the existing studies (Neykov et al. 2014, Ma et al. (2006). Therefore, we recommend a two-step methods when applying our proposed method to the ultrahigh dimensional censored data. We first adapt the SIS procedure (e.g. Zhang et al. 2018) to reduce the ultrahigh dimension model to a moderate scale $n/(3\log(n))$ and then apply our proposed method to analysis the reduced model. We evaluate the performance of proposed two-step method via the following example.

Example 3 The data is generated following the same model as in Example 2 but an ultra-high dimensional case with $p = 2500$, where the first six variables are important. The screening procedure by Zhang et al. (2018) is used to screening out the unrelated variables in the first step. Tables 5 reports the simulation results including the average of selected model size, MRME and the true positive rate (TPR) which is defined as the rate that the all the important variables are selected over 100 replications. 6 show the rate of each important variables being selected. It can be seen that the two-step methods works well for ultra-high dimensional data.

The theoretical results of proposed method is obtained under completely non-informative censoring assumption, which may be violated in practice. We conduct

Table 1 The rate that each component is selected as important variable and the relative model error in Example 1

Method	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	MRME
GSCAD	1	1	1	.08	.13	.14	.10	.11	.08	.11	.10	.20	.13	.13	.15	.273
GMCP	1	1	1	.03	.06	.12	.09	.08	.07	.05	.07	.14	.08	.12	.06	.279
LAND	1	.99	.56	.07	.06	.09	.07	.06	.05	.09	.05	.06	.06	.10	.07	.781
TRUE	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	–

GSCAD, GMCP: the proposed method with the group SCAD and group MCP respectively, TRUE: the oracle method with the model structure known, LAND: linear and nonlinear discoverer in Zhang et al. (2011), MRME: the median of the relative model error

Table 2 The rate that important components are selected as nonlinear effect in Example 1

Method	x_1	x_2	x_3
GSCAD	.09	.15	.94
GMCP	.10	.14	.91
LAND	.03	.02	.56
TRUE	0	0	1

GSCAD, GMCP: the proposed method with the group SCAD and group MCP respectively, TRUE: the oracle method with the model structure known, LAND: linear and nonlinear discoverer in Zhang et al. (2011)

additional simulations to examine the robustness of proposed method when this assumption is violated.

Example 4 We generate the data similar to Example 1, but the error ε is distributed by $N(0, \sigma^2)$, and the censoring time is generated from the uniform distribution $U[0, c]$ with $c = (2 + \exp(\epsilon))c_1$ if $X_1 < 0$ and $c = (3 + \exp(\epsilon))c_1$ otherwise, where c_1 is taken such that the censoring rate is 20% around. Thus the censoring time is correlated with the failure time via the covariate X_1 and the residual ϵ, where σ is chosen as 0.3, 0.5 and 1 to control the correlation between failure and censoring time.

Table 7 reports the rate of each component being selected as an important variable and MRME. Table 8 reports the structure identification results. It can be found that both the proposed methods and LAND method are robust when the assumption of the completely non-informative censoring is violated.

6 Applications

DLBCL is the gene-expression data set from diffuse large B-cell lymphomas published in Rosenwald (2002). There includes 3583 gene expression data from 112 tumors with the germinal center B-like phenotype and from 82 tumors with the activated B-like phenotype in R package DLBCL. In addition, survival information is available from 190 patients. To identify genes whose expression levels are significantly associated with survival, we first delete two samples with missing data and apply global normalization to gene expression levels so that they are comparable for different patients. To avoid the instability caused by a high dimensional matrix, we adapt SIS procedure to reduce the model to a moderate scale $n/4$.

We then apply the proposed method and competing LAND approach to identify model structure and select important variables using 188 patients data with 47 gene expressions. The analysis results are summarized in Fig. 3. All selectors identify that the gene expressions have only linear effects on the survival time, where group MCP selects 16 important linear effect gene expressions, group SCAD selects 23 important variables, and LAND identifies only one variable as important variable. Thus, LAND identifies much more sparser models than the proposed method, which is consistent with our simulation results.

Table 3 The rate that each component is selected as important variable and the relative model error in Example 2

Method	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	MRME
GSCAD	1	1	1	1	1	1	.30	.26	.20	.21	.22	.26	.19	.22	.13	.875
GMCP	1	1	1	1	1	1	.28	.21	.15	.16	.16	.17	.17	.20	.15	.881
LAND	.93	.97	.40	.90	.23	.85	.03	.03	.00	.07	.06	.09	.06	.06	.09	.867
TRUE	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	–

GSCAD, GMCP: the proposed method with the group SCAD and group MCP respectively, TRUE: the oracle method with the model structure known, LAND: linear and nonlinear discoverer in Zhang et al. (2011) respectively, MRME: the median of the relative model error

Table 4 The rate that important components are selected as nonlinear effect in Example 2

Method	x_1	x_2	x_3	x_4	x_5	x_6
GSCAD	.37	.34	.35	1	1	.71
GMCP	.35	.37	.32	1	1	.68
LAND	.05	.08	.01	.90	.23	.85
TRUE	0	0	0	1	1	1

GSCAD, GMCP: the proposed method with the group SCAD and group MCP respectively, TRUE: the oracle method with the model structure known, LAND: linear and nonlinear discoverer in Zhang et al. (2011) respectively

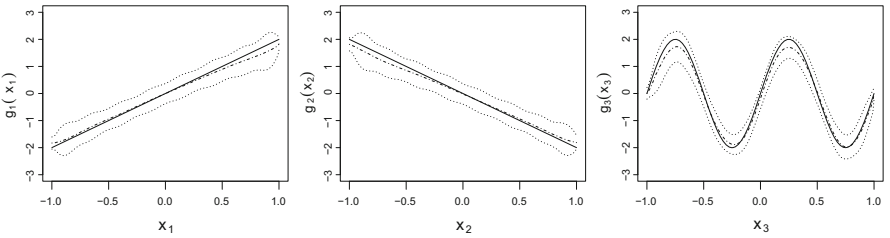

Fig. 1 Estimates of functions $g_j(x_j)$, $j = 1, \ldots, 3$ by group SCAD selector in Example 1. The solid line is the true function, the dot and dash line is the pointwise mean estimate, and the dotted lines are the 95% pointwise confidence intervals

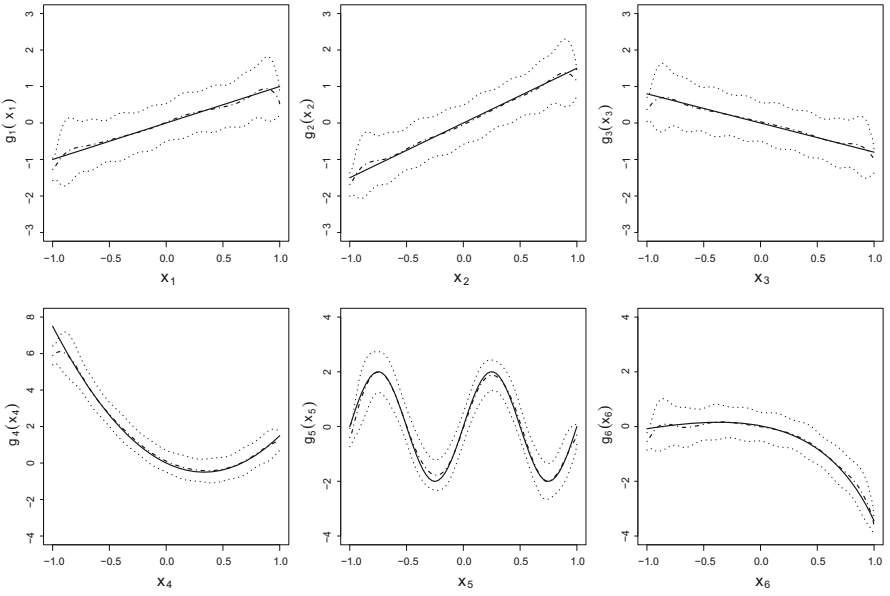

Fig. 2 Estimates of functions $g_j(x_j)$, $j = 1, \ldots, 6$ in Example 2 by group SCAD selector. The solid line is the true function, the dot and dash line is the pointwise mean estimate, and the dotted lines are the 95% pointwise confidence intervals

Table 5 The results of variable selection in Example 3

Method	#S	TPR	MRME
GSCAD	8.69	.992	.871
GMCP	8.32	.987	.856
TRUE	6	1	–

GSCAD, GMCP: the proposed method with the group SCAD and group MCP respectively. #S: the number of selected important variables; TPR: the true positive rate; MRME: the median of the relative model error;

Table 6 The rate that important components are selected as nonlinear effect in Example 3

Method	x_1	x_2	x_3	x_4	x_5	x_6
GSCAD	.39	.34	.39	1	1	.73
GMCP	.45	.36	.43	1	1	.75
TRUE	0	0	0	1	1	1

GSCAD, GMCP: the proposed method with the group SCAD and group MCP respectively

7 Concluding remarks

In this paper, we have developed a double penalized weighted least square procedure to automatically eliminate the coefficients associated with inactive variables, pursuit model structure, and estimate the nonzero effects simultaneously in additive AFT model. We show that the proposed approach can consistently identify the true model under mild assumptions and the estimates of the coefficients have the oracle property. An ADMM algorithm is applied to solve the optimization problem. Numerical calculations show that the proposed method performs well in selecting important variables, identifying the model structure and estimating the effects. We have a couple of cautionary notes on the limitations of the proposed method. Firstly, our method is based on Stute's weighted least square loss function. The validity of the proposed method relies on the assumption that the failure time and censoring time are independent. In the Buckey–James and rank based estimators, it requires only some conditional independent assumption. It is worthy to extend our proposed method to Buckey–James and rank based method. Secondly, to deal with more practical data sets, we can further consider a partially linear model

$$T_i = \sum_{j=1}^{p} \beta' Z_{ij} + \sum_{j=1}^{d} g_j(X_{ij}) + \epsilon_i, \tag{9}$$

where the covariates Z_j's are pre-known to have linear effects (e.g. the categorical variables). The proposed method is readily extended to this kind of model. Thirdly, our theoretical results are obtained under the assumption of non-informative censoring, which may not be true in most of practices. Extension of the proposed method to non-informative censoring assumption is worth of future studies. Finally, this paper tackles the theoretical problems under the scenario that the covariate dimension is the

Table 7 The rate that each component is selected as important variable and the relative model error in Example 4

Method	σ	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	MRME
GSCAD	0.3	1	1	1	.21	.05	.11	.11	.08	.07	.08	.14	.11	.16	.04	.09	.250
	0.5	1	1	1	.21	.13	.12	.09	.14	.08	.09	.12	.19	.20	.14	.17	.252
	1	1	1	1	.24	.27	.25	.30	.23	.21	.29	.28	.26	.32	.31	.25	.262
GMCP	0.3	1	1	1	.10	.06	.05	.07	.07	.07	.09	.13	.08	.12	.10	.10	.249
	0.5	1	1	1	.11	.13	.09	.05	.11	.06	.10	.12	.17	.17	.13	.11	.224
	1	1	1	1	.25	.21	.23	.21	.22	.12	.26	.25	.22	.24	.26	.20	.264
LAND	0.3	1	.99	.58	.02	.06	.03	.01	.08	.03	.05	.04	.04	.03	.02	.06	.753
	0.5	.99	.98	.43	.03	.05	.01	.08	.02	.01	.02	.04	.07	.04	.09	.05	.823
	1	.99	.91	.37	.03	.05	.03	.05	.06	.02	.04	.05	.06	.05	.05	.07	.887
TRUE	–	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	–

GSCAD, GMCP: the proposed method with the group SCAD and group MCP respectively, LAND: linear and nonlinear discoverer in Zhang et al. (2011), MRME: the median of the relative model error

Table 8 The rate that important components are selected as nonlinear effect in Example 4

Method	σ	x_1	x_2	x_3
GSCAD	0.3	.03	.13	.88
	0.5	.05	.11	.89
	1	.11	.10	.93
GMCP	0.3	.06	.15	.89
	0.5	.07	.12	.93
	1	.07	.15	.93
LAND	0.3	.11	.12	.58
	0.5	.14	.11	.43
	1	.17	.17	.37
TRUE	–	0	0	1

GSCAD, GMCP: the proposed method with the group SCAD and group MCP respectively, LAND: linear and nonlinear discoverer in Zhang et al. (2011)

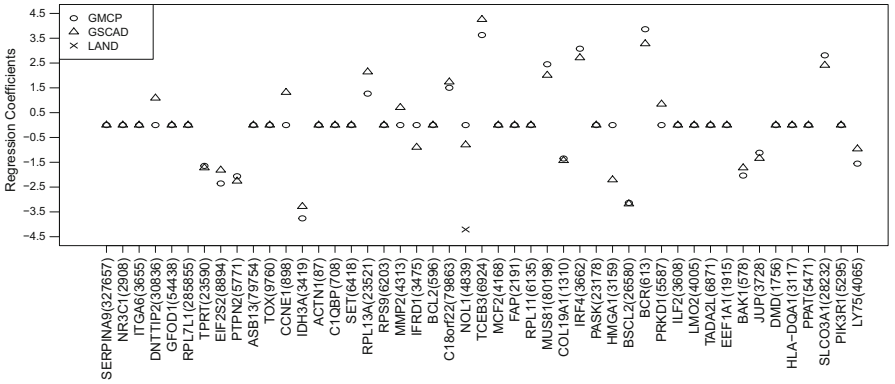

Fig. 3 The results of selection and estimation for the DLBCL data by using group MCP, group SCAD and LAND selectors

polynomial order of sample size. For ultrahigh dimensional data, we recommend a two-step method, i.e., reducing the dimension to moderate scale first by using of the existing screening methods (e.g. Liu et al. 2018; Zhang et al. 2018) and then applying our proposed method. The further research for the ultrahigh dimensional data is still in progress.

Acknowledgements The authors would like to thank the referees, the associate editor and the editor for their constructive and insightful comments and suggestions that greatly improved the paper. This research was partially supported by the National Nature Science Foundation of China (Nos. 11971362, 11571263 and 11771366). The work of J. Huang is supported in part by the NSF grant DMS-1916199.

Appendix: Proofs

Proof of Proposition 1. First, the fact that $\phi_j(x) = \beta_j + \tilde{\phi}_j(x)$ with $\beta_j = \int_{\alpha_1}^{\alpha_2} \phi_j(x)dx$ and $\tilde{\phi}_j(x) = \phi_j(x) - \beta_j$ for $j = 1, \ldots, p$ implies that the decomposition (4) holds. To show the uniqueness of the decomposition, we assume that there exist $(\beta_1^{(l)}, \ldots, \beta_{d_n}^{(l)})' \in \mathbb{R}^{d_n}$ and $(\tilde{\phi}_1^{(l)}, \ldots, \tilde{\phi}_{d_n}^{(l)})' \in \tilde{\mathcal{H}}^{d_n}$, $l = 1, 2$ such that

$$\sum_{j=1}^{d_n} x_j[\beta_j^{(1)} + \tilde{\phi}_j^{(1)}(x_j)] \equiv \sum_{j=1}^{d_n} x_j[\beta_j^{(2)} + \tilde{\phi}_j^{(2)}(x_j)]. \tag{10}$$

It suffices to prove that $\beta_j^{(1)} = \beta_j^{(2)}$ and $\tilde{\phi}_j^{(1)}(x) \equiv \tilde{\phi}_j^{(2)}(x)$ for each $j = 1, \ldots, d_n$. To the end, we note that (10) implies that

$$\sum_{j=1}^{d_n} x_j \Big(\Big[\beta_j^{(1)} - \beta_j^{(2)}\Big] + \Big[\tilde{\phi}_j^{(1)}(x_j) - \tilde{\phi}_j^{(2)}(x_j)\Big] \Big) \equiv 0.$$

When the covariates are not linearly dependent, by the Fubini's theorem, there exists $(x_1^0, \ldots, x_{j-1}^0, x_{j+1}^0, \ldots, x_{d_n}^0) \in [\alpha_1, \alpha_2]^{d_n - 1}$ such that

$$x_j \Big([\beta_j^{(1)} - \beta_j^{(2)}] + [\tilde{\phi}_j^{(1)}(x_j) - \tilde{\phi}_j^{(2)}(x_j)] \Big) \equiv - \sum_{i \neq j} x_i^0 \Big([\beta_i^{(1)} - \beta_i^{(2)}] + [\tilde{\phi}_i^{(1)}(x_i^0) - \tilde{\phi}_i^{(2)}(x_i^0)] \Big).$$

Writing $-\sum_{i \neq j} x_i^0 \Big([\beta_i^{(1)} - \beta_i^{(2)}] + [\tilde{\phi}_i^{(1)}(x_i^0) - \tilde{\phi}_i^{(2)}(x_i^0)] \Big)$ as C_j and using the condition that $E(\beta_j^{(l)} X_j + X_j \tilde{\phi}_j^{(l)}(X_j)) = E(X_j \phi(X_j))$ for $l = 1, 2$, we have

$$(\beta_j^{(1)} - \beta_j^{(2)}) + [\tilde{\phi}_j^{(1)}(x) - \tilde{\phi}_j^{(2)}(x)] \equiv 0 \tag{11}$$

for each $j = 1, \ldots, d_n$. Noting that $\tilde{\phi}_j^{(l)}(x) \in \tilde{\mathcal{H}}$, integrating two sides of (11) on variable x from α_1 to α_2 gives that

$$\beta_j^{(1)} = \beta_j^{(2)}.$$

Combining with (11), we get that

$$\tilde{\phi}_j^{(1)}(x) \equiv \tilde{\phi}_j^{(2)}(x).$$

□

Let $\mathbb{P}_n$ be the empirical measure of $\{(Y_i, \delta_i, \boldsymbol{X}_i) : i = 1, 2, \ldots, n\}$, and $\mathbb{P}$ be the probability measure of $(Y, \delta, \boldsymbol{X})$. Define $g_{nj}^*(X_j) = g_j^*(\phi_{nj}, X_j)$ and

$g_{0j}^*(X_j) = g_j^*(\phi_{0j}, X_j)$ for $\phi_{nj} \in \Omega_n$. Then denote $g_n(\boldsymbol{X}) = \sum_{j=1}^{d_n} X_j\phi_{nj}(X_j)$, $g_n^*(\boldsymbol{X}) = \sum_{j=1}^{d_n} g_{nj}^*(X_j)$ and $g_0^*(\boldsymbol{X}) = \sum_{j=1}^{d_n} g_{0j}^*(X_j)$. Define

$$\boldsymbol{C}_{\boldsymbol{\xi}} = \begin{pmatrix} -\frac{m}{u_{m+1}-u_1} & \frac{m}{u_{m+1}-u_1} & 0 & \cdots & 0 \\ 0 & -\frac{m}{u_{m+2}-u_2} & \frac{m}{u_{m+2}-u_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & -\frac{m}{u_{q_n+m-1}-u_{q_n-1}} & \frac{m}{u_{q_n+m-1}-u_{q_n-1}} \end{pmatrix}_{(q_n-1)\times q_n},$$

for $u_0 = \cdots = u_m = \xi_0$, $u_{m+1} = \xi_1, \ldots, u_{q_n-1} = \xi_{K_n-1}$, $u_{q_n} = \cdots = u_{q_n+m} = \xi_{K_n}$. Let $\xrightarrow{P}$ and $\xrightarrow{d}$ represent convergence in probability and in distribution, respectively, as $n \to \infty$ unless otherwise stated. Similar to Lemma A5 in Huang (1999), the following lemma can be established first.

Lemma 1 *Assume that Conditions (C1)–(C4) hold for any $1 \le j \le d_n$. Then there exists a function $\phi_{nj} \in \Omega_n$ such that*

$$\|g_n^* - g_0^*\|_\infty = \|g_n - g_0\|_\infty = O_p(d_n(n^{-\nu p} + n^{-(1-\nu)/2}))$$

with $\mathbb{P}_n\delta g_{nj} = 0$.

Proof According to Corollary 6.21 of Schumaker (1981), for any $1 \le j \le d_n$, there exists $\phi_{nj} \in \Omega_n$ such that $\|\phi_{nj} - \phi_{0j}\|_\infty = O(n^{-\nu p})$. We define $\widetilde{g}_{nj}(X_j) = X_j\phi_{nj}(X_j)$ and

$$g_{nj} = \widetilde{g}_{nj} - n_\delta^{-1}\mathbb{P}_n\delta\widetilde{g}_{nj},$$

where $n_\delta = \sum_{i=1}^n \delta_i/n$. Then it is easy to see that $\mathbb{P}_n\delta g_{nj} = 0$ for any $1 \le j \le d_n$. Furthermore, we note that

$$\|g_{nj} - g_{0j}\|_\infty \le \|g_{nj} - \widetilde{g}_{nj}\|_\infty + \|\widetilde{g}_{nj} - g_{0j}\|_\infty \triangleq I_{1n} + I_{2n}, \tag{12}$$

where

$$I_{1n} = \|g_{nj} - \widetilde{g}_{nj}\|_\infty \le c\|\mathbb{P}_n\delta\widetilde{g}_{nj}\|_\infty \le c(\|(\mathbb{P}_n - \mathbb{P})\delta\widetilde{g}_{nj}\|_\infty + \|\mathbb{P}(\delta\widetilde{g}_{nj} - \delta g_{0j})\|_\infty),$$

with c being a constant independent of n. By Lemma 3.4.2 in van der Vaart and Wellner (1996), we have $(\mathbb{P}_n - \mathbb{P})\delta\widetilde{g}_{nj} = O_p(n^{-1/2}n^{\nu/2})$. And the definition of ϕ_{nj} shows that $\|\mathbb{P}(\delta\widetilde{g}_{nj} - \delta g_{0j})\|_\infty \le E(\delta)\|\widetilde{g}_{nj} - g_{0j}\|_\infty = O(n^{-\nu p})$. Hence we have

$$I_{1n} = O_p(n^{-\nu p} + n^{-(1-\nu)/2}). \tag{13}$$

In addition,

$$I_{2n} = \|X_j\phi_{nj} - X_j\phi_{0j}\|_\infty = O_p(n^{-\nu p}). \tag{14}$$

Plugging (13) and (14) into (12), we can get $\|g_{nj} - g_{0j}\|_\infty = O_p(n^{-\nu p} + n^{-(1-\nu)/2})$. By using the property of Kaplan–Meier weights (Stute 1993) and Lemma 3.4.2 in van der Vaart and Wellner (1996), we have

$$\begin{aligned}
\|g^*_{nj} - g^*_{0j}\|_\infty &\le c_1\|(X_j\phi_{nj} - X_j\phi_{0j}) - (\overline{g}_{jw}(\phi_{nj}, X_j) - \overline{g}_{jw}(\phi_{0j}, X_j))\|_\infty \\
&\le c_1\|\widetilde{g}_{nj} - g_{0j}\|_\infty + c_2\Big\|\sum_{i=1}^{n}\omega_i X_{(i)j}\phi_{nj}(X_{(i)j}) - \delta\widetilde{g}_{nj}\Big\|_\infty \\
&\quad + c_3\Big\|\sum_{i=1}^{n}\omega_i X_{(i)j}\phi_{0j}(X_{(i)j}) - \delta g_{0j}\|_\infty + c_4\|\delta\widetilde{g}_{nj} - \delta g_{0j}\Big\|_\infty \\
&= O_p(n^{-\nu p}) + O_p(n^{-(1-\nu)/2}) + O_p(n^{-1/2}) + O_p(n^{-\nu p}) \\
&= O_p(n^{-\nu p} + n^{-(1-\nu)/2}),
\end{aligned}$$

where c_i's $i = 1, \ldots, 4$ are finite constants. Thus, we have

$$\|g^*_n - g^*_0\|_\infty = \|g_n - g_0\|_\infty = O_p(d_n(n^{-\nu p} + n^{-(1-\nu)/2})).$$

□

Define $\widehat{g}^*_{nj}(X_j) = g^*_j(\widehat{\phi}_{nj}, X_j)$ and $\widehat{g}^*_n(\boldsymbol{X}) = \sum_{j=1}^{d_n}\widehat{g}^*_{nj}(X_j)$, then we have the following lemma.

Lemma 2 *Assume that Conditions (C1)–(C7) hold. If* $0.25/p < \nu < 0.5$, *then* $\|\widehat{g}^*_n - g^*_n\|^2 = o_p(d_n^2 q_n^{-1})$ *and* $\Big\|\dfrac{1}{d_n}(\widehat{g}^*_n - g^*_n)\Big\|_\infty = o_p(1)$.

Proof Let $\eta_{nj} \in \Omega_n$ such that $\eta_{nj}(x) = \boldsymbol{\theta}^{*T}_{nj}\boldsymbol{\psi}_{q_n,m}(x)$ and $\|\eta_{nj}(x)\|^2 = O(q_n^{-1})$. Denote $h^*_n(\boldsymbol{X}) = \sum_{j=1}^{d_n} g^*_j(\eta_{nj}, X_j)$, then we have $\|\dfrac{1}{d_n}h^*_n(\boldsymbol{X})\|^2 = O_p(q_n^{-1})$. Define $H_n(\alpha) = Q_n(\boldsymbol{\theta}_n + \alpha\boldsymbol{\theta}^*_n)$. To prove this lemma, it is sufficient to show that for any $\alpha_0 > 0$, $H'_n(\alpha_0) > 0$ and $H'_n(-\alpha_0) < 0$ with probability tending to one.

Note that

$$\begin{aligned}
H_n(\alpha_0) = \frac{1}{2n}\|Y^* - (g^*_n + \alpha_0 h^*_n)(\boldsymbol{X})\|^2 &+ \sum_{j=1}^{d_n} P_1(\|\boldsymbol{C}_{\boldsymbol{\xi}}(\boldsymbol{\theta}_{nj} + \alpha_0\boldsymbol{\theta}^*_{nj})\|; \lambda_1) \\
&+ \sum_{j=1}^{d_n} P_2(\|\boldsymbol{\theta}_{nj} + \alpha_0\boldsymbol{\theta}^*_{nj}\|; \lambda_2).
\end{aligned}$$

Then

$$\begin{aligned}
H'_n(\alpha_0) = &-\mathbb{P}_n\Big[h^*_n\big(Y^* - g^*_n - \alpha_0 h^*_n\big)\Big] \\
&+ \sum_{j=1}^{d_n} P'_1(\|\boldsymbol{C}_{\boldsymbol{\xi}}(\boldsymbol{\theta}_{nj} + \alpha_0\boldsymbol{\theta}^*_{nj})\|; \lambda_1)\frac{(\boldsymbol{C}_{\boldsymbol{\xi}}\boldsymbol{\theta}^*_{nj})^T\boldsymbol{C}_{\boldsymbol{\xi}}(\boldsymbol{\theta}_{nj} + \alpha_0\boldsymbol{\theta}^*_{nj})}{\|\boldsymbol{C}_{\boldsymbol{\xi}}(\boldsymbol{\theta}_{nj} + \alpha_0\boldsymbol{\theta}^*_{nj})\|}
\end{aligned}$$

$$+\sum_{j=1}^{d_n} P_2'(\|\boldsymbol{\theta}_{nj}+\alpha_0\boldsymbol{\theta}_{nj}^*\|;\lambda_2)\frac{\boldsymbol{\theta}_{nj}^{*T}(\boldsymbol{\theta}_{nj}+\alpha_0\boldsymbol{\theta}_{nj}^*)}{\|\boldsymbol{\theta}_{nj}+\alpha_0\boldsymbol{\theta}_{nj}^*\|}$$
$$\triangleq H_1+H_2+H_3.$$

We consider the first part

$$\begin{aligned}H_1 &= -\mathbb{P}_n\big[h_n^*(Y^*-g_n^*)\big]+\alpha_0\mathbb{P}_n(h_n^*\cdot h_n^*)\\ &= -\mathbb{P}_n\big[h_n^*(Y^*-g_n^*)\big]+\alpha_0\mathbb{P}\|h_n^*(\boldsymbol{X})\|^2+\alpha_0(\mathbb{P}_n-\mathbb{P})(h_n^*\cdot h_n^*)\\ &= -\mathbb{P}_n\big[h_n^*(Y^*-g_n^*)\big]+c_0\alpha_0 d_n^2 n^{-\nu}+O_p(n^{-1/2}d_n^2),\end{aligned}$$

where $c_0>0$ is a constant and the first term

$$\begin{aligned}\mathbb{P}_n\big[h_n^*(Y^*-g_n^*)\big] &= (\mathbb{P}_n-\mathbb{P})\big[h_n^*(Y^*-g_n^*)\big]+\mathbb{P}\big[h_n^*(Y^*-g_n^*)\big]\\ &\triangleq J_{1n}+J_{2n}.\end{aligned}$$

In J_{1n}, $\|Y^*-g_n^*\|_\infty=\|Y^*-g_0^*+g_0^*-g_n^*\|_\infty\le O_p(1)+O_p(d_n(n^{-\nu p}+n^{-(1-\nu)/2}))$. Since $d_n^4/n\to 0$, $0.25/p<\nu<0.5$, we have $\|\frac{1}{d_n^2}h_n^*(Y^*-g_n^*)\|_\infty\le M_0$ with a constant M_0. Let

$$\mu_0(\eta)=\left\{\frac{1}{d_n^2}h_n^*\cdot(Y^*-g_n^*):\ \left\|\frac{1}{d_n}h_n^*\right\|\le\eta,\left\|\frac{1}{d_n}(g_n^*-g_0^*)\right\|\le\eta\right\}.$$

Then similar to Lemma A2 and Corollary A1 in Huang (1999), we have

$$\log N_\square(\varepsilon,\mu_0(\eta),L_2(\mathbb{P}))\le c_0 q_n\log(\eta/\varepsilon),$$

for any $\varepsilon<\eta$ with a constant c_0 and

$$J_\square(\eta,\mu_0,L_2(\mathbb{P}))\le c_0 q_n^{1/2}\eta.$$

Here we can take $\eta=q_n^{-1/2}$. Combining the results of Lemma 3.4.2 in van der Vaart and Wellner (1996) and Lemma A1 in Huang (1999), we get

$$J_{1n}=O_p(1)\cdot d_n^2\cdot n^{-1/2}\Big(q_n^{1/2}\eta+\frac{q_n}{\sqrt{n}}M_0\Big)=O_p\big(n^{-1/2}d_n^2\big).$$

We then consider J_{2n} as

$$J_{2n}=\mathbb{P}\big[h_n^*(g_n^*-g_0^*)\big]=d_n^2\cdot\mathbb{P}\Big[\frac{h_n^*}{d_n}\cdot\frac{g_0^*-g_n^*}{d_n}\Big],$$

which gives that

$$|J_{2n}| \le O_p(1)\cdot d_n^2\cdot \Big\|\frac{h_n^*}{d_n}\Big\|\cdot\Big\|\frac{g_0^*-g_n^*}{d_n}\Big\| = O_p\big(d_n^2(n^{-(1/2+p)\nu}+n^{-1/2})\big).$$

Therefore,

$$H_1 \ge c_0\alpha_0 d_n^2 n^{-\nu} + O_p(d_n^2 n^{-1/2}) + O_p(d_n^2(n^{-(1/2+p)\nu}+n^{-1/2})).$$

Next we focus on H_2 and H_3. Let $\boldsymbol{B}_j(X_j) = (\boldsymbol{\psi}_{q_n,m}(X_{1j}),\ldots,\boldsymbol{\psi}_{q_n,m}(X_{nj}))^T$. By Lemma 3 of Huang and Ma (2010), it follows that there are constants $0 < c_3 < c_4 < \infty$ such that

$$c_3 q_n^{-1} \le \Lambda_{\min}\Big(\frac{\boldsymbol{B}_j(X_j)^T\boldsymbol{B}_j(X_j)}{n}\Big) \le \Lambda_{\max}\Big(\frac{\boldsymbol{B}_j(X_j)^T\boldsymbol{B}_j(X_j)}{n}\Big) \le c_4 q_n^{-1}$$

with probability tending to one. Then we have $\|\boldsymbol{\theta}_{nj}^*\| = O_p(1)$ and $\|\boldsymbol{C}_{\boldsymbol{\xi}}\boldsymbol{\theta}_{nj}^*\| = O_p(1)$ by using of the fact that $\|\boldsymbol{\theta}_{nj}^{*T}\boldsymbol{\psi}_{q_n,m}(X_j)\| = O(q_n^{-1/2})$. Observing that

$$\begin{aligned}|H_2| &= \Bigg|\sum_{j=1}^{d_n} P_1'(\|\boldsymbol{C}_{\boldsymbol{\xi}}(\boldsymbol{\theta}_{nj}+\alpha_0\boldsymbol{\theta}_{nj}^*)\|;\lambda_1)\frac{(\boldsymbol{C}_{\boldsymbol{\xi}}\boldsymbol{\theta}_{nj}^*)^T\boldsymbol{C}_{\boldsymbol{\xi}}(\boldsymbol{\theta}_{nj}+\alpha_0\boldsymbol{\theta}_{nj}^*)}{\|\boldsymbol{C}_{\boldsymbol{\xi}}(\boldsymbol{\theta}_{nj}+\alpha_0\boldsymbol{\theta}_{nj}^*)\|}\Bigg| \\ &\le \sum_{j=1}^{d_n} P_1'(\|\boldsymbol{C}_{\boldsymbol{\xi}}(\boldsymbol{\theta}_{nj}+\alpha_0\boldsymbol{\theta}_{nj}^*)\|;\lambda_1)\frac{\big|(\boldsymbol{C}_{\boldsymbol{\xi}}\boldsymbol{\theta}_{nj}^*)^T\boldsymbol{C}_{\boldsymbol{\xi}}(\boldsymbol{\theta}_{nj}+\alpha_0\boldsymbol{\theta}_{nj}^*)\big|}{\|\boldsymbol{C}_{\boldsymbol{\xi}}(\boldsymbol{\theta}_{nj}+\alpha_0\boldsymbol{\theta}_{nj}^*)\|},\end{aligned}$$

by using of Condition (C9) and $\lambda_1 = o(d_n n^{-\nu})$, we have

$$|H_2| \le P_1'(0+;\lambda_1)\sum_{j=1}^{d_n}\|\boldsymbol{C}_{\boldsymbol{\xi}}\boldsymbol{\theta}_{nj}^*\| \le O(\lambda_1)O_p(d_n) = o_p(d_n^2 n^{-\nu}).$$

The same arguments as above give that $|H_3| \le o_p(d_n^2 n^{-\nu})$ if $\lambda_2 = o(d_n n^{-\nu})$.

Consequently, $H_n'(\alpha_0) \ge c_0\alpha_0 d_n^2 n^{-\nu} + o_p(d_n^2 n^{-\nu}) > 0$ with probability tending to one. Similarly, we can prove that $H_n'(-\alpha_0) < 0$ with probability tending to one. Therefore, the boundness of covariate $\boldsymbol{X}$ in Condition (C2) ensures that

$$\|\widehat{g}_n^* - g_n^*\|^2 = o_p(d_n^2 q_n^{-1}) = o_p(d_n^2 n^{-\nu}).$$

Subsequently, Lemma 7 of Stone (1986) yields that $\|\frac{1}{d_n}(\widehat{g}_n^* - g_n^*)\|_\infty = o_p(1)$. □

To verify the consistency of parameter estimation, we need the following lemma.

Lemma 3 *Define* $m_0(x, y^*; g^*) = (y^* - g^*(x))^2/d_n^2$. *Denote* $M_0 = \mathbb{P}m_0$ *and* $M_n = \mathbb{P}_n m_0 = \frac{1}{n}\|Y^* - g^*(\boldsymbol{X})\|^2/d_n^2$. *Under the conditions of Lemma 1, for any function* $g(\cdot)$ *satisfying* $E[\delta g(\boldsymbol{X})] = 0$, *there exists a constant* $c > 0$ *such that*

$$\mathbb{P}m_0(\cdot; g^*) - \mathbb{P}m_0(\cdot; g_n^*) = c\left\|\frac{1}{d_n}(g^* - g_n^*)\right\|^2 + O_p(n^{-2\nu p} + n^{-(1-\nu)}).$$

Proof Let $h^* = g^* - g_0^*$ and

$$\begin{aligned} L(s) &= \mathbb{P}m_0(\cdot; g_0^* + sh^*) - \mathbb{P}m_0(\cdot; g_0^*) \\ &= \frac{1}{d_n^2}\left[\mathbb{P}(Y^* - (g_0^* + sh^*))^2 - \mathbb{P}(Y^* - g_0^*)^2\right] \\ &= \frac{1}{d_n^2}\mathbb{P}(-2sY^*h^* + 2sg_0^*h^* + s^2h^{*2}). \end{aligned}$$

Since $L'(0) = 0$ and $L''(0) = 2\mathbb{P}(h^{*2})/d_n^2$, there exists a constant $c > 0$, such that $\mathbb{P}m_0(\cdot; g^*) - \mathbb{P}m_0(\cdot; g_0^*) = c\left\|\frac{1}{d_n}(g^* - g_0^*)\right\|^2$. Similarly, we have

$$\mathbb{P}m_0(\cdot; g_n^*) - \mathbb{P}m_0(\cdot; g_0^*) = O_p(1)\left\|\frac{1}{d_n}(g_n^* - g_0^*)\right\|^2.$$

By Lemma 1, $\mathbb{P}m_0(\cdot; g_n^*) - \mathbb{P}m_0(\cdot; g_0^*) = O_p(n^{-2\nu p} + n^{-(1-\nu)})$. Combining the following equality

$$\mathbb{P}m_0(\cdot; g^*) - \mathbb{P}m_0(\cdot; g_n^*) = \left(\mathbb{P}m_0(\cdot; g^*) - \mathbb{P}m_0(\cdot; g_0^*)\right) + \left(\mathbb{P}m_0(\cdot; g_0^*) - \mathbb{P}m_0(\cdot; g_n^*)\right)$$

with the triangle inequality

$$\|g^* - g_n^*\|^2 - \|g_n^* - g_0^*\|^2 \le \|g^* - g_0^*\|^2 \le \|g^* - g_n^*\|^2 + \|g_n^* - g_0^*\|^2,$$

we have

$$\mathbb{P}m_0(\cdot; g^*) - \mathbb{P}m_0(\cdot; g_n^*) = c\left\|\frac{1}{d_n}(g^* - g_n^*)\right\|^2 + O_p(n^{-2\nu p} + n^{-(1-\nu)}),$$

where $c > 0$ is a finite constant. □

Proof of Theorem 1. Let

$$\begin{aligned} V &= M_n(g^*) - M_n(g_n^*) - (M_0(g^*) - M_0(g_n^*)) \\ &= \mathbb{P}_n m_0(\cdot; g^*) - \mathbb{P}_n m_0(\cdot; g_n^*) - (\mathbb{P}m_0(\cdot; g^*) - \mathbb{P}m_0(\cdot; g_n^*)) \\ &= (\mathbb{P}_n - \mathbb{P})(m_0(\cdot; g^*) - m_0(\cdot; g_n^*)), \end{aligned}$$

By Lemma 3.4.2 of van der Vaart and Wellner (1996),

$$E \sup_{\left\|\frac{1}{d_n}(g^*-g_n^*)\right\|\le\eta} |V| = n^{-1/2}\eta q_n^{1/2}.$$

Then by Theorem 3.4.1 of van der Vaart and Wellner (1996), choosing the distance $d(\widehat{g}_n^*, g_n^*) = -[\mathbb{P}m_0(\cdot;\widehat{g}_n^*) - \mathbb{P}m_0(\cdot; g_n^*)]$ there, we have

$$-r_{1n}^2[\mathbb{P}m_0(\cdot;\widehat{g}_n^*) - \mathbb{P}m_0(\cdot; g_n^*)] = O_p(1),$$

where $r_{1n} = O(n^{1/2}q_n^{-1/2}) = O(n^{(1-\nu)/2})$. Therefore, $\mathbb{P}m_0(\cdot;\widehat{g}_n^*) - \mathbb{P}m_0(\cdot; g_n^*) = O_p(n^{-(1-\nu)})$. Thus Lemma 3 gives that $\|\frac{1}{d_n}(\widehat{g}_n^* - g_n^*)\|^2 = O_p(n^{-2\nu p} + n^{-(1-\nu)})$. Combining the result in Lemma 1 that $\|g_n^* - g_0^*\|_\infty^2 = O_p(d_n^2(n^{-2\nu p} + n^{-(1-\nu)}))$, we have

$$\|\widehat{g}_n^* - g_0^*\|^2 = O_p(d_n^2(n^{-2\nu p} + n^{-(1-\nu)})).$$

By Conditions (C2)–(C4), it follows that

$$E\delta\left\|\boldsymbol{X}_{M_1}(\widehat{\boldsymbol{\phi}}_n(\boldsymbol{X}_{M_1}) - \tilde{\boldsymbol{\phi}}_0(\boldsymbol{X}_{M_1})) + \boldsymbol{X}_{M_2}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)\right\|^2 = O(d_n^2(n^{-2\nu p} + n^{-(1-\nu)})).$$

Denoting the projection of $\boldsymbol{X}_{M_2}$ on $\boldsymbol{X}_{M_1}$ as W, we have

$$\begin{aligned} &E\delta\left\|(\boldsymbol{X}_{M_2} - W)(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) + W(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) + \boldsymbol{X}_{M_1}(\widehat{\boldsymbol{\phi}}_n - \tilde{\boldsymbol{\phi}}_0)\right\|^2 \\ &= E\delta\left\|(\boldsymbol{X}_{M_2} - W)(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)\right\|^2 + E\delta\left\|W(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) + \boldsymbol{X}_{M_1}(\widehat{\boldsymbol{\phi}}_n - \tilde{\boldsymbol{\phi}}_0)\right\|^2 \\ &= O(d_n^2(n^{-2\nu p} + n^{-(1-\nu)})). \end{aligned}$$

By Condition (C6), we obtain

$$\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|^2 = O_p(d_n^2(n^{-(1-\nu)} + n^{-2\nu p})).$$

This in turn implies $E\delta\left\|\boldsymbol{X}_{M_1}(\widehat{\boldsymbol{\phi}}_n - \tilde{\boldsymbol{\phi}}_0)\right\|^2 = O_p(d_n^2(n^{-(1-\nu)} + n^{-2\nu p}))$. Therefore,

$$\|\widehat{\boldsymbol{\phi}}_n - \tilde{\boldsymbol{\phi}}_0\|^2 = O_p(d_n^2(n^{-(1-\nu)} + n^{-2\nu p})).$$

This completes the proof of Theorem 1. □

Proof of Theorem 2. (i) First, we prove the selection consistency of the variables. Let $\widetilde{\boldsymbol{\theta}}_n = (\widetilde{\boldsymbol{\theta}}_{n1}^T, \ldots, \widetilde{\boldsymbol{\theta}}_{nd_n}^T)^T$ with

$$\widetilde{\boldsymbol{\theta}}_{nj} = \begin{cases} \widehat{\boldsymbol{\theta}}_{nj}, & \text{if } j \notin M_3, \\ 0, & \text{if } j \in M_3. \end{cases}$$

Note that $\widehat{\boldsymbol{\theta}}_n$ satisfies $\dfrac{\partial Q_n(\widehat{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}} = \mathbf{0}$. By the definition of $\widehat{\boldsymbol{\theta}}_n$ and $\widetilde{\boldsymbol{\theta}}_n$, we have

$$
\begin{aligned}
& Q_n(\widehat{\boldsymbol{\theta}}_n) - Q_n(\widetilde{\boldsymbol{\theta}}_n) \\
&= \frac{\partial Q_n(\widehat{\boldsymbol{\theta}}_n)^T}{\partial \boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}_n - \widetilde{\boldsymbol{\theta}}_n) - \frac{1}{2}(\widehat{\boldsymbol{\theta}}_n - \widetilde{\boldsymbol{\theta}}_n)^T \frac{\partial^2 Q_n(\boldsymbol{\theta}_n^*)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}(\widehat{\boldsymbol{\theta}}_n - \widetilde{\boldsymbol{\theta}}_n) \\
&= -\frac{1}{2}(\widehat{\boldsymbol{\theta}}_n - \widetilde{\boldsymbol{\theta}}_n)^T \frac{\partial^2 Q_n(\boldsymbol{\theta}_n^*)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}(\widehat{\boldsymbol{\theta}}_n - \widetilde{\boldsymbol{\theta}}_n) \\
&= -\frac{1}{2}\widehat{\boldsymbol{\theta}}_{nM_3}^T \frac{\partial^2 \ell_n(\boldsymbol{\theta}_n^*)}{\partial \boldsymbol{\theta}_{M_3}\partial \boldsymbol{\theta}_{M_3}^T}\widehat{\boldsymbol{\theta}}_{nM_3} - \frac{1}{2}\sum_{j\in M_3}(\boldsymbol{C}_{\boldsymbol{\xi}}\widehat{\boldsymbol{\theta}}_{nj})^T\big(P_1''(\|\boldsymbol{C}_{\boldsymbol{\xi}}\widehat{\boldsymbol{\theta}}_{nj}\|;\lambda_1) + o_p(1)\big)(\boldsymbol{C}_{\boldsymbol{\xi}}\widehat{\boldsymbol{\theta}}_{nj}) \\
&\quad -\frac{1}{2}\sum_{j\in M_3}\widehat{\boldsymbol{\theta}}_{nj}^T\big(P_2''(\|\widehat{\boldsymbol{\theta}}_{nj}\|;\lambda_2) + o_p(1)\big)\widehat{\boldsymbol{\theta}}_{nj},
\end{aligned}
$$

where $\boldsymbol{\theta}_n^*$ is between $\widehat{\boldsymbol{\theta}}_n$ and $\widetilde{\boldsymbol{\theta}}_n$.
Since $\widehat{\boldsymbol{\theta}}_n$ is the minimizer of $Q(\boldsymbol{\theta})$, we have $Q(\widehat{\boldsymbol{\theta}}_n) \le Q(\widetilde{\boldsymbol{\theta}}_n)$, which implies that

$$
\begin{aligned}
\frac{1}{2}\widehat{\boldsymbol{\theta}}_{nM_3}^T \frac{\partial^2 \ell_n(\boldsymbol{\theta}_n^*)}{\partial \boldsymbol{\theta}_{M_3}\partial \boldsymbol{\theta}_{M_3}^T}\widehat{\boldsymbol{\theta}}_{nM_3} \ge & -\frac{1}{2}\sum_{j\in M_3}(\boldsymbol{C}_{\boldsymbol{\xi}}\widehat{\boldsymbol{\theta}}_{nj})^T\big(P_1''(\|\boldsymbol{C}_{\boldsymbol{\xi}}\widehat{\boldsymbol{\theta}}_{nj}\|;\lambda_1) + o_p(1)\big)(\boldsymbol{C}_{\boldsymbol{\xi}}\widehat{\boldsymbol{\theta}}_{nj}) \\
& -\frac{1}{2}\sum_{j\in M_3}\widehat{\boldsymbol{\theta}}_{nj}^T\big(P_2''(\|\widehat{\boldsymbol{\theta}}_{nj}\|;\lambda_2) + o_p(1)\big)\widehat{\boldsymbol{\theta}}_{nj}. \qquad (15)
\end{aligned}
$$

Note that the left hand of Eq (15)

$$
I_1 \le c\widehat{\boldsymbol{\theta}}_{nM_3}^T E(X_{M_3}^T X_{M_3})\widehat{\boldsymbol{\theta}}_{nM_3} \le c\rho_n^*\|\widehat{\boldsymbol{\theta}}_{nM_3}\|^2
$$

for some constant c by the continuity of the B-spline functions and the definition of ρ_n^*. And using Condition (C9), there exist constants a, b and c such that the right hand of Eq (15)

$$
I_2 \ge c(\lambda_1^a + \lambda_2^a)\|\widehat{\boldsymbol{\theta}}_{nM_3}\|^{2-b}.
$$

Thus, by the results of Theorem 1, we obtain that

$$
O_p(1)(d_n^2(n^{-(1-\nu)} + n^{-2\nu p}))^{b/2} \ge \|\widehat{\boldsymbol{\theta}}_{nM_3}\|^b \ge O_p(1)\frac{\lambda_1^a + \lambda_2^a}{\rho_n^*}.
$$

This shows that under the condition that $\dfrac{\lambda_1^a}{\rho_n^*(d_n^2(n^{-(1-\nu)} + n^{-2\nu p}))^{b/2}}$ and $\dfrac{\lambda_2^a}{\rho_n^*(d_n^2(n^{-(1-\nu)} + n^{-2\nu p}))^{b/2}}$ goes to infinity,

$$
P(\|\widehat{\boldsymbol{\theta}}_{nM_3}\| > 0) \le P\Big(\frac{\lambda_1^a + \lambda_2^a}{\rho_n^*(d_n^2(n^{-(1-\nu)} + n^{-2\nu p}))^{b/2}} \le O_p(1)\Big) \to 0.
$$

Next, we prove the structure selection consistency. Assume that $\underset{\sim}{\boldsymbol{\theta}}_n = (\underset{\sim}{\boldsymbol{\theta}}_{n1}^T, \ldots, \underset{\sim}{\boldsymbol{\theta}}_{nd_n}^T)^T$ with

$$\underset{\sim}{\boldsymbol{\theta}}_{nM_2} = \begin{cases} \widehat{\boldsymbol{\theta}}_{nj}, & \text{if } j \notin M_2, \\ \underset{\sim}{\boldsymbol{\theta}}_{nj}, s.t.\ \boldsymbol{C}_{\boldsymbol{\xi}} \underset{\sim}{\boldsymbol{\theta}}_{nj} = 0, & \text{if } j \in M_2. \end{cases}$$

Then we have

$$\begin{aligned} &\frac{1}{2}\Big(\widehat{\boldsymbol{\theta}}_{nM_2} - \underset{\sim}{\boldsymbol{\theta}}_{nM_2}\Big)^T \frac{\partial^2 \ell_n(\boldsymbol{\theta}_n^0)}{\partial \boldsymbol{\theta}_{M_2} \partial \boldsymbol{\theta}_{M_2}^T}\Big(\widehat{\boldsymbol{\theta}}_{nM_2} - \underset{\sim}{\boldsymbol{\theta}}_{nM_2}\Big) \\ \geq\ & -\frac{1}{2}\sum_{j \in M_2} (\boldsymbol{C}_{\boldsymbol{\xi}} \widehat{\boldsymbol{\theta}}_{nj})^T \big(P_1''(\|\boldsymbol{C}_{\boldsymbol{\xi}} \widehat{\boldsymbol{\theta}}_{nj}\|; \lambda_1) + o_p(1)\big)(\boldsymbol{C}_{\boldsymbol{\xi}} \widehat{\boldsymbol{\theta}}_{nj}) \\ & -\frac{1}{2}\sum_{j \in M_2} \Big(\widehat{\boldsymbol{\theta}}_{nj} - \underset{\sim}{\boldsymbol{\theta}}_{nj}\Big)^T \big(P_2''(\|\widehat{\boldsymbol{\theta}}_{nj}\|; \lambda_2) + o_p(1)\big)\Big(\widehat{\boldsymbol{\theta}}_{nj} - \underset{\sim}{\boldsymbol{\theta}}_{nj}\Big), \end{aligned} \tag{16}$$

where $\boldsymbol{\theta}_n^0$ is between $\widehat{\boldsymbol{\theta}}_n$ and $\underset{\sim}{\boldsymbol{\theta}}_n$. The left hand of equation (16)

$$\begin{aligned} II_1 &\leq O_p(1)\Big(\boldsymbol{C}_\xi(\widehat{\boldsymbol{\theta}}_{nM_2} - \underset{\sim}{\boldsymbol{\theta}}_{nM_2})\Big)^T \cdot \frac{\partial^2 \ell_n(\boldsymbol{\theta}_n^0)}{\partial \boldsymbol{\theta}_{M_2} \partial \boldsymbol{\theta}_{M_2}^T} \cdot \Big(\boldsymbol{C}_\xi(\widehat{\boldsymbol{\theta}}_{nM_2} - \underset{\sim}{\boldsymbol{\theta}}_{nM_2})\Big) \\ &= O_p(1)(\boldsymbol{C}_\xi \widehat{\boldsymbol{\theta}}_{nM_2})^T \cdot \frac{\partial^2 \ell_n(\boldsymbol{\theta}_n^0)}{\partial \boldsymbol{\theta}_{M_2} \partial \boldsymbol{\theta}_{M_2}^T} \cdot (\boldsymbol{C}_\xi \widehat{\boldsymbol{\theta}}_{nM_2}) \\ &\leq O_p(1)\rho_n^* \|\widehat{\boldsymbol{\theta}}_{nM_2}\|^2. \end{aligned}$$

Similarly, we can obtain that the right hand of equation (16)

$$II_2 \geq c(\lambda_1^a + \lambda_2^a)\|\widehat{\boldsymbol{\theta}}_{nM_2}\|^{2-b}.$$

Therefore,

$$P(\|\widehat{\boldsymbol{\theta}}_{nM_2}\| > 0) \leq P\Big(\frac{\lambda_1^a + \lambda_2^a}{\rho_n^*(d_n^2(n^{-(1-\nu)} + n^{-2\nu p}))^{b/2}} \leq O_p(1)\Big) \to 0.$$

The selection consistency of variable and structure is concluded.

(ii) Let the column and row vectors of covariate matrix $\boldsymbol{X}^*$ are $X_1^*, \ldots, X_{d_n}^*$ and $X_{(1)}^*, \ldots, X_{(n)}^*$, respectively. Define

$$\overline{X}_w = \frac{\sum_{i=1}^n \omega_i X_{(i)}}{\sum_{i=1}^n \omega_i}, \quad X_{(i)}^* = (n\omega_i)^{1/2}(X_{(i)} - \overline{X}_w),$$

$$U(\boldsymbol{W};\boldsymbol{\beta},\widehat{\boldsymbol{\phi}}_n) \triangleq (-\boldsymbol{X}^*_{M_2})\Big(Y^* - \sum_{j\in M_1}\hat{g}^*_{nj}(X_j) - \boldsymbol{X}^*_{M_2}\boldsymbol{\beta}\Big),$$

$$\widehat{U}_n(\boldsymbol{\beta}) \triangleq \frac{1}{n}\sum_{i=1}^{n} U(\boldsymbol{W}_i;\boldsymbol{\beta},\widehat{\boldsymbol{\phi}}_n),$$

with $\boldsymbol{W} \triangleq (\omega, \boldsymbol{X}, Y)$. Then $\widehat{\boldsymbol{\beta}}_n$ satisfies the estimating equation $\widehat{U}_n(\widehat{\boldsymbol{\beta}}) = 0$ by the definition of $\widehat{\boldsymbol{\beta}}_n$ and $\widehat{\boldsymbol{\phi}}_n$.

Let $U_n(\boldsymbol{\beta}) \triangleq \frac{1}{n}\sum_{i=1}^{n} U(\boldsymbol{W}_i;\boldsymbol{\beta},\tilde{\boldsymbol{\phi}}_0)$ and $\widetilde{\boldsymbol{\beta}}_n$ be the root of $U_n(\boldsymbol{\beta}) = 0$. We then show that $\widehat{\boldsymbol{\beta}}_n$ has the same distribution with $\widetilde{\boldsymbol{\beta}}_n$. The Fréchet derivative of $U(\boldsymbol{W};\boldsymbol{\beta}_0,\boldsymbol{\phi})$ at $\tilde{\boldsymbol{\phi}}_0$ in the direction $\boldsymbol{h}$ is given as

$$\begin{aligned} D(\boldsymbol{W},\boldsymbol{h}) &= \lim_{\alpha\to 0}\frac{U(\boldsymbol{W};\boldsymbol{\beta}_0,\tilde{\boldsymbol{\phi}}_0+\alpha\boldsymbol{h}) - U(\boldsymbol{W};\boldsymbol{\beta}_0,\tilde{\boldsymbol{\phi}}_0)}{\alpha} \\ &= \boldsymbol{X}^{*T}_{M_2}\boldsymbol{X}^*_{M_1}\boldsymbol{h}, \end{aligned}$$

with $\boldsymbol{h} \in \{h_1 + \ldots + h_{|M_1|}, h_j \in \tilde{\mathcal{H}},\ j \in M_1\}$.

The relation $\|(\widehat{\boldsymbol{\phi}}_n - \tilde{\boldsymbol{\phi}}_0)/d_n\| = O_p(n^{-(1-\nu)/2} + n^{-\nu p}) = o_p(n^{-1/4})$ ensures that the linear assumption 5.1 in Newey (1994) is satisfied. Then by Lemma 3.4.2 of van der Vaart and Wellner (1996), we have

$$\sqrt{n}(\mathbb{P}_n - \mathbb{P})\{D(\boldsymbol{W};\widehat{\boldsymbol{\phi}}_n - \tilde{\boldsymbol{\phi}}_0)\} \xrightarrow{P} 0.$$

It follows that the stochastic equicontinuity assumption 5.2 holds. For $\boldsymbol{\phi}$ close enough to $\tilde{\boldsymbol{\phi}}_0$, a straightforward calculation yields that $ED(\boldsymbol{W};\boldsymbol{\phi} - \tilde{\boldsymbol{\phi}}_0) = 0$ by using Condition (C8). Then the mean square continuity assumption 5.3 holds with $\alpha(\boldsymbol{W}) = 0$. By Lemma 5.1 of Newey (1994), $\widehat{\boldsymbol{\beta}}_n$ and $\widetilde{\boldsymbol{\beta}}_n$ have the same distribution.

Next, we seek for the asymptotic distribution of $\widetilde{\boldsymbol{\beta}}_{nM_2}$. Let $\iota_n = n^{-1/2}$, $V_{1n}(\boldsymbol{a}) = Q_n(\boldsymbol{\beta}_0 + \iota_n(\boldsymbol{a}^T, \boldsymbol{0}^T)^T, \tilde{\boldsymbol{\phi}}_0) - Q_n(\boldsymbol{\beta}_0, \tilde{\boldsymbol{\phi}}_0)$, where $\boldsymbol{a} = (a_1, \ldots, a_{|M_2|})^T$ is a $|M_2|$-dimensional constant vector and $\boldsymbol{0}$ is a $|M_3|$-dimensional zero vector. By part (i) of Theorem 2, $\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0 = \iota_n(\widehat{\boldsymbol{a}}_n^T, \boldsymbol{0}^T)^T$ with probability converging to one, where $\widehat{\boldsymbol{a}}_n = \operatorname{argmin}\{V_{1n}(\boldsymbol{a}) : \boldsymbol{a} \in \mathbb{R}^{|M_2|}\}$. Letting $\widetilde{\boldsymbol{\theta}}_n$ be the estimator corresponding to $\widetilde{\boldsymbol{\beta}}_n$, then similar to Theorem 2 (i), we also have $\boldsymbol{C}_{\boldsymbol{\xi}}\widetilde{\boldsymbol{\theta}}_{nj} = 0$ $(j \in M_2)$ with probability converging to one.

Note that

$$\begin{aligned} V_{1n}(\boldsymbol{a}) &= Q_n(\boldsymbol{\beta}_0 + \iota_n(\boldsymbol{a}^T, \boldsymbol{0}^T)^T, \tilde{\boldsymbol{\phi}}_0) - Q_n(\boldsymbol{\beta}_0, \tilde{\boldsymbol{\phi}}_0) \\ &= \Big(\iota_n\boldsymbol{a}^T U_n(\boldsymbol{\beta}_0) + \frac{\iota_n^2}{2}\boldsymbol{a}^T U_n'(\boldsymbol{\beta}_0)\boldsymbol{a}\Big) \end{aligned}$$

$$
\begin{aligned}
&+\left(\sum_{j\in M_2} P_1(\|\boldsymbol{C}_{\boldsymbol{\xi}}\widetilde{\boldsymbol{\theta}}_{nj}\|;\lambda_1)-\sum_{j\in M_2} P_1(\|\boldsymbol{C}_{\boldsymbol{\xi}}\boldsymbol{\theta}_{0j}\|;\lambda_1)\right)\\
&+\left(\sum_{j\in M_2} P_2(\|\widetilde{\boldsymbol{\theta}}_{nj}\|;\lambda_{2n})-\sum_{j\in M_2} P_2(\|\boldsymbol{\theta}_{0j}\|;\lambda_{2n})\right)\\
&\triangleq A_{1n}(\boldsymbol{a})+A_{2n}(\boldsymbol{a})+A_{3n}(\boldsymbol{a}).
\end{aligned}
$$

Since $\widetilde{\boldsymbol{\beta}}_{nM_2}-\boldsymbol{\beta}_0=\iota_n\boldsymbol{a}=\boldsymbol{\psi}_{q_n,m}(\boldsymbol{X}_{M_2})^T(\widetilde{\boldsymbol{\theta}}_{nM_2}-\boldsymbol{\theta}_{0M_2})$, we have

$$
\widetilde{\boldsymbol{\theta}}_{nj}-\boldsymbol{\theta}_{0j}=(\boldsymbol{\psi}_{q_n,m}(X_j)\boldsymbol{\psi}_{q_n,m}(X_j)^T)^{-1}\boldsymbol{\psi}_{q_n,m}(X_j)\iota_n a_j,\ \ j=s_1+1,\ldots,s_2.
$$

It follows that

$$
\begin{aligned}
A_{3n}(\boldsymbol{a})&=\sum_{j\in M_2} P_2(\|\widetilde{\boldsymbol{\theta}}_{nj}\|;\lambda_{2n})-\sum_{j\in M_2} P_2(\|\boldsymbol{\theta}_{0j}\|;\lambda_{2n})\\
&=\sum_{j\in M_2}\left[P_2'(\|\boldsymbol{\theta}_{0j}\|;\lambda_{2n})\frac{\boldsymbol{\theta}_{0j}^T}{\|\boldsymbol{\theta}_{0j}\|}+o_p(1)\right]\\
&\quad\left[(\boldsymbol{\psi}_{q_n,m}(X_j)\boldsymbol{\psi}_{q_n,m}(X_j)^T)^{-1}\boldsymbol{\psi}_{q_n,m}(X_j)\iota_n a_j\right].
\end{aligned}
$$

By Condition (C7), we have

$$
\begin{aligned}
|A_{3n}(\boldsymbol{a})|&\le d_n P_2'(0+;\lambda_{2n})\sqrt{\|(\boldsymbol{\psi}_{q_{n,m}}(X_j)\boldsymbol{\psi}_{q_{n,m}}(X_j)^T)^{-1}\|}\iota_n a_j\\
&=O_p(d_n^2 n^{-\nu})O_p(n^{-(1-\nu)/2})=o_p(1).
\end{aligned}
$$

Similarly, we can get that $A_{2n}(\boldsymbol{a})\xrightarrow{p}0$.
Hence, $\widehat{\boldsymbol{a}}_n=\operatorname{argmin}\{V_{1n}(\boldsymbol{a}):\boldsymbol{a}\in\mathbb{R}^{|M_2|}\}=\operatorname{argmin}\{A_{1n}(\boldsymbol{a}):\boldsymbol{a}\in\mathbb{R}^{|M_2|}\}$ and so we only care about the minimum of $nA_{1n}(\boldsymbol{a})$. Similar to Huang et al. (2010), we have

$$
\begin{aligned}
nA_{1n}(\boldsymbol{a})&=\boldsymbol{a}^T\left(\sqrt{n}U_n(\boldsymbol{\beta}_0)\right)+\frac{1}{2}\boldsymbol{a}^T U_n'(\boldsymbol{\beta}_0)\boldsymbol{a}\\
&\triangleq\boldsymbol{a}^T T_1+\boldsymbol{a}^T T_2\boldsymbol{a}.
\end{aligned}
$$

It can be seen that $T_2\xrightarrow{p}\Sigma_2$ and $\boldsymbol{u}\Sigma_3^{-1/2}T_1$ is distributed asymptotically by $N(0,1)$ for any $\boldsymbol{u}\in\mathbb{R}^{|M_2|}$ with $\|\boldsymbol{u}\|=1$, where $\Sigma_3=Var(\delta\gamma_0(Y)(Y-g_0(\boldsymbol{X}))\boldsymbol{X}_{M_2}+(1-\delta)\gamma_1(Y)-\gamma_2(Y))$ with the following notations that

$$
\begin{aligned}
&\tilde{H}^{11}(\boldsymbol{x},y)=P(\boldsymbol{X}\le\boldsymbol{x},Y\le y,\delta=1),\quad \tilde{H}^0=P(Y\le y,\delta=0),\\
&\gamma_0(y)=\exp\left(\int_0^{y-}\frac{\tilde{H}^0(dw)}{1-H(w)}\right),
\end{aligned}
$$

$$\gamma_{1,j}(y) = \frac{1}{1-H(y)} \int I(z > y) e(z, \boldsymbol{x}) x_j \gamma_0(z) \tilde{H}^{11}(d\boldsymbol{x}, dz),$$
$$\gamma_{2,j}(y) = \iint \frac{I(v < y, v < z) e(z, \boldsymbol{x}) x_j \gamma_0(z)}{[1-H(v)]^2} \tilde{H}^0(dv) \tilde{H}^{11}(d\boldsymbol{x}, dz),$$
$$\gamma_l(y) = (\gamma_{l,j};\ j \in M_2),\ \ l = 1, 2.$$

Let $\hat{\boldsymbol{a}} = \text{argmin}\{V_1(\boldsymbol{a}) = \boldsymbol{a}^T T_1 + \frac{1}{2}\boldsymbol{a}^T \Sigma_2 \boldsymbol{a} : \boldsymbol{a} \in \mathbb{R}^{|M_2|}\}$. According to the continuous mapping theorem of Kim and Pollard (1990), $\sqrt{n}\boldsymbol{u}\Sigma^{-1/2}(\widehat{\boldsymbol{\beta}}_{nM_2} - \boldsymbol{\beta}_0)$ has the same asymptotical distribution as $\boldsymbol{u}\Sigma^{-1/2}\hat{\boldsymbol{a}} \xrightarrow{d} N(0, 1)$ for any $\boldsymbol{u} \in \mathbb{R}^{|M_2|}$ with $\|\boldsymbol{u}\| = 1$, where $\Sigma = \Sigma_2^{-1}\Sigma_3\Sigma_2^{-1}$. This completes the proof of Theorem 2. □

References

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Second international symposium on information theory, pp 267–281

Antoniadis A, Gijbels I, Lambert-Lacroix S (2014) Penalized estimation in additive varying coefficient models using grouped regularization. Stat Pap 55:727–750

Boyd S, Parikh N, Chu E, Peleato B, Eckstein J (2011) Distributed optimization and statistical learning via the alternating direction method of multipliers. Found Trends Mach Learn 3:1–122

Buckley J, James I (1979) Linear regression with censored data. Biometrika 66:429–436

Candes E, Tao T (2007) The Dantzig selector: statsitical estimation when p is much larger than n. Ann Stat 35:2313–2351

Cao Y, Huang J, Liu Y, Zhao X (2016) Sieve estimation of Cox models with latent structures. Biometrics 72:1086–1097

Chen K, Shen J, Ying Z (2005) Rank estimation in partial linear model with censored data. Stat Sin 15(3):767–779

Chen S, Zhou Y, Ji Y (2018) Nonparametric identification and estimation of sample selection models under symmetry. J Econom 202(2):148–160

Craven P, Wahba G (1979) Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. Numer Math 31:377–403

de Boor C (1978) A practical guide to splines. Applied Mathematical Sciences, vol 27, no 149. Springer, New York, pp 157

Fleming TR, Harrington DP (1991) Counting processes and survival analysis. Wiley, New York

Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc 96:1348–1360

Fan J, Li R (2002) Variable selection for Cox's proportional hazards model and frailty model. Ann Stat 30:74–99

Huang J (1999) Efficient estimation of the partly linear additive Cox model. Ann Stat 27:1536–1563

Huang J, Ma S (2010) Variable selection in the accelerated failure time model via the bridge method. Lifetime Data Anal 16:176–195

Huang J, Horowitz JL, Wei F (2010) Variable selection in nonparametric additive models. Ann Stat 38:2282–2313

Huang J, Wei F, Ma S (2012) Semiparametic regression pursuit. Stat Sin 22:1403–1426

Joseph A (2013) Variable selection in high-dimension with random designs and orthogonal matching pursuit. J Mach Learn Res 14:1771–1800

Kim J, Pollard DB (1990) Cube root asymptotics. Ann Stat 18:191–219

Lam C, Fan J (2009) Sparsitency and rates of convergence on large covariance matrix estimation. Ann Stat 37:4254–4278

Leng C, Ma S (2007) Accelerated failure time models with nonlinear covariates effects. Aust N Z J Stat 49:155–172

Lian H, Lai P, Liang H (2013) Partially linear structure selection in Cox models with varying coefficients. Biometrics 69:348–357
Liu Y, Zhang J, Zhao X (2018) A new nonparametric screening method for ultrahigh-dimensional survival data. Comput Stat Data Anal 119:74–85
Ma S, Du P (2012) Variable selection in partly linear regression model with diverging dimensions for right censored data. Stat Sin 22:1003–1020
Ma S, Kosorok MR, Fine JP (2006) Additive risk models for survival data with high-dimensional covariates. Biometrics 62:202–210
Newey WK (1994) The asymptotic variance of semiparametric estimators. Econometrica 62:1349–1382
Neykov NM, Filzmoser P, Neytchev PN (2014) Ultrahigh dimensional variable selection through the penalized maximum trimmed likelihood estimator. Stat Pap 55:187–207
Robert J, Gray (1992) Flexible methods for analyzing survival data using splines with applications to breast cancer prognosis. J Am Stat Assoc 8:942–951
Rosenwald A et al (2002) The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. N Engl J Med 346:1937–1947
Schumaker L (1981) Spline functions: basic theory. Wiley, New York
Schwartz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464
Stone C (1986) The dimensionality reduction principle for generalized additive models. Ann Stat 14:590–606
Stute W (1993) Consistent estimation under random censorship when covariables are available. J Multivar Anal 45:89–103
Stute W (1996) Distributional convergence under random censorship when covariables are present. Scand J Stat 23:461–471
Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc B 58:267–288
Tibshirani R (1997) The lasso method for variable selection in the Cox model. Stat Med 16:385–395
van der Vaart A, Wellner JA (1996) Weak convergence and empirical processes. Springer, New York
Wang K, Lin L (2019) Robust and efficient estimator for simultaneous model structure indentification and variable selection in generalized partial linear varying coefficient models with longitudinal data. Stat Pap 60:1649–1676
Wang S, Nan B, Zhu J, David GB (2008) Doubly penalized Buckley-James method for survival data with high-dimensional covariates. Biometrics 64:132–140
Wei LJ (1992) The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. Stat Med 11:1871–1879
Wu Y, Stefanski LA (2015) Automatic structure recovery for additive models. Biometrika 102:381–395
Zeng D, Lin D (2007) Efficient estimation for the accelerated failure time model. J Am Stat Assoc 102:1387–1396
Zhang C (2010) Nearly unbiased variable selection under minimax concave penalty. Ann Stat 38:894–942
Zhang HH, Lu W (2007) Adaptive Lasso for Cox's proportional hazards model. Biometrika 94:691–703
Zhang HH, Cheng G, Liu Y (2011) Linear or nonlinear? Automatic structure discovery for partially linear models. J Am Stat Assoc 106:1099–1112
Zhang J, Yin G, Liu Y, Wu Y (2018) Censored cumulative residual independent screening for ultrahigh-dimensional survival data. Lifetime Data Anal 24:273–292
Zou H (2006) The adaptive Lasso and its oracle properties. J Am Stat Assoc 101:1418–1429