# A Soft-Thresholding Operator for Sparse Time-Varying Effects in Survival Models

Yuan Yang, Jian Kang and Yi Li

**Abstract** We consider a class of Cox models with time-dependent effects that may be zero over certain unknown time regions or, in short, sparse time-varying effects. The model is particularly useful for biomedical studies as it conveniently depicts the gradual evolution of effects of risk factors on survival. Statistically, estimating and drawing inference on infinite dimensional functional parameters with sparsity (e.g., time-varying effects with zero-effect time intervals) present enormous challenges. To address them, we propose a new soft-thresholding operator for modeling sparse, piecewise smooth and continuous time-varying coefficients in a Cox time-varying effects model. Unlike the common regularized methods, our approach enables one to estimate non-zero time-varying effects and detect zero regions simultaneously, and construct a new type of sparse confidence intervals that accommodate zero regions. This leads to a more interpretable model with a straightforward inference procedure. We develop an efficient algorithm for inference in the target functional space, show that the proposed method enjoys desired theoretical properties, and present its finite sample performance by way of simulations. We apply the proposed method to analyze the data of the Boston Lung Cancer Survivor Cohort, an epidemiological cohort study investigating the impacts of risk factors on lung cancer survival, and obtain clinically useful results.

Yuan Yang
Parexel, Waltham: Anna.Yang@parexel.com

Jian Kang
University of Michigan, Ann Arbor: jiankang@umich.edu

Yi Li
University of Michigan, Ann Arbor: yili@umich.edu

# 1 Introduction

The Cox proportional hazards model, proposed by the late Sir D.R. Cox [7], has dominated the analysis of survival studies for decades and has also become an indispensable analytical tool in the era of precision medicine, because of its elegant estimation and inference framework and ease of interpretation [14]. One key assumption of the proportional hazards model is that the effect of a given covariate remains constant over time, which, however, may not always hold. In fact, non-proportionality has been commonly observed and sparked much interest [8, 15, 18, 19, 20, 29], which led to the development of time-dependent coefficients Cox models [12].

An often overlooked feature in time-dependent effects Cox models is the sparsity associated with time-varying effects, meaning that the covariate effects can be zero on some specific time intervals and non-zero but time-varying on the others. For example, Anderson and Gill (1982) [1] noticed the effects of some covariates disappeared in the later follow-up in a vulvar cancer study; Gore et al. (1982) [11] found that the influence of signs recorded at diagnosis waned with time in a breast cancer study; Tian et al. (2005) [25] noted sparsity in the edema effect during the early stage and also showed the effect of prothrombin on survival diminished over time in a biliary cirrhosis study. In the motivating Boston Lung Cancer Survivor Cohort [6], an epidemiological study investigating the impacts of clinical and molecular risk factors on lung cancer survival, chemotherapy and radiotherapy did not seem to increase or decrease patients' overall survival, leading to the notion of detecting zero-effect regions for these treatment options.

It is rather challenging to detect no-effects periods and estimate non-zero effects simultaneously, as the existing methods for fitting the time-dependent Cox models cannot achieve these goals. For example, the commonly used penalized spline models [13, 17, 31, 33] and kernel weighted likelihood approaches [3, 25] can detect or label covariates as time-varying or time-constant, but cannot detect no-effects periods within each covariate effect's trajectory.

We propose a new statistical method that can efficiently model sparse time-varying effects in a survival setting, by using a soft-thresholding operator to represent the time-varying effects in the Cox model. Both soft-thresholding and hard-thresholding approaches can be applied in this setting and have their own merits. However, we opt for soft-thresholding because it respects the continuity of the effect with respect to time and may conveniently depict the gradual evolution of effects of risk factors on survival. Indeed, the concept of soft thresholding was introduced by Donoho (1994, 1995) [9, 10], who applied this estimator to the coefficients of a wavelet transform of a function measured with noise. Since then, the use of soft-thresholding for effect shrinkaging has been flourishing: Chang et al. (2000) [4] proposed an adaptive, data-driven thresholding method for image denoising in a Bayesian framework; Tibshirani (1996) [26] pointed out that the Lasso estimator is a soft-thresholding estimator when the covariate matrix has an orthonormal design. Kang et al. (2018) [16] used a soft-thresholding operator for modeling sparse, continuous, and piecewise smooth functions in image data analysis; however, as their

method was not developed for survival analysis, it is unclear whether its extension to a survival setting is feasible.

We propose a soft-thresholding operator to model time-varying effects of covariates in a survival regression setting, and use the B-splines to approximate the nonparametric parts. Estimation is carried out by maximizing a penalized and smoothed partial likelihood. We prove the asymptotic properties of our proposed estimator, and introduce a new class of sparse confidence intervals for quantifying the uncertainty of the sparse functional estimates.

This chapter is organized as follows. In Section 2, we introduce the proposed soft-thresholding operator for a Cox model with sparse time-varying effects and derive an algorithm for fitting the model. Section 3 lists the theoretical properties of the method and proposes the sparse confidence intervals for inferring from the estimated sparse time-varying effects. We present simulation results in Section 4 to assess the finite sample performance of our methods, and analyze the aforementioned Boston Lung Cancer study in Section 5. Section 6 concludes this chapter with a brief summary. We defer all the technical proofs to the Appendix.

## 2 Methods

### *2.1 Model*

Let $T_i^u$ and $T_i^c$ represent the survival and censoring times, respectively, for the $i$th patient. Observed are $T_i = T_i^u \wedge T_i^c$ with $a \wedge b = \min\{a,b\}$, and the death indicators $\Delta_i = \mathscr{I}(T_i^u \leq T_i^c)$ with $\mathscr{I}(A)$ indicating whether condition $A$ holds ($=1$) or not ($=0$). Let $\boldsymbol{Z}_i = (Z_{i1},\ldots,Z_{ip})^\top$ be a $p$-dimensional covariate vector for sample $i$. The observed data consist of $n$ independent vectors, $(T_i,\Delta_i,\boldsymbol{Z}_i)$, which are identical and independently distributed (i.i.d.) copies of $(T,\Delta,\boldsymbol{Z})$. Further, $(T_i^u,T_i^c), i=1,\ldots,n$, are i.i.d. copies of $(T^u,T^c)$.

Denote by $\lambda(t|\boldsymbol{Z}_i)$ the hazard function at $t$ given $\boldsymbol{Z}_i$. A time-varying effects Cox model stipulates that

$$\lambda(t \mid \boldsymbol{Z}_i) = \lambda_0(t)\exp\{\boldsymbol{Z}_i^\top \boldsymbol{\beta}(t)\},$$

where $\lambda_0(t)$ is the baseline hazard, and $\boldsymbol{\beta}(t) = \{\beta_1(t),\ldots,\beta_p(t)\}^\top$ are the $p$ time-dependent coefficients corresponding to $\boldsymbol{Z}_i$.

The log partial likelihood with time-varying coefficients is

$$\mathrm{PL}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \Delta_i \left\{ \sum_{j=1}^{p} Z_{ij}\beta_j(T_i) - \log\left[ \sum_{l\in R_i} \exp\left\{ \sum_{j=1}^{p} Z_{lj}\beta_j(T_i) \right\} \right] \right\}, \qquad (1)$$

where $R_i = \{l : T_l > T_i\}$ is the risk set at $T_i$.

We assume that $\beta_j(\cdot), j=1,\ldots,p$, is continuous everywhere, with zero-effect regions ($\mathscr{R}_0$) consisting of at least one interval, and is smooth over regions (positive

$\mathscr{R}_+$ and negative $\mathscr{R}_-$) where its effect is non-zero. On each interval of the non-zero regions, the $d$th derivative of $\beta_j(t)$ exists and satisfies the Lipschitz condition. That is, for any $s,t$ in the interval, there exists a constant $C > 0$ such that

$$|\beta_j^{(d)}(s) - \beta_j^{(d)}(t)| \leq C|s-t|^w, \tag{2}$$

where $d$ is a non-negative integer, and $w \in (0,1]$ such that $m \equiv d + w > 0{\cdot}5$. Let $\mathbb{H}$ be the set of all such functions. Often, we use $d = 3$ (as in our later simulations and data analysis) corresponding to piecewise cubic functions. Let $\boldsymbol{\beta}_0(\cdot) = \{\beta_{01}(\cdot), \ldots, \beta_{0p}(\cdot)\}^\top$ be the true coefficient vector in the model that generates the observed data and $\beta_{0j} \in \mathbb{H}$.

We use the soft-thresholding operator $\zeta$ to represent a varying coefficient with zero regions:

$$\zeta\{\theta(t), \alpha\} = \{\theta(t) - \alpha\}\mathscr{I}\{\theta(t) > \alpha\} + \{\theta(t) + \alpha\}\mathscr{I}\{\theta(t) < -\alpha\},$$

where $\alpha > 0$ is the thresholding parameter and $\theta(t)$ is a real-valued function.

To avoid technicalities at the tail of the distribution of $T^u$, we estimate $\beta_j(\cdot)$ over a finite interval $(0,\tau)$ and base the estimation on the partial likelihood over the same interval, where $\tau$ is within the support of $T$. By doing so, we need to effectively replace $T_i$ and $\Delta_i$ in likelihood (1) (and also the modified partial likelihoods thereafter) by $T_i = \min(T_i^u, T_i^c, \tau)$ and $\Delta_i = I(T_i^u \leq T_i^c \wedge \tau)$. In practice, if $\tau$ is chosen to be the maximum observed survival time in the data, no such replacements are needed.

It can be shown that for any function $\beta(t) \in \mathbb{H}$ and any $\alpha > 0$, there exists at least one $\theta(t) \in \mathbb{F}_0$ such that $\beta(t) = \zeta\{\theta, \alpha\}(t)$, where $\mathbb{F}_0$ is the class of functions $\theta$ defined on $[0,\tau]$, with the $d$th derivative $\theta^{(d)}$ satisfying the Lipschitz condition (2). As such, we introduce a new penalized likelihood for estimation

$$\begin{aligned}&\mathrm{PL}(\boldsymbol{\theta})\\ &= \sum_{i=1}^n \Delta_i \left\{ \sum_{j=1}^p Z_{ij}\zeta\{\theta_j(T_i), \alpha_j\} - \log\left[\sum_{l\in R_i} \exp\left\{\sum_{j=1}^p Z_{lj}\zeta\{\theta_j(T_l), \alpha_j\}\right\}\right]\right\} - \rho||\boldsymbol{\theta}||_2^2,\end{aligned} \tag{3}$$

where $\boldsymbol{\theta}(t) = \{\theta_1(t), \ldots, \theta_p(t)\}^\top$ and $\rho > 0$ is the predetermined penalization coefficient.

With the soft-thresholding representation, we can convert the problem from estimating non-smooth functions to estimating smooth functions. Among many approaches for modeling smooth functions, we will utilize the B-spline basis approach because of its convenience and numerical stability [28]; other alternatives may include P-splines and smoothing splines. Let $\mathbb{F}$ be the B-spline function sieve space, $K = O(n^\nu)$ be an integer with $0 < \nu < 0{\cdot}5$, and $B_k(t)(1 \leq k \leq q$, and $q = K + d)$ be the B-spline basis functions of degree $d+1$ associated with the knots $0 = t_0 < t_1 < \cdots < t_{K-1} < t_K = 1$, satisfying $\max_{1\leq k\leq K}(t_k - t_{k-1}) = O(n^{-\nu})$. Let $\boldsymbol{B}(t) = \{B_1(t), \ldots, B_q(t)\}^\top$ be a functional vector of the B-spline bases; with $d = 3$, this corresponds to a vector of cubic B-spline bases. Then, we have

$$\mathbb{F} = \left\{ \theta(t) : \theta(t) = \sum_{k=1}^{q} \gamma_k B_k(t), t \in [0,\tau], \gamma_k \in \mathbb{R}, k = 1,\ldots,q \right\}.$$

For given $\alpha$ and $q$, we define the thresholding sieve space

$$\mathbb{S}_{q,\alpha} = \left\{ \beta(t) = \zeta\{\theta(t),\alpha\} : \theta(t) = \sum_{k=1}^{q} \gamma_k B_k(t), t \in [0,\tau], \gamma_k \in \mathbb{R}, k = 1,\ldots,q \right\}.$$

Let $\boldsymbol{\gamma}_j = (\gamma_{j1},\ldots,\gamma_{jq})^\top$ be the basis coefficients for $\theta_j(t)$. Then we represent $\theta_j(t) = \boldsymbol{B}(t)^\top \boldsymbol{\gamma}_j$. The penalized log partial likelihood can be written as

$$\begin{aligned} \mathrm{PL}(\boldsymbol{\gamma}) = &\sum_{i=1}^{n} \Delta_i \left\{ \sum_{j=1}^{p} Z_{ij} \zeta\{\boldsymbol{B}(T_i)^\top \boldsymbol{\gamma}_j, \alpha_j\} - \log \left[ \sum_{l \in R_i} \exp \left\{ \sum_{j=1}^{p} Z_{lj} \zeta\{\boldsymbol{B}(T_i)^\top \boldsymbol{\gamma}_j, \alpha_j\} \right\} \right] \right\} \\ &- \rho \sum_{j=1}^{p} \sum_{i=1}^{n} \{\boldsymbol{B}(T_i)^\top \boldsymbol{\gamma}_j\}^2, \end{aligned} \tag{4}$$

where $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1,\ldots,\boldsymbol{\gamma}_p)$.

## *2.2 Estimation*

It is challenging to directly maximize the likelihood function (4) as the thresholding operator $\zeta(\theta,\alpha)$ is non-smooth. We therefore consider a smooth approximation of $\zeta(\theta,\alpha)$:

$$\begin{aligned} h_\eta\{\theta(t),\alpha\} = \frac{1}{2} \Bigg( &\left[ 1 + \frac{2}{\pi} \arctan\{\theta_-(t)/\eta\} \right] \theta_-(t) + \\ &\left[ 1 - \frac{2}{\pi} \arctan\{\theta_+(t)/\eta\} \right] \theta_+(t) \Bigg), \end{aligned}$$

where $\alpha > 0$, $\eta > 0$ and $\theta_\pm(t) = \theta(t) \pm \alpha$. Noting $\lim_{\eta \to 0} h_\eta\{\theta(t),\alpha\} = \xi(\theta,\alpha)$, we define $h_0\{\theta(t),\alpha\} = \xi(\theta,\alpha)$. As such, $h_\eta\{\theta(t),\alpha\}$ is a sufficiently smooth function in $\eta$ and, in particular, in a small neighborhood, e.g., $\eta \in [0,\varepsilon)$ where $\varepsilon > 0$ is small. Taking a Taylor expansion of $h_\eta\{\theta(t),\alpha\}$ at $\eta = 0$ within this neighborhood, we can show that the approximation error between $h_\eta\{\theta(t),\alpha\}$ and $\zeta(\theta,\alpha)$ is bounded by $\eta + O(\eta^3)$. We drop $\eta$ hereafter for simplicity of notation. Then, we obtain a smoothed log partial likelihood function:

$$\mathrm{PL}(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \Delta_i \left\{ \sum_{j=1}^{p} Z_{ij} h\{\boldsymbol{B}(T_i)^\top \boldsymbol{\gamma}_j, \alpha_j\} - \log \left[ \sum_{l \in R_i} \exp \left\{ \sum_{j=1}^{p} Z_{lj} h\{\boldsymbol{B}(T_i)^\top \boldsymbol{\gamma}_j, \alpha_j\} \right\} \right] \right\} - \rho \sum_{j=1}^{p} \sum_{i=1}^{n} \{\boldsymbol{B}(T_i)^\top \boldsymbol{\gamma}_j\}^2, \quad (5)$$

forming the basis for estimation and inference.

Let $\tilde{\boldsymbol{\gamma}} = \arg\max_{\boldsymbol{\gamma}} \mathrm{E}_{T,\Delta,\boldsymbol{Z}} \mathrm{PL}(\boldsymbol{\gamma})$, where the expectation is taken with respect to the joint distribution of $T, \Delta$ and $\boldsymbol{Z}$ under the true parameter $\boldsymbol{\beta}_0(t)$. An estimate of $\tilde{\boldsymbol{\gamma}}$ is obtained by maximizing the likelihood (5) so that $\hat{\boldsymbol{\gamma}} = \arg\max_{\boldsymbol{\gamma}} \mathrm{PL}(\boldsymbol{\gamma})$. Then an estimate of $\boldsymbol{\beta}(t)$ is given by $\hat{\boldsymbol{\beta}}(t) = \{\hat{\beta}_1(t), \ldots, \hat{\beta}_p(t)\}^\top$ with $\hat{\beta}_j(t) = \zeta(\boldsymbol{B}(t)^\top \hat{\boldsymbol{\gamma}}_j, \alpha_j)$.

Optimizing $\mathrm{PL}(\boldsymbol{\gamma})$ can be implemented by gradient-based methods [5] and coordinate descent algorithms [30]. With appropriate initial values, global optimizers can be reached. Specifically, for each $j = 1, \ldots, p$, we obtain the non-varying coefficients $(a_1, \ldots, a_p)^\top$ from the Cox model, then we set the initial $\boldsymbol{\gamma}_j^{(0)}$ to be a vector of $a_j$ with length $q$. In practice, we recommend to vary the initial values and check the robustness of the final results. We choose the pre-specified parameters as follows. In theory, our method works for any $\alpha$; however, in practice, a value of $\alpha$ comparable to the scale of true coefficients works best. Thus, we set $\alpha_j$ to be $0.5 \times |a_j|$. The choices of $\eta$ and $\rho$ can be specified in accordance with Condition C6. The knots of B-spline are equally spaced over $[0, \tau]$. The number of basis functions, $q$, can be determined through $R$-fold cross-validation. That is, partition the full data $D$ into $R$ equal-sized groups, denoted by $D_r$, for $r = 1 \ldots, R$, and let $\hat{\boldsymbol{\beta}}_{-r}^{(q)}(t)$ be the estimate obtained with $q$ bases using all the data except for $D_r$. We obtain the optimal $q$ by minimizing the cross-validation error, which is the average of the negative objective function (1) evaluated at $\hat{\boldsymbol{\beta}}_{-r}^{(q)}(t)$ on $D_r$ with $r$ running from 1 to $R$.

## 3 Inference

We begin with some needed notation. First, for a $p_1 \times q_1$ matrix $A = (a_{ij})$ and a $p_2 \times q_2$ matrix $B = (b_{ij})$, their Kronecker product is defined to be

$$A \otimes B = \begin{pmatrix} a_{11}B & \ldots & a_{1q_1}B \\ \ldots & \ldots & \ldots \\ a_{p_1 1}B & \ldots & a_{p_1 q_1}B \end{pmatrix}.$$

With that, we define the following:

$$
\begin{aligned}
g(\boldsymbol{\beta},\boldsymbol{Z},t) &= \sum_{j=1}^{p} Z_j \beta_j(t), \\
g_n(\boldsymbol{\gamma},\boldsymbol{Z},t) &= \sum_{j=1}^{p} Z_j h_j(\boldsymbol{B}(t)\boldsymbol{\gamma}_j), \\
S_{0n}(\tilde{\boldsymbol{\gamma}},t) &= \frac{1}{n}\sum_{i=1}^{n} Y_i(t)\exp(g_n(\tilde{\boldsymbol{\gamma}},\boldsymbol{Z}_i,t)), \\
S_0(t) &= \mathrm{EY(t)}\exp(\mathrm{g}(\boldsymbol{\beta},\boldsymbol{Z},\mathrm{t})), \\
S_{1n}(\tilde{\boldsymbol{\gamma}},t) &= \frac{1}{n}\sum_{i=1}^{n} Y_i(t)\exp(g_n(\tilde{\boldsymbol{\gamma}},\boldsymbol{Z}_i,t))\boldsymbol{Z}_i \otimes \boldsymbol{B}_i, \\
S_1(t) &= \mathrm{EY(t)}\exp(\mathrm{g}(\boldsymbol{\beta},\boldsymbol{Z},\mathrm{t}))\boldsymbol{Z} \otimes \boldsymbol{B}, \\
S_{2n}(\tilde{\boldsymbol{\gamma}},t) &= \frac{1}{n}\sum_{i=1}^{n} Y_i(t)\exp(g_n(\tilde{\boldsymbol{\gamma}},\boldsymbol{Z}_i,t))(\boldsymbol{Z}_i\boldsymbol{Z}_i^{\top}) \otimes (\boldsymbol{B}_i\boldsymbol{B}_i^{\top}), \\
\text{and} \quad S_2(t) &= \mathrm{EY(t)}\exp(\mathrm{g}(\boldsymbol{\beta},\boldsymbol{Z},\mathrm{t}))(\boldsymbol{Z}\boldsymbol{Z}^{\top}) \otimes (\boldsymbol{B}\boldsymbol{B}^{\top}),
\end{aligned}
$$

followed by some key sufficient conditions that guarantee the properties of our estimator.

C1 The failure time $T^u$ and the censoring time $T^c$ are conditionally independent given the covariate $\boldsymbol{Z}$.

C2 $\tau$ is chosen so that $\Pr(T^u > \tau|\boldsymbol{Z}) > 0$ almost surely and $\tau < \infty$; at $\tau$, the baseline cumulative hazard function $\Lambda_0(\tau) \equiv \int_0^\tau \lambda_0(s)ds < \infty$.

C3 The covariates $\boldsymbol{Z}$ takes value in a bounded subset of $\mathbb{R}^p$ and $\Pr(Z_j = 0) < 1$. Also, $\sum_{j=1}^{p} |Z_j| = O_p(1)$.

C4 There exists a small positive constant $\varepsilon$ such that $\Pr(\Delta = 1|\boldsymbol{Z}) > \varepsilon$ and $\Pr(T^c > \tau|\boldsymbol{Z}) > \varepsilon$ almost surely.

C5 Let $0 < c_1 < c_2 < \infty$ be two constants. The joint density $f(t,\boldsymbol{z},\Delta = 1)$ of $(T,\boldsymbol{Z},\Delta = 1)$ satisfies $c_1 \le f(t,\boldsymbol{z},\Delta = 1) < c_2$ for all $(t,\boldsymbol{z}) \in [0,\tau] \times \mathbb{R}^p$.

C6 $\eta = o(q^{-m})$, $\rho = O(n^a)$ with $a \le -1$, and $q = o(n)$.

C7 There exist a neighborhood $\Theta$ of $\tilde{\boldsymbol{\gamma}}$ and scalar, vector and matrix functions $s_0$, $s_1$ and $s_2$ defined on $\boldsymbol{\gamma} \times [0,\tau]$ such that for $j = 0,1,2$,

$$\sup_{0\le t\le\tau,\boldsymbol{\gamma}\in\Theta} ||S_j(\boldsymbol{\gamma},t) - s_{(j)}(\boldsymbol{\gamma},t)|| \to_p 0.$$

C8 Let $\Theta$, $s_0$, $s_1$ and $s_2$ be as in Condition C7 and define $e = s_1/s_0$ and $v = s_2/s_0 - e^{\otimes 2}$. For all $\boldsymbol{\gamma} \in \Theta$, $t \in [0,\tau]$:

$$s_1(\boldsymbol{\gamma},t) = \frac{\partial}{\partial\boldsymbol{\gamma}} s_0(\boldsymbol{\gamma},t), \quad s_2(\boldsymbol{\gamma},t) = \frac{\partial^2}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^{\top}} s_0(\boldsymbol{\gamma},t),$$

$s_0(\cdot,t)$, $s_1(\cdot,t)$, $s_2(\cdot,t)$ are continuous functions of $\boldsymbol{\gamma} \in \Theta$, uniformly in $t \in [0,\tau]$, $s_0$, $s_1$, and $s_2$ are bounded on $\Theta \times [0,\tau]$, and the matrix

$$\Sigma(\tilde{\boldsymbol{\gamma}},\tau)=\int_0^{\tau} v(\tilde{\boldsymbol{\gamma}},t)s_0(\tilde{\boldsymbol{\gamma}},t)\tilde{\boldsymbol{\gamma}}(t)dt$$

is positive definite.

C9 There exists a $\delta>0$ such that

$$n^{-1/2}\sup_{i,t}||\boldsymbol{Z}_i||_\infty|Y_i(t)\mathscr{I}\{\boldsymbol{Z}_i^\top\boldsymbol{\beta}>-\delta||\boldsymbol{Z}_i||_\infty\}|\to_p 0.$$

Condition C1 is commonly assumed in survival analysis for non-informative censoring. The finite $\tau$ condition of C2 is assumed in many studies, including [1]. Condition C3 is often assumed in nonparametric regression and is reasonable in practical situations as we do not observe infinite covariates. Condition C4 controls the censoring rate so that the data have adequate information [21]. Condition C5 is needed for model identifiability and used in Huang (1999) [15]. Condition C6 controls estimation biases and ensures the convergence. Conditions C7, C8, and C9 are regularity conditions, which can be found in Anderson and Gill (1982) [1].

### *3.1 Asymptotic theory*

**Theorem 1.** *Suppose Conditions C1-C6 hold. If $\beta_{0j}(t)\in\mathbb{S}_{q,\alpha_j}$ for $j=1,\ldots,p$ with $q$ and $\alpha_j$ being the same as in* $\mathrm{PL}(\boldsymbol{\theta})$*, then*

$$||\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0||_2=O_p\left((q/n)^{1/2}\right);$$

*if $\beta_{0j}(t)\notin\mathbb{S}_{q,\alpha_j}$ for $j=1,\ldots,p$,*

$$||\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0||_2=O_p\left(r_n^{1/2}\right),$$

*where $r_n=q/n+q^{-2m}$.*

Theorem 1 implies convergence of $\hat{\boldsymbol{\beta}}$ by Condition C6 and $m>0{\cdot}5$. If the true curves are in the thresholding sieve space, there is no approximation error; and if $q$ is $O(1)$, Theorem 1 suggests root-$n$ consistency.

Let $\boldsymbol{e}_j$ be a directional vector of length $p$ with $j$th entry as 1 and others 0. For any $t\in[0,\tau]$, let $\boldsymbol{a}(t)=\boldsymbol{e}_j\otimes\boldsymbol{B}(t)$, then $\hat{\theta}_j(t)=\boldsymbol{a}(t)^\top\hat{\boldsymbol{\gamma}}$.

**Theorem 2.** *Under Conditions C1-C9, we have for any $t\in[0,\tau]$ and $j=1,\ldots,p$,*

$$\frac{\hat{\theta}_j(t)-\theta_j(t)}{\sigma_{nj}(t)}\to_d N(0,1),\quad \text{as } n\to\infty,$$

*where* $\sigma_{nj}^2(t)=n\boldsymbol{a}(t)^\top\left[\{\mathrm{PL}\}''(\tilde{\boldsymbol{\gamma}})\right]^{-1}\Sigma(\tilde{\boldsymbol{\gamma}},1)\left[\{\mathrm{PL}\}''(\tilde{\boldsymbol{\gamma}})\right]^{-1}\boldsymbol{a}(t)$.

With Theorem 2, we can then obtain the asymptotic distribution of $\hat{\beta}_j(t)$ based on $\hat{\beta}_j(t) = \zeta\{\hat{\theta}_j(t), \alpha_j\}$.

**Theorem 3.** *Under Conditions C1–C9, for any $t \in [0,\tau]$, the limiting distribution of $\hat{\beta}_j(t)$ ( $j = 1,\ldots,p$) satisfies*

$$\lim_{n\to\infty} \left| \Pr(\hat{\beta}_j(t) \le x) - G_{nj}(x) \right| \;=\; 0,$$

*where $G_{nj}(x) = \left[\Phi\left\{\frac{x+\alpha_j-\tilde{\theta}_j(t)}{\sigma_{nj}(t)}\right\}\mathscr{I}(x \ge 0) + \Phi\left\{\frac{x-\alpha_j-\tilde{\theta}_j(t)}{\sigma_{nj}(t)}\right\}\mathscr{I}(x < 0)\right]$ and $\Phi(\cdot)$ is the cumulative distribution function of $N(0,1)$.*

The limiting distribution in Theorem 3 guarantees that our proposed estimator can detect the zero-effect regions, because the probability of $\hat{\beta}_j(t) = 0$ can be greater than 0 even with a finite sample size.

## *3.2 Sparse confidence intervals*

We introduce the sparse confidence intervals to gauge the uncertainty of the point estimates and make valid statistical inferences on the selection and the zero-effect region detection.

Given a $\xi \in (0,1)$, for any $t \in [0,\tau]$ we construct a pointwise $(1-\xi)$ level asymptotic sparse confidence interval for $\beta_j(t)$, denoted by $[u_{nj}(t), v_{nj}(t)]$. Let $z_{\xi/2}$ and $\Phi$ be the $(1-\xi/2)$ quantile and the cumulative distribution function of $N(0,1)$, respectively. Let $P_+ = \Pr\{\hat{\beta}_j(t) > 0\}$ and $P_- = \Pr\{\hat{\beta}_j(t) < 0\}$, which can be estimated by $\hat{P}_+ = 1 - \Phi\{(\alpha_j - \hat{\theta}_j)/\hat{\sigma}_{nj}(t)\}$ and $\hat{P}_- = \Phi\{(-\alpha_j - \hat{\theta}_j)/\hat{\sigma}_{nj}(t)\}$ using Theorem 3. Here, $\hat{\sigma}_{nj}(t)$ is as defined in Theorem 2. We construct $[u_{nj}(t), v_{nj}(t)]$ as follows:

- if $\hat{P}_+ + \hat{P}_- \le \xi$, $u_{nj}(t) = v_{nj}(t) = 0$;
- else if $\hat{P}_+ < \xi/2$ and $\hat{P}_- < 1-\xi/2$, $[u_{nj}(t), v_{nj}(t)] = \left[\hat{\beta}_j(t) - \hat{\sigma}_{nj}(t)\hat{B}, 0\right]$ with $\hat{B} = \Phi^{-1}\left\{1 - \xi + \Phi(-\hat{\sigma}_{nj}^{-1}(t)\alpha_j + \hat{\sigma}_{nj}^{-1}(t)\hat{\theta}_j)\right\}$;
- else if $\hat{P}_- < \xi/2$ and $\hat{P}_+ < 1-\xi/2$, $[u_{nj}(t), v_{nj}(t)] = \left[0, \hat{\beta}_j(t) + \hat{\sigma}_{nj}(t)\hat{A}\right]$ with $\hat{A} = -\Phi^{-1}\left\{\xi - 1 + \Phi(\hat{\sigma}_{nj}^{-1}(t)\alpha_j + \hat{\sigma}_{nj}^{-1}(t)\hat{\theta}_j)\right\}$;
- else $[u_{nj}(t), v_{nj}(t)] = \left[\hat{\beta}_j(t) - \hat{\sigma}_{nj}(t)z_{\xi/2}, \hat{\beta}_j(t) + \hat{\sigma}_{nj}(t)z_{\xi/2}\right]$.

**Theorem 4.** *Under Conditions C1-C9, $[u_{nj}(t), v_{nj}(t)]$ is a $(1-\xi)$ level sparse confidence interval of $\beta_j(t)$ for $j = 1,\ldots,p$ and any $t \in [0,\tau]$.*

We omit its proof as it is a straightforward application of Theorem 3.

## 4 Simulations

We compare the proposed model with the regular time-varying effects Cox model. With $p = 3$, we design some special varying coefficient functions containing zero-effect regions as follows:

$$\begin{aligned} \beta_1(t) &= (-t^2+3)\mathscr{I}(t \le \sqrt{3}), \\ \beta_2(t) &= 2\log(t+0.01)\mathscr{I}(t \ge 1), \\ \text{and} \qquad \beta_3(t) &= \left(\frac{-6}{t+1}+2\right)\mathscr{I}(t \le 2). \end{aligned} \tag{6}$$

We first simulate $\boldsymbol{Z_i} = (Z_{i1}, Z_{i2}, Z_{i3})^\top \sim \mathrm{N}(\mathbf{0}, \Sigma)$, where $\Sigma$ has the following three structures: independent (Ind) with $\mathrm{cov}(Z_{ij}, Z_{ij^*}) = \mathscr{I}(j = j^*)$, autoregressive [AR(1)] with $\mathrm{cov}(Z_{ij}, Z_{ij^*}) = 0{\cdot}5^{|j-j^*|}$, and compound symmetry (CS) with $\mathrm{cov}(Z_{ij}, Z_{ij^*}) = \mathscr{I}(j = j^*) + 0{\cdot}5\mathscr{I}(j \ne j^*)$. We simulate $U_i \sim U(0,1)$, and solve $T_i^u$ using $U_i = 1 - \exp\left\{-\int_0^{T_i^u} \lambda_0(u)\exp(\sum_{j=1}^3 Z_j\beta_j(u))\mathrm{d}u\right\}$, where $\lambda_0(u)$ is set to be some constant in $(0,1)$. The censoring times $C_i$ are generated from $U(0,10)$, and $T_i^c = C_i \wedge 3$.

We choose sample sizes $n = 500$, $2{,}000$ and $5{,}000$, and generate 200 independent datasets for each setting. For implementing our proposed model, we set $d = 3$, $\eta = 0{\cdot}001$, and $\alpha_j$ to be half of the absolute values of the least-squares estimates. The number of knots, $K$, is selected via cross-validation. Specifically, we tune $K$ by conducting 10-fold cross-validation over a reasonable range such as $\{3, 5, 9, 13, 17, 21\}$. We note that $\rho$ can also be selected via cross-validation; however, that would increase the computational burden, given that $K$ needs to to be tuned. Based on our numerical experience, we have found that specifying $\rho = 1/n^2$ that satisfies Condition C6 would give a good performance. Therefore, we set $\rho = 1/n^2$ in simulations and our later data analysis.

For comparison, we fit time-varying effects Cox models by following [31, 33]. For evaluation criteria, we use the integrated squared errors (ISE) and the averaged integrated squared errors (AISE), defined as $\mathrm{ISE}(\beta_j) = n_g^{-1}\sum_{g=1}^{n_g}\{\hat{\beta}_j(t_g) - \beta_j(t_g)\}^2$ and $\mathrm{AISE} = p^{-1}\sum_{j=1}^p \mathrm{ISE}(\beta_j)$, respectively, where $t_g$ $(g = 1, \ldots, n_g)$ are the grid points on $(0,3)$. Table 1 shows that the soft-thresholding time-varying effects Cox model has a better accuracy than the regular time-varying effects Cox model by presenting smaller integrated squared errors and averaged integrated squared errors.

Figure 1, which plots the estimation curves and their median for the soft-thresholding time-varying effects Cox model and the regular time-varying effects Cox model, shows the medium estimation curves from the soft-thresholding time-varying effects Cox model coincide with the truth and the soft-thresholding approach has the zero-effect detection ability. In contrast, the regular time-varying effects Cox model fails to estimate zero effects.

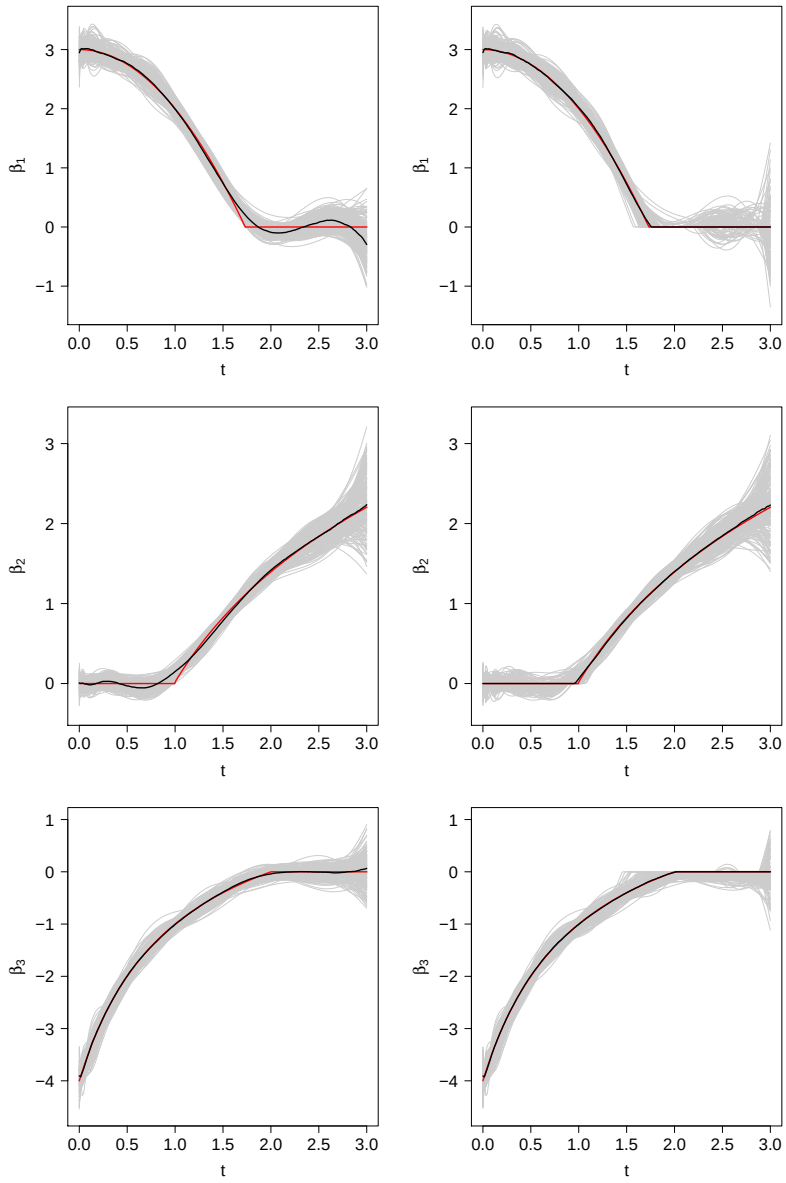


Fig. 1: Comparisons of the results obtained from the soft-thresholding time-varying effects Cox model (right panel) and the regular time-varying effects Cox model (left panel); the gray curves are 200 estimated curves based on 200 simulations, the black curves are the median estimates, and the red curves are the truth; the sample size is $5,000$ and the average censoring rate is 0.12.

Table 1: Comparisons of estimation accuracy for the soft-thresholding time-varying effects Cox model and the regular time-varying effects Cox model.

| Covariance | n | Model | ISE($\beta_1$) | ISE($\beta_2$) | ISE($\beta_3$) | AISE |
|---|---|---|---|---|---|---|
| Ind | 500 | STTV | 62.6 (77.1) | 53.1 (43.5) | 58.7 (59.5) | 58.1 (39.7) |
| | | RegTV | 75.5 (94.6) | 56.6 (44.3) | 61.9 (60.6) | 65.4 (46.2) |
| | 2000 | STTV | 12.4 ( 9.7) | 12.0 ( 8.5) | 13.1 (10.4) | 12.5 ( 5.7) |
| | | RegTV | 13.9 ( 8.2) | 11.8 ( 8.6) | 12.4 ( 8.8) | 12.7 ( 5.1) |
| | 5000 | STTV | 4.2 (3.2) | 4.1 (2.8) | 4.0 (2.7) | 4.1 (1.7) |
| | | RegTV | 5.6 (3.0) | 4.2 (2.8) | 4.5 (2.7) | 4.7 (1.6) |
| AR(1) | 500 | STTV | 16.2 (16.0) | 18.2 (47.1) | 15.2 (11.1) | 16.5 (18.7) |
| | | RegTV | 16.3 (14.1) | 20.9 (50.3) | 13.4 ( 8.2) | 16.9 (19.4) |
| | 2000 | STTV | 3.6 (2.2) | 2.6 (2.0) | 3.9 (2.3) | 3.3 (1.5) |
| | | RegTV | 3.7 (2.2) | 2.8 (2.5) | 3.1 (1.6) | 3.2 (1.4) |
| | 5000 | STTV | 1.3 (1.0) | 1.1 (0.9) | 1.2 (0.8) | 1.2 (0.6) |
| | | RegTV | 1.9 (0.9) | 1.3 (0.9) | 1.3 (0.8) | 1.5 (0.6) |
| CS | 500 | STTV | 18.9 (24.6) | 19.1 (30.2) | 16.5 (14.6) | 18.2 (16.2) |
| | | RegTV | 19.1 (15.5) | 20.4 (30.3) | 17.0 (12.2) | 18.8 (13.2) |
| | 2000 | STTV | 3.6 (2.6) | 2.7 (2.5) | 3.8 (2.7) | 3.4 (1.8) |
| | | RegTV | 4.0 (2.3) | 2.8 (2.4) | 3.2 (1.6) | 3.4 (1.4) |
| | 5000 | STTV | 1.2 (0.8) | 1.1 (0.9) | 1.0 (0.6) | 1.1 (0.5) |
| | | RegTV | 1.8 (0.7) | 1.1 (0.9) | 1.2 (0.7) | 1.4 (0.5) |

STTV: the soft-thresholding time-varying effects Cox model; RegTV: the regular time-varying effects Cox model; ISE: the integrated squared errors; AISE: the averaged integrated squared errors. All numbers are after being multiplied by 100.

Figure 2 compares the estimated coverage probabilities from the soft-thresholding time-varying effects Cox model and the regular time-varying effects Cox model, and shows that the soft-thresholding time-varying Cox model has a reasonable coverage probability in both zero-effect regions and non-zero-effect regions. In the region around the transition point, the soft thresholding time-varying effects Cox model has a higher coverage probability estimation than the regular time-varying effects Cox model. All of the results confirm that the soft-thresholding time-varying effects Cox model draws better inference than the regular time-varying effects Cox model.

With $|A|$ being the cardinality of set $A$, we next compare zero-effect region detection using the following criteria:

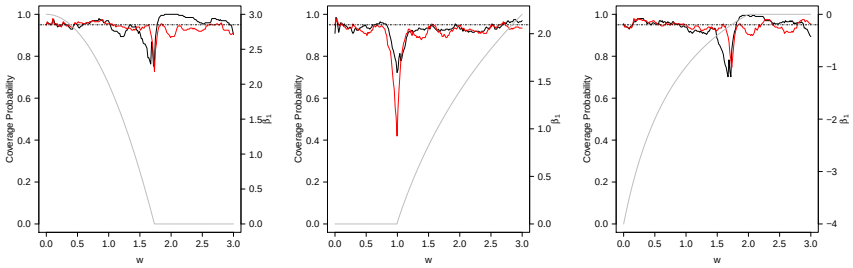


Fig. 2: Comparisons of coverage probability from the regular time-varying effects Cox model (RegTV) (the black curve) and the soft-thresholding time-varying effects Cox model (STTV) (the red curve). The sample size is $5,000$ and the average censoring rate is 0.12.

$$\text{Estimation-based true positive ratio: ETPR}(\beta) = \frac{|\{t : \hat{\beta}(t) \neq 0 \text{ and } \beta(t) \neq 0\}|}{|\{t : \beta(t) \neq 0\}|},$$

$$\text{Estimation-based true negative ratio: ETNR}(\beta) = \frac{|\{t : \hat{\beta}(t) = 0 \text{ and } \beta(t) = 0\}|}{|\{t : \beta(t) = 0\}|},$$

$$\text{Inference-based true positive ratio: ITPR}(\beta) = \frac{|\{t : 0 \notin \text{CI}\{\hat{\beta}(t)\} \text{ and } \beta(t) \neq 0\}|}{|\{t : \beta(t) \neq 0\}|},$$

and

$$\text{Inference-based true negative ratio: ITNR}(\beta) = \frac{|\{t : 0 \in \text{CI}\{\hat{\beta}(t)\} \text{ and } \beta(t) = 0\}|}{|\{t : \beta(t) = 0\}|},$$

where $\text{CI}\{\hat{\beta}(t)\}$ is the 95% confidence interval of $\beta(t)$.

We set a total of 100 grid points on $[0,3]$, counting the number of $t_g$ in each set as its cardinality. Table 2 shows that the soft-thresholding time-varying effects Cox model has a higher inference-based true negative ratio than the regular time-varying effects Cox model. Although the inference-based true positive and negative ratios are more reliable with controlled false discovery rates, their computational burden increases when the sample size increases. Therefore, the estimation-based true positive and negative ratios are favorable for large datasets as their calculation merely depends on the estimations. First, our method presents a better estimation-based true negative ratio ratio, indicating our method can detect zero-effect regions well. Second, as documented in Table 2, our method also presents a higher estimation-based true positive ratio than the inference-based true positive ratio, indicating a better performance of our method in inferring non-zero effects.

Table 2: Comparisons of true positive ratios and true negative ratios for zero-effect region detection

| | | | STTV | | | | RegTV | |
|---|---|---|---|---|---|---|---|---|
| | n | $\beta$ | ETPR | ETNR | ITPR | ITNR | ITPR | ITNR |
| Ind | 500 | $\beta_1$ | 0.96 (0.08) | 0.44 (0.25) | 0.81 (0.10) | 0.94 (0.11) | 0.81 (0.09) | 0.95 (0.11) |
| | | $\beta_2$ | 0.96 (0.05) | 0.22 (0.13) | 0.57 (0.17) | 0.94 (0.08) | 0.57 (0.18) | 0.94 (0.12) |
| | | $\beta_3$ | 0.95 (0.12) | 0.37 (0.28) | 0.55 (0.12) | 0.94 (0.12) | 0.55 (0.12) | 0.94 (0.12) |
| | 2000 | $\beta_1$ | 0.95 (0.06) | 0.61 (0.23) | 0.89 (0.07) | 0.95 (0.10) | 0.91 (0.06) | 0.94 (0.10) |
| | | $\beta_2$ | 0.97 (0.04) | 0.34 (0.16) | 0.85 (0.08) | 0.94 (0.08) | 0.86 (0.08) | 0.95 (0.08) |
| | | $\beta_3$ | 0.96 (0.12) | 0.50 (0.24) | 0.70 (0.10) | 0.94 (0.11) | 0.72 (0.10) | 0.95 (0.13) |
| | 5000 | $\beta_1$ | 0.98 (0.03) | 0.64 (0.27) | 0.95 (0.04) | 0.94 (0.10) | 0.96 (0.04) | 0.93 (0.11) |
| | | $\beta_2$ | 0.98 (0.03) | 0.46 (0.18) | 0.92 (0.05) | 0.94 (0.09) | 0.93 (0.04) | 0.94 (0.10) |
| | | $\beta_3$ | 0.97 (0.09) | 0.50 (0.31) | 0.81 (0.09) | 0.96 (0.10) | 0.80 (0.08) | 0.96 (0.10) |
| AR(1) | 500 | $\beta_1$ | 0.96 (0.05) | 0.60 (0.22) | 0.90 (0.07) | 0.95 (0.10) | 0.93 (0.06) | 0.93 (0.12) |
| | | $\beta_2$ | 0.98 (0.04) | 0.32 (0.18) | 0.85 (0.08) | 0.92 (0.13) | 0.86 (0.08) | 0.93 (0.12) |
| | | $\beta_3$ | 0.97 (0.14) | 0.51 (0.27) | 0.69 (0.14) | 0.95 (0.13) | 0.73 (0.12) | 0.95 (0.12) |
| | 2000 | $\beta_1$ | 0.97 (0.04) | 0.71 (0.19) | 0.94 (0.04) | 0.95 (0.08) | 0.99 (0.02) | 0.92 (0.11) |
| | | $\beta_2$ | 0.99 (0.02) | 0.49 (0.19) | 0.95 (0.04) | 0.94 (0.10) | 0.97 (0.03) | 0.93 (0.11) |
| | | $\beta_3$ | 0.96 (0.09) | 0.62 (0.25) | 0.77 (0.10) | 0.92 (0.13) | 0.86 (0.07) | 0.94 (0.13) |
| | 5000 | $\beta_1$ | 1.00 (0.01) | 0.79 (0.17) | 0.98 (0.02) | 0.96 (0.08) | 1.00 (0.00) | 0.85 (0.11) |
| | | $\beta_2$ | 1.00 (0.01) | 0.56 (0.17) | 0.98 (0.02) | 0.94 (0.09) | 1.00 (0.01) | 0.87 (0.11) |
| | | $\beta_3$ | 0.97 (0.05) | 0.63 (0.30) | 0.90 (0.05) | 0.97 (0.09) | 0.91 (0.05) | 0.96 (0.10) |
| CS | 500 | $\beta_1$ | 0.96 (0.06) | 0.58 (0.23) | 0.90 (0.07) | 0.96 (0.10) | 0.92 (0.07) | 0.94 (0.12) |
| | | $\beta_2$ | 0.98 (0.03) | 0.32 (0.19) | 0.85 (0.07) | 0.93 (0.12) | 0.86 (0.07) | 0.94 (0.13) |
| | | $\beta_3$ | 0.98 (0.13) | 0.51 (0.29) | 0.70 (0.13) | 0.96 (0.11) | 0.71 (0.12) | 0.95 (0.11) |
| | 2000 | $\beta_1$ | 0.97 (0.04) | 0.68 (0.21) | 0.94 (0.04) | 0.96 (0.08) | 0.98 (0.02) | 0.92 (0.12) |
| | | $\beta_2$ | 0.99 (0.02) | 0.48 (0.18) | 0.96 (0.04) | 0.94 (0.09) | 0.97 (0.03) | 0.94 (0.09) |
| | | $\beta_3$ | 0.97 (0.11) | 0.65 (0.25) | 0.78 (0.11) | 0.95 (0.09) | 0.86 (0.07) | 0.94 (0.13) |
| | 5000 | $\beta_1$ | 0.99 (0.01) | 0.73 (0.18) | 0.98 (0.02) | 0.96 (0.08) | 1.00 (0.01) | 0.87 (0.11) |
| | | $\beta_2$ | 1.00 (0.01) | 0.55 (0.16) | 0.98 (0.02) | 0.92 (0.10) | 1.00 (0.01) | 0.89 (0.10) |
| | | $\beta_3$ | 0.96 (0.06) | 0.66 (0.30) | 0.89 (0.06) | 0.97 (0.08) | 0.90 (0.05) | 0.96 (0.11) |

STTV: the soft-thresholding time-varying effects Cox model; RegTV: the regular time-varying effects Cox model.

## 5 Analysis of the Boston Lung Cancer Survivor Cohort

We apply our method to study a subset of the Boston Lung Cancer Survivor Cohort (BLCSC) [6]. The data consist of $n = 599$ individuals, among whom 148 (24.7%) were alive and 451 (75.3%) were dead by the end of the follow up. The primary endpoint was overall survival measuring the time lag from the diagnosis of lung cancer to death or the end of the study, which ever came first. The range of the observed survival time was from 6 days to 8584 days, and the restricted mean survival and censoring times at $\tau = 8584$ days were 2124 (SE: 105) and 4397 (SE: 187) days, respectively. The observed survival time was skewed to the right. Patients who were

alive were younger than those of those who died (average age in years: 55.4 vs. 61.2), and were slightly less likely to be Caucasian (89.9% vs. 95.8%). With early-stage lung cancer including stages lower than II, e.g., 1A, 1B, IIA, and IIB, 64.2% of the alive patients had early-stage lung cancer, slightly higher than those who died (62.3%). The percentage of the alive patients who had surgery was 83.8%, higher than that of the dead patients (63.0%). See Table 3 for more details.

Table 3: Summary of the patient characteristics

| | Alive ($n = 148$) | Dead ($n = 451$) |
|---|---|---|
| Age | 55.4 (10.1) | 61.2 (10.8) |
| Pack years | 34.4 (29.7) | 51.6 (38.5) |
| Race | | |
| White (ref) | 133 (89.9%) | 432 (95.8%) |
| Others | 15 (10.1%) | 19 (4.2%) |
| Education | | |
| Under high school (ref) | 10 (6.8%) | 72 (16%) |
| High school graduate | 30 (20.3%) | 113 (25.1%) |
| Above high school | 108 (73.0%) | 266 (59.0%) |
| Sex | | |
| Female (ref) | 113 (76.4%) | 256 (56.8%) |
| Male | 35 (23.6%) | 195 (43.2%) |
| Smoking status | | |
| Ever or never (ref) | 96 (64.9%) | 281 (62.3%) |
| Current | 52 (35.1%) | 170 (37.7%) |
| Cancer stage | | |
| Early (ref) | 95 (64.2%) | 190 (42.1%) |
| Late | 53 (35.8%) | 261 (57.9%) |
| Surgery | 124 (83.8%) | 284 (63.0%) |
| Chemotherapy | 48 (32.4%) | 206 (45.7%) |
| Radiotherapy | 35 (23.6%) | 184 (40.8%) |

Continuous variables are presented in mean (standard deviation), and categorical variables are presented in count (percentage). Due to rounding, some summations of percentages for one variable are not one. Reference groups are marked.

Included in our analysis are age, race, education, sex, smoking status, cancer stage, and treatments received (surgery, chemotherapy, and radiotherapy). For comparisons, we fit the data by using the Cox proportional hazards model, the regular time-varying effects Cox model (RegTV) and the soft-thresholding time-varying effects Cox model (STTV). When implementing STTV, we set the needed parameters such as $\rho, \alpha_j$ and $K$ as in done in the simulation section. In particular, with respect to the choice of $K$ and $\rho$, we have determined that $K = 5$ by minimizing a 10-fold

cross-validation error over a candidate set of $\{3, 5, 9, 13, 17, 21\}$, while setting the penalty parameter $\rho$ to be $1/n^2$. We fit RegTV by using the penalized B-spline approach of [31, 33]. See the results as summarized in Figures 3, 4 and 5.

Compared with the regular time-varying effects Cox model, the soft-thresholding time-varying effects Cox model agrees more to the Cox proportional hazards model. For some non-significant coefficients in the constant effect Cox model, STTV estimates those to be all zero over the time, such as for chemotherapy and radiotherapy. The results seem to be reasonable: insofar as surgery had a strong protective effect for this group of lung cancer patients, adding chemotherapy or radiotherapy did not seem to be associated with additional protective or harmful impacts on lung cancer patients' survival. This is consistent with the clinical practice that surgery is often the first line therapy for operable lung cancer patients [32]. Interestingly, smoking at diagnosis was associated with higher short-term (in the first 3 years post-diagnosis) and long-term (after 7 years post-diagnosis) mortality, but was not significantly associated with mortality between 3 and 7 years post-diagnosis, possibly a stabilization period for patients. The result highlights the importance of early cessation of smoking [2].

The other results are equally interesting. Adjusting for all the other factors, the expected hazard was significantly higher among male patients than female patients; older patients had a significantly higher hazard than younger patients; non-white patients had a lower hazard than white patients; a later cancer stage was strongly associated with worse lung cancer mortality. However, there were no significant associations between education levels and lung cancer mortality. In conclusion, the results of STTV are consistent with those obtained by using the Cox model, but STTV can more accurately capture the time-varying effect of each factor.

## 6 Discussion

To address the challenge of modeling time-varying coefficients with zero-effect regions in survival analysis, we proposed a new soft-thresholding time-varying coefficient model, where the varying coefficients are piecewise smooth with zero-effect regions. To quantify uncertainty of the estimates, we have designed a new type of sparse confidence intervals, which extend classical confidence intervals by accommodating exact zero estimates. Our framework enables us to estimate non-zero time-varying effects and detect zero-effect regions simultaneously, extending the already-widely-used Cox models to a new territory. The work pays tribute to Sir D.R. Cox, whose work has fundamentally influenced modern biomedical research.

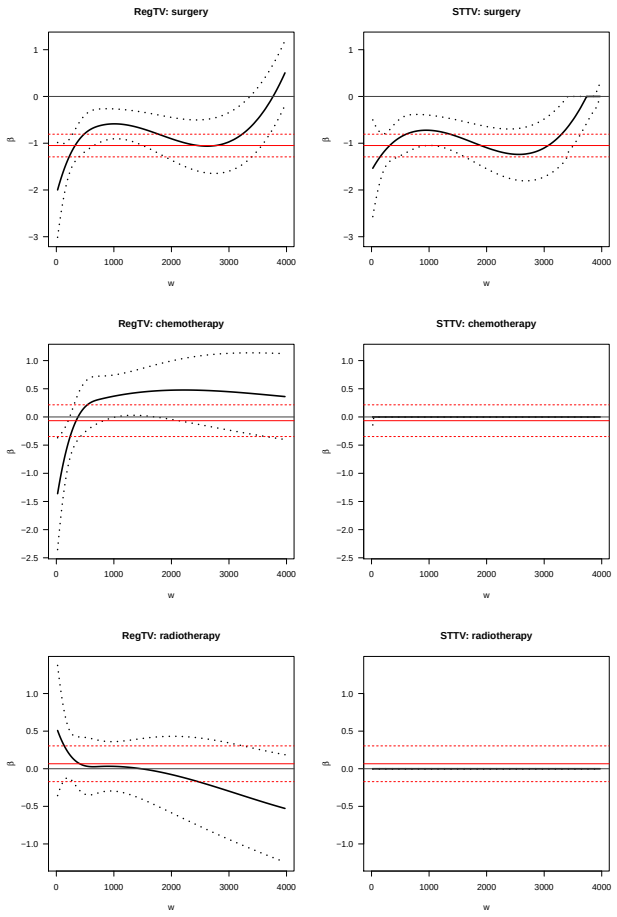


Fig. 3: Estimation results (part I) for the BLCSC data using the regular time-varying effects Cox model (RegTV) and the soft-thresholding time-varying effects Cox model (STTV): the solid lines are the estimated coefficient function curves; the dotted lines are the pointwise (sparse) confidence intervals; black lines are from varying coefficient models; red lines are from the constant effect Cox model.

## Acknowledgements

We thank Dr. David Christiani for providing the BLCSC data. We thank two referees for their insightful comments that have improved the presentation and the quality of the submission. The work is partially supported by grants from NIH (R01CA249096 and U01CA209414).

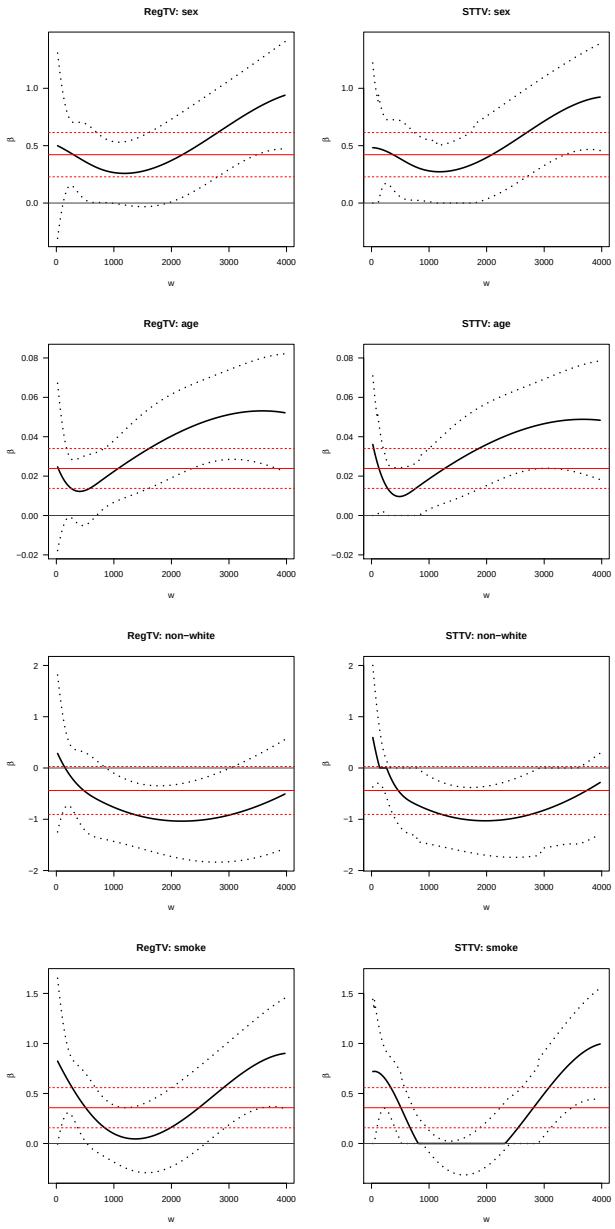


Fig. 4: Estimation results (part II) for the BLCSC data using the regular time-varying effects Cox model (RegTV) and the soft-thresholding time-varying effects Cox model (STTV): the solid lines are the estimated coefficient function curves; the dotted lines are the pointwise (sparse) confidence intervals; the black lines are from varying coefficient models and the red lines are from the Cox proportional hazards model.

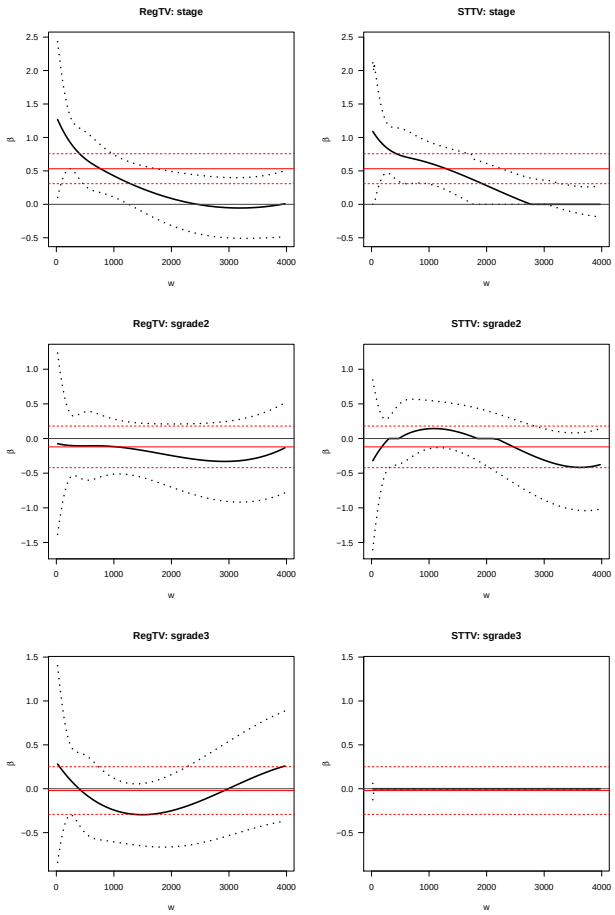


Fig. 5: Estimation results (part III) for the BLCSC data using the regular time-varying effects Cox model (RegTV) and the soft-thresholding time-varying effects Cox model (STTV): the solid lines are the estimated coefficient function curves; the dotted lines are the pointwise (sparse) confidence intervals; the black lines are from varying coefficient models and the red lines are from the Cox proportional hazards model.

# Appendix

**Proof of Theorem 1**:

For every $f \in \mathbb{F}_0$, by Corollary 6.21 of Schumaker (2007) [22] , there exists an $f_n \in \mathbb{F}$, $||f_n - f||_\infty = O(q^{-m})$. For any $\delta_1 > 0$ and $\delta_2 > 0$, there exists an $\eta$ (when constructing $h$) such that $||h(f_n) - \zeta(f_n)|| < \delta_1$ and $||h(f) - \zeta(f)|| < \delta_2$. Then we

have,

$$\begin{aligned}||h(f_n)-h(f)|| &< ||h(f_n)-\zeta(f_n)|| + ||\zeta(f_n)-\zeta(f)|| + ||\zeta(f)-h(f)|| \\ &= A_1 + A_2 + A_3.\end{aligned}$$

Let $\delta_1 = O(q^{-m})$ and $\delta_2 = O(q^{-m})$, then $A_1 < \delta_1 = O(q^{-m})$ and $A_2 < \delta_2 = O(q^{-m})$. We have $A_3 \leq ||f_n - f|| = O(q^{-m})$ because the Lipschitz continuous property in Lemma 1 of Kang et al. (2018) [16]. Therefore, $||h(f_n)-h(f)||_\infty = O(q^{-m})$. For simplicity of notation, let $h_{nj}$ denote $h(f_{nj})$ and $h_{0j}$ denote $h(f_{0j})$.

Let $g_n = \sum_{j=1}^p Z_j h_{nj}$. Then, given $\boldsymbol{Z}$, $|g_n - g_0| = |\sum_{j=1}^p Z_j(h_{nj}-h_{0j})| \leq \sum_{j=1}^p |Z_j||(h_{nj}-h_{0j})| = O_p(q^{-m})$. Thus, we have $||g_n - g_0|| = O_p(q^{-m})$.

By Lemma 5.1 of Huang (1999) [15], $||\hat{g}_n - g_n||_2^2 = o_p(1)$. We then only need to prove

$$\mathrm{E}\sup_{\delta/2<||g-g_n||\leq\delta} |\mathrm{M_n(g)-M_n(g_n)-(M_0(g)-M_0(g_n))}| = \mathrm{O_p(n^{-\frac{1}{2}}\delta(q^{\frac{1}{2}}+\log^{\frac{1}{2}}(1/\delta)))}. \tag{7}$$

It follows that

$$\begin{aligned}&M_n(g)-M_n(g_n)-\{M_0(g)-M_0(g_n)\}\\ =&P_{\Delta_n}m_n(\cdot,g)-P_{\Delta_n}m_n(\cdot,g_n)-P_\Delta m_0(\cdot,g)+P_\Delta m_0(\cdot,g_n)\\ =&P_{\Delta_n}m_n(\cdot,g)-P_\Delta m_0(\cdot,g)-P_{\Delta_n}m_n(\cdot,g_n)+P_\Delta m_0(\cdot,g_n)\\ =&P_{\Delta_n}\{\log S_{0n}(\cdot,g)-\log S_0(\cdot,g)\}-P_{\Delta_n}\{\log S_0(\cdot,g_n)-\log S_0(\cdot,g_n)\}\\ =&J_{1n}+J_{2n}.\end{aligned}$$

For any $\beta \in \mathbb{H}_n$ and any $\alpha > 0$, we can find at least one $f \in \mathbb{F}_n$ such that $\beta = \zeta(f,\alpha)$, then $\log N_{[]}(\varepsilon,\mathbb{H}_n,\delta) \leq \log N_{[]}(\varepsilon,\mathbb{F}_n,\delta) \lesssim c_1 q\log(\delta/\varepsilon)$ by calculation in [23]. Therefore, we can also obtain $\log N_{[]}(\varepsilon,\mathbb{H}_n,\delta) \lesssim c_2 q\log(\delta/\varepsilon)$ according to its construction. Because both exp and log are monotone functions, we have $\log N_{[]}(\varepsilon,E_{n,\delta},\delta) \lesssim c_2 q\log(\delta/\varepsilon) + c_3 q\log(\delta/\varepsilon) \lesssim c_4 q\log(\delta/\varepsilon)$, where $c_4 = \max(c_2,c_3)$.

Therefore, $J_{[]}(\delta,\varepsilon_{n,\delta},\rho) = \int_0^\delta \sqrt{1+\log N_{[]}(\varepsilon,E_{n,\delta},\rho)}\mathrm{d}\varepsilon \lesssim \delta q^{\frac{1}{2}}$. By Lemma 3.4.2 of Van Der Vaart and Wellner (1996) [27], we have

$$\mathrm{E}||\mathrm{J_{1n}}|| \lesssim \mathrm{n^{-\frac{1}{2}}q^{\frac{1}{2}}\delta(1+\frac{q^{\frac{1}{2}}\delta}{\delta^2\sqrt{n}}c_5) = O(n^{-\frac{1}{2}}q^{\frac{1}{2}}\delta)}. \tag{8}$$

On the other hand, we have

$$\begin{aligned}\sup_{||g-g_n||\le\delta}|J_{2n}| &\le 2\sup_{0\le t\le\tau,||g-g_n||\le\delta}\left|log\frac{S_{0n}(\cdot,g)}{S_{0n}(\cdot,g_n)}-log\frac{S_0(\cdot,g)}{S_0(\cdot,g_n)}\right|\\ &\lesssim \sup_{0\le t\le\tau,||g-g_n||\le\delta}\left|\frac{S_{0n}(\cdot,g)}{S_{0n}(\cdot,g_n)}-\frac{S_0(\cdot,g)}{S_0(\cdot,g_n)}\right|\\ &\lesssim \sup_{0\le t\le\tau,||g-g_n||\le\delta}\left|\frac{S_{0n}(\cdot,g)S_0(\cdot,g_n)-S_{0n}(\cdot,g_n)S_0(\cdot,g)}{S_{0n}(\cdot,g_n)S_0(\cdot,g_n)}\right|.\end{aligned}$$

Since the denominator is bounded away from 0 with probability approaching to 1, we only need to consider the numerator. It follows that

$$\begin{aligned}&S_{0n}(\cdot,g)S_0(\cdot,g_n)-S_{0n}(\cdot,g_n)S_0(\cdot,g)\\ =&S_0(t,g_n)\{S_{0n}(t,g)-S_{0n}(t,g_n)-S_0(t,g)+S_0(t,g_n)\}-\\ &\{S_{0n}(t,g_n)-S_0(t,g_n)\}\{S_0(t,g)-S_{0n}(t,g)\}\\ =&I_{1n}-I_{2n}.\end{aligned}$$

Since $I_{1n}=S_0(t,g_n)Y(t)[\exp(g(z))-\exp(g_n(z))]$, we consider the class of function $Y(t)\exp(g(z))$. Since exp is monotone and the entropy of the class of indicator function $Y(t)=I[0\le t\le\tau]$ is $\delta\log^{\frac{1}{2}}(1/\delta)$, we have that the entropy of the class of function $Y(t)exp(g(z))$ is $\delta(q^{\frac{1}{2}}+log^{\frac{1}{2}}(1/\delta))$. By Lemma 3.4.2 of Van Der Vaart and Wellner (1996) [27], $I_{1n}\lesssim n^{-\frac{1}{2}}\delta(q^{\frac{1}{2}}+\log^{\frac{1}{2}}(1/\delta))$.

By Taylor's expansion and Jensen's inequality, we have

$$\begin{aligned}|S_0(t,g)-S_0(t,g_n)| &\le E(Y(t)[\exp(g)-\exp(g_n)])\\ &\le \mathrm{E}(\exp(\mathrm{g_n})|\mathrm{g}-\mathrm{g_n}|)\\ &\lesssim (\mathrm{E}(\mathrm{g}-\mathrm{g_n})^2)^{\frac{1}{2}}=\mathrm{O_p}(\delta).\end{aligned}$$

Since $S_n(t,g_n)-S_0(t,g_n)=O_p(n^{-\frac{1}{2}}q^{\frac{1}{2}})$, we obtain $I_{2n}=O_p(n^{-\frac{1}{2}}q^{\frac{1}{2}}\delta)$.

Therefore, $\sup_{||g-g_n||\le\delta}|J_{2n}|\lesssim n^{-\frac{1}{2}}\delta(q^{\frac{1}{2}}+\log^{\frac{1}{2}}(1/\delta))$. Thus, we have $M_n(g)-M_n(g_n)-\{M_0(g)-M_0(g_n)\}=O_p(n^{-\frac{1}{2}}\delta(q^{\frac{1}{2}}+\log^{\frac{1}{2}}(1/\delta)))$.

By Theorem 3.4.1 of Van Der Vaart and Wellner (1996) [27], the key function $\phi(\delta)$ takes the form of $\phi_n(\delta)=\delta(q^{\frac{1}{2}}+log^{\frac{1}{2}}(1/\delta))$. Therefore, $||(\hat{g}_n-g_n)||_2=O_p((q/n)^{\frac{1}{2}})$. Therefore, we have

$$\begin{aligned}||\hat{g}_n-g_0||_2^2 &\le ||\hat{g}_n-g_n||_2^2+||g_n-g_0||_2^2\\ &\le O_p(q/n)+O_P(q^{-2m})\\ &\le O_p(r_n),\end{aligned}\tag{9}$$

where $r_n=q/n+q^{-2m}$.

Then by Lemma 1 of Stone (1985) [24], we have

$$E(Z_j\hat{h}_j(t)-Z_jh_j(t))^2=O_p(r_n),\quad 1\le j\le p.\tag{10}$$

By Condition C3, there exists $\delta, \varepsilon > 0$, $\Pr(|Z_j| > \delta) > \varepsilon$. Then

$$\begin{aligned} E(Z_j\hat{h}_j(t) - Z_j h_j(t))^2 &> \Pr(|Z_j| > \delta)\delta^2(\hat{h}_j(t) - h_j(t))^2 \\ &> \varepsilon\delta^2(\hat{h}_j(t) - h_j(t))^2. \end{aligned} \tag{11}$$

Therefore for any $t$, we have $(\hat{\beta}_j(t) - \beta_j(t))^2 = O_p(r_n)$, i.e. $|\hat{\beta}_j(t) - \beta_j(t)| = O_p(r_n^{1/2})$. Then we have $||\hat{\beta}_j - \beta_j||_\infty = O_p(r_n^{1/2})$ for $j = 1, \ldots, p$. ■

**Proof of Theorem 2**:

We show Theorem 2 is true when $\tau = 1$. The extension to any $\tau < \infty$ satisfying condition C2 is straightforward and is omitted.

Following the counting process notation in Anderson and Gill (1982) [1], we let

$$C(\boldsymbol{\gamma}, t) = \sum_{i=1}^{n} \int_0^\top \sum_{j=1}^{p} Z_{ij} h_j(\boldsymbol{\gamma}_j, s) \mathrm{d}N_i(s) - \int_0^\top \log\left\{ \sum_{i=1}^{n} Y_i(s) \exp\{ \sum_{j=1}^{p} Z_{ij}(s) h_j(\boldsymbol{\gamma}_j, s) \} \right\} \mathrm{d}\bar{N}(s),$$

then we have,

$$\mathrm{PL}(\boldsymbol{\gamma}) = C(\boldsymbol{\gamma}, 1) - \rho ||\boldsymbol{\theta}||_2^2.$$

Then for any $\boldsymbol{\gamma}$,

$$\mathrm{PL}^{'}(\boldsymbol{\gamma}) = C^{'}(\boldsymbol{\gamma}, 1) - \rho \sum_{i=1}^{n} \boldsymbol{\theta} \otimes \boldsymbol{B}(T_i).$$

By Taylor's expansion, we have that

$$\{\mathrm{PL}\}^{'}(\hat{\boldsymbol{\gamma}}) - \mathrm{PL}^{'}(\tilde{\boldsymbol{\gamma}}) = \{\mathrm{PL}\}^{''}(\boldsymbol{\gamma}^*)(\boldsymbol{\gamma} - \tilde{\boldsymbol{\gamma}}),$$

where $\boldsymbol{\gamma}^*$ is on the line segment between $\hat{\boldsymbol{\gamma}}$ and $\tilde{\boldsymbol{\gamma}}$. Since $\{\mathrm{PL}\}^{'}(\hat{\boldsymbol{\gamma}}) = 0$, we have

$$\begin{aligned} \boldsymbol{\gamma} - \tilde{\boldsymbol{\gamma}} &= -\left[\{\mathrm{PL}\}^{''}(\boldsymbol{\gamma}^*)\right]^{-1} \mathrm{PL}^{'}(\tilde{\boldsymbol{\gamma}}) \\ &= -\left[\{\mathrm{PL}\}^{''}(\boldsymbol{\gamma}^*)\right]^{-1} \left\{ C^{'}(\tilde{\boldsymbol{\gamma}}, 1) - \rho \sum_{i=1}^{n} \boldsymbol{\theta}_0 \otimes \boldsymbol{B}(T_i) \right\} \\ &= -\left[\{\mathrm{PL}\}^{''}(\boldsymbol{\gamma}^*)\right]^{-1} C^{'}(\tilde{\boldsymbol{\gamma}}, 1) + \rho \left[\{\mathrm{PL}\}^{''}(\boldsymbol{\gamma}^*)\right]^{-1} \sum_{i=1}^{n} \boldsymbol{\theta} \otimes \boldsymbol{B}(T_i). \end{aligned}$$

The goal is to prove that for any non-zero $\boldsymbol{a}$,

$$\frac{\boldsymbol{a}^\top(\hat{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}})}{\hat{\sigma}(\boldsymbol{a})} \to_d N(0, 1),$$

where $\hat{\sigma}(\boldsymbol{a}) = n\boldsymbol{a}^\top \left[\{\mathrm{PL}\}^{''}(\tilde{\boldsymbol{\gamma}})\right]^{-1} \Sigma(\tilde{\boldsymbol{\gamma}}, 1) \left[\{\mathrm{PL}\}^{''}(\tilde{\boldsymbol{\gamma}})\right]^{-1} \boldsymbol{a}$.

We claim that

$$\frac{\boldsymbol{a}^\top \left[-\{\mathrm{PL}\}''(\boldsymbol{\gamma}^*)\right]^{-1} C'(\tilde{\boldsymbol{\gamma}},1)}{\hat{\sigma}(\boldsymbol{a})} \to_d N(0,1) \tag{12}$$

and

$$\rho \boldsymbol{a}^\top \left[\{\mathrm{PL}\}''(\boldsymbol{\gamma}^*)\right]^{-1} \sum_{i=1}^{n} \boldsymbol{\theta} \otimes \boldsymbol{B}(T_i)/\hat{\sigma}(\boldsymbol{a}) \to_p 0. \tag{13}$$

To show (12), we will utilize the martingale theories in Anderson and Gill (1982) [1] to prove that $\boldsymbol{a}^\top \left[-\{\mathrm{PL}\}''(\boldsymbol{\gamma}^*)\right]^{-1} C'(\tilde{\boldsymbol{\gamma}},t)/\hat{\sigma}(\boldsymbol{a})$ is converging to a Gaussian process. Indeed,

$$C'(\tilde{\boldsymbol{\gamma}},t) = \sum_{i=1}^{n} \int_0^\top \{A_i(\tilde{\boldsymbol{\gamma}},s) - E(\tilde{\boldsymbol{\gamma}},s)\}\, dM_i(s),$$

where $A_i(\tilde{\boldsymbol{\gamma}},s) = \boldsymbol{U}_i \otimes \boldsymbol{B}_i$ and $E(\tilde{\boldsymbol{\gamma}},s) = S_1(\tilde{\boldsymbol{\gamma}},s)/S_0(\tilde{\boldsymbol{\gamma}},s)$. Then we have

$$\frac{\boldsymbol{a}^\top \left[-\{\mathrm{PL}\}''(\boldsymbol{\gamma}^*)\right]^{-1}}{\hat{\sigma}(\boldsymbol{a})} C'(\tilde{\boldsymbol{\gamma}},t) = \sum_{i=1}^{n} \int_0^\top \frac{\boldsymbol{a}^\top \left[-\{\mathrm{PL}\}''(\boldsymbol{\gamma}^*)\right]^{-1}}{\hat{\sigma}(\boldsymbol{a})} \{A_i(\tilde{\boldsymbol{\gamma}},s) - E(\tilde{\boldsymbol{\gamma}},s)\}\, dM_i(s).$$

Let

$$H_i(s) = \frac{\boldsymbol{a}^\top \left[-\{\mathrm{PL}\}''(\boldsymbol{\gamma}^*)\right]^{-1}}{\hat{\sigma}(\boldsymbol{a})} \{A_i(\tilde{\boldsymbol{\gamma}},s) - E(\tilde{\boldsymbol{\gamma}},s)\},$$

we then can show claim 12 is true by applying Theorem I.2 in Anderson and Gill (1982) [1]. Condition (I.3) of Theorem I.2 is valid, because by Conditions C2, C7 and C8, we have

$$\begin{aligned}
\int_0^\top \sum_{i=1}^{n} H_i^2(s)\lambda_i(s)\mathrm{d}s =& \boldsymbol{a}^\top \left[-\{\mathrm{PL}\}''(\boldsymbol{\gamma}^*)\right]^{-1} \cdot \\
& \int_0^\top \sum_{i=1}^{n} \{A_i(\tilde{\boldsymbol{\gamma}},s) - E(\tilde{\boldsymbol{\gamma}},s)\} \{A_i(\tilde{\boldsymbol{\gamma}},s) - E(\tilde{\boldsymbol{\gamma}},s)\}^\top \lambda_i(s)\mathrm{d}s \cdot \\
& \left[-\{\mathrm{PL}\}''(\boldsymbol{\gamma}^*)\right]^{-1} \boldsymbol{a}/\hat{\sigma}^2(\boldsymbol{a}) \\
\to_p & \, r(t),
\end{aligned}$$

where $r(t)$ is some positive function of $t$ and $r(1) = 1$.

By similar arguments in Anderson and Gill (1982) [1], condition (I.4) of Theorem I.2 is true by Conditions C2, C7, and C9. Then claim (12) is valid.

Claim (13) is valid because

$$\rho \left| \boldsymbol{a}^\top \left[\{\mathrm{PL}\}''(\boldsymbol{\gamma}^*)\right]^{-1} \sum_{i=1}^{n} \boldsymbol{\theta} \otimes \boldsymbol{B}(T_i)/\hat{\sigma}(\boldsymbol{a}) \right| \leq O_p(n\rho) \to_p 0 \tag{14}$$

by Condition C6. Therefore, for any non-zero $\boldsymbol{a}$,

$$\frac{\boldsymbol{a}^\top(\hat{\boldsymbol{\gamma}}-\tilde{\boldsymbol{\gamma}})}{\hat{\sigma}(\boldsymbol{a})} \to_d N(0,1),$$

where $\hat{\sigma}(\boldsymbol{a}) = n\boldsymbol{a}^\top \left[\{\mathrm{PL}\}''(\tilde{\boldsymbol{\gamma}})\right]^{-1} \Sigma(\tilde{\boldsymbol{\gamma}},1) \left[\{\mathrm{PL}\}''(\tilde{\boldsymbol{\gamma}})\right]^{-1} \boldsymbol{a}$.

Since for any $t \in [0,\tau]$, $\hat{\theta}_j(t) = (\boldsymbol{e}_j \otimes \boldsymbol{B}(t))^\top \hat{\boldsymbol{\gamma}}$, then let $\boldsymbol{a} = \boldsymbol{e}_j \otimes \boldsymbol{B}(t)$, we have for any $t \in [0,\tau]$,

$$\frac{\hat{\theta}_j(t)-\theta_j(t)}{\sigma_{nj}(t)} \to_d N(0,1),$$

where $\sigma^2_{nj}(t) = n\{\boldsymbol{e}_j \otimes \boldsymbol{B}(t)\}^\top \left[-\{\mathrm{PL}\}''(\boldsymbol{\gamma}^*)\right]^{-1} \Sigma(\tilde{\boldsymbol{\gamma}},1) \left[-\{\mathrm{PL}\}''(\boldsymbol{\gamma}^*)\right]^{-1} \{\boldsymbol{e}_j \otimes \boldsymbol{B}(t)\}$.

Finally, denote by $\hat{\sigma}^2_{nj}(t) = n\{\boldsymbol{e}_j \otimes \boldsymbol{B}(t)\}^\top \left[-\{\mathrm{PL}\}''(\hat{\boldsymbol{\gamma}})\right]^{-1} \Sigma(\hat{\boldsymbol{\gamma}},1) \left[-\{\mathrm{PL}\}''(\hat{\boldsymbol{\gamma}})\right]^{-1} \{\boldsymbol{e}_j \otimes \boldsymbol{B}(t)\}$. As $||\hat{\boldsymbol{\gamma}}-\tilde{\boldsymbol{\gamma}}||_2 \to_p 0$, it follows that $\hat{\sigma}^2_{nj}(t)/\sigma^2_{nj}(t) \to_p 1$ for $t > 0$. Hence, by the Slutsky theorem,

$$\frac{\hat{\theta}_j(t)-\theta_j(t)}{\hat{\sigma}_{nj}(t)} \to_d N(0,1),$$

which justifies the use of $\hat{\sigma}_{nj}(t)$ as a consistent estimate of the variance. ■

## References


[1] Anderson JA, Senthilselvan A (1982) A Two-Step Regression Model for Hazard Functions. Journal of the Royal Statistical Society Series C (Applied Statistics) 31(1):44–51, DOI 10.2307/2347073, URL http://www.jstor.org/stable/2347073

[2] Barbeau EM, Li Y, Calderon P, Hartman C, Quinn M, Markkanen P, Roelofs C, Frazier L, Levenstein C (2006) Results of a union-based smoking cessation intervention for apprentice iron workers (United States). Cancer Causes & Control 17(1):53–61

[3] Cai Z, Sun Y (2003) Local Linear Estimation for Time-Dependent Coefficients in Cox's Regression Models. Scandinavian Journal of Statistics 30(1):93–111, DOI 10.1111/1467-9469.00320, URL papers2://publication/uuid/14A8538C-13F5-4307-8491-4505CAFCB8F9

[4] Chang SG, Yu B, Vetterli M (2000) Adaptive wavelet thresholding for image denoising and compression. IEEE transactions on image processing 9(9):1532–1546

[5] Chau M, Fu MC, Qu H, Ryzhov IO (2014) Simulation optimization: a tutorial overview and recent developments in gradient-based methods. In: Proceedings of the Winter Simulation Conference 2014, IEEE, pp 21–35

[6] Christiani DC (2017) The Boston lung cancer survival cohort. Tech. rep., NIH, URL http://grantome. com/grant/NIH/U01-CA209414-01A1

[7] Cox DR (1972) Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological) 34(2):187–202
[8] D Schoenfeld (1982) Partial Residuals for The Proportional Hazards Regression Model. Biometrika 69(1):239–241, URL http://www.jstor.org/stable/2335876
[9] Donoho DL (1995) De-Noising by Soft-Thresholding. IEEE Transactions on Information Theory 41(3):613–627, DOI 10.1109/18.382009, 0611061v2
[10] Donoho DL, Johnstone IM (1994) Ideal spatial adapatation by wavelet shrinkage. Biometrika 81(3):425–455
[11] Gore SM, Pocock SJ, Kerr GR (1984) Regression Models and Non-Proportional Hazards in the Analysis of Breast Cancer Survival. Journal of the Royal Statistical Society Series C (Applied Statistics) 33(2):176–195
[12] Hastie TJ, Tibshirani R (1993) Varying-coefficient Models. Journal of the royal statistical society 55(4):757–796, DOI 10.2307/2345993
[13] He K, Yang Y, yan L, Zhu J, Li Y (2017) Modeling Time-varying Effects with Large-scale Survival Data : An Efficient Quasi-Newton Approach. Journal of Computational and Graphical Statistics 26(3)
[14] Hong HG, Christiani DC, Li Y (2019) Quantile regression for survival data in modern cancer research: expanding statistical tools for precision medicine. Precision clinical medicine 2(2):90–99
[15] Huang J (1999) Efficient estimation of the partly linear additive Cox model. Annals of Statistics 27(5):1536–1563
[16] Kang J, Reich BJ, Staicu AM (2018) Scalar-on-image regression via the soft-thresholded gaussian process. Biometrika 105(1):165–184
[17] Lian H, Lai P, Liang H (2013) Partially linear structure selection in cox models with varying coefficients. Biometrics 69(2):348–357, DOI 10.1111/biom.12024
[18] Martinussen T, Scheike TH, Skovgaard IM (2002) Efficient estimation of fixed and time-varying covariate effects in multiplicative intensity models. Scandinavian Journal of Statistics 29(1):57–74, DOI 10.1111/1467-9469.00060
[19] Marzec L (1997) On fitting Cox's regression model with time-dependent coefficients. Biometrika 84(4):901–908, DOI 10.1093/biomet/84.4.901, URL http://biomet.oupjournals.org/cgi/doi/10.1093/biomet/84.4.901
[20] Murphy S (1993) Testing for a time dependent coefficient in Cox's regression model. Scandinavian Journal of Statistics 20(1):35–50, URL http://www.jstor.org/stable/10.2307/4616258
[21] Sasieni P (1992) Non-orthogonal projections and their application to calculating the information in a partly linear cox model. Scandinavian Journal of Statistics pp 215–233
[22] Schumaker L (2007) Spline Functions: Basic Theory. Cambridge: Cambridge University Press
[23] Shen X, Wong WH (1994) Convergence rate of sieve estimates. The Annals of Statistics 22:580–615
[24] Stone CJ (1985) Additive regression and other nonparametric models. The Annals of Statistics 13(2):689–705

[25] Tian L, Zucker D, Wei LJ (2005) On the Cox Model With Time-Varying Regression Coefficients. Journal of the American Statistical Association 100(469):172–183, DOI 10.1198/016214504000000845

[26] Tibshirani R (1996) Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology 58(1):267–288, DOI 10.1111/j.1467-9868.2011.00771.x, 11/73273

[27] Van Der Vaart AW, Wellner JA (1996) Weak convergence. In: Weak convergence and empirical processes, Springer, pp 16–28

[28] Wahba G (1980) Ill posed problems: Numerical and statistical methods for mildly, moderately and severely ill posed problems with noisy data. Tech. rep., WISCONSIN UNIV-MADISON DEPT OF STATISTICS

[29] Winnett A, Sasieni P (2003) Iterated residuals and time-varying covariate effects in Cox regression. Journal of the Royal Statistical Society Series B: Statistical Methodology 65(2):473–488, DOI 10.1111/1467-9868.00397

[30] Wright SJ (2015) Coordinate descent algorithms. Mathematical Programming 151(1):3–34

[31] Yan J, Huang J (2012) Model Selection for Cox Models with Time-Varying Coefficients. Biometrics 68(2):419–428, DOI 10.1111/j.1541-0420.2011.01692.x, NIHMS150003

[32] Zheng D, Ye T, Hu H, Zhang Y, Sun Y, Xiang J, Chen H (2018) Upfront surgery as first-line therapy in selected patients with stage iiia non–small cell lung cancer. The Journal of thoracic and cardiovascular surgery 155(4):1814–1822

[33] Zucker DM, Karr AF (1990) Nonparametric Survival Analysis with Time-Dependent Covariate Effects: A Penalized Partial Likelihood Approach. The Annals of Statistics 18(1):329–353, DOI 10.1214/aos/1176347503