Published in final edited form as:
Lifetime Data Anal. 2022 April ; 28(2): 194–218. doi:10.1007/s10985-021-09544-2.

Scalable Proximal Methods for Cause-Specific Hazard Modeling with Time-Varying Coefficients

Wenbo Wu,
Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109-2029, USA

Jeremy M. G. Taylor,
Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109-2029, USA

Andrew F. Brouwer,
Department of Epidemiology, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109-2029, USA

Lingfeng Luo,
Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109-2029, USA

Jian Kang,
Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109-2029, USA

Hui Jiang,
Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109-2029, USA

Kevin He[*]
Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109-2029, USA

Abstract

Survival modeling with time-varying coefficients has proven useful in analyzing time-to-event data with one or more distinct failure types. When studying the cause-specific etiology of breast and prostate cancers using the large-scale data from the Surveillance, Epidemiology, and End Results (SEER) Program, we encountered two major challenges that existing methods for estimating time-varying coefficients cannot tackle. First, these methods, dependent on expanding the original data in a repeated measurement format, result in formidable time and memory consumption as the sample size escalates to over one million. In this case, even a well-configured workstation cannot accommodate their implementations. Second, when the large-scale data under analysis include binary predictors with near-zero variance (e.g., only 0.6% of patients in our

[*] corresponding author Kevin He, kevinhe@umich.edu.

Conflict of interest
The authors have no conflicts of interest to declare that are relevant to the content of this article.

SEER prostate cancer data had tumors regional to the lymph nodes), existing methods suffer from numerical instability due to ill-conditioned second-order information. The estimation accuracy deteriorates further with multiple competing risks. To address these issues, we propose a proximal Newton algorithm with a shared-memory parallelization scheme and tests of significance and nonproportionality for the time-varying effects. A simulation study shows that our scalable approach reduces the time and memory costs by orders of magnitude and enjoys improved estimation accuracy compared with alternative approaches. Applications to the SEER cancer data demonstrate the real-world performance of the proximal Newton algorithm.

Keywords

Kronecker product; B-spline; Proximal algorithm; Parallel computing; Breast cancer; Prostate cancer

1 Introduction

The temporal variation in the effects of interventions or risk factors is a common phenomenon in time-to-event data (Wolfe et al., 1999; Thior et al., 2006; Dekker et al., 2008). To allow the effects to vary with time when analyzing the data, an important extension of the Cox model is often used—the relative risk model with time-varying coefficients. As remarked in Kalbfleisch and Prentice (2002, §4.1), this extended model is not only instrumental for testing the proportional hazards relationship, but also allows a concise description of a useful class of covariate effects. When the event of interest involves several distinct types, the time-varying effects can be similarly incorporated into a competing risk framework.

Our endeavors here were motivated by studying the cause-specific etiology of breast and prostate cancers using data from the National Cancer Institute Surveillance, Epidemiology, and End Results (SEER) Program. Different from most analyses assuming constant effects of prognostic factors for survival, our purpose was to account for how the effects change with time. Early evidence from breast cancer patients (Bellera et al., 2010; Baulies et al., 2015) suggested that tumor grade had a significant time-varying effect. As a more recent example, Brouwer et al. (2020) studied the cause-specific survival of patients diagnosed with squamous cell carcinomas (head and neck cancers) and found that the effects of age and sex were strongest at the time of diagnosis, but attenuated dramatically over the first few years. Ignoring the dynamic nature of a time-varying effect may weaken the internal validity of the study and cloud its implications for risk prediction, treatment development, and health care policy.

Along with the rising need for time-varying effect modeling in a cause-specific context, the growing volume and complexity of data pose overarching challenges to existing analytic frameworks. To name a few examples, Zucker and Karr (1990) established a nonparametric penalized partial likelihood approach, which was revisited in Hastie and Tibshirani (1993) with a cubic spline penalty. Gray (1992, 1994) instead used the cubic B-spline bases (de Boor, 2001) with a small number of knots to parameterize the penalty function. Alternatively, Verweij and van Houwelingen (1995) and Tutz and Binder (2004) adopted as penalty the sum of squared pairwise differences of effect estimates at adjacent time points. In terms of implementation, these methods expand the original data in a repeated measurement format (Therneau et al., 2020) using existing software such as the survival package (Therneau, 2020), and perform well when the input data set is relatively small. As the data under analysis escalate in size, however, fitting a cause-specific hazard model with time-varying coefficients becomes formidably time-consuming and memory inefficient.

To illustrate this issue, we benchmarked the cause-specific hazard model fitting to simulated data sets (details in Section 5) using the function coxph of survival, called hereafter the Naive Newton (NaiveN) method. As shown in Figure 1, increasing the number of observations from 1,000 to 10,000 leads to substantial growth in the runtime and memory usage of NaiveN, whereas the runtime and memory consumption of our proposed algorithm slightly increase. If the sample size is further scaled up to over 100,000, as in Brouwer et al. (2020), even a well-configured workstation with 500 GB of RAM can barely accommodate the execution. The SEER breast cancer data we used consist of over 1 million patients, rendering any data-expansion-based method infeasible.

In the literature, some analyses have attempted to address this computational challenge. Inspired by a Kronecker product-based routine of Perperoglou et al. (2006), He et al. (2017) and He et al. (2021) respectively considered the Quasi-Newton (QuasiN) and minorize-maximization-based steepest ascent (MMSA) methods. Taking advantage of the large number of small strata in their settings, both methods demonstrated improved computation compared with the NaiveN, but were unable to handle an unstratified risk set with over one million subjects as in our cancer applications. Since gradient-based methods such as the MMSA only utilize first-order information, they often lead to appreciably more iterations than Newton-type methods. As will be seen in Table 1, the QuasiN may also produce highly biased estimates due to poor Hessian matrix approximation.

In addition to the computational burden, numerical instability often arises from ill-conditioned second-order information in large-scale cause-specific hazard modeling. Specifically, when the data under analysis include a number of binary covariates with near-zero variation (e.g., in the SEER prostate cancer data, only 0.6% of the 716,553 patients had their tumors regional to the lymph nodes), the associated observed information matrix of a Newton-type method may have its minimum eigenvalue close to zero with a large condition number. Inverting such a nearly singular matrix is numerically unstable and the corresponding Newton updates are likely to be confined within a small neighborhood of the initial value, causing the estimates to be far from the optimal solutions. When multiple failure types are present, the issue of inaccurate estimation can be further exacerbated using existing methods (Section 5.1).

To achieve computational efficiency and reduce numerical instability, we propose a spline-based Newton-type method, which we term the proximal Newton (ProxiN) algorithm. This algorithm originates from the so-called proximal algorithms (Parikh and Boyd, 2014), and bears some resemblance to the more generic proximal Newton-type methods (Lee et al., 2012, 2014). Compared with the data-expansion-based NaiveN, the ProxiN reduces the execution time and memory consumption by orders of magnitude. As shown in Figure 1, the runtime and memory curves of the ProxiN stand in sharp contrast with those of the NaiveN and demonstrate the superiority of our proposed approach. Moreover, a shared-memory parallelization scheme further adds to the computational efficiency of the ProxiN with mild hardware requirements. As will be seen in Section 5.1, the ProxiN also leads to improved estimation accuracy compared to the NaiveN and QuasiN methods. The R and C++ code implementing the ProxiN and the parallelization scheme is available online at https://github.com/UM-KevinHe/surtiver.

The rest of this article proceeds as follows: Section 2 lays out a cause-specific hazard model with time-varying coefficients. Section 3 presents the proximal Newton algorithm, its convergence properties, and the parallelization scheme. Section 4 introduces testing procedures. Simulation results are discussed in Section 5. In Section 6, the proposed method is applied to two large-scale cancer databases of SEER. Section 7 concludes with a discussion.

2 Model

For the ith subject, $i = 1, \dots, n$, let T_i, C_i and $X_i := T_i \wedge C_i$ denote the failure, censoring and observed time, respectively, where n denotes the total number of subjects and $a \wedge b :=$ min{a, b}. Let $\mathbf{Z}_i := (Z_{i1}, \dots, Z_{ip})^\mathrm{T}$ denote a vector of p covariates for risk adjustment. Let J_i be a random variable such that $J_i = j$ if subject i has a failure of type j, $j = 1, \dots, m$, and $J_i = 0$ if subject i has a censoring event. Let $\Delta_{ij} := I(T_i \le C_i, J_i = j)$ be an indicator of type j failure, where $I(\cdot)$ is an indicator function. We assume that conditional on $\mathbf{Z}_i$, T_i is independently censored by C_i.

To model competing risks, we consider a Cox relative risk model

$$\lambda_j(t \mid \mathbf{Z}_i) := \lambda_{0j}(t) \exp[\mathbf{Z}_i^\mathrm{T} \boldsymbol{\beta}_j(t)], \quad j = 1, \dots, m, \tag{1}$$

where for failure type j, $\lambda_j(t \mid \mathbf{Z}_i)$ denotes the cause-specific hazard function, $\lambda_{0j}(t)$ denotes the baseline hazard, and $\boldsymbol{\beta}_j(t) := [\beta_{j1}(t), \dots, \beta_{jp}(t)]^\mathrm{T}$ is a p-dimensional vector of potentially time-varying coefficients. To estimate $\boldsymbol{\beta}_j(t)$ at time t, we span $\boldsymbol{\beta}_j(\cdot)$ by a set of K B-spline basis functions. Specifically, for $l = 1, \dots, p$, $\beta_{jl}(\cdot)$ is formulated as a linear combination

$$\beta_{jl}(t) := \boldsymbol{\gamma}_{jl}^\mathrm{T} \mathbf{B}(t) = \sum_{k=1}^{K} B_k(t) \gamma_{jlk}, \tag{2}$$

where $\mathbf{B}(t) := [B_1(t), \dots, B_K(t)]^\mathrm{T}$ forms a basis, and $\boldsymbol{\gamma}_{jl} := [\gamma_{jl1}, \dots, \gamma_{jlK}]^\mathrm{T}$ is a vector of K unknown parameters for the lth time-varying coefficient $\beta_{jl}(\cdot)$ of failure type j. The time points at which pieces of B-spline polynomials meet are called knots and may be chosen based on the quantiles of the failure time points (Gray, 1992; He et al., 2017, 2021). For ease of notation, we only consider a fixed number of K basis functions across different time-varying effects $\beta_{jl}(t)$; the general case of a varying number of basis functions is discussed in Section 5. Letting $\Gamma_j := [\boldsymbol{\gamma}_{j1}, \dots, \boldsymbol{\gamma}_{jp}]^\mathrm{T}$, we define $\boldsymbol{\gamma}_j := \mathrm{vec}(\boldsymbol{\Gamma}_j^\mathrm{T})$, a vectorization of $\boldsymbol{\Gamma}_j^\mathrm{T}$, by stacking its columns on top of each other, and $\boldsymbol{\gamma} := [\boldsymbol{\gamma}_1^\mathrm{T}, \dots, \boldsymbol{\gamma}_m^\mathrm{T}]^\mathrm{T}$. Then model (1) leads to a log-partial likelihood given by

$$\ell(\boldsymbol{\gamma}) = \sum_{j=1}^{m} \ell_j(\boldsymbol{\gamma}_j), \tag{3}$$

in which

$$\begin{aligned}\ell_j(\boldsymbol{\gamma}_j) &:= \frac{1}{n}\sum_{i=1}^{n} \Delta_{ij}\left[\mathbf{Z}_i^{\mathsf{T}}\boldsymbol{\Gamma}_j\mathbf{B}(X_i) - \log\left\{\sum_{r \in R(X_i)} \exp\left(\mathbf{Z}_r^{\mathsf{T}}\boldsymbol{\Gamma}_j\mathbf{B}(X_i)\right)\right\}\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} \Delta_{ij}\left[\mathbf{D}_i^{\mathsf{T}}(X_i)\boldsymbol{\gamma}_j - \log\left\{\sum_{r \in R(X_i)} \exp\left(\mathbf{D}_r^{\mathsf{T}}(X_i)\gamma_j\right)\right\}\right],\end{aligned} \tag{4}$$

where $R(X_i) := \{r \in \{1, \ldots, n\} : X_r \geq X_i\}$ denotes the risk set of subject i, $\mathbf{D}_r(X_i) := \mathbf{Z}_r \otimes \mathbf{B}(X_i)$, and $\otimes$ denotes the Kronecker product.

Observe that $\ell(\boldsymbol{\gamma})$ is twice continuously differentiable and concave since a log-sum-exp function is convex (Boyd and Vandenberghe, 2004, §3.1.5, pp.74). In addition, $\ell(\boldsymbol{\gamma})$ can be optimized by maximizing each $\ell_j(\boldsymbol{\gamma}_j)$ separately with respect to $\boldsymbol{\gamma}_j$. The gradient $\nabla\ell_j(\boldsymbol{\gamma}_j)$ and Hessian matrix $\nabla^2\ell_j(\boldsymbol{\gamma}_j)$ of $\ell_j(\boldsymbol{\gamma}_j)$ are available in the Appendix.

3 Estimation

3.1 Proximal Newton algorithm

As discussed in Section 1, the classical Newton-type methods tend to provide unstable estimation, especially when the information matrix is nearly singular. Our proposed solution to this numerical instability has its roots in the proximal algorithm. For completeness, we start by reviewing this technique as well as its affinity to the traditional Newton approach. Interested readers are referred to Parikh and Boyd (2014) for a detailed account.

Let $\ell : \mathbb{R}^d \to \mathbb{R}$ be a closed and concave function; that is, its hypograph $\mathrm{hyp}(\ell) := \{(\boldsymbol{\gamma}, s) \in \mathbb{R}^{d+1} : \ell(\boldsymbol{\gamma}) \geq s\}$ is a nonempty closed convex set. For any $\lambda > 0$, a proximal operator of $\lambda\ell$ denoted as $\mathbf{prox}_{\lambda\ell}$ is defined as

$$\mathbb{R}^d \ni \mathbf{v} \to \mathbf{prox}_{\lambda\ell}(\mathbf{v}) := \underset{\boldsymbol{\gamma}}{\operatorname{argmax}}\,\{\ell(\boldsymbol{\gamma}) - \|\boldsymbol{\gamma} - \mathbf{v}\|_2^2 / (2\lambda)\} \in \mathbb{R}^d, \tag{5}$$

where $\|\cdot\|_2$ denotes the Euclidean norm for vectors, or the induced L_2 norm for matrices. The use of argmax is justified by Proposition 1 in the Supplementary Information.

To reveal the connection between the proximal operator (5) and Newton approach, note that if ℓ is twice continuously differentiable, its second-order Taylor approximation $\hat{\ell}_{\mathbf{v}}(\boldsymbol{\gamma})$ at $\mathbf{v}$ is $\hat{\ell}_{\mathbf{v}}(\boldsymbol{\gamma}) := \ell(\mathbf{v}) + \nabla\ell^{\mathsf{T}}(\mathbf{v})(\boldsymbol{\gamma} - \mathbf{v}) + (\boldsymbol{\gamma} - \mathbf{v})^{\mathsf{T}}\nabla^2\ell(\mathbf{v})(\boldsymbol{\gamma} - \mathbf{v}) / 2$. To derive the proximal operator of $\lambda\hat{\ell}_{\mathbf{v}}(\boldsymbol{\gamma})$, observe that the corresponding maximand is

$$\ell(\mathbf{v}) + \nabla\ell^{\mathsf{T}}(\mathbf{v})(\boldsymbol{\gamma} - \mathbf{v}) + (\boldsymbol{\gamma} - \mathbf{v})^{\mathsf{T}}\left(\nabla^2\ell(\mathbf{v}) - \mathbf{I}/\lambda\right)(\boldsymbol{\gamma} - \mathbf{v}) / 2,$$

where $\nabla^2\ell(\mathbf{v}) - \mathbf{I}/\lambda$ is a negative definite matrix with $\mathbf{I}$ being a $d \times d$ identity matrix. Maximizing the above quadratic maximand yields

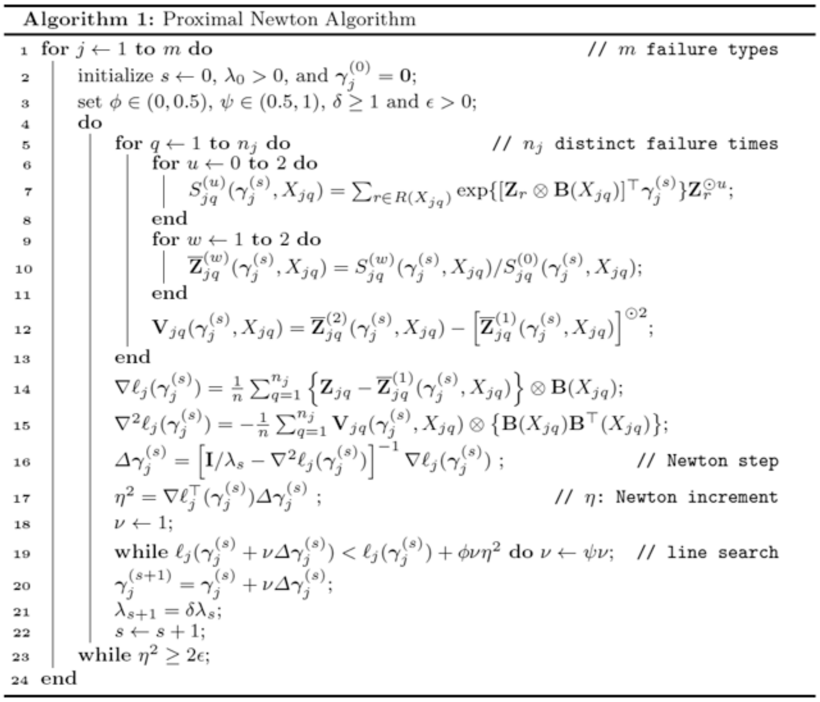
$$\mathbf{prox}_{\lambda\hat{\ell}_{\mathbf{v}}}(\mathbf{v}) = \mathbf{v} + \left(\mathbf{I} / \lambda - \nabla^2\ell(\mathbf{v})\right)^{-1}\nabla\ell(\mathbf{v}), \tag{6}$$

which is a Levenberg–Marquardt step (Levenberg, 1944; Marquardt, 1963), or a Newton step with a modified Hessian matrix (Nocedal and Wright, 2006).

As noted in Section 2, the log-partial likelihood $\ell(\boldsymbol{\gamma})$ in (3) is twice continuously differentiable and concave. Since a function is upper semi-continuous if and only if its hypograph is closed (Rockafellar, 1970, Theorem 7.1), (3) is also a closed function. Applying (6) to the second-order Taylor approximation of (3), we have the proximal Newton algorithm sketched as Algorithm 1, where $X_{j1} < \cdots < X_{jn_j}$ denote the n_j distinct times of type j failures, $j = 1, \ldots, m$, and $\mathbf{Z}_{jq}$ denotes the vector $\mathbf{Z}_i$ such that $\Delta_{ij} = 1$ and $X_i = X_{jq}$, $q = 1, \ldots, n_j$.

3.2 Convergence of the proximal Newton algorithm

The proposed proximal Newton algorithm, as a likelihood maximization approach, includes particular features to ensure convergence in most practical settings. First, the Newton step $\Delta\boldsymbol{\gamma}_j^{(s)}$ on Line 16 of Algorithm 1 is an ascent direction of $\ell_j(\boldsymbol{\gamma}_j^{(s)})$ at $\boldsymbol{\gamma}_j^{(s)}$, which is defined as follows:



```
Algorithm 1: Proximal Newton Algorithm
1  for $j \leftarrow 1$ to $m$ do                                              // $m$ failure types
2      initialize $s \leftarrow 0$, $\lambda_0 > 0$, and $\gamma_j^{(0)} = \mathbf{0}$;
3      set $\phi \in (0, 0.5)$, $\psi \in (0.5, 1)$, $\delta \geq 1$ and $\epsilon > 0$;
4      do
5          for $q \leftarrow 1$ to $n_j$ do                                   // $n_j$ distinct failure times
6              for $u \leftarrow 0$ to $2$ do
7                  $S_{jq}^{(u)}(\gamma_j^{(s)}, X_{jq}) = \sum_{r \in R(X_{jq})} \exp\{[\mathbf{Z}_r \otimes \mathbf{B}(X_{jq})]^\top \gamma_j^{(s)}\}\mathbf{Z}_r^{\odot u}$;
8              end
9              for $w \leftarrow 1$ to $2$ do
10                 $\overline{\mathbf{Z}}_{jq}^{(w)}(\gamma_j^{(s)}, X_{jq}) = S_{jq}^{(w)}(\gamma_j^{(s)}, X_{jq}) / S_{jq}^{(0)}(\gamma_j^{(s)}, X_{jq})$;
11             end
12             $\mathbf{V}_{jq}(\gamma_j^{(s)}, X_{jq}) = \overline{\mathbf{Z}}_{jq}^{(2)}(\gamma_j^{(s)}, X_{jq}) - \left[\overline{\mathbf{Z}}_{jq}^{(1)}(\gamma_j^{(s)}, X_{jq})\right]^{\odot 2}$;
13         end
14         $\nabla\ell_j(\gamma_j^{(s)}) = \frac{1}{n}\sum_{q=1}^{n_j}\left\{\mathbf{Z}_{jq} - \overline{\mathbf{Z}}_{jq}^{(1)}(\gamma_j^{(s)}, X_{jq})\right\} \otimes \mathbf{B}(X_{jq})$;
15         $\nabla^2\ell_j(\gamma_j^{(s)}) = -\frac{1}{n}\sum_{q=1}^{n_j}\mathbf{V}_{jq}(\gamma_j^{(s)}, X_{jq}) \otimes \{\mathbf{B}(X_{jq})\mathbf{B}^\top(X_{jq})\}$;
16         $\Delta\gamma_j^{(s)} = \left[\mathbf{I}/\lambda_s - \nabla^2\ell_j(\gamma_j^{(s)})\right]^{-1}\nabla\ell_j(\gamma_j^{(s)})$ ;          // Newton step
17         $\eta^2 = \nabla\ell_j^\top(\gamma_j^{(s)})\Delta\gamma_j^{(s)}$ ;                      // $\eta$: Newton increment
18         $\nu \leftarrow 1$;
19         while $\ell_j(\gamma_j^{(s)} + \nu\Delta\gamma_j^{(s)}) < \ell_j(\gamma_j^{(s)}) + \phi\nu\eta^2$ do $\nu \leftarrow \psi\nu$;   // line search
20         $\gamma_j^{(s+1)} = \gamma_j^{(s)} + \nu\Delta\gamma_j^{(s)}$;
21         $\lambda_{s+1} = \delta\lambda_s$;
22         $s \leftarrow s + 1$;
23     while $\eta^2 \geq 2\epsilon$;
24 end
```

Definition 1 A direction $\boldsymbol{\mu} \in \mathbb{R}^d$ is an ascent direction of a function $\ell : \mathbb{R}^d \to \mathbb{R}$ at a point $\boldsymbol{\gamma} \in \mathbb{R}^d$ if $\exists\, \bar{\nu} > 0$ such that $\forall\, \nu \in (0, \bar{\nu}]$, $\ell(\boldsymbol{\gamma} + \nu\boldsymbol{\mu}) > \ell(\boldsymbol{\gamma})$.

Using the concept of directional derivative, Definition 1 implies that $\boldsymbol{\mu} \in \mathbb{R}^d$ is an ascent direction of a differentiable function ℓ at $\boldsymbol{\gamma}$ if

$$\lim_{\nu \to 0} \frac{\ell(\boldsymbol{\gamma} + \nu\boldsymbol{\mu}) - \ell(\boldsymbol{\gamma})}{\nu} = \nabla \ell^{\top}(\boldsymbol{\gamma})\boldsymbol{\mu} > 0 .$$

An equivalent condition is provided in the following Lemma 1, which shows that $\Delta\boldsymbol{\gamma}_j^{(s)}$ on Line 16 is an ascent direction of $\ell_j(\boldsymbol{\gamma}_j^{(s)})$ at $\boldsymbol{\gamma}_j^{(s)}$ $(\mathbf{I} / \lambda_s - \nabla^2 \ell_j(\boldsymbol{\gamma}_j^{(s)})$ is positive definite for any $\lambda_s > 0)$. The proof of Lemma 1 is available in the Supplementary Information.

Lemma 1 *Let* $\ell : \mathbb{R}^d \to \mathbb{R}$ *be a differentiable function. Then a direction* $\boldsymbol{\mu} \in \mathbb{R}^d$ *satisfies* $\nabla \ell^{\top}(\boldsymbol{\gamma})\boldsymbol{\mu} > 0$ *at* $\boldsymbol{\gamma}$ *if and only if there exists a symmetric and positive definite matrix* $\mathbf{M}$ *such that* $\boldsymbol{\mu} = \mathbf{M}^{-1}\nabla \ell(\boldsymbol{\gamma})$.

Second, the backtracking line search on Line 19 of Algorithm 1 constitutes a practical implementation of the Armijo–Goldstein conditions

$$\ell_j(\boldsymbol{\gamma}_j^{(s)} + \nu\Delta\boldsymbol{\gamma}_j^{(s)}) \geq \ell_j(\boldsymbol{\gamma}_j^{(s)}) + \phi\nu\nabla \ell_j^{\top}(\boldsymbol{\gamma}_j^{(s)})\Delta\boldsymbol{\gamma}_j^{(s)}, \tag{7}$$

$$\ell_j(\boldsymbol{\gamma}_j^{(s)} + \nu\Delta\boldsymbol{\gamma}_j^{(s)}) \leq \ell_j(\boldsymbol{\gamma}_j^{(s)}) + \psi\nu\nabla \ell_j^{\top}(\boldsymbol{\gamma}_j^{(s)})\Delta\boldsymbol{\gamma}_j^{(s)}, \tag{8}$$

$\phi \in (0, 0.5)$, $\psi \in (0.5, 1)$, based on which the step length ν is determined. Condition (7), known as the Armijo condition (Armijo, 1966), explicitly requires a sufficient increase in ℓ_j proportional to step length ν and directional derivative $\nabla \ell^{\top}(\boldsymbol{\gamma}_j^{(s)})\Delta\boldsymbol{\gamma}_j^{(s)}$ before the line search is terminated. However, (7) alone does not guarantee convergence since ϕ can be arbitrarily small. Condition (8), known as the Goldstein condition (Goldstein, 1967), imposes a lower bound on ν so that $\boldsymbol{\gamma}_j^{(s)}$ cannot be very close to $\boldsymbol{\gamma}_j^{(s)} + \nu\Delta\boldsymbol{\gamma}_j^{(s)}$.

We present below three assumptions through which the convergence properties of the proximal Newton algorithm are achieved.

Assumption 1 *The log-partial likelihood component* $\ell_j(\boldsymbol{\gamma}_j)$ *of* (4) *is coercive, i.e.,* $\lim_{\|\boldsymbol{\gamma}_j\|_2 \to \infty} \ell_j(\boldsymbol{\gamma}_j) = -\infty$, $j = 1, \ldots, m$.

As discussed in Lange (2013, §12.3, pp.298), this assumption along with the continuity and concavity of ℓ_j guarantees that the superlevel set $\{\boldsymbol{\gamma}_j \in \mathbb{R}^{pK} : \ell_j(\boldsymbol{\gamma}_j) \geq \ell_j(\boldsymbol{\gamma}_j^{(0)})\}$ is convex and compact.

Assumption 2 *The matrix* $\mathbf{I} / \lambda_s - \nabla^2 \ell_j(\boldsymbol{\gamma}_j^{(s)})$ *on Line 16 of Algorithm 1 has a bounded condition number, i.e.,* $\exists \kappa > 0$, *such that*

$$\mathbf{I} / \lambda_s - \nabla^2 \ell_j(\boldsymbol{\gamma}_j^{(s)}) \leq \kappa, \quad j = 1, \ldots, m, \tag{9}$$

where for any invertible matrix $\mathbf{M}$, $\kappa_2(\mathbf{M}) := \|\mathbf{M}\|_2\|\mathbf{M}^{-1}\|_2$.

Assumption 3 *The sequence* $\{\lambda_s\}_{s=0}^{\infty}$ *of positive tuning parameters monotonically approaches infinity as* $s \to \infty$ *i.e.,* $\lim_{s\to\infty} \lambda_s = \infty$.

The following theorem provides a set of convergence characterizations of Algorithm 1. The proof is included in the Supplementary Information.

Theorem 1 *Let* ℓ_j *assume* (4) *with an initial iterate* $\boldsymbol{\gamma}_j^{(0)}$, *and let* $\{\boldsymbol{\gamma}_j^{(s)}\}_{s=1}^{\infty}$ *be a sequence of iterates defined by Line 20 of Algorithm 1, where* $\Delta\boldsymbol{\gamma}_j^{(s)}$ *is given by Line 16, and* $\nu > 0$ *is determined by* (7) and (8) *with* $\phi \in (0, 0.5)$ *and* $\psi \in (0.5, 1)$. *If Assumptions 1 and 2 hold, then* $\{\ell_j(\boldsymbol{\gamma}_j^{(s)})\}_{s=0}^{\infty}$ *converges and* $\lim_{s\to\infty}\|\nabla\ell_j(\boldsymbol{\gamma}_j^{(s)})\|_2 = 0$.

Note that Theorem 1 does not conclude with the convergence of $\{\boldsymbol{\gamma}_j^{(s)}\}_{s=0}^{\infty}$. However, given the fact that $\boldsymbol{\gamma}_j^*$ is a global maximizer of the concave and differentiable function ℓ_j if and only if $\nabla\ell_j(\boldsymbol{\gamma}_j^*) = \mathbf{0}$, the ultimate iterate from Algorithm 1 should be close enough to the optimal solution with a sufficiently small tolerance ϵ in most practical situations.

With a priori assumptions on the optimal solution $\boldsymbol{\gamma}_j^*$, requiring $\phi \in (0, 0.5)$ and $\psi \in (0.5, 1)$ allows a step length ν equal to 1 to ultimately satisfy (7) and (8), and enables Algorithm 1 to achieve superlinear convergence as defined below. A formal statement is given in Theorem 2, with the proof available in the Supplementary Information.

Definition 2 A sequence $\{\boldsymbol{\gamma}^{(s)}\}_{s=1}^{\infty} \subset \mathbb{R}^d$ converges superlinearly to $\boldsymbol{\gamma}^* \in \mathbb{R}^d$ if there exists a sequence $\{\xi_s\}_{s=1}^{\infty}$ of positive real numbers with $\lim_{s\to\infty} \xi_s = 0$ such that $\forall\, s \in \mathbb{N}$, $\|\boldsymbol{\gamma}^{(s+1)} - \boldsymbol{\gamma}^*\|_2 \le \xi_s \|\boldsymbol{\gamma}^{(s)} - \boldsymbol{\gamma}^*\|_2$.

Theorem 2 *Let* ℓ_j *assume* (4) *with an initial iterate* $\boldsymbol{\gamma}_j^{(0)}$, *and let* $\{\boldsymbol{\gamma}_j^{(s)}\}_{s=1}^{\infty}$ *be a sequence of iterates defined by Line 20 of Algorithm 1, where* $\Delta\boldsymbol{\gamma}_j^{(s)}$ *is given by Line 16, and* $\nu > 0$ *is determined by* (7) and (8) *with* $\phi \in (0, 0.5)$ *and* $\psi \in (0.5, 1)$. *In addition, assume that* $\{\boldsymbol{\gamma}_j^{(s)}\}_{s=1}^{\infty}$ *converges to* $\boldsymbol{\gamma}_j^*$ *with a negative definite* $\nabla^2\ell_j(\boldsymbol{\gamma}_j^*)$. *If Assumptions 1 and 3 hold, then (1)* $\exists\, s_0 \in \mathbb{N}$ *such that* $\forall\, s > s_0$, $\nu = 1$ *satisfies* (7) and (8); *(2)* $\nabla\ell_j(\boldsymbol{\gamma}_j^*) = \mathbf{0}$; *and (3)* $\{\boldsymbol{\gamma}_j^{(s)}\}_{s=0}^{\infty}$ *converges superlinearly to* $\boldsymbol{\gamma}_j^*$ *provided that* $\forall\ s \ge s_0$, $\nu = 1$ *for some* $s_0 \in \mathbb{N}$.

3.3 Shared-memory parallelization

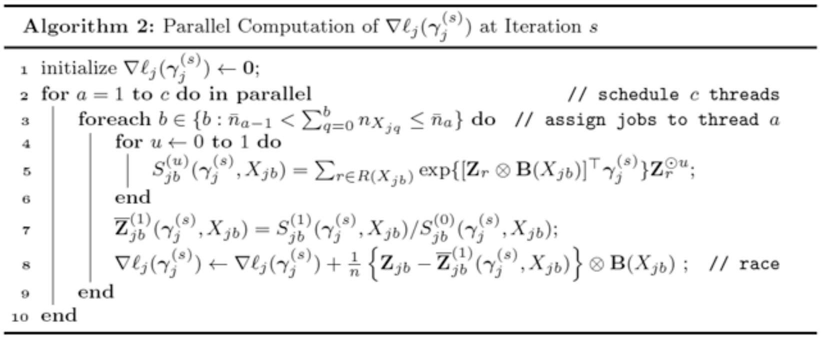
In the literature, various parallel computing schemes have been proposed to boost computational efficiency in generalized linear models (GLMs) (Peng et al., 2013; Do and Poulet, 2015; Jyothi and Babu, 2020), Bayesian inference (Goudie et al., 2020), and random forests (Wright and Ziegler, 2017), among other instances. Despite the widespread recognition from the statistics community (Eddelbuettel, 2021), there is a paucity of research on the application of parallel computing to large-scale time-to-event data, especially in a shared-memory context. The utmost reason is that modeling survival outcomes often involves risk-set-specific calculation tasks at all failure times. These tasks, unlike the observation-specific calculations in GLMs, are not equally costly in terms of computational complexity, since the size of the risk set $R(X_i)$ (defined in Section 2) varies with the observed time X_i. The unequal-sized risk sets resulting from an increasing sequence of failure times pose a challenge to load balancing, i.e., the distribution of tasks over a set of computing units (threads).

Following a distributed-memory framework, Lu et al. (2015) bypassed this issue by sample stratification so that risk sets can only be formed within a certain stratum. However, their approach becomes infeasible if stratification is not possible. Moreover, as the sample size escalates (as in our cancer applications), the distributed-memory approach becomes less appealing since having multiple copies of a large data set concurrently is not memory-efficient.

In addition to load balancing arising from unequal-sized risk sets, the presence of time-varying coefficients poses a second challenge to parallel computing. When $\boldsymbol{\beta}_j(t)$ is time-invariant, i.e., $\boldsymbol{\beta}_j(t) = \boldsymbol{\beta}_j$, one may approach the problem by first calculating $\{\exp(\mathbf{Z}_i^\top \boldsymbol{\beta}_j)\}_{i=1}^n$ and then obtaining the cumulative sums of $\{\exp(\mathbf{Z}_i^\top \boldsymbol{\beta}_j)\}_{i=1}^n$ in parallel by means of the prefix sum algorithm (Casanova et al., 2008). When $\boldsymbol{\beta}_j(t)$ varies with time t, however, $\exp[\mathbf{Z}_i^\top \boldsymbol{\beta}_j(t)]$ has to be re-evaluated for different risk sets, making the aforementioned approach infeasible.

To tackle the issue of load balancing in the presence of massive data and time-varying coefficients, we propose a shared-memory paradigm that optimizes workload allocation among a given number c of available threads where $c \geq 2$. For time X_{jq} of failure type j, let $n_{X_{jq}} := |R(X_{jq})|$, i.e., the number of elements of the risk set $R(X_{jq})$. For failure type j, Algorithm 1 culminates in the calculations of $\ell_j(\boldsymbol{\gamma}_j^{(s)})$, $\nabla \ell_j(\boldsymbol{\gamma}_j^{(s)})$ and $\nabla^2 \ell_j(\boldsymbol{\gamma}_j^{(s)})$ at iteration s, which in turn depend upon $S_{jq}^{(u)}(\boldsymbol{\gamma}_j^{(s)}, X_{jq})$. An analysis of time complexity reveals that computing $S_{jq}^{(u)}(\boldsymbol{\gamma}_j^{(s)}, X_{jq})$ costs $O(pKn_{X_{jq}})$, $O(p(K+1)n_{X_{jq}})$ and $O(p(4K+3p+3)n_{X_{jq}})$, respectively, for $u = 0, 1, 2$. The linearity with respect to $n_{X_{jq}}$ suggests using as cutoffs the c-quantiles $\{\bar{n}_a\}_{a=1}^{c-1}$ of the cumulative sums of $\{n_{X_{jq}}\}_{q=0}^{n_j}$ (with $n_{X_{j0}} = 0$) to partition the collection of n_j risk sets into c subcollections of nearly equal computational costs.

Let $\bar{n}_c$ denote the sum of $\{n_{X_{jq}}\}_{q=0}^{n_j}$ and let $\bar{n}_0 = 0$. Algorithm 2 presents the parallelization of computing $\nabla \ell_j(\boldsymbol{\gamma}_j^{(s)})$ at iteration s (Lines 5–14 of Algorithm 1), in which Line 8 is a race condition requiring execution on one thread at a time (nonparallel). The other two quantities can be obtained similarly. Evidence in the Supplementary Information using the SEER breast and prostate cancer data demonstrates the speedup and efficiency of the proposed parallelization scheme.



Algorithm 2: Parallel Computation of $\nabla \ell_j(\gamma_j^{(s)})$ at Iteration s

```
initialize $\nabla \ell_j(\gamma_j^{(s)}) \leftarrow \mathbf{0}$;
for $a = 1$ to $c$ do in parallel                                    // schedule $c$ threads
    foreach $b \in \{b : \bar{n}_{a-1} < \sum_{q=0}^{b} n_{X_{jq}} \leq \bar{n}_a\}$ do  // assign jobs to thread $a$
        for $u \leftarrow 0$ to $1$ do
            $S_{jb}^{(u)}(\gamma_j^{(s)}, X_{jb}) = \sum_{r \in R(X_{jb})} \exp\{[\mathbf{Z}_r \otimes \mathbf{B}(X_{jb})]^\top \gamma_j^{(s)}\}\mathbf{Z}_r^{\odot u}$;
        end
        $\overline{\mathbf{Z}}_{jb}^{(1)}(\gamma_j^{(s)}, X_{jb}) = S_{jb}^{(1)}(\gamma_j^{(s)}, X_{jb}) / S_{jb}^{(0)}(\gamma_j^{(s)}, X_{jb})$;
        $\nabla \ell_j(\gamma_j^{(s)}) \leftarrow \nabla \ell_j(\gamma_j^{(s)}) + \frac{1}{n}\left\{\mathbf{Z}_{jb} - \overline{\mathbf{Z}}_{jb}^{(1)}(\gamma_j^{(s)}, X_{jb})\right\} \otimes \mathbf{B}(X_{jb})$ ;  // race
    end
end
```

4 Hypothesis testing

Inferential attempts regarding the significance of the time-varying effects $\boldsymbol{\beta}_j(t)$ for type j failure can be formulated as the linear hypothesis $H_{01} : \mathbf{C}\boldsymbol{\beta}_j(t) = \mathbf{0}$, where $\mathbf{C}$ is a contrast matrix with full row rank r. Our penalty-free spline-based modeling and estimation lay the groundwork for a straightforward Wald-type significance test. By (2), the null H_{01} can be rewritten as $H_{01} : [\mathbf{C} \otimes \mathbf{B}^\top(t)]\boldsymbol{\gamma}_j = \mathbf{0}$, and a Wald test statistic is given by

$$\hat{\boldsymbol{\gamma}}_j^\top [\mathbf{C}^\top \otimes \mathbf{B}(t)] n \{[\mathbf{C} \otimes \mathbf{B}^\top(t)][\mathbf{I} / \lambda - \nabla^2 \ell_j(\hat{\boldsymbol{\gamma}}_j)]^{-1} [\mathbf{C}^\top \otimes \mathbf{B}(t)]\}^{-1} [\mathbf{C} \otimes \mathbf{B}^\top(t)] \hat{\boldsymbol{\gamma}}_j,$$

where $\hat{\boldsymbol{\gamma}}_j$ is the estimate of $\boldsymbol{\gamma}_j$. Under the null H_{01}, the test statistic approximately follows a chi-square distribution with r degrees of freedom. Pointwise confidence intervals across time are readily obtainable via test inversion. For instance, if one wants to test whether $\beta_{jl}(t) = 0$, where $\beta_{jl}(t)$ is the lth component of $\boldsymbol{\beta}_j(t)$, $l = 1, \ldots, p$, then $\mathbf{C} = [0, \ldots, 1, \ldots, 0]$, where only the lth element equals 1.

A second test of particular interest is to examine whether a certain effect $\beta_{jl}(t)$ is constant over time. In the literature, various procedures have been proposed to address this inference issue. As the default check for nonproportionality in the R package survival (Therneau, 2020), Grambsch and Therneau (1994) suggested a generalized least squares test on the scaled Schoenfeld residuals. Assuming $\beta_{jl}(t) = \beta_{jl} + \theta_{jl} g_{jl}(t)$ with unknown constants β_{jl} and θ_{jl}, and a possibly unknown function $g_{jl}(\cdot)$, the residuals are based on a one-step Newton-Raphson estimator $\hat{\theta}_{jl}$ of θ_{jl} and an estimator $\hat{\beta}_{jl}$ of β_{jl} from the Cox proportional hazards model. This approach provides a fast and easy check for nonproportionality without the need to fit a model of time-varying effects. Relying on a one-term Taylor approximation, however, using the scaled Schoenfeld residuals may lead to inflated type-I error when $|\beta_{jl}(t) - \beta_{jl}|$ is large. In addition, the residual calculation may be unstable, particularly near the end of follow-up (Therneau and Grambsch, 2000, pp.133).

To test whether the effect $\beta_{jl}(t)$ is time-invariant, our approach amounts to a Wald test on the control points. Similar to He et al. (2017), we observe that if $\gamma_{jl1} = \cdots = \gamma_{jlK} = \bar{\gamma}$, then

$$\beta_{jl}(t) = \bar{\gamma} \sum_{k=1}^{K} B_k(t) = \bar{\gamma},$$

in which we utilize the property of the B-spline basis that $\sum_{k=1}^{K} B_k(t) = 1$ for any t. This leads to the null hypothesis $H_{02l} : \bar{\mathbf{L}}\boldsymbol{\gamma}_{jl} = \mathbf{0}$, where $\bar{\mathbf{L}}$ is a $(K - 1) \times K$ matrix given by

$$\bar{\mathbf{L}} = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & -1 \end{bmatrix}.$$

A Wald test statistic can thus be constructed as

$$\hat{\boldsymbol{\gamma}}_{jl}^{\top}\bar{\mathbf{L}}^{\top}(\bar{\mathbf{L}}\mathbf{M}_{jl}\bar{\mathbf{L}}^{\top})^{-1}\bar{\mathbf{L}}\hat{\boldsymbol{\gamma}}_{jl},$$

where $\mathbf{M}_{jl}$ denotes the lth diagonal $K \times K$ block of $[\mathbf{I} / \lambda - \nabla^2 \ell_j(\hat{\boldsymbol{\gamma}}_j)]^{-1} / n$ for $l = 1, \ldots, p$. Under the null H_{02l}, this test statistic approximately follows a chi-square distribution with $K - 1$ degrees of freedom.

Once the time-varying effects are distinguished from the time-independent ones through tests of nonproportionality, a cause-specific hazard model with time-variant and -invariant coefficients can be fit via an equality constrained maximization problem. Suppose $\beta_{jl_1}(t), \ldots, \beta_{jl_{\bar{p}}}(t)$ are flagged as time-variant effects. Let $\mathbf{L}$ be a $p(K - 1) \times pK$ matrix whose lth $(K - 1) \times K$ submatrix on the diagonal equals $\bar{\mathbf{L}}$ if $\beta_{jl}(t)$ is time-variant or $\mathbf{0}$ otherwise, and all off-diagonal blocks equal $\mathbf{0}$. Solving the following problem

$$\begin{aligned} \underset{\Delta\boldsymbol{\gamma}_j \in \mathbb{R}^{pK}}{\text{maximize}} \quad & \nabla \ell_j^{\top}(\boldsymbol{\gamma}_j)\Delta\boldsymbol{\gamma}_j + \Delta\boldsymbol{\gamma}_j^{\top}[\nabla^2 \ell_j(\boldsymbol{\gamma}_j) - \mathbf{I} / \lambda]\Delta\boldsymbol{\gamma}_j / 2 \\ \text{subject to} \quad & \mathbf{L}\Delta\boldsymbol{\gamma}_j = \mathbf{0}, \end{aligned}$$

in which $\boldsymbol{\gamma}_j$ is a feasible point satisfying $\mathbf{L}_{\boldsymbol{\gamma}_j} = \mathbf{0}$ (e.g., $\boldsymbol{\gamma}_j = \mathbf{0}$), we can obtain the Newton step

$$\Delta\boldsymbol{\gamma}_j^{*} = \mathbf{U}\left[\mathbf{U}^{\top}\{\mathbf{I} / \lambda - \nabla^2 \ell_j(\boldsymbol{\gamma}_j)\}\mathbf{U}\right]^{-1}\mathbf{U}^{\top}\nabla \ell_j(\boldsymbol{\gamma}_j)$$

at each iteration (to replace Line 16 of Algorithm 1), where $\mathbf{U}$ is a $pK \times \bar{p}$ matrix, whose range (column space) is the null space of $\mathbf{L}$.

5 Simulation Study

To compare the proximal Newton algorithm with the NaiveN and QuasiN methods, we conducted a series of simulation experiments. The NaiveN was implemented via the function coxph in the R package survival, and the QuasiN was implemented using the base R function optim (the BFGS algorithm, Nocedal and Wright, 2006, §6.1). Since the estimation and inference with respect to different failure types can be handled separately within a cause-specific hazard framework, we focused primarily on a single failure type and dropped the index j to simplify notation.

In each simulation scenario, a number of independent data replicates were generated with the sample size n ranging from 1,000 to 10,000. We considered $p = 5$ covariates $\mathbf{Z}_i$ drawn from a multivariate normal distribution with zero mean, unit variance and an AR(1) correlation structure with parameter $\rho = 0.6$. To introduce numerical singularity, the continuous covariates were then dichotomized into binary variables, with the probability of being one uniformly varying from 0.8 to 0.9. This treatment intended to mimic our application setting where the Hessian matrix had a large condition number even when the number of observations was large. A constant baseline hazard $\lambda_0(t) = 0.5$ was used with covariate parameters calibrated as $\beta(t) = [1, \sin(3\pi t/4), -1, -1, 1]^T$. Failure times were generated from the survivor function of (1), and censoring times were drawn from a uniform distribution between 0 and 3. Observed times were determined as the minimum of the failure and censoring time pairs.

5.1 Estimation accuracy

To assess the estimation accuracy of the proposed ProxiN, Table 1 presents the integrated mean squared error (IMSE), average bias, and average variance associated with the three algorithms. Model fitting was performed by treating all coefficients as time-dependent. Using a uniform distribution, we sampled 1,000 distinct time points from the interval between 0 and 3. At each time t, the mean estimates of $\beta_1(t)$ and $\beta_2(t)$ across 100 data replicates were used to calculate the mean squared error and variance, the difference of which is the squared bias. Taking the average across the 1,000 time points, we obtained the IMSE, average squared bias, and average variance. The average bias is simply the squared root of the average squared bias.

Panel A of Table 1 displays the three measures of estimation accuracy for $\beta_1(t)$. Since the nearly singular Hessian matrix was inaccurately approximated by a matrix in the BFGS algorithm, the QuasiN had consistently much higher IMSE than the other two methods. Of these two, the ProxiN had lower IMSE, bias and variance, especially when the sample size equaled 1,000 or 5,000. As a side observation, the IMSE was largely due to the variance component for all three methods. When it comes to the estimation accuracy of $\beta_2(t)$, the ProxiN overall outperformed the alternatives and the performance of QuasiN was even worse than that for $\beta_1(t)$. The difference in the accuracy measures among the first two approaches shrunk as the sample size increased. To explore the impact of different censoring schemes, we varied the uniform distribution with different ranges of support (from [0, 3] to [1.5, 3]), and used the exponential distribution with different rates (from 0.2 to 1.0) as an alternative scheme. In addition, we also considered the performance of the ProxiN in settings where the sample size was of a similar order as in our cancer applications. Results are also available in the Supplementary Information.

As noted in Section 2, it is conceptually desirable that the number of B-spline basis functions is allowed to vary across different time-varying coefficients. Although a systematic investigation into such a general case is absent in the literature and beyond the scope of this article, we conducted simulation experiments (results available in the Supplementary Information) to shed the first light on knot selection based on the variation of a covariate. The bottom line is that as the covariate variation shrinks toward zero, fewer knots should be applied to expanding a time-varying coefficient, so that the effect can be estimated with sufficient accuracy.

With the sample size equal to 1,000, Figure 2 displays the true value along with the pointwise mean of estimates $\hat{\beta}_1(t)$ and $\hat{\beta}_2(t)$ across 100 data replicates, where $\beta_4(t) = t^2 \exp(t/2)/9$ and $\beta_5(t) = \exp(-1.5t)$. The QuasiN was not included due to its poor performance. For $\beta_1(t)$, the estimate curve of ProxiN had much smaller deviance from the true value curve than the NaiveN, the deviance of which was in the opposite direction. As for the time-varying $\beta_2(t)$, the estimate curve of ProxiN varied closely along the true value curve, whereas the estimate curve of the NaiveN deviated from the true one when $t > 2$.

Given a 95% confidence level, Figure 3 compares the coverage probability (CP) of estimates $\hat{\beta}_1(t)$ and $\hat{\beta}_2(t)$ resulting from the ProxiN and NaiveN, with time t varying from 0 to 3. As time increases, the CP curve for $\hat{\beta}_1(t)$ from the ProxiN algorithm fluctuates more closely around 0.95 than the NaiveN, though the CP curve of ProxiN drops sharply near the end of follow-up ($t = 3$) when $n = 5{,}000$ or 10,000. The QuasiN approach was not included as it often led to a singular Hessian matrix.

To illustrate the performance of ProxiN with more than one cause of failure, we compared the estimation accuracy of ProxiN, NaiveN and QuasiN with different sample sizes and two causes of failure (Table 2 and Supplementary Table 10). With the notation in Section 2, we set $\beta_{11}(t) = 1$, $\beta_{12}(t) = \sin(3\pi t/4)$, $\beta_{13}(t) = -1$, $\beta_{14}(t) = -1$, $\beta_{15}(t) = 1$ for the first failure type, and set $\beta_{21}(t) = -1$, $\beta_{22}(t) = \cos(3\pi t/4)$, $\beta_{23}(t) = 1$, $\beta_{24}(t) = 1$, $\beta_{25}(t) = -1$ for the second failure type. Failure times and types were determined based on Beyersmann et al. (2009, §3.1). Censoring times were generated from a uniform distribution between 0 and 3. As in the case with only one cause of failure, the ProxiN outperformed the other two methods in terms of the IMSE, average bias, and average variance. A larger sample generally led to more accurate estimation of the true effects.

5.2 Testing for time-varying effects

The assessment of the test of nonproportionality is reported in Figure 4, where the average type-I error rate regarding a test of the time-invariant $\beta_1(t)$, and the average power regarding a test of the time-variant $\beta_2(t)$ across 1,000 data replicates are plotted against different levels of sample size, with a 5% significance level. When $\beta_2(t) = \sin(3\pi t/4)$ (top two panels), the ProxiN had a lower error curve for $\beta_1(t)$ and a higher power curve for $\beta_2(t)$. When the magnitude of $\beta_2(t)$ was tripled (bottom two panels), i.e., $\beta_2(t) = 3\sin(3\pi t/4)$, the NaiveN had much inflated error and power curves, both of which approached one as the sample size grew. By contrast, the proposed ProxiN maintained a controlled error curve around 5% as well as a high-level power line at one.

6 Applications

To demonstrate the real-world performance of the proposed estimation and testing procedures, we applied these methods to the nationwide breast and prostate cancer survival database administered by the U.S. Surveillance, Epidemiology, and End Results (SEER) Program (Surveillance, Epidemiology, and End Results Program, 2017, 2019).

6.1 SEER breast cancer data

For our study, 1,093,192 female patients first diagnosed with breast cancer between 1973 and 2015 were selected and their cause-specific deaths (cancer or other, see Brouwer et al., 2020), if not censored, were recorded. In the analysis, we considered three risk factors: age, race and tumor stage at the time of diagnosis. Among all the patients, 24.21% were younger than 50 at diagnosis, 24.02% aged 50 to 59, 23.68% aged 60 to 69, and 28.09% were at least 70; 9.75% were black, 82.37% were white (including Hispanic), 7.42% belonged to other racial groups (American Indian, Alaska Native, Asian, Pacific Islander), and the remaining 0.46% were unknown. As for tumor staging, 60.02% had localized tumors, 31.39% had regionalized tumors, 6.13% had distant tumors, and 2.46% had their tumors recorded as unstaged. Event times (time to cancer death, other deaths or censoring) ranged from 1 month to 515 months, with a median of 80 months since diagnosis.

Treating cancer and other deaths as two distinct types of failure, we fit two cause-specific hazard models to the SEER breast cancer data with time-varying coefficients via Algorithm 1. Effect estimates as well as pointwise 95% confidence intervals are displayed in Figure 5 with a 20-year presentation. Treating the localized stage as the reference level and the other three as covariates, the top two panels display the overall shrinking staging effects on the two causes of death. Each of the three stages had a larger effect on cancer death than that on other deaths. As expected, an advanced stage had a stronger effect on cancer death than an early stage. Relative to the white cohort, black breast cancer patients were more likely to die as a result of either cancer or other causes. They had an initial increase in the hazard of cancer death, followed by a gradual decrease to nearly zero. In contrast, the shrinkage of race effects on other deaths was slower. The three effects of age groups on cancer death immediately declined after diagnosis and then either remained stable (older than 70) or gradually increased (younger than 60). Age effects on other deaths remained relatively flat as time passed. The speedup and efficiency of the parallelized ProxiN is discussed in detail in the Supplementary Information.

6.2 SEER prostate cancer data

In the prostate cancer data, 716,553 patients with a first diagnosis of prostate cancer between 2004 and 2017 were chosen and their cause-specific deaths or censorings were recorded. Similarly as in the analysis of breast cancer data, we examined age, race and tumor stage at the time of diagnosis. Among all the patients, 2.79% were younger than 50 at diagnosis, 20.83% aged 50 to 59, 40.74% aged 60 to 69, and 35.64% were at least 70; 14.58% were black, 69.44% were non-Hispanic white, 8.81% were Hispanic, and the remaining 7.17% belonged to other racial groups. (Since this data were collected only starting in 2004, the registry used different ethnic groupings than the breast cancer data, which started in 1973.) In terms of summary staging, 82.41% had localized tumors, 11.32% had regionalized tumors by direct extension, 0.6% had regional tumors to lymph nodes, 1.12% had their tumors as regional both by direct extension and lymph nodes, and 4.54% had tumors of unknown stage. Event times ranged from 1 month to 167 months, with a median of 6 years since diagnosis.

As in the application of breast cancer, we fit two cause-specific hazard models with time-varying coefficients to the SEER prostate cancer data. Estimates and confidence intervals are displayed in Figure 6 with a 10-year presentation. With the localized stage as the reference group and the other four as covariates, the top two panels reveal different patterns of staging effects on the two types of death. Overall, an advanced tumor stage led to a considerably higher hazard ratio of cancer death than the hazard ratio of other deaths. While the effects of regional both and regional by direct extension on cancer death were significantly positive, their effects on other deaths were negative. Nonproportionality tests with 5% size of the staging effects on cancer death indicated that they should all be viewed as time-variant. Relative to the white cohort, black prostate cancer patients were more likely to die as a result of either cancer or other causes. As expected, older patients had a higher hazard of dying from any cause than younger patients.

7 Discussion

The increasing availability of large-scale and complex data has the potential to vastly improve our understanding of important real-world problems such as cancer survival, but only with methodological and computational advances. Existing data-expansion- or gradient-based methods impose formidable computational costs and numerical instability to model fitting. To facilitate efficient and accurate statistical analysis in this context, we propose the proximal Newton algorithm along with a shared-memory parallelization paradigm and testing procedures. Simulation analyses demonstrate superior scalability, efficiency and estimation accuracy compared to alternative approaches. Applications to the SEER breast and prostate cancer data confirm the excellent real-world performance of our proposed approach.

Although developed for analyzing cancer data, the proposed technique can be used in many other applications that involve time-varying effect analysis. In kidney transplantation, for example, the relative risk of death among recipients relative to those on dialysis is known to initially increase due to surgery, but the subsequent decrease eventually leads to an overall survival benefit (Wolfe et al., 1999). Similarly, when comparing two infant feeding strategies for preventing mother-to-child human immunodeficiency virus transmission, evidence from a randomized trial showed that, although breastfeeding with prophylaxis was associated with lower infant mortality at 7 months relative to formula feeding, this difference shrunk to insignificance through age 18 months (Thior et al., 2006). Obesity, a well-known risk factor of mortality in the general population, was found among dialysis patients to have a short-term protective effect on survival and an increased risk of death after a long-term exposure (Kalantar-Zadeh et al., 2003; Kalantar-Zadeh, 2005; de Mutsert et al., 2007; Dekker et al., 2008). In all these instances, our proposed methods would have undoubtedly contributed to a better understanding of the changes in effects over time.

Depending on specific analytic needs, the proximal Newton algorithm can also be applied to a more general setting with stratum-specific baseline hazards. In a head and neck cancer application, for instance, there was evidence of substantial differences in the baseline hazards by tumor stage (Brouwer et al., 2020). A stratified analysis taking account of the stage-wise variation may better reflect the effect evolution of prognostic factors. As another example, the analysis of electronic health records often involves integrating data from multiple health care providers. Stratification by providers can alleviate the mediation between provider-specific effects and the effects of risk factors. In either case, our proposed method can readily handle the less demanding computational burdens with reduced risk sets.

As for the determination of the number and location of knots in the cause-specific hazard model, we followed the rules by Gray (1992), that is, a small number of knots (e.g., 10) chosen to include an equal number of events within each time interval. Although using this early suggestion yields stable estimation in our applications, a systematic guideline on this issue is beyond the current endeavors. In addition, it is worth further exploration into the use of the penalized B-spline to alleviate overfitting and increase smoothness in coefficient estimation. Moreover, when the dimension of the parameter space is very high, existing model selection techniques such as Yan and Huang (2012) would no longer be feasible. This necessitates in-depth investigation into high-dimensional variable selection methods with time-varying effects. Fortunately, the superb performance of the proposed algorithm paves the way for possible advances along these paths in a large-scale cause-specific setting.

In the top right panel of Figure 6, the effect curve of lymph on other deaths has more variation than the curve of ext especially for the initial 2.5 years since diagnosis, but the test of nonproportionality identified the effect of ext as time-dependent rather than the effect of lymph. This suggests that the effect of lymph on other deaths may not be nonzero everywhere. Although addressing this issue systematically is beyond the aims of the current article, more analytical effort is worthwhile on accounting for zero-effect regions in competing risk models with time-varying effects. Currently, there is a paucity of studies in the survival literature on time-varying effect modeling with zero-effect regions. For a relevant account on varying coefficients with zero-effect regions in the context of generalized linear models, we refer to a recent work by Yang (2020).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors would like to thank Dr. Kirsten F. Herold (University of Michigan), the Associate Editor and two referees for helpful comments on the manuscript. This work was partially supported by The University of Michigan Office of Research, the University of Michigan Rogel Cancer Center (Project Number P30CA046592) and the National Institutes of Health (Grant Number UL1TR002240).

Appendix

This appendix is devoted to the derivation of the gradient $\nabla \ell_j(\boldsymbol{\gamma}_j)$ and Hessian matrix $\nabla^2 \ell_j(\boldsymbol{\gamma}_j)$ of $\ell_j(\boldsymbol{\gamma}_j)$ as in (4). We define

$$S_{ij}^{(u)}(\boldsymbol{\gamma}_j, X_i) := \sum_{r \in R(X_i)} \exp\{[\mathbf{Z}_r \otimes \mathbf{B}(X_i)]^\top \boldsymbol{\gamma}_j\}\mathbf{Z}_r^{\odot u}, \quad u = 0, 1, 2,$$

where for a vector $\mathbf{v} \in \mathbb{R}^p$, $\mathbf{v}^{\odot 0} := 1$, $\mathbf{v}^{\odot 1} := \mathbf{v}$, and $\mathbf{v}^{\odot 2} := \mathbf{v}\mathbf{v}^\top$. The gradient $\nabla \ell_j(\boldsymbol{\gamma}_j)$ and Hessian $\nabla^2 \ell_j(\boldsymbol{\gamma}_j)$ of $\ell_j(\boldsymbol{\gamma}_j)$ are hence given by

$$\nabla \ell_j(\boldsymbol{\gamma}_j) = \frac{1}{n} \sum_{i=1}^{n} \Delta_{ij} \{\mathbf{Z}_i - \overline{\mathbf{Z}}_{ij}(\boldsymbol{\gamma}_j, X_i)\} \otimes \mathbf{B}(X_i), \tag{10}$$

$$\nabla^2 \ell_j(\boldsymbol{\gamma}_j) = -\frac{1}{n} \sum_{i=1}^{n} \Delta_{ij} \mathbf{V}_{ij}(\boldsymbol{\gamma}_j, X_i) \otimes \{\mathbf{B}(X_i)\mathbf{B}^\top(X_i)\}, \tag{11}$$

in which

$$\overline{\mathbf{Z}}_{ij}(\boldsymbol{\gamma}_j, X_i) := \frac{S_{ij}^{(1)}(\boldsymbol{\gamma}_j, X_i)}{S_{ij}^{(0)}(\boldsymbol{\gamma}_j, X_i)}, \quad \mathbf{V}_{ij}(\boldsymbol{\gamma}_j, X_i) := \frac{S_{ij}^{(2)}(\boldsymbol{\gamma}_j, X_i)}{S_{ij}^{(0)}(\boldsymbol{\gamma}_j, X_i)} - \overline{\mathbf{Z}}_{ij}^{\odot 2}(\boldsymbol{\gamma}_j, X_i).$$

References

Armijo L (1966). Minimization of functions having Lipschitz continuous first partial derivatives. Pacific Journal of Mathematics, 16(1):1–3.

Baulies S, Belin L, Mallon P, Senechal C, Pierga J, Cottu P, Sablin M, Sastre X, Asselain B, Rouzier R, et al. (2015). Time-varying effect and long-term survival analysis in breast cancer patients treated with neoadjuvant chemotherapy. British Journal of Cancer, 113(1):30–36. [PubMed: 26079300]

Bellera CA, MacGrogan G, Debled M, de Lara CT, Brouste V, and Mathoulin-Pélissier S (2010). Variables with time-varying effects and the Cox model: some statistical concepts illustrated with a prognostic factor study in breast cancer. BMC Medical Research Methodology, 10(1):1–12. [PubMed: 20053272]

Beyersmann J, Latouche A, Buchholz A, and Schumacher M (2009). Simulating competing risks data in survival analysis. Statistics in Medicine, 28(6):956–971. [PubMed: 19125387]

Boyd S and Vandenberghe L (2004). Convex Optimization. Cambridge University Press.

Brouwer AF, He K, Chinn SB, Mondul AM, Chapman CH, Ryser MD, Banerjee M, Eisenberg MC, Meza R, and Taylor JMG (2020). Time-varying survival effects for squamous cell carcinomas at oropharyngeal and nonoropharyngeal head and neck sites in the United States, 1973–2015. Cancer, 126(23):5137–5146. [PubMed: 32888317]

Casanova H, Legrand A, and Robert Y (2008). Parallel Algorithms. CRC Press.

de Boor C (2001). A Practical Guide to Splines. Springer, Revised edition.

de Mutsert R, Snijder MB, van der Sman-de Beer F, Seidell JC, Boeschoten EW, Krediet RT, Dekker JM, Vandenbroucke JP, Dekker FW, et al. (2007). Association between body mass index and mortality is similar in the hemodialysis population and the general population at high age and equal duration of follow-up. Journal of the American Society of Nephrology, 18(3):967–974. [PubMed: 17267739]

Dekker FW, de Mutsert R, Van Dijk PC, Zoccali C, and Jager KJ (2008). Survival analysis: time-dependent effects and time-varying risk factors. Kidney International, 74(8):994–997. [PubMed: 18633346]

Do T-N and Poulet F (2015). Parallel multiclass logistic regression for classifying large scale image datasets. In Advanced Computational Methods for Knowledge Engineering, pages 255–266. Springer.

Eddelbuettel D (2021). CRAN Task View: High-Performance and Parallel Computing with R. https://cran.r-project.org/web/views/HighPerformanceComputing.html. Accessed: 2021–01-26.

Eddelbuettel D and Balamuta JJ (2018). Extending R with C++: A brief introduction to Rcpp. The American Statistician, 72(1):28–36.

Eddelbuettel D and François R (2011). Rcpp: Seamless R and C++ integration. Journal of Statistical Software, 40(8):1–18.

Eddelbuettel D and Sanderson C (2014). RcppArmadillo: Accelerating R with high-performance C++ linear algebra. Computational Statistics & Data Analysis, 71:1054–1063.

Goldstein AA (1967). Constructive Real Analysis. Harper & Row.

Goudie RJ, Turner RM, De Angelis D, and Thomas A (2020). Multi-BUGS: A parallel implementation of the BUGS modelling framework for faster Bayesian inference. Journal of Statistical Software, 95(7):1–20.

Grambsch PM and Therneau TM (1994). Proportional hazards tests and diagnostics based on weighted residuals. Biometrika, 81(3):515–526.

Gray RJ (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 RJ (1994). Spline-based tests in survival analysis. Biometrics, 50(3):640–652. [PubMed: 7981391]

Hastie T and Tibshirani R (1993). Varying-coefficient models. Journal of the Royal Statistical Society: Series B, 55(4):757–779.

He K, Yang Y, Li Y, 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, Zhu J, Kang J, and Li Y (2021). Stratified cox models with time-varying effects for national kidney transplant patients: A new block-wise steepest ascent method. Biometrics.

Hester J and Schmidt D (2020). bench: High Precision Timing of R Expressions. https://cran.r-project.org/package=bench. R package version 1.1.1.

Jyothi R and Babu P (2020). Piano: A fast parallel iterative algorithm for multinomial and sparse multinomial logistic regression. https://arxiv.org/abs/2002.09133. Accessed: 2021–09-14.

Kalantar-Zadeh K (2005). Causes and consequences of the reverse epidemiology of body mass index in dialysis patients. Journal of Renal Nutrition, 15(1):142–147. [PubMed: 15648024]

Kalantar-Zadeh K, Block G, Humphreys MH, and Kopple JD (2003). Reverse epidemiology of cardiovascular risk factors in maintenance dialysis patients. Kidney International, 63(3):793–808. [PubMed: 12631061]

Kalbfleisch JD and Prentice RL (2002). The Statistical Analysis of Failure Time Data. John Wiley & Sons, Second edition.

Lange K (2013). Optimization. Springer Science & Business Media, second edition.

Lee JD, Sun Y, and Saunders M (2012). Proximal Newton-type methods for convex optimization. Advances in Neural Information Processing Systems, 25:827–835.

Lee JD, Sun Y, and Saunders MA (2014). Proximal Newton-type methods for minimizing composite functions. SIAM Journal on Optimization, 24(3):1420–1443.

Levenberg K (1944). A method for the solution of certain non-linear problems in least squares. Quarterly of Applied Mathematics, 2(2):164–168.

Lu C-L, Wang S, Ji Z, Wu Y, Xiong L, Jiang X, and Ohno-Machado L (2015). WebDISCO: a web service for distributed Cox model learning without patient-level data sharing. Journal of the American Medical Informatics Association, 22(6):1212–1219. [PubMed: 26159465]

Marquardt DW (1963). An algorithm for least-squares estimation of non-linear parameters. Journal of the Society for Industrial and Applied Mathematics, 11(2):431–441.

Nocedal J and Wright S (2006). Numerical Optimization. Springer Science & Business Media.

Parikh N and Boyd S (2014). Proximal algorithms. Foundations and Trends® in Optimization, 1(3):127–239.

Peng H, Liang D, and Choi C (2013). Evaluating parallel logistic regression models. In 2013 IEEE International Conference on Big Data, pages 119–126. IEEE.

Perperoglou A, le Cessie S, and van Houwelingen HC (2006). A fast routine for fitting Cox models with time varying effects of the covariates. Computer Methods and Programs in Biomedicine, 81(2):154–161. [PubMed: 16426701]

Rockafellar RT (1970). Convex Analysis. Princeton University Press.

Surveillance, Epidemiology, and End Results Program (2017). Incidence - SEER 9 Regs Research Data, Nov 2017 Sub (1973–2015) <Katrina/Rita Population Adjustment>. https://seer.cancer.gov/data-software/documentation/seerstat/nov2017. Accessed: 2021–1-26.

Surveillance, Epidemiology, and End Results Program (2019). Incidence - SEER Research Data, 18 Registries, Nov 2019 Sub (2000–2017). https://seer.cancer.gov/data-software/documentation/seerstat/nov2019. Accessed: 2021–1-26.

Therneau T, Crowson C, and Atkinson E (2020). Using Time Dependent Covariates and Time Dependent Coefficients in the Cox Model. https://cran.r-project.org/web/packages/survival/vignettes/timedep.pdf. Accessed: 2021–01-26.

Therneau TM (2020). A Package for Survival Analysis in R. R package version 3.2–7.

Therneau TM and Grambsch PM (2000). Modeling survival data: Extending the Cox model. Springer.

Thior I, Lockman S, Smeaton LM, Shapiro RL, Wester C, Heymann SJ, Gilbert PB, Stevens L, Peter T, Kim S, et al. (2006). Breastfeeding plus infant zidovudine prophylaxis for 6 months vs formula feeding plus infant zidovudine for 1 month to reduce mother-to-child HIV transmission in Botswana. JAMA, 296(7):794–805. [PubMed: 16905785]

Tutz G and Binder H (2004). Flexible modelling of discrete failure time including time-varying smooth effects. Statistics in Medicine, 23(15):2445–2461. [PubMed: 15273958]

Verweij PJM and van Houwelingen HC (1995). Time-dependent effects of fixed covariates in Cox regression. Biometrics, 51(4):1550–1556. [PubMed: 8589239]

Wolfe RA, Ashby VB, Milford EL, Ojo AO, Ettenger RE, Agodoa LY, Held PJ, and Port FK (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. [PubMed: 10580071]

Wright MN and Ziegler A (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. Journal of Statistical Software, 95(7).

Yan J and Huang J (2012). Model selection for Cox models with time-varying coefficients. Biometrics, 68(2):419–428. [PubMed: 22506825]

Yang Y (2020). Novel Methods for Estimation and Inference in Varying Coefficient Models. PhD thesis, University of Michigan, ProQuest LLC, 789 East Eisenhower Parkway, P.O. Box 1346, Ann Arbor, MI 48106–1346. https://deepblue.lib.umich.edu/bitstream/handle/2027.42/163251/yuanyang_1.pdf?sequence=1.

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

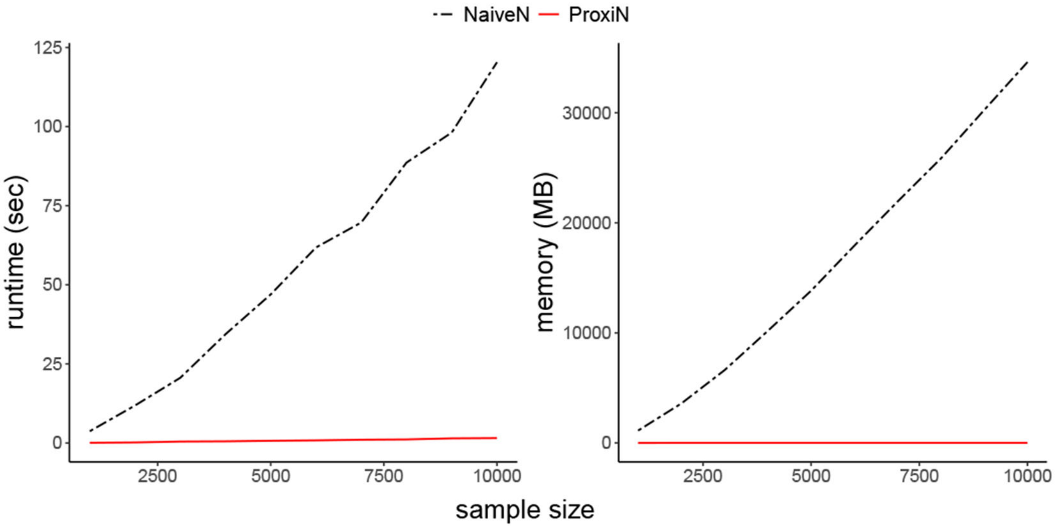


Fig. 1.

Runtime and memory usage of proximal Newton (ProxiN) and naive Newton (NaiveN) with sample sizes varying from 1,000 to 10,000. In each scenario, 10 data replicates were generated, and a fixed number of $K = 10$ knots were used for model fitting. Dichotomization was not applied to covariates. A tolerance level $\epsilon = 10^{-10}$ was used. The vertical axis displays average runtime (in seconds) across the 10 simulated data sets. Experiments were conducted on an Intel® Xeon® Gold 6254 quad-processor with max frequency 4 GHz and RAM 576 GB. ProxiN was implemented using Rcpp (Eddelbuettel and François, 2011; Eddelbuettel and Balamuta, 2018) and RcppArmadillo (Eddelbuettel and Sanderson, 2014). Runtime and memory usage were measured using bench (Hester and Schmidt, 2020).

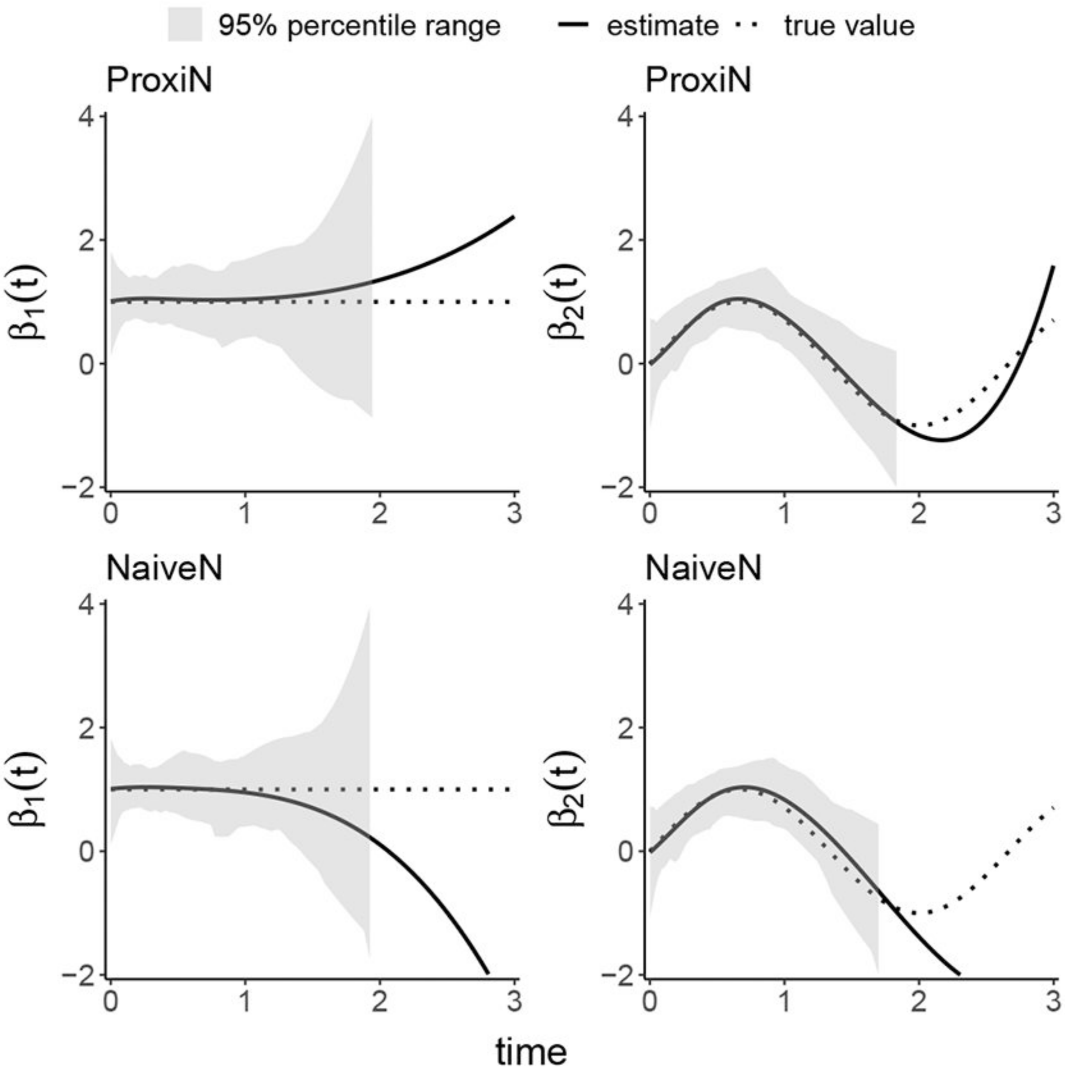


Fig. 2.

Mean of estimates $\hat{\beta}_1(t)$ and $\hat{\beta}_2(t)$ at each time t using the proximal Newton (ProxiN) and naive Newton (NaiveN) methods, with a 95% percentile range (2.5th and 97.5th percentiles as lower and upper limits). In each scenario, 100 data replicates were generated with sample size equal to 1,000. A fixed number of $K = 5$ knots were used for model fitting. True values were $\beta_1(t) = 1$ and $\beta_2(t) = \sin(3\pi t/4)$, with $\beta_3(t) = -1$, $\beta_4(t) = t^2 \exp(t/2)/9$, $\beta_5(t) = \exp(-1.5t)$.

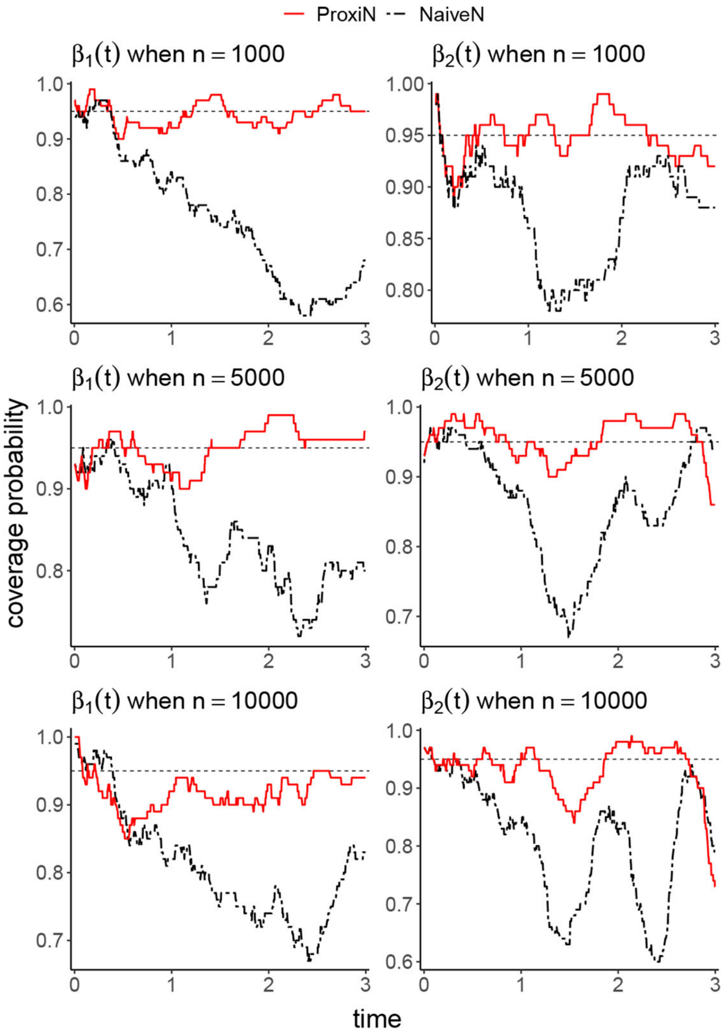


Fig. 3.

Coverage probability (CP) of estimates $\hat{\beta}_1(t)$ and $\hat{\beta}_2(t)$ at each time t using the proximal Newton (ProxiN) and naive Newton (NaiveN) methods, with a 95% confidence level. In each scenario, 100 data replicates were generated and a fixed number of $K = 5$ knots were used for model fitting. True values were $\beta_1(t) = 1$ and $\beta_2(t) = \sin(3\pi t/4)$.

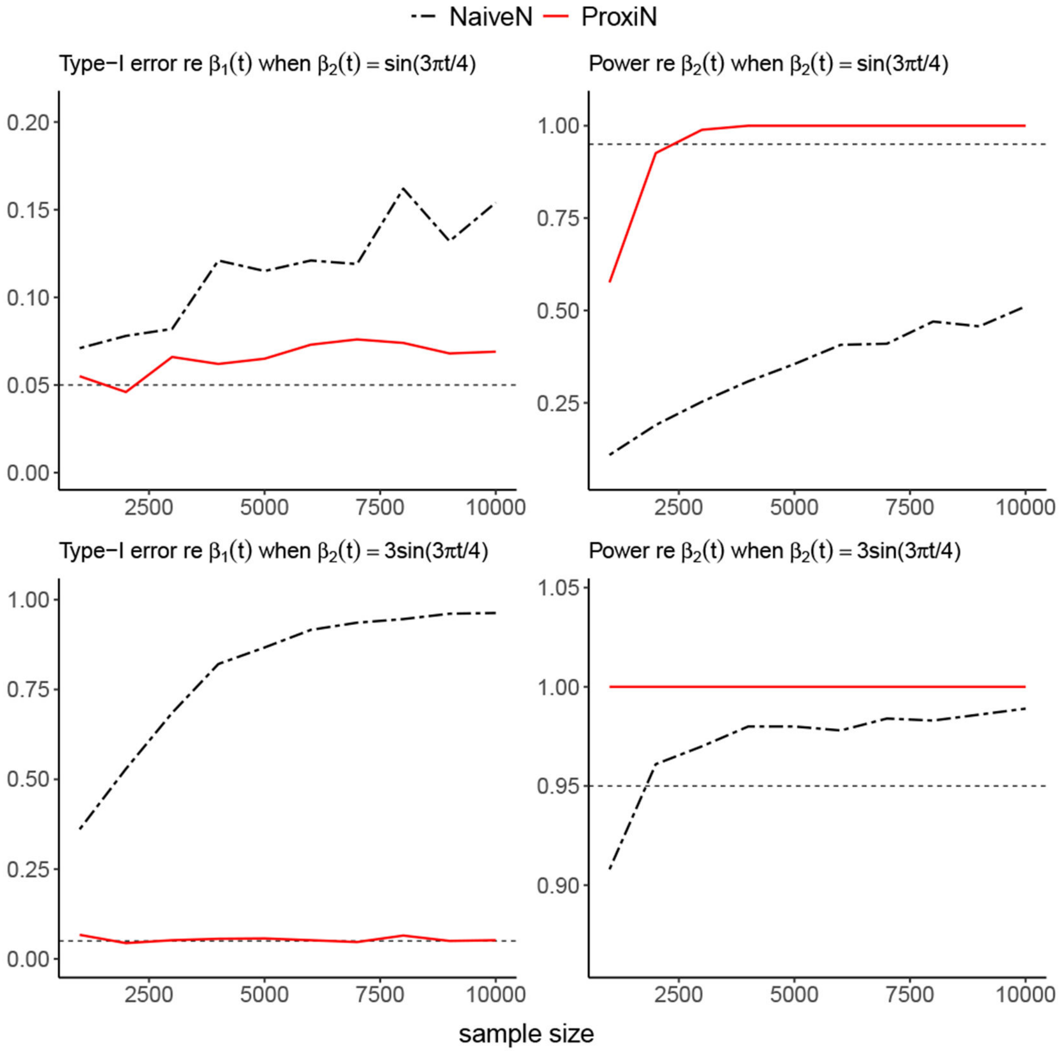


Fig. 4.

Type-I error rate and power regarding $\beta_1(t)$ and $\beta_2(t)$ using the proximal Newton (ProxiN) and naive Newton (NaiveN) methods with varying sample sizes. In each scenario, 1,000 data replicates were generated, and a fixed number of $K = 5$ knots were used for model fitting. In the first row, true values were $\beta_1(t) = 1$ and $\beta_2(t) = \sin(3\pi t/4)$, while in the second row, true values were $\beta_1(t) = 1$ and $\beta_2(t) = 3\sin(3\pi t/4)$.

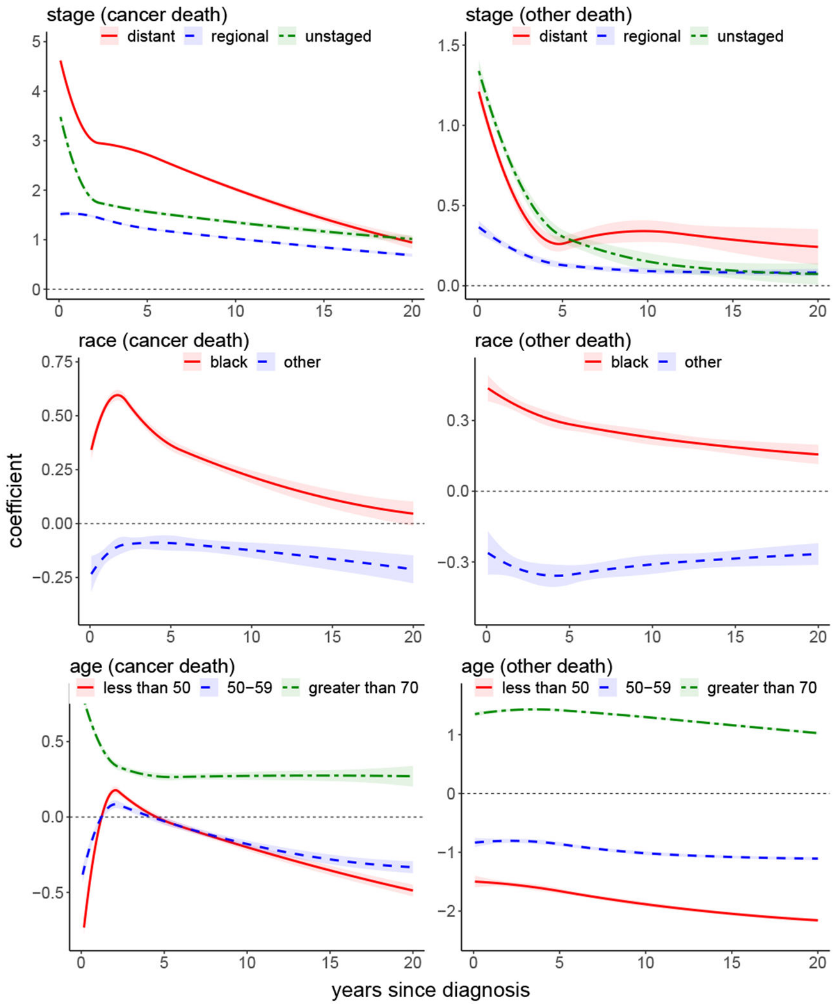


Fig. 5.

Estimates of the time-varying effects of tumor stage, race and age on death (due to cancer or other causes) as a function of time since diagnosis using the SEER breast cancer data. Quadratic B-splines were applied throughout the analysis with $K = 5$ knots. The ribbons in all panels represent 95% pointwise confidence intervals for the time-varying coefficients. At a 5% level, all effects on cancer death or other deaths were significantly time-dependent using the testing procedure in Section 4.

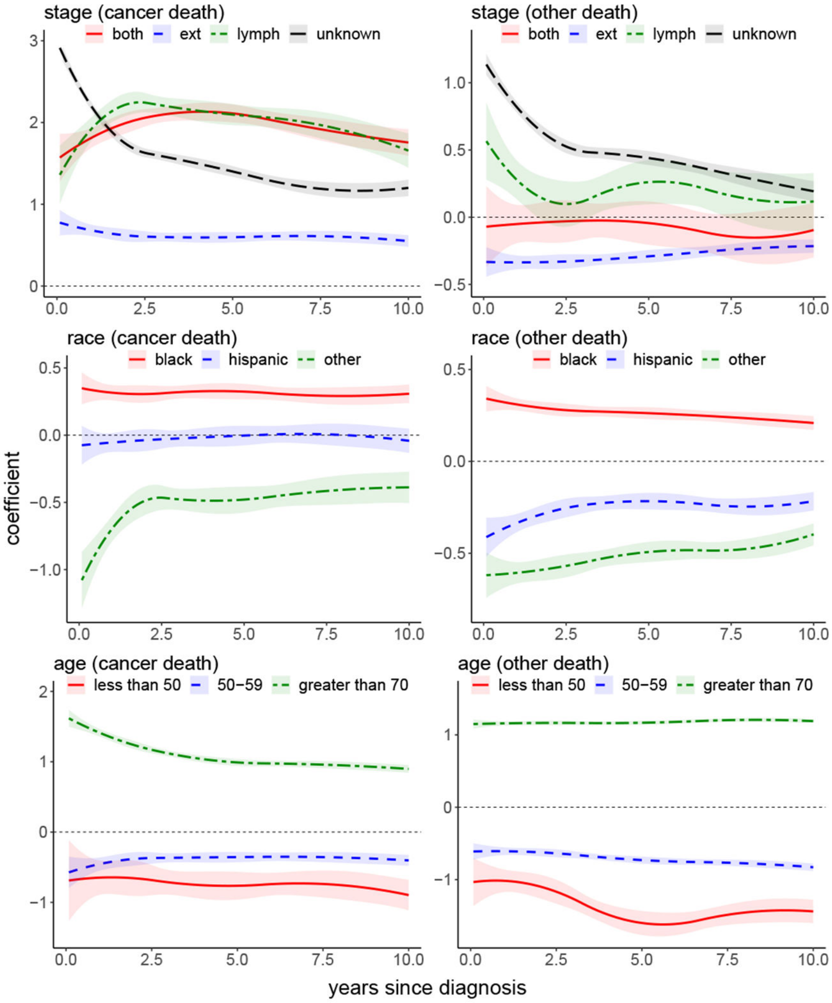


Fig. 6.

Estimates of the time-varying effects of tumor stage, race and age on death (due to cancer or other causes) as a function of time since diagnosis using the SEER prostate cancer data. Quadratic B-splines were applied throughout the analysis with $K = 5$ knots. The ribbons in all panels represent 95% pointwise confidence intervals for the time-varying coefficients. The four stages displayed in the legends are regional both by direct extension and lymph nodes (both), regional by direct extension (ext), regional by lymph nodes (lymph) and unknown. At a 5% level, significant time-varying effects on cancer death included age greater than 70, other races and the four stage effects. All effects on other deaths were significantly time-dependent except both and lymph.

Table 1

Integrated mean squared error (IMSE), average bias, and average variance of estimates $\hat{\beta}_1(t)$ and $\hat{\beta}_2(t)$ using the proximal Newton (ProxiN), naive Newton (NaiveN), and quasi-Newton (QuasiN) methods with varying sample sizes. In each scenario, 100 data replicates were generated, and a fixed number of $K = 5$ knots were used for model fitting. True values were $\beta_1(t) = 1$ and $\beta_2(t) = \sin(3\pi t/4)$.

method	size	IMSE	bias	variance
		Panel A: $\beta_1(t)$		
ProxiN	1000	3.60	0.24	3.55
	5000	0.25	0.02	0.25
	10000	0.15	0.04	0.15
NaiveN	1000	35.82	0.99	34.84
	5000	0.26	0.03	0.26
	10000	0.15	0.05	0.15
QuasiN	1000	6772.12	69.37	1960.34
	5000	4870.17	40.94	3194.03
	10000	3969.22	44.47	1991.35
		Panel B: $\beta_2(t)$		
ProxiN	1000	1.82	0.28	1.74
	5000	0.18	0.20	0.14
	10000	0.14	0.23	0.09
NaiveN	1000	20.94	1.41	18.95
	5000	0.25	0.23	0.20
	10000	0.13	0.20	0.09
QuasiN	1000	72892.15	237.68	16400.41
	5000	41906.34	106.80	30499.53
	10000	26924.89	107.92	15279.30

Table 2

Integrated mean squared error (IMSE), average bias, and average variance of estimates $\hat{\beta}_{11}(t)$ and $\hat{\beta}_{12}(t)$ (corresponding to the first cause of failure) using the proximal Newton (ProxiN), naive Newton (NaiveN), and quasi-Newton (QuasiN) methods with varying sample sizes. In each scenario, 100 data replicates were generated and a fixed number of $K = 5$ knots were used for model fitting. True values were $\beta_{11}(t) = 1$, $\beta_{12}(t) = \sin(3\pi t/4)$, $\beta_{13}(t) = -1$, $\beta_{14}(t) = -1$, and $\beta_{15}(t) = 1$.

method	size	IMSE	bias	variance
		Panel A: $\beta_{11}(t)$		
ProxiN	1000	2.41	0.22	2.36
	5000	0.61	0.08	0.60
	10000	0.45	0.08	0.44
NaiveN	1000	7.72	0.68	7.26
	5000	3.75	0.06	3.74
	10000	2.63	0.35	2.51
QuasiN	1000	2830.04	41.82	1081.30
	5000	3715.09	34.98	2491.78
	10000	1700.60	28.43	892.08
		Panel B: $\beta_{12}(t)$		
ProxiN	1000	2.47	0.22	2.42
	5000	1.02	0.25	0.96
	10000	0.71	0.17	0.68
NaiveN	1000	195.44	2.06	191.19
	5000	79.27	0.90	78.47
	10000	22.60	1.18	21.21
QuasiN	1000	111975.71	303.89	19627.88
	5000	61091.58	143.38	40532.74
	10000	17822.45	92.46	9274.30