Published in final edited form as:

Stat Med. 2018 February 28; 37(5): 789–800. doi:10.1002/sim.7544.

Sparse Boosting for High-Dimensional Survival Data with Varying Coefficients

Mu Yue^{a,*†}, Jialiang Li^{a,b,c}, and Shuangge Ma^d

^aDepartment of Statistics and Applied Probability, National University of Singapore, Singapore

^bDuke–NUS Graduate Medical School, Singapore

^cSingapore Eye Research Institute, Singapore

^dSchool of Public Health, Yale University, 60 College ST, LEPH 206, 06520, New Haven, CT, USA

Abstract

Motivated by high-throughput profiling studies in biomedical research, variable selection methods have been a focus for biostatisticians. In this paper we consider semiparametric varying-coefficient accelerated failure time models for right censored survival data with high-dimensional covariates. Instead of adopting the traditional regularization approaches, we offer a novel sparse boosting (Sparse L_2 Boosting) algorithm to conduct model-based prediction and variable selection. One main advantage of this new method is that we do not need to perform the time-consuming selection of tuning parameters. Extensive simulations are carried out to examine the performance of our sparse boosting feature selection techniques. We further illustrate our methods using a lung cancer data analysis.

Keywords

boosting; accelerated failure time model; varying-coefficient model; high-dimensional data; minimum description length

1. Introduction

Data science is an ever-expanding field. In the past decades, analysts have generated a great deal of research outcomes for high-dimensional data analysis where the feature dimension p grows exponentially or even non-polynomially fast with sample size n , i.e., $\log(p) = O(n^\alpha)$ for some $\alpha \in (0, 1/2)$. It is well known that including irrelevant features in the statistical model could induce unstable estimation and undesirable computing issues. Variable screening and variable selection techniques have been developed to address the challenges. Among all developments, penalized approaches have been favored by researchers. Familiar penalty functions include the LASSO [1], smoothly clipped absolute deviation (SCAD) [2], MCP [3] and their various extensions [4–6]. Nevertheless, those regularized estimation

*Correspondence to: Mu Yue, Department of Statistics and Applied Probability, National University of Singapore, Singapore.

†E-mail: yuemu.moon@gmail.com

procedures all involve some tuning parameters and need to be chosen by computationally intensive methods such as K-fold cross-validation.

Recently, boosting methods have become very popular for high-dimensional data settings since they can substantially improve the selection accuracy and reduce the probability of selecting irrelevant variables. The original version of boosting algorithm is an ensemble method that iteratively combines weaker learners to minimize the expected loss. Freund and Schapire [7] proposed AdaBoost with the exponential loss function while Bühlmann and Yu [8] and Schmid and Hothorn [9] considered L_2 Boosting with the squared error loss. Although L_2 Boosting is demonstrated to be an effective method to construct a sparse model, it tends to select many ineffective variables. Bühlmann and Yu [10] proposed Sparse L_2 Boosting which replaces the squared loss in L_2 Boosting with the penalized loss function. Bühlmann and Hothorn [11] proposed Twin Boosting, which consists of two rounds of boosting only and Wang [12] proposed HingeBoost to optimize the weighted hinge loss. Hofner et al. [13] investigated the impact of base learner specification on the performance of boosting as a model selection procedure. Boosting for a variety of regression and classification models have been implemented by [14] in R package **mboost**. Boosting algorithms enjoy many advantages such as relatively faster computation speed, lower risk of over-fitting as well as easier adaptations to include additional constraints, e.g., monotonicity, by simple algorithmic adaptations. We will propose a new sparse boosting method for high-dimensional varying-coefficient models in this paper.

Varying-coefficient models include an important class of nonparametric models to accommodate the dynamic covariate effects. They have wide applications in many scientific disciplines, such as epidemiology, medical science, finance, and economics. Parametric models with constant coefficients impose assumptions that are often unrealistic in practice. They are efficient only if being correctly specified, and otherwise can be seriously biased. In contrast, fully nonparametric statistical inference makes little structure assumption on the functional form of the covariate effects and may be rather inefficient. Varying-coefficient models are flexible alternative approaches to parametric and nonparametric models, especially when complicated interaction effects are present. For example, to analyze infant mortality in China, Cheng, Zhang and Chen [15] considered the dynamic impacts varying over time since China made a dramatic change to its child policy since 1949. Another example is [16] where the cure rate and survival rate of the decompression sickness patients both depend on covariates with effects modified by their body mass index. In both examples, the covariate effects cannot be simply regarded as constant, and varying coefficient models lead to more meaningful findings. There are many other examples in the literature of time series, lifetime data and longitudinal data. See [17] for a comprehensive review. We note that there is relatively limited development for high-dimensional varying-coefficient models ([18,19]). To model the varying-coefficients for high-dimensional covariates we usually have to make a group sparsity or fused sparsity assumption ([20]). We will address this problem using a novel boosting approach. In particular, we focus on right censored survival time response in this paper.

Survival time data are also referred to as failure time data or time-to-event data. They are usually subject to censoring. Identifying the risk factors and prediction of survival time can

be quite useful for patient treatment selection, prevention strategy or disease management in evidence-based medicine. A popular regression model is the Cox proportional hazards (PH) model [21] which assumes multiplicative covariate effects on hazard function. An alternative model to Cox model is the accelerated failure time (AFT) model [22] which assumes that the effect of a covariate is to accelerate or decelerate the life time of a disease. The regression coefficients in this model have the direct interpretation of the covariate effects on the mean survival time. Moreover, some authors [23,24] showed that, under special conditions, the AFT model leads to more efficient parameter estimation than the Cox model. See [25] and references therein for more detailed accounts. Penalized estimation has been applied widely in the context of survival analysis. For instance, Tsai [26] adopted the LASSO method to Cox models; Liu and Zeng [27] developed an adaptive LASSO procedure for transformation models with right-censored data; Fan and Li [28] extended their SCAD-penalized likelihood approach to Cox and frailty models; Cai et al. [29] applied the SCAD penalty method to multivariate failure time data; Xia et al. [30] considered low-dimensional confounder adjustment in addition to high-dimensional estimation, among many others. Recently, researchers developed boosting methods to analyze survival data. For example, Schmid and Hothorn [31] proposed a flexible boosting for parametric AFT models, and Wang and Wang [32] proposed Buckley-James Boosting for right censored survival data with high dimensionality.

In this paper, we develop a novel sparse boosting method to fit high-dimensional varying-coefficient models. Particularly, we examine the sparse boosting techniques for right censored survival data. Extending known boosting methods for such a complicated model setting proves to be technically challenging and requires non-trivial development. The detailed development in this paper can contribute significantly for such data analysis. We demonstrate using numerical examples that our methods perform competitively with respect to available approaches.

The rest of the article is arranged as follows. In section 2, we formulate the varying-coefficient AFT model for survival data and propose a detailed Sparse L_2 Boosting algorithm to fit the model. In section 3, we propose a structure identification procedure to identify constant parameters. In section 4, we evaluate the proposed sparse boosting and structure identification techniques through simulation studies. In section 5, the performance of sparse boosting is examined and compared with that of some other variable selection techniques by a real data example. Finally, section 6 provides conclusion and discussion.

2. Sparse Boosting for Survival Data

Before we present the new sparse boosting algorithm, let us first review the generic boosting method. Suppose we have data (X_i, Y_i) , $i = 1, \dots, n$, where $X_i \in \mathbb{R}^p$ denotes a p -dimensional predictor variable and Y_i is a univariate response. We usually assume that (X_i, Y_i) are i.i.d. or from a stationary process. The goal is to predict Y with the regression function f such that $Y = f(X) + \epsilon$, with ϵ being random noise with mean 0 and finite variance. Denote $L(Y, f)$ as the loss function. Boosting is based on minimizing the empirical risk $n^{-1} \sum_{i=1}^n L(Y_i, f(X_i))$

and following iterative steepest descent steps in a function searching space. We present the following generic boosting algorithm [33, 34]:

Generic Boosting Algorithm

-
- a. Initialization. Set $k = 0$ and let $f^{[k]} = \bar{Y}$, i.e. the sample mean.
 - b. Iteration. $k = k+1$. Compute the residuals, defined as the negative gradients of loss function $R_i^{[k]} = -\frac{\partial L(Y_i, f)}{\partial f} \Big|_{f=f^{[k-1]}(X_i)}$ for $i = 1, \dots, n$ and fit a base learner $g(R, X)$ to the current residual R_i with covariates X_i . The fit is denoted by $g^{[k]}(\cdot)$.
 - c. Update. $f^{[k]}(X) = f^{[k-1]}(X) + \nu g^{[k]}(\cdot)$, where $0 < \nu \leq 1$ is the step size.
 - d. Repeat step (b)–(c) until a stopping criterion is satisfied.
-

For L_2 Boosting, the squared loss function $L = (Y_i - f(X_i))^2/2$ is adopted, and the residual in step k is $R_i^{[k]} = Y_i - f^{[k-1]}(X_i)$. Similarly, if we consider nonnegative weight $w = (w_1, \dots, w_n)$, the weighted squares loss function $L = w_i(Y_i - f(X_i))^2/2$ may be adopted, and the residual in iteration k is given as $R_i^{[k]} = w_i(Y_i - f^{[k-1]}(X_i))$.

2.1. Model and Estimation

Let T_i and C_i be the logarithm of survival time and censoring time for the i th subject in a random sample of size n respectively. In practice we observe $Y_i = \min\{T_i, C_i\}$ and the censoring indicator $\delta_i = I(T_i \leq C_i)$ ([35]). Denote $\mathbf{X}_i = (X_{i1}, \dots, X_{i,p-1})$ to be the corresponding $(p-1)$ -dimensional covariates such as biomarkers or gene expressions for the i th subject and U_i to be the univariate index variable. Our observed data set $\{(\mathbf{X}_i, \delta_i, Y_i, U_i) : \mathbf{X}_i \in \mathbb{R}^{p-1}, \delta_i \in \{0, 1\}, Y_i \in \mathbb{R}, U_i \in \mathbb{R}, i = 1, 2, \dots, n\}$ is an independently and identically distributed random sample from $(\mathbf{X}, \delta, Y, U)$. We consider the varying-coefficient AFT model:

$$T_i = \beta_0(U_i) + \sum_{j=1}^{p-1} X_{i,j} \beta_j(U_i) + \varepsilon_i, \quad i = 1, \dots, n, \quad (1)$$

where $\beta_0(\cdot), \beta_1(\cdot), \dots, \beta_{p-1}(\cdot)$ are the unknown varying-coefficient functions of confounder U and ε_i is the random error with $E(\varepsilon_i | \mathbf{X}_i, U_i) = 0$. The distribution of the error is unspecified.

The index variable U is usually chosen by practitioners based on the particular medical applications. For example, [15] considered using calendar time as the index variable and [16] considered using body mass index (BMI) as the index variable. In the lung cancer data analysis of this paper, we will consider age as the index variable as it is well-known that age interacts with many clinical variables for cancer studies ([36]). In practice one may need to seek prior information or consult subject matter experts to determine a proper index variable.

We utilize a weighted least squares estimation approach. Let w_i 's be the Kaplan-Meier weights [37], which are the jumps in the Kaplan-Meier estimator computed as $w_1 = \frac{\delta_{(1)}}{n}$ and $w_i = \frac{\delta_{(i)}}{n - i + 1} \prod_{j=1}^{i-1} \left(\frac{n - j}{n - j + 1} \right)^{\delta_{(j)}}$, $i = 2, \dots, n$. Let $Y_{(1)} \leq \dots \leq Y_{(n)}$ be the order statistics of Y_i 's, $\delta_{(1)}, \dots, \delta_{(n)}$ be the order statistics of δ_i 's, and $U_{(1)} \leq \dots \leq U_{(n)}$ be the order statistics of U_i 's.

$\dots, \delta_{(n)}$ be the associated censoring indicators of the ordered Y_i 's, and $X_{(1),j}, \dots, X_{(n),j}, j=1, \dots, p-1$ and $U_{(1)}, \dots, U_{(n)}$ are defined similarly. The weighed least squares loss function is

$$\sum_{i=1}^n w_i (Y_{(i)} - \beta_0(U_{(i)}) - \sum_{j=1}^{p-1} X_{(i),j} \beta_j(U_{(i)}))^2. \quad (2)$$

The smoothing splines technique is usually used to estimate the unknown functions and it shows very steady performance in practice. We will consider the B-spline basis functions to approximate the unknown coefficient functions $\beta_j(\cdot), j=0, \dots, p-1$ in this paper. Other basis choices include Fourier basis ([38]) and wavelets, among others.

Let $B(\cdot) = (B_1(\cdot), \dots, B_L(\cdot))^T$ be an equal-spaced B-spline basis, where L is the dimension of the basis. Under appropriate smoothness assumptions, i.e., polynomials when restricted to knot intervals and there are continuous $(m-\mu)$ th derivatives at a knot of multiplicity μ , where m is the order of B-spline, the Curry-Schonberg theorem [39] implies that for every smooth function $\beta_j(\cdot)$, we can approximate it by

$$\beta_j(\cdot) \approx B^T(\cdot) \gamma_j, \quad j=0, \dots, p-1, \quad (3)$$

where γ_j is a vector of length L . Then the weighted least squares loss function (2) can be approximated by

$$\sum_{i=1}^n w_i (Y_{(i)} - B^T(U_{(i)}) \gamma_0 - \sum_{j=1}^{p-1} X_{(i),j} B^T(U_{(i)}) \gamma_j)^2. \quad (4)$$

The smoothness conditions are usually satisfied for many functional dependence studies where the underlying curve and its derivatives do not have any unnatural jump or break. Many applications adopt such methods ([40,41]) and assembling polynomial functions in an interval provides a sensible approximation to the functional effects.

Denote by $\tilde{Y} = (Y_{(1)}, \dots, Y_{(n)})^T$, $X_{(i),0} = 1$ for $i=1, \dots, n$, $\tilde{\mathbf{X}}_j = (B(U_{(1)})X_{(1),j}, \dots, B(U_{(n)})X_{(n),j})^T$, $\tilde{\mathbf{X}} = (\tilde{\mathbf{X}}_0, \dots, \tilde{\mathbf{X}}_{p-1})$, $W = \text{diag}(w_1, \dots, w_n)$ and $\gamma = (\gamma_0^T, \dots, \gamma_{p-1}^T)^T$. Then the objective function (4) may be written in the following matrix form:

$$(\tilde{Y} - \tilde{\mathbf{X}}\gamma)^T W (\tilde{Y} - \tilde{\mathbf{X}}\gamma). \quad (5)$$

The estimation thus may produce close-form solution for the spline coefficients when p is small or moderate. With high-dimensional p the solution cannot be easily attained. Let

$\gamma^{[\hat{K}]} = ((\gamma_0^{[\hat{K}]})^T, \dots, (\gamma_{p-1}^{[\hat{K}]})^T)^T$ be the estimator of γ using the sparse boosting approach with weighted square loss function (5), where $\hat{K}$ is the estimated stopping iterations. Then the coefficient function estimates are given by

$$\hat{\beta}_j(u) = B^T(u) \gamma_j^{[\hat{K}]}, \quad j=0, \dots, p-1. \quad (6)$$

Our practical goal is to identify which genes in the high-dimensional microarray data sets are cancer-associated as well as to estimate the corresponding coefficient functions in an AFT model. Instead of using the penalized estimation approach, we propose in the following subsection Sparse L_2 Boosting techniques to estimate $\gamma^{[K]}$.

2.2. Simultaneous Fitting and Selection by Sparse L_2 Boosting Techniques

The main idea of Sparse L_2 Boosting is to replace the empirical risk function in L_2 Boosting with the penalized empirical risk function which is a combination of squared loss and a measure of complexity of boosting, and then iteratively perform gradient descent in a function space. Thus Sparse L_2 Boosting produces sparser models than L_2 Boosting. The g-prior minimum description length (gMDL) proposed by [42] can be used as the penalized empirical risk function and also to estimate the stopping criterion. The gMDL takes the form:

$$\text{gMDL}(RSS, \text{trace}(\mathcal{B})) = \log(S) + \frac{\text{trace}(\mathcal{B})}{n} \log\left(\frac{\tilde{Y}^T \tilde{Y} - RSS}{\text{trace}(\mathcal{B}) \times S}\right), \quad S = \frac{RSS}{n - \text{trace}(\mathcal{B})}. \quad (7)$$

Here RSS is the residual sum of squares and $\mathcal{B}$ is the boosting operator. Bühlmann [43] suggested to use the trace of the boosting operator as the degree of freedom in the context of boosting and as the indicator of the complexity of model. The principle of gMDL is to choose the model that gives the shortest description of data. The advantage is that it does not involve any tuning parameter since it has a data-dependent penalty for each dimension and is explicitly given as a function of data only.

We describe the Sparse L_2 Boosting procedure in details. The initial value of γ is set to be a zero vector, i.e. $\gamma^{[k]} = \mathbf{0}$ for $k=0$, while in each of the k th iteration ($1 \leq k \leq K$ for K being the total number of iterations) we only use the current residual $R^{[k]} = \tilde{Y} - \tilde{\mathbf{X}} \gamma^{[k-1]}$ to regress each of the j th working component $\tilde{\mathbf{X}}_j, j=0, \dots, p-1$. The fit denoted by $\hat{\lambda}_j^{[k]}$ can be obtained by minimizing the weighted squared loss function $(R^{[k]} - \tilde{\mathbf{X}}_j \lambda)^T W (R^{[k]} - \tilde{\mathbf{X}}_j \lambda)$ with respect to λ . Hence the weighted least squared estimate is $\hat{\lambda}_j^{[k]} = [(\tilde{\mathbf{X}}_j)^T W (\tilde{\mathbf{X}}_j)]^{-1} (\tilde{\mathbf{X}}_j)^T W R^{[k]}$, the corresponding hat matrix is $\mathcal{H}_j = (\tilde{\mathbf{X}}_j) [(\tilde{\mathbf{X}}_j)^T W (\tilde{\mathbf{X}}_j)]^{-1} (\tilde{\mathbf{X}}_j)^T W$ and the weighted residual sum of squares is $RSS_j^{[k]} = (R^{[k]} - \tilde{\mathbf{X}}_j \hat{\lambda}_j^{[k]})^T W (R^{[k]} - \tilde{\mathbf{X}}_j \hat{\lambda}_j^{[k]})$. The selected component $\hat{s}_k$ can be obtained by:

$$\hat{s}_k = \operatorname{argmin}_{0 \leq j \leq p-1} \text{gMDL}(RSS_j^{[k]}, \text{trace}(\mathcal{B}_j^{[k]})), \quad (8)$$

where $\mathcal{B}_j^{[1]} = \mathcal{H}_j$ and $\mathcal{B}_j^{[k]} = I - (I - \mathcal{H}_j)(I - \nu \mathcal{H}_{\hat{s}_{k-1}}) \cdots (I - \nu \mathcal{H}_{\hat{s}_1})$ for $k > 1$ is the boosting operator for selecting j th component in the k th iteration. Therefore, at each iteration there is only one working component $\hat{\mathbf{X}}_{\hat{s}_k}$ to be chosen, and only the corresponding coefficient vector $\gamma_{\hat{s}_k}^{[k]}$ changes, i.e. $\gamma_{\hat{s}_k}^{[k]} = \gamma_{\hat{s}_k}^{[k-1]} + \nu \hat{\lambda}_{\hat{s}_k}^{[k]}$, where ν is the step size, while all the other $\gamma_j^{[k]}$ for $j \neq \hat{s}_k$ remain the same. We continue this process for K iterations and estimate the stopping iteration K by

$$\hat{K} = \operatorname{argmin}_{1 \leq k \leq K} \text{gMDL}(RSS_{\hat{s}_k}^{[k]}, \text{trace}(\mathcal{B}^{[k]})), \quad (9)$$

where $\mathcal{B}^{[k]} = I - (I - \nu \mathcal{H}_{\hat{s}_k}) \cdots (I - \nu \mathcal{H}_{\hat{s}_1})$.

From this Sparse L_2 Boosting procedure, we obtain the estimator of γ as

$\gamma^{[\hat{K}]} = ((\gamma_0^{[\hat{K}]})^T, \dots, (\gamma_{p-1}^{[\hat{K}]})^T)^T$. Now we can present the Sparse L_2 Boosting algorithm for the varying-coefficient AFT model as follows.

Sparse L_2 Boosting Algorithm for Varying-Coefficient AFT Model

- a. Initialization. Set $k = 0$ and $\gamma_0^{[k]} = \mathbf{0}, \dots, \gamma_{p-1}^{[k]} = \mathbf{0}$ (component-wise).
- b. Iteration. $k = k+1$. Compute $\hat{s}_k = \operatorname{argmin}_{0 \leq j \leq p-1} \text{gMDL}(RSS_j^{[k]}, \text{trace}(\mathcal{B}_j^{[k]}))$, where $\mathcal{B}_j^{[1]} = \mathcal{H}_j$ and $\mathcal{B}_j^{[k]} = I - (I - \mathcal{H}_j)(I - \nu \mathcal{H}_{\hat{s}_{k-1}}) \cdots (I - \nu \mathcal{H}_{\hat{s}_1})$ for $k > 1$.
- c. Update. $\gamma_{\hat{s}_k}^{[k]} = \gamma_{\hat{s}_k}^{[k-1]}$ for $j \neq \hat{s}_k$ and $\gamma_{\hat{s}_k}^{[k]} = \gamma_{\hat{s}_k}^{[k-1]} + \nu \hat{\lambda}_{\hat{s}_k}^{[k]}$, where ν is the step size.
- d. Iteration. Repeat step (b)–(c) for K iterations.
- e. Stopping. Estimate $\hat{K} = \operatorname{argmin}_{1 \leq k \leq K} \text{gMDL}(RSS_{\hat{s}_k}^{[k]}, \text{trace}(\mathcal{B}^{[k]}))$, where $\mathcal{B}^{[k]} = I - (I - \nu \mathcal{H}_{\hat{s}_k}) \cdots (I - \nu \mathcal{H}_{\hat{s}_1})$. Thus, $\gamma^{[\hat{K}]} = ((\gamma_0^{[\hat{K}]})^T, \dots, (\gamma_{p-1}^{[\hat{K}]})^T)^T$ is the estimator for γ and $\hat{\beta}_j(u) = B^T(u) \gamma_j^{[\hat{K}]}$, $j = 0, \dots, p-1$ are the varying coefficient estimates. The final estimate of $\tilde{Y}$ is $\hat{Y}^{[\hat{K}]} = \hat{\mathbf{X}} \gamma^{[\hat{K}]}$.

As suggested by [10] and references therein, the choice of step size ν is not critical as long as it is small. A smaller value of ν requires a larger number of boosting iterations and more computing time, while the predictive accuracy has been found to be potentially better for a smaller ν . We set $\nu = 0.1$ as suggested by [34] in this paper.

3. Identification of Model Structure

The Sparse L_2 Boosting technique enables us to select important covariates with a high probability. After variable selection, we need to decide which coefficients are indeed varying and which are constant. Identifying nonzero constant coefficients is important since treating a constant coefficient as varying will yield a slower convergence rate and less efficient estimation results. Consider the partly linear semivarying-coefficient model

$$T_i = \beta_0(U_i) + \sum_{j=1}^{p_1} \alpha_j X_{i,j} + \sum_{d=1}^{p_2} \beta_d(U_i) X_{i,p_1+d} + \varepsilon_i, \quad i=1, \dots, n, \quad (10)$$

where $\mathbf{X}_i = (X_{i,1}, \dots, X_{i,p_1+p_2})$ are the $(p_1 + p_2)$ -dimensional covariates for individual i , $i = 1, \dots, n$. $\alpha_1, \dots, \alpha_{p_1}$ are the constant coefficients, $\beta_0(\cdot), \beta_1(\cdot), \dots, \beta_{p_2}(\cdot)$ are the varying-coefficient functions of confounder and ε_i is the random error with $E(\varepsilon_i | \mathbf{X}_i, U_i) = 0$. If we use L -dimensional B-spline $B(\cdot) = (B_1(\cdot), \dots, B_L(\cdot))^T$ to approximate $\beta_d(\cdot)$, i.e., $\beta_d(\cdot) \approx B^T(\cdot) \gamma_d$ where $\gamma_d = (\gamma_{d1}, \dots, \gamma_{dL})^T$ is a vector of length L , then the problem of estimating $(\alpha_1, \dots, \alpha_{p_1}, \beta_0(\cdot), \beta_1(\cdot), \dots, \beta_{p_2}(\cdot))$ in (10) reduces to estimating $(\alpha_1, \dots, \alpha_{p_1}, \gamma_0, \gamma_1, \dots, \gamma_{p_2})$ in the linear model

$$T_i = B^T(U_i) \gamma_0 + \sum_{j=1}^{p_1} \alpha_j X_{i,j} + \sum_{d=1}^{p_2} B^T(U_i) \gamma_d X_{i,p_1+d} + \varepsilon_i, \quad i=1, \dots, n. \quad (11)$$

As suggested by [44], the AIC [45] criterion for model (11) is

$$AIC = -2 \log \text{-likelihood} + 2p, \quad (12)$$

where $p = p_1 + (p_2 + 1) \times L$ is the number of parameters to be estimated. Then we propose the following iterative procedure similar to [15] to identify the constant parameters. We start with the varying-coefficient model $\mathcal{M}_0$ (i.e., all coefficients are varying) as the candidate model and at the k th step of the iteration we examine whether one of the functional coefficients in the candidate model $\mathcal{M}_k$ can be further reduced to a constant.

Model Structure Identification Algorithm

-
- a. Initialization. Set $k = 0$ and calculate $AIC(\mathcal{M}_0)$.
 - b. Iteration. For each of the unknown functions in the candidate model $\mathcal{M}_k$, say $\beta_d(\cdot)$, that could be reduced to a constant, calculate

$$S_d = \sum_{i=1}^n (\hat{\beta}_d(U_i) - \bar{\beta}_d)^2, \quad \bar{\beta}_d = n^{-1} \sum_{i=1}^n \hat{\beta}_d(U_i).$$

Changing the function $\beta_d(\cdot)$ in $\mathcal{M}_k$ that has the smallest S_d to a constant parameter results in a new model $\mathcal{M}_{k+1}$.

- c. Update. Compute $\text{AIC}(\mathcal{M}_{k+1})$. If $\text{AIC}(\mathcal{M}_k) < \text{AIC}(\mathcal{M}_{k+1})$, then $\mathcal{M}_k$ is the chosen model, and the model selection is completed. Otherwise, $\mathcal{M}_{k+1}$ becomes the candidate model; $k = k+1$.
- d. Iteration. Repeat step (b)–(c) until $\text{AIC}(\mathcal{M}_k) < \text{AIC}(\mathcal{M}_{k+1})$ for some $k < p_1 + p_2 + 1$ (i.e., $\mathcal{M}_k$ is the chosen model) or until $k = p_1 + p_2 + 1$ (i.e., all of the parameters are reduced to constants in $\mathcal{M}_k$), then $\mathcal{M}_{p_1+p_2+1}$ is the chosen model, and model selection is completed.

4. Simulation

In this section, we illustrate the proposed Sparse L_2 Boosting algorithm by studying their performance on simulated data. We compare (a) L_2 Boosting with (b) Sparse L_2 Boosting in their performance of variable selection and function estimation. Method (b) is what we propose in this paper while Method (a) is a relatively simpler version and may not yield sparse solution in general. In both simulation and real data analysis, we adopt cubic B-spline basis with 3 equally-spaced inner knots and without intercept, thus the dimension of the B-spline basis is 6. All results are based on 500 Monte Carlo simulations.

We consider the following true model:

$$T_i = \beta_0(U_i) + \sum_{j=1}^5 X_{i,j} \beta_j(U_i) + \varepsilon_i, \quad i = 1, \dots, n,$$

where $\mathbf{X}_i = (X_{i,1}, \dots, X_{i,p-1})$, $i = 1, \dots, n$ are i.i.d. from the multivariate normal distribution with mean $\mathbf{0}_{p-1}$, and autoregression correlation $\text{Corr}(X_{i,j}, X_{i,k}) = \rho^{|j-k|}$, where ρ is set to be 0.2 and 0.8 respectively. The error term ε_i 's are generated independently from standard normal distribution thus the survival time follows a log-normal distribution. The logarithm of censoring time C_i 's are generated independently from normal distribution with mean 3 and variance 1 such that the censoring rate is about 30%. The confounder U_i 's are generated independently from uniform distribution on $[0,1]$. We consider $n = 400$ with $p = 20,000$. The sample size and number of covariates considered here are comparable to the most medical studies where usually tens of thousands gene expressions with hundreds of subjects are screened for a particular disease.

In face of ultra-high dimensional data, pre-screening is necessary to reduce the dimensionality to a moderate size. In particular, we adopt marginal screening [46] to screen out irrelevant covariates. More specifically, we regress the logarithm of survival time on each covariate to construct the marginal AFT model $T_i = \beta_0 + \beta_j X_{i,j} + e_j$, $j = 1, \dots, p-1$. Assuming that error term follows the normal distribution, the Wald P -value for testing null hypothesis $\beta_j = 0$ can be obtained by **survreg** function in R. After the sorting 20,000 P -value in an ascending order, the top markers in the ranked list reflect the strongest covariate effects on the mean survival time. Following the recommendation by [46] and other references [18], [47], [48], we keep the top $\left\lfloor \frac{n}{\log(n)} \right\rfloor$ variables in the ranked list and screen out the rest of the covariates. Thus, the first 66 variables in the ranked list are selected to conduct downstream

analysis. Thereafter, we proceed to use variable selection approaches to build up the final parsimonious model. In our implementation, the maximum number of boosting iterations is limited to $K = 500$. We investigate 5 scenarios for coefficient functions:

Case I Let

$$\begin{aligned}\beta_0(u) &= (1-4u^2)I_{\{u \leq 0.5\}} + (\frac{4}{3}u^2 - \frac{1}{3})I_{\{u > 0.5\}}, \beta_1(u) = (1-4u^2)I_{\{u \leq 0.5\}} + (\frac{4}{3}u^2 - \frac{1}{3})I_{\{u > 0.5\}} \\ \beta_2(u) &= (1-2u)I_{\{u \leq 0.5\}} + (2u-1)I_{\{u > 0.5\}}, \\ \beta_3(u) &= (1-4u^2)I_{\{u \leq 0.5\}} + (\frac{4}{3}u^2 - \frac{1}{3})I_{\{u > 0.5\}}, \beta_4(u) = (1-4u^2)I_{\{u \leq 0.5\}} + (2u-1)I_{\{u > 0.5\}}, \\ \beta_5(u) &= (1-4u^2)I_{\{u \leq 0.5\}} + (\frac{4}{3}u^2 - \frac{1}{3})I_{\{u > 0.5\}}.\end{aligned}$$

Case II Let $\beta_0(u) = 1.5, \beta_1(u) = 1.2, \beta_2(u) = 1, \beta_3(u) = 1.5, \beta_4(u) = 1.2, \beta_5(u) = 1$.

Case III Let $\beta_0(u) = -1.5, \beta_1(u) = 1.2, \beta_2(u) = (3u-1)^2, \beta_3(u) = 2\exp(-(3u-1)^2) + \exp(-4(u-3)^2), \beta_4(u) = 0.2\sin(2\pi u) + 0.2\cos(2\pi u) + 0.3\sin(2\pi u)^2 + 0.4\cos(2\pi u)^2 + 0.5\sin(2\pi u)^3, \beta_5(u) = 3\exp(-4u)$.

Case IV Let $\beta_0(u) = -1.5, \beta_1(u) = 1, \beta_2(u) = 2u, \beta_3(u) = 3\exp(-4u), \beta_4(u) = 3u_3 + 0.5, \beta_5(u) = 3u^2$.

Case V Let $\beta_0(u) = 2u, \beta_1(u) = 3\exp(-4u), \beta_2(u) = 3u^3 + 0.5u, \beta_3(u) = 2\exp(-(3u-1)^2) + \exp(-4(u-3)^2), \beta_4(u) = (3u-1)^2, \beta_5(u) = 1.5u$.

We design the settings for various scenarios such that for case I, all of the coefficient functions are varying but none of them are smooth; for case II, all of the coefficients are nonzero constant; for cases III and IV, the first two coefficients are constant while the remaining are varying with u and for case V, all of the coefficients are varying with u . We assess the performance of variable screening, variable selection and structure identification by reporting the following quantities over 500 replications in Table 1 :

$\mathcal{S}$: coverage probability that the top $\lfloor \frac{n}{\log(n)} \rfloor$ covariates after screening includes all important covariates;

TP: the median of true positives;

FP: the median of false positives;

Size: the median of model sizes;

ISPE: the average of in-sample prediction errors defined as

$$\sum_{i=1}^n \delta_i (Y_i - Y_i^{[\hat{K}]})^2 / \sum_{i=1}^n \delta_i;$$

RMISE: the average of root mean integrated squared errors defined as

$$\sqrt{\frac{1}{n} \sum_{j=0}^5 \sum_{i=1}^n (\beta_j(u_i) - \hat{\beta}_j(u_i))^2};$$

C_{var} : the mean proportion of variables with varying coefficients that are selected;

C_{fix} : the mean proportion of variables with constant nonzero coefficients that are selected.

From Table 1, it follows that the screener perform quite well with the coverage probability close to one. Both L_2 Boosting and Sparse L_2 Boosting produce true positives equal to the true

number of important variables while Sparse L_2 Boosting produces much fewer false positives than L_2 Boosting. Although ISPE using L_2 Boosting is a little bit smaller than using Sparse L_2 Boosting due to the fact that the former involves much more irrelevant covariates, RMISE using Sparse L_2 Boosting is much smaller than that using L_2 Boosting. Moreover, the results for low correlation are superior to high correlation. When the coefficient functions violate the smoothness assumption in Curry-Schonberg theorem (case I), the variable selection performance remains good. In general, Sparse L_2 Boosting outperforms L_2 Boosting in terms of variable selection and parameter estimation. Last but not least, through the structure identification procedure, the nonzero constant coefficients can be identified with a high probability under most cases, especially using the covariates selected by Sparse L_2 Boosting.

5. Lung Cancer Data Analysis

Lung cancer is the prime cancer killer for people in the U.S. About one quarter cancer deaths are from lung cancer. Every year, more people die from lung cancer than from colon, breast, and prostate cancers combined [49]. Analyzing gene expressions in lung cancer and identification of relevant gene expressions are important for treatment and prevention.

Our example comes from a large, training-testing, multi-site blinded validation study [50] with 442 lung adenocarcinomas. Gene expression data are generated by four different laboratories under a common protocol and possess high quality. Xie et al. [51] and Li et al. [52] studied the same data set. In this analysis, we use age as the potential confounder, since it is strongly associated with survival time [53]. After removing missing confounder or measurements in overall survival, a total of 439 patients are included in downstream analysis. The median follow-up time is 46 months (range: 0.03 to 204 months) and the median age at diagnosis is 65 years (range: 33 to 87 years). The overall censoring rate is 46.47 %. For each subject, 22,283 gene expressions are available.

To deal with ultra-high dimensionality, we first adopt a marginal screening procedure to screen out irrelevant genes. After sorting 22,283 P -values in an ascending order, the first

$\left\lfloor \frac{n}{\log(n)} \right\rfloor = 72$ gene expressions in this list are kept to carry out downstream analysis.

Thereafter, we proceed to use variable selection approaches to select important genes associated with lung cancer.

For the purpose of comparison, in addition to L_2 Boosting and the proposed Sparse L_2 Boosting, we also consider some other existing variable selection approaches for constant-coefficient AFT models: Buckley-James Boosting with linear least squares (BJ-LS) [32], Buckley-James Twin Boosting with linear least squares (Twin BJ-LS) [32], Buckley-James regression with elastic net penalty (BJ-EN) [54] and Buckley-James regression with SCAD penalty (BJ-SCAD).

In our implementation, the maximum number of boosting iterations in all cases are set to 500. Except L_2 Boosting and Sparse L_2 Boosting, all other methods are executed using R package **bujar** written by Zhu Wang based on [32] and the tuning parameters are chosen by 5-fold cross-validation. The results of variable selection, estimation and prediction are

summarized in Table 2. In addition to the number of selected genes, we also report in-sample prediction error (ISPE) $\sum_{i=1}^n \delta_i (Y_i - \hat{Y}_i)^2 / \sum_{i=1}^n \delta_i$ and out-of-sample prediction error (OSPE) measured by 5-fold cross-validation. We observe from Table 2 that L_2 Boosting and Sparse L_2 Boosting for varying-coefficient AFT model not only possess smaller ISPE and OSPE, but also produce relatively sparser model than the four methods for constant-coefficient AFT model. Again, Sparse L_2 Boosting produce sparser model than L_2 Boosting. In summary, compared with varying-coefficient and constant-coefficient AFT model, we observe that including age in the varying coefficients could lead to more accurate estimate and our proposed Sparse L_2 Boosting method for varying-coefficient AFT model performs very well in terms of sparsity, estimation and prediction.

Moreover, we also present the Venn diagram showing overlapping genes identified by all different methods in Figure 1. We observe that L_2 Boosting and Sparse L_2 Boosting select 4 genes in common and the other four Buckley-James methods have 10 genes in common. However, L_2 Boosting and Sparse L_2 Boosting produce distinct list of selected genes from the other Buckley-James methods. There are only 2 genes selected by all of the methods.

Applying Sparse L_2 Boosting in the varying-coefficient AFT model, 11 genes are selected: PIK3R1, BSDC1, MAT2A, GPR116, BMPY, AGFG1, ETV5, STEAP1, RFC5, PGK1 and FAM117A. To implement the proposed structure identification procedure, we fit the varying-coefficient AFT model with normal error term and identify that genes STEAP1, ETV5, PGK1, PIK3R1, MAT2A, BMPY, RFC5, and AGFG1 have constant coefficients and the other 3 genes have varying coefficients. Finally, we fit the semi-varying coefficient AFT model and obtain the refined estimators with the standard errors for the eight constant coefficients: -0.071 (0.145) for STEAP1, 0.367 (0.273) for ETV5, -0.171 (0.364) for PGK1, 0.356 (0.250) for PIK3R1, 0.049 (0.324) for MAT2A, 0.264 (0.557) for BMPY, -0.009 (0.254) for RFC5 and 0.087 (0.274) for AGFG1. We plot the estimated curves of varying coefficients for three genes with their 95% confidence bands constructed by 500 bootstrap resamples in Figure 2. All of the functions are quite different from a straight line and this suggest that the varying-coefficient model is more appropriate to describe the covariate effects on survival time in our data. Using more sophisticated semiparametric model specification may provide more accurate model for the high-dimensional analysis. Furthermore, we observe different functional forms for the three genes. However, it is interesting that the overall patterns are similar, with an U (or reversed U) shape. The “interactions” between age and gene expressions in cancer prognosis have been reported in multiple studies. The exact mechanisms for the three identified genes will need to be examined by functional experiments. We observe a small “bump” around age=70 for gene BSDC1. With a limited sample size, such a subtle structure needs to be taken with cautions.

Our literature search suggests that the identified genes may have important biological implications. For example, gene PIK3R1 belongs to the PI3K pathway, whose deregulation is frequent in tumor cells. The underexpression of PIK3R1 has been identified as an independent prognostic marker in multiple cancers. The protein encoded by gene MAT2A catalyzes the production of S-adenosylmethionine (AdoMet) from methionine and ATP, with AdoMet being the key methyl donor in cellular processes. The dysregulation of gene

MAT2A has been observed in multiple cancer types. The analysis of GEO (Gene Expression Omnibus) and other data has suggested gene GPR116 as a regulator of cancer metastasis, which is achieved by driving EMT in cells through the AKT/ERK signaling pathway. The dysregulation of this gene has been associated with a poorer prognosis of breast cancer, colorectal carcinoma, and others. The protein encoded by gene AGFG1 is related to nucleoporins, a class of proteins that mediate nucleocytoplasmic transport, which is required for the nuclear export of Rev-directed RNAs. Gene ETV5 has been extensively studied. Altered expression of its encoded proteins has been identified in lung adenocarcinoma. Its gene expression has been identified as an independent adverse prognostic factor for overall survival of breast cancer and lymphoma patients. Gene STEAP has been observed as expressed in common malignancies including prostate, bladder, colon, and ovarian carcinomas, and in Ewing's sarcoma, suggesting that may function as an almost universal tumor antigen. Gene RFC5 encodes the smallest subunit of the replication factor C complex, which consists of five subunits and is required for DNA replication, a fundamental mechanism in cell function. The protein encoded by gene PGK1 is secreted by tumor cells where it participates in angiogenesis by functioning to reduce disulfide bonds in the serine protease, plasmin, which consequently leads to the release of the tumor blood vessel inhibitor angiostatin.

6. Discussion

In this paper, we proposed the sparse boosting procedure for nonparametric varying-coefficient models that can simultaneously perform variable selection and estimation of functional coefficients. Our methods are illustrated by analyzing right censored survival data. Numerical studies indicate that sparse boosting is effective in selecting important covariates and estimating the smooth coefficient functions. Also, sparse boosting has stronger ability to exclude redundant variables and produce sparser model than L_2 Boosting.

In the above construction, we adopted B-spline basis functions to approximate the varying coefficients due to their success in interpolating polynomials and their usefulness in providing a smooth approximation to a covariate function of unspecified form. Numerical studies showed that it works well even if the smoothness assumption is violated. However, it should be noted that other types of basis functions can be employed to approximate the coefficient functions as well. Also, although the models and procedure considered in this paper are demonstrated by analyzing the AFT model with right censored survival data, the proposed sparse boosting procedure can be easily extended to other regression models such as the Cox model, model for longitudinal data and logistic model with varying coefficients.

The method for varying coefficients is limited to a scalar confounder U . Some applications may involve a group of index variables such as age, gender and other low-dimensional clinical variables. It is not a trivial task to smooth more than one confounder in the varying-coefficient functions due to the curse of dimensionality. In order to solve this problem, one may need to assume certain functional structure for the multiple confounder effects. Model selection and model validation are necessary for such a goal.

In summary, we proposed a sparse boosting procedure for variable selection in varying-coefficient models and demonstrated its usefulness by analyzing right censored survival data. We particularly considered the application of such a sparse boosting procedure to selection of varying coefficient in high-dimensional data setting. Results from analysis of the lung cancer data set show that the procedure can effectively select important genes associated with lung cancer. However, careful study of the theoretical properties of this procedure is left for future research.

Acknowledgments

We thank the editor and reviewers for their careful review and insightful comments. This study has been partly supported by awards CA204120 and CA191383 from NIH and Academic Research Funding R-155-000-174-114 in Singapore.

References

1. Robert, Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*. 1996;267–288.
2. Jianqing, Fan, Runze, Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*. 2001; 96(456):1348–1360.
3. Cun-Hui, Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*. 2010:894–942.
4. Hui, Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*. 2006; 101(476):1418–1429.
5. Hui, Zou, Trevor, Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2005; 67(2):301–320.
6. Ming, Yuan, Yi, Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2006; 68(1):49–67.
7. Yoav, Freund, Schapire Robert, E. A decision-theoretic generalization of on-line learning and an application to boosting. Springer; 1995. p. 23–37.
8. Peter, Bühlmann, Bin, Yu. Boosting with the L 2 loss: regression and classification. *Journal of the American Statistical Association*. 2003; 98(462):324–339.
9. Matthias, Schmid, Torsten, Hothorn. Boosting additive models using component-wise P-splines. *Computational Statistics & Data Analysis*. 2008; 53(2):298–311.
10. Peter, Bühlmann, Bin, Yu. Sparse boosting. *Journal of Machine Learning Research*. 2006; 7(Jun): 1001–1024.
11. Peter, Bühlmann, Torsten, Hothorn. Twin boosting: improved feature selection and prediction. *Statistics and Computing*. 2010; 20(2):119–138.
12. Zhu, Wang. HingeBoost: ROC-based boost for classification and variable selection. *The International Journal of Biostatistics*. 2011; 7(1):1–30.
13. Benjamin, Hofner, Torsten, Hothorn, Thomas, Kneib, Matthias, Schmid. A framework for unbiased model selection based on boosting. *Journal of Computational and Graphical Statistics*. 2011; 20(4):956–971.
14. Torsten, Hothorn, Peter, Buehlmann, Thomas, Kneib, Matthias, Schmid, Benjamin, Hofner. mboost: Model-based boosting. R package version. 2012:2–1.
15. Ming-Yen, Cheng, Wenyang, Zhang, Lu-Hung, Chen. Statistical estimation in generalized multiparameter likelihood models. *Journal of the American Statistical Association*. 2009; 104(487):1179–1191.
16. Shao F, Li J, Ma S, Lee M-LT. Semiparametric Varying-coefficient Model for Interval Censored Data with a Cured proportion. *Statistics in Medicine*. 2014; 33:1700–1712. [PubMed: 24302535]
17. Jianqing, Fan, Wenyang, Zhang. Statistical methods with varying coefficient models. *Statistics and its Interface*. 2008; 1(1):179. [PubMed: 18978950]

18. Ming-Yen, Cheng, Toshio, Honda, Jialiang, Li, Heng, Peng. Nonparametric independence screening and structure identification for ultra-high dimensional longitudinal data. *The Annals of Statistics*. 2014; 42(5):1819–1849.
19. Xia X, Yang H, Li J. Feature screening for generalized varying coefficient models with application to dichotomous responses. *Computational Statistics and Data Analysis*. 2016; 102:85–97.
20. Giraud, C. *Introduction to High-dimensional Statistics*. CRC Press; 2015.
21. David Cox R. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society*. 1972; 34:187–220.
22. Lee-Jen, Wei. The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. *Statistics in medicine*. 1992; 11(14–15):1871–1879. [PubMed: 1480879]
23. Bradley, Efron. The efficiency of Cox's likelihood function for censored data. *Journal of the American statistical Association*. 1977; 72(359):557–565.
24. David, Oakes. The asymptotic information in censored survival data. *Biometrika*. 1977:441–448.
25. Miller Rupert, G, Jr. *Survival analysis*. John Wiley & Sons; 2011.
26. Yann, Tsai Wei. Pseudo-partial likelihood for proportional hazards models with biased-sampling data. *Biometrika*. 2009; 96(3):601–615. [PubMed: 22422175]
27. Xiaoxi, Liu, Donglin, Zeng. Variable selection in semiparametric transformation models for right-censored data. *Biometrika*. 2013; 100(4):859–876.
28. Jianqing, Fan, Runze, Li. Variable selection for Cox's proportional hazards model and frailty model. *Annals of Statistics*. 2002:74–99.
29. Jianwen, Cai, Jianqing, Fan, Runze, Li, Haibo, Zhou. Variable selection for multivariate failure time data. *Biometrika*. 2005; 92(2):303–316. [PubMed: 19458784]
30. Xiaochao, Xia, Binyan, Jiang, Jialiang, Li, Wenyang, Zhang. Low-dimensional confounder adjustment and high-dimensional penalized estimation for survival analysis. *Lifetime data analysis*. 2016; 22:547–569. [PubMed: 26463818]
31. Matthias, Schmid, Torsten, Hothorn. Flexible boosting of accelerated failure time models. *BMC bioinformatics*. 2008; 9(1):269. [PubMed: 18538026]
32. Zhu, Wang, Wang, CY., et al. Buckley-James boosting for survival analysis with high-dimensional biomarker data. *Statistical Applications in Genetics and Molecular Biology*. 2010; 9(1):24.
33. Jerome, Friedman, Trevor, Hastie, Robert, Tibshirani, et al. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*. 2000; 28(2):337–407.
34. Friedman Jerome H. Greedy function approximation: a gradient boosting machine. *Annals of statistics*. 2001:1189–1232.
35. Li, J., Ma, S. *Survival Analysis in Medicine and Genetics*. CRC Press; 2013.
36. Leon, Gordis. *Epidemiology* Saunders. 2009.
37. Winfried, Stute. Consistent estimation under random censorship when covariables are present. *Journal of Multivariate Analysis*. 1993; 45(1):89–103.
38. Xu M, Li J, Chen Y. Varying coefficient functional autoregressive model with application to the US Treasuries. *Journal of Multivariate Analysis*. 2017; 159:169–183.
39. Curry Haskell B, Schoenberg Isaac J. On Pólya frequency functions IV: the fundamental spline functions and their limits. *Journal d'analyse mathématique*. 1966; 17(1):71–107.
40. Ming-Yen, Cheng, Honda, T., Li, J., Peng, H. Nonparametric independence screening and structure identification for ultra-high dimensional longitudinal data. *The Annals of Statistics*. 2014; 42(5): 1819–1849.
41. Cheng M, Honda T, Li J. Efficient estimation in semivarying coefficient models for longitudinal/ clustered data. *The Annals of Statistics*. 2016; 44(5):1988–2017.
42. Hansen Mark H, Bin Yu. Model selection and the principle of minimum description length. *Journal of the American Statistical Association*. 2001; 96(454):746–774.
43. Peter, Buehlmann. Boosting for high-dimensional linear models. *The Annals of Statistics*. 2006:559–583.
44. Huang Jianhua Z, Haipeng Shen. Functional Coefficient Regression Models for Non-linear Time Series: A Polynomial Spline Approach. *Scandinavian journal of statistics*. 2004; 31(4):515–534.

45. Hirotugu, Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*. 1974; 19(6):716–723.
46. Jianqing, Fan, Jinchi, Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2008; 70(5):849–911. [PubMed: 19603084]
47. Yue M, Li J. Improvement Screening for Ultra-High Dimensional Data with Censored Survival Outcomes and Varying Coefficients. *International Journal of Biostatistics*. 2017
48. Xia X, Li J, Fu B. Conditional quantile correlation learning for ultrahigh dimensional varying coefficient models and its application in survival analysis. *Statistica Sinica*. 2017
49. Yan, Lu, William, Lemon, Peng-Yuan, Liu, et al. A gene expression signature predicts survival of patients with stage I non-small cell lung cancer. *PLoS Med*. 2006; 3(12):e467. [PubMed: 17194181]
50. Kerby, Shedden, Taylor Jeremy, MG., Enkemann Steven, A., et al. Gene expression–based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nature medicine*. 2008; 14(8):822–827.
51. Yang, Xie, Guanghua, Xiao, Coombes Kevin, R., et al. Robust gene expression signature from formalin-fixed paraffin-embedded samples predicts prognosis of non–small-cell lung cancer patients. *Clinical Cancer Research*. 2011; 17(17):5705–5714. [PubMed: 21742808]
52. Jialiang, Li, Qi, Zheng, Limin, Peng, Zhipeng, Huang. Survival impact index and ultrahigh-dimensional model-free screening with survival outcomes. *Biometrics*. 2016
53. Dario, Consonni, Alberto, Bertazzi Pier, Carlo, Zocchetti. Why and how to control for age in occupational epidemiology. *Occupational and environmental medicine*. 1997; 54(11):772–776. [PubMed: 9538347]
54. Sijian, Wang, Bin, Nan, Ji, Zhu, Beer David, G. Doubly Penalized Buckley–James Method for Survival Data with High-Dimensional Covariates. *Biometrics*. 2008; 64(1):132–140. [PubMed: 17680828]

Figure 1. Venn diagram showing overlapping TFs selected among all methods. Red: L_2 Boosting; Yellow: Sparse L_2 Boosting; Gray: BJ-LS; Green: Twin BJ-LS; Orange: BJ-EN; Purple: BJ-SCAD.

Figure 2. Estimated function of $\beta_j(\text{age})$ with their 95% confidence bands based on 500 bootstraps for the 3 gene expressions in the final semivarying-coefficient model.

Table 1

Variable screening, variable selection and structure identification performance in simulation.

Case	ρ	Method	S	TP	FP	Size	ISPE	RMISE	C_{var}	C_{fix}
I	0.2	(a)	1.00 (0.00)	6 (0.00)	23 (6.08)	29 (6.08)	0.70 (0.10)	0.81 (0.06)	0.97 (0.08)	-
		(b)	1.00 (0.00)	6 (0.00)	2 (2.33)	8 (2.33)	0.92 (0.08)	0.67 (0.06)	0.63 (0.14)	-
	0.8	(a)	1.00 (0.00)	6 (0.00)	9 (4.30)	15 (4.30)	1.05 (0.15)	0.81 (0.10)	0.67 (0.23)	-
		(b)	1.00 (0.00)	6 (0.52)	0 (1.36)	6 (1.41)	1.16 (0.15)	0.83 (0.14)	0.39 (0.15)	-
II	0.2	(a)	1.00 (0.00)	6 (0.00)	32 (6.59)	38 (6.59)	0.76 (0.19)	1.34 (0.16)	-	0.04 (0.08)
		(b)	1.00 (0.00)	6 (0.00)	2 (2.49)	8 (2.49)	0.96 (0.12)	0.90 (0.13)	-	0.80 (0.27)
	0.8	(a)	1.00 (0.00)	6 (0.00)	15 (4.41)	21 (4.41)	1.17 (0.21)	1.32 (0.19)	-	0.40 (0.26)
		(b)	1.00 (0.00)	6 (0.00)	1 (1.86)	7 (1.86)	1.31 (0.21)	1.16 (0.22)	-	0.76 (0.27)
III	0.2	(a)	0.99 (0.10)	6 (0.14)	13 (4.56)	19 (4.53)	1.31 (0.25)	0.93 (0.08)	0.88 (0.17)	0.64 (0.41)
		(b)	0.99 (0.10)	6 (0.14)	0 (0.65)	6 (0.64)	1.35 (0.23)	0.79 (0.07)	0.60 (0.19)	0.99 (0.05)
	0.8	(a)	1.00 (0.00)	6 (0.10)	7 (3.53)	13 (3.55)	1.85 (0.39)	1.44 (0.09)	0.72 (0.35)	0.87 (0.26)
		(b)	1.00 (0.00)	6 (0.20)	0 (0.56)	6 (0.58)	1.75 (0.36)	1.25 (0.15)	0.24 (0.21)	0.99 (0.05)
IV	0.2	(a)	1.00 (0.00)	6 (0.00)	16 (4.78)	22 (4.78)	1.20 (0.19)	0.97 (0.11)	0.90 (0.21)	0.55 (0.41)
		(b)	1.00 (0.00)	6 (0.00)	0 (0.39)	6 (0.39)	1.22 (0.16)	0.77 (0.09)	0.61 (0.20)	0.98 (0.10)
	0.8	(a)	1.00 (0.00)	6 (0.00)	8 (3.72)	14 (3.72)	1.51 (0.24)	1.32 (0.13)	0.84 (0.28)	0.87 (0.27)
		(b)	1.00 (0.00)	6 (0.10)	0 (0.60)	6 (0.60)	1.49 (0.23)	1.24 (0.18)	0.32 (0.32)	0.98 (0.10)
V	0.2	(a)	1.00 (0.00)	6 (0.00)	31 (5.43)	37 (5.44)	0.89 (0.18)	1.59 (0.15)	0.96 (0.04)	-
		(b)	1.00 (0.00)	6 (0.00)	3 (2.46)	9 (2.46)	1.04 (0.15)	0.97 (0.18)	0.74 (0.15)	-
	0.8	(a)	1.00 (0.00)	6 (0.00)	16 (4.87)	22 (4.87)	1.32 (0.28)	1.93 (0.15)	0.90 (0.16)	-
		(b)	1.00 (0.00)	6 (0.22)	1 (1.84)	7 (1.83)	1.42 (0.29)	1.67 (0.27)	0.57 (0.18)	-

Method: (a) L_2 Boosting (b) Sparse L_2 Boosting; S: coverage probability that the top $\left\lfloor \frac{n}{\log(n)} \right\rfloor$ covariates after screening includes all important covariates; TP: the median of true positive; FP: the median of false positive; Size: the median of model sizes; ISPE: the average of mean square error; RMISE: the average of root mean integrated squared error; C_{var} : the mean proportion of variables with varying coefficients that are selected; C_{fix} : the mean proportion of variables with constant nonzero coefficients that are selected; The values in the parentheses are the robust standard deviations; Simulation based on 500 replicates.

Table 2

Variable selection, estimation and prediction performance in lung cancer example.

Model	Method	No.	ISPE	OSPE
varying-coefficient AFT	L_2 Boosting	13	1.19	1.74
	Sparse L_2 Boosting	11	1.29	1.75
constant-coefficient AFT	BJ-LS	26	1.38	1.88
	Twin BJ-LS	18	1.40	1.94
	BJ-EN	32	1.33	1.90
	BJ-SCAD	14	1.51	1.89

No.: Number of genes selected; ISPE: In-sample prediction error $\sum_{i=1}^n \delta_i (Y_i - \hat{Y}_i)^2 / \sum_{i=1}^n \delta_i$; OSPE: Out-of-sample prediction error by 5-fold cross-validation.