Stratified Cox Models with Time-Varying Effects for National Kidney Transplant Patients: A New Block-Wise Steepest Ascent Method

Kevin He[1,*], Ji Zhu[2], Jian Kang[1] and Yi Li[1]

[1]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, US

[2]Department of Statistics, University of Michigan, Ann Arbor, Michigan, US

[*]*email:* kevinhe@umich.edu

Summary: Analyzing the national transplant database, which contains about 300,000 kidney transplant patients treated in over 290 transplant centers, may guide the disease management, and inform the policy of kidney transplantation. Cox models stratified by centers provide a convenient means to account for the clustered data structure, while studying more than 160 predictors with effects that may vary over time. As fitting a time-varying effect model with such a large sample size may defy any existing software, we propose a block-wise steepest ascent procedure by leveraging the block structure of parameters inherent from the basis expansions for each coefficient function. The algorithm iteratively updates the optimal block-wise search direction, along which the increment of the partial likelihood is maximized. The proposed method can be interpreted from the perspective of the Minorization-Maximization algorithm and increases the partial likelihood until convergence. We further propose a Wald statistic to test whether the effects are indeed time varying. We evaluate the utility of the proposed method via simulations. Finally, we apply the method to analyze the national kidney transplant data and detect the time-varying nature of the effects of various risk factors.

Key words: Kidney transplant; Steepest ascent; Stratified model; Survival analysis; Time-varying effects.

1. Introduction

End-stage renal disease (ESRD) is one of the most deadly and costly diseases in the United States (Saran et al., 2018), and kidney transplantation is the most preferred treatment (Wolfe et al., 1999). Despite much effort to improve survival, the mortality of kidney transplant recipients is still thrice higher than that of the general population. Identifying risk factors associated with post-transplant mortality is pivotal in prolonging the survival of transplant patients and optimizing organ allocations (Snyder et al., 2016). The widely used proportional hazards model (Cox, 1972) assumes that the effects of covariates are constant over time, which is often violated. For example, contrary to the common belief that obesity is a risk factor for mortality, Kalantar (2005) and Dekker et al. (2008) showed obesity has a short-term protective effect, but is a risk factor in the long run. Models that feature time-varying effects provide valuable clinical information. The national kidney transplant data, obtained from the U.S. Organ Procurement and Transplantation Network (OPTN, 2013), contains more than 160 predictors for over 300,000 patients who underwent transplantation between 1988 and 2012. Analyzing this dataset may guide the disease management and inform the transplantation policy. Existing statistical methods that perform well for moderate sample sizes and small-dimensional data do not scale to this data because of the large size of the involved at-risk sets (He et al., 2017a). Of special interest is how time-varying effect models can be extended to accommodate large-scale time-to-event data.

Another important aspect of our motivating example is that patients came from multiple transplant centers. In the absence of adjustment for center effects, the estimation of covariate effects may be biased due to uncontrolled confounding by centers (Pan, 2002; Kalbfleisch and Wolfe, 2013; He and Schaubel, 2015). One could estimate the center effects through frailty models (He et al., 2017b). However, the commonly frailty approach assumes that the center effects are constant over time, which is often violated (He and Schaubel, 2014, 2015) and much work is needed to implement time-varying frailty models that can be applicable to the national kidney transplant database. We propose to adopt a stratified model with stratum-specific baseline hazards, which avoids modeling the center effects explicitly and simplifies the computation of the partial likelihood by downsizing the at-risk sets.

Methods have been proposed for relative risk models with time-varying effects: Zucker and Karr (1990) conducted a nonparametric estimation of the time-varying effects; a specialized algorithm for this problem was provided by Hastie and Tibshirani (1993); Gray (1992, 1994) proposed using fixed knots spline functions. He et al. (2017a) implemented a quasi-Newton algorithm; He et al. (2017b) further considered a frailty model with time-varying effects. Kernel-based partial likelihood approaches have also been developed (Tian et al., 2005). Some recent studies (Honda and Härdle, 2014; Yan and Huang, 2012) have proposed variable selection of time-varying effects using penalized methods such as adaptive lasso (Zou, 2006; Zhang and Lu, 2006). Xiao et al. (2016) combined the ideas of local polynomial smoothing and group non-negative garrote to achieve these goals. Alternatively, Hofner et al. (2013) proposed a component-wise likelihood boosting algorithm for survival data that permits the inclusion of both parametric and nonparametric time-varying effects.

These methods may not be applicable to studies with large sample sizes or many covariates. When implementing them, datasets are usually expanded in a repeated measurement format, where the time is divided into small intervals which contain a distinctive event. The covariate values and outcomes for all at-risk subjects at each interval are stacked to form a working dataset, which becomes infeasible for a large sample size. As a remedy, a routine based on the Kronecker product has been suggested (Perperoglou et al., 2006). Even with this tool, for large-scale kidney transplant data, existing methods easily overwhelm powerful computers.

Moreover, time-varying effects are often represented by basis expansions using B-splines. The parameter vector, consisting of coefficients of the bases, possesses a block structure, of which the dimension increases quickly as the number of predictors grows. This leads to unstable estimates for the commonly used Newton (Perperoglou et al., 2006) and quasi-Newton (He et al., 2017a) methods. To see this, we conducted a simulation (Setting B of Section 4) to assess the biases of the Newton approach, gradient ascent, and stochastic gradient ascent implemented by adaptive moment estimation (ADAM) (Diederik et al., 2015); see Figure 1. Alternative stochastic gradient approaches such as Annealing (Robbins and Monro, 1951), Momentum (Qian, 1999), Adagrad (Duchi et al., 2011) and Adadelta (Zeiler, 2012) were also conducted; their performances were worse than ADAM and not shown. The Newton approach introduces large biases, and Gradient-based methods are less efficient by overlooking Hessian matrices. The issue becomes more exacerbated for the analysis of the kidney transplant database, wherein many comorbidities have rare frequencies.

We propose a block-wise steepest ascent (BSA) procedure for stratified time-varying effect models, which makes the following contributions. First, BSA iteratively updates the optimal block-wise search direction, avoids complicated computation of inverting the observed information matrix and, hence, is computationally efficient for large-scale problems. Second, BSA converts a high-dimensional optimization problem into a sequence of low-dimensional ones. Simplicity is achieved by substituting a surrogate function that is separable for different blocks of parameters. Third, BSA can be interpreted from the perspective of the Minorization-Maximization (MM) algorithm (Lange, 2012). The updated estimates ensure the increment of likelihood. Fourth, unlike the classical gradient-based procedures, which typically rely on a first order approximation and a large number of iterations, the proposed BSA utilizes a block-wise second order approximation and achieves faster convergence; see Figure 1. Finally, choosing a proper learning rate for classical gradient-based methods can be cumbersome, whereas BSA is less sensitive to the choice of learning rates and our numerical properties help clarify the required learning rates and their roles in various methods.

The remainder of this article is organized as follows. We describe the proposed BSA procedure and testing algorithm for time-varying effects in Section 2. Convergence properties are considered in Section 3. Numerical properties are examined in Section 4 through simulations. We apply BSA to analyze the national kidney transplant data in Section 5. The article concludes with a discussion in Section 6.

2. Method

2.1 *Stratified Time-Varying Effect Model*

Let D_{ij} denote the time lag from transplantation to death and C_{ij} be the censoring time for patient i in center j, $i = 1, \ldots, n_j$, and $j = 1, \ldots, J$. Here n_j is the sample size in center j, and J is the number of centers. The total number of patients is $N = \sum_{j=1}^{J} n_j$, the observed time is $T_{ij} = \min\{D_{ij}, C_{ij}\}$, and the death indicator is given by $\delta_{ij} = I(D_{ij} \leqslant C_{ij})$. Let $\mathbf{X}_{ij} = (X_{ij1}, \ldots, X_{ijP})^T$ be a P-dimensional covariate vector. We assume that D_{ij} is independent from C_{ij} given $\mathbf{X}_{ij}$. Consider a stratum-specific hazard function

$$\lambda(t|\mathbf{X}_{ij}) = \lambda_{0j}(t) \exp\{\mathbf{X}_{ij}^T \boldsymbol{\beta}(t)\},$$

where $\lambda_{0j}(t)$ is the baseline hazard for stratum j. To estimate the time-varying coefficients $\boldsymbol{\beta}(t) = \{\beta_1(t), \ldots, \beta_P(t)\}$, we span $\boldsymbol{\beta}(\cdot)$ by a set of cubic B-splines defined on a given number of knots:

$$\beta_p(t) = \boldsymbol{\theta}_p^T \mathbf{B}(t) = \sum_{k=1}^{K} \theta_{pk} B_k(t), \quad p = 1, \ldots, P,$$

where $\mathbf{B}(t) = \{B_1(t), \ldots, B_K(t)\}^T$ forms a basis, K is the number of basis functions, and $\boldsymbol{\theta}_p = (\theta_{p1}, \ldots, \theta_{pK})^T$ is a vector of coefficients with θ_{pk} being the coefficient for the k-th basis of the p-th covariate. With a length-PK parameter vector $\boldsymbol{\theta} = vec(\boldsymbol{\Theta})$, the vectorization of the coefficient matrix $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_P)^T$ by row, the log-partial likelihood function is

$$\ell(\boldsymbol{\theta}) = \sum_{j=1}^{J} \sum_{i=1}^{n_j} \delta_{ij} \left[\mathbf{X}_{ij}^T \boldsymbol{\Theta} \mathbf{B}(T_{ij}) - \log \left\{ \sum_{i' \in R_{ij}} \exp\{\mathbf{X}_{i'j}^T \boldsymbol{\Theta} \mathbf{B}(T_{ij})\} \right\} \right], \qquad (1)$$

where $R_{ij} = \{i' : 1 \leqslant i' \leqslant n_j,\ T_{i'j} \geqslant T_{ij}\}$ is the at-risk set for stratum j. That $\boldsymbol{\theta}$ has P "blocks" of subvectors, i.e. $\boldsymbol{\theta}_p, p = 1, \ldots, P$, each corresponding to a covariate, will inform the development of our proposed block-wise steepest ascent algorithm.

2.2 *Review of Newton Approach*

When both N and P are moderate, maximization of (1) can be achieved by a Newton approach, which requires computation of the gradient and Hessian matrix, given by $\nabla \ell(\boldsymbol{\theta}) = \sum_j \sum_i \Psi_{ij}(\boldsymbol{\theta})$ and

$$\nabla^2 \ell(\boldsymbol{\theta}) = -\sum_{j=1}^{J} \sum_{i=1}^{n_j} \delta_{ij} \mathbf{V}_{ij}(\boldsymbol{\Theta}, T_{ij}) \otimes \left\{\mathbf{B}(T_{ij})\mathbf{B}^T(T_{ij})\right\}, \tag{2}$$

respectively. Here $\otimes$ is the Kronecker product, and

$$\Psi_{ij}(\boldsymbol{\theta}) = \delta_{ij} \left\{ \mathbf{X}_{ij} - \frac{\mathbf{S}_{ij}^{(1)}(\boldsymbol{\Theta}, T_{ij})}{\mathbf{S}_{ij}^{(0)}(\boldsymbol{\Theta}, T_{ij})} \right\} \otimes \mathbf{B}(T_{ij}), \tag{3}$$

where

$$\mathbf{V}_{ij}(\boldsymbol{\Theta}, T_{ij}) = \frac{\mathbf{S}_{ij}^{(2)}(\boldsymbol{\Theta}, T_{ij}) S_{ij}^{(0)}(\boldsymbol{\Theta}, T_{ij}) - \{\mathbf{S}_{ij}^{(1)}(\boldsymbol{\Theta}, T_{ij})\}^{\otimes 2}}{\{S_{ij}^{(0)}(\boldsymbol{\Theta}, T_{ij})\}^2},$$

$$\mathbf{S}_{ij}^{(r)}(\boldsymbol{\Theta}, T_{ij}) = \sum_{i' \in R_{ij}} \exp\{\mathbf{X}_{i'j}^T \boldsymbol{\Theta} \mathbf{B}(T_{ij})\} \mathbf{X}_{i'j}^{\otimes r},$$

for $r = 0, 1, 2$. For a column vector $\mathbf{v}$, $\mathbf{v}^{\otimes 0} = 1$, $\mathbf{v}^{\otimes 1} = \mathbf{v}$ and $\mathbf{v}^{\otimes 2} = \mathbf{v}\mathbf{v}^T$.

Computational burden mainly comes from two sources. First, summations across all the risk sets are cumbersome, especially when N is large. Second, with a large P, inversions of Hessian matrices are costly. In summary, the computation complexities of the Newton method for the un-stratified time-varying effect model and the stratified time-varying effect model are at the order of $O(N^2P^2K^2 + P^3K^3)$ and $O(N^2P^2K^2/J + P^3K^3)$, respectively. Though stratified models reduce the first term by a factor of J, the Newton approach is still numerically challenging or even impractical for large sample and high-dimensional problems. This motivates us to propose a feasible approach that reduces the computation complexity to an order of $O(N^2PK^2/J + PK^3)$; see the next section.

2.3 *Proposed Block-Wise Steepest Ascent*

Given a current estimate $\widehat{\boldsymbol{\theta}}$, we consider a first-order Taylor's expansion:

$$\ell(\widehat{\boldsymbol{\theta}} + \alpha\boldsymbol{\mu}) = \ell(\widehat{\boldsymbol{\theta}}) + \alpha\nabla\ell(\widehat{\boldsymbol{\theta}})^T\boldsymbol{\mu} + \frac{1}{2}\alpha^2\boldsymbol{\mu}^T\nabla^2\ell(\widehat{\boldsymbol{\theta}} + w\boldsymbol{\mu})\boldsymbol{\mu},$$

where $\boldsymbol{\mu}$ is the update direction of $\boldsymbol{\theta}$, α is a small positive value, $w \in [0, \alpha]$, and the term $\nabla\ell(\widehat{\boldsymbol{\theta}})^T\boldsymbol{\mu}$ is the directional derivative along $\boldsymbol{\mu}$. If $\nabla\ell(\widehat{\boldsymbol{\theta}})^T\boldsymbol{\mu} > 0$, the direction $\boldsymbol{\mu}$ is an ascent direction of $\boldsymbol{\theta}$ to increase $\ell(\boldsymbol{\theta})$. We identify an update direction (with a unit norm), along which $\ell(\boldsymbol{\theta})$ ascends most rapidly. This motivates us to find a steepest ascent direction,

$$\boldsymbol{\mu}^\star = \underset{\boldsymbol{\mu}}{\operatorname{argmax}}\{\nabla\ell(\widehat{\boldsymbol{\theta}})^T\boldsymbol{\mu} \mid ||\boldsymbol{\mu}||_\dagger = 1\}, \tag{4}$$

where $||\cdot||_\dagger$ is a vector norm on $\mathbb{R}^{PK}$. As the choice of norm $||\boldsymbol{\mu}||_\dagger$ plays a crucial role in computational efficiency and numerical stability, we propose to use a block-quadratic norm by leveraging the block structure of the parameter vector $\boldsymbol{\theta}$,

$$||\boldsymbol{\mu}||_\dagger = \sum_{p=1}^{P} ||\boldsymbol{\mu}_p||_{\mathbf{H}_p(\widehat{\boldsymbol{\theta}})}, \tag{5}$$

where $||\boldsymbol{\mu}_p||_{\mathbf{H}_p(\widehat{\boldsymbol{\theta}})}$ is a quadratic norm, defined as $||\boldsymbol{\mu}_p||_{\mathbf{A}} = \left(\boldsymbol{\mu}_p^T\mathbf{A}\boldsymbol{\mu}_p\right)^{1/2}$ for a positive semi-definite matrix $\mathbf{A}$. Here $\boldsymbol{\mu}_p$ is a K-dimensional vector corresponding to the p-th block of $\boldsymbol{\mu}$, and $\mathbf{H}_p(\widehat{\boldsymbol{\theta}})$ is a $K \times K$-dimensional matrix.

A simple choice is to set $\mathbf{H}_p(\widehat{\boldsymbol{\theta}})$ as an identity matrix, leading to a block-wise gradient ascent method with low computation cost at each iteration; however, its convergence can be slow, especially when the condition numbers of the observed information matrix are large; see Section 3. To address this problem, for $p = 1, \ldots, P$, we choose

$$\mathbf{H}_p(\widehat{\boldsymbol{\theta}}) = -\left[\nabla\ell(\widehat{\boldsymbol{\theta}})_p^T\{-\nabla^2\ell(\widehat{\boldsymbol{\theta}})_p\}^{-1}\nabla\ell(\widehat{\boldsymbol{\theta}})_p\right]\nabla^2\ell(\widehat{\boldsymbol{\theta}})_p, \tag{6}$$

where $\nabla\ell(\widehat{\boldsymbol{\theta}})_p$ is the p-th block of the gradient vector and $\nabla^2\ell(\widehat{\boldsymbol{\theta}})_p$ is the block diagonal of the Hessian matrix defined in (2), corresponding to the p-th variable. Here the scalar $\nabla\ell(\widehat{\boldsymbol{\theta}})_p^T\{-\nabla^2\ell(\widehat{\boldsymbol{\theta}})_p\}^{-1}\nabla\ell(\widehat{\boldsymbol{\theta}})_p$ is a normalization factor.

With the Cauchy-Schwarz inequality,

$$\nabla \ell(\widehat{\boldsymbol{\theta}})^T \boldsymbol{\mu} \leqslant \sum_{p=1}^{P} ||\nabla \ell(\widehat{\boldsymbol{\theta}})_p||_{\mathbf{H}_p^{-1}(\widehat{\boldsymbol{\theta}})} ||\boldsymbol{\mu}_p||_{\mathbf{H}_p(\widehat{\boldsymbol{\theta}})} \leqslant \left\{ \max_p \left(||\nabla \ell(\widehat{\boldsymbol{\theta}})_p||_{\mathbf{H}_p^{-1}(\widehat{\boldsymbol{\theta}})} \right) \right\} \sum_{p=1}^{P} ||\boldsymbol{\mu}_p||_{\mathbf{H}_p(\widehat{\boldsymbol{\theta}})}.$$

With $\boldsymbol{\mu}$ satisfying $\sum_p ||\boldsymbol{\mu}_p||_{\mathbf{H}_p(\widehat{\boldsymbol{\theta}})} = 1$, we have

$$\nabla \ell(\widehat{\boldsymbol{\theta}})^T \boldsymbol{\mu} \leqslant \max_p \left(||\nabla \ell(\widehat{\boldsymbol{\theta}})_p||_{\mathbf{H}_p^{-1}(\widehat{\boldsymbol{\theta}})} \right).$$

The resulting block-wise steepest ascent direction

$$\boldsymbol{\mu}^\star = \operatorname*{argmax}_{\boldsymbol{\mu}} \{ \nabla \ell(\widehat{\boldsymbol{\theta}})^T \boldsymbol{\mu} \mid ||\boldsymbol{\mu}||_\dagger = 1 \} = (0, \ldots, 0, \widetilde{\boldsymbol{\mu}}_{p^\star}^T, 0, \ldots, 0)^T, \tag{7}$$

maximizes the directional derivative, i.e.

$$\nabla \ell(\widehat{\boldsymbol{\theta}})^T \boldsymbol{\mu}^\star = \max_p \left(||\nabla \ell(\widehat{\boldsymbol{\theta}})_p||_{\mathbf{H}_p^{-1}(\widehat{\boldsymbol{\theta}})} \right),$$

and let

$$p^\star = \operatorname*{argmax}_p \left(||\nabla \ell(\widehat{\boldsymbol{\theta}})_p||_{\mathbf{H}_p^{-1}(\widehat{\boldsymbol{\theta}})} \right), \tag{8}$$

with $\widetilde{\boldsymbol{\mu}}_{p^\star}$ given by

$$\widetilde{\boldsymbol{\mu}}_{p^\star} = \left\{ \mathbf{H}_{p^\star}(\widehat{\boldsymbol{\theta}}) \right\}^{-1} \nabla \ell(\widehat{\boldsymbol{\theta}})_{p^\star}. \tag{9}$$

We summarize the proposed algorithm as follows:

BSA Algorithm

(a) Initialize $\widehat{\boldsymbol{\theta}}^{(0)} = \mathbf{0}$. For $m = 1, 2, 3, \ldots$, identify $p^\star$ as in (8).

(b) Update the estimate by $\widehat{\boldsymbol{\theta}}_{p^\star}^{(m)} = \widehat{\boldsymbol{\theta}}_{p^\star}^{(m-1)} + \nu \; \widetilde{\boldsymbol{\mu}}_{p^\star}$.

(c) The iteration continues until the directional derivative $\nabla \ell(\widehat{\boldsymbol{\theta}}^{(m)})^T \boldsymbol{\mu}^\star$ or the relative change in the log-partial likelihood is less than a convergence threshold (e.g. 10^{-6}).

We comment that the block-wise algorithm ranks the importance of each predictor and measures how fast the log-partial likelihood would increase by including each predictor; the proposed algorithm converts a difficult optimization problem into a simpler surrogate function that is separable across blocks of the parameter vector and avoids iterative inversions of high dimensional Hessian matrices; the learning rate, ν, can be chosen to be a small positive value, e.g. 0.05. Further clarification for the choice of ν is provided in Section 3.

2.4 *Minorization-Maximization-Based Interpretation*

The proposed method can be interpreted from the perspective of the MM algorithm, which reaffirms the ascent property and helps clarify the numerical advantage of the proposed procedure. To see this, we note the block-quadratic norm considered in (5) leads to a minority surrogate function

$$g(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) = \ell(\widehat{\boldsymbol{\theta}}) + \nabla\ell(\widehat{\boldsymbol{\theta}})^T(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}) - \frac{1}{2\nu}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^T\mathbf{H}(\widehat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}),$$

where ν is a small positive value to be specified and $\mathbf{H}(\widehat{\boldsymbol{\theta}}) = \text{diag}\{\mathbf{H}_1(\widehat{\boldsymbol{\theta}}), \mathbf{H}_2(\widehat{\boldsymbol{\theta}}), \cdots, \mathbf{H}_P(\widehat{\boldsymbol{\theta}})\}$ is a block-diagonal matrix, and $\mathbf{H}_p(\widehat{\boldsymbol{\theta}})$ is defined in (6). Here, the blocks correspond to the basis expansions for each variable. With $g(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}) = \ell(\widehat{\boldsymbol{\theta}})$, Proposition 1 in Section 3 shows that, given a suitable ν, $g(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) \leqslant \ell(\boldsymbol{\theta})$ for all $\boldsymbol{\theta}$. Thus, $g(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})$ serves as a minority surrogate function of $\ell(\boldsymbol{\theta})$. Leveraging the block-diagonal structure of $\mathbf{H}(\widehat{\boldsymbol{\theta}})$, the minority surrogate function $g(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})$ is separable across the blocks of parameters. Therefore, this "minorization" step reduces a high-dimensional optimization problem to simpler ones.

The block-wise update [as in (8) and (9)] maximizes $g(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})$ subject to the constraint that only one variable is updated at each iteration. This "Maximization" step, coupled with the previous "minorization" step, is essentially a Minorization-Maximization-based steepest ascent procedure, which iteratively pursues the optimal block-wise update direction.

2.5 *Connection with Existing Optimization Approaches*

It is instructive to assess several commonly used norms for (4) and tie them to the existing steepest ascent approaches. For example, an ℓ_2 norm corresponds to the gradient ascent method:

$$\boldsymbol{\mu}^\star = \underset{\boldsymbol{\mu}}{\text{argmax}}\{\nabla\ell(\widehat{\boldsymbol{\theta}})^T\boldsymbol{\mu} \mid ||\boldsymbol{\mu}||_2 = 1\} = \nabla\ell(\widehat{\boldsymbol{\theta}})/||\nabla\ell(\widehat{\boldsymbol{\theta}})||_2.$$

As illustrated in Figure 1 and Web Figures S1 and S2 in the supporting information, the convergence of gradient-based methods is slow, especially when the observed information matrix is ill-conditioned (i.e. near singular). One may consider a quadratic norm $||\boldsymbol{\mu}||_{\mathbf{A}} = \left(\boldsymbol{\mu}^T \mathbf{A} \boldsymbol{\mu}\right)^{1/2}$ with $\mathbf{A} = -\nabla^2 \ell(\widehat{\boldsymbol{\theta}})$, with the update direction coinciding with the Newton update, which becomes numerically unstable or even impractical for large-scale data. Also, an ℓ_1 norm leads to the coordinate-wise gradient boosting procedure (Bühlmann and Yu, 2003, 2006; He et al., 2016). However, this procedure does not take into account the group structure and will lead to sparse basis presentations, and is not suitable for estimating time-varying effects.

2.6 *Testing for Time-Varying Effects*

To test whether the effects are time-varying, we use the constant property of B-splines, that is, if $\theta_{p1} = \cdots = \theta_{pK}$, the corresponding covariate effect is time-independent. Specify a matrix $\mathbf{C}_p$ such that $\mathbf{C}_p \boldsymbol{\theta} = \mathbf{0}$ corresponds to the contrast that $\theta_{p1} = \cdots = \theta_{pK}$. Following He et al. (2017a), a Wald statistic can be constructed by

$$(\mathbf{C}_p \widehat{\boldsymbol{\theta}})^T \left[\mathbf{C}_p \{-\nabla^2 \ell(\widehat{\boldsymbol{\theta}})\}^{-1} \mathbf{C}_p^T\right]^{-1} (\mathbf{C}_p \widehat{\boldsymbol{\theta}}),$$

where $\widehat{\boldsymbol{\theta}}$ is obtained through the proposed BSA.

In the kidney transplant database with large N and P, computation of the observed information matrix is infeasible as discussed in Section 2.2, though gradients are easier to compute. We consider a modified statistic

$$S_p = (\mathbf{C}_p \widehat{\boldsymbol{\theta}})^T \{\mathbf{C}_p \mathbf{V}^{-1}(\widehat{\boldsymbol{\theta}}) \mathbf{C}_p^T\}^{-1} (\mathbf{C}_p \widehat{\boldsymbol{\theta}}), \tag{10}$$

where $\mathbf{V}(\widehat{\boldsymbol{\theta}}) = \sum_{j=1}^{J} \sum_{i=1}^{n_j} \Psi_{ij}(\widehat{\boldsymbol{\theta}}) \Psi_{ij}(\widehat{\boldsymbol{\theta}})^T$ is an approximation of the empirical information matrix (McLachlan and Krishnan, 2007), with Ψ_{ij} defined in (3). Under the null hypothesis that the effect is time-independent, S_p is asymptotically chi-square distributed with $K - 1$ degrees of freedom. To incorporate potential correlations among patients within strata, a robust inference procedure (Lin and Wei, 1989; Schaubel and Cai, 2005) can be adopted.

2.7 *Variable Selection with High-Dimensional Covariates*

Our proposed BSA algorithm can also be extended to accommodate a large P small N problem. Specifically, BSA is a group-wise procedure. With only one variable updated at each iteration, variable selection can be achieved if the procedure is set to stop at a finite number of steps. Effectively, the step number is a tuning parameter and can be determined by cross-validation. Compared with the penalized methods, BSA is flexible and easily implemented without the need to apply constrained optimizations, and the parallel computing algorithms can be integrated with separable minority surrogate functions. Further discussion and empirical results are provided in the supporting information.

3. Convergence Properties

We impose the following conditions: (A) For any initial value $\boldsymbol{\theta}^{(0)}$, the matrices, $\mathbf{H}_p(\boldsymbol{\theta})$, $p = 1, \ldots, P$, are positive definite in the super-level set $\{\boldsymbol{\theta} : \ell(\boldsymbol{\theta}) \geqslant \ell(\boldsymbol{\theta}^{(0)})\}$; (B) The negative log-partial likelihood function satisfies $\lim_{||\boldsymbol{\theta}||_2 \to \infty} -\ell(\boldsymbol{\theta}) = \infty$.

Condition (A) guarantees the existence of the BSA update; Condition (B) ensures that the super-level set is compact and the maximum value of $\ell(\boldsymbol{\theta})$ is attained, and a cluster point of BSA exists. We show that there exists a learning rate ν such that the proposed algorithm satisfies the ascent property.

Proposition 1 (Ascent Property)

Suppose Conditions (A) and (B) hold. For $\nu > 0$ satisfying

$$\sup_{\{\boldsymbol{\theta}:\ell(\boldsymbol{\theta})\geqslant\ell(\boldsymbol{\theta}^{(0)})\}} \left(\lambda_{max}\left[\{\mathbf{H}(\widehat{\boldsymbol{\theta}}^{(m-1)})\}^{-1/2}\{-\nabla^2\ell(\boldsymbol{\theta})\}\{\mathbf{H}(\widehat{\boldsymbol{\theta}}^{(m-1)})\}^{-1/2}\right]\right) < 1/\nu, \qquad (11)$$

then $g(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(m-1)}) \leqslant \ell(\boldsymbol{\theta})$ for all $\boldsymbol{\theta}$, where $\lambda_{max}(\cdot)$ represent the largest eigenvalues.

Proposition 1 shows that $g(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(m-1)})$ serves as a minority surrogate function of $\ell(\boldsymbol{\theta})$. Thus, the resulting estimates $\widehat{\boldsymbol{\theta}}^{(m)}$ from the BSA ensure the ascent property,

$$\ell(\widehat{\boldsymbol{\theta}}^{(m)}) \geqslant g(\widehat{\boldsymbol{\theta}}^{(m)}|\widehat{\boldsymbol{\theta}}^{(m-1)}) \geqslant g(\widehat{\boldsymbol{\theta}}^{(m-1)}|\widehat{\boldsymbol{\theta}}^{(m-1)}) = \ell(\widehat{\boldsymbol{\theta}}^{(m-1)}).$$

Proposition 1 also informs the choice of the learning rate ν and gives an upper bound of ν (which is small) to ensure the ascent property. For example, in classical gradient-based procedures, $\mathbf{H}(\widehat{\boldsymbol{\theta}}^{(m-1)})$ equals an identity matrix and the updates at each iteration are computed based on gradient information only. When the conditional number of the observed information matrix is large, a sufficiently small learning rate is needed in Proposition 1 to ensure that the estimates in each iteration of the gradient-based procedure serve as refinements of the previous step, which requires a large number of iterations and more computation time. Thus, empirically we find that the performance of gradient-based methods is more sensitive to the choice of the learning rate. In contrast, the proposed BSA is based on the block diagonal of the observed information matrix, which is an improved approximation compared to the identity matrix used in the gradient methods. Thus, a learning rate of 0.05 typically ensures the inequality in Proposition 1. Our numerical experience also indicates that BSA is less sensitive to the choice of the learning rate.

Proposition 2 (Numerical Convergence)

Suppose Conditions (A) and (B) hold. Then every cluster point of the iterates $\widehat{\boldsymbol{\theta}}^{(m)} = M(\widehat{\boldsymbol{\theta}}^{(m-1)})$ *generated by the iteration map* $M(\boldsymbol{\theta})$ *of the BSA algorithm is a stationary point of* $\ell(\boldsymbol{\theta})$*. Furthermore, the set of stationary points* $\mathcal{F}$ *is closed, and the limit of the distance function is zero:*

$$\lim_{m\to\infty} \inf_{\boldsymbol{\theta}\in\mathcal{F}} ||\widehat{\boldsymbol{\theta}}^{(m)} - \boldsymbol{\theta}||_2 = 0.$$

Moreover, if the observed information matrix $-\nabla^2\ell(\boldsymbol{\theta})$ *is positive definite in the super-level set defined in Condition (A), any sequence of* $\widehat{\boldsymbol{\theta}}^{(m)}$ *possesses a limit,* $\widehat{\boldsymbol{\theta}}$*, and this limit is a stationary point and hence maximizes the log-partial likelihood in* (1)*.*

The convergence mode involved in this proposition is with respect to a sequence of real vectors, and not embedded in a probability space. All technical proofs have been deferred to the supporting information.

4. Simulations

We compare the computational speed and parameter estimation of the proposed BSA with various methods, and then assess the performance of the proposed testing procedure for time-varying effects. Ten knots are used in all settings in Sections 4.1 and 4.3. In section 4.2, we vary the numbers of knots to assess its influence on the performance. Following the suggestion by Gray (1992), the locations of knots in further analyses are chosen to include an equal number of events within each time interval.

4.1 *Evaluation of Computational Speed*

We first consider the following simulation setting (termed Setting A). Death times are generated from an exponential model with a baseline hazard 0.5. Censoring times are generated from the Uniform distribution over (0,3), with a censoring proportion of approximately $20-30\%$. Continuous predictors are generated from a multivariate normal distribution, with mean zero and an AR1 covariance matrix with an auto-correlation parameter of 0.6. We vary the sample size between $N = 10,000$ (from 10 center) and $N = 351,719$ (from 290 centers). The number of covariates varies from $P = 10$ to $P = 164$. We choose $\beta_2(t) = \sin\{3(\pi t/4)\}$ and $\beta_4(t) = -(t/3)^2 \exp(t/2)$ to represent time-varying effects. The remaining covariate coefficients are set to be 1. For each data configuration, 100 data are generated.

With $N = 10,000$ and $P = 10$, the computation time for the Newton method (implemented by R *Survival* package), the quasi-Newton method (implemented in Rcpp through R package *RcppArmadillo* in combination with the R function *optim*), the likelihood-based boosting (implemented by R package $COX_{flex}Boost$) and the proposed BSA is 0.17 minutes, 15.43 minutes, 10.36 hours and 0.12 minutes, respectively. The original quasi-Newton work of He et al. (2017a) was implemented in R, but we re-implement it in Rcpp for improved speed. With $N = 351,719$ and $P = 164$ as in the motivating example, all of the aforementioned competing methods fail due to their intensive computation, and the proposed method takes

11.64 hours. The experiments are conducted on a HP workstation with 4-core 3.50-GHz Intel Core E5-1620v3 processor and 32GB RAM.

4.2 *Estimation of Time-Varying Effects*

To mimic the motivating real data, we consider a simulation setting (termed Setting B) and generate binary covariates (0 or 1) with means between 0.05 and 0.2. The number of covariates varies from 5, 10, 20 to 50, and the sample size is chosen to be $N = 10,000$ from 10 center. The remaining set-ups are the same as Setting A.

Table 1 compares the average computing time, the average biases and the average integrated mean square error (IMSE) over the simulated time points for the Newton approach, the gradient ascent, the stochastic gradient ascent with step size determined by ADAM algorithm (Diederik et al., 2015), and the proposed BSA, under simulation Setting B with $N = 10,000$ and various numbers of covariates.

Table 1 shows that the Newton approach incurs large biases and IMSE; the gradient ascent and the stochastic gradient ascent improve upon the biases, but converge slowly; the proposed BSA is computationally efficient and achieves the smallest biases in all scenarios. Web Figure S2 further compares the average estimated coefficients across various iterations of the proposed method and the gradient ascent, using simulation setting B. Compared with gradient-based procedure, the proposed BSA is less sensitive to the choice of learning rate, which confirms the numerical properties provided in Section 3. Figure 2 compares the average estimates and the 95% empirical percentiles over 100 simulation replications for the conventional Newton approach and the BSA algorithm. We vary the number of basis functions from 5 to 10. The simulation set-up is based on Setting A with 10 variables. The performance of the Newton is more sensitive to the number of basis function, which can be explained in part as follows: in the late stage of the follow-up period, the at-risk set is small, causing unstable estimation of the Hessian matrix. The proposed BSA is less sensitive to the number of basis functions, achieving more stable results. Web Table S1 compares the biases and IMSE for various approaches which select the number of basis functions based on the simulation setting B. Five-fold cross-validation achieves the smallest estimation biases in all scenarios and outperforms alternative approaches such as AIC and BIC. Web Table S2 and Figures S3 and S4 compare the performance of various methods under the simulation setting with heterogeneous center effects. Web Table S2 and Figure S4 further assess a simulation setting with a high censoring proportion (approximately between 50% and 60%).

4.3 *Testing for Time-Varying Effects*

Finally, to assess the testing performance for time-varying effects, we consider a simulation setting (termed Setting C) with two continuous predictors. The corresponding coefficients are set to be $\beta_1 = 1$ and $\beta_2(t) = \gamma \text{sin}\{3(\pi t/4)\}$ with γ varying between 0 and 3, representing the magnitude of the time-varying effects. We vary the number of centers from 10 (with $1,000$ subjects per center) to 100 (with 100 subjects per center). The remaining set-ups are the same as Setting A.

Comparing the proposed testing algorithm with the test based on the scaled Schoenfeld residuals (implemented by R *Survival* package), Figure 3 reports the empirical Type-I error and the empirical power based on Setting C. The proposed algorithm (10) outperforms the Schoenfeld method with a higher power and a smaller Type-I error. Web Table S3 further assesses the empirical Type-I error and the empirical power for the robust inference procedure, using simulation setting C with 100 centers.

5. Analysis of the National Kidney Transplant Dataset

Data are obtained from the U.S. Organ Procurement and Transplantation Network (OPTN). Included in our analysis are $351,719$ patients (from 293 centers) who underwent kidney transplantation between January 1988 and December 2012. Failure time is defined as the time from transplantation to graft failure or death, whichever occurred first. To study the 10-year post-transplant survival, patient survival is censored 10 year post-transplant or at the end of study in 2012. The overall censoring rate was 62%. Covariates ($P = 164$) in this study include baseline recipient characteristics such as age, race, gender, BMI, time on dialysis, indicator of previous kidney transplant, immunosuppression, and cormorbidity conditions (e.g. glomerulonephritis, polycystic kidney disease, diabetes, and hypertension), and donor characteristics such as blood type, cold ischemic time and donor type. Race is categorized as White, African American, Asian, and the other. Cold ischemia time is categorized as low (20 hours or less) and high (longer than 20 hours). Donors are categorized as living, of standard criteria, and of expanded criteria. Waiting time on dialysis is categorized as low (less than 1 years), medium (1-5 years) and high (greater than 5 years). More details are in Table 2.

To determine the number of basis functions, we perform 5-fold cross-validation (Verweij and van Houwelingen, 1993) and choose 10 basis functions for further analysis. Our proposed test identifies a total of 12 variables with significant time-varying effects; see Figure 4 with 95% point-wise confidence intervals (dashed lines) as well as additional results provided in Web Figure S5. Figures 4a and 4b show that anti-viral therapies and anti-rejection immunosuppressant medications have a strong protective effect shortly after transplantation, but the association weakens over time. One possible explanation is that these therapies prevent rejection of new kidneys and declining rates of acute rejection have led to improvements in short term kidney transplant survival, but the effects may wane over time (Muntean and Lucan, 2013). Figure 4c supports the previous findings (Meier et al., 2000) that long waiting on dialysis (greater than 5 years) negatively impacts post-transplant survival. Figure 4d indicates that the effects of stroke, the most frequent donor cause of death, varies over time, showing an increased risk of worsening recipient outcomes initially, followed by a slightly weakening association over time. Though stroke is a predictor for worse survival for kidney transplantation, it is associated with a low rate of rejection immediately after the renal transplantation (Frohnert et al., 1997), which may lead to time-varying associations.

Figure 4e indicates that the survival of African Americans continues to be poorer than that of non-African Americans. The change of its covariate effect after transplant may be partly due to higher immunological risk among American Africans, leading to higher acute rejection rates and graft loss (Harding et al., 20017). We have detected some novel signals. For example, polycystic kidney disease (PKD) is the most common genetic kidney disease and is present among 2% to 9% of ESRD patients (Rozanski et al., 2005). With conflicting reports of renal allograft outcomes for PKD patients (Hadimeri et al., 1997), Figure 4g suggests time-varying associations of PKD with survival; thus, accounting for time-varying effects provides valuable clinical information that could have been missed otherwise. Finally, Web Figure S5c shows that male recipients is a protective factor immediately after the renal transplantation and then has a much worse prognosis than female. One possible explanation is that women have better immunosuppressant compliance than men, and females undergo follow-up visits and show more concern to protect graft function (Puoti et al., 2016).

6. Discussion

Detecting and accounting for time-varying effects are particularly important in the context of clinical studies (Dekker et al., 2008; Yu et al., 2014; Chen et al., 2015; Estes, 2018). However, in survival analysis, the computational burden to model time-varying effects increases quickly as the sample size or the number of predictors grows. We propose a block-wise steepest ascent method, which iteratively updates the optimal block-wise direction along which the directional derivative is maximized and, hence, the approximate increment in log-partial likelihood is greatest. Numerical results show that the proposed algorithm provides sufficient and rapid updates, achieving much computational efficiency.

Acknowledgements

We thank the Editor, the AE and the referees for providing constructive comments that have improved the manuscript. We thank Dr. Kirsten Herold for her helpful suggestions to improve the presentation of the manuscript and thank Dr. Abhijit Naik for clinical discussion. The work is partially supported by grants from NIH.

Data Availability Statement

The data that support the findings in this paper can be accessed through the Organ Procurement and Transplantation Network at https://optn.transplant.hrsa.gov/data/.

References

Boyd, S. and Vandenberghe, L. (2004) Convex Optimization. *Cambridge University Press*, New York.

Bühlmann, P. and Yu, B. (2003) Boosting with the $L2$ loss: regression and classification. *Journal of the American Statistical Association*, 98(462), 324–339.

Bühlmann, P. and Yu, B. (2006) Boosting for high-dimensional linear models. *Annals of Statistics*, 34(2), 559–583.

Chen, X., Sun, J. and Liu, L. (2015) Semiparametric partial Linear quantile regression of longitudinal data with time varying coefficients and informative observation times. *Statistica Sinica*, 25(4), 1437–1458.

Cox, D.R. (1972) Regression models and life tables (with Discussion). *Journal of the Royal Statistical Society, Series B*, 34(2), 187–200.

Dekker, F.W., Mutsert, R., Dijk, P.C., Zoccali, C. and Jager, K.J. (2008) Survival analysis: time-dependent effects and time-varying risk factors. *Kidney International*, 74, 994–997.

Duchi, J., Hazan, E. and Singer, Y. (2011) Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12, 2121–2159.

Diederik, P.K. and Jimmy, L.B. (2015) Adam: a method for stochastic optimization. *International Conference on Learning Representations*, 1–13.

Harding, K., Mersha, T.B., Pham, P.T., Waterman, A.D., Webb, F.A., Vassalotti, J.A. and Nicholas, S.B. (2017) Health Disparities in Kidney Transplantation for African Americans. *American Journal of Nephrology*, 46(2), 165–175.

Estes, J.P., Nguyen, D.V., Chen, Y., Dalrymple, L.S., Rhee, C.M., Kalantar-Zadeh, K. and Sentürk, D. (2018) Time-dynamic profiling with application to hospital readmission among patients on dialysis. *Biometrics*, 74(4), 1383–1394.

Frohnert, P.P., Donadio, J.V.Jr, Velosa, J.A., Holley, K.E. and Sterioff, S. (1997) The fate of renal transplants in patients with IgA nephropathy. *Clinical Transplant*, 11(2), 127–133.

Grambsch, P. and Therneau, T. (1994) Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81(3), 515–526.

Gray, R.J. (1992) Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association*, 87(420), 942–951.

Gray, R.J. (1994) Spline-based tests in survival analysis. *Biometrics*, 50(3), 640–652.

Hadimeri, H., Norden, G., Friman, S. and Nyberg, G. (1997) Autosomal dominant polycystic kidney disease in a kidney transplant population. *Nephrol Dial Transplant*, 12, 1431–1436.

Hastie, T. and Tibshirani, R.J. (1993) Varying-coefficient models. *Journal of the Royal Statistical Society, Series B*, 55(4), 757–796.

He, K. and Schaubel, D.E. (2014) Methods for estimating center effects in survival analysis using direct standardization. *Statistics in Medicine*, 33(12), 2048–2061.

He, K. and Schaubel, D.E. (2015) Standardized mortality ratio for evaluating center-specific mortality: assessment and alternative. *Statistics in Biosciences*, 7, 296–321.

He, K., Yang, Y., Li, Y.M., Zhu, J. and 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), 635–645.

He, K., Li, Y.M., Wei, Q.Y. and Li, Y. (2017) Computationally efficient approach for modeling complex and big survival data. *Big and Complex Data Analysis: Statistical Methodologies and Applications, Edited volume by Springer*, 193–207.

He, K., Li, Y.M., Zhu, J., Liu, H.L., Lee, J.E., Amos, C.I., Hyslop, T., Jin, J.S., Lin, H.Z., Wei, Q.Y. and Li, Y. (2016) Component-wise gradient boosting and false discovery control in survival analysis with high-dimensional covariates. *Bioinformatics*, 32(1), 50–57.

Hofner, B., Hothorn, T. and Kneib, T. (2013) Variable selection and model choice in structured survival models. *Computational Statistics*, 28, 1079–1101.

Honda, T. and Härdle, W.K. (2014) Variable selection in Cox regression models with varying coefficients. *Journal of Statistical Planning and Inference*, 148, 67–81.

Kalantar-Zadeh, K. (2005) Causes and consequences of the reverse epidemiology of body mass index in dialysis patients. *Journal of Renal Nutrition*, 15, 142–147.

Kalbfleisch, J.D. and Wolfe, R.A. (2013) On monitoring outcomes of medical providers. *Statistics in Biosciences*, 2, 286–302.

Lange, K. (2012) Optimization. Second Edition. *Springer*, New York.

Lin, D. and Wei, L. (1989) The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association*, 84(408), 1074–1078.

McLachlan, G.J. and Krishnan, T. (2007) The EM Algorithm and Extensions. Second Edition. *John Wiley and Sons*, New Jersey.

Meier-Kriesche, H.U., Port, F.K., Ojo, A.O., Rudich, S.M., Hanson, J.A., Cibrik, D.M., Leichtman, A.B. and Kaplan, B. (2000) Effect of waiting time on renal transplant outcome. *Kidney International*, 58(3), 1311–1317.

Muntean, A. and Lucan, M. (2013) Immunosuppression in kidney transplantation. *Clujul*

Medical, 86, 177–180.

OPTN. https://optn.transplant.hrsa.gov/data/ (accessed August 2013).

Pan, W. (2002) A note on the use of marginal likelihood and conditional likelihood in analyzing clustered data. *The American Statistician*, 56, 171–174.

Perperoglou, A., le Cessie, S. and van Houwelingen, H.C. (2006) A fast routine for fitting Cox models with time varying effects of the covariates. *Computer Methods and Programs in Biomedicine*, 25, 154–161.

Puoti, F., Ricci, A., Nanni-Costa, A., Ricciardi, W., Malorni, W. and Ortona, E. (2016) Organ transplantation and gender differences: a paradigmatic example of intertwining between biological and sociocultural determinants. *Biology of Sex Differences*, 7(35).

Qian, N. (1999) On the momentum term in gradient descent learning algorithms. *Neural Networks*, 12(1), 145–151.

Robbins, H. and Monro, S. (1951) A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3), 400–407.

Rozanski, J., Kozlowska, I. and Myslak, M. (2005) Pretransplant nephrectomy in patients with autosomal dominant polycystic kidney disease. *Transplant Proc*, 37, 666–668.

Ruder, S. (2016) An overview of gradient descent optimization algorithm. *arXiv preprint*

Saran, R., Robinson, B. and Abbott, K.C. et al. (2018) US Renal Data System 2017 Annual Data Report: Epidemiology of Kidney Disease in the United States. *American Journal of Kidney Diseases*, 71(3), S1–S672.

Schaubel, D.E. and Cai, J. (2005) Semiparametric methods for clustered recurrent event data. *Lifetime Data Analysis*, 11(3), 405–425.

Snyder, J.J., Salkowski, N., Kim, S.J., Zaun, D., Xiong, H., Israni, A.K. and Kasiske, B.L. (2016) Developing statistical models to assess transplant outcomes using national registries: the process in the United States. *Transplantation*, 100(2), 288–294.

Tian, L., Zucker, D. and Wei, L. (2005) On the Cox model with time-varying regression coefficients. *Journal of the American Statistical Association*, 100(469), 72–183.

Verweij, P.J.M. and van Houwelingen, H.C. (1993) Cross-validation in survival analysis. *Statistics in Medicine*, 12(24), 2305–2314.

Wolfe, R.A., Ashby, V.B., Milford, E.L., Ojo, A.O., Ettenger, R.E., Agodoa, L.Y., Held, P.J. and Port, F.K. (1999) Comparison of mortality in all patients on dialysis, patients on dialysis awaiting transplantation and recipients of a first cadaveric transplant. *New England Journal of Medicine*, 341(23), 1725–1730.

Journal of the American Statistical Association, 106(493), 296–305.

Xiao, W., Lu, W. and Zhang, H.H. (2016) Joint structure selection and estimation in the time-varying coefficient Cox model. *Statistica Sinica*, 26(2), 547–567.

Yan, J. and Huang, J. (2012) Model selection for Cox models with time-varying coefficients. *Biometrics*, 68(2), 419–428.

Yu, Z., Liu, L., Bravata, D.M. and Williams, L.S. (2014) Joint model of recurrent events and a terminal event with time-varying coefficients. *Biometrical Journal*, 56, 183–197.

Zeiler, M.D. (2012) ADADELTA: An adaptive learning rate method. *arXiv preprint*. arXiv:1212.5701.

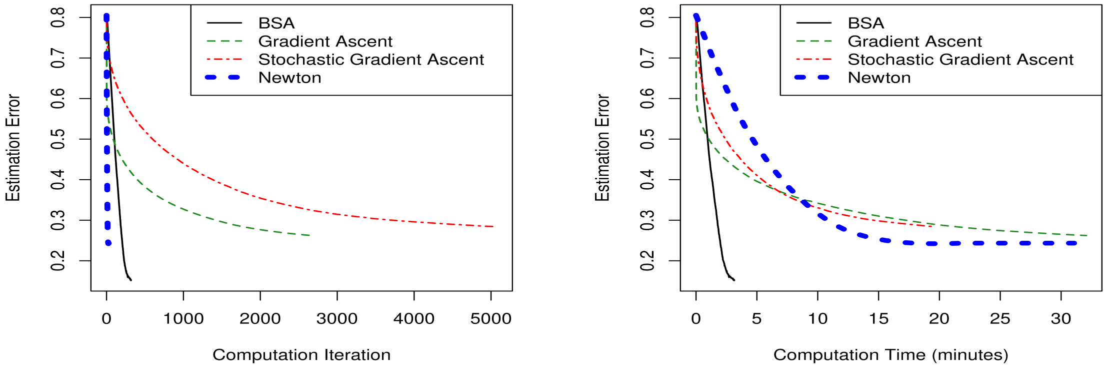
Zou, H. (2006) The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476), 1418–1429.

Zhang, H. and Lu, W. (2006) Adaptive lasso for Cox's proportional hazards model. *Biometrika*, 94(3), 691–703.

Zucker, D.M. and Karr, A.F. (1990) Nonparametric survival analysis with time-dependent covariate effects: a penalized partial likelihood approach. *Annals of Statistics*, 18(1), 329–353.

Supporting Information

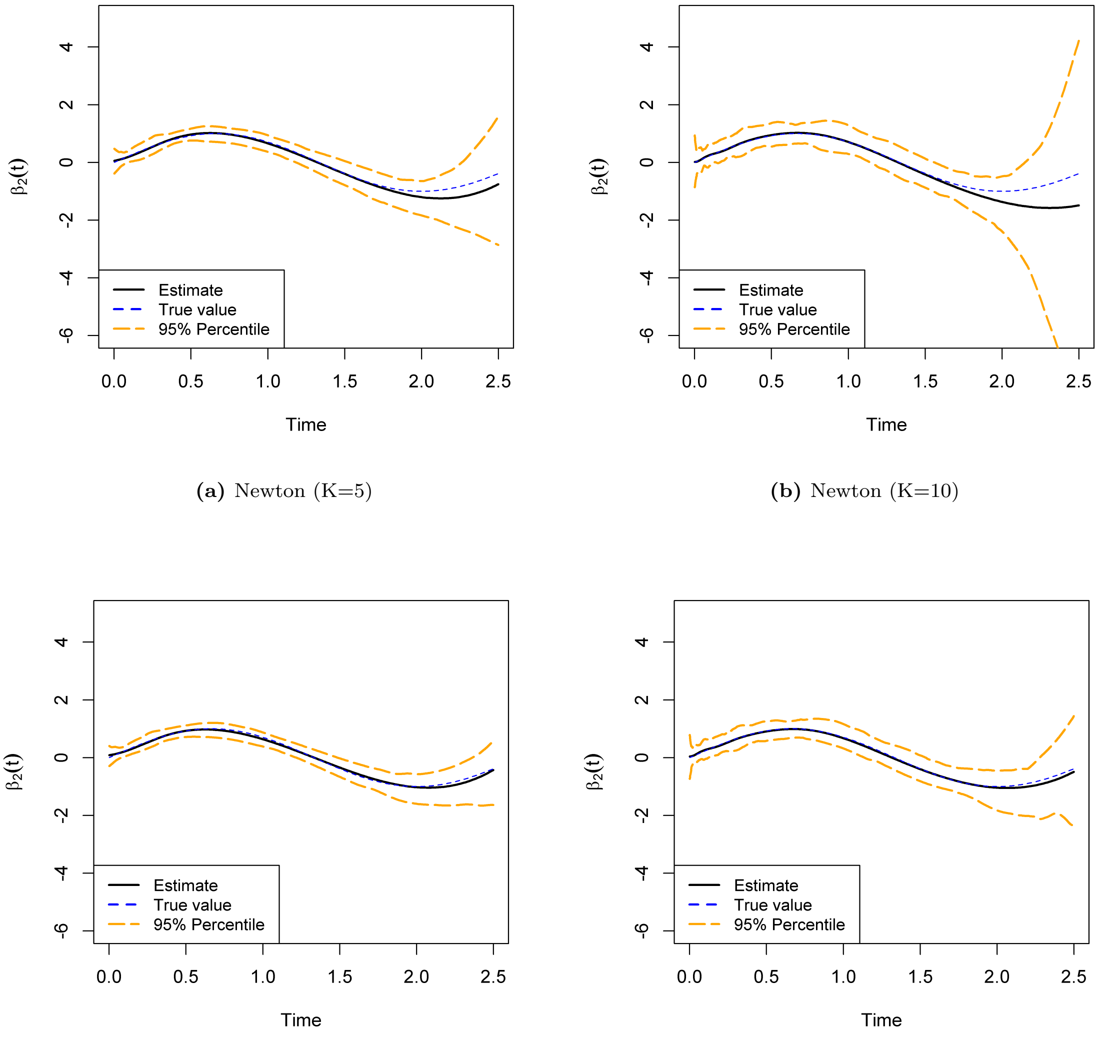
Web Appendices, Tables, and Figures referenced in Sections 3, 4.2, 4.3 and 5, along with example R codes, are available with this paper at the Biometrics website on Wiley Online Library. We also provide a publicly available R package *BSATV*, hosted on the *GitHub* (https://github.com/UM-KevinHe/TimeVaryingCox).



(a) Computation Iterations

(b) Computation Time

Figure 1: Comparisons of iterations, computation time and biases; Setting B with N=10,000 and P=5; the Newton approach is implemented by R *Survival* package, the stochastic gradient ascent is implemented by the adaptive moment estimation (ADAM) approach. The timings were taken on a HP workstation with 4-core 3.50-GHz Intel Core E5-1620v3 processor and 32GB RAM. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.



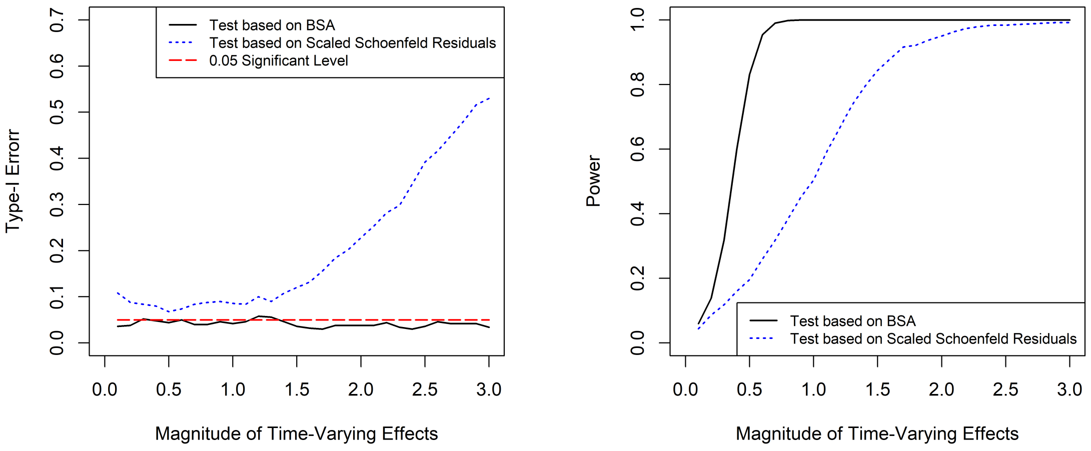
(a) Newton (K=5)

(b) Newton (K=10)

(c) BSA (K=5)

(d) BSA (K=10)

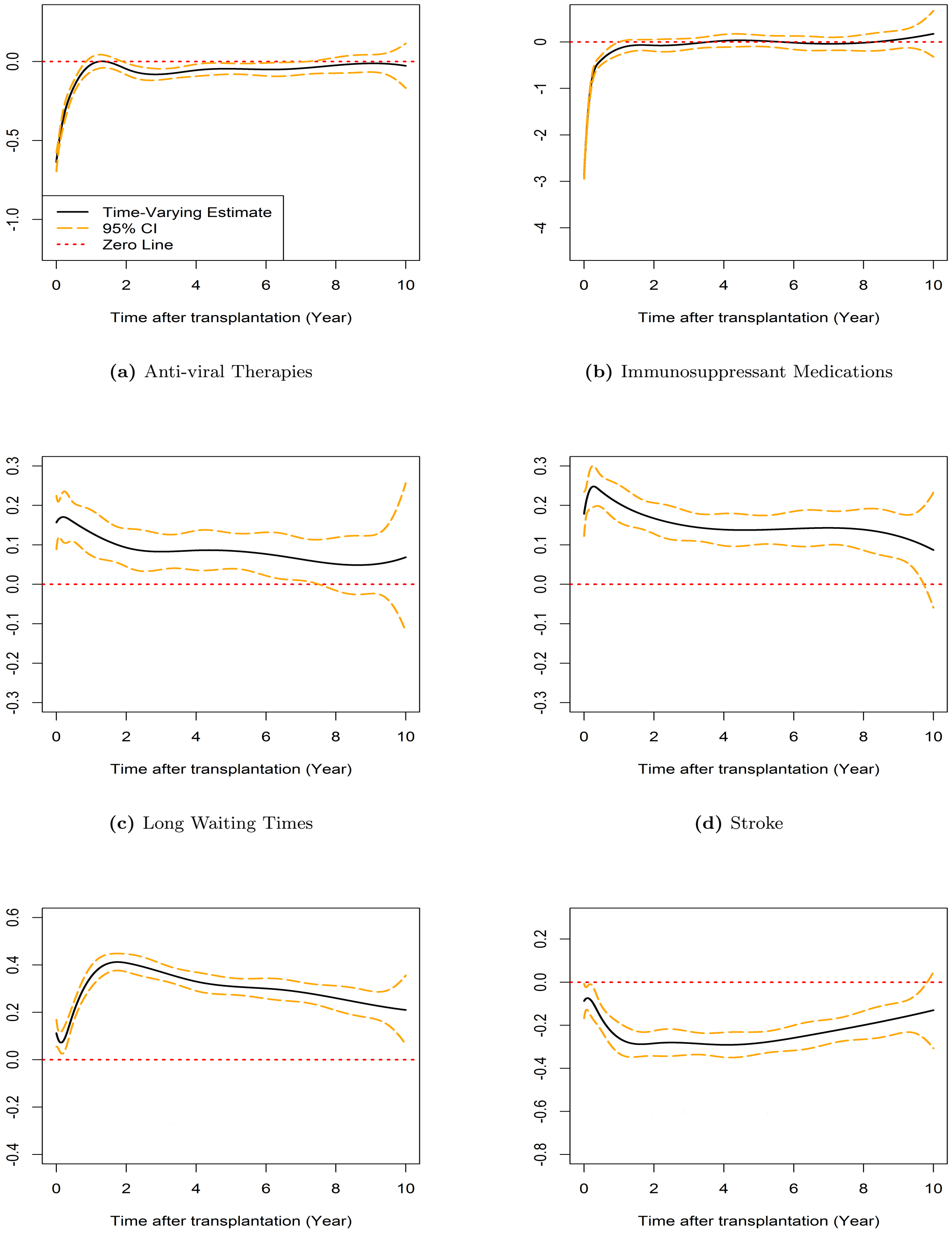
Figure 2: Average estimated coefficient functions (solid lines) and 95% empirical percentiles (dashed lines) for different number of spline basis functions; 100 simulation iterations; Setting A with N=10,000 and P=10. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.



(a) Type I Error

(b) Power

Figure 3: Comparisons of Type-I error and Power for testing of time-varying effects at significance level 0.05; Setting C with N=2,000; Two continuous covariates are generated with coefficients $\beta_1 = 1$ and $\beta_2(t) = \gamma\text{sin}\{3(\pi t/4)\}$, where γ varies between 0 and 3, representing the magnitude of the time-varying effects; The average type-I error rate is only evaluated for the time-invariant $\beta_1(t)$, and the average power is only evaluated for the time-variant $\beta_2(t)$. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.



(a) Anti-viral Therapies

(b) Immunosuppressant Medications

(c) Long Waiting Times

(d) Stroke

(e) African American

(f) Polycystic Kidney Disease

Figure 4: Data analysis results: estimated coefficient functions (solid lines) and 95% pointwise confidence interval (dashed lines) for time-varying effects. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

P	Method	Time	Bias	IMSE
5	Newton	0.22	0.646	0.592
	Gradient Ascent	169.33	0.249	0.202
	Stochastic Gradient Ascent	183.51	0.290	0.256
	BSA	25.60	0.156	0.136
20	Newton	1.10	0.305	0.169
	Gradient Ascent	687.15	0.136	0.058
	Stochastic Gradient Ascent	415.43	0.140	0.070
	BSA	43.48	0.075	0.055
50	Newton	9.05	0.147	0.086
	Gradient Ascent	1620.21	0.150	0.050
	Stochastic Gradient Ascent	757.07	0.118	0.038
	BSA	95.27	0.064	0.030

Table 1: Average computation time (in seconds), average estimation error (Bias) and average integrated mean square error (IMSE) for various methods; based on Setting B with N=10,000.

Variable	Categories	Counts	Proportions
Donor Type	Deceased	229,465	65.2%
	Living	122,254	34.8%
Recipient Gender	Male	211,880	60.2%
	Female	139,839	39.8%
Recipient Race	White	248,254	70.6%
	Black	82,816	23.5%
	Asian	15,347	4.4%
	Other	5,302	1.5%
Recipient BMI	Underweight	16,866	4.8%
	Normal	109,385	31.1%
	Overweight	152,765	43.4%
	Obesity	72,703	20.7%
Recipient Age	< 10 years	6,596	1.9%
	$[10, 18)$ years	12,405	3.5%
	$[18, 25)$ years	18,059	5.1%
	$[25, 35)$ years	47,894	13.6%
	$[35, 45)$ years	68,963	19.6%
	$[45, 55)$ years	84,151	23.9%
	$[55, 65)$ years	76,081	21.6%
	$[65, 75)$ years	34,281	9.7%
	>= 75 years	3,289	0.9%
Recipient Anti-viral Therapies	Yes	169,037	48.1%
	No	182,682	51.9%
Recipient Immunosuppressant Medications	Yes	341,677	97.1%
	No	10,042	2.9%
Recipient: Polycystic Kidney Disease	Yes	31,558	9.0%
	No	320,161	91.0%
Waiting Time on Dialysis	Short (< 1 years)	135,585	38.5%
	Medium ($1 - 5$ years)	165,012	46.9%
	Long (> 5 years)	51,122	14.5%
Cold Ischemia Time	High (> 20 hours)	91,861	26.1%
	Low (<= 20 hours)	259,858	73.9%
Expanded Criteria Donor	Yes	31,126	8.8%
	No	320,593	91.2%
Donor Cause of Death: Stroke	Yes	82,474	23.4%
	No	269,245	76.6%

Table 2: Baseline characteristics of kidney transplantation data.