Statistica Sinica **26** (2016), 547-567
doi:http://dx.doi.org/10.5705/ss.2013.076

JOINT STRUCTURE SELECTION AND ESTIMATION IN THE TIME-VARYING COEFFICIENT COX MODEL

Wei Xiao[1], Wenbin Lu[1] and Hao Helen Zhang[2]

[1]*North Carolina State University and* [2]*University of Arizona*

Abstract: The time-varying coefficient Cox model has been widely studied and popularly used in survival data analysis due to its flexibility for modeling covariate effects. It is of great practical interest to accurately identify the structure of covariate effects in a time-varying coefficient Cox model, covariates with null effect, constant effect and truly time-varying effect, and estimate the corresponding regression coefficients. Combining the ideas of local polynomial smoothing and group nonnegative garrote, we develop a new penalization approach to achieve such goals. Our method is able to identify the underlying true model structure with probability tending to one and can simultaneously estimate the time-varying coefficients consistently. The asymptotic normalities of the resulting estimators are established. We demonstrate the performance of our method using simulations and an application to the primary biliary cirrhosis data.

Key words and phrases: Group nonnegative garrote, local polynomial smoothing, model selection, time-varying coefficient Cox model.

1. Introduction

Cox proportional hazards model (Cox (1972)) is the standard semiparametric model in survival analysis due to its nice hazard interpretation and easy estimation based on partial likelihood principle with elegant counting process-based martingale theory (Andersen and Gill (1982)). However, one main limitation of the standard Cox model is to assume that the hazard ratios stay constant over time, which may be unrealistic in practical applications. Many alternatives have been proposed to relax the proportional hazards assumption. Among them, the time-varying coefficient Cox model is a natural extension that allows temporal effects of covariates; it has been widely studied in the literature (e.g. Zucker and Karr (1990); Cai and Sun (2003); Tian, Zucker, and Wei (2005)).

An important issue in fitting a time-varying coefficient Cox model is to distinguish covariates with null effect, constant effect, or truly time-varying effect, since this can help to build a more accurate risk model. Excluding covariates with null effect can greatly reduce the dimension of model, which is important, especially when p is large. Distinguishing covariates with constant effect and truly time-varying effect, a more easily interpreted semiparametric model results. Thus, Yu and Lin (2010), in considering data from a western Kenya parasitemia study, found that exposure to mosquito bites (BITE), age, and gender have constant effects on time to onset of parasitemia, while baseline parasitemia density (BPD) has a time-varying effect, and so considered a semiparametric time-varying coefficient model for better risk prediction. See Zhang, Lee, and Song (2002); Fan and Huang (2005); Ahmad, Leelahanon, and Li (2005); Wang, Zhu, and Zhou (2009) for other demonstration of the benefits of semiparametric varying-coefficient models.

Model selection has been extensively studied in the past few decades. Traditional model selection techniques, such as best-subset selection, coupled with C_p (Mallows (1973)), AIC (Akaike (1973)) and BIC (Schwarz (1978)), separate model selection and model estimation steps and are generally unstable due to the their inherent discreteness (Breiman (1995)) and stochastic errors (Fan and Li (2001)). They also lack asymptotic selection consistency, a desirable asymptotic property. More importantly, they are not computationally feasible for data sets with moderate to large dimensions. To overcome these difficulties, various penalization methods have been introduced, including, the nonnegative garrote (Breiman (1995)), LASSO (Tibshirani (1996, 1997)), SCAD (Fan and Li (2001, 2002)) and adaptive LASSO (Zou (2006); Zhang and Lu (2007)). These methods provide competing performances for simultaneously selecting important variables and estimating their effects, but most of them focus on variable selection for simple linear regression models. Less has been studied for such model structure selection methods as the identification of linear/nonlinear structure in partially linear regression models or time-invariant/time-varying coefficients in regression models with time-varying coefficients. Zhang, Cheng, and Liu (2011) proposed a novel penalization approach in the frame of smoothing spline ANOVA for automatically discovering covariates with null, linear, and nonlinear effects in a partially linear model. For censored data, Yan and Huang (2012) proposed an adaptive group LASSO (AGLASSO) method based on a penalized B-spline approach for model structure selection in a time-varying coefficient Cox model.

In this paper, we propose a method for automatic model structure selection and coefficient estimation in a time-varying coefficient Cox model by coupling the kernel-weighted partial likelihood estimation (Cai and Sun (2003); Tian, Zucker, and Wei (2005)) with a group nonnegative garrote penalty. Compared with the spline method proposed in Yan and Huang (2012), our method is better able to capture some local features of time-varying coefficient functions. By using the local kernel estimation, we can rigorously study the asymptotic properties of the proposed estimators for both constant and time-varying coefficients, whereas these properties have not been established for existing approaches like that of

Yan and Huang (2012). The proposed method also provides an automatic and effective way to conduct structure selection for a time-varying coefficient Cox model, that can deal with relatively large dimensions in contrast with such existing methods based on hypothesis testing as those studied in Huang, Wu, and Zhou (2002), Fan and Huang (2005), Tian, Zucker, and Wei (2005), and Liu, Lu, Shore, and Zeleniuch-Jacquotte (2010). The remainder of the paper is organized as follows.

Our proposed kernel group nonnegative garrote (KGNG) method and its variant (KGNG2) are introduced in Section 2. Asymptotic properties of KGNG and KGNG2 estimators are presented in Section 3. Section 4 is devoted to simulations and an application to primary biliary cirrhosis data. The proofs are relegated to an online supplementary appendix available at `http://www.stat.sinica.edu.tw/statistica`.

2. Structure Selection with Kernel Group Nonnegative Garrote

2.1. Methods

Consider a random sample of n individuals. Let T_i be the failure time, C_i be the censoring time, and $\boldsymbol{Z}_i$ be a p-vector of covariates for subject i. Conditional on $\boldsymbol{Z}_i$, T_i and C_i are assumed independent. Take $\widetilde{T}_i = \min(T_i, C_i)$ and $\Delta_i = \mathbb{1}(T_i \leq C_i)$. The data consist of the triplets $(\widetilde{T}_i, \boldsymbol{Z}_i, \Delta_i)$, $i = 1, \ldots, n$. The time-varying coefficient Cox model has

$$\alpha(t|\boldsymbol{Z}_i) = \alpha_0(t) e^{\boldsymbol{\beta}_0^{\intercal}(t)\boldsymbol{Z}_i}, \tag{2.1}$$

where $\alpha(\cdot|\boldsymbol{Z}_i)$ is the conditional hazard function given covariates, $\alpha_0(\cdot)$ is a completely unspecified baseline hazard function, and $\boldsymbol{\beta}_0(t) = (\beta_{01}(t), \ldots, \beta_{0p}(t))^{\intercal}$ is a p-dimensional smooth function of t.

Without loss of generality, we assume $\boldsymbol{\beta}_0(t) = (\boldsymbol{\beta}_O^{\intercal}(t), \boldsymbol{\beta}_C^{\intercal}(t), \boldsymbol{\beta}_{NC}^{\intercal}(t))^{\intercal}$, where $\boldsymbol{\beta}_O(t) \in \mathbb{R}^{p_1}$, $\boldsymbol{\beta}_C(t) \in \mathbb{R}^{p_2}$, and $\boldsymbol{\beta}_{NC}(t) \in \mathbb{R}^{p_3}$ correspond to covariates with null effect, constant effect, and time-varying effect, respectively, with $p = p_1 + p_2 + p_3$. Denote the corresponding index sets of the three classes by I_O, I_C, and I_{NC}, and let $\mathrm{I} = \{1, \ldots, p\} = \{\mathrm{I}_O \bigcup \mathrm{I}_C \bigcup \mathrm{I}_{NC}\}$. Our method consists of two steps.

In Step 1, for any fixed t, we obtain the initial estimator $\widetilde{\boldsymbol{\beta}}(t) = (\widetilde{\beta}_1(t), \ldots, \widetilde{\beta}_p(t))^{\intercal} \in \mathbb{R}^p$ using the kernel-weighted partial likelihood estimation (Cai and Sun (2003); Tian, Zucker, and Wei (2005)): maximize the local partial likelihood

$$L_{1n}(\boldsymbol{\beta}, t) = \frac{1}{n} \sum_{i=1}^{n} \int_0^{\tau} K_h(s-t) \Big[\boldsymbol{\beta}^{\intercal} \boldsymbol{Z}_i - \log\Big(\sum_{j=1}^{n} Y_j(s) e^{\boldsymbol{\beta}^{\intercal} \boldsymbol{Z}_j}\Big)\Big] dN_i(s), \tag{2.2}$$

with respect to $\boldsymbol{\beta}$, where $K_h(\cdot) = K(\cdot/h)/h$ with $K(\cdot)$ a symmetric kernel density function, h a bandwidth parameter, $Y_j(t) = \mathbb{1}(\widetilde{T}_j \geq t)$, $N_i(t) = \mathbb{1}(\widetilde{T}_i \leq t, \Delta_i = 1)$, and τ a pre-specified constant such that $P(\widetilde{T}_i > \tau) > 0$. We decompose the initial estimator $\widetilde{\boldsymbol{\beta}}(t)$ into a mean part $\widetilde{\boldsymbol{m}} = (\widetilde{m}_1, \ldots, \widetilde{m}_p)^{\top}$ and a deviation part $\widetilde{\boldsymbol{\beta}}^*(t) = (\widetilde{\beta}_1^*(t), \ldots, \widetilde{\beta}_p^*(t))^{\top}$, where $\widetilde{m}_k = \tau^{-1}\int_0^{\tau} \widetilde{\beta}_k(u)du$ and $\widetilde{\beta}_k^*(t) = \widetilde{\beta}_k(t) - \widetilde{m}_k$ for $k = 1, \ldots, p$. Practically, we can choose M grid points $\mathcal{T}_M = \{t_1, \ldots, t_M\}$, equally spaced between 0 and τ, where M is a large positive integer. We then let $\widetilde{m}_k = \sum_{i=1}^M \widetilde{\beta}_k(t_i)/M$ and $\widetilde{\beta}_k^*(t) = \widetilde{\beta}_k(t_j) - \widetilde{m}_k$, where $j = \text{argmin}_k |t_k - t|$ and $t \in [0, \tau]$. In our numerical studies, we set M to be 100 which, based on our experiment, is large enough to allow good approximations of $\widetilde{m}_k$ and $\widetilde{\beta}_k^*(t)$.

In Step 2, we adapt group nonnegative garrote penalties for structure selection. Let $\boldsymbol{\lambda}_1 = (\lambda_{11}, \ldots, \lambda_{1p})^{\top}$, $\boldsymbol{\lambda}_2 = (\lambda_{21}, \ldots, \lambda_{2p})^{\top}$ be p-dimensional vectors. We obtain $\widehat{\boldsymbol{\lambda}}_1$ and $\widehat{\boldsymbol{\lambda}}_2$ by minimizing

$$\begin{aligned} Q_{2n}(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) = &- \sum_{i=1}^n \int_0^{\tau} \Bigg[\Big(\widetilde{\boldsymbol{m}} \circ \boldsymbol{\lambda}_1 + \widetilde{\boldsymbol{\beta}}^*(s) \circ \boldsymbol{\lambda}_2\Big)^{\top} \boldsymbol{Z}_i \\ &- \log\Big(\sum_{j=1}^n Y_j(s) e^{\left(\widetilde{\boldsymbol{m}} \circ \boldsymbol{\lambda}_1 + \widetilde{\boldsymbol{\beta}}^*(s) \circ \boldsymbol{\lambda}_2\right)^{\top} \boldsymbol{Z}_j}\Big)\Bigg] dN_i(s) \\ &+ \theta_1 \sum_{j=1}^p \lambda_{1j} + \theta_2 \sum_{j=1}^p \lambda_{2j} \end{aligned} \tag{2.3}$$

subject to $\lambda_{1j} \geq 0$ and $\lambda_{2j} \geq 0$, $j = 1, \ldots, p$, where $\boldsymbol{\theta} = (\theta_1,\ \theta_2)$ are two-dimensional nonnegative tuning parameters, and $a \circ b$ denotes the Hadamard (element-wise) product of vectors a and b. The proposed KGNG estimator of $\beta_{0k}(t)$ is

$$\widehat{\beta}_k(t) = \widehat{\lambda}_{1k}\widetilde{m}_k + \widehat{\lambda}_{2k}\widetilde{\beta}_k^*(t),\ k = 1, \ldots, p,\ t \in [0, \tau]. \tag{2.4}$$

The automatic structure selection is achieved by shrinking some components of $\widehat{\boldsymbol{\lambda}}_1$ and $\widehat{\boldsymbol{\lambda}}_2$ to zero. Specifically, we take $\widehat{\mathrm{I}}_O = \{k \in \mathrm{I} : \widehat{\lambda}_{1k} = 0, \widehat{\lambda}_{2k} = 0\}$, $\widehat{\mathrm{I}}_C = \{k \in \mathrm{I} : \widehat{\lambda}_{1k} \neq 0, \widehat{\lambda}_{2k} = 0\}$, and $\widehat{\mathrm{I}}_{NC} = \{k \in \mathrm{I} : \widehat{\lambda}_{2k} \neq 0\}$ as estimated index sets for I_O, I_C and I_{NC}, respectively.

When the number of covariates is large, the inclusion of many noise variables (covariates with null effect) at Step 1 may affect the structure selection performance. Accordingly, we propose a variant of the KGNG estimator. We add a preliminary step (Step 0) to exclude all noise variables prior to structure selection. To do this, we conduct a standard group nonnegative garrote estimation by minimizing

$$-\sum_{i=1}^n \int_0^{\tau} \Big[\Big(\widetilde{\boldsymbol{\beta}}(s) \circ \boldsymbol{\lambda}^*\Big)^{\top} \boldsymbol{Z}_i - \log\Big(\sum_{j=1}^n Y_j(s) e^{\left(\widetilde{\boldsymbol{\beta}}(s) \circ \boldsymbol{\lambda}^*\right)^{\top} \boldsymbol{Z}_j}\Big)\Big] dN_i(s) + \theta^* \sum_{j=1}^p \lambda_j^*$$

with respect to $\boldsymbol{\lambda}^* = (\lambda_1^*, \ldots, \lambda_p^*)^\intercal$, where $\lambda_j^* \geq 0$, $j = 1, \ldots, p$, and θ^* is a nonnegative tuning parameter. Let $\widehat{\boldsymbol{\lambda}}^* = (\widehat{\lambda}_1^*, \ldots, \widehat{\lambda}_p^*)^\intercal$ denote the resulting minimizer. We exclude the kth covariate if $\widehat{\lambda}_k^* = 0$. Let $\underline{\boldsymbol{Z}}_i$ denote the remaining sub-vector of $\boldsymbol{Z}_i$ by keeping all important covariates. We then implement Steps 1 and 2 with $\boldsymbol{Z}_i$ replaced by $\underline{\boldsymbol{Z}}_i$ for further structure selection. The resulting estimator is denoted by KGNG2.

2.2. Computational aspects

We implemented the proposed method in R, and the corresponding code can be downloaded from the author's web page (`http://www4.ncsu.edu/~wxiao/`). In Step 1, $L_{1n}(\boldsymbol{\beta}, t)$ is strictly concave with probability one and thus has a unique solution. The maximization can be realized based on a regular Newton-Raphson iteration or an efficient iterative algorithm proposed in Cai, Fan, and Li (2000). In Step 2, after proper transformations, the minimization problem (2.3) is equivalent to finding the lasso solution for a Cox model with time-dependent covariates $\widetilde{\boldsymbol{Z}}_i(s)$ under the nonnegative constraint of regression parameters, where

$$\widetilde{\boldsymbol{Z}}_i(s) = \begin{pmatrix} \widetilde{\boldsymbol{m}} \circ \boldsymbol{Z}_i \\ \widetilde{\boldsymbol{\beta}}^*(s) \circ \boldsymbol{Z}_i \end{pmatrix}.$$

We used the R package "penalized" (Goeman (2010)) for this. The algorithm is based on a combination of gradient ascent optimization and the Newton-Raphson algorithm, which can also incorporate nonnegative constraints on the parameters (Goeman (2010)). As well, the minimization in the preliminary Step 0 is equivalent to finding a lasso solution for a Cox model with time-dependent covariates $\widetilde{\boldsymbol{\beta}}(s) \circ \boldsymbol{Z}_i$, and can be computed similarly with existing R packages.

2.3. Tuning procedure

For computing KGNG, we need to choose the bandwidth h at the maximum local partial likelihood estimation step (2.2) and (θ_1, θ_2) at the group nonnegative garrote estimation step (2.3). To choose h, we use a K-fold cross-validation method as suggested in Tian, Zucker, and Wei (2005). We randomly split the data set into K roughly equal-sized parts, and for each $k = 1, \ldots, K$, we delete the kth part and fit the time-varying coefficient Cox model with the other $K-1$ parts. Then we compute the prediction error $PE_k(h)$, which measures how well the fitted model predicts the kth part of the data. Here,

$$PE_k(h) = -\sum_{i \in \mathrm{I}_k} \int_0^\tau \Big[\widetilde{\boldsymbol{\beta}}^{(-k)}(s)^\intercal \boldsymbol{Z}_i - \log\Big(\sum_{j \in \mathrm{I}_k} Y_j(s) e^{\widetilde{\boldsymbol{\beta}}^{(-k)}(s)^\intercal \boldsymbol{Z}_j} \Big) \Big] dN_i(s),$$

where I_k is the index set for the kth part of the data and $\widetilde{\boldsymbol{\beta}}^{(-k)}(t)$ is the maximum local partial likelihood estimator calculated with the kth part of the data deleted. Last, the optimal h is obtained by minimizing the total prediction error $PE(h) = \sum_{k=1}^{K} PE_k(h)$.

For (θ_1, θ_2), we consider a set of bivariate grid values, and choose the optimal (θ_1, θ_2) by minimizing

$$\mathbf{BIC} = -\frac{2\log(\textbf{partial likelihood})}{n} + \frac{\log n}{n} \times df_1 + \log\frac{nh^*}{nh^*} \times df_2,$$

where df_1 and df_2 are the number of nonzero components in $\widehat{\boldsymbol{\lambda}}_1$ and $\widehat{\boldsymbol{\lambda}}_2$, respectively, and $h^* = h/\tau$ is the effective bandwidth when we scale τ to 1. The effective sample size nh^* is used here for the time-varying components instead of the original n to account for the fact that $\beta_i^*(t)$ is estimated locally. A similar strategy was adopted in Wang and Xia (2009) and Hu and Xia (2012).

Similarly, θ^* in the preliminary Step 0 can be chosen by minimizing

$$\mathbf{BIC} = -\frac{2\log(\textbf{partial likelihood})}{n} + \log\frac{nh^*}{nh^*} \times df_3,$$

where df_3 is the number of covariates with nonzero effect.

3. Theoretical Properties

3.1. Asymptotic properties of initial estimators

Denote the true mean and deviation part of $\boldsymbol{\beta}_0(t)$ as $\boldsymbol{m}_0 = (m_{01}, \ldots, m_{0p})^\intercal$ and $\boldsymbol{\beta}_0^*(t) = (\beta_{01}^*(t), \ldots, \beta_{0p}^*(t))^\intercal$, respectively, where

$$m_{0k} = \tau^{-1}\int_0^\tau \beta_{0k}(u)du, \quad \beta_{0k}^*(t) = \beta_{0k}(t) - m_{0k},$$

for $k = 1, \ldots, p$. Let $I(\boldsymbol{\beta}, t) = -\partial^2 L_{1n}(\boldsymbol{\beta}, t)/\partial\boldsymbol{\beta}^2 = n^{-1}\sum_{i=1}^n \int_0^\tau V(\boldsymbol{\beta}, s)K_{h_n}(s-t)dN_i(s)$, where

$$V(\boldsymbol{\beta}, t) = \frac{S^{(2)}(\boldsymbol{\beta}, t)}{S^{(0)}(\boldsymbol{\beta}, t)} - \left(\frac{S^{(1)}(\boldsymbol{\beta}, t)}{S^{(0)}(\boldsymbol{\beta}, t)}\right)^{\otimes 2},$$

$$S^{(r)}(\boldsymbol{\beta}, t) = n^{-1}\sum_{i=1}^n Y_i(t)\boldsymbol{Z}_i^{\otimes r}e^{\boldsymbol{\beta}'\boldsymbol{Z}_i}, r = 0, 1, 2,$$

with $\otimes$ denoting the outer product. Let $\boldsymbol{E}(\boldsymbol{\beta}, t) = S^{(1)}(\boldsymbol{\beta}, t)/S^{(0)}(\boldsymbol{\beta}, t)$, $P(t|\boldsymbol{z}) = \mathrm{P}(\widetilde{T} \geq t|\boldsymbol{Z} = \boldsymbol{z})$, $Q_0(t) = \mathrm{E}[P(t|\boldsymbol{Z})\alpha(t|\boldsymbol{Z})]$, $Q_1(t) = \mathrm{E}[P(t|\boldsymbol{Z})\alpha(t|\boldsymbol{Z})\boldsymbol{Z}]$, and $Q_2(t) = \mathrm{E}[P(t|\boldsymbol{Z})\alpha(t|\boldsymbol{Z})\boldsymbol{Z}^{\otimes 2}]$. Define $\Sigma(t) = Q_2(t) - Q_1(t)Q_1(t)^\intercal/Q_0(t)$. Let $s^{(r)}(\boldsymbol{\beta}, t)$ denote the limits of $S^{(r)}(\boldsymbol{\beta}, t)$, $r = 0, 1, 2$, as $n \to \infty$. Let $\mathcal{N}(t, \epsilon)$ be an ϵ-neighborhood of t, for $\epsilon > 0$ and $t \in [0, \tau]$, and $\mathcal{B}$ be a compact set of $\mathrm{I\!R}^p$ that includes a neighborhood of $\boldsymbol{\beta}_0(t)$ for $t \in [0, \tau]$. We need certain regularity conditions.

(A.1) The kernel function $K(\cdot)$ is a bounded and symmetric density with a bounded support [-1,1].

(A.2) For $t \in [0, \tau]$

$$\mathrm{E}\Big[\exp\Big\{2\Big(\sup_{u \in \mathcal{N}(t,\epsilon)} |\boldsymbol{\beta}_0(u)| + \boldsymbol{\beta}_0^{'}(t) + 3\Big)|Z|\Big\}\Big] < \infty.$$

(A.3) $Q_0(t) > 0$, $Q_1(t)$, and $Q_2(t)$ are continuous for $t \in [0, \tau]$.

(A.4) $\alpha_0(t)$ is positive and continuous, $P(t|\boldsymbol{z}) > 0$, and the coefficient functions $\{\beta_{0j}(t)\}$ have a continuous second derivatives for $t \in [0, \tau]$.

(A.5) The matrix $\Sigma(t)$ is positive definite for $t \in [0, \tau]$.

(A.6) $s^{(r)}(\boldsymbol{\beta}, t)$ is uniformly continuous with respect to $(\boldsymbol{\beta}^\intercal, t)^\intercal \in \mathcal{B} \times [0, \tau]$ for $t \in [0, \tau]$.

Lemma 1. *If $h_n = O(n^{-\nu})$ with $\nu \in [1/5, 1)$, $\widetilde{\boldsymbol{\beta}}(t) \xrightarrow{p} \boldsymbol{\beta}_0(t)$, $0 \le t \le \tau$. If $1/5 < \nu < 1$, we have, for fixed $t \in (0, \tau)$,*

$$(nh_n)^{1/2}\left(\widetilde{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}_0(t)\right) \xrightarrow{d} N\left\{0,\ \Sigma^{-1}(t)\int_{-1}^{1} K^2(s)ds\right\},$$

where $\Sigma(t)$ can be consistently estimated by $I(\widetilde{\boldsymbol{\beta}}(t), t)$.

Lemma 2. *If $h_n = O(n^{-\nu})$ with $1/4 < \nu < 1/2$, $n^{1/2}\left(\widetilde{\boldsymbol{m}} - \boldsymbol{m}_0\right) \xrightarrow{d} N\left(0,\ \Sigma_m\right)$, where $\Sigma_m = \int_0^\tau \Sigma^{-1}(u)du/\tau^2$ can be consistently estimated by*

$$\widehat{\Sigma}_m = \int_h^{\tau-h} \frac{I^{-1}(\widetilde{\beta}(u), u)du}{(\tau - 2h)^2}.$$

The proof of Lemma 1 follows Cai and Sun (2003) and the proof of Lemma 2 follows steps given in Section 5 of Tian, Zucker, and Wei (2005).

3.2. Asymptotic properties of KGNG estimators

Let $\boldsymbol{\lambda}_{01}$ and $\boldsymbol{\lambda}_{02}$ denote the true values of $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$, respectively, and partition them as: $\boldsymbol{\lambda}_{01} = (\boldsymbol{\lambda}_{01}^{O\top}, \boldsymbol{\lambda}_{01}^{C\top}, \boldsymbol{\lambda}_{01}^{NC\top})^\intercal$ and $\boldsymbol{\lambda}_{02} = (\boldsymbol{\lambda}_{02}^{O\top}, \boldsymbol{\lambda}_{02}^{C\top}, \boldsymbol{\lambda}_{02}^{NC\top})^\intercal$, according to the true index sets I_O, I_C and I_{NC}, respectively. Then

$$\boldsymbol{\lambda}_{01}^{O} = \mathbf{0}_{p_1},\ \ \boldsymbol{\lambda}_{01}^{C} = \mathbf{1}_{p_2},\ \ \boldsymbol{\lambda}_{02}^{O} = \mathbf{0}_{p_1},\ \ \boldsymbol{\lambda}_{02}^{C} = \mathbf{0}_{p_2},\ \ \boldsymbol{\lambda}_{02}^{NC} = \mathbf{1}_{p_3},$$

and $\lambda_{01,j}^{NC} = 0$ or 1 for $j \in \mathrm{I}_{NC}$, where $\lambda_{01,j}^{NC}$ is the jth component of $\boldsymbol{\lambda}_{01}^{NC}$ and corresponds to the mean part of the jth covariate with time-varying effect. If $\lambda_{01,j}^{NC} = 0$, the jth time-varying effect has zero mean effect; otherwise, it has nonzero mean effect. As we do not distinguish between the two types of time-varying effects, without loss of generality, we assume $\boldsymbol{\lambda}_{01}^{NC} = \mathbf{1}_{p_3}$ in our theoretical derivations. Let $\boldsymbol{\lambda}_0 = (\boldsymbol{\lambda}_{01}^\top, \boldsymbol{\lambda}_{02}^\top)^\top$. We further partition $\boldsymbol{\lambda}_0$ as $\boldsymbol{\lambda}_0^{(1)}$ representing all the ones and $\boldsymbol{\lambda}_0^{(0)}$ representing all the zeros, where $\boldsymbol{\lambda}_0^{(1)} = (\boldsymbol{\lambda}_{01}^{C\top}, \boldsymbol{\lambda}_{01}^{NC\top}, \boldsymbol{\lambda}_{02}^{NC\top})^\top = \mathbf{1}_{p_2+2p_3}$ and $\boldsymbol{\lambda}_0^{(0)} = (\boldsymbol{\lambda}_{01}^{O\top}, \boldsymbol{\lambda}_{02}^{O\top}, \boldsymbol{\lambda}_{02}^{C\top})^\top = \mathbf{0}_{2p_1+p_2}$. In a similar manner we define $\boldsymbol{\lambda}$, $\boldsymbol{\lambda}^{(1)}$, $\boldsymbol{\lambda}^{(0)}$, $\widehat{\boldsymbol{\lambda}}$, $\widehat{\boldsymbol{\lambda}}^{(1)}$ and $\widehat{\boldsymbol{\lambda}}^{(0)}$.

Note that the KGNG estimator defined in (2.4) takes the form $\widehat{\boldsymbol{\beta}}(t) = \widehat{\boldsymbol{\lambda}}_1 \circ \widetilde{\boldsymbol{m}} + \widehat{\boldsymbol{\lambda}}_2 \circ \widetilde{\boldsymbol{\beta}}^*(t)$. To derive its asymptotic properties, we need to study the asymptotic properties of $\widetilde{\boldsymbol{m}}$, $\widetilde{\boldsymbol{\beta}}^*(t)$, and $\widehat{\boldsymbol{\lambda}}$. In Lemmas 1 and 2, we have established the asymptotic properties of $\widetilde{\boldsymbol{m}}$ and $\widetilde{\boldsymbol{\beta}}^*(t)$. In the following, we derive the asymptotic properties of $\widehat{\boldsymbol{\lambda}}^{(1)}$ and $\widehat{\boldsymbol{\lambda}}^{(0)}$. Specifically, we establish the root-n consistency of $\widehat{\boldsymbol{\lambda}}^{(1)}$ in Theorem 1 and the sparsity property of $\widehat{\boldsymbol{\lambda}}^{(0)}$ in Theorem 2.

Theorem 1. *Let* $h_n = O(n^{-\nu})$ *with* $1/4 < \nu < 1/2$. *Under* (A1)−(A6), *if* $\max(\theta_1, \theta_2)/\sqrt{n}$ *is bounded, then* $\|\widehat{\boldsymbol{\lambda}}^{(1)} - \boldsymbol{\lambda}_0^{(1)}\| = O_p(n^{-1/2})$.

Theorem 2. *Let* $h_n = O(n^{-\nu})$ *with* $1/4 < \nu < 1/2$. *Under* (A1)−(A6), *if* $\|\widehat{\boldsymbol{\lambda}}^{(1)} - \boldsymbol{\lambda}_0^{(1)}\| = O_p(n^{-1/2})$ *and* $h_n^{1/2} \min(\theta_1, \theta_2) \to \infty$, *then* $P(\widehat{\boldsymbol{\lambda}}^{(0)} = \mathbf{0}) \to 1$.

Combining Theorems 1 and 2, we can prove the selection consistency of the KGNG estimator, summarized in part (a) of Theorem 3. We further establish the asymptotic normality of the KGNG estimators for both nonzero constant and time-varying regression coefficients in Theorem 3. With our standard partitioning, we write $\boldsymbol{m}_0 = (\boldsymbol{m}_O^\top, \boldsymbol{m}_C^\top, \boldsymbol{m}_{NC}^\top)^\top$ and $\boldsymbol{\beta}_0^*(t) = (\boldsymbol{\beta}_O^{*\top}, \boldsymbol{\beta}_C^{*\top}, \boldsymbol{\beta}_{NC}^{*\top}(t))^\top$, and partition $\widehat{\boldsymbol{\beta}}(t)$, $\widetilde{\boldsymbol{\beta}}(t)$, $\widetilde{\boldsymbol{m}}$, and $\widetilde{\boldsymbol{\beta}}^*(t)$ accordingly.

Theorem 3. *Let* $h_n = O(n^{-\nu})$ *with* $1/4 < \nu < 1/2$. *Under (A1)-(A6), if* $\max(\theta_1, \theta_2)/\sqrt{n}$ *is bounded and* $h_n^{1/2} \min(\theta_1, \theta_2) \to \infty$, *then*

(a) **(Selection consistency)** *with probability tending to one,* $\widehat{\mathrm{I}}_O = \mathrm{I}_O$, $\widehat{\mathrm{I}}_C = \mathrm{I}_C$ *and* $\widehat{\mathrm{I}}_{NC} = \mathrm{I}_{NC}$*;*

(b) **(Root-n consistency of $\widehat{\boldsymbol{\beta}}_C$)** $\widehat{\boldsymbol{\beta}}_C$ *is a root-n consistent estimator for* $\boldsymbol{\beta}_C$.

(c) **(Asymptotic normality of $\widehat{\boldsymbol{\beta}}_C$)** *If we further assume* $\max(\theta_1, \theta_2)/\sqrt{n} \to 0$,

$$(n)^{1/2}\left(\widehat{\boldsymbol{\beta}}_C - \boldsymbol{\beta}_C\right) \xrightarrow{d} N\left\{0,\ \Sigma_m^F\right\},$$

where

$$\Sigma_m^F = \int_0^\tau \left(D(u) + \frac{1}{\tau} B_{10} \Sigma^{-1}(u)\right) \Sigma(u) \left(D(u) + \frac{1}{\tau} B_{10} \Sigma^{-1}(u)\right)^\top du. \quad (3.1)$$

Here $D(u)$ *is a* $p_2 \times p$ *matrix given in* (S3.8) *in the Supplement and* $B_{10} = (\mathbf{0}_{p_2\times p_1}|\boldsymbol{I}_{p_2}|\mathbf{0}_{p_2\times p_3})$.

(d) **(Asymptotic normality of $\widehat{\boldsymbol{\beta}}_{NC}(t)$)**

$$(nh_n)^{1/2}\left(\widehat{\boldsymbol{\beta}}_{NC}(t) - \boldsymbol{\beta}_{NC}(t)\right) \xrightarrow{d} N\left\{0,\ \{\Sigma^{-1}(t)\}_{NC,NC}\int_{-1}^{1} K^2(s)ds\right\},$$

where $\{\Sigma^{-1}(t)\}_{NC,NC}$ *is the submatrix of* $\Sigma^{-1}(t)$ *corresponding to* I_{NC}.

The asymptotic variance-covariance matrices given in parts (c) and (d) can be consistently estimated by the usual plug-in method. We note that the limiting distribution of the KGNG estimator for the time-varying coefficients given in part (d) is the same as that of the corresponding initial estimator. Actually in the proof of part (d), we show that the difference between $\widehat{\boldsymbol{\beta}}_{NC}(t)$ and $\widetilde{\boldsymbol{\beta}}_{NC}(t)$ is uniformly asymptotically negligible in t. Thus we can construct confidence bands of $\widehat{\boldsymbol{\beta}}_{NC}(t)$ using the resampling technique proposed in Tian, Zucker, and Wei (2005).

3.3. Asymptotic properties of KGNG2 estimator

Let $\boldsymbol{Z} = (\boldsymbol{Z}_O^\top, \boldsymbol{Z}_C^\top, \boldsymbol{Z}_{NC}^\top)^\top$ and $\underline{\boldsymbol{Z}} = (\boldsymbol{Z}_C^\top, \boldsymbol{Z}_{NC}^\top)^\top$ be the subvector of $\boldsymbol{Z}$ with only important covariates kept. Let $\underline{\boldsymbol{\beta}}_0(t) = (\boldsymbol{\beta}_C^\top(t), \boldsymbol{\beta}_{NC}^\top(t))^\top$, $\alpha(t|\underline{\boldsymbol{Z}}) = \alpha_0(t)e^{\underline{\boldsymbol{\beta}}_0(t)^\top\underline{\boldsymbol{Z}}}$, $P(t|\underline{\boldsymbol{Z}}) = P(Y \geq t|\underline{\boldsymbol{Z}} = \underline{\boldsymbol{z}})$, $\underline{Q}_0(t) = E[P(t|\underline{\boldsymbol{Z}})\alpha(t|\underline{\boldsymbol{Z}})]$, $\underline{Q}_1(t) = E[P(t|\underline{\boldsymbol{Z}})\alpha(t|\underline{\boldsymbol{Z}})\underline{\boldsymbol{Z}}]$, and $\underline{Q}_2(t) = E[P(t|\underline{\boldsymbol{Z}})\alpha(t|\underline{\boldsymbol{Z}})\underline{\boldsymbol{Z}}^{\otimes 2}]$. Define $\underline{\Sigma}(t) = \underline{Q}_2(t) - \underline{Q}_1(t)\underline{Q}_1(t)^\top/\underline{Q}_0(t)$. We add "*" to distinguish the KGNG2 estimator from the KGNG estimator. We summarize the asymptotic properties of KGNG2 estimator in Theorem 4.

Theorem 4. *Let* $h_n = O(n^{-\nu})$ *with* $1/4 < \nu < 1/2$. *Under* (A1)$-$(A6), *if* $\theta^*/\sqrt{n}$ *is bounded,* $h_n^{1/2}\theta^* \to \infty$, $\max(\theta_1, \theta_2)/\sqrt{n}$ *is bounded, and* $h_n^{1/2}\min(\theta_1, \theta_2) \to \infty$, *then*

(a) **(Selection consistency of preliminary Step 0)** *with probability tending to one,* $\widehat{\lambda}_k^* = 0$, *for* $k \in \mathrm{I}_O$ *and* $\widehat{\lambda}_k^* \neq 0$, *for* $k \in \mathrm{I}_C \bigcup \mathrm{I}_{NC}$*;*

(b) **(Selection consistency)** *with probability tending to one,* $\widehat{\mathrm{I}}_O^* = \mathrm{I}_O$, $\widehat{\mathrm{I}}_C^* = \mathrm{I}_C$ *and* $\widehat{\mathrm{I}}_{NC}^* = \mathrm{I}_{NC}$*;*

(c) **(Root-n consistency of $\widehat{\boldsymbol{\beta}}_C^*$)** $\widehat{\boldsymbol{\beta}}_C^*(t)$ *is a root-n consistent estimator for* $\boldsymbol{\beta}_C$*;*

(d) **(Asymptotic normality of $\widehat{\boldsymbol{\beta}}_C^*$)** *if we further assume* $\max(\theta_1, \theta_2)/\sqrt{n} \to 0$,

$$(n)^{1/2}\left(\widehat{\boldsymbol{\beta}}_C^* - \boldsymbol{\beta}_C\right) \xrightarrow{d} N\left\{0,\ \underline{\Sigma}_m^F\right\},$$

where $\underline{\Sigma}_m^F$ *can be computed following* (3.1) *with some obvious changes;*

(e) **(Asymptotic normality of $\widehat{\boldsymbol{\beta}}^*_{NC}(t)$)**

$$(nh_n)^{1/2}\left(\widehat{\boldsymbol{\beta}}^*_{NC}(t)-\boldsymbol{\beta}_{NC}(t)\right)\xrightarrow{d}N\left\{0,\ \{\underline{\Sigma}^{-1}(t)\}_{NC,NC}\int_{-1}^{1}K^2(s)ds\right\},$$

where $\{\underline{\Sigma}^{-1}(t)\}_{NC,NC}$ is the submatrix of $\underline{\Sigma}^{-1}(t)$ corresponding to I_{NC}.

The proof of Theorem 4 is similar to that of Theorem 3 and is omitted. Based on Theorem 3 and 4, $\widehat{\boldsymbol{\beta}}^*_{NC}(t)$ is strictly more efficient than $\widehat{\boldsymbol{\beta}}_{NC}(t)$ if I_O is not empty. However, there is no clear order between the efficiencies of $\widehat{\boldsymbol{\beta}}^*_{C}$ and $\widehat{\boldsymbol{\beta}}_{C}$.

4. Numerical Studies

4.1. Simulation studies

We generated failure times from the varying-coefficient Cox model (2.1). The covariate vector $\boldsymbol{Z}$ was generated from a multivariate normal with mean 0, variance 0.5 and correlation coefficient $0.5^{|j-k|}$ for any pair (j,k). We considered both the low-dimensional and the high-dimensional cases, with $p=10$ and 50. There are three nonzero coefficients in $\boldsymbol{\beta}_0(t)$: $\beta_{02}(t)=-\{1+\cos(\pi t)\}\mathbb{1}(0<t<1)$, $\beta_{03}(t)=1.5\{\cos(\pi t/2)\}$, and $\beta_{08}(t)=-1$. Thus two covariates with time-varying effects, one with constant effect, and all others with null effect. The baseline hazard function $\alpha_0(t)=\exp\{-\cos(\pi t/2)\}$. We considered cases with censoring times dependent and independent of the covariates. When $p=10$, the censoring times of ith subject were mixtures of W and a point mass at 2, where $W=\min(\exp(Z_{i2}-Z_{i5}),\ \mathrm{Unif}(0,2))$. When $p=50$, we generated censoring times from a mixture of $\mathrm{Unif}(0,2))$ and a point mass at 2. In both cases the mixing probability was chosen to have the censoring proportion $c_p=20\%$ or 40%. For each scenario, we conducted 100 simulation runs with sample size $n=200$ and 400. We compared the proposed KGNG and KGNG2 with the AGLASSO of Yan and Huang (2012). For our estimators, we used the Epanechnikov kernel $K(x)=3(1-x^2)/4$, $-1\le x\le 1$. The bandwidth h was chosen using 5-fold cross validation as discussed in Section 2.3. For KGNG2, the same bandwidth was used for Step 1 as for Step 0. We used the proposed BIC criterion in Section 2.3 to tune (θ_1,θ_2) and θ^*. In addition, we compared the proposed methods with a conventional method based on the confidence bands, denoted as the CB method. Specifically, in the CB method, we first constructed confidence bands based on the initial estimates of time-varying coefficients as done in Tian, Zucker, and Wei (2005). When the zero-line was contained in the estimated confidence band, we classified the corresponding covariate as a null-effect covariate. If a constant-line but not the zero-line, was contained in the estimated confidence

Table 1. Variable selection and estimation results for $p = 10$. MSE stands for mean-squared error. Standard deviations of the Monte Carlo estimates are given in parentheses.

n	c_p	method	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8	Z_9	Z_{10}	MSE (SD)
200	20	KGNG	7	100	99	9	5	10	10	100	11	7	0.180 (0.107)
		KGNG2	3	100	100	4	2	5	4	100	6	1	0.172 (0.126)
		AGLASSO	6	100	73	5	2	2	8	100	5	4	0.444 (0.103)
		CB	0	87	92	1	0	0	0	88	1	1	
400	20	KGNG	1	100	100	4	2	3	6	100	4	3	0.114 (0.053)
		KGNG2	0	100	100	0	1	0	1	100	4	0	0.114 (0.054)
		AGLASSO	0	100	100	2	2	0	1	100	2	0	0.268 (0.055)
		CB	3	100	100	1	1	0	5	100	2	0	
200	40	KGNG	10	100	98	14	10	9	11	100	14	5	0.244 (0.151)
		KGNG2	7	100	100	7	5	3	5	100	11	3	0.227 (0.176)
		AGLASSO	7	100	82	9	4	4	6	100	6	1	0.492 (0.122)
		CB	0	33	26	0	0	0	0	29	0	0	
400	40	KGNG	5	100	100	3	3	4	4	100	7	2	0.117 (0.080)
		KGNG2	2	100	100	1	1	0	1	100	3	1	0.110 (0.126)
		AGLASSO	1	100	100	2	0	0	3	100	5	1	0.303 (0.096)
		CB	0	100	98	0	0	0	2	98	1	0	

band, we classified the corresponding covariate as a constant-effect covariate. If these conditions did not hold, we classified the corresponding covariate as a time-varying-effect covariate.

Tables 1 and 2 summarize the mean squared errors and variable selection results for $p = 10$ and 50, respectively. The selection frequency of each variable over 100 runs is reported, where the important covariates are Z_2, Z_3, and Z_8. The mean squared error (MSE) was calculated as

$$\frac{1}{100}\sum_{i=1}^{100}\left\{\widehat{\boldsymbol{\beta}}(t_i)-\boldsymbol{\beta}_0(t_i)\right\}^{\top} V\left\{\widehat{\boldsymbol{\beta}}(t_i)-\boldsymbol{\beta}_0(t_i)\right\},$$

where $\{t_1, \ldots, t_{100}\}$ are 100 equally-spaced grid points in the time interval $(0, 2)$ and V is the population covariance matrix of covariates.

From Tables 1 and 2, we make the following observations. First, KGNG2 shows the best performance in terms of variable selection and MSE in almost all scenarios, especially for the high-dimension case. This is expected since KGNG2 is a two-stage approach, first excluding the noise variables. KGNG and KGNG2 outperform AGLASSO with deduction in MSE as large as 60%, and they select Z_2, Z_3, and Z_8 as important covariates nearly all the times while AGLASSO misses Z_3 occasionally, especially when the sample size is small. When $p = 10$, the CB method overall has comparable performance as the proposed methods except for $n = 200$ and $c_p = 40\%$, where the CB method misses the important

Table 2. Variable selection and estimation results for $p = 50$. MSE stands for mean-squared error. Standard deviations of the Monte Carlo estimates are given in parentheses.

n	c_p	Method	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8	Z_9	Z_{10}	$Z_{11} - Z_{50}$	MSE (SD)
200	20	KGNG	5	100	100	3	3	3	1	100	3	3	2.9	0.297 (0.095)
		KGNG2	3	99	87	3	3	3	3	100	3	2	2.4	0.291 (0.130)
		AGLASSO	10	100	79	4	5	3	7	100	7	2	3.3	0.425 (0.099)
		CB	0	0	0	0	0	0	0	0	0	0	0.0	
400	20	KGNG	1	100	100	2	1	2	0	100	3	2	2.4	0.168 (0.056)
		KGNG2	0	100	100	2	0	1	0	100	1	0	0.6	0.134 (0.053)
		AGLASSO	3	100	100	2	1	1	4	100	0	2	1.2	0.261 (0.053)
		CB	0	2	0	0	0	0	0	1	0	0	0.0	
200	40	KGNG	2	97	92	3	5	4	2	100	5	3	3.9	0.403 (0.147)
		KGNG2	2	98	87	4	3	2	2	100	5	1	2.0	0.367 (0.222)
		AGLASSO	12	100	91	12	11	6	10	100	11	9	6.2	0.494 (0.137)
		CB	0	0	0	0	0	0	0	2	0	1	0.0	
400	40	KGNG	0	100	100	4	0	1	2	100	2	5	2.1	0.195 (0.058)
		KGNG2	0	100	100	1	0	0	1	100	1	0	1.3	0.181 (0.076)
		AGLASSO	5	100	100	4	1	2	3	100	6	2	4.3	0.292 (0.079)
		CB	0	0	0	0	0	0	0	0	0	0	0.0	

covariates frequently over simulations. When $p = 50$, the performance of the CB method is very bad since it cannot identify any important covariates in almost all simulations.

Table 3 summarizes the structure selection result of the three covariates with nonzero effect. We report the frequencies of each covariate being classified into the categories I_O, I_C and I_{NC}. In summary, the proposed KGNG and KGNG2 outperform AGLASSO for all covariates, exhibiting the most significant improvement for Z_2. The AGLASSO tends to falsely select the time-varying-effect Z_2 as a constant-effect covariate. For example, when $p = 10$, $n = 200$, and $c_p = 20\%$, AGLASSO correctly classifies Z_2 only 34 times out of 100, while KGNG and KGNG2 classify Z_2 correctly more than 80 times. Similarly, the CB method has very poor structure selection results when $p = 50$.

We plot the initial estimator, KGNG, and AGLASSO for the nonzero coefficients and their pointwise 95% confidence intervals based on 100 simulations for $p = 10$. The plots for $c_p = 20\%$ and $c_p = 40\%$ are given in Figures 1 and 2, respectively. Here the performance of KGNG and AGLASSO improve substantially when the sample size increases. KGNG has smaller biases in the estimation of time-varying coefficients than does AGLASSO in all cases, while KGNG and AGLASSO give comparable estimates for the constant coefficient. The improvement of KGNG over the initial estimator for time-varying coefficients is not obvious, but, the improvement for the constant coefficient is significant. This agrees with our Theorem 3, parts (c) and (d).

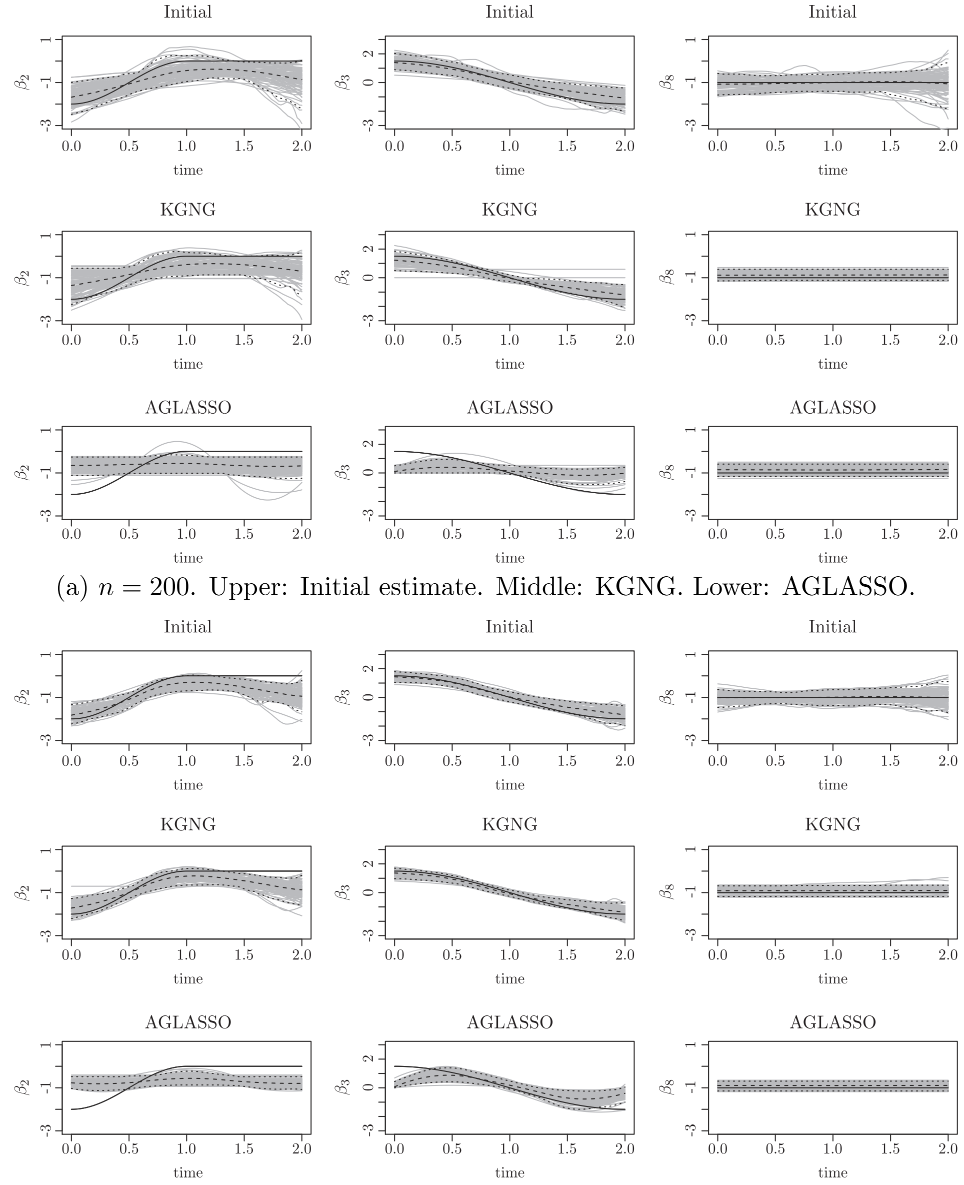
Table 3. Structure selection results for covariates 2, 3 and 8. Here O is for null effect, C for constant effect, and NC for time-varying effect.

			p=10									**p=50**								
			Z_2			Z_3			Z_8			Z_2			Z_3			Z_8		
n	c_p	method	O	C	NC	O	C	NC	O	C	NC	O	C	NC	O	C	NC	O	C	NC
200	20	KGNG	0	19	81	1	1	98	0	97	3	0	62	38	0	1	99	0	99	1
		KGNG2	0	10	90	0	1	99	0	94	6	1	38	61	13	0	87	0	95	5
		AGLASSO	0	66	34	27	2	71	0	86	14	0	56	44	21	0	79	0	92	8
		CB	13	79	8	8	42	50	12	88	0	100	0	0	100	0	0	100	0	0
400	20	KGNG	0	1	99	0	0	100	0	97	3	0	1	99	0	0	100	0	97	3
		KGNG2	0	1	99	0	0	100	0	91	9	0	0	100	0	0	100	0	97	3
		AGLASSO	0	51	49	1	0	99	0	96	4	0	41	59	0	0	100	0	95	5
		CB	0	8	92	0	0	100	0	100	0	98	2	0	100	0	0	99	1	0
200	40	KGNG	0	31	69	2	4	94	0	97	3	3	57	40	8	11	81	0	98	2
		KGNG2	0	23	77	0	5	95	0	89	11	2	27	71	13	1	86	0	92	8
		AGLASSO	0	75	25	18	10	72	0	84	16	0	50	50	9	9	82	0	91	9
		CB	67	33	0	74	23	3	71	29	0	100	0	0	100	0	0	98	2	0
400	40	KGNG	0	0	100	0	0	100	0	97	3	0	16	84	0	0	100	0	99	1
		KGNG2	0	0	100	0	0	100	0	95	5	0	5	95	0	0	100	0	98	2
		AGLASSO	0	35	65	0	0	100	0	94	6	0	4	96	0	0	100	0	94	6
		CB	0	55	45	2	13	85	2	98	0	100	0	0	100	0	0	100	0	0

4.2. Analysis of primary biliary cirrhosis (PBC) data

We applied our KGNG and KGNG2 methods to analyze the PBC data (Fleming and Harrington (1991)). The data is from the Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver conducted between 1974 and 1984. The primary biliary cirrhosis is a chronic disease in which the bile ducts in one's liver are slowly destroyed. In the study, 312 out of 424 patients who participated in the randomized trial were eligible for the analysis. There are 17 covariates: trtmt=treatment (Yes/No), age (in 10 years), gender=female/male, ascites=presence of ascites (Yes/No), hypato=presence of hepatomegaly (Yes/No), spiders=presence of spiders, edema=severity of oedema (0 denotes no oedema, 0.5 denotes untreated or successfully treated oedema and 1 denotes unsuccessfully treated oedema), logbili=logarithm of serum bilirubin (mg/dl), chol=serum cholesterol (mg/dl), logalb=logarithm of albumin (gm/dl), copper=urine copper (mg/day), alk=alkaline phosphatase (U/liter), sgot=liver enzyme (U/ml), trig=triglicerides (mg/dl), platelet=platelets per 10^{-3} ml^3, logprotime=logarithm of prothrombin time (seconds), stage=histologic stage of disease (category: 1, 2, 3 or 4).

Model selection here has been previously studied in the context of the Cox model with time-independent coefficients (Tibshirani (1997); Zhang and Lu (2007)) and Cox model with time-varying coefficients (Tian, Zucker, and Wei (2005); Yan and Huang (2012)). To ease the comparison, we analyzed the data of 276 patients

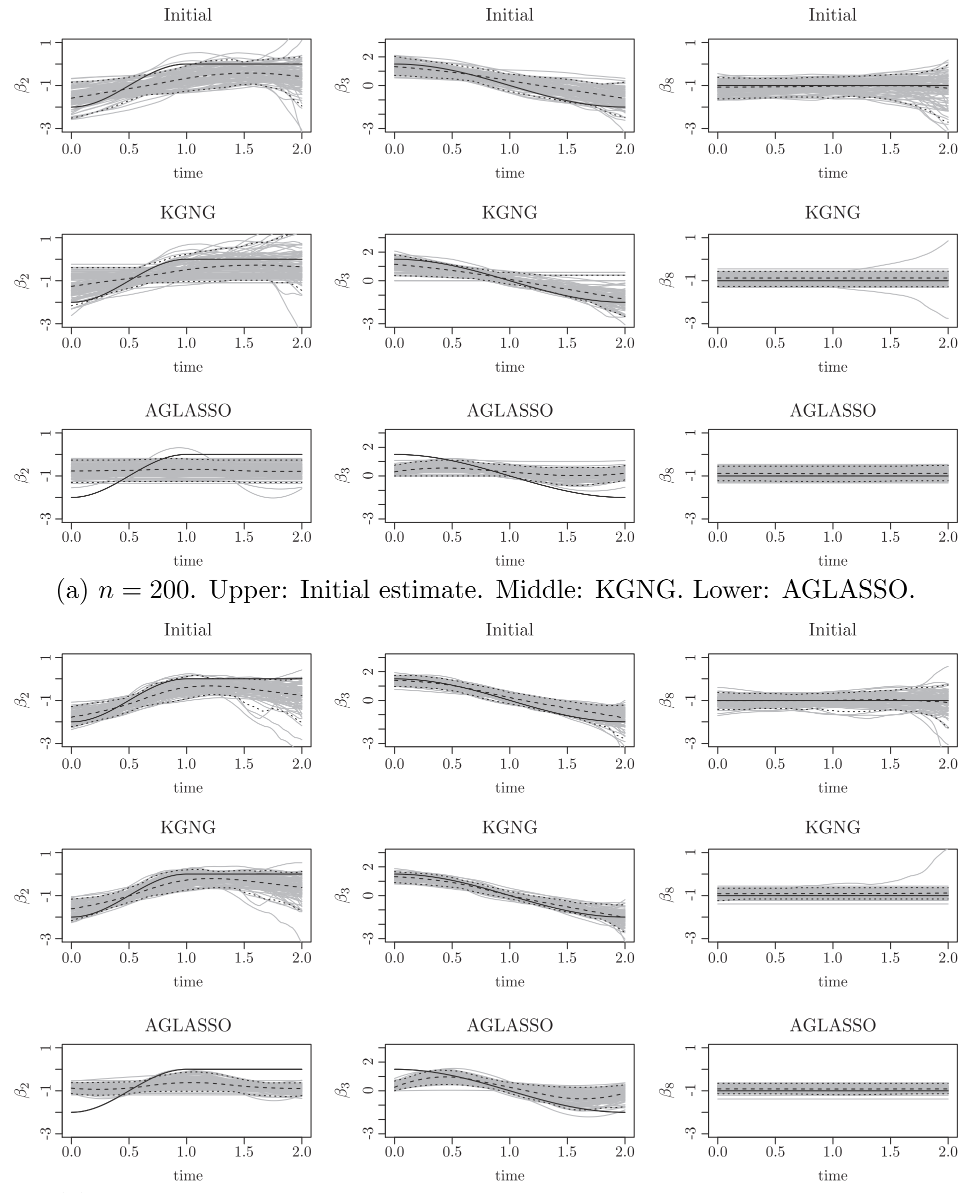


(a) $n = 200$. Upper: Initial estimate. Middle: KGNG. Lower: AGLASSO.

(b) $n = 400$. Upper: Initial estimate. Middle: KGNG. Lower: AGLASSO.

Figure 1. Estimated curves (gray) of the three nonzero coefficients from 100 replicates when $c_p = 20\%$ and $p = 10$. The dark lines are the true curves. The dashed lines are the average of 100 estimates. The dotted lines are the simulation-based pointwise 95% confidence intervals.

with no missingness in covariates and took log transformation of serum bilirubin, albumin, and prothrombin time, as did Yan and Huang (2012). We used 10-fold



(a) $n = 200$. Upper: Initial estimate. Middle: KGNG. Lower: AGLASSO.

(b) $n = 400$. Upper: Initial estimate. Middle: KGNG. Lower: AGLASSO.

Figure 2. Estimated curves (gray) of the three nonzero coefficients from 100 replicates when $c_p = 40\%$ and $p = 10$. The dark lines are the true curves. The dashed lines are the average of 100 estimates. The dotted lines are the simulation-based pointwise 95% confidence intervals.

cross validation to find the optimal bandwidth in the initial estimator, 2,000 (days). We chose $\tau = 3{,}200$, which covers around 90% of the observed survival

Table 4. Analysis results for PBC data. TV stands for time-varying coefficients.

Covariate	MPLE	ALASSO	AGLASSO	KGNG	KGNG2
trtmt	-0.062 (0.211)				
age	0.261 (0.113)	0.270 (0.124)	0.263 (0.126)	0.211 (0.111)	0.253 (0.113)
gender	-0.256 (0.317)				
ascites	0.162 (0.381)				
hypato	-0.100 (0.254)				
spiders	0.049 (0.243)				
edema	0.926 (0.378)	0.842 (0.410)	0.932 (0.443)	TV	TV
logbili	0.723 (0.162)	0.699 (0.115)	TV	0.716 (0.096)	0.718 (0.116)
chol	0.000 (0.000)				
logalb	-2.270 (0.947)	-2.538 (0.762)	-2.440(0.789)	-2.173 (0.934)	-2.294 (1.183)
copper	1.694 (1.251)	2.218 (1.236)	2.089(1.261)	TV	1.379 (1.419)
alk	0.000 (0.000)				
sgot	0.003 (0.002)				
trig	-0.002 (0.001)				
platelet	0.001 (0.001)				
logprotime	2.335 (1.321)	2.099 (1.241)	1.822 (1.249)	TV	TV
stage	0.381 (0.176)	0.274 (0.140)	0.278 (0.143)	0.244 (0.110)	0.218 (0.130)

times. Table 4 gives the estimates of coefficients by five methods: the maximum partial likelihood estimator (MPLE), the adaptive LASSO (ALASSO) estimator of Zhang and Lu (2007) based on a standard Cox model, the AGLASSO, and KGNG and KGNG2 based on a time-varying coefficient Cox model. The numbers given in parenthesis are the estimated standard errors for important constant coefficients selected by each method. The results for ALASSO and AGLASSO are copied directly from Yan and Huang (2012). Here ALASSO, AGLASSO, KGNG, and KGNG2 all select the same seven important covariates: age, cooper, edema, logbili, logalb, logprotime, and stage. KGNG identifies three covariates with time-varying coefficients and KGNG2 identifies two, in which edema and logprotime are the common covariates. On the other hand, AGLASSO only selects logbilli as the covariate with time-varying coefficient. These results partly agree with the findings of Tian, Zucker, and Wei (2005), where only 5 covariates age, edema, logbili, logalb and logprotime were considered in their time-varying coefficient Cox model, and three covariates edema, logprotime and logbili were identified as having time-varying coefficients. In Figures 3−4, we plotted the estimated coefficients by the initial step (maximum local partial likelihood estimator), KGNG, and KGNG2 for the seven important covariates and their associated 95% pointwise confidence intervals and simultaneous confidence bands.

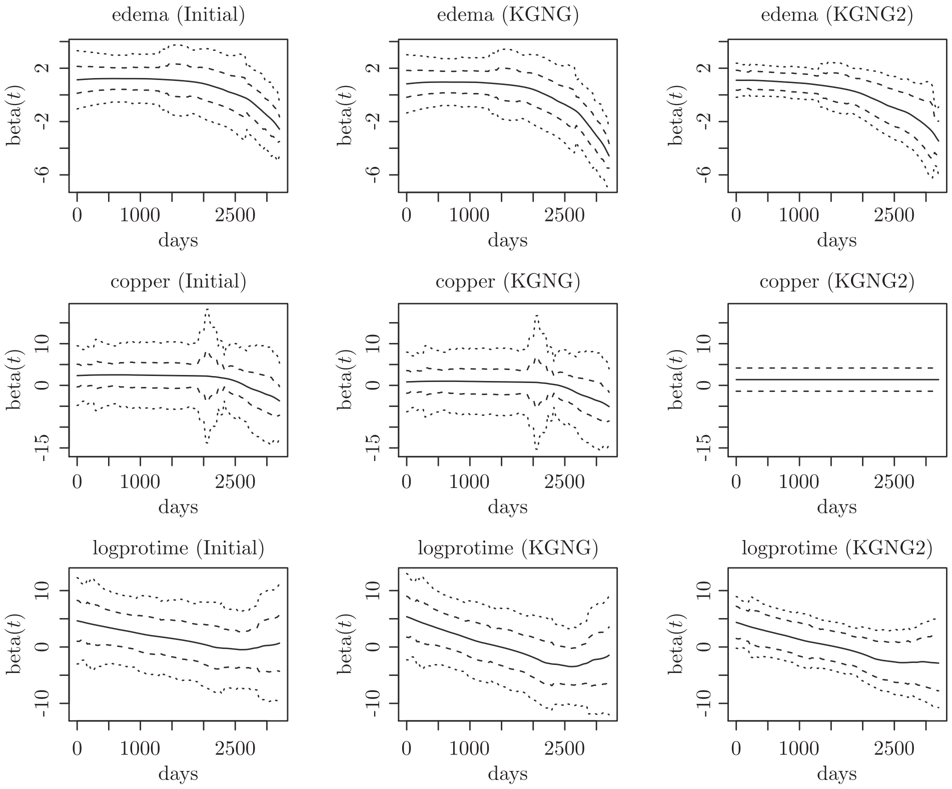


Figure 3. Estimated coefficients for covariates: edema, copper, and logprotime. Left panel: initial estimator in KGNG; Middle panel: KGNG; Right panel: KGNG2. Solid lines: estimated curves; Dashed lines: 95% pointwise confidence intervals; Dotted lines: 95% simultaneous confidence bands.

5. Discussion

We propose a kernel group nonnegative garrote (KGNG) estimation method and its variant (KGNG2) for automatic structure selection and coefficient estimation in a time-varying coefficient Cox model. We establish the asymptotic properties, including structure selection consistency and asymptotic distributions, of our estimators for both constant and time-varying coefficients. Numerical studies have shown the competitive performance of the proposed methods compared with existing approaches.

We have focused on the case with fixed dimension p, with p smaller than n. For the $p > n$ case, a penalty term needs to be added to (2.2) to get reasonable initial estimates of the coefficient functions. Then Step 2 and 3 can follow, as proposed in this paper. If $p \gg n$, a screening procedure can be utilized to remove the noisy covariates beforehand. Then the dimension of the model can be decreased to a value that can be handled directly. However, a screening

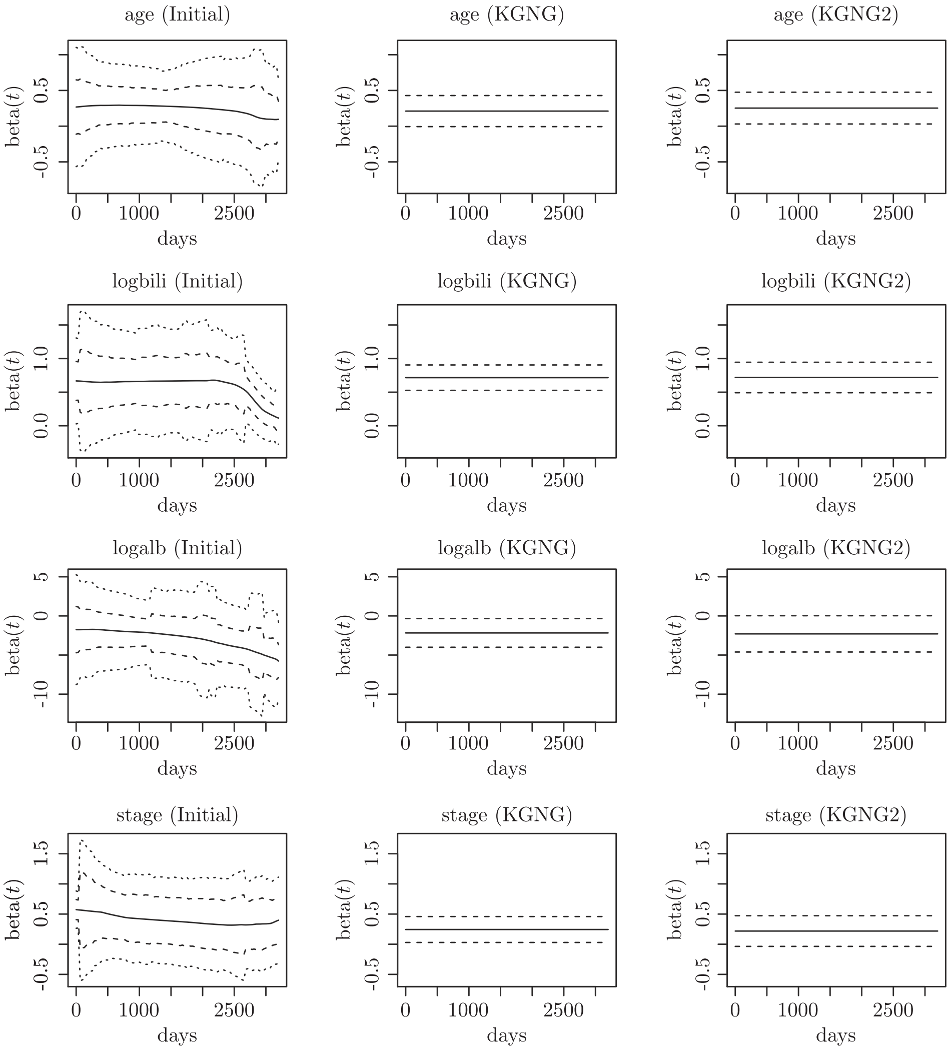


Figure 4. Estimated coefficients for covariates: age, logbili, logalb and stage. Left panel: initial estimator in KGNG; Middle panel: KGNG; Right panel: KGNG2. Solid lines: estimated curves; Dashed lines: 95% pointwise confidence intervals; Dotted lines: 95% simultaneous confidence bands.

procedure for the time-varying coefficient Cox model has yet to be developed; this needs further investigation.

Since the proposed procedure depends on a large number of tuning parameters, it is worthwhile to develop a statistical test to check the goodness-of-fit of the final estimated model. We think that a cumulative sums of martingale residuals-based goodness-of-fit test can be derived for the final estimated model, following the techniques of Lin, Wei, and Ying (1993). A similar goodness-of-fit test procedure was developed for the Dantzig selector in Cox's proportional hazards model (Antoniadis, Fryzlewicz, and Letué (2010)). This is an interesting topic that needs further investigation.

Another interesting problem, as suggested by a referee, is to extend the proposed methods a procedure that estimates the coefficients as zero on parts of the time domain, and as nonzero (and time-varying) on the remaining parts. A simple solution would be to chop the coefficient functions evenly into small pieces on the study time domain and then apply the group nonnegative garrote penalty to identify the significant pieces. However, when there are a large number of covariates, this approach is computationally challenging. The proposed KGNG/KGNG2 methods for structure selection can be regarded as a preliminary step for domain selection since they can help to remove all the covariates with null or constant effects and thus achieve effective dimension reduction. Then, the domain selection can focus only on the selected covariates with truly time-varying coefficients. This is an extension that warrants future research.

Supplementary Materials

The online supplementary materials contain the proofs of Theorem 1, 2, and 3 in the main paper.

Acknowledgement

We thank the Editor, an associate editor, and two reviewers for their constructive comments on improving the paper. We acknowledge support from the National Institutes of Health grant RO1 CA140632 and National Science Foundation grants DMS-1418172, DMS-1309507, and DBI-1261830.

References

Ahmad, I., Leelahanon, S. and Li, Q. (2005). Efficient estimation of a semiparametric partially linear varying coefficient model. *Ann. Statist.* **33**, 258-283.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* (Edited by Tsahkadsor), 267-281. Akadémiai Kiadó, Budapest.

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *Ann. Statist.* **10**, 1100-1120.

Antoniadis, A., Fryzlewicz, P., and Letué, F. (2010). The Dantzig selector in Cox's proportional hazards model. *Scand. J. Stat.* **37**, 531–552.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics* **37**, 373-384.

Cai, Z., Fan, J. and Li, R. (2000). Efficient estimation and inferences for varying-coefficient models. *J. Amer. Statist. Assoc.* **95**, 888-902.

Cai, Z. and Sun, Y. (2003). Local linear estimation for time-dependent coefficients in Cox's regression models. *Scand. J. Statist.* **30**, 93-111.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *J. Roy. Statist. Soc. B* **34**, 187-220.

Fan, J. and Huang, T. (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli* **11**, 1031-1057.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348-1360.

Fan, J. and Li, R. (2002). Variable selection for Cox proportional hazards model and frailty model. *Ann. Statist.* **30**, 74-99.

Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis.* John Wiley, New York.

Goeman, J. J. (2010). L1 penalized estimation in the Cox proportional hazards model. *Biometrical Journal* **52**, 70-84.

Huang, J. Z., Wu, C. O. and Zhou, L. (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika* **89**, 111-128.

Hu, T. and Xia, Y. (2012). Adaptive semi-varying coefficient model selection. *Statist. Sinica* **22**, 575-599.

Lin, D. Y., Wei, L. J. and Ying, Z. (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika* **80**, 557-572.

Liu, M., Lu, W., Shore, R. E. and Zeleniuch-Jacquotte, A. (2010). Cox regression model with time-varying coefficients in nested case-control studies. *Biostatistics* **11**, 693-706.

Mallows, C. L. (1973). Some comments on Cp. *Technometrics* **15**, 661–675.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-464.

Tian, L., Zucker, D. and Wei, L. J. (2005). On the Cox model with time-varying regression coefficients. *J. Amer. Statist. Assoc.* **100**, 172-183.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statist. Medicine* **16**, 385-395.

Wang, H. and Xia, Y. (2009). Shrinkage estimation of the varying coefficient model. *J. Amer. Statist. Assoc.* **104**, 747-757.

Wang, H. J., Zhu, Z. and Zhou, J. (2009). Quantile regression in partially linear varying coefficient models. *Ann. Statist.* **37**, 3841C-3866.

Yan, J. and Huang, J. (2012). Model selection for Cox models with time-varying coefficients. *Biometrics* **68**, 419-428.

Yu, Z. and Lin, X. (2010). Semiparametric regression with time-dependent coefficients for failure time data analysis. *Statist. Sinica* **20**, 853-869.

Zhang, H. H., Cheng, G. and Liu, Y. (2011). Linear or nonlinear? Automatic structure discovery for partially linear models. *J. Amer. Statist. Assoc.* **106**, 1099-1112.

Zhang, H. H. and Lu, W. (2007). Adaptive Lasso for Cox's proportional hazards model. *Biometrika* **94**, 691-703.

Zhang, W., Lee, S. Y. and Song, X. (2002). Local polynomial fitting in semivarying coefficient model. *J. Multivariate Anal.* **82**, 166-188.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418-1429.

Zucker, D. M. and Karr, A. F. (1990). Nonparametric survival analysis with time-dependent covariate effects: a penalized partial likelihood approach. *Ann. Statist.* **18**, 329-353.

Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203, U.S.A.

E-mail: wxiao@ncsu.edu

Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203, U.S.A.

E-mail: lu@stat.ncsu.edu

Department of Mathematics and Statistics Interdisciplinary Program, University of Arizona, Tucson, AZ 85721-0089, U.S.A.

E-mail: hzhang@math.arizona.edu

(Received April 2013; accepted March 2015)